Online nonnegative CP tensor factorization for Markovian data

Christopher Strohmeier C.STROHMEIER@MATH.UCLA.EDU
Hanbaek Lyu HLYU@MATH.UCLA.EDU
Deanna Needell DEANNA@MATH.UCLA.EDU
University of California, Los Angeles

Abstract

We introduce a novel online nonnegative tensor factorization (NTF) algorithm that learns a CANDECOMP/PARAFAC (CP) basis from a given stream of tensor-valued data under general constraints. In particular, using nonnegativity constraints, the learned CP modes also give localized dictionary atoms that respect the tensor structure in multi-model data. On the theoretical side, we prove that our algorithm converges to the set of stationary points of the objective function under the hypothesis that the sequence of data tensors have functional Markovian dependence. This assumption covers a wide range of application contexts including data streams generated by independent or MCMC sampling. On the application side, we demonstrate the efficiency of our online algorithm against standard offline algorithms on both synthetic and real-world tensor data, and also illustrate the advantage of being able to flexibly reshape multi-modal tensor data and learn CP-dictionary atoms for any desired groups of modes jointly through video data applications.

1. Introduction

In modern applications, there is often a critical need to analyze and understand data that is high-dimensional (many variables), large-scale (many samples), and multi-modal (many attributes). A *tensor* is a multi-way array that is a natural generalization of a matrix (which is itself a 2-mode tensor) and is suitable in representing multi-model data. As matrix factorization is for unimodal data, *tensor factorization* (TF) provides a powerful and versatile tool that can extract useful latent information out of multi-model data tensors. As a result, tensor factorization methods have witnessed increasing popularity and adoption in modern data science [39, 43, 46, 49, 50].

Besides being multi-modal, another unavoidable characteristic of modern data is their enormous volume and the rate at which new data are generated. *Online learning* algorithms permit incremental processing that overcomes the sample complexity bottleneck inherent to batch processing, which is especially important when storing the entire data set is cumbersome. Not only do online algorithms address capacity and accessibility, but they also have the ability to learn qualitatively different information than offline algorithms for data that admit such a “sequential" structure (see e.g. [27]). In the literature, many “online" variants of more classical “offline" algorithms have been extensively studied — nonnegative matrix factorization (NMF) [18, 29, 33], TF [12, 19, 44, 50, 51], and dictionary learning [3, 4, 21, 38]. Online *nonnegative* TF (NTF) algorithms can serve as valuable tools that can extract interpretable features from multi-modal data.

We roughly divide the literature on TF into two classes depending on *structured* or *unstructured* TF problems. The *structured TF problem* concerns recovering exact loading matrices of a tensor, where a structured tensor decomposition with loading matrices satisfying some incoherence or sparsity conditions is assumed. A number of works address this problem in the offline setting [2, 5, 30, 41, 42, 46, 47]. Recently, [39] addresses an online structured TF problem by reducing it to an online MF problem using sparsity constraints on all but one loading matrices.

On the other hand, the *unstructured TF problem* seeks to find n loading matrices of a given n-mode tensor possibly with some constraints (e.g., nonnegativity) but without any structural assumptions on the true decomposition. In this case convergence to a globally optimal solution cannot be expected, and global convergence to stationary points of the objective function is desired. For offline problems, global convergence to stationary points of the block coordinate descent method is known to hold under some regularity assumptions on the objective function [8, 16, 17]. The recent works [12, 19, 44, 50, 51] on online TF focus on computational considerations and do not provide convergence guarantee. For online NMF, almost sure convergence to stationary points of a stochastic majorization-minimization (SMM) algorithm under i.i.d. data assumption is well-known [33], which is recently extended to the Markovian case in [29]. Similar convergence for online TF is not known even under the i.i.d. assumption. The main difficulty of extending a similar approach to online TF is that the recursively constructed surrogate loss function is nonconvex and cannot be jointly minimized in all n loading matrices when $n \geq 2$.

Contribution. In this work, we develop a novel algorithm and theory for the problem of *online nonnegative CP decomposition* (or online NTF), where the goal is to progressively learn an 'average' nonnegative CP decomposition of a stream of tensor data. Namely, given a sequence of n-mode nonnegative tensors $(\mathbf{X}_t)_{t\geq 0}$, we seek to find a single set of nonnegative loading matrices $U_1,\dots,U_n$ such that they give an approximate nonnegative CP decomposition of each $\mathbf{X}_t$ up to suitable nonnegative linear combination. Our main result shows that our online algorithm produces a sequence of loading matrices that converge almost surely to the set of stationary points of the objective function. This result holds for arbitrary number of modes in the tensor data, arbitrary convex constraints in place of the nonnegativity constraint, data samples with Markovian dependence (including the i.i.d. setting), and with a sparsity regularization in linear coefficients.

2. Background and problem formulation

An n-mode tensor $\mathbf{X}$ of shape $I_1\times\dots,I_n$ is a map $(i_1,\dots,i_n)\mapsto \mathbf{X}(i_1,\dots,i_n)\in\mathbb{R}$ from the multi-index set $\{1,\dots,I_1\}\times\cdots\times\{1,\dots,I_n\}$ into the real line $\mathbb{R}$. Suppose we have N observed n-mode tensor-valued signals $\mathbf{X}_1,\dots,\mathbf{X}_N\in\mathbb{R}_{\geq 0}^{I_1\times\cdots\times I_n}$. Fix an integer $R\geq 1$ and consider the following approximate factorization problem: ($\times_{n+1}$ denotes mode $(n+1)$-product [20])

$$\begin{cases} [\mathbf{X}_1,\dots,\mathbf{X}_N] \approx \mathtt{Out}(U_1,\dots,U_n)\times_{n+1} H \qquad \in \mathbb{R}_{\geq 0}^{I_1\times\cdots\times I_n\times N}; \\ \mathtt{Out}(U_1,\dots,U_n) := \left[\bigotimes_{k=1}^{n} U_k(:,1), \bigotimes_{k=1}^{n} U_k(:,2),\dots,\bigotimes_{k=1}^{n} U_k(:,R)\right] \in \mathbb{R}_{\geq 0}^{I_1\times\cdots\times I_n\times R}, \end{cases} \tag{1}$$

where $U_k(:,j)$ denotes the j^{th} column of the $I_k\times R$ matrix U_k and $\otimes$ denotes the outer product. Here $U_1,\dots,U_n$ are called the *loading matrices*. Such an approximate factorization learns R *dictionary atoms* $\mathbf{D}_1,\dots,\mathbf{D}_R$ in $\mathtt{Out}(U_1,\dots,U_n)$ that together can approximate each observed signal $\mathbf{X}_j$ by using the nonnegative linear coefficients in the j^{th} column of the *code* $H\in\mathbb{R}_{\geq 0}^{R\times N}$. When we have a single observed tensor ($N=1$), by absorbing the coordinates of H into the loading matrices, (1) reduces to

$$\mathbf{X}\approx \sum \mathtt{Out}(U_1,\dots,U_n) := \sum_{i=1}^{R}\bigotimes_{k=1}^{n} U_k(:,i), \tag{2}$$

which is the nonnegative CANDECOMP/PARAFAC (CP) decomposition problem [43, 49]. On the other hand, if $n = 1$ so that $\mathbf{X}_j$ are vector-valued signals, then it reduces to the classical dictionary learning problem [13, 14, 23, 25, 36] as well as the *nonnegative matrix factorization* (NMF) problem, where the use of nonnegativity constraint is crucial in obtaining "parts-based" representation of the input signals [22]. For these reasons, we refer (1) as the *CP dictionary learning* (CPDL) problem. We call the $(I_1 \times \cdots \times I_n \times R)$-mode tensor $\texttt{Out}(U_1,\dots,U_n) = [\mathbf{D}_1,\dots,\mathbf{D}_R]$ a *CP-dictionary* and the matrix $H \in \mathbb{R}_{\geq 0}^{R\times N}$ the *code* of the dataset $\mathscr{X} = [\mathbf{X}_1,\dots,\mathbf{X}_N]$, respectively. Here we call the rank-1 tensors $\mathbf{D}_j$ the *atoms* of the CP-dictionary.

In this paper, we consider an *online* version of the CPDL problem we introduced above. Namely, given a stream of data tensors $(\mathbf{X}_t)_{t\geq 0}$, can we find a CP-dictionary such that *all* observed signals $\mathbf{X}_t$ can be approximated as a suitable nonnegative linear combination of CP-dictionary atoms? This online problem can be explicitly formulated by an expected risk minimization problem, as follows. Suppose we have a probability distribution π on the set of data tensors $\mathbb{R}_{\geq 0}^{I_1\times\cdots\times I_n}$. The *online CP-dictionary learning* problem is the following stochastic program

$$\operatorname*{arg\,min}_{U_1,\dots,U_n}\left(f(U_1,\dots,U_n) := \mathbb{E}_{\mathbf{X}\sim\pi}\left[\inf_{h\in\mathbb{R}_{\geq 0}^{R\times 1}} \|\mathbf{X} - \texttt{Out}(U_1,\dots,U_n)\times_{n+1} h\|_F^2 + \lambda\|h\|_1 \right]\right), \tag{3}$$

where the minimization is over all $[U_1,\dots,U_n] \in \mathbb{R}_{\geq 0}^{I_1\times R}\times\cdots\times\mathbb{R}_{\geq 0}^{I_n\times R}$, the *random* data tensor $\mathbf{X}$ is sampled from the distribution π and $\lambda \geq 0$ is a sparsity regularizer. For each realization of $\mathbf{X}$, the optimal choice of $h \in \mathbb{R}_{\geq 0}^{R\times 1}$ gives the nonnegative coefficients to combine the atoms in the CP-dictionary $\texttt{Out}(U_1,\dots,U_n)$.

3. Algorithm: Online CPDL

In this section, we give a high-level description of our main algorithm (Algorithm 1 in the appendix). For simplicity, we will consider only the case of $n = 3$ in this section. Our algorithm for the Online CPDL problem (3) takes the following form: Suppose we have learned n loading matrices $\mathscr{D}_{t-1} := [U_1^{(t-1)},\dots,U_n^{(t-1)}]$ from the sequence $\mathbf{X}_1,\dots,\mathbf{X}_{t-1}$ of data tensors. Then:

$$\begin{cases} h_t & \leftarrow \operatorname*{arg\,min}_{h\in\mathbb{R}_{\geq 0}^{R\times 1}} \left[\|\mathbf{X}_s - \texttt{Out}(\mathscr{D}_{t-1})\times_{n+1} h\|_F^2 + \lambda\|h\|_1\right] \\ \hat{f}_t(\mathscr{D}) & \leftarrow (1-w_t)\hat{f}_{t-1}(\mathscr{D}) + w_t\left(\|\mathbf{X}_t - \texttt{Out}(\mathscr{D})\times_{n+1} h_t\|_F^2 + \lambda\|h_t\|_1\right) \\ \text{For } i = 1,\dots,n: & \\ \quad U_i^{(t)} & \leftarrow \operatorname*{arg\,min}_{U\in\mathbb{R}_{\geq 0}^{I_i\times R},\, \|U-U_i^{(t-1)}\|_F\leq c' w_t} \hat{f}_t(U_1^{(t)},\dots,U_{i-1}^{(t)},U,U_{i+1}^{(t-1)},\dots,U_n^{(t-1)}), \end{cases} \tag{4}$$

where $c' > 0$ and $\lambda \geq 0$ are fixed constants. The recursively defined function $\hat{f}_t : \mathscr{D} = [U_1,\dots,U_n] \mapsto [0,\infty)$ is called the *surrogate loss function*, which is quadratic in each factor U_j but not jointly convex for $n \geq 2$. When the new tensor data $\mathbf{X}_t$ arrives, one computes the code h_t for $\mathbf{X}_t$ with respect to the tuple $\mathscr{D}_{t-1}$ of previous loading matrices and updates the surrogate loss function $\hat{f}_t$, and then *sequentially* minimizes it to find updated loading matrices under shrinking search radius $c' w_t$. Our algorithm (4) uses the general scheme of stochastic majorization-minimization (SMM) [32], and it can be shown to reduce to the classical online NMF algorithm in [33] when $n = 1$ and c' is a large constant. The use of cyclic block coordinate descent and search radius restriction in the minimization step of $\hat{f}_t$ are two crucial ingredients in handling the multi-modal case $n \geq 2$. In Appendix A, we state an implementation of (4) that processes $b \geq 1$ tensor-valued signals at once (minibatch extension) and that avoids computing the full surrogate reconstruction error function $\hat{f}_t$ by carefully updating 'aggregate tensors' of fixed sizes.

4. Statement of main results

Here we state our main convergence result concerning the Online CPDL problem (3) and algorithm (4). We first lay out all technical assumptions required for our convergence results to hold.

(A1) *The observed data tensors* $\mathbf{X}_t$ *are given by* $\mathbf{X}_t = \varphi(Y_t)$, *where* Y_t *is an irreducible and aperiodic Markov chain defined on a finite state space* Ω *and* $\varphi : \Omega \to \mathbb{R}^{I_1 \times \cdots \times I_n}$ *is a bounded function.*

(A2) *For each* $1 \leq j \leq n$, *the* j*th loading matrices for CP-dictionaries* $\mathcal{D}_t$ *are constrained to a compact and convex subset* $\mathcal{C}_j^{\text{dict}} \subseteq \mathbb{R}_{\geq 0}^{I_j \times R}$.

(A3) *The expected loss function* f *defined in* (3) *and the loss function inside the expectation in* (3) *are continuously differentiable and have Lipschitz gradient.*

Assumptions (A2)-(A3) are standard in the literature of online dictionary learning and [32–34]. It is also standard to assume that the sequence of signals are drawn from a distribution π in an independent fashion [31, 33]. However, this is not feasible when the signals have to be sampled from some complicated or unknown distribution, and one instead often uses Markov Chain Monte Carlo (MCMC) sampling algorithms (e.g., sampling from the posterior in Bayesian methods [48] or motif sampling from sparse graphs [26]). Our assumption on input signals in (A1) is general enough to handle such situations. Markovian extension of the classical online NMF algorithm developed in [29] has applications in dictionary learning, denoising, and edge inference problems for network data [28]. Note that (A1) implies the underlying Markov chain Y_t mixes exponentially fast (see [24]).

The main result in this paper, which is stated below in Theorem 1, states that the sequence $\mathcal{D}_t$ of CP-dictionaries produced by Algorithm 1 converges to the set of stationary points of the expected loss function f defined in (3). To the best of our knowledge, Theorem 1 is the first convergence guarantee for any online *constrained* dictionary learning algorithm for tensor-valued signals or as an online *unstructured* CP-factorization algorithm, which have not been available even under the classical i.i.d. assumption on input signals.

Theorem 1 *Suppose (A1)-(A3). Let* $(\mathcal{D}_t)_{t\geq 1}$ *be an output of algorithm* (4). *Then the distance (measured by element-wise Frobenius distance) between* $\mathcal{D}_t$ *and the set of stationary points of* f *over* $\mathcal{C}_1^{\text{dict}} \times \cdots \times \mathcal{C}_n^{\text{dict}}$ *converges to zero almost surely.*

The biggest difficulty in the convergence analysis is that the iterate $\mathcal{D}_t$ is *not* a local minimum of the surrogate loss function $\hat{f}_t$, which is only convex in each loading matrices not not jointly convex for $n \geq 2$. This causes one to lose essential properties of SMM algorithms that are granted in the matrix factorization ($n = 1$) setting. One of our innovations is to use the restricted search radius for updating loading matrices to ensure stability of estimates, which plays a similar role of diminishing step sizes in gradient descent algorithms [9]. We show the effect of this additional constraint vanishes in the limit. Another difficulty is that, under the Markovian dependence in (A1), the theory of quasi-martingales [15, 40], which is a key ingredient in convergence analysis under i.i.d input in [1, 32, 33], cannot be used. Instead, we use the recently developed technique of "conditioning on distant past" in [29] to overcome this issue of dependence in data samples.

5. Experimental validation and application

In Figure 1, we show the efficiency of our proposed algorithm in (4) against two most popular algorithms for nonnegative CP decomposition (also known as nonnegative tensor factorization) on tasks of Alternating Least Squares (ALS) and Multiplicative Update (MU) (see [43]).

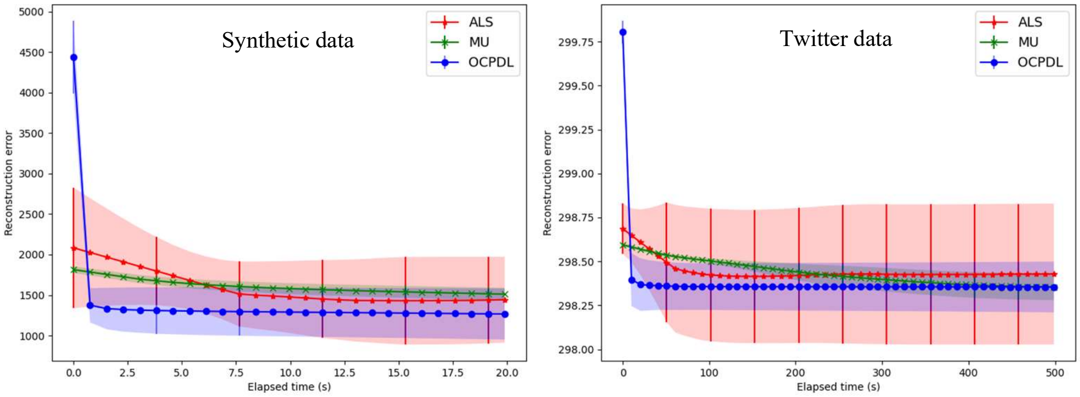

Figure 1: Comparison of performance by run time of Online CPDL (this work) for the nonnegative tensor factorization problem for Alternating Least Squares (ALS) and Multiplicative Update (MU). For the synthetic and Twitter tensor data [37] of shape (100, 100, 5000) and (90, 5000, 1000), respectively, we apply each algorithm ten times to find nonnegative loading matrices U_1, U_2, U_3 of $R = 5$ columns. The average reconstruction error with 1 standard deviation are shown by the solid lines and shaded regions of respective colors.

In Figure 2, we demonstrate our method on video data of brain activity across a mouse cortex, and how our Online CPDL learns dictionaries for the spatial and temporal activation patterns simultaneously. The original video is due to Barson et al. [6]. In order to learn periodic activation patterns occurring within 2-seconds, we applied algorithm (4) with $w_t = 1/t$, $\lambda = 2$, and $c' = 10^5$ for 200 random samples of 2-second-long clips, reshaped into 2-way tensors where one mode combines space and color modes and the other mode is time. Due to the nonnegativity constraint, spatial activation atoms representing localized activation regions in the cortex are learned, while the darker ones represent the background brain shape without activation.

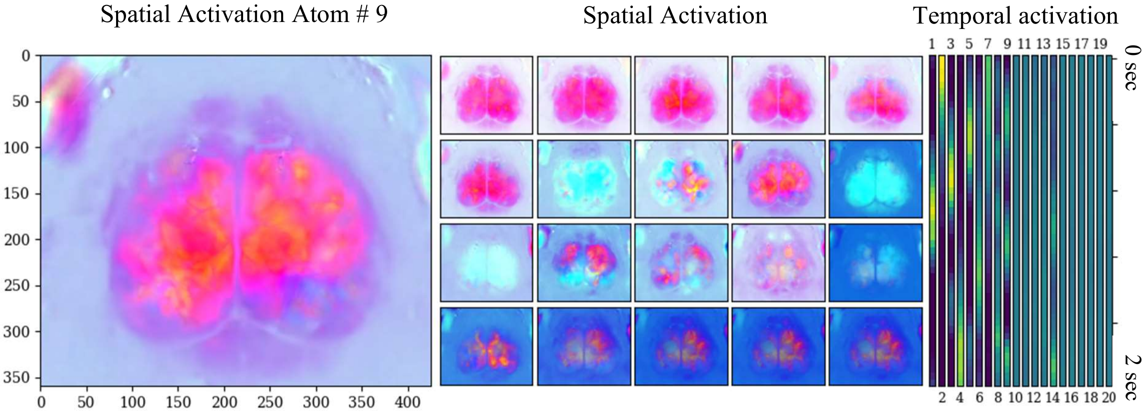

Figure 2: Learning 20 CP-dictionary patches from video frames on brain activity across the mouse cortex. Original tensor has shape (time, width, height, color) = (1501, 360, 426, 3) with total duration of 60 seconds. Blue= 0 and brighter color indicates larger values.

Acknowledgements

HL is partially supported by NSF DMS-2010035. DN is grateful to and was partially supported by NSF DMS-2011140 and NSF BIGDATA 1740325.

References

[1] Abhishek Agarwal, Jianhao Peng, and Olgica Milenkovic. Online convex dictionary learning. *arXiv preprint arXiv:1904.02580*, 2019.

[2] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Learning overcomplete latent variable models through tensor methods. In *Conference on Learning Theory*, pages 36–112, 2015.

[3] Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Conference on Learning Theory*, pages 779–806, 2014.

[4] Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. 2015.

[5] Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 143–151, 2015.

[6] Daniel Barson, Ali S Hamodi, Xilin Shen, Gyorgy Lur, R Todd Constable, Jessica A Cardin, Michael C Crair, and Michael J Higley. Simultaneous mesoscopic and two-photon imaging of neuronal activity in cortical circuits. *Nature methods*, 17(1):107–113, 2020.

[7] Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

[8] Dimitri P Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

[9] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[10] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[11] Sourav Chatterjee. A generalization of the lindeberg principle. *The Annals of Probability*, 34 (6):2061–2076, 2006.

[12] Yishuai Du, Yimin Zheng, Kuang-chih Lee, and Shandian Zhe. Probabilistic streaming tensor decomposition. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 99–108. IEEE, 2018.

[13] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

[14] Kjersti Engan, Sven Ole Aase, and John Hakon Husoy. Frame based signal compression using method of optimal directions (mod). In *ISCAS'99. Proceedings of the 1999 IEEE International Symposium on Circuits and Systems VLSI (Cat. No. 99CH36349)*, volume 4, pages 1–4. IEEE, 1999.

[15] Donald L Fisk. Quasi-martingales. *Transactions of the American Mathematical Society*, 120 (3):369–389, 1965.

[16] Luigi Grippo and Marco Sciandrone. On the convergence of the block nonlinear gauss–seidel method under convex constraints. *Operations research letters*, 26(3):127–136, 2000.

[17] Luigi Grippof and Marco Sciandrone. Globally convergent block-coordinate techniques for unconstrained optimization. *Optimization methods and software*, 10(4):587–637, 1999.

[18] Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan. Online nonnegative matrix factorization with robust stochastic approximation. *IEEE Transactions on Neural Networks and Learning Systems*, 23(7):1087–1099, 2012.

[19] Furong Huang, UN Niranjan, Mohammad Umar Hakeem, and Animashree Anandkumar. Online tensor methods for learning latent variable models. *The Journal of Machine Learning Research*, 16(1):2797–2835, 2015.

[20] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[21] Alec Koppel, Garrett Warnell, Ethan Stump, and Alejandro Ribeiro. D4l: Decentralized dynamic discriminative dictionary learning. *IEEE Transactions on Signal and Information Processing over Networks*, 3(4):728–743, 2017.

[22] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[23] Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on pattern analysis and machine intelligence*, 27(5):684–698, 2005.

[24] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[25] Michael S Lewicki and Terrence J Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000.

[26] H Lyu, F Memoli, and D Sivakoff. Sampling random graph homomorphisms and applications to network data analysis. *arXiv:1910.09483*, 2019.

[27] Hanbaek Lyu, Christopher Strohmeier, Georg Menz, and Deanna Needell. Applications of online nonnegative matrix factorization to image and time-series data. *Submitted. Preprint available upon request.*, 2019.

[28] Hanbaek Lyu, Yacoub Kureh, , Josh Vendrow, and Mason Porter. Learning low-rank latent mesoscale structures in networks. *In preparation*, 2020.

[29] Hanbaek Lyu, Deanna Needell, and Laura Balzano. Online matrix factorization for markovian data and applications to network dictionary learning. *Journal of Machine Learning Research 21 (To appear)*, 2020.

[30] Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 438–446. IEEE, 2016.

[31] Julien Mairal. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pages 783–791, 2013.

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

[33] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(Jan):19–60, 2010.

[34] Arthur Mensch, Julien Mairal, Bertrand Thirion, and Gaël Varoquaux. Stochastic subsampling for factorizing huge matrices. *IEEE Transactions on Signal Processing*, 66(1):113–128, 2017.

[35] Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5, 1998.

[36] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[37] Anand Rajaraman and Jeffrey David Ullman. *Mining of massive datasets.* Cambridge University Press, 2011.

[38] Sirisha Rambhatla, Xingguo Li, and Jarvis Haupt. Noodl: Provable online dictionary learning and sparse coding. *In 7th International Conference on Learning Representations, ICLR 2019*, 2019.

[39] Sirisha Rambhatla, Xingguo Li, and Jarvis Haupt. Provable online cp/parafac decomposition of a structured tensor via dictionary learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[40] K Murali Rao. Quasi-martingales. *Mathematica Scandinavica*, 24(1):79–92, 1969.

[41] Tselil Schramm and David Steurer. Fast and robust tensor decomposition with applications to dictionary learning. *arXiv preprint arXiv:1706.08672*, 2017.

[42] Vatsal Sharan and Gregory Valiant. Orthogonalized als: A theoretically principled tensor decomposition algorithm for practical use. *arXiv preprint arXiv:1703.01804*, 2017.

[43] Amnon Shashua and Tamir Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning*, pages 792–799. ACM, 2005.

[44] Shaden Smith, Kejun Huang, Nicholas D Sidiropoulos, and George Karypis. Streaming tensor factorization for infinite data sources. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 81–89. SIAM, 2018.

[45] Christopher Strohmeier, Hanbaek Lyu, and Deanna Needell. Online nonnegative tensor factorization and cp-dictionary learning for markovian data. *arXiv preprint arXiv:2009.07612*, 2020.

[46] Will Wei Sun, Junwei Lu, Han Liu, and Guang Cheng. Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 3(79):899–916, 2017.

[47] Gongguo Tang and Parikshit Shah. Guaranteed tensor decomposition: A moment approach. In *International Conference on Machine Learning*, pages 1491–1500, 2015.

[48] Don Van Ravenzwaaij, Pete Cassey, and Scott D Brown. A simple introduction to markov chain monte–carlo sampling. *Psychonomic bulletin & review*, 25(1):143–154, 2018.

[49] Stefanos Zafeiriou. Algorithms for nonnegative tensor factorization. In *Tensors in Image Processing and Computer Vision*, pages 105–124. Springer, 2009.

[50] Shuo Zhou, Nguyen Xuan Vinh, James Bailey, Yunzhe Jia, and Ian Davidson. Accelerating online cp decompositions for higher order tensors. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1375–1384, 2016.

[51] Shuo Zhou, Sarah Erfani, and James Bailey. Online cp decomposition for sparse tensors. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1458–1463. IEEE, 2018.

Appendix A. Bounded memory implementation of Online CPDL algorithm

In this section, we introduce an alternative implementation of algorithm (4) that uses bounded memory that is independent of the number T of minibatches of data tensors being processed. This will be done by replacing the step for computing the surrogate loss function $\hat{f}_t$ with computing two 'aggregate tensors' based on our deterministic analysis in Proposition 2. The total amount of information fed in to the algorithm is $O(T\prod_{i=1}^{n} I_n)$ and $T \to \infty$, whereas Algorithm 1 stores only $O(R\prod_{i=1}^{n} I_n)$ (recall that R is the number of dictionary atoms to be learned and T is the number of minibatches of data tensors that have arrived). This is an inherent memory efficiency of online algorithms against non-online algorithms (see, e.g., [33]).

For given n-mode tensors $\mathbf{A}$ and $\mathbf{B}$, denote by $\mathbf{A} \odot \mathbf{B}$ and $\mathbf{A} \otimes_{kr} \mathbf{B}$ their Hadamard (pointwise) product and Katri-Rao product, respectively. When B is a matrix, for each $1 \leq j \leq n$, we also denote their mode-j product by $\mathbf{A} \times_j B$.

Algorithm 1 Online CP-Dictionary Learning (Bounded Memory Implementation)

```
Input: $(\mathscr{X}_t)_{1\leq t\leq T}$ (minibatches of data tensors in $\mathbb{R}_{\geq 0}^{I_1\times\cdots\times I_n\times b}$); $[U_1^{(0)},\dots,U_n^{(0)}] \in \mathbb{R}_{\geq 0}^{I_1\times R} \times\cdots\times \mathbb{R}_{\geq 0}^{I_n\times R}$ (initial loading matrices)
Constraints: $\mathscr{C}_i^{\mathrm{dict}} \subseteq \mathbb{R}^{I_i\times R}$, $1\leq j\leq n$, $\mathscr{C}^{\mathrm{code}} \subseteq \mathbb{R}^{R\times b}$ (e.g., nonnegativity constraints)
Parameters: $R\in\mathbb{N}$ (# of dictionary atoms); $\lambda\geq 0$ ($\ell_1$-regularizer); $c'>0$ (search radius constant);
  Initialize aggregate tensors $A_0\in\mathbb{R}^{R\times R}$, $B_0\in\mathbb{R}^{I_1\times\cdots\times I_n\times R}$;
  For $t=1,\dots,T$ do:
    Coding: Compute the optimal code matrix
        $H_t \leftarrow \operatorname*{arg\,min}_{H\in\mathscr{C}^{\mathrm{code}}\subseteq\mathbb{R}^{R\times b}} \left[ \left\| \mathbf{X}_s - \mathtt{Out}(U_1^{(t-1)},\dots,U_n^{(t-1)}) \times_{n+1} H \right\|_F^2 + \lambda\|H\|_1 \right];$   (5)
    Update aggregate tensors:
        $A_t \leftarrow (1-w_t)A_{t-1} + w_t H_t H_t^T \in \mathbb{R}^{R\times R};$   (6)
        $B_t \leftarrow (1-w_t)B_{t-1} + w_t(\mathscr{X}_t \times_{n+1} H_t^T) \in \mathbb{R}^{I_1\times\cdots\times I_n\times R};$
    Update dictionary:
      $[U_1,\dots,U_1] \leftarrow [U_1^{(t-1)},\dots,U_1^{(t-1)}]$;
      For $j=1,\dots,n$ do:
        $\overline{A}_{t;j}\in\mathbb{R}^{R\times R}$, $\overline{B}_{t;j}\in\mathbb{R}^{I_i\times R} \leftarrow$ Algorithm 2 with input $A_t, B_t, U_1,\dots,U_n, j$;
        $U_i' \leftarrow \operatorname{arg\,min}_{U\in\mathscr{C}_i,\, \|U-U_i\|_F\leq c'w_t} \left[ \mathrm{tr}(U\overline{A}_{t;j}U^T) - 2\mathrm{tr}(U\overline{B}_{t;j}^T) \right]$ ;
        $U_i \leftarrow U_i'$;
      End for
    $[U_1^{(t)},\dots,U_n^{(t)}] \leftarrow [U_1',\dots,U_n']$;
  End for
Return: $[U_1^{(T)},\dots,U_n^{(T)}] \in \mathscr{C}_1^{\mathrm{dict}}\times\cdots\times\mathscr{C}_n^{\mathrm{dict}}$;
```

Algorithm 2 Intermediate Aggregation

```
Input: $A \in \mathbb{R}^{R\times R}$, $B \in \mathbb{R}^{I_1\times\cdots\times I_n\times R}$, $[U_1,\ldots,U_n] \in \mathbb{R}^{I_1\times R}\times\ldots\times\mathbb{R}^{I_n\times R}$, $1 \le j \le n$
Do: (Notation: $\mathbf{A}\odot\mathbf{B}$ =Hadamard (pointwise) product; $\mathbf{A}\otimes_{kr}\mathbf{B}$ =Katri-Rao product; When $B$ is a matrix, $\mathbf{A}\times_j B$ =mode-$j$ product. (See [20])
    $\overline{A}_i = A \odot U_1^T U_1 \odot \ldots \odot U_{i-1}^T U_{i-1} \odot U_{i+1}^T U_{i+1} \odot \ldots \odot U_n^T U_n \in \mathbb{R}^{R\times R}$    (7)
    For $r = 1,\ldots,R$ do:
        $B(,r) :=$ mode-$(n+1)$ slice of $B$ at coordinate $r$
        $b_{i;r} := B(,r)\times_1 U_1(:,r)\times_2\cdots\times_{i-1}U_{i-1}(:,r)\times_{i+1}U_{i+1}(:,r)\times_{i+2}\cdots\times_n U_n(:,r) \in \mathbb{R}^{I_i}$
        $\overline{B}_{t;j} := I_i \times R$ matrix whose $r$th column is $b_{i;r}$
    End for
Return:
    $\overline{A}_i = \overline{A}_i(A, U_1,\ldots,U_{i-1},U_{i+1},\ldots,U_n)$
    $\overline{B}_i = \overline{B}_i(B, U_1,\ldots,U_{i-1},U_{i+1},\ldots,U_n)$
```

In order to see why Algorithm 1 is equivalent to (4), first, consider the following block optimization problem

$$\text{Upon arrival of } \mathscr{X}_t\text{:}\quad \begin{cases} H_t = \operatorname{arg\,min}_{H\in\mathscr{C}^{\text{code}}\subseteq\mathbb{R}^{R\times b}}\ \ell(\mathscr{X}_t,\mathscr{D}_{t-1},H) \\ A_t = (1-w_t)A_{t-1} + w_t H_t H_t^T \\ B_t = (1-w_t)B_{t-1} + w_t(\mathscr{X}_t \times_{n+1} H_t^T) \\ \mathscr{D}_t = \underset{\substack{\mathscr{D}=[U_1,\ldots,U_n]\in\mathscr{C}^{\text{dict}} \\ \|U_i - U_i^{(t-1)}\|_F \le c' w_t\ \forall i}}{\operatorname{arg\,min}}\ \hat{g}_t(\mathscr{D}) \end{cases}, \tag{8}$$

where for each $\mathscr{D} = [U_1,\ldots,U_n] \in \mathscr{C}^{\text{dict}}$ and $B_t^{(n+1)}$ denoting the mode-$(n-1)$ unfolding of B_t,

$$\hat{g}_t(\mathscr{D}) := \operatorname{tr}(A_t\,(U_n^T U_n \odot \ldots \odot U_1^T U_1)) - 2\operatorname{tr}\left(B_t^{(n+1)}(U_n \otimes_{kr} \ldots \otimes_{kr} U_1)^T\right). \tag{9}$$

The following proposition reformulates (4) into Algorithm 8.

Proposition 2 *The following holds:*

(i) *Let $\hat{f}_t$ be as in (4) and $\hat{g}_t$ as above. Then*

$$\hat{f}_t(U_1,\ldots,U_n) = \hat{g}_t(U_1,\ldots,U_n) + \sum_{s=1}^{t}\operatorname{tr}\left(\operatorname{MAT}(\mathscr{X}_s)\operatorname{MAT}(\mathscr{X}_s)^T\right) + \lambda\sum_{s=1}^{t}\|H_s\|_1, \tag{10}$$

(ii) *For each $1 \le j \le n$, we can rewrite $\hat{g}_t(\mathscr{D}) = \hat{g}_t(U_1,\ldots,U_n)$ in (9) as*

$$\hat{g}_t(U_1,\ldots,U_n) = \operatorname{tr}\left(U_i\overline{A}_{t;j}U_i^T\right) - 2\operatorname{tr}\left(U_i\overline{B}_{t;j}^T\right), \tag{11}$$

where $\overline{A}_{t;j} \in \mathbb{R}^{R\times R}$, $\overline{B}_{t;j} \in \mathbb{R}^{I_i\times R}$ are computed by Algorithm 2 with input $A_t, B_t, U_1,\ldots,U_n$, and j.

Proof See Subsection B.2. ■

Now We describe how Algorithm 1 is derived and why it is equivalent to Algorithm (4). By the time that the new data tensor $\mathscr{X}_t$ arrives, the algorithm have computed previous loading matrices $U_1^{(t-1)},\dots,U_n^{(t-1)}$ and two aggregate tensors $A_{t-1}\in\mathbb{R}^{R\times R}$ and $B_{t-1}\in\mathbb{R}^{I_1\times\cdots\times I_n\times b}$. Then one computes the code matrix $H_t\in\mathscr{C}^{\text{code}}\subseteq\mathbb{R}^{R\times b}$ by solving the convex optimization problem in (5), and then updates the aggregate tensors $A_t\leftarrow A_{t-1}$ and $B_t\leftarrow B_{t-1}$. In order to perform the block coordinate descent to update the loading matrices $U_i^{(t)}$ in line 12 of Algorithm (4), we appropriately recompute intermediate aggregate matrices $\overline{A}_i$ and $\overline{B}_i$ using Algorithm 2 so that we are correctly minimizing the surrogate loss function $\hat{f}_t$ in (4) marginally according to Proposition 2 **(ii)**.

Appendix B. Proof of Theorem 1

In this section, we provide a sketch the proof of Theorem 1. In [45], a more general version of Theorem 1 (replacing w_t in Algorithm 1 with a general weight w_t under certain condition) is shown with more detailed analysis. Throughout this section, we assume the code matrices H_t and loading matrices $U_i^{(t)}$ belong to convex and compact constraint sets $H_t\in\mathscr{C}^{\text{code}}\subseteq\mathbb{R}^{R\times b}$, $U_i^{(t)}\in\mathscr{C}^{\text{dict}}\subseteq\mathbb{R}^{I_i\times R}$ and denote $\mathscr{C}^{\text{dict}}=\mathscr{C}_1^{\text{dict}}\times\cdots\times\mathscr{C}_n^{\text{dict}}\subseteq\mathbb{R}^{I_1\times R}\times\cdots\times\mathbb{R}^{I_n\times R}$. For each $\mathscr{X}\in\mathbb{R}_{\geq 0}^{I_1\times\cdots\times I_n\times b}$, $\mathscr{D}=[U_1,\dots,U_n]\in\mathbb{R}^{I_1\times R}\times\cdots\times\mathbb{R}^{I_n\times R}$, $H\in\mathbb{R}^{R\times b}$, define

$$\ell(\mathscr{X},\mathscr{D},H):=\|\mathscr{X}-\texttt{Out}(\mathscr{D})\times_{n+1}H\|_F^2+\lambda\|H\|_1; \tag{12}$$

$$\ell(\mathscr{X},\mathscr{D}):=\inf_{H\in\mathscr{C}^{\text{code}}}\ell(\mathscr{X},\mathscr{D},H), \tag{13}$$

where $\lambda\geq 0$ is a sparsity regularizer. Also, we introduce the *empirical loss function* f_t associated with algorithm 4, which is defined recursively as

$$f_t(\mathscr{D})=(1-w_t)f_{t-1}(\mathscr{D})+w_t\,\ell(\mathscr{X},\mathscr{D}), \tag{14}$$

By comparing with the definition of the surrogate loss function $\hat{f}_t$ in (4), it is clear that $\hat{f}_t\geq f_t$ for all $t\geq 0$ given that we initialize $\hat{f}_0\geq f_0$. Also it is important to note that $\|f_t-f\|\to 0$ almost surely as $t\to\infty$ by Lemma 17, where f is the expected loss function (main objective function) in (3).

B.1. Key lemmas to the proof of Theorem 1.

In this subsection, we state the key lemmas we use to prove Theorem 1 and illustrate our contribution in techniques for convergence analysis. As we mentioned in Section 3, there is a significant amount of difficulty in convergence analysis in the multi-modal case $n\geq 2$, as the surrogate loss functions $\hat{f}_t$ computed by (4) are non-convex.

Next, we list the key properties of Stochastic Majorization-Minimization (SMM) scheme in the unimodal case $n=1$ that have been critically used in convergence analysis in related works [29, 32–34].

1 (Surrogate Optimality) $\mathscr{D}_t$ is a minimizer of $\hat{f}_t$ over $\mathscr{C}^{\text{dict}}$.

2 (Forward Monotonicity) $\hat{f}_t(\mathscr{D}_{t-1})\geq\hat{f}_t(\mathscr{D}_t)$.

3 (Backward Monotonicity) $\hat{f}_{t-1}(\mathscr{D}_{t-1})\leq\hat{f}_{t-1}(\mathscr{D}_t)$.

4 (Second-Order Growth Property) $\hat{f}_t(\mathscr{D}_{t-1}) - \hat{f}_t(\mathscr{D}_t) \geq c\|\mathscr{D}_t - \mathscr{D}_{t-1}\|_F^2$ for some constant $c > 0$.

5 (Stability of Estimates) $\|\mathscr{D}_t - \mathscr{D}_{t-1}\|_F = O(w_t)$.

6 (Stability of Errors) For $h_t := \hat{f}_t - f_t \geq 0$, $|h_t(\mathscr{D}_t) - h_{t-1}(\mathscr{D}_{t-1})| = O(w_t)$.

For $n = 1$, it is crucial that $\hat{f}_t$ is convex so that $\mathscr{D}_t$ is a minimizer of $\hat{f}_t$ in the convex constraint set $\mathscr{C}^{\text{dict}}$, as stated in **1**. From this the monotonicity properties **2** and **3** follow immediately. The second-order growth property **4** requires additional assumption that the surrogates $\hat{f}_t$ are strongly convex uniformly in t. Then **3** and **4** imply **5**, which then implies **6**. Lastly, **1** is also crucially used to conclude that every limit point of $(\mathscr{D}_t)_{t\geq 1}$ is a stationary point of f over $\mathscr{C}^{\text{dict}}$. Now in the multi-modal case $n \geq 2$, we do not have **1** so all of the implications mentioned above are not guaranteed. Hence the analysis in the multi-modal case requires a significant amount of technical innovation.

Now we state our key lemma that handles the non-convexity of the surrogate loss $\hat{f}_t$ in the general multi-modal case $n \geq 1$.

Lemma 3 (Key Lemma) *Assume (A2) and (A3). Let $(\mathscr{D}_t)_{t\geq 1}$ be an output of Algorithm 1. Denote $h_t := \hat{f}_t - f_t$. Then for all $t \geq 1$, the following hold:*

(i) *(Forward Monotonicity)* $\hat{f}_t(\mathscr{D}_{t-1}) \geq \hat{f}_t(\mathscr{D}_t)$;

(ii) *(Stability of Estimates)* $\|\mathscr{D}_t - \mathscr{D}_{t-1}\|_F = O(w_t)$;

(iii) *(Stability of Errors)* $|h_t(\mathscr{D}_t) - h_{t-1}(\mathscr{D}_{t-1})| = O(w_t)$.

(iv) *(Asymptotic Surrogate Stationarity) Further assume (A1), $\sum_{t=1}^{\infty} w_t = \infty$, and $\sum_{t=1}^{\infty} w_t^2\sqrt{t} < \infty$. Let $(t_k)_{k\geq 1}$ be any sequence such that $\mathscr{D}_{t_k}$ and $\hat{f}_{t_k}$ converges almost surely. Then $\mathscr{D}_\infty = \lim_{k\to\infty} \mathscr{D}_{t_k}$ is almost surely a stationary point of $\hat{f}_\infty = \lim_{k\to\infty} \hat{f}_{t_k}$ over $\mathscr{C}^{\text{dict}}$.*

We show Lemma 3 **(i)** using a monotonicity property of block coordinate descent and Lindeberg's replacement trick [11], which is often used in the probability literature (see the proof of Lemma 3). One of our key observations is that we can directly ensure the stability properties **5** and **6** (Lemma 3 **(ii)** and **(iii)**) by using a search radius restriction (see line 12 of Algorithm 4). In turn, we do not need the properties **3** and **4**. In particular, our analysis does not require strong convexity of the surrogate loss $\hat{f}_t$ in each loading matrices as opposed to the existing analysis (see, e.g., [33, Assumption **B**] and [32, Def. 2.1]). Lastly, our analysis requires that estimates $\mathscr{D}_t$ are only asymptotically stationary to the limiting surrogate loss function along convergent subsequences, as stated in Lemma 3 **(iv)**. The proof of this statement is nontrivial and requires a substantial work. Roughly speaking, we show that the effect of search radius restriction by $O(w_t)$ vanishes in the limit and the gradient $\nabla\hat{f}_\infty(\mathscr{D}_\infty)$ is in the normal cone of $\mathscr{C}^{\text{dict}}$ at $\mathscr{D}_\infty$.

The second technical challenge is to handle dependence in input signals, as stated in (A1). The theory of quasi-martingales [15, 40] is a key ingredient in convergence analysis under i.i.d input in [1, 32, 33]. However, dependent signals do not induce quasi-martingale since conditional on the information $\mathscr{F}_t$ at time t, the following signal $\mathscr{X}_{t+1}$ could be heavily biased. We use the recently developed technique in [29] to overcome this issue of data dependence. The key insight is to condition on "distant past" $\mathscr{F}_{t-N}$, not on the present $\mathscr{F}_t$, in order to allow the underlying Markov chain to mix close enough to the stationary distribution π for N iterations. This allows us to control positive variations of the one-step error of the algorithm using Markov chain mixing, as stated in Lemma 4.

Lemma 4 (Convergence of Positive Variation) *Let $(\mathscr{D}_t)_{t\ge 1}$ be an output of Algorithm 4. Suppose (A1) and (A2) holds.*

(i) *Let $(a_t)_{t\ge 0}$ be a sequence of non-decreasing non-negative integers such that $a_t \in o(t)$. Then there exists absolute constants $C_1, C_2, C_3 > 0$ such that for all sufficiently large $t \ge 0$,*

$$\begin{aligned}&\mathbb{E}\left[\left(\mathbb{E}\left[w_{t+1}\big(\ell(\mathscr{X}_{t+1},\mathscr{D}_t)-f_t(\mathscr{D}_t)\big)\,\middle|\,\mathscr{F}_{t-a_t}\right]\right)^+\right]\\ &\qquad\le C_1(t-a_t)^{-2}\sqrt{t}+C_2\,t^{-2}a_t+C_3\,w_t\sup_{\mathbf{y}\in\Omega}\|P^{a_t+1}(\mathbf{y},\cdot)-\pi\|_{TV}.\end{aligned}$$

(ii) $\displaystyle\sum_{t=0}^{\infty}\left(\mathbb{E}\left[\hat{f}_{t+1}(\mathscr{D}_{t+1})-\hat{f}_t(\mathscr{D}_t)\right]\right)^+ \le \sum_{t=0}^{\infty} w_{t+1}\left(\mathbb{E}\left[\big(\ell(\mathscr{X}_{t+1},\mathscr{D}_t)-f_t(\mathscr{D}_t)\big)\right]\right)^+ < \infty.$

B.2. Deterministic analysis

We first provide some deterministic analysis of our online algorithm (Algorithm 1), which are foundational to the forthcoming stochastic analysis. The first three results are original in this work and handle most of the difficulties unique to the tensor-valued signals.

We first derive Proposition 2.

Proof [**Proof of Proposition 2**] Let $\mathrm{MAT}(\mathscr{X}_s)=[\mathrm{vec}(\mathbf{X}_{s;1}),\dots,\mathrm{vec}(\mathbf{X}_{s;b})]\in\mathbb{R}^{(I_1\dots I_n)\times b}$ denote the matrix whose i^{th} column is the vectorization $\mathrm{vec}(\mathbf{X}_{s;j})$ of the tensor $\mathbf{X}_{s;j}\in\mathbb{R}^{I_1\times\cdots\times I_n}$. The first assertion follows easily from observing that, for each $[U_1,\dots,U_n]\in\mathscr{C}^{\text{dict}}$ and $H\in\mathbb{R}^{R\times b}$

$$\begin{aligned}&\|\mathscr{X}_s-\mathtt{Out}(U_1,\dots,U_n)\times_{n+1}H\|_F^2\\ &\quad=\|\mathrm{MAT}(\mathscr{X}_s)-(U_n\otimes_{kr}\dots\otimes_{kr}U_1)H\|_F^2\\ &\quad=\mathrm{tr}\left((U_n\otimes_{kr}\dots\otimes_{kr}U_1)HH^T(U_n\otimes_{kr}\dots\otimes_{kr}U_1)^T\right)\\ &\qquad\quad-2\mathrm{tr}\left(\mathrm{MAT}(\mathscr{X}_s)H^T(U_n\otimes_{kr}\dots\otimes_{kr}U_1)^T\right)+\mathrm{tr}\left(\mathrm{MAT}(\mathscr{X}_s)\,\mathrm{MAT}(\mathscr{X}_s)^T\right),\end{aligned}$$

and also noting that

$$\begin{aligned}&\mathrm{tr}\left((U_n\otimes_{kr}\dots\otimes_{kr}U_1)HH^T(U_n\otimes_{kr}\dots\otimes_{kr}U_1)^T\right)\\ &\qquad=\mathrm{tr}(HH^T(U_n\otimes_{kr}\dots\otimes_{kr}U_1)^T(U_n\otimes_{kr}\dots\otimes_{kr}U_1))\\ &\qquad=\mathrm{tr}(HH^T\,(U_n^TU_n\odot\dots\odot U_1^TU_1)).\end{aligned}$$

Indeed, this and the definition of A_t and B_t together with the linearity of trace show

$$\begin{aligned}\hat{f}_t(U_1,\dots,U_n)&=\mathrm{tr}(A_t\,(U_n^TU_n\odot\dots\odot U_1^TU_1))-2\mathrm{tr}\left(\widetilde{B}_t(U_n\otimes_{kr}\dots\otimes_{kr}U_1)^T\right)\\ &\qquad+\sum_{s=1}^{t}\mathrm{tr}\left(\mathrm{MAT}(\mathscr{X}_s)\,\mathrm{MAT}(\mathscr{X}_s)^T\right)+\lambda\sum_{s=1}^{t}\|H_s\|_1,\end{aligned}\tag{15}$$

where the matrix $\widetilde{B}_t\in\mathbb{R}^{(I_1\times\cdots\times I_n)\times b}$ is defined recursively by

$$\widetilde{B}_s=(1-w_t)\widetilde{B}_{s-1}+w_t\,\mathrm{MAT}(\mathscr{X}_s)H_s^T.$$

By an induction, one can show that $\widetilde{B}_t$ equals the mode-$(n+1)$ unfolding $B_t^{(n+1)}$ of B_t. This shows **(i)**.

For **(ii)**, first note that

$$\begin{aligned}&\mathrm{tr}(A\,(U_n^T U_n \odot \ldots \odot U_1^T U_1))\\ &\quad = \mathrm{tr}((A \odot U_1^T U_1 \odot \ldots U_{i-1}^T U_{i-1} \odot U_{i+1}^T U_{i+1} \odot \ldots \odot U_n^T U_n)\, U_j^T U_j)\\ &\quad = \mathrm{tr}(U_j\,(A \odot U_1^T U_1 \odot \ldots U_{i-1}^T U_{i-1} \odot U_{i+1}^T U_{i+1} \odot \ldots \odot U_n^T U_n)\, U_j^T)\\ &\quad = \mathrm{tr}(U_j\, \overline{A}_{t;j}\, U_j^T).\end{aligned}$$

Also, recall that B_t and $U_n \otimes_{kr} \ldots \otimes_{kr} U_1$ are $\left(\prod_{i=1}^n I_j\right) \times R$ matrices. We note that

$$\begin{aligned}&\mathrm{tr}\left(B_t^{(n+1)} (U_n \otimes_{kr} \ldots \otimes_{kr} U_1)^T\right)\\ &= \sum_{r=1}^R \mathrm{tr}\,(B_t(,r) \times_1 U_1(:,r) \times_2 \cdots \times_{i-1} U_{i-1}(:,r) \times_i U_i(:,r) \times_{i+1} U_{i+1}(:,r) \times_{i+2} \cdots \times_n U_n(:,r))\\ &= \mathrm{tr}\left(\sum_{r=1}^R [B_t(,r) \times_1 U_1(:,r) \times_2 \cdots \times_{i-1} U_{i-1}(:,r) \times_{i+1} U_{i+1}(:,r) \times_{i+2} \cdots \times_n U_n(:,r)]\, U_i(:,r)^T\right)\\ &= \mathrm{tr}\left(U_i \overline{B}_{t;j}^T\right),\end{aligned}$$

where $B_t(,r) \in \mathbb{R}^{I_1 \times \cdots \times I_n}$ denotes the r^{th} mode-$(n+1)$ slice of B_t. Then the assertion follows. ■

Proof [**Proof of Lemma 3 (i)-(iii)**] First, we show **(i)**. Write $\mathcal{D}_{t-1} = [U_1, \ldots, U_n]$ and $\mathcal{D}_t = [U_1', \ldots, U_n']$. Using Proposition 2 **(i)** and Lindeberg's replacement trick, we write

$$\begin{aligned}\hat{f}_t(\mathcal{D}_{t-1}) - \hat{f}_t(\mathcal{D}_t) &= \hat{f}_t([U_1, \ldots, U_n]) - \hat{f}_t([U_1', \ldots, U_n'])\\ &= \sum_{i=1}^n \hat{f}_t([U_1', \ldots, U_{i-1}', U_i, U_{i+1}, \ldots, U_n]) - \hat{f}_t([U_1', \ldots, U_{i-1}', U_i', U_{i+1}, \ldots, U_n]).\end{aligned}$$

Recall that U_i' is a minimizer of the function $U \mapsto \hat{f}_t([U_1', \ldots, U_{i-1}', U, U_{i+1}, \ldots, U_n])$ (which is convex by Proposition 2) over the convex set $\mathcal{C}_i$ defined in Algorithm 1. Also, U_i' belongs to $\mathcal{C}_i$. Hence each summand in the last expression above is nonnegative. This shows $\hat{f}_t(\mathcal{D}_{t-1}) - \hat{f}_t(\mathcal{D}_t) \geq 0$, as desired.

(ii) is clear from the algorithm (4).

Lastly, we show **(iii)**. Both $\hat{f}_t$ and f_t are uniformly bounded and Lipschitz by Lemma 15. Hence $h_t = \hat{f}_t - f_t$ is also Lipschitz with some constant $C'' > 0$ independent of t. Then by the recursive definitions of $\hat{f}_t$ and f_t (see (4) and (14)) and noting that $\ell(\mathcal{X}_t, \mathcal{D}_{t-1}, H_t) = \ell(\mathcal{X}_t, \mathcal{D}_{t-1})$, we have

$$\begin{aligned}|h_t(\mathcal{D}_t) - h_{t-1}(\mathcal{D}_{t-1})| &\leq |h_t(\mathcal{D}_t) - h_t(\mathcal{D}_{t-1})| + |h_t(\mathcal{D}_{t-1}) - h_{t-1}(\mathcal{D}_{t-1})| \qquad (16)\\ &\leq C\|\mathcal{D}_t - \mathcal{D}_{t-1}\|_F + \left|\left(\hat{f}_t(\mathcal{D}_{t-1}) - \hat{f}_{t-1}(\mathcal{D}_{t-1})\right) - \left(f_t(\mathcal{D}_{t-1}) - f_{t-1}(\mathcal{D}_{t-1})\right)\right|\\ &= C\|\mathcal{D}_t - \mathcal{D}_{t-1}\|_F + w_t|\hat{f}_{t-1}(\mathcal{D}_{t-1}) - f_{t-1}(\mathcal{D}_{t-1})|.\end{aligned}$$

If $\delta_t \equiv 1$, then this and **(i)** show $|h_t(\mathcal{D}_t) - h_{t-1}(\mathcal{D}_{t-1})| = O(w_t)$. For the more general case when $\delta_t \in [1 - w_t, 1]$, taking expectation and **(i)** shows $\mathbb{E}\,[|h_t(\mathcal{D}_t) - h_{t-1}(\mathcal{D}_{t-1})|] = O(w_t)$, as desired. ■

Next, we establish two elementary yet important inequalities connecting the empirical and surrogate loss functions. This is trivial in the case of vector-valued signals, in which case we can directly minimize $\hat{f}_t$ over a convex constraint set $\mathscr{C}^{\text{dict}}$ to find $\mathscr{D}_t$ so we have the the 'forward monotonicity' $\hat{f}_t(\mathscr{D}_t) \le \hat{f}_t(\mathscr{D}_{t-1})$ immediately from the algorithm design. In the tensor case, this still holds since we use block coordinate descent to progressively minimize $\hat{f}_t$ in each loading matrix. Also, one can deduce the forward monotonicity from Lemma 3 **(i)**.

Proposition 5 *Let $(\mathscr{D}_t)_{t\ge 1}$ be an output of Algorithm 1. Then for each $t \ge 0$, the following hold:*

(i) $\hat{f}_{t+1}(\mathscr{D}_{t+1}) - \hat{f}_t(\mathscr{D}_t) \le w_{t+1}\left(\ell(\mathscr{X}_{t+1}, \mathscr{D}_t) - f_t(\mathscr{D}_t)\right)$.

(ii) $0 \le w_{t+1}\left(\hat{f}_t(\mathscr{D}_t) - f_t(\mathscr{D}_t)\right) \le w_{t+1}\left(\ell(\mathscr{X}_{t+1}, \mathscr{D}_t) - f_t(\mathscr{D}_t)\right) + \hat{f}_t(\mathscr{D}_t) - \hat{f}_{t+1}(\mathscr{D}_{t+1})$.

Proof We begin by observing that

$$\hat{f}_{t+1}(\mathscr{D}_t) = (1 - w_{t+1})\hat{f}_t(\mathscr{D}_t) + w_{t+1}\ell_{t+1}(\mathscr{X}_{t+1}, \mathscr{D}_t) \tag{17}$$

for all $t \ge 0$. The first equality above uses the definition of $\hat{f}_t$, and the second equality uses the fact that H_{t+1} is a minimizer of $\ell(\mathscr{X}_{t+1}, \mathscr{D}_t, H)$ over $\mathscr{C}^{\text{code}}$. Hence

$$\begin{aligned} &\hat{f}_{t+1}(\mathscr{D}_{t+1}) - \hat{f}_t(\mathscr{D}_t) \qquad (18)\\ &\quad = \hat{f}_{t+1}(\mathscr{D}_{t+1}) - \hat{f}_{t+1}(\mathscr{D}_t) + \hat{f}_{t+1}(\mathscr{D}_t) - \hat{f}_t(\mathscr{D}_t)\\ &\quad = \hat{f}_{t+1}(\mathscr{D}_{t+1}) - \hat{f}_{t+1}(\mathscr{D}_t) + (1 - w_{t+1})\hat{f}_t(\mathscr{D}_t) + w_{t+1}\ell(\mathscr{X}_{t+1}, \mathscr{D}_t) - \hat{f}_t(\mathscr{D}_t)\\ &\quad = \hat{f}_{t+1}(\mathscr{D}_{t+1}) - \hat{f}_{t+1}(\mathscr{D}_t) + w_{t+1}(\ell(\mathscr{X}_{t+1}, \mathscr{D}_t) - f_t(\mathscr{D}_t)) + w_{t+1}(f_t(\mathscr{D}_t) - \hat{f}_t(\mathscr{D}_t)). \end{aligned}$$

Now note that $\hat{f}_{t+1}(\mathscr{D}_{t+1}) - \hat{f}_{t+1}(\mathscr{D}_t) \le 0$ by assumption and $f_t \le \hat{f}_t$ by definition. Furthermore, $\hat{f}_{t+1}(\mathscr{D}_{t+1}) - \hat{f}_{t+1}(\mathscr{D}_t) \le 0$ by Lemma 3 **(i)**, so the above inequalities apply for $\mathscr{D}_{t+1} = \mathscr{D}_{t+1}$. Thus the inequalities in both **(i)** and **(ii)** follow. ■

B.3. Stochastic analysis

In this section, we develop stochastic analysis on our online algorithm, a major portion of which is devoted to handle functional Markovian dependence in signals as stated in assumption (A1) (which generalizes (A1)). The analysis here is verbatim as the one developed in [29] for the vector-valued signal (or matrix factorization) case, which we present some of the important arguments in details here for the sake of completeness. However, the results in this subsection crucially relies on the deterministic analysis in the previous section that was necessary to handle difficulties arising in the tensor-valued signal case.

Recall that under our assumption (A1), the signals $(\mathscr{X}_t)_{t\ge 0}$ are modulated by an underlying Markov chain $(Y_t)_{t\ge 0}$ as $\mathscr{X}_t = \varphi(Y_t)$ for a fixed function φ. We would like to establish convergence of our online dictionary learning algorithm for tensor-valued signals in this general setting. Note that Proposition 5 gives a bound on the change in surrogate loss $\hat{f}_t(\mathscr{D}_t)$ in one iteration that allows to control its *positive variation* in terms of difference $\ell(\mathscr{X}_{t+1}, \mathscr{D}_t) - f_t(\mathscr{D}_t)$. The core of the stochastic analysis in this subsection is to get a good bound on this quantity. In the classical setting when Y_t's are i.i.d., our signals $\mathscr{X}_t = \varphi(Y_t)$ are also i.i.d., so we can condition on the information $\mathscr{F}_t$ up to time t so that

$$\mathbb{E}\left[\ell(\mathscr{X}_{t+1}, \mathscr{D}_t) - f_t(\mathscr{D}_t) \,\middle|\, \mathscr{F}_t\right] = f(\mathscr{D}_t) - f_t(\mathscr{D}_t). \tag{19}$$

Note that for each fixed $\mathscr{D} \in \mathscr{C}^{\text{dict}}$, $f_t(W) \to f(W)$ almost surely as $t \to \infty$ by the strong law of large numbers. To handle time dependence of the evolving dictionaries $\mathscr{D}_t$, one can instead look that the convergence of the supremum $\|f_t - f\|_\infty$ over the compact set $\mathscr{C}^{\text{dict}}$, which is provided by the classical Glivenko-Cantelli theorem. This is the approach taken in [32, 33] for i.i.d. input.

However, the same approach is not applicable for dependent signals, for instance, when $(Y_t)_{t\ge 0}$ is a Markov chain. This is because, in this case, conditional on $\mathscr{F}_t$, the distribution of Y_{t+1} is not necessarily the stationary distribution π. In fact, when Y_t's form a Markov chain with transition matrix P, Y_t given Y_{t-1} has distribution $P(Y_{t-1}, \cdot)$, and this conditional distribution is a constant distance away from the stationary distribution π. (For instance, consider the case when Y_t alternates between two matrices. Then $\pi = [1/2, 1/2]$ and π_t is either $[1,0]$ or $[0,1]$ for all $t \ge 1$.)

To handle dependence in data samples, we adopt the strategy developed in [29] in order to handle a similar issue for vector-valued signals (or matrix factorization). The key insight in [29] is that, while the 1-step conditional distribution $P(X_{t-1}, \cdot)$ may be far from the stationary distribution π, the N-step conditional distribution $P^N(X_{t-N}, \cdot)$ is exponentially close to π under mild conditions. Hence we can condition much early on – at time $t - N$ for some suitable $N = N(t)$. Then the Markov chain runs $N+1$ steps up to time $t+1$, so if N is large enough for the chain to mix to its stationary distribution π, then the distribution of Y_{t+1} conditional on $\mathscr{F}_{t-N}$ is close to π. The error of approximating the stationary distribution by the $N+1$ step distribution can be controlled using total variation distance and Markov chain mixing bound. This is stated more precisely in the proposition below.

Proposition 6 *Suppose (A1) hold. Fix a CP-dictionary $\mathscr{D}$. Then for each $t \ge 0$ and $0 \le N < t$, conditional on the information $\mathscr{F}_{t-N}$ up to time $t - N$,*

$$\left(\mathbb{E}\left[\ell(\mathscr{X}_{t+1}, \mathscr{D}) - f_t(\mathscr{D}) \,\middle|\, \mathscr{F}_{t-N}\right]\right)^+ \le \left|f(\mathscr{D}) - f_{t-N}(\mathscr{D})\right| + N w_t f_{t-N}(\mathscr{D}) \tag{20}$$

$$+ 2\|\ell(\cdot, \mathscr{D})\|_\infty \sup_{\mathbf{y}\in\Omega} \|P^{N+1}(\mathbf{y}, \cdot) - \pi\|_{TV}. \tag{21}$$

Proof Proof is identical to that of [29, Prop. 7.5]. ∎

Lemma 7 *Let $(\mathscr{D}_t)_{t\ge 1}$ be the output of Algorithm 1. Suppose (A1) and (A2) hold. Then the following hold.*

(i) $\sum_{t=0}^{\infty} \mathbb{E}\left[w_{t+1}\left(\ell(\mathscr{X}_{t+1}, \mathscr{D}_t) - f_t(\mathscr{D}_t)\right)\right]^+ < \infty$*;*

(ii) $\mathbb{E}[\hat{f}_t(\mathscr{D}_t)]$ *converges as $t \to \infty$;*

(iii) $\mathbb{E}\left[\sum_{t=0}^{\infty} w_{t+1}\left(\hat{f}_t(\mathscr{D}_t) - f_t(\mathscr{D}_t)\right)\right] = \sum_{t=0}^{\infty} w_{t+1}\left(\mathbb{E}[\hat{f}_t(\mathscr{D}_t)] - \mathbb{E}[f_t(\mathscr{D}_t)]\right) < \infty$*;*

(iv) $\sum_{t=0}^{\infty} w_{t+1}\left(\hat{f}_t(\mathscr{D}_t) - f_t(\mathscr{D}_t)\right) < \infty$ *almost surely.*

Proof Part **(i)** can be derived from Proposition 6 and Jensen's inequality. See the proof of [29, Lem. 12 **(ii)**] for details. Parts **(ii)**-**(iv)** can be shown by using Propositions 5, 6, and part **(i)**. See the proof of [29, Lem. 13] for details. ∎

B.4. Proof of Lemma 3 (iv)

In this subsection, we prove Lemma 3 **(iv)**, which is one of the most nontrivial arguments we give in this work. Throughout this subsection, we will denote by $(\mathcal{D}_t)_{t\ge 1}$ the output of Algorithm 1 and $\Lambda := \{\mathcal{D}_t \,|\, t\ge 1\} \subseteq \mathcal{C}^{\text{dict}}$. Note that by Proposition 2, $\hat{f}_{t_k}$ converges almost surely if and only if $A_{t_k}, B_{t_k}, \mathcal{X}_{t_k}, H_{t_k}$ converge a.s. as $k\to\infty$. In what follows, we say $\mathcal{D}_\infty \in \mathcal{C}^{\text{dict}}$ a *stationary point* of Λ if it is a limit point $\mathcal{D}_\infty$ of Λ and there exists a sequence $t_k\to\infty$ such that $\mathcal{D}_{t_k}\to\mathcal{D}_\infty$ and $\hat{f}_\infty := \lim_{k\to\infty}\hat{f}_{t_k}$ exists almost surely and $\mathcal{D}_\infty$ is a stationary point of $\hat{f}_\infty$ over $\mathcal{C}^{\text{dict}}$. Our goal is to show that every limit point of Λ is stationary.

The following observation is a key to our argument.

Proposition 8 *Assume (A1), (A2), and $\sum_{t=1}^\infty w_t^2\sqrt{t}<\infty$. Let $(\mathcal{D}_t)_{t\ge1}$ be an output of Algorithm 1. Then almost surely,*

$$\sum_{t=1}^{\infty}\left|\mathrm{tr}\left(\nabla\hat{f}_{t+1}(\mathcal{D}_{t+1})^T(\mathcal{D}_t-\mathcal{D}_{t+1})\right)\right|<\infty.$$

Proof Since $\mathcal{C}^{\text{dict}}$ is compact by (A2) and the aggregate tensors A_t, B_t are uniformly bounded by Lemma 14, we can see from Proposition 2 that $\nabla\hat{f}_{t+1}$ over $\mathcal{C}^{\text{dict}}$ is Lipschitz with some uniform constant $L>0$. Hence by Lemma 13, for all $t\ge1$,

$$\left|\hat{f}_{t+1}(\mathcal{D}_t)-\hat{f}_{t+1}(\mathcal{D}_{t+1})-\mathrm{tr}\left(\nabla\hat{f}_{t+1}(\mathcal{D}_{t+1})^T(\mathcal{D}_t-\mathcal{D}_{t+1})\right)\right|\le\frac{L}{2}\|\mathcal{D}_t-\mathcal{D}_{t+1}\|_F^2.$$

Also note that $\hat{f}_{t+1}(\mathcal{D}_t)\ge\hat{f}_{t+1}(\mathcal{D}_{t+1})$ by Lemma 3 **(i)**. Hence it follows that

$$\left|\mathrm{tr}\left(\nabla\hat{f}_{t+1}(\mathcal{D}_{t+1})^T(\mathcal{D}_t-\mathcal{D}_{t+1})\right)\right|\le\frac{L}{2}\|\mathcal{D}_t-\mathcal{D}_{t+1}\|_F^2+\hat{f}_{t+1}(\mathcal{D}_t)-\hat{f}_{t+1}(\mathcal{D}_{t+1}) \tag{22}$$

On the other hand, (18) and $\hat{f}_t\ge f_t$ yields

$$0\le\hat{f}_{t+1}(\mathcal{D}_t)-\hat{f}_{t+1}(\mathcal{D}_{t+1})\le\hat{f}_t(\mathcal{D}_t)-\hat{f}_{t+1}(\mathcal{D}_{t+1})+w_{t+1}(\ell(\mathcal{X}_{t+1},\mathcal{D}_t)-f_t(\mathcal{D}_t)).$$

Hence using Lemma 7, we have

$$\sum_{t=1}^{\infty}\mathbb{E}\left[\hat{f}_{t+1}(\mathcal{D}_t)-\hat{f}_{t+1}(\mathcal{D}_{t+1})\right]<\infty.$$

Then from (22) and noting that $\|\mathcal{D}_t-\mathcal{D}_{t+1}\|_F^2=O(w_{t+1}^2)$ and $\sum_{t=1}^\infty w_t^2<\infty$, it follows that

$$\begin{aligned}\sum_{t=1}^{\infty}\mathbb{E}\left[\left|\mathrm{tr}\left(\nabla\hat{f}_{t+1}(\mathcal{D}_{t+1})^T(\mathcal{D}_t-\mathcal{D}_{t+1})\right)\right|\right]&=\frac{L}{2}\sum_{t=1}^{\infty}\mathbb{E}\left[\|\mathcal{D}_t-\mathcal{D}_{t+1}\|_F^2\right]\\&\quad+\sum_{t=1}^{\infty}\mathbb{E}\left[\hat{f}_{t+1}(\mathcal{D}_t)-\hat{f}_{t+1}(\mathcal{D}_{t+1})\right]<\infty.\end{aligned}$$

Then the assertion follows by Fubini's theorem and the fact that $\mathbb{E}[|X|]<\infty$ implies $|X|<\infty$ almost surely for any random variable X. ■

Next, we show that the block coordinate descent we use to obtain $\mathcal{D}_{t+1}$ should always give the optimal first order descent up to an additive error of order $O(w_{t+1})$.

Proposition 9 *Assume (A1) and (A2). Then there exists a constant $c > 0$ such that for all $t \geq 1$,*

$$\operatorname{tr}\left(\nabla \hat{f}_{t+1}(\mathscr{D}_t)^T \frac{(\mathscr{D}_{t+1} - \mathscr{D}_t)}{\|\mathscr{D}_{t+1} - \mathscr{D}_t\|_F}\right) \leq \inf_{\mathscr{D} \in \mathscr{C}^{\mathrm{dict}}} \operatorname{tr}\left(\nabla \hat{f}_{t+1}(\mathscr{D}_t)^T \frac{(\mathscr{D} - \mathscr{D}_t)}{\|\mathscr{D} - \mathscr{D}_t\|_F}\right) + c w_{t+1}. \tag{23}$$

Proof Write $\mathscr{D}_t = [U_1^{(t)}, \ldots, U_n^{(t)}]$ and denote $\hat{f}_{t+1;i} : U \mapsto \hat{f}_{t+1}(U_1^{(t+1)}, \ldots, U_{i-1}^{(t+1)}, U, U_{i+1}^{(t)}, \ldots, U_n^{(t)})$ for $U \in \mathbb{R}^{I_i \times R}$ and $i = 1, \ldots, n$. Recall that $U_i^{(t+1)}$ is a minimizer of $\hat{f}_{t+1;i}$ over the convex set of $\mathscr{C}_i^{\mathrm{dict}}$ intersected with $\{U \colon \|U - U_i^{(t)}\| \leq c' w_{t+1}\}$. By the convexity, note that for each $U_i \in \mathscr{C}_i^{\mathrm{dict}}$, $U_i^{(t)} + a(U_i - U_i^{(t)}) \in \mathscr{C}_i^{\mathrm{dict}}$ for all $a \in [0,1]$. Hence denoting $\|\mathscr{C}_i^{\mathrm{dict}}\|_F := \sup_{U, U' \in \mathscr{C}_i} \|U - U'\|_F < \infty$,

$$\hat{f}_{t+1;i}(U_i^{(t+1)}) \leq \hat{f}_{t+1;i}\left(U_i^{(t)} + \frac{c w_{t+1}}{\|\mathscr{C}_i^{\mathrm{dict}}\|_F}(U_i - U_i^{(t)})\right) \tag{24}$$

for all $t \geq 1$. Recall that $\nabla \hat{f}_{t+1} = [\nabla \hat{f}_{t+1;1}, \ldots, \nabla \hat{f}_{t+1;n}]$ is Lipschitz with uniform Lipschitz constant $L > 0$ by Lemma 15. Hence by Lemma 13,

$$\operatorname{tr}\left(\nabla \hat{f}_{t+1;i}(U_i^{(t)})^T \frac{(U_i^{(t+1)} - U_i^{(t)})}{\|U_i^{(t+1)} - U_i^{(t)}\|_F}\right) \leq \operatorname{tr}\left(\nabla \hat{f}_{t+1;i}(U_i^{(t)})^T \frac{(U_i - U_i^{(t)})}{\|U_i - U_i^{(t)}\|_F}\right) + L c' w_{t+1}. \tag{25}$$

Adding up these inequalities for $i = 1, \ldots, n$ and writing $\mathscr{D} = [U_1, \ldots, U_n] \in \mathscr{C}^{\mathrm{dict}}$, we get

$$\begin{aligned}\operatorname{tr}\left(\nabla \hat{f}_{t+1}(\mathscr{D}_t)^T \frac{(\mathscr{D}_{t+1} - \mathscr{D}_t)}{\|\mathscr{D}_{t+1} - \mathscr{D}_t\|_F}\right) &\leq \operatorname{tr}\left(\left[\nabla \hat{f}_{t+1;1}(U_1^{(t)}), \ldots, \hat{f}_{t+1;n}(U_n^{(t)})\right]^T \frac{(\mathscr{D} - \mathscr{D}_t)}{\|\mathscr{D} - \mathscr{D}_t\|_F}\right) + nLc' w_{t+1} &(26)\\ &\leq \operatorname{tr}\left(\nabla \hat{f}_{t+1}(\mathscr{D}_t)^T \frac{(\mathscr{D} - \mathscr{D}_t)}{\|\mathscr{D} - \mathscr{D}_t\|_F}\right) + n(L+1)c' w_{t+1}, &(27)\end{aligned}$$

where the second inequality follows from Lipschitz continuity of $\nabla \hat{f}_{t+1}$. Since $\mathscr{D} \in \mathscr{C}^{\mathrm{dict}}$ was arbitrary, this shows the assertion. ■

Proposition 10 *Assume (A1), (A2), and $\sum_{t=1}^{\infty} w_t^2 \sqrt{t} < \infty$. Suppose there exists a subsequence $(\mathscr{D}_{t_k})_{k \geq 1}$ such that either*

$$\sum_{k=1}^{\infty} w_{t_k+1} = \infty \qquad \textit{or} \qquad \liminf_{k \to \infty} \left|\operatorname{tr}\left(\nabla \hat{f}_{t_k+1}(\mathscr{D}_{t_k+1})^T \frac{\mathscr{D}_{t_k} - \mathscr{D}_{t_k+1}}{\|\mathscr{D}_{t_k} - \mathscr{D}_{t_k+1}\|_F}\right)\right| = 0. \tag{28}$$

There exists a further subsequence $(s_k)_{k \geq 1}$ of $(t_k)_{k \geq 1}$ such that $\mathscr{D}_\infty := \lim_{k \to \infty} \mathscr{D}_{s_k}$ exists and is a stationary point of Λ.

Proof By Proposition 8, we have

$$\sum_{k=1}^{\infty} w_{t_k+1} \left|\operatorname{tr}\left(\nabla \hat{f}_{t_k+1}(\mathscr{D}_{t_k+1})^T \frac{\mathscr{D}_{t_k} - \mathscr{D}_{t_k+1}}{\|\mathscr{D}_{t_k} - \mathscr{D}_{t_k+1}\|_F}\right)\right| < \infty. \tag{29}$$

Hence if $\sum_{k=1}^{\infty} w_{t_k+1} = \infty$, then the latter condition in (28) holds. Thus it suffices to show that this latter condition implies the assertion. Assume this condition, and let $(s_k)_{k \geq 1}$ be a subsequence of $(t_k)_{k \geq 1}$ for which the liminf in (28) is achieved. By taking a subsequence, we may assume that $\mathscr{D}'_\infty = \lim_{k \to \infty} \mathscr{D}_{s_k}$ and $\hat{f}_\infty := \lim_{k \to \infty} \hat{f}_{s_k}$ exist.

Now suppose for contradiction that $\mathscr{D}_\infty$ is not a stationary point of $\hat{f}_\infty$ over $\mathscr{C}^{\text{dict}}$. Then there exists $\mathscr{D}^\star \in \mathscr{C}^{\text{dict}}$ and $\delta > 0$ such that

$$\operatorname{tr}\left(\nabla \hat{f}_\infty(\mathscr{D}_\infty)^T(\mathscr{D}^\star - \mathscr{D}_\infty)\right) < -\delta < 0. \tag{30}$$

By triangle inequality, write

$$\begin{aligned} &\|\nabla \hat{f}_{s_k+1}(\mathscr{D}_{s_k})^T(\mathscr{D}^\star - \mathscr{D}_{s_k}) - \nabla \hat{f}_\infty(\mathscr{D}_\infty)^T(\mathscr{D}^\star - \mathscr{D}_\infty)\|_F && (31)\\ &\quad \le \|\nabla \hat{f}_{s_k+1}(\mathscr{D}_{s_k}) - \nabla \hat{f}_\infty(\mathscr{D}_\infty)\|_F \cdot \|\mathscr{D}^\star - \mathscr{D}_{s_k}\|_F + \|\nabla \hat{f}_\infty(\mathscr{D}_\infty)\|_F \cdot \|\mathscr{D}_\infty - \mathscr{D}_{s_k}\|_F. && (32)\end{aligned}$$

Noting that $\|\mathscr{D}_t - \mathscr{D}_{t-1}\|_F = O(w_t) = o(1)$, we see that the right hand side goes to zero as $k \to \infty$. Hence for all sufficiently large $k \ge 1$, we have

$$\operatorname{tr}\left(\nabla \hat{f}_{s_k+1}(\mathscr{D}_{s_k})^T(\mathscr{D}^\star - \mathscr{D}_{s_k})\right) < -\delta/2. \tag{33}$$

Then by Proposition 9, denoting $\|\mathscr{C}^{\text{dict}}\|_F := \sup_{\mathscr{D},\mathscr{D}' \in \mathscr{C}^{\text{dict}}} \|\mathscr{D} - \mathscr{D}'\|_F < \infty$,

$$\liminf_{k\to\infty} \operatorname{tr}\left(\nabla \hat{f}_{s_k+1}(\mathscr{D}_{s_k+1})^T \frac{\mathscr{D}_{s_k} - \mathscr{D}_{s_k+1}}{\|\mathscr{D}_{s_k} - \mathscr{D}_{s_k+1}\|_F}\right) \le -\frac{\delta}{2\|\mathscr{C}^{\text{dict}}\|_F} < 0, \tag{34}$$

which contradicts the choice of the subsequence $(\mathscr{D}_{s_k})_{k\ge1}$. This shows the assertion. ■

Recall that during the update $\mathscr{D}_{t-1} \mapsto \mathscr{D}_t$ each factor matrix of $\mathscr{D}_{t-1}$ changes by at most $c' w_t$ in Frobenius norm. For each $t \ge 1$, we say $\mathscr{D}_t$ is a *long point* if none of the factor matrices of $\mathscr{D}_{t-1}$ change by $c' w_t$ in Frobenius norm and *short point* otherwise. Observe that if $\mathscr{D}_t$ is a long point, then imposing the search radius restriction in 12 has no effect and $\mathscr{D}_t$ is obtained from $\mathscr{D}_{t-1}$ by a single cycle of block coordinate descent on $\hat{f}_t$ over $\mathscr{C}^{\text{dict}}$.

Proposition 11 *Assume (A1) and (A2). If $(\mathscr{D}_{t_k})_{k\ge1}$ is a convergent subsequence of $(\mathscr{D}_t)_{t\ge1}$ consisting of long points, then the $\mathscr{D}_\infty = \lim_{k\to\infty} \mathscr{D}_{s_k}$ is stationary.*

Proof For each $A \in \mathbb{R}^{R\times R}$, $B \in \mathbb{R}^{I_1\times\cdots\times I_n\times b}$, $\mathscr{D} = [U_1,\dots,U_n] \in \mathbb{R}^{I_1\times R}\times\cdots\times\mathbb{R}^{I_n\times R}$, define

$$\hat{g}(A,B,\mathscr{D}) = \operatorname{tr}(A\,(U_n^T U_n \odot \ldots \odot U_1^T U_1)) - 2\operatorname{tr}\left(B^{(n+1)}(U_n \otimes_{kr} \ldots \otimes_{kr} U_1)^T\right). \tag{35}$$

By taking a subsequence of $(t_k)_{k\ge1}$, we may assume that $A_\infty := \lim_{k\to\infty} A_{t_k}$ and $B_\infty := \lim_{k\to\infty} B_{t_k}$ exist. Hence the function $\hat{g}_\infty := \lim_{k\to\infty} \hat{g}_{t_k} = \hat{g}(A_\infty, B_\infty, \cdot)$ is well-defined. Noting that $\nabla \hat{f}_t = \nabla \hat{g}_t$ for all $t \ge 1$, it suffices to show that $\mathscr{D}_\infty$ is a stationary point of $\hat{g}_\infty$ over $\mathscr{C}^{\text{dict}}$ almost surely.

The argument is similar to that of [7, Prop. 2.7.1]. However, here we do not need to assume uniqueness of solutions to minimization problems of $\hat{f}_t$ in each block coordinate due to the added search radius restriction. Namely, write $\mathscr{D}_\infty = [U_1^{(\infty)},\dots,U_n^{(\infty)}]$. Then for each $k \ge 1$,

$$\hat{g}_{t_k}(U_1^{(t_k)}, U_2^{(t_k-1)},\dots,U_n^{(t_k-1)}) \le \hat{g}_{t_k}(U_1, U_2^{(t_k-1)},\dots,U_n^{(t_k-1)}) \tag{36}$$

for all $U_1 \in \mathscr{C}_1^{\text{dict}} \cap \{U\colon \|U - U_1^{(t_k-1)}\|_F \le c' w_{t_k}\}$. In fact, since $\mathscr{D}_{t_k}$ is a long point by the assumption, (36) holds for all $U_1 \in \mathscr{C}_1^{\text{dict}}$. Taking $k \to \infty$ and using the fact that $\|U_1^{(t_k)} - U_1^{(t_k-1)}\|_F \le c' w_{t_k}$,

$$\hat{g}_\infty(U_1^{(\infty)}, U_2^{(\infty)},\dots,U_n^{(\infty)}) \le \hat{g}_\infty(U_1, U_2^{(\infty)},\dots,U_n^{(\infty)}) \quad \text{for all } U_1 \in \mathscr{C}_1^{\text{dict}}. \tag{37}$$

Since $\mathscr{C}_1^{\text{dict}}$ is convex, it follows that

$$\nabla_1 \hat{g}_\infty(\mathscr{D}_\infty)^T (U_1 - U_1^{(\infty)}) \geq 0 \qquad \text{for all } U_1 \in \mathscr{C}_1^{\text{dict}}. \tag{38}$$

By using a similar argument for other coordinates of $\mathscr{D}_\infty$, it follows that $\nabla \hat{g}_\infty(\mathscr{D}_\infty)^T(\mathscr{D} - \mathscr{D}_\infty) \geq 0$ for all $\mathscr{D} \in \mathscr{C}^{\text{dict}}$. This shows the assertion. ■

Proposition 12 *Assume (A1), (A2), $\sum_{t=1}^\infty w_t = \infty$, and $\sum_{t=1}^\infty w_t^2 \sqrt{t} < \infty$. Suppose there exists a non-stationary limit point $\mathscr{D}_\infty$ of Λ. Then there exists $\varepsilon > 0$ such that the ε-neighborhood $B_\varepsilon(\mathscr{D}_\infty) := \{\mathscr{D} \in \mathscr{C}^{\text{dict}} \,|\, \|\mathscr{D} - \mathscr{D}_\infty\|_F < \varepsilon\}$ with the following properties:*

(a) *$B_\varepsilon(\mathscr{D}_\infty)$ does not contain any stationary points of Λ.*

(b) *There exists infinitely many $\mathscr{D}_t$'s outside of $B_\varepsilon(\mathscr{D}_\infty)$.*

Proof We will first show that there exists an ε-neighborhood $B_\varepsilon(\mathscr{D}_\infty)$ of $\mathscr{D}_\infty$ that does not contain any long points of Λ. Suppose for contradiction that for each $\varepsilon > 0$, there exists a long point Λ in $B_\varepsilon(\mathscr{D}_\infty)$. Then one can construct a sequence of long points converging to $\mathscr{D}_\infty$. But then by Proposition 11, $\mathscr{D}_\infty$ is a stationary point, a contradiction.

Next, we show that there exists $\varepsilon > 0$ such that $B_\varepsilon(\mathscr{D}_\infty)$ satisfies **(a)**. Suppose for contradiction that there exists no such $\varepsilon > 0$. Then we have a sequence $(\mathscr{D}_{\infty;k})_{k\geq 1}$ of stationary points of Λ that converges to $\mathscr{D}_\infty$. Denote the limiting surrogate loss function associated with $\mathscr{D}_{\infty;k}$ by $\hat{f}_{\infty;k}$. Recall that each $\hat{f}_{\infty;k}$ is parameterized by elements in a compact set (see (A1), Proposition 2, and Lemma 15). Hence by choosing a subsequence, we may assume that $\hat{f}_\infty := \lim_{k\to\infty} \hat{f}_{\infty;k}$ is well-defined. Fix $\mathscr{D} \in \mathscr{C}^{\text{dict}}$ note that by Cauchy-Schwarz inequality,

$$\begin{aligned}\nabla \hat{f}_\infty(\mathscr{D}_\infty)^T(\mathscr{D} - \mathscr{D}_\infty) \geq &-\|\nabla \hat{f}_\infty(\mathscr{D}_\infty) - \nabla \hat{f}_{\infty;k}(\mathscr{D}_{\infty;k})\|_F \cdot \|\mathscr{D} - \mathscr{D}_\infty\|_F \qquad (39)\\ &- \|\nabla \hat{f}_{\infty;k}(\mathscr{D}_{\infty;k})\|_F \cdot \|\mathscr{D}_\infty - \mathscr{D}_{\infty;k}\|_F + \nabla \hat{f}_{\infty;k}(\mathscr{D}_{\infty;k})^T(\mathscr{D} - \mathscr{D}_{\infty;k}). \qquad (40)\end{aligned}$$

Note that $\nabla \hat{f}_{\infty;k}(\mathscr{D}_{\infty;k})^T(\mathscr{D} - \mathscr{D}_{\infty;k}) \geq 0$ since $\mathscr{D}_{\infty;k}$ is a stationary point of $\hat{f}_{\infty;k}$ over $\mathscr{C}^{\text{dict}}$. Hence by taking $k \to \infty$, this shows $\nabla \hat{f}_\infty(\mathscr{D}_\infty)^T(\mathscr{D} - \mathscr{D}_\infty) \geq 0$. Since $\mathscr{D} \in \mathscr{D}^{\text{dict}}$ was arbitrary, this shows that $\mathscr{D}_\infty$ is a stationary point of $\hat{f}_\infty$ over $\mathscr{C}^{\text{dict}}$, a contradiction.

Lastly, from the earlier results, we can choose $\varepsilon > 0$ such that $B_\varepsilon(\mathscr{D}_\infty)$ has no long points of Λ and also satisfies **(b)**. We will show that $B_{\varepsilon/2}(\mathscr{D}_\infty)$ satisfies **(c)**. Then $B_{\varepsilon/2}(\mathscr{D}_\infty)$ satisfies **(a)**-**(b)**, as desired. Suppose for contradiction there are only finitely many $\mathscr{D}_t$'s outside of $B_{\varepsilon/2}(\mathscr{D}_\infty)$. Then there exists an integer $M \geq 1$ such that $\mathscr{D}_t \in B_{\varepsilon/2}(\mathscr{D}_\infty)$ for all $t \geq M$. Then each $\mathscr{D}_t$ for $t \geq M$ is a short point of Λ. By definition, it follows that $\|\mathscr{D}_{t-1} - \mathscr{D}\|_F \geq c' w_t$ for all $t \geq M$. Then by Proposition 8, we have

$$\sum_{t\geq M} w_{t+1} \left| \operatorname{tr}\left(\nabla \hat{f}_{t+1}(\mathscr{D}_{t+1})^T \frac{\mathscr{D}_t - \mathscr{D}_{t+1}}{\|\mathscr{D}_t - \mathscr{D}_{t+1}\|_F} \right) \right| < \infty. \tag{41}$$

Since $\sum_{t=1}^\infty w_t = \infty$, there exists a subsequence $(s_k)_{k\geq 1}$ such that $\mathscr{D}'_\infty := \lim_{k\to\infty} \mathscr{D}_{t_k}$ exists and is stationary. But since $\mathscr{D}'_\infty \in B_\varepsilon(\mathscr{D})$, this contradicts **(a)** for $B_\varepsilon(\mathscr{D})$. This shows the assertion. ■

We are now ready to give a proof of Lemma 3 **(iv)**.

Proof [**Proof of Lemma 3 (iv)**] Assume (A1), (A2), $\sum_{t=1}^{\infty} w_t = \infty$, and $\sum_{t=1}^{\infty} w_t^2 \sqrt{t} < \infty$. Suppose there exists a non-stationary limit point $\mathscr{D}_\infty$ of Λ. By Proposition 12, we may choose $\varepsilon > 0$ such that $B_\varepsilon(\mathscr{D}_\infty)$ satisfies the conditions **(a)**-**(b)** of Proposition 12. Choose $M \geq 1$ large enough so that $w_t < \varepsilon/4$ whenever $t \geq M$. We call an integer interval $I := [\ell, \ell']$ a *crossing* if $\mathscr{D}_\ell \in B_{\varepsilon/3}(\mathscr{D}_\infty)$, $\mathscr{D}_{\ell'} \in B_{2\varepsilon/3}(\mathscr{D}_\infty)$, and no proper subset of I satisfies both of these conditions. By definition, two distinct crossings have empty intersection. Fix a crossing $I = [\ell, \ell']$, it follows that by triangle inequality,

$$c'\sqrt{n} \sum_{t=\ell}^{\ell'-1} w_{t+1} \geq \sum_{t=\ell}^{\ell'-1} \|\mathscr{D}_{t+1} - \mathscr{D}_t\|_F \geq \|\mathscr{D}_{\ell'} - \mathscr{D}_\ell\|_F \geq \varepsilon/3. \tag{42}$$

Note that since $\mathscr{D}_\infty$ is a limit point of Λ, $\mathscr{D}_t$ visits $B_{\varepsilon/3}(\mathscr{D}_\infty)$ infinitely often. Moreover, by condition **(a)** of Proposition 12, $\mathscr{D}_t$ also exits $B_\varepsilon(\mathscr{D}_\infty)$ infinitely often. It follows that there are infinitely many crossings. Let t_k denote the k^{th} smallest integer that appears in some crossing. Then $t_k \to \infty$ as $k \to \infty$, and by (42) (recall that n denotes the number of modes in tensors and is fixed),

$$\sum_{k=1}^{\infty} w_{t_k+1} \geq (\text{\# of crossings}) \frac{c'\varepsilon}{3\sqrt{n}} = \infty. \tag{43}$$

Then by Proposition 10, there exists a further subsequence $(s_k)_{k\geq 1}$ of $(t_k)_{k\geq 1}$ such that $\mathscr{D}'_\infty := \lim_{k\to\infty} \mathscr{D}_{s_k}$ exists and is stationary. However, since $\mathscr{D}_{t_k} \in B_{2\varepsilon/3}(\mathscr{D}_\infty)$, we have $\mathscr{D}'_\infty \in B_\varepsilon(\mathscr{D}_\infty)$. This contradicts condition **(b)** of Proposition 12 for $B_\varepsilon(\mathscr{D}_\infty)$ that it cannot contain any stationary point of Λ. This shows the assertion. ■

B.5. Proof of the main result

Now we prove the first main result in this paper, Theorem 1.

Proof [**Proof of Theorem 1**] Suppose (A1) and (A2) hold. We first show **(i)**. Recall that $\mathbb{E}[\hat{f}_t(\mathscr{D}_t)]$ converges by Lemma 7. Jensen's inequality and Lemma 3 **(iv)** imply

$$|\mathbb{E}[h_{t+1}(\mathscr{D}_{t+1})] - \mathbb{E}[h_t(\mathscr{D}_t)]| \leq \mathbb{E}[|h_{t+1}(\mathscr{D}_{t+1}) - h_t(\mathscr{D}_t)|] = O(w_{t+1}). \tag{44}$$

Since $\mathbb{E}[\hat{f}_t(\mathscr{D}_t)] \geq \mathbb{E}[f_t(\mathscr{D}_t)]$, Lemma 7 **(ii)**-**(iii)** and Lemma 16 give

$$\lim_{t\to\infty} \mathbb{E}[f_t(\mathscr{D}_t)] = \lim_{t\to\infty} \mathbb{E}[\hat{f}_t(\mathscr{D}_t)] + \lim_{t\to\infty} \left(\mathbb{E}[f_t(\mathscr{D}_t)] - \mathbb{E}[\hat{f}_t(\mathscr{D}_t)]\right) = \lim_{t\to\infty} \mathbb{E}[\hat{f}_t(\mathscr{D}_t)] \in (1, \infty). \tag{45}$$

This shows **(i)**.

Next, we show **(ii)**. Triangle inequality gives

$$|f(\mathscr{D}_t) - \hat{f}_t(\mathscr{D}_t)| \leq \left(\sup_{\mathscr{D}\in\mathscr{C}^{\text{dict}}} |f(\mathscr{D}) - f_t(\mathscr{D})|\right) - h_t(\mathscr{D}_t). \tag{46}$$

Note that $|h_{t+1}(\mathscr{D}_{t+1}) - h_t(\mathscr{D}_t)| = O(w_{t+1})$ by Lemma 3 **(iii)**. Hence Lemma 7 **(iv)** and Lemma 16 show that $h_t(\mathscr{D}_t) \to 0$ almost surely. Furthermore, (46) and Lemma 17 show that $|f(\mathscr{D}_t) - \hat{f}_t(\mathscr{D}_t)| \to 0$ almost surely. This completes the proof of **(ii)**.

Lastly, we show **(iii)**. Let $\mathscr{D}_\infty \in \mathscr{C}^{\text{dict}}$ be an arbitrary limit point of the sequence $(\mathscr{D}_t)_{t\geq 1}$. Recall that $\Sigma_t := (\mathscr{D}_t, A_t, B_t, r_t)_{t\geq 0}$ is bounded by Lemma 14 and (A1) and (A2). Hence we may choose a random subsequence $(t_k)_{k\geq 1}$ so that $\mathscr{D}_{t_k} \to \mathscr{D}_\infty$. By taking a further subsequence, we may also assume that Σ_{t_k} converges to some random element $(\mathscr{D}_\infty, A_\infty, B_\infty, r_\infty)$ a.s. as $k\to\infty$. Then $\hat{f}_\infty := \lim_{k\to\infty} \hat{f}_{t_k}$ exists almost surely. It is important to note that $\mathscr{D}_\infty$ is a stationary point of $\hat{f}_\infty$ over $\mathscr{C}^{\text{dict}}$ by Lemma 3 **(iv)**.

Recall that $\hat{f}_t(\mathscr{D}_t) - f_t(\mathscr{D}_t) \to 0$ as $t\to\infty$ almost surely by part **(ii)**. By using continuity of $\hat{f}_t$, f_t, f in parameters (see Assumption (A3)d), it follows that

$$\left|\hat{f}_\infty(\mathscr{D}_\infty) - f(\mathscr{D}_\infty)\right| = \lim_{k\to\infty} \left|\hat{f}_{t_k}(\mathscr{D}_{t_k}) - f_{t_k}(\mathscr{D}_{t_k})\right| \leq \lim_{k\to\infty} \left(\sup_{\mathscr{D}\in\mathscr{C}^{\text{dict}}} \left|f - f_{t_k}(\mathscr{D})\right| - h_{t_k}(\mathscr{D}_{t_k}) \right) = 0, \tag{47}$$

where the last equality also uses Lemma 17.

Fix $\varepsilon > 0$ and $\mathscr{D} \in \mathbb{R}^{I_1\times R} \times \cdots \times \mathbb{R}^{I_n\times R}$. Hence, almost surely,

$$\hat{f}_\infty(\mathscr{D}_\infty + \mathscr{D}) = \lim_{k\to\infty} \hat{f}_{s_k}(\mathscr{D}_{s_k} + \mathscr{D}) \geq \lim_{k\to\infty} f_{s_k}(\mathscr{D}_{s_k} + \mathscr{D}) = f(\mathscr{D}_\infty + \mathscr{D}), \tag{48}$$

where the last equality follows from Lemma 17. Using first order Taylor expansion, write

$$\hat{f}_\infty(\mathscr{D}_\infty + \varepsilon\mathscr{D}) = \hat{f}_\infty(\mathscr{D}_\infty) + \operatorname{tr}\left(\nabla \hat{f}_\infty(\mathscr{D}_\infty)^T(\varepsilon\mathscr{D})\right) + O\left(\varepsilon^2\|\mathscr{D}\|_F^2\right), \tag{49}$$

$$f(\mathscr{D}_\infty + \varepsilon\mathscr{D}) = f(\mathscr{D}_\infty) + \operatorname{tr}\left(\nabla f(\mathscr{D}_\infty)^T(\varepsilon\mathscr{D})\right) + O\left(\varepsilon^2\|\mathscr{D}\|_F^2\right). \tag{50}$$

Recall that $\hat{f}_\infty(\mathscr{D}_\infty) = f(\mathscr{D}_\infty)$ a.s. by (47). Hence it follows that there exists some constant $c > 0$ such that almost surely

$$\operatorname{tr}\left(\left(\nabla\hat{f}_\infty(\mathscr{D}_\infty) - \nabla f(\mathscr{D}_\infty)\right)^T(\varepsilon\mathscr{D})\right) \geq -c\varepsilon^2\|\mathscr{D}\|_F^2. \tag{51}$$

After canceling out $\varepsilon > 0$ and letting $\varepsilon \searrow 0$ in (51),

$$\operatorname{tr}\left(\left(\nabla\hat{f}_\infty(\mathscr{D}_\infty) - \nabla f(\mathscr{D}_\infty)\right)^T\mathscr{D}\right) \geq 0 \qquad \text{a.s.} \tag{52}$$

Since this holds for all $\mathscr{D} \in \mathbb{R}^{I_1\times R}\cdots\times\mathbb{R}^{I_n\times R}$, it follows that $\nabla\hat{f}_\infty(\mathscr{D}_\infty) = \nabla f(\mathscr{D}_\infty)$ almost surely. But since $\mathscr{D}_\infty$ is a stationary point of $\hat{f}_\infty$ over $\mathscr{C}^{\text{dict}}$ by Lemma 3 **(iv)**, it follows that $\nabla\hat{f}_\infty(\mathscr{D}_\infty)$ is in the normal cone of $\mathscr{C}^{\text{dict}}$ at $\mathscr{D}_\infty$ (see., e.g., [10]). The same holds for $\nabla f(\mathscr{D}_\infty)$. This means that $\mathscr{D}_\infty$ is a stationary point of f over $\mathscr{C}^{\text{dict}}$. Since $\mathscr{D}_\infty$ is an arbitrary limit point of $\mathscr{D}_t$, the desired conclusion follows. ■

Appendix C. Auxiliary lemmas

Lemma 13 (Convex Surrogate for Functions with Lipschitz Gradient) *Let $f:\mathbb{R}^p\to\mathbb{R}$ be differentiable and ∇f be L-Lipschitz continuous. Then for each $\theta, \theta' \in \mathbb{R}^p$,*

$$\left|f(\theta') - f(\theta) - \nabla f(\theta)^T(\theta'-\theta)\right| \leq \frac{L}{2}\|\theta - \theta'\|_F^2. \tag{53}$$

Proof This is a classical Lemma. See [35, Lem 1.2.3]. ■

For each $\mathscr{X} \in \mathbb{R}^{I_1 \times \cdots \times I_n \times b}$ and $\mathscr{D} \in \mathbb{R}^{I_1 \times R} \times \ldots \mathbb{R}^{I_n \times R}$, denote

$$H^{\star}(\mathscr{X}, \mathscr{D}) \in \operatorname*{arg\,min}_{H \in \mathscr{C}^{\text{code}}} \ell(\mathscr{X}, \mathscr{D}, H). \tag{54}$$

Recall Assumption (A1). Denote $\|\varphi(\Omega)\|_F = \sup_{Y \in \Omega} \|\varphi(Y)\|_F$. The following boundedness results for the codes H_t and aggregate tensors A_t, B_t are easy to derive.

Lemma 14 *Assume (A1) and (A2). Then the following hold:*

(i) *For all $\mathscr{X} \in \mathbb{R}^{I_1 \times \cdots \times I_n \times b}$ and $\mathscr{D} \in \mathscr{C}^{\text{dict}}$,*

$$\|H^{\star}(\mathscr{X}, \mathscr{D})\|_F^2 \le \lambda^{-2} \|\varphi(\Omega)\|_F^4 < \infty. \tag{55}$$

(ii) *For any sequences $(\mathscr{X}_t)_{t \ge 1}$ in $\mathbb{R}^{I_1 \times \cdots \times I_n \times b}$ and $(\mathscr{D}_t)_{t \ge 1}$ in $\mathscr{C}$, define A_t and B_t recursively as in (8). Then for all $t \ge 1$, we have*

$$\|A_t\|_F \le \lambda^{-2} \|\varphi(\Omega)\|_F^4, \qquad \|B_t\|_F \le \lambda^{-1} \|\varphi(\Omega)\|_F^3. \tag{56}$$

Proof Omitted. See [29, Prop. 7.2]. ■

The following lemma shows Lipschitz continuity of the loss function $\ell(\varphi(\cdot), \cdot)$ defined in (13). Since Ω and $\mathscr{C}^{\text{code}}$ are both compact, this also implies that $\mathscr{D} \mapsto \hat{f}_t(\mathscr{D})$ and $\mathscr{D} \mapsto f_t(\mathscr{D})$ are L-Lipschitz for some $L > 0$ uniformly for all $t \ge 0$.

Lemma 15 *Suppose (A1) and (A2) hold, and let $M = 2\|\varphi(\Omega)\|_F + 2\|\mathscr{C}\|_F \|\varphi(\Omega)\|_F^2 / \lambda$. Then for each $Y_1, Y_2 \in \Omega$ and $\mathscr{D}_1, \mathscr{D}_2 \in \mathscr{C}^{\text{dict}}$,*

$$|\ell(\varphi(Y_1), \mathscr{D}_1) - \ell(\varphi(Y_2), \mathscr{D}_2)| \le M \left(\|Y_1 - Y_2\|_F + \lambda^{-1} \|\varphi(\Omega)\|_F \|\mathscr{D}_1 - \mathscr{D}_2\|_F \right). \tag{57}$$

Proof Omitted. See [29, Prop. 7.3]. ■

The following deterministic statement on converging sequences is due to [33].

Lemma 16 *Let $(a_n)_{n \ge 0}$ and $(b_n)_{\ge 0}$ be non-negative real sequences such that*

$$\sum_{n=0}^{\infty} a_n = \infty, \qquad \sum_{n=0}^{\infty} a_n b_n < \infty, \qquad |b_{n+1} - b_n| = O(a_n). \tag{58}$$

Then $\lim_{n \to \infty} b_n = 0$.

Proof Omitted. See [32, Lem. A.5]. ■

Lemma 17 *Under the assumptions (A1) and (A2),*

$$\mathbb{E}\left[\sup_{W \in \mathscr{C}^{\text{dict}}} \sqrt{t} \left| f(\mathscr{D}) - \frac{1}{t} \sum_{s=1}^{t} \ell(\mathscr{X}_s, \mathscr{D}) \right| \right] = O(1). \tag{59}$$

Furthermore, $\sup_{W \in \mathscr{C}} \left| f(\mathscr{D}) - \frac{1}{t} \sum_{s=1}^{t} \ell(\mathscr{X}_s, \mathscr{D}) \right| \to 0$ *almost surely as* $t \to \infty$.

Proof Omitted. See [29, Lem. 7.8]. ■