Complexity of Block Coordinate Descent with Proximal Regularization and Applications to Wasserstein CP-dictionary Learning

Dohyun Kwon [*1] **Hanbaek Lyu** [*2]

Abstract

We consider the block coordinate descent methods of Gauss-Seidel type with proximal regularization (BCD-PR), which is a classical method of minimizing general nonconvex objectives under constraints that has a wide range of practical applications. We theoretically establish the worst-case complexity bound for this algorithm. Namely, we show that for general nonconvex smooth objective with block-wise constraints, the classical BCD-PR algorithm converges to an ε-stationary point within $\widetilde{O}(\varepsilon^{-1})$ iterations. Under a mild condition, this result still holds even if the algorithm is executed inexactly in each step. As an application, we propose a provable and efficient algorithm for 'Wasserstein CP-dictionary learning', which seeks a set of elementary probability distributions that can well-approximate a given set of d-dimensional joint probability distributions. Our algorithm is a version of BCD-PR that operates in the dual space, where the primal problem is regularized both entropically and proximally.

1. Introduction

Consider the minimization of a continuous function $f : \mathbb{R}^{I_1} \times \cdots \times R^{I_m} \to [0, \infty)$ on a cartesian product of convex sets $\mathbf{\Theta} = \Theta^{(1)} \times \cdots \times \Theta^{(m)}$:

$$\boldsymbol{\theta}^* \in \underset{\boldsymbol{\theta}=[\theta_1,\ldots,\theta_m]\in\mathbf{\Theta}}{\arg\min} f(\theta_1, \ldots, \theta_m). \tag{1}$$

When the objective function f is nonconvex, the convergence of any algorithm for solving (1) to a globally optimal solution can hardly be expected. Instead, global convergence to stationary points of the objective function is desired, and in some problem classes, stationary points could be as good as global optimizers either practically as well as theoretically (see (Mairal et al., 2010; Sun et al., 2015)).

In order to solve (1), we will consider the *block coordinate descent* (BCD) methods of Gauss–Seidel type, which seeks to minimize the objective function restricted to a subset (block) of coordinates (Wright, 2015), often following the cyclic order of blocks. For the minimization problem (1) we refer to the set of coordinates in each $\Theta^{(i)}$, $i = 1, \ldots, m$, a block coordinate. Namely, let $\theta_n^{(i)}$ denote the ith block of the parameter after n updates. Write

$$\boldsymbol{\theta}_n^{(i-1)} := (\theta_n^{(1)}, \cdots, \theta_n^{(i)}, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}), \tag{2}$$
$$f_n^{(i)}(\theta) := f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right).$$

The algorithm we consider in this work updates $\boldsymbol{\theta}_n^{(i-1)}$ to $\boldsymbol{\theta}_n^{(i)}$ by updating its ith block by minimizing the marginal loss function $g_n^{(i)}$ over the ith block $\Theta^{(i)}$:

$$\theta_n^{(i)} \leftarrow \underset{\theta\in\Theta^{(i)}}{\arg\min}\, g_n^{(i)}(\theta) := f_n^{(i)}(\theta) + \frac{\lambda_n}{2}\|\theta - \theta_{n-1}^{(i)}\|^2, \tag{3}$$

where $\lambda_n \geq 0$ is called proximal regularization coefficient and $\|\cdot\|$ denotes the Frobenius norm. The proximal regularzer $\lambda_n\|\theta - \theta_{n-1}^{(i)}\|^2$ ensures that the next block iterate $\theta_n^{(i)}$ is not too far from the previous iterate $\theta_{n-1}^{(i)}$. The above update is applied cyclicly for $i = 1, \ldots, m$. We call the algorithm (3) BCD-PR for block coordinate descent with proximal regularization.

Due to its simplicity, BCD type algorithms have been applied to a wide range of nonconvex problems (Bottou, 2010), including matrix and tensor decomposition problems such as nonnegative matrix factorization (Lee & Seung, 1999; 2001; Wang & Zhang, 2012) and nonnegative CANDECOMP/PARAFAC (CP) decomposition (Tucker, 1966; Harshman, 1970; Carroll & Chang, 1970). Notably, all these decomposition problems enjoy block multi-convex structure, wherein the objective function is convex when restricted on each block coordinate so that each convex sub-problems can be solved via standard convex optimization algorithms (Boyd et al., 2004). However, such multi-convexity is not

[1]Department of Mathematics, University of Seoul, Seoul, Republic of Korea [2]Department of Mathematics, University of Wisconsin - Madison, Wisconsin, United States. Correspondence to: Dohyun Kwon <dh.dohyun.kwon@gmail.com>, Hanbaek Lyu <hlyu@math.wisc.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

required to apply BCD, as simple coordinate-wise gradient descent can be applied to find the approximate minimizer of the sub-problems (Wright, 2015).

It is known that vanilla BCD ((3) with $\lambda_n \equiv 0$) does not always converge to the stationary points of the non-convex objective function that is convex in each block coordinate (Powell, 1973; Grippo & Sciandrone, 2000). It is known that BCD-PR with $\lambda_n \equiv Const.$ is guaranteed to converge to the set of stationary points (Grippo & Sciandrone, 2000). Under a more general condition, BCD-PR and its prox-linear variant are shown to converge to Nash equilibria. Local convergence result with rate is known for these algorithms under the stronger condition of Kurdyka-Łojasiewicz (Attouch et al., 2010; Xu & Yin, 2013; Bolte et al., 2014). For convex objectives, iteration complexity of $O(\varepsilon^{-1})$ is established in Hong et al. (2017). The BCD method has been drawing attention as an alternative method for training Deep Neural Network (DNN) models. In Zhang & Brand (2017), a BCD method is shown to converge to stationary points for Tikhonov regularized DNN models. In Zeng et al. (2019), BCD-PR for training DNNs with general activation functions is shown to have iteration complexity of $O(\varepsilon^{-1})$.

Contribution. While being one of the fundamental nonconvex optimization methods, the worst-case iteration complexity of BCD-PR (3) for general objectives under constraints has not been established in the literature. We intend to fill this gap with contributions summarized below:

- Global convergence to stationary points of BCD-PR for L-smooth objective f under constraints;
- Worst-case bound of $O(\varepsilon^{-1}(\log \varepsilon^{-1})^2)$ on the number of iterations to achieve ε-approximate stationary points;
- Robustness of the aforementioned results under inexact execution of the algorithm.

To our best knowledge, we believe our work provides the first result on the global rate of convergence and worst-case iteration complexity of BCD-PR for the general smooth objectives, especially with the additional robustness result. For gradient descent methods with unconstrained nonconvex objective, it is known that such rate of convergence cannot be faster than $O(\varepsilon^{-1})$ (Cartis et al., 2010), so our rate bound matches the optimal result up to a $(\log \varepsilon^{-1})^2$ factor. We emphasize that the above result does not claim that BCD-PR is provably faster than existing non-convex optimization algorithms. Instead, our novel analysis confirms that the classic and practical algorithm of BCD-PR is guaranteed to converge as fast as existing algorithms in the worst case.

The works (Attouch et al., 2010) and (Bolte et al., 2014) assume that the objective function satisfies KL property at every point in the parameter space and obtains a global rate of convergence to a stationary point for block proximal Gauss-Seidel (equivalent to our Algorithm 1) and block proximal alternating linearized minimization. On the other hand, Xu & Yin (2013) assumed local KL property and obtained a local rate of convergence to a stationary point for both types of BCD methods. In our work, we do not assume KL property at any point and still obtain a global convergence rate for block proximal Gauss-Seidel.

Application to Wasserstein CP-dictionary learning. In order to motivate our theoretical underpinning of BCD-PR, we consider the problem of Wasserstein CP-dictionary learning for d-dimensional joint distributions, which seeks a set of elementary probability distributions that can well-approximate a given set of d-dimensional joint probability distributions represented as d-mode tensors.

- We propose the *Wasserstein CP-dictionary learning* (WCPDL) framework for learning elementary probability distributions that reconstruct d-dimensional joint probability distributions represented as d-mode tensors.
- We propose an algorithm for WCPDL based on BCD-PR, where the sub-problems of Wasserstein reconstruction error minimization are handled by using entropic regularization and dual formulation for computational efficiency.
- We establish worst-case bound of $O(\varepsilon^{-1}(\log \varepsilon^{-1})^2)$ on the number of iterations to achieve ε-approximate stationary points for WCPDL.

We also demonstrate the advantage of the Wasserstein formulation for distribution-valued dictionary learning through a number of experiments and applications.

2. Preliminaries

Before stating our main results in the following sections, let us recall a list of definitions for (1). We say $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ is a *stationary point* of a function f over $\boldsymbol{\Theta}$ if

$$\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \langle \nabla f(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \geq 0, \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot project on $\mathbb{R}^{I_1 + \cdots + I_m} \supseteq \boldsymbol{\Theta}$. This is equivalent to saying that $-\nabla f(\boldsymbol{\theta}^*)$ is in the normal cone of $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}^*$. If $\boldsymbol{\theta}^*$ is in the interior of $\boldsymbol{\Theta}$, then it implies $\|\nabla f(\boldsymbol{\theta}^*)\| = 0$. For iterative algorithms, such a first-order optimality condition may hardly be satisfied exactly in a finite number of iterations, so it is more important to know how the worst-case number of iterations required to achieve an ε-approximate solution scales with the desired precision ε. More precisely, we say $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ is an *ε-approxiate*

stationary point of f over $\boldsymbol{\Theta}$ if

$$-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}^*), \frac{(\boldsymbol{\theta}-\boldsymbol{\theta}^*)}{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|} \right\rangle \leq \sqrt{\varepsilon}. \tag{5}$$

This notion of ε-approximate solution is consistent with the corresponding notion for unconstrained problems. Indeed, if $\boldsymbol{\theta}^*$ is an interior point of $\boldsymbol{\Theta}$, then (5) reduces to $\|\nabla f(\boldsymbol{\theta}^*)\|^2 \leq \varepsilon$. It is also equivalent to a similar notion in Def. 1 in Nesterov (2013), which is stated for non-smooth objectives using subdifferentials instead of gradients as in (5). Next, for each $\varepsilon > 0$ we define the *worst-case iteration complexity* N_ε of an algorithm computing $(\boldsymbol{\theta}_n)_{n\geq 1}$ for solving (1) as

$$N_\varepsilon := \sup_{\boldsymbol{\theta}_0\in\boldsymbol{\Theta}} \inf \left\{ n \,|\, \begin{matrix} \boldsymbol{\theta}_n \text{ is an } \varepsilon\text{-approximate} \\ \text{stationary point of } f \text{ over } \boldsymbol{\Theta} \end{matrix} \right\}, \tag{6}$$

where $(\boldsymbol{\theta}_n)_{n\geq 0}$ is a sequence of estimates produced by the algorithm with an initial estimate $\boldsymbol{\theta}_0$. Note that N_ε gives the *worst-case* bound on the number of iterations for an algorithm to achieve an ε-approximate solution due to the supremum over the initialization $\boldsymbol{\theta}_0$ in (6).

3. Statement of the results

We state the main result, Theorem 3.4. To our best knowledge, this gives the first worst-case rate of convergence and iteration complexity of BCD-type algorithms with proximal regularization in the literature. We impose the following two mild conditions for our theoretical analysis of BCD-PR (3).

Assumption 3.1. For each $i = 1, 2, \cdots, m$, there exists a constant $L^{(i)} > 0$ such that the function $f : \boldsymbol{\Theta} = \Theta^{(1)} \times \cdots \times \Theta^{(m)} \to [0, \infty)$ is $L^{(i)}$-smooth in each block coordinate i, that is, the function $\theta \mapsto \nabla f(\theta^{(1)}, \cdots, \theta^{(i-1)}, \theta, \theta^{(i+1)}, \cdots, \theta^{(m)})$ is $L^{(i)}$-Lipschitz in $\Theta^{(i)}$ for any $\theta^{(j)} \in \Theta^{(j)}, j = 1, 2, \cdots, i-1, i+1, \cdots, m$.

Assumption 3.2. The constraint sets $\Theta^{(i)} \subseteq \mathbb{R}^{I_i}$, $i = 1, \ldots, m$ are convex. Furthermore, the sub-level sets $f^{-1}((-\infty, a)) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : f(\boldsymbol{\theta}) \leq a\}$ are compact for each $a \in \mathbb{R}$.

We also allow an inexact computation of the solution to the sub-problem (3). For a quantitative statement, for each $n \geq 1$, we define the *optimality gap* Δ_n by

$$\Delta_n := \max_{1\leq i\leq m} \left(g_n^{(i)}(\theta_n^{(i)}) - \inf_{\theta\in\Theta^{(i)}} g_n^{(i)}(\theta) \right), \tag{7}$$

where $g_n^{(i)}$ is in (3). For our convergence results to hold, we require the optimality gaps to decay sufficiently fast so that they are summable:

Assumption 3.3. The optimality gaps Δ_n are summable, that is, $\sum_{n=1}^{\infty} \Delta_n < \infty$.

We now state our main result for BCD-PR.

Theorem 3.4. *Let $(\boldsymbol{\theta}_n)_{n\geq 0}$ be an inexct output of (3). Suppose that Assumptions 3.1-3.3 hold. Let $L^{(i)} > 0$ be such that ∇f is $L^{(i)}$-Lipschitz in each block coordinate and suppose the proximal regularizers $(\tau_n^{(i)})_{n\geq 1}$ satisfy $\tau_n^{(i)} > L^{(i)}$ for $n \geq 1$ and $\tau_n = O(1)$. Then the following hold:*

(i) *(Global convergence to stationary points) Every limit point of $(\boldsymbol{\theta}_n)_{n\geq 0}$ is a stationary point of f over $\boldsymbol{\Theta}$.*

(ii) *(Worst-case rate of convergence) There exists a constant M independent of $\boldsymbol{\theta}_0$ such that for $n \geq 1$,*

$$\min_{1\leq k\leq n} \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_k), \frac{(\boldsymbol{\theta}-\boldsymbol{\theta}_k)}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|} \right\rangle \right]^2 \leq \frac{M + 2m\sum_{n=1}^{\infty} \Delta_n}{n/(\log n)^2}. \tag{8}$$

(iii) *(Worst-case iteration complexity) Suppose the optimality gaps are uniformly summable, that is, $\sup_{\boldsymbol{\theta}_0\in\boldsymbol{\Theta}} \sum_{n=1}^{\infty} \Delta_n < \infty$. Then the worst-case iteration complexity N_ε for BCD-PR (3) satisfies $N_\varepsilon = O(\varepsilon^{-1}(\log \varepsilon^{-1})^2)$ if $\tau_n \equiv 1$.*

4. Application to d-dimensional Wasserstein dictionary learning

We apply our optimization method of BCD-PR (3) to solve *d-dimensional Wasserstein dictionary learning*, where the goal is to learn a dictionary of product probability distributions from a set of joint distributions. Namely, given d-dimensional joint probability distributions $(\mathbf{X}_k)_{1\leq k\leq N}$, we seek to find a set of product distributions such that each $\mathbf{X}_k$ can be approximated by a suitable mixture of the product distributions.

4.1. Dictionary learning for distribution-valued signals

For N observed d-mode tensor-valued signals $\mathbf{X}_1, \ldots, \mathbf{X}_N$ in $\mathbb{R}^{I_1\times\cdots\times I_d}$, we are interested in extracting r 'features' from this set, where each feature again takes the form of d-mode tensors in $\mathbb{R}^{I_1\times\cdots\times I_d}$. In other words, we seek to learn a 'dictionary' $\mathcal{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_r] \in \mathbb{R}^{I_1\times\cdots\times I_d\times r}$ of r 'atoms' so that each data tensor $\mathbf{X}_i$ can be linearly approximated by the atoms $\mathbf{D}_1, \ldots, \mathbf{D}_r$ in the dictionary $\mathcal{D}$. Namely, there exists a suitable 'code matrix' $\Lambda \in \mathbb{R}^{r\times N}$ such that we have the following approximate factorization:

$$\begin{aligned} [\mathbf{X}_1, \ldots, \mathbf{X}_N] &\approx [\mathbf{D}_1, \ldots, \mathbf{D}_r] \times_{d+1} \Lambda \\ \iff \quad \mathcal{X} &\approx \mathcal{D} \times_{d+1} \Lambda, \end{aligned} \tag{9}$$

where $\times_{d+1}$ denotes the mode $(d+1)$ tensor-matrix product (see (Kolda & Bader, 2009)) and $\mathcal{X} := [\mathbf{X}_1, \ldots, \mathbf{X}_n]$ denotes the $(d+1)$-mode tensor in $\mathbb{R}^{I_1\times\cdots\times I_d\times N}$ that concatenates the tensor-valued signals $\mathbf{X}_1, \ldots, \mathbf{X}_n$ in $\mathbb{R}^{I_1\times\cdots\times I_d}$

along the last mode. As a special case, suppose $d = 1$ so that the signals $\mathbf{X}_1, \dots, \mathbf{X}_n$ are in fact I_1-dimensional vectors. Then (9) becomes the usual matrix factorization formulation for factorizing the data matrix $\mathcal{X} \in \mathbb{R}^{I_1 \times N}$ into the (matrix) product of a dictionary matrix $\mathcal{D} \in \mathbb{R}^{I_1 \times r}$ and the code matrix $\Lambda \in \mathbb{R}^{r \times N}$ (Lee & Seung, 1999; Elad & Aharon, 2006; Mairal et al., 2007; Peyré, 2009).

As a more precise optimization formulation of (9), we consider

$$\min_{\mathcal{D} \in \mathbb{R}^{I_1 \times \cdots \times I_d \times r}, \Lambda \in \mathbb{R}^{r \times N}} \delta\Big(\mathcal{X},\, \mathcal{D} \times_{d+1} \Lambda\Big), \tag{10}$$

where $\delta : (\mathbb{R}^{I_1 \times \cdots \times I_d \times N})^2 \to [0, \infty)$ is a 'dissimilarity function' that maps a pair of tensors $(\mathcal{X}, \mathcal{X}')$ to a nonnegative number $\delta(\mathcal{X}, \mathcal{X}')$. This function is used to measure the difference between the data tensor $\mathcal{X}$ and the 'reconstruction' $\mathcal{D} \times_{d+1} \Lambda$. For $d = 1$, standard choices of δ include the distance function induced by the Frobenius norm and the KL divergence.

4.2. Wasserstein distance between d-dimensional probability distributions

A natural notion of dissimilarity between two probability distributions on the same probability space is the *p-Wasserstein distance*, which is a central notion in this paper, which we will define below.

Define the *cost tensor* $\mathbf{M} \in \mathbb{R}^{I_1 \times \cdots \times I_d} \times \mathbb{R}^{I_1 \times \cdots \times I_d}$ for d-mode tensors to be the tensor defined by $\mathbf{M}(J_1, J_2) = \|J_1 - J_2\|_2$ for all multi-indices $J_1, J_2 \in [I_1] \times \cdots \times [I_d]$.

One can regard $\mathbf{M}$ as giving weights on the difference between the J_1- and the J_2-entry of two tensors. For instance, if $d = 1$, then the dissimilarity between the two random variables Y_1 and Y_2 depends not only on the probability that they differ but also on the actual value $|Y_1 - Y_2|$. The cost matrix $\mathbf{M}$, in this case, measures the probabilistic 'cost' of having different probability mass on coordinates J_1 and J_2. Next, for two one-dimensional probability mass functions $p_1 \in \mathbb{R}^m$, $p_2 \in \mathbb{R}^n$, we call a two-dimensional joint distribution $T \in \Sigma_{m,n}$ a *coupling* between p_1 and p_2 if its row (resp., column) sums agree with p_1 (resp., p_2). We denote by

$$U(p_1, p_2) := \Bigg\{ T \in \Sigma_{m,n} \,\Bigg|\, p_1(i) = \sum_{j=1}^{n} T(i,j), p_2(j) = \sum_{i=1}^{m} T(i,j)\, \forall i \in \{1, \dots, m\},\, j \in \{1, \dots, n\} \Bigg\}$$

the set of all couplings between p_1 and p_2.

Now, we can define the Wasserstein distance. Fix a cost tensor $\mathbf{M} \in \mathbb{R}^{I_1 \times \cdots \times I_d} \times \mathbb{R}^{I_1 \times \cdots \times I_d}$ and let $\mathbf{M}^2 \in \mathbb{R}^{(I_1 \cdots I_d) \times (I_1 \cdots I_d)}$ denote its matricization (see (Kolda & Bader, 2009)). Fix a parameter $\gamma \geq 0$. For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{I_1 \times \cdots \times I_d}$, define

$$\begin{aligned} W_\gamma(\mathbf{A}, \mathbf{B}) &:= W_\gamma(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\mathbf{B})) \\ &:= \min_{T \in U(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\mathbf{B}))} \left\langle \mathbf{M}^2, T \right\rangle + \gamma \left\langle T, \log T \right\rangle, \end{aligned} \tag{11}$$

where $\mathrm{vec}(\mathbf{A})$ and $\mathrm{vec}(\mathbf{B})$ denote the vectorization of $\mathbf{A}$ and $\mathbf{B}$, respectively. When $\gamma = 0$, W_γ above is known as the Wasserstein distance. The additional term $\gamma\langle T, \log T\rangle$ is known as the *entropic regularization* of Wasserstein distance (Cuturi, 2013).

4.3. d-dimensional Wasserstein dictionary learning

We are interested in the case that the tensor-valued signals $\mathbf{X}_1, \dots, \mathbf{X}_N$ describe d-dimensional probability mass functions. Namely, we denote

$$\Sigma_{I_1, \dots, I_d} := \left\{ \mathbf{X} \in \mathbb{R}^{I_1 \times \cdots \times I_d}_{\geq 0} \,\middle|\, \sum_{i_1, \dots, i_d} \mathbf{X}[i_1, \dots, i_d] = 1 \right\}.$$

We can think of an element $\mathbf{X}$ of $\Sigma_{I_1, \dots, I_d}$ as the joint probability mass function of d discrete random variables $(Y_1, \dots, Y_d)$ where each Y_i takes values from $\{1, \dots, I_i\}$. For this reason, we will call an element of $\Sigma_{I_1, \dots, I_d}$ simply as a 'd-dimensional joint distribution'. We also denote by $\Sigma^N_{I_1, \dots, I_d}$ the N-fold product of $\Sigma_{I_1, \dots, I_d}$, which we identify as a subset of $\mathbb{R}^{I_1 \times \cdots \times I_d \times N}$ in the usual way.

When each d-mode tensor $\mathbf{X}_i$ subject to the factorization in (10) is a d-dimensional joint distribution, then the dissimilarity function δ in (10) should measure the dissimilarity between two tuples of d-dimensional joint distribution. By using the *entropy-regularized Wasserstein distance* W_γ (see (11)), we formulate the *d-dimensional Wasserstein Dictionary Learning* (dWDL) as (9), where the dictionary atoms $\mathbf{D}_1, \dots, \mathbf{D}_r$ are taken to be d-dimensional joint distributions (elements of $\Sigma_{I_1, \dots, I_d}$) and the dissimilarity function $\delta : \Sigma^N_{I_1, \dots, I_d} \times \Sigma^N_{I_1, \dots, I_d} \to [0, \infty)$ is

$$\delta([\mathbf{X}_1, \dots, \mathbf{X}_N],\, [\mathbf{X}'_1, \dots, \mathbf{X}'_N]) := \sum_{i=1}^{N} W_\gamma(\mathbf{X}_i, \mathbf{X}'_i).$$

Equivalently, we formulate our problem (dWDL) as below:

$$\textbf{(dWDL)} \qquad \min_{\substack{\mathcal{D} = [\mathbf{D}_1, \dots, \mathbf{D}_r] \in \Sigma^r_{I_1, \dots, I_d} \\ \Lambda \in \Sigma^N_r}} f_W(\mathcal{D}, \Lambda), \tag{12}$$

$$\text{where } f_W(\mathcal{D}, \Lambda) := \sum_{i=1}^{N} W_\gamma\left(\mathbf{X}_i,\, \mathcal{D} \times_{d+1} \Lambda[:, i]\right).$$

For $d = 1$, this formulation (12) has been discussed in the study of Wasserstein dictionary learning, including (Sandler & Lindenbaum, 2011), (Zen et al., 2014), and (Rolet et al., 2016).

4.4. Algorithm (dWDL)

Given the previous estimate $(\Lambda_{n-1}, \mathcal{D}_{n-1})$, we compute the updated estimate $(\Lambda_n, \mathcal{D}_n)$ by solving convex sub-problems as follows:

$$\Lambda_n \in \underset{\Lambda \in \Sigma_r^N}{\arg\min}\ f_W(\mathcal{D}_{n-1}, \Lambda) + \frac{\tau_n}{2}\|\Lambda - \Lambda_{n-1}\|_F^2 \tag{13}$$

$$\mathcal{D}_n \in \underset{\mathcal{D} \in \Sigma^r_{I_1 \times \cdots \times I_d}}{\arg\min}\ f_W(\mathcal{D}, \Lambda_n) + \frac{\tau_n}{2}\|\mathcal{D} - \mathcal{D}_{n-1}\|_F^2. \tag{14}$$

For the standard nonnegative matrix factorization using the Frobenius norm instead of the Wasserstein norm, solving the corresponding convex sub-problems amounts to solving standard nonnegative least squares problem, which can be done by applying standard projected gradient descent. However, solving convex sub-problems in (13) and (14) is computationally demanding since one is required to compute N Wasserstein distances W_γ, each of which involves finding an optimal transport plan by solving a separate optimization problem. Below, we propose a computationally efficient algorithm where one is only required to solve a single and simple subproblem (instead of N) for each block coordinate descent step.

Algorithm 1 dWDL (12)

1: **Input:** $\boldsymbol{\theta}_0 = (\mathcal{D}_0, \Lambda_0) \in \Sigma^r_{I_1 \times \cdots \times I_d} \times \Sigma_r^N$ (initial estimate); N (number of iterations); $(\tau_n)_{n \geq 1}$, (non-decreasing sequence in $[1, \infty)$);
2: **for** $n = 1, \ldots, N-1$ **do**:
3: Update estimate $\boldsymbol{\theta}_{n-1} = (\mathcal{D}_{n-1}, \Lambda_{n-1})$ by

$$\Lambda_n \leftarrow \text{Algorithm 2 with input } (\mathcal{D}_{n-1}, \Lambda_{n-1}) \tag{15}$$
$$\mathcal{D}_n \leftarrow \text{Algorithm 3 with input } (\mathcal{D}_{n-1}, \Lambda_n) \tag{16}$$

4: **end for**
5: **output:** $\boldsymbol{\theta}_N$

We now describe Algorithms 2 and 3 that solve the convex sub-problems in (13) and (14). To solve the primal problem (13), we consider its dual problem. For simplicity, denote the distance function and the proximal term by for $\mathbf{X}, y \in \Sigma_{I_1 \times \cdots \times I_d}$ and for given $\lambda_0 \in \Sigma_r$,

$$H_{\mathbf{X}}(y) := W_\gamma(\mathbf{X}, y) \text{ and}$$
$$F_{\lambda_0}(\lambda) := \begin{cases} \frac{1}{2}\|\lambda - \lambda_0\|_F^2 & \text{for } \lambda \in \Sigma_r, \\ +\infty & \text{otherwise .} \end{cases} \tag{17}$$

Then, the primal problem (13) can be re-written as

$$\min_{\Lambda \in \mathbb{R}^{r \times N}} \sum_{i=1}^N \{H_{\mathbf{X}_i}(\mathcal{D}_{n-1} \times_{d+1} \Lambda[:, i]) + \tau_n F_{\Lambda_{n-1}[:,i]}(\Lambda[:, i])\}. \tag{18}$$

Here, the condition $\Lambda \in \Sigma_r^N$ is enforced by F in the second term.

Note that the above is a convex minimization problem but solving it directly is computationally expensive since simply evaluating the function $H_{\mathbf{X}_i}$ above involves finding an optimal transport map $T \in U(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\mathbf{B}))$. In order to overcome this issue, we consider the dual problem of (18) reminiscent of Cuturi (2013). Introducing a dual variable $G \in \mathbb{R}^{I_1 \times \cdots \times I_d \times N}$, we obtain the dual problem:

$$\min_{G \in \mathbb{R}^{I_1 \times \cdots \times I_d \times N}} \sum_{i=1}^N \{H^*_{\mathbf{X}_i}(-G[:, i]) + \tau_n F^*_{\Lambda_{n-1}[:,i]}(\mathcal{D}_{n-1} \times_{\leq d} G[:, i]/\tau_n)\}. \tag{19}$$

Here, the *conjugate* f^* of f is defined as

$$f^* : \mathbb{R}^d \to [-\infty, +\infty] : u \mapsto \sup_x(\langle x, u\rangle - f(x)). \tag{20}$$

This dual problem can be solved without having to deal with a matrix-scaling problem, as in the primal one (see (Cuturi & Peyré, 2016)). We postpone further discussion about the conjugate functions H^* and F^* to the subsequent sections.

Algorithm 2 Solving for Λ

1: **Input:** $\boldsymbol{\theta}_{n-1} = (\mathcal{D}_{n-1}, \Lambda_{n-1}) \in \Sigma^r_{I_1 \times \cdots \times I_d} \times \Sigma_r^N$ (current estimate); $(\tau_n)_{n \geq 1}$;
2: Update estimate Λ_{n-1} by

$$G_n^\circ \leftarrow \text{ the minimizer of (19)}$$
$$\Lambda_n \leftarrow \left(\Lambda_{n-1} + \frac{\mathcal{D}_{n-1} \times_{\leq d} G_n^\circ}{\tau_n} - J^\circ \otimes c_n^\circ\right)_+$$

where $c_n^\circ \in \mathbb{R}^{N \times 1}$ is chosen to satisfy $\Lambda_n \in \Sigma_r^N$ and all entries of $J^\circ \in \mathbb{R}^{r \times 1}$ are one.
3: **output:** $\boldsymbol{\theta}_{n-\frac{1}{2}} = (\mathcal{D}_{n-1}, \Lambda_n)$

Here, the $1, 2, \cdots, d$-mode product $\mathcal{D} \times_{\leq d} \Lambda$ of $\mathcal{D} \in \mathbb{R}^{I_1 \times \cdots \times I_d \times N}$ with a tensor $\Lambda \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_d \times J}$ is

$$(\mathcal{D} \times_{\leq d} \Lambda)[j] := \sum_{i_1, i_2, \cdots, i_d} \mathcal{D}[i_1, i_2, \cdots, i_d] \times \Lambda[i_1, i_2, \cdots, i_d, j]. \tag{21}$$

Based on similar arguments, the dual problem of (14) can be derived as follows:

$$\min_{G \in \mathbb{R}^{I_1 \times \cdots \times I_d \times N}} \left\{ \left(\sum_{i=1}^N H^*_{\mathbf{X}_i}(-G[:, i]) \right) + \tau_n F^*_{\mathcal{D}_{n-1}}(G \times_{d+1} \Lambda_n^T/\tau_n) \right\}. \tag{22}$$

Algorithm 3 Solving for $\mathcal{D}$

1: **Input:** $\boldsymbol{\theta}_{n-\frac{1}{2}} = (\mathcal{D}_{n-1}, \Lambda_n) \in \Sigma^r_{I_1 \times \cdots \times I_d} \times \Sigma^N_r$ (current estimate); $(\tau_n)_{n \geq 1}$;
2: Update estimate $\mathcal{D}_{n-1}$ by

$$G_n^\dagger \leftarrow \text{ the minimizer of (22)}$$
$$\mathcal{D}_n \leftarrow \left(\mathcal{D}_{n-1} + \frac{G_n^\dagger \times_{d+1} \Lambda_n^T}{\tau_n} - J^\dagger \otimes c_n^\dagger \right)_+$$

where $c_n^\dagger \in \mathbb{R}^{r \times 1}$ is chosen to satisfy $\mathcal{D}_n \in \Sigma^r_{I_1 \times \cdots \times I_d}$ and all entries of $J^\dagger \in \mathbb{R}^{I_1 I_2 \cdots I_d \times 1}$ are one.
3: **output:** $\boldsymbol{\theta}_n = (\mathcal{D}_n, \Lambda_n)$

The per-iteration cost of Algorithms 2 and 3 is given by $O((I_1 \dots I_d)^2 N)$.

5. Theoretical guarantees of Wasserstein dictionary learning

We prove that our computationally efficient algorithm, Algorithm 1, is actually solving BCD with proximal regularization for our main problem (12). The proof of Theorem 5.1 can be found in Appendix B.

Theorem 5.1. *(Per-iteration correctness) Algorithm 1 solves (13) and (14).*

Formally speaking, the dual problem (19) is derived from the primal problem (18) as follows: for given $(\mathcal{D}_{n-1}, \Lambda_{n-1}) \in \Sigma^r_{I_1 \times \cdots \times I_d} \times \Sigma^N_r$ and $\tau_n > 0$,

$$\begin{aligned}
&\min_{\Lambda \in \Sigma_r^N} H_{\mathbf{X}_i}\left(\mathcal{D}_{n-1} \times_{d+1} \Lambda[:,i]\right) + \tau_n F_{\Lambda_{n-1}[:,i]}(\Lambda[:,i]), \\
&= \min_{\substack{\Lambda \in \Sigma_r^N, \\ Q \in \Sigma^N_{I_1 \times \cdots \times I_d}}} \max_{G \in \mathbb{R}^{I_1 \times \cdots \times I_d \times N}} H_{\mathbf{X}_i}(Q[:,i]) \\
&\quad + \tau_n F_{\Lambda_{n-1}[:,i]}(\Lambda[:,i]) \\
&\quad + \langle Q[:,i] - \mathcal{D}_{n-1} \times_{d+1} \Lambda[:,i], G[:,i] \rangle, \\
&= - \min_{G \in \mathbb{R}^{I_1 \times \cdots \times I_d \times N}} H^*_{\mathbf{X}_i}(-G[:,i]) \\
&\quad + \tau_n F^*_{\Lambda_{n-1}[:,i]}(\mathcal{D}_{n-1} \times_{\leq d} G_n[:,i]/\tau_n).
\end{aligned}$$

The above derivation is standard in the classical theory of convex optimization. However, solving Algorithm 1 requires us to find the optimizers of the primal problem (13) and (14) in terms of the inputs and their dual solutions. Due to the constraints, $\mathcal{D} \in \Sigma^r_{I_1 \times \cdots \times I_d}$ and $\Lambda \in \Sigma^N_r$, this does not directly follows.

To establish the correctness rigorously, we consider a general minimization problem of a bivariate function under inequality constraints in Lemma B.4: for given functions $f : \mathcal{K} \to (-\infty, +\infty]$, $h : \mathcal{H} \to (-\infty, +\infty]$, and $R : \mathcal{H} \to \mathcal{K}$,

$$\min_{x \in \mathcal{H}, Rx \in K} f(Rx) + h(x). \tag{23}$$

Here, $\mathcal{H}$ and $\mathcal{K}$ are real Hilbert spaces with inner product $\langle \cdot, \cdot \rangle$, and K is a nonempty closed convex cone in $\mathcal{K}$. The key idea is based on Propositions 19.18 and 19.23 in Bauschke et al. (2011), but we provide the proof in Appendix B for the sake of completeness.

Now we can obtain a convergence and complexity result for Algorithm 1 using Theorems 5.1 and3.4.

Theorem 5.2. *Suppose that Assumption 3.3 holds, the proximal regularizers $(\tau_n)_{n \geq 1}$ satisfy $\tau_n > 1/\gamma$ for $n \geq 1$ and $\tau_n = O(1)$. For a output $(\boldsymbol{\theta}_n)_{n \geq 0}$ of Algorithm 1, the following hold:*

(i) *(Global convergence to stationary points) Every limit point of $(\boldsymbol{\theta}_n)_{n \geq 0}$ is a stationary point of f_W over $\boldsymbol{\Theta} := \Sigma^r_{I_1 \times \cdots \times I_d} \times \Sigma^N_r$.*

(ii) *(Worst-case rate of convergence) There exists a constant M independent of $\boldsymbol{\theta}_0$ such that for $n \geq 1$, (8) in Theorem 3.4 holds.*

(iii) *(Worst-case complexity) The worst-case iteration complexity N_ε for Algorithm 1 satisfies $N_\varepsilon = O(\varepsilon^{-1}(\log \varepsilon^{-1})^2)$. Furthermore, the worst-case complexity of Algorithm 1 is*

$$\begin{aligned}
&O(N_\varepsilon \cdot (\text{worst-case cost of solving sub-problems})) \\
&\quad = O(N_\varepsilon \cdot \log N_\varepsilon \cdot (\text{cost of PGD step for dual})) \\
&\quad = O(\varepsilon^{-1}(\log \varepsilon^{-1})^3 (I_1 \times \cdots \times I_d)^2 N).
\end{aligned}$$

Proof of Theorem 5.2. Let us first show that Algorithm 1 satisfies Assumptions 3.1, and 3.2. Then, (i) and (ii) follow from Theorem 3.4. The conjugate function of $H_{\mathbf{X}}$ given in (17) has a closed form (Cuturi & Peyré, 2016): for $g \in \mathbb{R}^{I_1 \times \cdots \times I_d}$ and given $\mathbf{X} \in \Sigma_{I_1 \times \cdots \times I_d}$,

$$\begin{aligned}
H^*_{\mathbf{X}}(g; \Sigma_{I_1 \times \cdots \times I_d}) &:= \sup_{y \in \Sigma_{I_1 \times \cdots \times I_d}} \langle g, y \rangle - H_{\mathbf{X}}(y), \\
&= \gamma \left(\langle \mathbf{X}, \log \mathbf{X} \rangle + \langle \mathbf{X}, \ \log(K\alpha) \rangle \right).
\end{aligned}$$

Here, $K = \exp(-M/\gamma) \in \left(\mathbb{R}^{I_1 \times \cdots \times I_d}\right)^2$, $\alpha = \exp(g/\gamma) \in \mathbb{R}^{I_1 \times \cdots \times I_d}$, and $M \in \left(\mathbb{R}^{I_1 \times \cdots \times I_d}\right)^2$ is a given cost matrix. It is known from Theorem 2.4 in Cuturi & Peyré (2016) that this dual function is C^∞. In addition, its gradient function is $1/\gamma$ Lipschitz, and it is explicitly given as

$$\nabla H^*_{\mathbf{X}}(g) = \alpha \circ \left(K \frac{\mathbf{X}}{K\alpha} \right) \in \Sigma_{I_1 \times \cdots \times I_d}. \tag{24}$$

Therefore, Assumption 3.1 is satisfied. Furthermore, the constraint set $\Sigma^r_{I_1\times\cdots\times I_d}$ and Σ^N_r satisfy Assumption 3.2.

Next, we compute the per-iteration cost of Algorithms 2 and 3. The dual function of F_{λ_0} is given by for $g \in \mathbb{R}^r$

$$F^*_{\lambda_0}(g) := \sup_{\lambda\in\Sigma_r} \langle g, \lambda\rangle - \frac{1}{2}\|\lambda - \lambda_0\|_F^2.$$

From Lemma D.1, the optimizer of the above is given as

$$\lambda^* = (g + \lambda_0 - c1_r)_+ \tag{25}$$

where c is a constant chosen to satisfy $\lambda \in \Sigma_r$, and thus

$$F^*_{\lambda_0}(g) = \frac{1}{2}(g + \lambda_0 - c1_r)_+(g + \lambda_0 + c1_r) - \frac{1}{2}\|\lambda_0\|_F^2.$$

By the duality as in Lem. 7.15 in Santambrogio (2015), its gradient is given as the optimizer (25): $\nabla F^*_{\lambda_0}(g) = \lambda = (g + \lambda_0 - c1_r)_+ \in \Sigma_r$. Therefore, each gradient descent step to solve (19) or (22) requires $O((I_1 \dots I_d)^2 N)$. Lastly, (19) and (22) are convex problems, we conclude (iii). □

6. Extension to Wasserstein CP-dictionary learning

While it is possible to vectorize general d-mode tensor-valued signals to reduce to the case of dictionary learning for vector-valued signals, it would be more beneficial to tailor the d-dimensional dictionary learning problem (10) to exploit particular tensor structures that one desires to respect. One such approach is to constrain further the type of dictionary atoms $\mathbf{D}_1, \dots, \mathbf{D}_r$ that we allow. Namely, the CONDECOMP/PARAFAC (CP)-dictionary learning (Lyu et al., 2020) assumes that each $\mathbf{D}_i$ is a rank-1 tensor in the sense that it is the outer product of some 1-dimensional vectors. Also, exploiting Tucker-decomposition structure on the dictionary atoms has been studied recently in Shakeri et al. (2016); Ghassemi et al. (2017).

6.1. Wasserstein CP-dictionary learning

Suppose a data tensor $\mathbf{X} \in \mathbb{R}^{I_1\times\cdots\times I_d}$ is given and fix an integer $r \geq 1$. In the CANDECOMP/PARAFAC (CP) decomposition of $\mathbf{X}$ (Kolda & Bader, 2009), we would like to find r *loading matrices* $U^{(i)} \in \mathbb{R}^{I_i\times r}$ for $i = 1, \dots, d$ such that the sum of the outer products of their respective columns approximate $\mathbf{X}$:

$$\mathbf{X} \approx \sum_{k=1}^r \bigotimes_{i=1}^d U^{(i)}[:, k] =: [\![U^{(1)}, U^2, \dots, U^{(d)}]\!]$$

where $U^{(i)}[:, k]$ denotes the k^{th} column of the $I_i \times r$ loading matrix matrix $U^{(i)}$ and $\bigotimes$ denotes the outer product. We have also introduced the bracket operation $[\![\cdot]\!]$.

As an optimization problem, the above CP decomposition model can be formulated as the following the *constrained CP-decomposition* problem:

$$\operatorname*{arg\,min}_{U^{(1)}\in\Theta^{(1)},\dots,U^{(d)}\in\Theta^{(d)}} f_{\text{CP}}(U^{(1)}, \dots, U^{(d)}) \tag{26}$$

where

$$f_{\text{CP}}(U^{(1)}, \dots, U^{(d)}) := \left\|\mathbf{X} - [\![U^{(1)}, U^2, \dots, U^{(d)}]\!]\right\|_F^2$$

and $\Theta^{(i)} \subseteq \mathbb{R}^{I_i\times r}$ denotes a compact and convex constraint set and $\lambda_i \geq 0$ is a ℓ_1-regularizer for the i^{th} loading matrix $U^{(i)}$ for $i = 1, \dots, d$. In particular, by taking $\lambda_i = 0$ and $\Theta^{(i)}$ to be the set of nonnegative $I_i \times r$ matrices with bounded norm for $i = 1, \dots, d$, (26) reduces to the *nonnegative CP decomposition* (NCPD) (Shashua & Hazan, 2005; Zafeiriou, 2009). Also, it is easy to see that f_{CP} is equal to

$$\left\|\mathbf{X} - \texttt{Out}(U^{(1)}, \dots, U^{(d-1)}) \times_d (U^{(d)})^T\right\|_F^2, \tag{27}$$

which is the *CP-dictionary-learning* problem introduced in Lyu et al. (2020). Here $\times_d$ denotes the mode-d product (see (Kolda & Bader, 2009)) the outer product of loading matrices $U^{(1)}, \dots, U^{(m)}$ is defined as

$$\texttt{Out}(U^{(1)}, \dots, U^{(d)}) := \left[\bigotimes_{k=1}^d U^{(k)}[:, 1], \bigotimes_{k=1}^d U^{(1)}[:, 2], \dots, \bigotimes_{k=1}^d U^{(k)}[:, r]\right] \tag{28}$$

Namely, we can think of the d-mode tensor $\mathbf{X}$ as I_d observations of $(d-1)$-mode tensors, and the R rank-1 tensors in $\texttt{Out}(U^{(1)}, \dots, U^{(d)})$ serve as dictionary atoms, whereas the transpose of the last loading matrix $U^{(d)}$ can be regarded as the code matrix.

The Wasserstein formulation of the *CP-dictionary-learning* problem (26) is given as follows. As in the setting of (12), we suppose that each d-mode tensor $\mathbf{X}_i$ is a d-dimensional joint distribution. We aim to represent each data tensor X_i based on the product distributions of d one-dimensional distributions, $U^{(i)} \in \Sigma^r_{I_i}$ for $i = 1, \cdots, d$:

$$[\mathbf{X}_1, \dots, \mathbf{X}_N] \approx \texttt{Out}(U^{(1)}, \dots, U^{(d)}) \times_{d+1} \Lambda \tag{29}$$

for some code matrix $\Lambda \in \Sigma^N_r$ where $\texttt{Out}$ is given in (28). Comparing the Wasserstein distance between each X_i and the corresponding distribution, we formulate our main problem of Wasserstein CP-dictionary Learning (WCPDL):

$$\operatorname*{arg\,min}_{\substack{U^{(1)}\in\Sigma^r_{I_1},\dots,U^{(d)}\in\Sigma^r_{I_d},\\ \Lambda\in\Sigma^N_r}} f_{\text{WCP}}(U^{(1)}, \dots, U^{(d)}, \Lambda) \tag{30}$$

where

$$f_{\text{WCP}}(U^{(1)}, \dots, U^{(d)}, \Lambda) := \sum_{i=1}^N W_\gamma\left(\mathbf{X}_i, \texttt{Out}(U^{(1)}, \dots, U^{(d)}) \times_{d+1} \Lambda[:, i]\right).$$

Algorithm 4 WCPDL (30)

Input: $\boldsymbol{\theta}_0 = (U_0^{(1)}, \dots, U_0^{(d)}, \Lambda_0) \in \Sigma_{I_1}^r \times \cdots \times \Sigma_{I_d}^r \times \Sigma_r^N$ (initial estimate); N (number of iterations); $(\tau_n)_{n\geq 1}$, (non-decreasing sequence in $[1, \infty)$);
  for $n = 1, \dots, N-1$ **do**:
    Update estimate $\boldsymbol{\theta}_{n-1} = (U_{n-1}^{(1)}, \dots, U_{n-1}^{(d)}, \Lambda_{n-1})$ by

$$\mathcal{D} \leftarrow \texttt{Out}(U_{n-1}^{(1)}, \dots, U_{n-1}^{(d)})$$
$$\Lambda_n \leftarrow \text{Output of Algorithm 2 with input } (\mathcal{D}, \Lambda_{n-1})\ ;$$

    for $k = 1, \dots, d$ **do**:
      Update estimate $U_{n-1}^{(k)}$ by

$$\overline{\Lambda} \leftarrow \texttt{Out}(U_n^{(1)}, \dots, U_n^{(k-1)}, U_{n-1}^{(k+1)}, \dots, U_{n-1}^{(d)}, \Lambda_n^T)$$
$$\overline{\Lambda} \leftarrow \text{ Inserting the last mode of } \overline{\Lambda} \text{ into the } k\text{th mode}$$
$$U_n^{(k)} \leftarrow \text{Output of Algorithm 3 with input } (U_{n-1}^{(k)}, \overline{\Lambda})$$

    end for
  end for
output: $\boldsymbol{\theta}_N$

6.2. Algorithm (WCPDL)

We state our algorithm to solve Wasserstein CP-dictionary Learning (30). Given the previous estimates $U_{n-1}^{(1)}, \dots, U_{n-1}^{(d)}$ and Λ_{n-1}, we compute the updated estimate $U_n^{(1)}, \dots, U_n^{(d)}$ and Λ_n by solving convex subproblems, iteratively, as follows.

First, let $\mathcal{D}_{n-1}$ be $\texttt{Out}(U_{n-1}^{(1)}, \dots, U_{n-1}^{(d)}) \in \Sigma^r_{I_1 \times I_2 \times \cdots \times I_d}$. For a given data tensor $\mathbf{X} \in \Sigma^N_{I_1 \times I_2 \times \cdots \times I_d}$, $\tau_n > 0$, and the previous estimates above, the code matrix is updated as follows:

$$\Lambda_n \in \underset{\Lambda \in \Sigma_r^N}{\arg\min} \left(\sum_{i=1}^N W_\gamma \left(\mathbf{X}_i, \left(\mathcal{D}_{n-1} \times_{d+1} \Lambda \right)[:, i] \right) \right) + \frac{\tau_n}{2} \|\Lambda - \Lambda_{n-1}\|_F^2. \tag{31}$$

Next, for each $k \in \{1, 2, \cdots, d\}$, let $\overline{\Lambda} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_{k-1} \times r \times I_{k+1} \times \cdots \times I_d \times N}$ be obtained from

$$\texttt{Out}(U_n^{(1)}, \dots, U_n^{(k-1)}, U_{n-1}^{(k+1)}, \dots, U_{n-1}^{(d)}, \Lambda_n^T)$$

in $\mathbb{R}^{I_1 \times I_2 \times \cdots \times I_{k-1} \times I_{k+1} \times \cdots \times I_d \times N \times r}$ by inserting the last mode into the kth mode. Given $\overline{\Lambda}$, the dictionaries are updated as follows:

$$U_n^{(k)} \in \underset{U \in \Sigma_{I_k}^r}{\arg\min} \left(\sum_{i=1}^N W_\gamma \left(\mathbf{X}_i, \overline{\Lambda}[:, i] \times_k U^T \right) \right) + \frac{\tau_n}{2} \|U^{(k)} - U_{n-1}^{(k)}\|_F^2. \tag{32}$$

Theorem 6.1. *(Per-iteration correctness) Algorithm 4 solves (31) and (32).*

7. Experiments

7.1. Wasserstein barycenter problem

We first provide the simplest example when $r = 1$. In this case, $\Lambda \in \Sigma_1^N$ and thus all entries of Λ are 1's, which corresponds to the Wasserstein barycenter problem with equal weights: $\min_{\mathbf{D} \in \Sigma_{I_1, \dots, I_d}} \sum_{i=1}^N W_\gamma (\mathbf{X}_i, \mathbf{D})$.

For data living in the space of probability distributions, using the Wasserstein metric instead of the Euclidean metric may provide a better representation. Figure 1 provides the barycenter with respect to Wasserstein distance and the Frobenius norm when $d = 1$, $r = 1$, and $N = 3$.

As shown in the figure, the Wasserstein barycenter of three Gaussian distributions is close to the Gaussian distribution, while the Frobenius one is given as the vertical average of three distributions, which shows a significant difference between the two formulations.

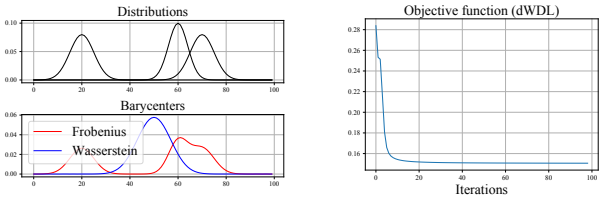

Figure 1. Finding the barycenter of three Gaussian distributions with respect to Wasserstein distance and the Frobenius norm

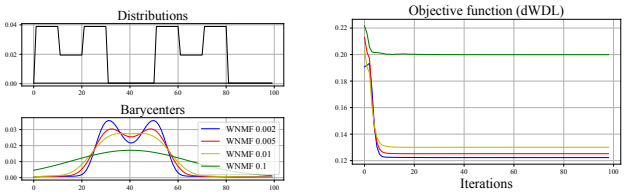

Figure 2. Finding the barycenter of two ⊔-shaped distributions with respect to Wasserstein distance for different γ's

As defined in (11), the regularized Wasserstein distance W_γ depends on the parameter $\gamma > 0$. In Figure 2, we solve the Wasserstein barycenter problem for different γ's and two ⊔-shaped distributions. While two peaks appear in $\gamma = 0.002$ and $\gamma = 0.005$, the distribution is getting close to Gaussian. This illustrates the importance of choosing appropriate γ to find out the geometric property of data sets.

7.2. Wasserstein dictionary learning

The additional knowledge of the underlying spaces can be utilized in Wasserstein dictionary learning. To illustrate this, we consider a sequence of figures generated by John Conway's Game of Life, which has a periodic domain. We solve the problems of Wasserstein dictionary learning with

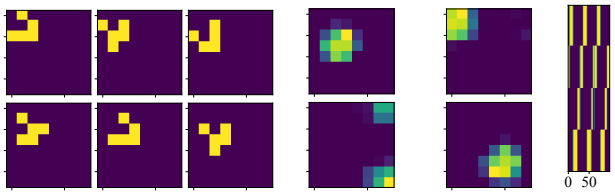

Figure 3. Wasserstein dictionary learning with $r = 4$, $N = 100$, and the Euclidean distance; a sequence of images (left), dictionaries (middle), code matrices (right)

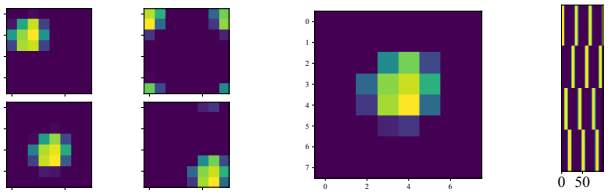

Figure 4. Wasserstein dictionary learning with $r = 4$, $N = 100$, and the distance on a torus; dictionaries (left), the translated top right dictionary (middle) code matrices (right)

two different ground metrics: the usual Euclidian distance in Figure 3 and the distance on a torus in Figure 4. It can be seen in Figure 4 that all dictionaries are similar up to translations.

The results for Wasserstein dictionary learning on MNIST for different r's are given as follows.

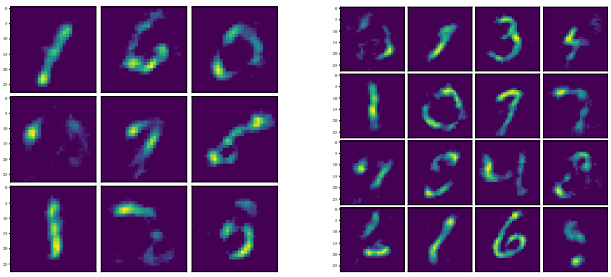

Figure 5. Wasserstein dictionary learning on MNIST; $r = 9$ (left) and $r = 16$ (right)

In Figure 6, we provide a numerical simulation of Algorithm 4 for Wasserstein CP-dictionary learning and verify our theoretical convergence results in Theorems 3.4 and 5.2. We observe faster convergence with the presence of proximal regularization with a suitable regularization coefficient.

8. Conclusion

We provide a theoretical analysis of the block coordinate descent methods with proximal regularization. The global convergence to the stationary points and the worst-case bound are obtained. We provide Wasserstein CP-dictionary learning as an application of our method.

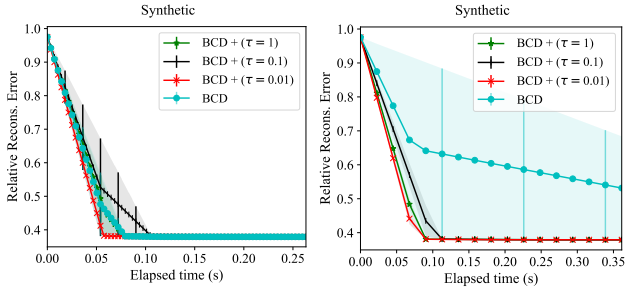

Figure 6. Plot of relative reconstruction error vs. time for Wasserstein CP-dictionary learning using Algorithm 4 with various choices of proximal regularization coefficient $\tau \in \{0, 0.1, 0.01, 1\}$. The tensor on the left and right has sizes $(100, 100, 500)$ and $(100, 100, 1000)$, respectively. Data tensors are generated by taking the outer product of randomly generated factor matrices of 10 columns plus i.i.d. noise of Uniform$(0, 10)$.

Acknowledgements

DK was supported by the 2023 Research Fund of the University of Seoul. HL was partially supported by the National Science Foundation through grants DMS-2206296 and DMS-2010035.

References

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of operations research*, 35(2):438–457, 2010.

Bauschke, H. H., Combettes, P. L., et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Bolte, J., Sabach, S., and Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer, 2010.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Carroll, J. D. and Chang, J.-J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

Cartis, C., Gould, N. I., and Toint, P. L. On the complexity of steepest descent, newton's and regularized newton's

methods for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

Cuturi, M. and Peyré, G. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

Daniilidis, A. and Malick, J. Filling the gap between lower-c1 and lower-c2 functions. *Journal of Convex Analysis*, 12(2):315–329, 2005.

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

Ghassemi, M., Shakeri, Z., Sarwate, A. D., and Bajwa, W. U. Stark: Structured dictionary learning through rank-one tensor recovery. In *Proceedings of the IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pp. 1–5. IEEE, 2017.

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

Harshman, R. A. Foundations of the parafac procedure: Models and conditions for an "explanatory" multimodal factor analysis. 1970.

Hong, M., Wang, X., Razaviyayn, M., and Luo, Z.-Q. Iteration complexity analysis of block coordinate descent methods. *Mathematical Programming*, 163(1):85–114, 2017.

Kolda, T. G. and Bader, B. W. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

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

Lee, D. D. and Seung, H. S. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 556–562, 2001.

Lyu, H. Convergence and complexity of stochastic block majorization-minimization. *arXiv preprint arXiv:2201.01652*, 2022.

Lyu, H., Strohmeier, C., and Needell, D. Online tensor factorization and cp-dictionary learning for markovian data. *arXiv preprint arXiv:2009.07612*, 2020.

Mairal, J. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pp. 783–791, 2013.

Mairal, J., Elad, M., and Sapiro, G. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69, 2007.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(Jan):19–60, 2010.

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

Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

Peyré, G. Sparse modeling of textures. *Journal of Mathematical Imaging and Vision*, 34(1):17–31, 2009.

Powell, M. J. On search directions for minimization algorithms. *Mathematical programming*, 4(1):193–201, 1973.

Rolet, A., Cuturi, M., and Peyré, G. Fast dictionary learning with a smoothed wasserstein loss. In *Artificial Intelligence and Statistics*, pp. 630–638. PMLR, 2016.

Sandler, R. and Lindenbaum, M. Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1590–1602, 2011.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Shakeri, Z., Bajwa, W. U., and Sarwate, A. D. Minimax lower bounds for kronecker-structured dictionary learning. In *IEEE International Symposium on Information Theory*, pp. 1148–1152. IEEE, 2016.

Shashua, A. and Hazan, T. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning*, pp. 792–799. ACM, 2005.

Sun, J., Qu, Q., and Wright, J. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

Tucker, L. R. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

Wang, Y.-X. and Zhang, Y.-J. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on knowledge and data engineering*, 25(6):1336–1353, 2012.

Wright, S. J. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

Xu, Y. and Yin, W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013.

Zafeiriou, S. Algorithms for nonnegative tensor factorization. In *Tensors in Image Processing and Computer Vision*, pp. 105–124. Springer, 2009.

Zen, G., Ricci, E., and Sebe, N. Simultaneous ground metric learning and matrix factorization with earth mover's distance. In *2014 22nd International Conference on Pattern Recognition*, pp. 3690–3695. IEEE, 2014.

Zeng, J., Lau, T. T.-K., Lin, S., and Yao, Y. Global convergence of block coordinate descent in deep learning. In *International Conference on Machine Learning*, pp. 7313–7323. PMLR, 2019.

Zhang, Z. and Brand, M. Convergent block coordinate descent for training tikhonov regularized deep neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

A. Block Coordinate Descent with proximal regularization

A.1. Proof of Theorem 3.4

Throughout this section, we let $(\boldsymbol{\theta}_n)_{n\geq 1}$ denote an inexact output of Algorithm (3) and write $\boldsymbol{\theta}_n = [\theta_n^{(1)}, \dots, \theta_n^{(m)}]$ for each $n \geq 1$. For each $n \geq 1$ and $i = 1, \dots, m$, denote

$$f_n^{(i)} : \theta \mapsto f(\theta_n^{(1)}, \dots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \dots, \theta_{n-1}^{(m)}), \tag{33}$$

which is L-smooth under Assumption 3.1. By Lemma D.6, it is also L-weakly convex. From this, it is easy to see that $g_n^{(i)}(\theta) = f_n^{(i)}(\theta) + \frac{\tau_n^{(i)}}{2}\|\theta - \theta_{n-1}^{(i)}\|^2$ is $(\tau_n^{(i)} - L^{(i)})$-strongly convex. Also, denote

$$\tau_n^- := \min_{i=1,\dots,m} \tau_n^{(i)}, \quad \tau_n := \max_{i=1,\dots,m} \tau_n^{(i)} \text{ for all } n \geq 1, \quad L := \max_{i=1,\dots,m} L^{(i)}. \tag{34}$$

We will use the notations above as well as this observation throughout this section.

Proposition A.1 (Forward monotonicity)**.** *Suppose Assumptions 3.1-3.3. Then the following hold:*

(i) $f(\boldsymbol{\theta}_{n-1}) - f(\boldsymbol{\theta}_n) \geq \frac{\tau_n^-}{2}\|\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}_n\|^2 - m\Delta_n$;

(ii) $\sum_{n=1}^{\infty} \tau_n^- \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 < \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} f(\boldsymbol{\theta}) + m\sum_{n=1}^{\infty}\Delta_n < \infty$.

Proof. Fix $i \in \{1, \dots, m\}$. Let $\theta_n^{(i\star)}$ be the exact minimizer of the $(\tau_n^{(i)} - L^{(i)})$-strongly convex function $g_n^{(i)}(\theta)$ over the convex set $\Theta^{(i)}$. Then $g_n^{(i)}(\theta_n^{(i)}) \leq f_n^{(i)}(\theta_{n-1}^{(i)}) = g_n^{(i)}(\theta_{n-1}^{(i)})$, for $n \geq 1$. Hence we deduce

$$f_n^{(i)}(\theta_{n-1}^{(i)}) - f_n^{(i)}(\theta_n^{(i)}) = g_n^{(i)}(\theta_{n-1}^{(i)}) - g_n^{(i)}(\theta_n^{(i)}) + g_n^{(i)}(\theta_n^{(i)}) - f_n^{(i)}(\theta_n^{(i)}) \geq -\Delta_n + \frac{\tau_n^{(i)}}{2}\|\theta_n^{(i)} - \theta_{n-1}^{(i)}\|^2. \tag{35}$$

It follows that

$$\begin{aligned}
&f(\boldsymbol{\theta}_{n-1}) - f(\boldsymbol{\theta}_n) && (36)\\
&= \sum_{i=1}^{n} f([\theta_n^{(1)}, \dots, \theta_n^{(i-1)}, \theta_{n-1}^{(i)}, \theta_{n-1}^{(i+1)}, \dots, \theta_{n-1}^{(m)}]) - f([\theta_n^{(1)}, \dots, \theta_n^{(i-1)}, \theta_n^{(i)}, \theta_{n-1}^{(i+1)}, \dots, \theta_{n-1}^{(m)}]) && (37)\\
&= \sum_{i=1}^{n} f_n^{(i)}(\theta_{n-1}^{(i)}) - f_n^{(i)}(\theta_n^{(i)}) && (38)\\
&\geq \sum_{i=1}^{m}\left(\frac{\tau_n^{(i)}}{2}\|\theta_n^{(i)} - \theta_{n-1}^{(i)}\|^2 - \Delta_n\right) = \frac{\tau_n^-}{2}\|\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}_n\|^2 - m\Delta_n. && (39)
\end{aligned}$$

This shows **(i)**.

Next, to show **(ii)**, adding up the above inequality,

$$\sum_{k=1}^{n}\frac{\tau_k^-}{2}\|\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}_k\|^2 \leq \left(\sum_{k=1}^{n} f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}_k)\right) + m\sum_{n=1}^{\infty}\Delta_n = f(\boldsymbol{\theta}_0) + m\sum_{n=1}^{\infty}\Delta_n < \infty, \tag{40}$$

where we have used the fact that $\sum_{n=1}^{\infty}\Delta_n < \infty$ due to Assumption 3.3. □

Proposition A.2 (Finite first-order variation)**.** *Suppose Assumptions 3.1-3.2. Also assume $\tau_n^- \geq 1$ for all $n \geq 1$. Suppose that $\sum_{n=1}^{\infty}\Delta_n < \infty$. Then*

$$\sum_{n=1}^{\infty}|\langle\nabla f(\boldsymbol{\theta}_{n+1}),\, \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\rangle| < \frac{L+2}{2}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} f(\boldsymbol{\theta}) + 3m\sum_{n=1}^{\infty}\Delta_n < \infty.$$

Proof. According to Assumptions 3.1 and 3.2, it follows that ∇f over $\boldsymbol{\Theta}$ is Lipschitz with Lipshitz constant L. Hence by Lemma D.3, for all $t \geq 1$,

$$|f(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_{n+1}) - \langle \nabla f(\boldsymbol{\theta}_{n+1}),\, \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\rangle| \leq \frac{L}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|_F^2.$$

Using Proposition A.1, it follows that

$$|f(\boldsymbol{\theta}_{n-1}) - f(\boldsymbol{\theta}_n)| \leq f(\boldsymbol{\theta}_{n-1}) - f(\boldsymbol{\theta}_n) + 2m\Delta_n. \tag{41}$$

Hence this yields

$$|\langle \nabla f(\boldsymbol{\theta}_{n+1}),\, \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\rangle| \leq \frac{L}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|_F^2 + |f(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_{n+1})| \tag{42}$$

$$\leq \frac{L}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|_F^2 + f(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_{n+1}) + 2m\Delta_n \tag{43}$$

for $n \geq 1$. Also note that $\sum_{t=1}^n f(\boldsymbol{\theta}_t) - f(\boldsymbol{\theta}_{t+1}) = f(\boldsymbol{\theta}_1) - f(\boldsymbol{\theta}_{n+1}) \leq f(\boldsymbol{\theta}_1)$. Hence

$$\begin{aligned}
\sum_{n=0}^{\infty} |\langle \nabla f(\boldsymbol{\theta}_{n+1}),\, \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\rangle| &\leq \frac{L}{2}\left(\sum_{n=0}^{\infty}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|_F^2\right) + f(\boldsymbol{\theta}_0) + 2m\sum_{n=1}^{\infty}\Delta_n \\
&\leq \frac{L}{2}\left(\sum_{n=0}^{\infty}\tau_n^-\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|_F^2\right) + f(\boldsymbol{\theta}_0) + 2m\sum_{n=0}^{\infty}\Delta_n \\
&\leq 2f(\boldsymbol{\theta}_0) + 3m\sum_{n=1}^{\infty}\Delta_n < \infty,
\end{aligned}$$

where we have used Proposition A.1 **(ii)**. □

Proposition A.3 (Boundedness of iterates)**.** *Under Assumptions 3.2 and 3.3, the set* $\{\boldsymbol{\theta}_n : n \geq 1\}$ *is bounded.*

Proof. Let $T := m\sum_{k=1}^{\infty}\Delta_k$, which is finite by Assumption 3.3. Recall that by Proposition A.1, we have

$$\sup_{n\geq 1} f(\boldsymbol{\theta}_n) \leq f(\boldsymbol{\theta}_1) + T < \infty. \tag{44}$$

Then we can conclude by using Assumption 3.2. □

Proposition A.4 (Asymptotic first-order optimality)**.** *Suppose Assumptions 3.1-3.3. Fix a sequence* $(b_n)_{n\geq 1}$ *with* $b_n > 0$ *for* $n \geq 1$*. Then there exists constants* $c_1, c_2 > 0$ *independent of* $\boldsymbol{\theta}_0 \in \boldsymbol{\theta}$ *such that for all* $n \geq 1$*,*

$$\langle \nabla f(\boldsymbol{\theta}_{n+1}),\, \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\rangle \leq b_{n+1}\inf_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\left\langle \nabla f(\boldsymbol{\theta}_n),\, \frac{\boldsymbol{\theta} - \boldsymbol{\theta}_n}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|}\right\rangle + c_0 b_{n+1}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\| \tag{45}$$

$$+ c_1\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 + c_2(L + \tau_{n+1})b_{n+1}^2 + \Delta_{n+1}. \tag{46}$$

Proof. Fix arbitrary $\boldsymbol{\theta} = [\theta^{(1)}, \ldots, \theta^{(m)}] \in \boldsymbol{\Theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq b_{n+1}$. By convexity of $\Theta^{(i)}$, $\theta_n^{(i)} + a(\theta^{(i)} - \theta_n^{(i)}) \in \Theta^{(i)}$ for all $a \in [0,1]$. Let $\theta_{n+1}^{(i\star)}$ denote the exact minimizer of $g_{n+1}^{(i)}$ over $\Theta^{(i)}$. Then we have

$$f_{n+1}^{(i)}(\theta_{n+1}^{(i)}) + \frac{\tau_{n+1}^{(i)}}{2}\|\theta_{n+1}^{(i)} - \theta_n^{(i)}\|^2 - \Delta_{n+1} \leq f_{n+1}^{(i)}(\theta_{n+1}^{(i\star)}) + \tau_{n+1}^{(i)}\|\theta_{n+1}^{(i\star)} - \theta_n^{(i)}\|^2 \tag{47}$$

$$\leq f_{n+1}^{(i)}\left(\theta_n^{(i)} + a(\theta^{(i)} - \theta_n^{(i)})\right) + \frac{\tau_{n+1}^{(i)}a^2}{2}\|\theta^{(i)} - \theta_n^{(i)}\|^2. \tag{48}$$

Recall that each $f_{n+1}^{(i)}$ is $L^{(i)}$-smooth by Assumption 3.1. Hence by subtracting $f_{n+1}^{(i)}(\theta_n^{(i)})$ from both sides and using Lemma D.3, we get

$$\left\langle \nabla f_{n+1}^{(i)}(\theta_n^{(i)}),\, \theta_{n+1}^{(i)} - \theta_n^{(i)}\right\rangle \leq a\left\langle \nabla f_{n+1}^{(i)}(\theta_n^{(i)}),\, \theta^{(i)} - \theta_n^{(i)}\right\rangle \tag{49}$$

$$+ \frac{L^{(i)}}{2}\|\theta_{n+1}^{(i)} - \theta_n^{(i)}\|^2 + \frac{L^{(i)}}{2}\|\theta^{(i)} - \theta_n^{(i)}\|^2 + \frac{\tau_{n+1}^{(i)}a^2}{2}\|\theta^{(i)} - \theta_n^{(i)}\|^2 + \Delta_{n+1}. \tag{50}$$

Adding up these inequalities for $i = 1, \dots, m$,

$$\left\langle \left[\nabla f_{n+1}^{(1)}(\theta_n^{(1)}), \dots, \nabla f_{n+1}^{(m)}(\theta_n^{(m)})\right], \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \right\rangle \le a \left\langle \left[\nabla f_{n+1}^{(1)}(\theta_n^{(1)}), \dots, \nabla f_{n+1}^{(m)}(\theta_n^{(m)})\right], \boldsymbol{\theta} - \boldsymbol{\theta}_n \right\rangle \tag{51}$$

$$+ \frac{L}{2}\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 + \frac{(L + \tau_{n+1}a^2)}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|^2 + \Delta_{n+1}. \tag{52}$$

Since for each $i = 1, \dots, m$ ∇f is $L^{(i)}$-Lipschits in the ith block coordinate, we have

$$\|\nabla_i f(\theta_n^{(1)}, \dots, \theta_n^{(m)}) - \nabla f_{n+1}^{(i)}(\theta_n^{(i)})\| \le L^{(i)}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|. \tag{53}$$

Hence there exists constants $c_1, c_2 > 0$ independent of $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, such that

$$\langle \nabla f(\boldsymbol{\theta}_{n+1}),\, \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle \le a \langle \nabla f(\boldsymbol{\theta}_n),\, \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle + amL\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\| \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \tag{54}$$

$$+ c_1\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 + c_2(L + \tau_{n+1}a^2)\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|^2 + \Delta_{n+1}. \tag{55}$$

The above inequality holds for all $a \in [0, 1]$.

Viewing the right hand side as a quadratic function in a, the only possibly negative term is the linear term $a\langle \nabla f(\boldsymbol{\theta}_n),\, \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle$, whose absolute value is bounded above by $a\|\nabla f(\boldsymbol{\theta}_n)\|\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|$. By Proposition A.3 and Assumption 3.3, $\|\nabla f(\boldsymbol{\theta}_n)\|$ is uniformly bounded, so this is bounded above by $ac_3\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|$ for some constant $c_3 > 0$. Hence we may choose $c_2 > 0$ large enough so that the right hand side above is non-increasing in a. Thus the inequality above holds for all $a \ge 0$. In particular, we can choose $a = b_{n+1}/\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|$. This and using $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \le b_{n+1}$ yield

$$\langle \nabla f(\boldsymbol{\theta}_{n+1}),\, \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle \le b_{n+1} \left\langle \nabla f(\boldsymbol{\theta}_n),\, \frac{\boldsymbol{\theta} - \boldsymbol{\theta}_n}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|} \right\rangle + c_0\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|b_{n+1} \tag{56}$$

$$+ c_1\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 + c_2(L + \tau_{n+1})b_{n+1}^2 + \Delta_{n+1}, \tag{57}$$

where we wrote $c_0 := mL$.

We have shown that the above holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \le b_{n+1}$. It remains to argue that (56) also holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \ge b_{n+1}$. Indeed, for such $\boldsymbol{\theta}$, let $\boldsymbol{\theta}'$ be the point in the secant line between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_n$ such that $\|\boldsymbol{\theta}' - \boldsymbol{\theta}_n\| \le b_{n+1}$. Then $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$ and (56) holds for $\boldsymbol{\theta}$ replaced with $\boldsymbol{\theta}'$. However, the right hand side is unchanged when replacing $\boldsymbol{\theta}$ with any point on the line passing through $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_n$. Thus (56) holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. This shows the assertion. □

Proposition A.5 (Optimality gap for iterates)**.** *For each $n \ge 1$ and $i \in \{1, \dots, m\}$, let $\theta_n^{(i\star)}$ be the exact minimizer of the $(\tau_n^{(i)} - L^{(i)})$-strongly convex function $\theta \mapsto g_n^{(i)}(\theta)$ in (3) over the convex set $\Theta^{(i)}$. Then*

$$\frac{\tau_n^{(i)} - L^{(i)}}{2}\|\theta_n^{(i\star)} - \theta_n^{(i)}\|^2 \le \Delta_n. \tag{58}$$

Proof. The assertion follows from

$$\frac{\tau_n^{(i)} - L^{(i)}}{2}\|\theta_n^{(i\star)} - \theta_n^{(i)}\|^2 \le g_n^{(i)}(\theta_n^{(i)}) - g_n^{(i)}(\theta_n^{(i\star)}) \le \Delta_n \tag{59}$$

for $n \ge 1$. Indeed, the first inequality follows from the second-order growth property (see Lemma D.4) since $g_n^{(n)}$ is $(\tau_n^{(i)} - L^{(i)})$-strongly convex minimized at $\theta_n^{(i)}$, and the second inequality follows from the definition of optimality gap Δ_n in (7). □

We are now ready to give a proof of Theorem 3.4.

Proof of Theorem 3.4. Suppose Assumptions 3.1-3.3 and $\tau_n^{(i)} > L^{(i)} + \delta$ for $n \ge 1$ for some $\delta > 0$. Also assume $\tau_n^{(i)} = O(1)$. We first show **(i)**. Fix a convergent subsequence $(\boldsymbol{\theta}_{n_k})_{k\ge1}$ of $(\boldsymbol{\theta}_n)_{n\ge1}$. We wish to show that $\boldsymbol{\theta}_\infty = \lim_{k\to\infty} \boldsymbol{\theta}_{n_k}$

is a stationary point of f over $\boldsymbol{\Theta}$. To this end, for each $i \in \{1, \cdots, m\}$, let $\theta_n^{(i\star)}$ denote the exact minimizer of the $(\tau_n^{(i)} - L^{(i)})$-strongly convex function $g_n^{(i)}$ defined in (3). By using the first-order optimality of $\theta_n^{(i\star)}$, we have

$$\left\langle \nabla g_n^{(i)}(\theta_n^{(i\star)}),\, \theta - \theta_n^{(i\star)} \right\rangle = \left\langle \nabla f_n^{(i)}(\theta_n^{(i\star)}) + \tau_n^{(i)}(\theta_n^{(i\star)} - \theta_{n-1}^{(i)}),\, \theta - \theta_n^{(i\star)} \right\rangle \geq 0 \qquad \forall \theta \in \Theta^{(i)}. \tag{60}$$

Let $T := m \sum_{k=1}^{\infty} \Delta_k$, which is finite by Assumption 3.3. Recall that by Proposition A.1, we have

$$\sup_{n \geq 1} f(\boldsymbol{\theta}_n) \leq f(\boldsymbol{\theta}_1) + T < \infty. \tag{61}$$

Let $K := \{\boldsymbol{\theta} \,:\, f(\boldsymbol{\theta}) \leq f(\boldsymbol{\theta}_1) + T\}$ and let $K(T) := \{\boldsymbol{\theta} \,:\, \exists \boldsymbol{\theta}' \in K \text{ s.t. } \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \leq T\}$ denote the T-neighborhood of K. By Assumption 3.2, K is compact, so $K(T)$ is also compact. Since f is $L^{(i)}$-smooth in its ith block coordinate, $\|\nabla g_n^{(i)}\|$ is uniformly bounded over $\boldsymbol{\theta} \in K(T)$ by some constant, say, $L_K > 0$. Now observe that

$$\begin{aligned}
\left\langle \nabla g_n^{(i)}(\theta_n^{(i\star)}),\, \theta - \theta_n^{(i)} \right\rangle &\geq \left\langle \nabla g_n^{(i)}(\theta_n^{(i\star)}),\, \theta - \theta_n^{(i\star)} \right\rangle - \left| \left\langle \nabla g_n^{(i)}(\theta_n^{(i\star)}),\, \theta_n^{(i)} - \theta_n^{(i\star)} \right\rangle \right| && (62)\\
&\geq -\|\nabla g_n^{(i)}(\theta_n^{(i\star)})\|\, \|\theta_n^{(i)} - \theta_n^{(i\star)}\| && (63)\\
&\geq -L_K \Delta_n. && (64)
\end{aligned}$$

Next, using $L^{(i)}$-Lipschitzness of ∇f in the ith block coordinate and Proposition A.5, we have

$$\begin{aligned}
&\left| \left\langle \nabla g_n^{(i)}(\theta_n^{(i\star)}),\, \theta - \theta_n^{(i)} \right\rangle - \left\langle \nabla g_n^{(i)}(\theta_n^{(i)}),\, \theta - \theta_n^{(i)} \right\rangle \right| && (65)\\
&\quad \leq \left| \left\langle \nabla f_n^{(i)}(\theta_n^{(i\star)}) - \nabla f_n^{(i)}(\theta_n^{(i)}) + \tau_n^{(i)}(\theta_n^{(i\star)} - \theta_n^{(i)}),\, \theta - \theta_n^{(i)} \right\rangle \right| && (66)\\
&\quad \leq \left(\|\nabla f_n^{(i)}(\theta_n^{(i\star)}) - \nabla f_n^{(i)}(\theta_n^{(i)})\| + \tau_n^{(i)} \|\theta_n^{(i\star)} - \theta_n^{(i)}\| \right) \|\theta - \theta_n^{(i)}\| && (67)\\
&\quad \leq (L^{(i)} + \tau_n^{(i)}) \|\theta - \theta_n^{(i)}\|\, \|\theta_n^{(i\star)} - \theta_n^{(i)}\| && (68)\\
&\quad \leq (L^{(i)} + \tau_n^{(i)}) \|\theta - \theta_n^{(i)}\| \sqrt{\frac{2\Delta_n}{\tau_n^{(i)} - L^{(i)}}} = \|\theta - \theta_n^{(i)}\| \sqrt{\frac{8\tau_n^{(i)} \Delta_n}{1 - L^{(i)}/\tau_n^{(i)}}}, && (69)
\end{aligned}$$

where for the last equality we have used that $\tau_n^{(i)} > L^{(i)}$ for $n \geq 1$. From Assumption 3.3, we can deduce $\Delta_n = o(1)$. Using the hypotheses $\tau_n^{(i)} > L^{(i)} + \delta$ for $n \geq 1$ for some $\delta > 0$ (see Algorithm (3)), and $\tau_n^{(i)} = O(1)$, we see that the term inside the square root in the last expression is $o(1)$. Furthermore, $\|\theta - \theta_{n_k}^{(i)}\|$ is uniformly bounded in k since $\theta_{n_k}^{(i)}$ converges as $k \to \infty$. Hence

$$\liminf_{k \to \infty} \left\langle \nabla g_{n_k}^{(i)}(\theta_{n_k}^{(i)}),\, \theta - \theta_{n_k}^{(i)} \right\rangle \geq 0 \qquad \forall \theta \in \Theta^{(i)}. \tag{70}$$

Note that by Proposition A.1 **(ii)** and $\tau_n^{(i)} = O(1)$, we get $\tau_n^{(i)} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = o(1)$. So if we write $\boldsymbol{\theta}_\infty = [\theta_\infty^{(1)}, \ldots, \theta_\infty^{(m)}]$, For each $\theta \in \Theta^{(i)}$, by the hypothesis, we get

$$\begin{aligned}
&\lim_{k \to \infty} \left| \left\langle \nabla f_{n_k}^{(i)}(\theta_{n_k}^{(i)}) + 2\tau_{n_k}^{(i)}(\theta_{n_k}^{(i)} - \theta_{n_k-1}^{(i)}),\, \theta - \theta_{n_k}^{(i)} \right\rangle - \left\langle \nabla f_{n_k}^{(i)}(\theta_{n_k}^{(i)}),\, \theta - \theta_{n_k}^{(i)} \right\rangle \right| && (71)\\
&\quad \leq \lim_{k \to \infty} 2\tau_{n_k}^{(i)} \|\theta_{n_k}^{(i)} - \theta_{n_k-1}^{(i)}\|\, \|\theta - \theta_{n_k-1}^{(i)}\| && (72)\\
&\quad = 2\|\theta - \theta_\infty^{(i)}\| \lim_{k \to \infty} \tau_{n_k}^{(i)} \|\theta_{n_k}^{(i)} - \theta_{n_k-1}^{(i)}\| = 0. && (73)
\end{aligned}$$

It follows that, for each $\theta \in \Theta^{(i)}$, using the continuity of ∇f in Assumption 3.2,

$$\left\langle \nabla_i f(\theta_\infty^{(1)}, \ldots, \theta_\infty^{(i-1)}, \theta_\infty^{(i)}, \theta_\infty^{(i+1)}, \ldots, \theta_\infty^{(m)}),\, \theta - \theta_\infty^{(i)} \right\rangle = \lim_{k \to \infty} \left\langle \nabla f_{n_k}^{(i)}(\theta_{n_k}^{(i)}),\, \theta - \theta_{n_k}^{(i)} \right\rangle \geq 0. \tag{74}$$

This holds for all $i = 1, \ldots, m$. Therefore we verify $\langle \nabla f(\boldsymbol{\theta}_\infty),\, \boldsymbol{\theta} - \boldsymbol{\theta}_\infty \rangle \geq 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, which means that $\boldsymbol{\theta}_\infty$ is a stationary point of f over $\boldsymbol{\Theta}$, as desired. This shows **(i)**.

Next, we show **(ii)**. Let b_n be any square-summable sequence of positive numbers. By Cauchy-Schwarz inequality,

$$\sum_{k=1}^{n} b_k \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}\| \le \left(\sum_{k=1}^{n} b_k^2\right)^{1/2} \left(\sum_{k=1}^{n} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}\|^2\right)^{1/2}. \tag{75}$$

Then by Proposition A.1, the right hand side is uniformly bounded in $n \ge 1$, so we see that the left hand side is also uniformly bounded in n. Hence using Propositions A.1 and A.2,

$$\sum_{n=1}^{\infty} b_{n+1} \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_n), \frac{\boldsymbol{\theta} - \boldsymbol{\theta}_n}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|} \right\rangle \right] \le C \left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} f(\boldsymbol{\theta}) + \sum_{n=1}^{\infty} \Delta_n(\boldsymbol{\theta}_0) + \sum_{n=1}^{\infty} b_n^2 + \sum_{n=1}^{\infty} b_n \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\| \right) \tag{76}$$

for some constant $C > 0$ independent of $\boldsymbol{\theta}_0$, and the right hand side is finite. Thus by taking $b_n = 1/(\sqrt{n}\log n)$, using Lemma D.2, we deduce

$$\min_{1\le k\le n} \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_k), \frac{\boldsymbol{\theta} - \boldsymbol{\theta}_k}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_k\|} \right\rangle \right] \le \frac{M + c\sum_{n=1}^{\infty} \Delta_n(\boldsymbol{\theta}_0)}{\sqrt{n}/\log n} \tag{77}$$

for some constants $M, c > 0$ independent of $\boldsymbol{\theta}_0$. This shows **(ii)**.

Lastly, we show **(iii)**. Assume $\sup_{\boldsymbol{\theta}_0\in\boldsymbol{\Theta}} \sum_{n=1}^{\infty} \Delta_n(\boldsymbol{\theta}_0) < \infty$. Then the above implies that for some constant $M' > 0$ independent of $\boldsymbol{\theta}_0$,

$$\min_{1\le k\le n} \sup_{\boldsymbol{\theta}_0\in\boldsymbol{\Theta}} \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_k), \frac{\boldsymbol{\theta} - \boldsymbol{\theta}_k}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_k\|} \right\rangle \right]^2 \le \frac{M'(\log n)^2}{n}. \tag{78}$$

Then one can conclude **(iii)** by using the fact that $n \ge 2\varepsilon^{-1}(\log \varepsilon^{-1})^2$ implies $(\log n)^2/n \le \varepsilon$ for all sufficiently small $\varepsilon > 0$. This completes the proof. □

B. Proof of Theorem 5.1

In this section, we establish Theorem 5.1, the per-iteration correctness of Algorithm 1. This directly follows from Propositions B.1 and B.2 below.

Proposition B.1. *For given $(\mathcal{D}_{n-1}, \Lambda_{n-1}) \in \Sigma^r_{I_1\times\cdots\times I_d} \times \Sigma^N_r$ and $\tau_n > 0$, let $\Lambda_n \in \Sigma^N_r$ be a solution of (18). Suppose each fiber of $\mathcal{D}_{n-1}$ along the last mode is not identically zero. Then, Λ_n is uniquely determined by*

$$\Lambda_n = \left(\Lambda_{n-1} + \frac{\mathcal{D}_{n-1} \times_{\le d} G_n^\circ}{\tau_n} - J^\circ \otimes c_n^\circ \right)_+. \tag{79}$$

Here, $G_n^\circ \in \mathbb{R}^N$ is defined as the unique solution of the dual problem (19), $c_n^\circ \in \mathbb{R}^{N\times 1}$ is chosen to satisfy $\Lambda_n \in \Sigma^N_r$ and all entries of $J^\circ \in \mathbb{R}^{r\times 1}$ are one.

As shown later in the proof, the assumption on $\mathcal{D}_{n-1}$ in the above proposition is required to ensure the above derivation. It is worth pointing out that it can be easily achieved in the algorithm by adding small noise, if necessary.

Proposition B.2. *For given $(\mathcal{D}_{n-1}, \Lambda_n) \in \Sigma^r_{I_1\times\cdots\times I_d} \times \Sigma^N_r$, let $\mathcal{D}_n \in \Sigma^r_{I_1\times\cdots\times I_d}$ be a solution of (14). Suppose each fiber of $\Lambda_n \in \Sigma^N_r$ along the 2nd mode is not identically zero. Then $\mathcal{D}_n$ is uniquely determined by*

$$\mathcal{D}_n = \left(\mathcal{D}_{n-1} + \frac{G_n^\dagger \times_{d+1} \Lambda_n^T}{\tau_n} - J^\dagger \otimes c_n^\dagger \right)_+. \tag{80}$$

Here, $G_n^\dagger$ is defined as the unique solution of the dual problem (22), $c_n^\dagger \in \mathbb{R}^{r\times 1}$ is chosen to satisfy $\mathcal{D}_n \in \Sigma^r_{I_1\times\cdots\times I_d}$ and all entries of $J^\dagger \in \mathbb{R}^{I_1 I_2\cdots I_d\times 1}$ are one.

The following definitions are taken from (Bauschke et al., 2011).

Definition B.3. (Bauschke et al., 2011)(Definitions 9.12, 19.10, 19.15 & 19.22)

- For a nonempty closed convex cone $K \subset \mathcal{K}$, we say that $R : \mathcal{H} \to \mathcal{K}$ is convex with respect to K if
$$R(\alpha x + (1-\alpha)y) - \alpha Rx - (1-\alpha)Ry \in K$$
for all $x, y \in \mathcal{H}$ and $\alpha \in (0,1)$.
- The set of proper lower semicontinuous convex functions from $\mathcal{H}$ to $(-\infty, +\infty]$ is denoted by $\Gamma_0(\mathcal{H})$.
- The Lagrangian of $\mathcal{J} : \mathcal{H} \times \mathcal{K} \to (-\infty, +\infty]$ is a function given as
$$\mathcal{L} : \mathcal{H} \times \mathcal{K} \to [-\infty, +\infty] : (x, v) \mapsto \inf_{y \in \mathcal{K}} \left(\mathcal{J}(x, y) + \langle y, v \rangle \right). \tag{81}$$
Moreover, $(x, v) \in \mathcal{H} \times \mathcal{K}$ is a saddle point of $\mathcal{L}$ if
$$\mathcal{L}(x, v) = \sup \mathcal{L}(x, \mathcal{K}) = \inf \mathcal{L}(\mathcal{H}, v).$$
- The primal problem and the dual problem of $\mathcal{J} : \mathcal{H} \times \mathcal{K} \to (-\infty, +\infty]$ are respectively given as
$$\min_{x \in \mathcal{H}} \mathcal{J}(x, 0), \quad \text{and} \quad \min_{v \in \mathcal{K}} \mathcal{J}^*(0, v). \tag{82}$$

We first observe that the primal problem of
$$\mathcal{J} : \mathcal{H} \times \mathcal{K} \to (-\infty, +\infty] : (x, y) \mapsto \begin{cases} f(Rx - y) + h(x), & \text{if } Rx \in y + K, \\ +\infty, & \text{if } Rx \notin y + K, \end{cases} \tag{83}$$
is the minimization problem (23). Its dual problem, the Lagrangian of $\mathcal{J}$, and the saddle point are given in the following lemma.

Lemma B.4 (Characterization of saddle point for general coding problem)**.** *Let $f \in \Gamma_0(\mathcal{K})$, $h \in \Gamma_0(\mathcal{H})$, and K be a nonempty closed convex cone in $\mathcal{K}$. Let $R : \mathcal{H} \to \mathcal{K}$ be continuous, convex with respect to K such that $K \cap R(\mathrm{dom} h) \neq \emptyset$. For $\mathcal{J}$ given in (83), the following hold:*

1. *The dual problem of $\mathcal{J}$ is given as*
$$\min_{v \in \mathcal{K}} f^*(-v; K) + h^*(R^* v) \tag{84}$$
where $f^*(\cdot; K) = \sup_{z \in K} \langle z, \cdot \rangle - f(z)$.
2. *The Lagrangian $\mathcal{L} : \mathcal{H} \times \mathcal{K} \to [-\infty, +\infty]$ is given as*
$$\mathcal{L}(x, v) = \begin{cases} -\infty & \text{if } x \in \mathrm{dom} h \text{ and } v \notin \mathrm{dom} f^*(\cdot; K); \\ -f^*(v; K) + h(x) + \langle Rx, v \rangle & \text{if } x \in \mathrm{dom} h \text{ and } v \in \mathrm{dom} f^*(\cdot; K); \\ +\infty & \text{if } x \notin \mathrm{dom} h. \end{cases} \tag{85}$$
3. *Suppose that the optimal values μ and μ^* of the primal problem and the dual problem satisfy the strong duality $\mu = -\mu^*$. Then, $(x^\circ, -v^\circ) \in \mathcal{H} \times \mathcal{K}$ is a saddle point of $\mathcal{L}$ if and only if*
$$x^\circ \in \mathrm{dom} h, \;\; Rx^\circ \in K, \;\; -v^\circ \in \mathrm{dom} f^*(\cdot; K), \\ R^* v^\circ \in \partial h(x^\circ) \text{ and } -v^\circ \in \partial f(Rx^\circ).$$

Proof. (1): For any $v \in \mathcal{K}$, it holds that
$$\begin{aligned} \mathcal{J}^*(0, v) &= \sup_{(x,y) \in \mathcal{H} \times \mathcal{K}} \langle y, v \rangle - \mathcal{J}(x, y), \\ &= \sup_{(x,y) \in \mathcal{H} \times \mathcal{K} \text{ s.t. } Rx - y \in K} \langle y, v \rangle - h(x) - f(Rx - y), \\ &= \sup_{(x,z) \in \mathcal{H} \times K} \langle x, R^* v \rangle - h(x) + \langle z, -v \rangle - f(z), \\ &= h^*(R^* v) + f^*(-v; K). \end{aligned}$$

From the definition of the dual problem, we conclude.

(2): If $x \notin \mathrm{dom} h$, then $h(x) = \infty$. As $f \in \Gamma_0(\mathcal{K})$, we have $\mathcal{J}(x, v) = \infty$ and thus the Lagrangian $\mathcal{L}(x, v) = \infty$. For $x \in \mathrm{dom} h$ and $v \in \mathrm{dom} f^*(\cdot; K)$, we have

$$\begin{aligned}\mathcal{L}(x, v) &= h(x) + \inf_{y \in \mathcal{K} \text{ s.t. } Rx - y \in K} f(Rx - y) + \langle y, v\rangle, \\ &= h(x) + \langle Rx, v\rangle + \inf_{z \in K} f(z) - \langle z, v\rangle, \\ &= h(x) + \langle Rx, v\rangle - \sup_{z \in K} \langle z, v\rangle - f(z), \\ &= h(x) + \langle Rx, v\rangle - f^*(v; K).\end{aligned}$$

If $x \in \mathrm{dom} h$ and $v \notin \mathrm{dom} f^*(\cdot; K)$, the above relation yields that $\mathcal{L}(x, v) = -\infty$.

(3): From $f \in \Gamma_0(\mathcal{K})$, $h \in \Gamma_0(\mathcal{H})$, and the convexity of R with respect to K, we have that $\mathcal{J} \in \Gamma_0(\mathcal{H} \times \mathcal{K})$. Applying Corollary 19.17 in (Bauschke et al., 2011), we obtain that $(x^\circ, -v^\circ)$ is a saddle point of $\mathcal{L}$ if and only if x° is a solution of the primal problem (23) and v° is a solution of the dual problem (84).

As $\mu = -\mu^*$, the equivalence in Corollary 19.1 from (Bauschke et al., 2011) concludes our claim. □

Now we are ready to prove Proposition B.1.

***Proof of Proposition** B.1.* The primal problem (18) for updating Λ has convex objective function and is strictly feasible under the hypothesis that $\mathcal{D}_{n-1}$ consists of nonzero tensor slices $\mathbf{D}_i$. Hence the primal problem (18) obtains strong duality (see, e.g., (Boyd et al., 2004)).

Let Λ_n and G_n be the optimizers of the primal problem (18) and the dual problem (19), respectively. In what follows, we will apply Lemma B.4. For $K = \Sigma^N_{I_1 \times \cdots \times I_d}$, let us consider

$$\begin{aligned}&\mathcal{J} : \mathbb{R}^{r \times N} \times \mathbb{R}^{I_1 \times \cdots \times I_d \times N} \to (-\infty, +\infty] : \\ &(\Lambda, Y) \mapsto \begin{cases} \sum_{i=1}^N \left\{ H_{\mathbf{X}_i}\left(\mathcal{D}_{n-1} \times_{d+1} \Lambda[:, i] - Y[:, i]\right) + \tau_n F_{\Lambda_{n-1}[:,i]}(\Lambda[:, i])\right\} & \text{if } \mathcal{D}_{n-1} \times_{d+1} \Lambda \in Y + K, \\ +\infty, & \text{if } \mathcal{D}_{n-1} \times_{d+1} \Lambda \notin Y + K. \end{cases}\end{aligned}$$

Then, (18) and (19) are the primal problem and the dual problem of $\mathcal{J}$, respectively.

From Cor. 19.17 in (Bauschke et al., 2011), $(\Lambda_n, -G_n)$ is a saddle point of the Lagrangian associated with $\mathcal{J}$. Applying Lemma B.4(3), we get $\mathcal{D}_{n-1} \times_{\leq d} G_n \in \tau_n \partial F(\Lambda_n)$. Note that $\xi + J^\circ \otimes c_n^\circ \in \partial F(\Lambda_n)$ for any $c_n^\circ \in \mathbb{R}^{N \times 1}$, where all entries of $J^\circ \in \mathbb{R}^{r \times 1}$ are one, and $\xi \in \mathbb{R}^{r \times N}$ satisfies

$$\begin{cases} \xi[i, j] = \Lambda_n[i, j] - \Lambda_{n-1}[i, j] & \text{if } \Lambda_n[i, j] > 0, \\ \xi[i, j] \in (-\infty, -\Lambda_{n-1}[i, j]] & \text{if } \Lambda_n[i, j] = 0, \end{cases} \tag{86}$$

for all $i = 1, 2, \cdots, r$, and $j = 1, 2, \cdots, N$. Hence

$$\mathcal{D}_{n-1} \times_{\leq d} G_n / \tau_n = \xi + J^\circ \otimes c_n^\circ \tag{87}$$

for some ξ satisfying (86) and $c_n^\circ \in \mathbb{R}^{N \times 1}$. Now combining (87) and (86) yields (79). Finally, since we must have $\Lambda_n \in \mathrm{dom}(F)$, $c_n^\circ \in \mathbb{R}^{N \times 1}$ should be such that Λ_n in (79) satisfies $\Lambda_n \in \Sigma^N_r$. □

C. Proof of Theorem 6.1

Here, we only prove the following proposition. The rest of arguments is parallel to the proof of Theorem 5.1

Proposition C.1. *For each $k \in \{1, 2, \cdots, d\}$, let $\overline{\Lambda} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_{k-1} \times r \times I_{k+1} \times \cdots \times I_d \times N}$ be obtained from*

$$\mathit{Out}(U_n^{(1)}, \ldots, U_n^{(k-1)}, U_{n-1}^{(k+1)}, \ldots, U_{n-1}^{(d)}, \Lambda_n^T) \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_{k-1} \times I_{k+1} \times \cdots \times I_d \times N \times r} \tag{88}$$

by inserting the last mode into the kth mode. Let $U_n^{(k)} \in \Sigma_{I_k}^r$ be a solution of (14). Suppose each fiber of $\Lambda_n \in \Sigma_r^N$ along the kth mode is not identically zero. Then $U_n^{(k)}$ is uniquely determined by

$$U_n^{(k)} = \left(U_{n-1}^{(k)} + \frac{G_n^\dagger \times_{\neq k} \overline{\Lambda}}{\tau_n} - J^\dagger \otimes c_n^\dagger \right)_+ . \tag{89}$$

Here, $G_n^\dagger$ is defined as the unique solution of the dual problem (22), $c_n^\dagger \in \mathbb{R}^{r\times 1}$ is chosen to satisfy $U_n^{(k)} \in \Sigma_{I_k}^r$ and all entries of $J^\dagger \in \mathbb{R}^{I_k \times 1}$ are one.

Proof. We first obtain the dual of (32):

$$\begin{aligned}
&\min_{U\in\Sigma_{I_k}^r} \left(\sum_{i=1}^N H_{\mathbf{X}_i}(\overline{\Lambda}[:,i] \times_k U^T) \right) + \tau_n F_{U_{n-1}^{(k)}}(U) \\
&= \min_{U\in\Sigma_{I_k}^r, Q\in\Sigma_{I_1\times\cdots\times I_d}^N, Q[:,i]=\overline{\Lambda}[:,i]\times_k U^T} \left(\sum_{i=1}^N H_{\mathbf{X}_i}(Q[:,i]) \right) + \tau_n F_{U_{n-1}^{(k)}}(U), \\
&= \min_{U\in\Sigma_{I_k}^r, Q\in\Sigma_{I_1\times\cdots\times I_d}^N} \max_{G\in\mathbb{R}^{I_1\times\cdots\times I_d\times N}} \sum_{i=1}^N \left\{ H_{\mathbf{X}_i}(Q[:,i]) + \langle Q[:,i] - \overline{\Lambda}[:,i] \times_k U^T, G[:,i]\rangle \right\} + \tau_n F_{U_{n-1}^{(k)}}(U), \\
&\overset{(a)}{=} \max_{G\in\mathbb{R}^{I_1\times\cdots\times I_d\times N}} \min_{U\in\Sigma_{I_k}^r, Q\in\Sigma_{I_1\times\cdots\times I_d}^N} \sum_{i=1}^N \left\{ H_{\mathbf{X}_i}(Q[:,i]) + \langle Q[:,i] - \overline{\Lambda}[:,i] \times_k U^T, G[:,i]\rangle \right\} + \tau_n F_{U_{n-1}^{(k)}}(U), \\
&= \max_{G\in\mathbb{R}^{I_1\times\cdots\times I_d\times N}} \sum_{i=1}^N - \left\{ \max_{Q\in\Sigma_{I_1\times\cdots\times I_d}^N} \langle Q[:,i], -G[:,i]\rangle - H_{\mathbf{X}_i}(Q[:,i]) \right\} \\
&\qquad + \min_{U\in\Sigma_{I_k}^r} \tau_n F_{U_{n-1}^{(k)}}(U) - \sum_{i=1}^n \langle \overline{\Lambda}[:,i] \times_k U^T, G[:,i]\rangle, \\
&\overset{(b)}{=} - \min_{G\in\mathbb{R}^{I_1\times\cdots\times I_d\times N}} \left[\sum_{i=1}^N \left\{ H_{\mathbf{X}_i}^*(-G[:,i]) \right\} + \max_{U\in\Sigma_{I_k}^r} \left\{ \langle \overline{\Lambda} \times_k U^T, G\rangle - \tau_n F_{U_{n-1}^{(k)}}(U) \right\} \right], \\
&\overset{(c)}{=} - \min_{G\in\mathbb{R}^{I_1\times\cdots\times I_d\times N}} \sum_{i=1}^N \left\{ H_{\mathbf{X}_i}^*(-G[:,i]) \right\} + \tau_n F_{U_{n-1}^{(k)}}^*(G \times_{\neq k} \overline{\Lambda}/\tau_n).
\end{aligned}$$

Here, (a) uses strong duality for convex objectives; (b) uses the fact that

$$(\overline{\Lambda} \times_k U^T) = [\overline{\Lambda}[:,1], \ldots, \overline{\Lambda}[:,N]] \times_k U^T = \left[\overline{\Lambda}[:,1] \times_k U^T, \ldots, \overline{\Lambda}[:,N] \times_k U^T \right], \tag{90}$$

and (c) follows from the identity $\langle \overline{\Lambda} \times_k U^T, G\rangle = \langle U, G \times_{\neq k} \overline{\Lambda}\rangle$, which is easily verified from the definition. Then we can conclude similarly as in the proof of Proposition B.1 by using Lemma B.4 with $K = \Sigma_{I_1\times\cdots\times I_d}^N$ and

$$\begin{aligned}
&\mathcal{J} : \mathbb{R}^{I_k\times r} \times \mathbb{R}^{I_1\times\cdots\times I_d\times N} \to (-\infty, +\infty] : \\
&(U, Y) \mapsto \begin{cases} \sum_{i=1}^N H_{\mathbf{X}_i}\left(\overline{\Lambda}[:,i] \times_k U^T - Y[:,i]\right) + \tau_n F_{U_{n-1}^{(k)}}(U) & \text{if } \overline{\Lambda} \times_k U^T \in Y + K, \\ +\infty, & \text{if } \overline{\Lambda} \times_k U^T \notin Y + K. \end{cases}
\end{aligned}$$

□

D. Auxiliary lemmas

Lemma D.1. *Fix $g \in \mathbb{R}^r$ and let $\Sigma_r := \{(x_1, \ldots, x_r) \in \mathbb{R}_{\geq 0}^r : \sum_{i=1}^r x_i = 1\}$. The optimality condition of the problem*

$$\sup_{\lambda\in\Sigma_r} \langle g, \lambda\rangle - \frac{1}{2}\|\lambda - \lambda_0\|_F^2 \tag{91}$$

is given as

$$\lambda^* = (g + \lambda_0 - c1_r)_+ \tag{92}$$

where c is a constant chosen to satisfy $\lambda^* \in \Sigma_r$.

Proof. As the cost function of (91) is strictly concave and Σ_r is a closed set, there exists a unique maximizer $\lambda^* \in \Sigma_r$ of (91). For any $\epsilon \in [0,1]$ and $\lambda \in \Sigma_r$, consider

$$h(\epsilon) := \langle g, \lambda^* + \epsilon(\lambda - \lambda^*)\rangle - \frac{1}{2}\|\lambda^* + \epsilon(\lambda - \lambda^*) - \lambda_0\|_F^2.$$

Noting that $\lambda^* + \epsilon(\lambda - \lambda^*)$ is also in Σ_r for any $\epsilon \in [0,1]$.

As $h(\epsilon)$ attains its maximum at $\epsilon = 0$, we have that for all $\lambda \in \Sigma_r$

$$0 \geq h'(0) = \langle g - \lambda^* + \lambda_0, \lambda - \lambda^*\rangle. \tag{93}$$

For $I_1 := \{i \in \{1,2,\cdots,r\} : \lambda^*[i] > 0\}$ and $I_2 := \{i \in \{1,2,\cdots,r\} : \lambda^*[i] = 0\}$, we obtain

$$0 \geq \sum_{i\in I_1}(g - \lambda^* + \lambda_0)[i] \times (\lambda - \lambda^*)[i] + \sum_{i\in I_2}(g - \lambda^* + \lambda_0)[i] \times \lambda[i]. \tag{94}$$

As $\lambda \in \Sigma_r$ is arbitrary, there exists a constant $c \in \mathbb{R}$ such that for $i \in I_1$

$$(g - \lambda^* + \lambda_0)[i] = c. \tag{95}$$

This yields that

$$0 \geq \sum_{i\in I_1} c \times (\lambda - \lambda^*)[i] + \sum_{i\in I_2}(g - \lambda^* + \lambda_0)[i] \times \lambda[i], \tag{96}$$

$$= \sum_{i\in I_2}(g - \lambda^* + \lambda_0[i] - c) \times \lambda[i]. \tag{97}$$

The last equality is due to $\lambda, \lambda^* \in \Sigma_r$, As a consequence, $(g - \lambda^* + \lambda_0)[i] \leq c$ and we conclude. □

Lemma D.2. *Let $(a_n)_{n\geq 0}$ and $(b_n)_{n\geq 0}$ be sequences of nonnegative real numbers such that $\sum_{n=0}^{\infty} a_n b_n < \infty$. Then*

$$\min_{1\leq k\leq n} b_k \leq \frac{\sum_{k=0}^{\infty} a_k b_k}{\sum_{k=1}^{n} a_k} = O\left(\left(\sum_{k=1}^{n} a_k\right)^{-1}\right). \tag{98}$$

Proof. The assertion follows from noting that

$$\left(\sum_{k=1}^{n} a_k\right)\min_{1\leq k\leq n} b_k \leq \sum_{k=1}^{n} a_k b_k \leq \sum_{k=1}^{\infty} a_k b_k < \infty. \tag{99}$$

□

Lemma D.3 (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 $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p$,*

$$\left|f(\boldsymbol{\theta}') - f(\boldsymbol{\theta}) - \langle \nabla f(\boldsymbol{\theta}),\, \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle\right| \leq \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2. \tag{100}$$

Proof. This is a classical Lemma (see, e.g., Lem 1.2.3 in (Nesterov, 1998)). We include a proof of this statement for completeness. First write

$$f(\boldsymbol{\theta}') - f(\boldsymbol{\theta}) = \int_0^1 \left\langle \nabla f\left(\boldsymbol{\theta} + s(\boldsymbol{\theta}' - \boldsymbol{\theta})\right),\, \boldsymbol{\theta}' - \boldsymbol{\theta}\right\rangle ds. \tag{101}$$

By Cauchy-Schwarz inequality and L-Lipscthizness of ∇f,

$$\left| \int_0^1 \left\langle \nabla f\left(\boldsymbol{\theta} + s(\boldsymbol{\theta}' - \boldsymbol{\theta})\right), \boldsymbol{\theta}' - \boldsymbol{\theta} \right\rangle - \int_0^1 \left\langle \nabla f\left(\boldsymbol{\theta}\right), \boldsymbol{\theta}' - \boldsymbol{\theta} \right\rangle ds \right| \leq \int_0^1 \left\| \nabla f\left(\boldsymbol{\theta} + s(\boldsymbol{\theta}' - \boldsymbol{\theta})\right) - \nabla f\left(\boldsymbol{\theta}\right) \right\| \left\| \boldsymbol{\theta}' - \boldsymbol{\theta} \right\| ds \tag{102}$$

$$\leq \int_0^1 Ls \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 \, ds \tag{103}$$

$$= \frac{L}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2. \tag{104}$$

Then the assertion follows. □

Lemma D.4 (Second-Order Growth Property)**.** *Let $g : \mathbb{R}^p \to [0, \infty)$ be μ-strongly convex and let $\boldsymbol{\Theta}$ is a convex subset of $\mathbb{R}^p$. Let $\boldsymbol{\theta}^*$ denote the minimizer of g over $\boldsymbol{\theta}$. Then for all $\boldsymbol{\theta} \in \boldsymbol{\theta}$,*

$$g(\boldsymbol{\theta}) \geq g(\boldsymbol{\theta}^*) + \frac{\mu}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2. \tag{105}$$

Proof. See Lem. B.5 in (Mairal, 2013). □

Lemma D.5 (Characterization of weak convexity)**.** *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a smooth function. Fix a convex set $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ and $\rho > 0$. The following conditions are equivalent.*

(i) *(Weak convexity) $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}) + \frac{\rho}{2}\|\boldsymbol{\theta}\|^2$ is convex on $\boldsymbol{\Theta}$;*

(ii) *(Hypermonotonicity) $\langle \nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle \geq -\rho \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$;*

(iii) *(Quadratic lower bound) $f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') \geq \langle \nabla f(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle - \frac{\rho}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$.*

Proof. See Lem. B.2 in (Lyu, 2022). See also Thm. 7 in (Daniilidis & Malick, 2005) for an equivalent statement for a more general case of locally Lipschitz functions. □

Lemma D.6. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a function such that ∇f is L-Lipscthiz for some $L > 0$. Then f is L-weakly convex, that is, $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}) + \frac{L}{2}\|\boldsymbol{\theta}\|^2$ is convex.*

Proof. Follows immediately by Lemmas D.3 and D.5. □