# Non-Negative Matrix Factorization Meets Time-Inhomogeneous Markov Chains

**Ievgen Redko** NAME.SURNAME@UNIV-ST-ETIENNE.FR
**Marc Sebban**
**Amaury Habrard**
*Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d Optique Graduate School*
*Laboratoire Hubert Curien UMR 5516*
*F-42023, Saint-Etienne, France*

## Abstract

Non-negative matrix factorization (NMF) [22] is a popular unsupervised learning approach that allows to obtain part-based representations of non-negative data samples and provide soft-clustering assignments for them. The optimization problem behind NMF is often solved using multiplicative update rules (MUR) that are known to exhibit several flaws related to their convergence and the uniqueness of the obtained solutions. In this paper, we provide a novel theoretical analysis of this optimization procedure by showing its equivalence to a time inhomogeneous Markov chain. This equivalence allows us to (1) derive sufficient conditions required for convergence to a non-negative solution regardless the initialization to take place and (2) to characterize the speed of this convergence. In general, we argue that the established results are negative and lead to an incentive of solving NMF with optimization strategies other than MUR.

## 1. Introduction

Non-negative matrix factorization (NMF) [22] is a popular learning method that is widely applied in many real-world applications such as times series analysis [26], clustering [31], topic modeling [24], recommender systems [2] and music analysis [9] due to its capacity of providing meaningful non-negative part-based data representations. Despite its widespread use, NMF represents a non-convex optimization problem with several major drawbacks. First, the factorization obtained by NMF is not unique in general so that one may obtain an alternative matrix decomposition for the same data matrix. This issue was studied in several theoretical contributions showing how one can ensure a uniqueness of the factorization through data preprocessing [5, 11, 18, 21], by imposing priors on the obtained matrices or by adding suitable regularization terms to the objective function [17, 19]. Second, several studies showed that multiplicative update rules (MUR) introduced in [22] for optimizing the NMF objective function may fail to converge to a local minimum [12] or even to a stationary point [23]. Despite these findings, there were no theoretical studies that analyzed analytically the convergence properties of the original MUR. Instead, a common solution adopted by many authors to ensure the convergence was to replace them with a more computationally expensive projected gradient and non-negative least-squares approaches [8].

In this paper, we provide several negative results for MUR within a standard NMF model. First, we prove that MUR is equivalent to a finite space time inhomogeneous Markov chain and show that this equivalence is not bijective in general. This latter presents a generalization of a traditional finite space Markov chain where transition matrices describing the probability of moving from one state to another are allowed to vary over time. We argue that it explains the convergence of MUR to distinct solutions for the same data matrix. Second, we derive sufficient conditions required for MUR to admit the same non-negative convergence point for any initialization and show that satisfying them requires solving a very difficult algebraic problem. Finally, we characterize the speed of the convergence and show that it depends on the spectral properties of the matrices involved in the optimization procedure. To the best of our knowledge, our results are the first of their kind both in terms of the used approach and the guarantees that they provide.

## 2. Preliminary knowledge

**Non-negative matrix factorization** A standard NMF [22] is represented as the following optimization problem:

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

Multiplicative update rules (MUR) used to optimize $J(\mathbf{W},\mathbf{H})$ were first introduced in [22] and can be summarized by the following iterative procedure:

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

where for the sake of simplicity we omit the superscripts $(i)$ for $\mathbf{H}$ and $\mathbf{W}$ when they are fixed in (2) and $\circ$ and $\div$ denote entrywise multiplication (also called Hadamard product) and division, respectively. MUR given in (2) are guaranteed to not increase the objective function in [22] and, due to their simplicity, are widely used in the NMF community [20, 27, 28]. As mentioned in [8], the popularity of NMF with MUR remains quite high despite several empirical results showing that the sequence $\{\mathbf{W}^{(i)},\mathbf{H}^{(i)}\}_{i=0}^{\infty}$ generated by (2) may fail to converge to a local minimum [12] or even a stationary point [23].

**Markov chains** A Markov chain [25] is defined by a countable set of states $S = \{s_1, s_2, \ldots, s_c\}$ and a conditional probability distribution (CPD) $P(S_t|S_{t-1})$ representing the probability of transitioning to a state $S_t \in S$ given the previous state $S_{t-1} \in S$. This CPD is summarized in the form of a non-negative row stochastic transition matrix $\{\mathbf{P}(S_i,S_j)\}_{i,j=1}^{c} = P(S_j|S_i)$ that can be used to generate a sequence of stochastic vectors $\{\mathbf{x}_j\}_{j=1}^{\infty}$ ( $\sum_{i=1}^{c}\mathbf{x}_j^i = 1, \forall j\ \mathbf{x}_j^i \geq 0$) starting from some $\mathbf{x}_0$ as follows:

$$\mathbf{x}_t = \mathbf{x}_0\mathbf{P}^t, \tag{3}$$

where $\mathbf{P}^t$ denotes the $t^{\text{th}}$ power of $\mathbf{P}$. We further give the following definitions.

**Definition 1** *A distribution $\pi$ supported on $S$ is called a* **stationary** *(also called* ***invariant****)* ***distribution*** *of a Markov chain with a transition matrix $\mathbf{P}$ if $\pi\mathbf{P} = \pi$.*

**Definition 2** *A Markov chain is* **irreducible** *if for all states $S_i, S_j \in S$, there exists a $t \geq 0$ such that $\mathbf{P}^t(S_i,S_j) > 0$.*

**Definition 3** *Let $\mathcal{T}(S_i) = \{t \geq 1 : \mathbf{P}^t(S_i, S_i) > 0\}$ be the set of all time steps for which a Markov chain can start and end in a state $S_i$. An irreducible Markov chain is* **aperiodic** *if the greatest common divisor (gcd) of $\mathcal{T}(S_i)$ is $1, \forall S_i \in S$.*

With these definitions, we now recall the convergence theorem for Markov chains [4, Theorem 1.9].

**Theorem 4** *If the Markov chain is irreducible and aperiodic, then there is a unique stationary distribution $\pi$. In this case, $\mathbf{P}^t$ converges to $\pi$ as follows, $\lim_{t\to\infty} \mathbf{P}^t = \mathbf{1}\pi$.*

The time homogeneous Markov chain introduced above can be also extended to a time inhomogeneous case where the transition matrix changes at each step. In this case, (3) becomes:

$$\mathbf{x}_t = \mathbf{x}_0 \prod_{i=1}^{t} \mathbf{P}_i, \text{ with } \mathbf{x}_{i+1} = \mathbf{x}_i \mathbf{P}_i. \tag{4}$$

In what follows, we denote by $\mathbf{P}^{(t,t')}$ the transition matrix between steps $t$ and $t'$ $(t < t')$, i.e., $\mathbf{P}^{(t,t')} = \prod_{j=t}^{t'-1} \mathbf{P}_j$. In this situation, each of the transition matrices $\mathbf{P}_i$ can be characterized individually in terms of reducibility and periodicity and has its own stationary distribution. The convergence of time inhomogeneous Markov chains was studied in several works [13, 29] where different assumptions regarding the properties of transition matrices were made. In this paper, we use the result from [29, Theorem 3.3] that links the convergence of inhomogeneous Markov chains to the spectra of the transition matrices and the general convergence theorem of [13] given below.

**Theorem 5** *Let $\{\mathbf{P}_i\}_{i=1}^{\infty}$ be a sequence of Markov transition matrices on $S$ admitting $\pi$ as an invariant distribution. For each $i$, let $\sigma_j(\mathbf{P}_i), j = 0, \ldots, |S| - 1$, be the singular values of $\mathbf{P}_i$ ordered in the decreasing order. Then, we have*

$$1. \lim_{t\to\infty} \prod_{i=1}^{t} \mathbf{P}_i(l,\cdot) - \pi = 0, \quad 2. \| \prod_{i=1}^{t} \mathbf{P}_i(l,\cdot) - \pi \|_2 \leq (\pi(\mathbf{x}) - 1)^{\frac{1}{2}} \prod_{i=1}^{t} \sigma_1(\mathbf{P}_i)$$

*where $\prod_{i=1}^{t} \mathbf{P}_i(l,\cdot)$ denotes the $l^{th}$ line of the product of matrices $\mathbf{P}_i$.*

When different transition matrices are not required to have the same stationary distribution, we can use the following more general result from [13].

**Theorem 6** *Let $\{\mathbf{P}_i\}_{i=1}^{\infty}$ be a sequence of Markov transition matrices on $S$ admitting for all $i$, $\pi_i$ as an invariant distribution. Assume that $\sum_{i=1}^{\infty} \|\pi_i - \pi_{i+1}\| < \infty$ and that there exists $n_0 < n_1 < n_2 \ldots$, such that $\sum_{k=1}^{\infty} (1 - \delta(\mathbf{P}^{(n_k,n_{k+1})}) = \infty$ with $\delta(\mathbf{P}) = \sup_{i,j\in S} \|\mathbf{P}(i,\cdot) - \mathbf{P}(j,\cdot)\|$. Then, for $\pi^* = \lim_{i\to\infty} \pi_i$ the following holds*

$$\forall t, \lim_{t'\to\infty} \sup_{l\in S} \left\| \mathbf{P}^{(t,t')}(l,\cdot) - \pi^* \right\| = 0.$$

Markov chains, both time homogeneous and inhomogeneous, have found their application in a wide variety of scientific fields including the modelling of complex processes in computer science (e.g., information retrieval and speech recognition), statistics and physics and thus present a topic of ongoing interest for research community.

We now proceed to the presentation of our main results.

## 3. Key results

Our plan of attack for analyzing NMF with MUR is to prove that the sequence of matrices that they generate can be equivalently obtained by moving along at least one time inhomogeneous Markov chain. This equivalence is further used to apply several results from the Markov chains' theory to obtain the convergence guarantees as well as the conditions when MUR converge to the same solution regardless the initialization.

The similarity between MUR given in Equation (2) and Equation (4) characterizing a Markov chain naturally raises a question on whether one can be shown to be equivalent to the other. In order to answer to this question, we first introduce the definition of the Soules matrices and Soules basis matrices [30] used in the proof of our equivalence result.

**Definition 7** *An orthogonal matrix $\mathbf{U}$ with non-negative first column is called a Soules basis matrix and $\mathbf{P}$ is called a Soules matrix if for every diagonal matrix $\mathbf{D} = diag(\lambda_1, \lambda_2, \dots, \lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$, the matrix $\mathbf{P} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ is non-negative.*

The construction of Soules basis matrices $\mathbf{U}$ was presented in [30] in order to solve the non-negative inverse eigenvalue problem (NIEP) [6] that consists in finding a matrix $\mathbf{P}$ with the desired spectrum given by $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$. We further call two matrices $\mathbf{P}$ and $\mathbf{S}$ similar if there exists an invertible matrix $\mathbf{D}$ such that $\mathbf{S} = \mathbf{D}^{-1}\mathbf{P}\mathbf{D}$ and note that similar matrices have the same list of eigenvalues and that their eigenvectors are related through $\mathbf{D}$. As the stationary distribution of the Markov chain is given by the dominant left eigenvector of the transition matrix, the similarity relationship allows us to use similar matrices as transition matrices interchangeably. We now formalize the link between MUR and Markov chains through the following theorem[1].

**Theorem 8** *Consider the NMF problem given in (1) and MUR updates given in (2). Let $\{\mathbf{P}_i\}_{i=0}^{\infty}$ be Soules matrices with eigenvalues given by vec $\left(\mathbf{X}\mathbf{H}^T/\mathbf{W}^{(i)}\mathbf{H}\mathbf{H}^T\right)$ and let $\mathbf{S}_i$ be a row stochastic matrix similar to $\mathbf{P}_i, \forall i$. Then, matrices $\{\mathbf{W}^{(i)}\}_{i=0}^{\infty}$ can be (up to a scaling factor) generated by a time inhomogeneous Markov chain with a finite state space of cardinality $mk$ and transition matrices $\{\mathbf{S}_i\}_{i=0}^{\infty}$.*

Theorem 8 establishes that the factors generated using MUR can be equivalently obtained by a time inhomogeneous Markov chain with a finite state space of cardinality $mk$. The states of this Markov chain correspond to the product vec$(\mathbf{W}^{(i)})\mathbf{U}^T$ implying that the elements of matrix $\mathbf{W}_i$ at each iteration can be retrieved via a simple multiplication by matrix $\mathbf{U}$ and reshaping of the obtained vector to the desired size. Note that there may exist more than one Markov chain with such properties implying that the proved equivalence is not bijective in general. This statement follows from the results on the NIEP that we use in a proof in order to construct matrices $\{\mathbf{P}_i\}_{i=1}^{\infty}$. Indeed, it is easy to see that the solution of the NIEP may not be unique, once it exists, since there are $mk$ given eigenvalues with respect to $\frac{m^2k^2}{2}$ unknown variables constituting each matrix $\mathbf{P}_i$. Furthermore, matrices $\{\mathbf{P}_i\}_{i=1}^{\infty}$ with the required spectrum can be potentially obtained using any other Soules basis matrix $\mathbf{U}' \neq \mathbf{U}$. As we argue above, constructing different such matrices may potentially lead to different Markov chains with transition matrices having different spectral properties (see Supplementary for an example)). Theorem 8 allows us to analyze MUR as a time inhomogeneous Markov chain so that we can now establish the conditions that one has to fulfill in

[1] We provide all proofs in the Supplementary material and analyze here only the update rules for matrix $\mathbf{W}$ as the same reasoning applies to matrix $\mathbf{H}$ as well.

order to expect MUR to converge to the same solution regardless the initialization where the same solution is unique up to a permutation and rescaling.

**Theorem 9** *Under the assumptions of Theorem 8, assume that $\|\mathbf{U}\| \leq M$ for some $M < +\infty$ and that $\forall i,\ \mathbf{S}_i$ is irreducible, aperiodic and satisfies the strong ergodicity conditions of Theorem 6. Let $P^{(i)}$ be the space of all Soules matrices with eigenvalues given by vec $\left(\mathbf{XH}^T/\mathbf{W}^{(i)}\mathbf{HH}^T\right)$, i.e.,*

$$P^{(i)} = \{\mathbf{P} \in \mathbb{R}_+^{mk\times mk} : \mathbf{P} = \mathbf{U_P}\mathit{diag}\left(\mathit{vec}\left(\frac{\mathbf{XH}^T}{\mathbf{W}^{(i)}\mathbf{HH}^T}\right)\right)\mathbf{U}_\mathbf{P}^T \text{ for some } \mathbf{U_P}\}.$$

*Then, MUR converge to the same non-negative solution regardless the initial initialization when one of the following conditions are verified:*

1. $\forall i, |P^{(i)}| = 1$;
2. $\forall i$ *and* $\forall \mathbf{P}, \mathbf{P}' \in P^{(i)}$, *transition matrices* $\mathbf{S}, \mathbf{S}'$ *similar to* $\mathbf{P}, \mathbf{P}'$ *have the same stationary distribution.*

The second assumption is very difficult to satisfy in practice for two reasons. First, restricting the space of all diagonalizable non-negative matrices to only one element for a given set of eigenvalues cannot be achieved without introducing structural constraints on the matrix $\mathbf{P}_i$. Indeed, to the best of our knowledge, the only variation of NIEP problem that was proved to admit a unique solution is that related to realizing a set of eigenvalues with an anti-bidiagonal or tridiagonal Jacobi matrices [16, Theorem 1]. This, in some sense, is in line with those algorithmic contributions on NMF where uniqueness of the factorization is achieved by enforcing sparsity, minimum polytope volume or orthogonality constraints on factor matrices (see [11, Section 1.2] for more details).

In practice, however, we are often interested in understanding on what particular properties of the data sample or on what initialization the speed of the convergence established in Theorem 9 depends. To this end, we provide below a corollary that quantifies the speed of convergence of MUR in a data-dependent way.

**Corollary 10** *Let $\{\mathbf{S}^{(i)}\}_{i=1}^{\infty}$ be as in Theorem 8 where for each $i$, we let $\sigma_1(\mathbf{S}_i)$ to be the second largest singular value of $\mathbf{S}_i$. If $\|\mathbf{U}\| \leq M$ for some $0 < M < +\infty$ and $\exists \pi^* : \forall i,\ \pi^*\mathbf{S}_i = \pi^*$ then, $\forall j \in [1, \dots, mk]$, we have*

$$\|\prod_{i=1}^{n} \mathbf{S}_i(j, \cdot) - \pi^*\|_2 \leq (\pi^*(j) - 1)^{\frac{1}{2}} \prod_{i=1}^{n} \sigma_1(\mathbf{S}_i).$$

In a nutshell, this result reveals a surprising dependence of the convergence of MUR on the product of the second largest singular values of matrices $\mathbf{S}_i$ (and to that of $\mathbf{P}_i$ due to the similarity). It is worth noting that the dependence of the convergence rate of an optimization scheme on the second largest eigenvalue of an involved quantity was also established, for instance, for genetic [10] and Google PageRank algorithms [15] but, to the best of our knowledge, no such results were proved for NMF problem before (see Supplementary material for an illustration on several datasets). We conclude this section by noting that the established equivalence between MUR and NMF in this case provides us with the first result of its kind, as the convergence rate of MUR was only studied previously in a very restrictive setting of supervised factorization in [1]. This, in its turn, shows the versatility and the complementarity of our approach to analyzing the NMF problem with this particular optimization scheme.

## References

[1] Roland Badeau, Nancy Bertin, and Emmanuel Vincent. Stability analysis of multiplicative update algorithms for non-negative matrix factorization. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[2] Yang Bao, Hui Fang, and Jie Zhang. Topicmf: Simultaneously exploiting ratings and reviews for recommendation. In *AAAI*, pages 2–8, 2014.

[3] C. Boutsidis and E. Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recogn.*, 41(4):1350–1362, 2008. ISSN 0031-3203.

[4] Joseph T. Chang. Stochastic Processes. https://iid.yale.edu/sites/default/files/files/chang-notes.pdf, 2007. Online; accessed November 4, 2020.

[5] David Donoho and Victoria Stodden. When does NMF give a correct decomposition into parts? In *NIPS*, pages 1141–1148, 2004.

[6] Patricia D Egleston, Terry D Lenker, and Sivaram K Narayan. The nonnegative inverse eigenvalue problem. *Linear Algebra and its Applications*, 379:475 – 490, 2004.

[7] L. Elsner, R. Nabben, and M. Neumann. On single and double soules matrices. *Linear Algebra Appl.*, 271:323–343, 1998.

[8] Igor Fedorov, Alican Nalci, Ritwik Giri, Bhaskar D. Rao, Truong Q. Nguyen, and Harinath Garudadri. A unified framework for sparse non-negative least squares using multiplicative updates and the NMF problem. *Signal Processing*, 146:79–91, 2018.

[9] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. NMF with the itakura-saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[10] Steven T. Garren and Richard L. Smith. Estimating the second largest eigenvalue of a markov transition matrix. *Bernoulli*, 6(2):215–242, 2000.

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

[12] Edward F. Gonzalez and Yin Zhang. Accelerating the lee-seung algorithm for NMF. *Technical report*, Rice University, 2005.

[13] J. Hajnal and M. S. Bartlett. The ergodic properties of non-homogeneous finite Markov chains. *Mathematical Proceedings of the Cambridge Philosophical Society*, 52(1):67–77, 1956.

[14] D.J. Hartfiel and J. W. Spellmann. Diagonal similarity of irreducible matrices to row stochastic matrices. *Pacific J. Math.*, 40(1):97–99, 1972.

[15] Taher H. Haveliwala, Sepandar D. Kamvar, and Ar D. Kamvar. The second eigenvalue of the google matrix, 2003.

[16] Olga Holtz. The inverse eigenvalue problem for symmetric anti-bidiagonal matrices. *Linear Algebra and its Applications*, 408:268–274, 2005.

[17] Patrik O. Hoyer and Peter Dayan. NMF with sparseness constraints. *J. Mach. Learn. Res.*, 5: 1457–1469, 2004.

[18] Kejun Huang, Student Member, Nicholas D. Sidiropoulos, and Ananthram Swami. NMF revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Trans. Signal Processing*, pages 211–224, 2014.

[19] Hyunsoo Kim and Haesun Park. Sparse NMF via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[20] Minje Kim and Paris Smaragdis. Mixtures of local dictionaries for unsupervised speech enhancement. *IEEE Signal Process. Lett.*, 22(3):288–292, 2015.

[21] Hans Laurberg, Mads Græsbøll Christensen, Mark D. Plumbley, Lars Kai Hansen, and Søren Holdt Jensen. Theorems on positive data: On the uniqueness of NMF. *Comp. Int. and Neurosc.*, 2008.

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

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

[24] Minnan Luo, Feiping Nie, Xiaojun Chang, Yi Yang, Alexander G. Hauptmann, and Qinghua Zheng. Probabilistic NMF and its robust extensions for topic modeling. In *AAAI*, pages 2308–2314, 2017.

[25] A. Markov. Extension of the Limit Theorems of Probability Theory to a Sum of Variables Connected in a Chain. In *Dynamic Probabilistic Systems (Volume I: Markov Models)*, pages 552–577. John Wiley & Sons, Inc., 1971.

[26] Jiali Mei, Yohann de Castro, Yannig Goude, and Georges Hébrail. NMF for time series recovery from a few temporal aggregates. In *ICML*, volume 70, pages 2382–2390, 2017.

[27] Lopamudra Mukherjee, Sathya N. Ravi, Vamsi K. Ithapu, Tyler Holmes, and Vikas Singh. An NMF perspective on binary hashing. In *ICCV*, pages 4184–4192, 2015.

[28] Jonathan Le Roux, John R. Hershey, and Felix Weninger. Deep NMF for speech separation. In *ICASSP*, pages 66–70, 2015.

[29] L. Saloff-Coste and J. Zúñiga. Convergence of some time inhomogeneous Markov chains via spectral techniques. *Stochastic Processes and their Applications*, 117(8):961–979, 2007.

[30] G.W. Soules. Constructing symmetric nonnegative matrices. *Linear and Multilinear Algebra*, 13:241–251, 1983.

[31] De Wang, Feiping Nie, and Heng Huang. Fast robust NMF for large-scale human action data clustering. In *IJCAI*, pages 2104–2110, 2016.

## Supplementary material

### Proof of Theorem 4

**Proof** Let us first consider Equation (2) and rewrite it using the relationship between the Hadamard product and the vectorization operation as follows:

$$\begin{aligned}
\mathbf{W}^{(i+1)} &= \mathbf{W}^{(i)} \circ \frac{\mathbf{X}\mathbf{H}^T}{\mathbf{W}^{(i)}\mathbf{H}\mathbf{H}^T} \implies \\
\text{vec}\left(\mathbf{W}^{(i+1)}\right) &= \text{vec}\left(\mathbf{W}^{(i)}\right) \circ \text{vec}\left(\frac{\mathbf{X}\mathbf{H}^T}{\mathbf{W}^{(i)}\mathbf{H}\mathbf{H}^T}\right) \implies \\
\text{vec}\left(\mathbf{W}^{(i+1)}\right) &= \text{vec}\left(\mathbf{W}^{(i)}\right) \text{diag}\left(\text{vec}\left(\frac{\mathbf{X}\mathbf{H}^T}{\mathbf{W}^{(i)}\mathbf{H}\mathbf{H}^T}\right)\right),
\end{aligned} \tag{5}$$

where diag$(\mathbf{x})$ stands for a diagonal matrix whose diagonal elements are given by the elements of vector $\mathbf{x}$ and the row vector vec $(\mathbf{A})$ denotes the vectorization of matrix $\mathbf{A}$.

Let us now sort the values of vec $\left(\frac{\mathbf{X}\mathbf{H}^T}{\mathbf{W}^{(i)}\mathbf{H}\mathbf{H}^T}\right)$ using a sorting operator $\mathfrak{s} : \mathbb{R}^{mk} \to \mathbb{R}^{mk}$ so that

$$\mathfrak{s}_i\left(\text{vec}\left(\frac{\mathbf{X}\mathbf{H}^T}{\mathbf{W}^{(i)}\mathbf{H}\mathbf{H}^T}\right)\right) \geq \mathfrak{s}_{(i+1)}\left(\text{vec}\left(\frac{\mathbf{X}\mathbf{H}^T}{\mathbf{W}^{(i)}\mathbf{H}\mathbf{H}^T}\right)\right)$$

for all $i = 1, \ldots, mk-1$. As all elements of vec $\left(\frac{\mathbf{X}\mathbf{H}^T}{\mathbf{W}^{(i)}\mathbf{H}\mathbf{H}^T}\right)$ are non-negative by construction, we can build a Soules matrix $\mathbf{P}_i$ having the following eigendecomposition

$$\mathbf{P}_i = \mathbf{U}\text{diag}\left(\mathfrak{s}\left(\text{vec}\left(\frac{\mathbf{X}\mathbf{H}^T}{\mathbf{W}^{(i)}\mathbf{H}\mathbf{H}^T}\right)\right)\right)\mathbf{U}^T,$$

so that

$$\text{diag}\left(\mathfrak{s}\left(\text{vec}\left(\frac{\mathbf{X}\mathbf{H}^T}{\mathbf{W}^{(i)}\mathbf{H}\mathbf{H}^T}\right)\right)\right) = \mathbf{U}^T\mathbf{P}_i\mathbf{U},$$

where $\mathbf{U}$ is a Soules basis matrix. Note that we construct Soules matrices $\mathbf{P}_i$ using the same Soules basis matrix $\mathbf{U}$ for all $i$ as this latter can be used for any list of non-negative eigenvalues. Using the properties of the eigendecomposition, we can further reestablish the initial order of the values given in vec $\left(\mathbf{X}\mathbf{H}^T/\mathbf{W}^{(i)}\mathbf{H}\mathbf{H}^T\right)$ by applying the inverse of $\mathfrak{s}$ to $\mathfrak{s}\left(\text{vec}\left(\mathbf{X}\mathbf{H}^T\mathbf{W}^{(i)}\mathbf{H}\mathbf{H}^T\right)\right)$ and by permuting the eigenvectors of $\mathbf{U}$ and $\mathbf{U}^T$ accordingly. In what follows, we denote by $\mathbf{P}_i$ a symmetric matrix constructed in this way for an unsorted list of eigenvalues. We can now rewrite (5) in the following form

$$\text{vec}\left(\mathbf{W}^{(i+1)}\right) = \text{vec}\left(\mathbf{W}^{(i)}\right)\mathbf{U}^T\mathbf{P}_i\mathbf{U}$$

implying

$$\text{vec}\left(\mathbf{W}^{(i+1)}\right)\mathbf{U}^T = \text{vec}\left(\mathbf{W}^{(i)}\right)\mathbf{U}^T\mathbf{P}_i. \tag{6}$$

Finally, denoting vec$(\mathbf{W}^{(i)})\mathbf{U}^T$ by $\mu_i$ allows us to express Equation (6) as $\mu_{i+1} = \mu_i\mathbf{P}_i$ which in its turn implies

$$\mu_n = \mu_0 \prod_{i=1}^{n} \mathbf{P}_i.$$

At this point we note that having $\forall i, \mathbf{U}_i = \mathbf{U}$ is not a simplification but a mandatory condition to fulfill as otherwise $\mu_{i+1} = \mu_i \mathbf{P}_i$ would imply $\mathbf{U}_i = \mathbf{U}_{i+1} = \mathbf{I}$ and in this case matrix $\mathbf{P}_i$ would be diagonal which presents little interest for our purpose. Without loss of generality, we now assume that vector $\mu_0$ lies in the probability simplex and that each $\mathbf{P}_i$ is normalized to be row stochastic via a similarity transformation. Note that while the effect of scaling by a constant $\sum \mu_0$ is negligible, the normalization of $\mathbf{P}_i$ requires some extra care. To this end, we use the result from [14] showing that each non-negative matrix is similar to a row-stochastic matrix up to a constant factor $r$, i.e., $\exists$ a diagonal matrix $\mathbf{D} : \mathbf{D}\mathbf{P}_i\mathbf{D}^{-1} = r\mathbf{S}_i$. As before, we omit the constant factor $r$ and note that the similarity relation ensures that $\mathbf{P}_i$ and $\mathbf{S}_i$ have the same eigenvalues and that their eigenvectors are tied through the diagonal scaling $\mathbf{D}$. The obtained construction is then a time inhomogeneous Markov chain characterized by transition matrices $\{\mathbf{S}_i\}_{i=1}^{\infty}$. This completes the proof. ■

### Example of Soules matrices leading to different transition matrices

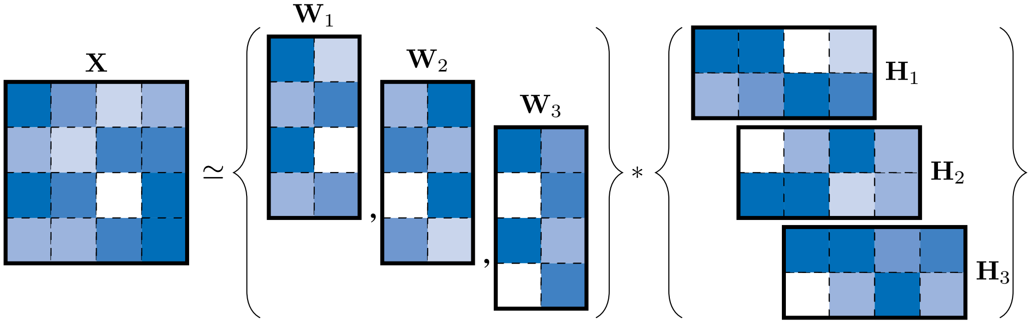

Figure 1: Illustration of 3 different solutions obtained using MUR with random initializations of $\mathbf{W}$ and $\mathbf{H}$. Here different degrees of gray correspond to numbers between 0 (white) and 5 (dark blue). The true numerical values for all matrices and the code to reproduce the experiment are given as part of the Supplementary material.

Let us consider the factorization of $\mathbf{X}$ from Figure 1 and show two different Soules basis matrices $\mathbf{U}$ and $\mathbf{U}'$ that can be build at the first iteration of MUR leading to two different transition matrices $\mathbf{P}$ and $\mathbf{P}'$. To this end, the first matrix given in Figure 2 (upper row, left) represents the Soules basis matrix constructed using the original approach of [30], while the second one (Figure 2 (upper row, middle left)) is build using a rooted binary tree splitting method proposed in [7]. We note that the two matrices have a very different structure and lead to different transition matrices $\mathbf{P}$ and $\mathbf{P}'$ as shown in Figure 2 (upper row, middle right) and (upper row, right). Moreover, the dominant left eigenvectors of these transition matrices (calculated after normalizing them to be row stochastic) are clearly different as shown in Figure 2 (bottom row). While with more iterations the two may eventually converge to the same stationary distribution, this example shows that it may as well not be the case in general.

### Proof of Theorem 5

**Proof** From Theorem 4, we know that the state space of a Markov chain generating NMF factors obtained with MUR is given by a sequence $\{\text{vec}(\mathbf{W}^{(i)})\mathbf{U}^T\}_{i=1}^{\infty}$. This latter represents a product of

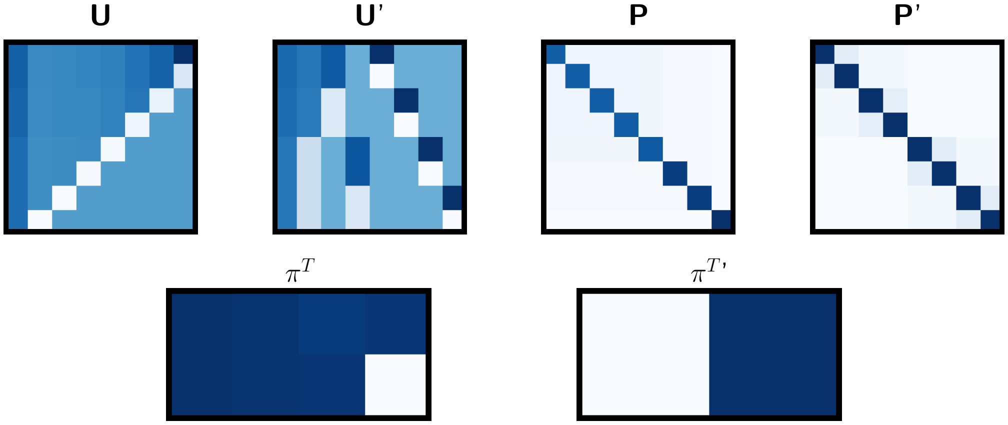

Figure 2: **(Top row)** Illustration of two Soules basis matrices $\mathbf{U}$, $\mathbf{U}'$ and their corresponding Soules matrices $\mathbf{P}$, $\mathbf{P}'$; **(Bottom row)** Stationary distributions $\pi$ and $\pi'$ (transposed for the sake of visibility) calculated as dominant left eigenvectors of $\mathbf{P}$, $\mathbf{P}'$ normalized to be row stochastic, respectively. The code to reproduce this figure is given in the Supplementary material.

a sequence multiplied by a constant matrix $\mathbf{U}^T$ so that its convergence implies the convergence of the sequence of interest $\{\mathbf{W}^{(i)}\}_{i=1}^{\infty}$ when $\mathbf{U}^T$ is bounded by some positive constant $M > 0$. Once verified, this condition allows us to analyze the convergence of the corresponding Markov chain and be sure that it implies the convergence of $\{\mathbf{W}^{(i)}\}_{i=1}^{\infty}$ as well. As for the convergence to the same solution for different initializations, we analyze two different cases given below.

**Case 1.** If $\forall i, |P^{(i)}| = 1$ then $\exists!$ Soules matrix $\mathbf{P}_i$ with eigenvalues given by vec $\left(\mathbf{XH}^T/\mathbf{W}^{(i)}\mathbf{HH}^T\right)$ that can be used to construct a similar row-stochastic matrix $\mathbf{S}_i$. This implies that there exists exactly one time-inhomogeneous Markov chain that generates factors identical to those obtained by MUR. In this case, the solution of the NIEP problem used to construct matrices $\mathbf{P}_i$ is also unique and we only need to ensure that all matrices $\mathbf{S}_i$ similar to $\mathbf{P}_i$ satisfy the conditions of Theorem 3 and its respective hypotheses to obtain the desired convergence guarantee. It should be noted that, in general, one can always take an arbitrary permutation matrix $\mathbf{B}$ to obtain a rearranged matrix $\mathbf{BPB}^T$ having the same eigenvalues as $\mathbf{P}$. However, here we require the existence of a unique such matrix for a sorted list of eigenvalues as usually done in NIEP problems.

**Case 2.** Using the property of the transition matrix of the Markov chain, we note that $\forall i, \pi_i\mathbf{S}_i = \pi_i$ so that $\pi_i$ is the left eigenvector corresponding to the largest eigenvalue, i.e., the maximum over all elements of vector vec $\left(\mathbf{XH}^T/\mathbf{W}^{(i)}\mathbf{HH}^T\right)$. If for any two distinct matrices $\mathbf{P}, \mathbf{P}' \in P^{(i)}$, transition matrices $\mathbf{S}, \mathbf{S}'$ similar to $\mathbf{P}, \mathbf{P}'$ are such that

$$\pi\mathbf{S} = \pi,\ \pi'\mathbf{S}' = \pi' \text{ and } \pi = \pi' = \pi_i$$

then any matrix $\mathbf{P} \in P^{(i)}$ can be picked at iteration $i$ to construct the transition matrix $\mathbf{S}_i$ of the desired Markov chain. Enforcing the conditions from Theorem 3 on these matrices gives the final result. ■

### Proof of Corollary 6

**Proof** The proof follows from Theorem 5 and Theorem 8. ■

### Experimental study

Apart from providing important insights regarding the behaviour of NMF with MUR, our theoretical results also suggest a practical way of comparing different initialization strategies proposed in the literature on NMF problem and their impact on the convergence speed based on Theorem 10. To this end, we consider a popular NNDSVD initialization method proposed in [3] and the pre-processing proposed in [11] that are both known for improving the convergence speed. We compare these methods to a baseline given by the random initialization of the NMF factors in two different settings where in one setting we factorize a matrix that provably admits a unique factorization, while in the other we factorize a randomly generated matrix constructed as a mixture of 5 Gaussian distributions representing the clusters. The goal of the comparison is two-fold: first, we want to verify whether the result provided in Theorem 10 is confirmed by our observations regarding the speed of convergence of different methods in practice; second, we want to assess the correlation between the quality of the factorization when measured by the reconstruction error for different initialization techniques as well as their ability to reach the global optimum when the corresponding NMF problem admits this latter.

**Unique factorization** In this setting, we consider a data matrix constructed using the following factors $\mathbf{H}$ and $\mathbf{W}$

$$\mathbf{H} = \begin{pmatrix} \alpha & 1 & 1 & \alpha & 0 & 0 \\ 1 & \alpha & 0 & 0 & \alpha & 1 \\ 0 & 0 & \alpha & 1 & 1 & \alpha \end{pmatrix}, \quad \mathbf{W} = \mathbf{H}^T$$

for some $\alpha \in (0,1)$. The data matrix $\mathbf{X} = \mathbf{WH}$ constructed in such way was shown to have a unique non-negative factorization for $k = 3$ in [21] when $\alpha = \{0.1, 0.3\}$. Thus, we set $\alpha = 0.3$ and run the considered baselines on $\mathbf{X}$ in order to see whether the different initialization techniques allow to recover the optimal factorization or whether they tend to sacrifice quality for speed by leading to a worse solution in fewer iterations.

**Mixture of Gaussian distributions** In this setting, we run the NMF with MUR on a data matrix having 20000 and 200 instances[2] for NNDSVD and Gillis' pre-processing, respectively and vary the dimensions $d$ of this latter from 10 to 100. The data is generated as a mixture of 5 isotropic Gaussian distributions centered in the hypercube defined over the interval $[10, 20]$ with variance $\mathbf{I}_d$ along all dimensions. Note that in this case, contrary to the setting considered above, the data matrix is not guaranteed to admit a unique factorization and thus random initialization can lead to potentially very different obtained factors.

**Results** The results of this comparison presenting the evolution of the product of second largest singular values of the transition matrices obtained at each iteration and the corresponding reconstruction error for both cases are given in Figures 3 and 4. From these figures, we observe that both non-random initialization techniques have a faster convergence rate in all cases considered as confirmed by both the product of their second largest singular values of the transition matrix and by the obtained reconstruction error that tends to stop decreasing earlier than in case of random initialization. This latter, however, appears to have a very different behaviour depending on the existence of a unique factorization and the pre-processing used to improve the convergence. Indeed, we see that for a data matrix admitting a unique factorization both pre-processing techniques

[2] We restrict our study to a matrix of a smaller size in the second case due to a prohibitively high computational complexity of Gillis' method scaling as $\mathcal{O}(n^{4.5})$.

converge quickly but to a solution of a lower quality that has a high reconstruction error. This is rather surprising as random initialization in this case recovers the optimal solution leading to an almost perfect factorization. Such a behaviour can be explained by the fact that both pre-processing techniques tend to provide a starting point close to a local minimum so that the algorithm struggles then to escape it and converge to a higher quality global minimum. As for the second scenario, we see that both methods once again converge much quicker than the random initialization with Gillis' pre-processing once again leading to a higher reconstruction error. In this case, such a behaviour is explained by a high sparsity of the factors obtained via Gillis' pre-processing and was also observed in the original paper [11, Table 2]. To summarize, we conclude by saying that the established link between the Markov chains and MUR scheme can be efficiently used in practice in order to analyze and compare different initialization strategies as in all cases it reflects correctly the convergence rate of each of them. As for the reconstruction error, we note that the effect of pre-processing leading to a unique factorization does not necessarily imply that this latter recovers the optimal solution but merely a solution to which one can converge reasonably quickly without much variance among the different runs.

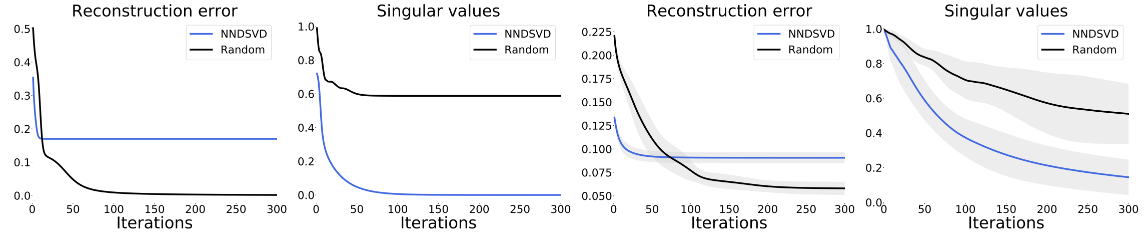
Reconstruction error
Singular values
Reconstruction error
Singular values

Figure 3: 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 $d \in \{10, \ldots, 100\}$.

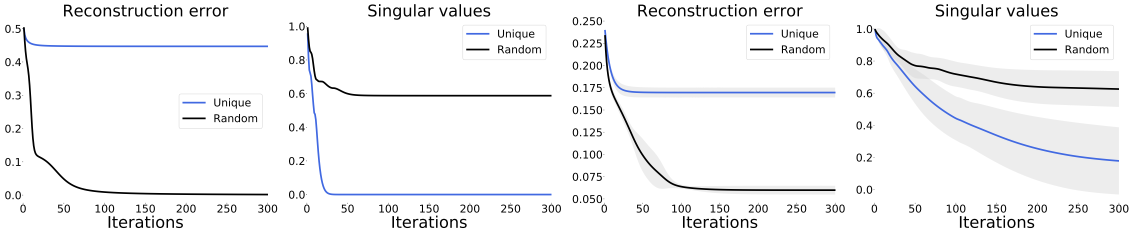
Reconstruction error
Singular values
Reconstruction error
Singular values

Figure 4: 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.