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 \(O(\varepsilon^{-1})\) 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 \(\Theta\) within \(\varepsilon\)-stationary point within \(O(\varepsilon^{-1})\) iterations.

1. Introduction
Consider the minimization of a continuous function \(f : \mathbb{R}^1 \times \cdots \times \mathbb{R}^d \rightarrow [0, \infty)\) on a cartesian product of convex sets \(\Theta = \Theta(1) \times \cdots \times \Theta(m)\):

\[
\theta^* \in \arg \min_{\theta \in \Theta} f(\theta_1, \ldots, \theta_m). \tag{1}
\]

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 \(i\)th block of the parameter after \(n\) updates. Write

\[
\begin{align*}
\theta_n^{(i-1)} &:= (\theta_n^{(1)}, \ldots, \theta_n^{(i)}, \theta_n^{(i+1)}, \ldots, \theta_n^{(m)}), \\
\theta_n^{(i)} &:= (\theta_n^{(1)}, \ldots, \theta_n^{(i-1)}, \theta_n^{(i+1)}, \ldots, \theta_n^{(m)}), \\
f_n^{(i)}(\theta) &:= f(\theta_n^{(1)}, \ldots, \theta_n^{(i-1)}, \theta_n^{(i+1)}, \ldots, \theta_n^{(m)}).
\end{align*}
\]

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

\[
\theta_n^{(i)} \leftarrow \arg \min_{\theta \in \Theta(i)} g_n^{(i)}(\theta) := f_n^{(i)}(\theta) + \frac{\lambda_n}{2} \|\theta - \theta_n^{(i-1)}\|^2, \tag{3}
\]

where \(\lambda_n \geq 0\) is called proximal regularization coefficient and \(\|\cdot\|\) denotes the Frobenius norm. The proximal regularizer \(\lambda_n\|\theta - \theta_n^{(i-1)}\|^2\) ensures that the next block iterate \(\theta_n^{(i)}\) is not too far from the previous iterate \(\theta_n^{(i-1)}\). 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
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 \( \lambda_n \equiv \text{Const.} \) 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 \( \varepsilon \)-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 \( \varepsilon \)-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 \( \theta^* \in \Theta \) is a *stationary point* of a function \( f \) over \( \Theta \) if

\[
\inf_{\theta \in \Theta} \langle \nabla f(\theta^*), \theta - \theta^* \rangle \geq 0,
\]

where \( \langle \cdot, \cdot \rangle \) denotes the dot product on \( \mathbb{R}^{I_1 + \cdots + I_m} \supseteq \Theta \). This is equivalent to saying that \( -\nabla f(\theta^*) \) is in the normal cone of \( \Theta \) at \( \theta^* \). If \( \theta^* \) is in the interior of \( \Theta \), then it implies \( \| \nabla f(\theta^*) \| = 0 \). 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 \( \varepsilon \)-approximate solution scales with the desired precision \( \varepsilon \). More precisely, we say \( \theta^* \in \Theta \) is an \( \varepsilon \)-approximate
stationary point of $f$ over $\Theta$ if

$$\inf_{\theta \in \Theta} \left\langle \nabla f(\theta^*), \frac{\theta - \theta^*}{\|\theta - \theta^*\|} \right\rangle \leq \sqrt{\varepsilon}.$$  \hspace{1cm} (5)

This notion of $\varepsilon$-approximate solution is consistent with the corresponding notion for unconstrained problems. Indeed, if $\theta^*$ is an interior point of $\Theta$, then (5) reduces to $\|\nabla f(\theta^*)\|^2 \leq \varepsilon$. 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 $N_\varepsilon$ of an algorithm computing $(\theta_n)_{n \geq 1}$ for solving (1) as

$$N_\varepsilon := \sup_{\theta_0 \in \Theta} \inf_{n} \left\{n \mid \theta_n \text{ is an } \varepsilon\text{-approximate stationary point of } f \text{ over } \Theta \right\},$$  \hspace{1cm} (6)

where $(\theta_n)_{n \geq 0}$ is a sequence of estimates produced by the algorithm with an initial estimate $\theta_0$. Note that $N_\varepsilon$ gives the worst-case bound on the number of iterations for an algorithm to achieve an $\varepsilon$-approximate solution due to the supremum over the initialization $\theta_0$ 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 $i = 1, 2, \ldots, m$, there exists a constant $L^{(i)} > 0$ such that the function $f : \Theta = \Theta^{(1)} \times \cdots \times \Theta^{(m)} \to [0, \infty)$ is $L^{(i)}$-smooth in each block coordinate $i$, that is, the function $\theta \mapsto \nabla f(\theta^{(1)}, \ldots, \theta^{(i-1)}, \theta^i, \theta^{(i+1)}, \ldots, \theta^{(m)})$ is $L^{(i)}$-Lipschitz in $\Theta^{(i)}$ for any $\theta^{(j)} \in \Theta^{(j)}$, $j = 1, 2, \ldots, i - 1, i + 1, \ldots, m$.

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

$$\Delta_n := \max_{1 \leq i \leq m} \left( g_n^{(i)}(\theta_n^{(i)}) - \inf_{\theta \in \Theta^{(i)}} g_n^{(i)}(\theta) \right),$$  \hspace{1cm} (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 $\Delta_n$ are summable, that is, $\sum_{n=1}^{\infty} \Delta_n < \infty$.

We now state our main result for BCD-PR.

Theorem 3.4. Let $(\theta_n)_{n \geq 0}$ be an inexact output of (3). Suppose that Assumptions 3.1-3.3 hold. Let $L^{(i)} > 0$ be such that $\nabla f$ 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 $(\theta_n)_{n \geq 0}$ is a stationary point of $f$ over $\Theta$.

(ii) (Worst-case rate of convergence) There exists a constant $M$ independent of $\theta_0$ such that for $n \geq 1$,

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

(iii) (Worst-case iteration complexity) Suppose the optimality gaps are uniformly summable, that is, $\sup_{\theta_0 \in \Theta} \sum_{n=1}^{\infty} \Delta_n < \infty$. Then the worst-case iteration complexity $N_\varepsilon$ 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 $(X_k)_{k \leq n}$, we seek to find a set of product distributions such that each $X_k$ 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 $X_1, \ldots, X_N$ 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’ $D = [D_1, \ldots, D_r] \in \mathbb{R}^{I_1 \times \cdots \times I_d \times r}$ of $r$ ‘atoms’ so that each data tensor $X_i$ can be linearly approximated by the atoms $D_1, \ldots, D_r$ in the dictionary $D$. Namely, there exists a suitable ‘code matrix’ $\Lambda \in \mathbb{R}^{r \times N}$ such that we have the following approximate factorization:

$$[X_1, \ldots, X_N] \approx [D_1, \ldots, D_r] \times_{d+1} \Lambda$$  \hspace{1cm} (9)

$$\iff X \approx D \times_{d+1} \Lambda,$$

where $\times_{d+1}$ denotes the mode $(d+1)$ tensor-matrix product (see (Kolda & Bader, 2009)) and $X := [X_1, \ldots, X_N]$ denotes the $(d+1)$-mode tensor in $\mathbb{R}^{I_1 \times \cdots \times I_d \times N}$ that concatenates the tensor-valued signals $X_1, \ldots, X_n$ 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 \( \mathbf{X}_1, \ldots, \mathbf{X}_n \) are in fact \( f_1 \)-dimensional vectors. Then (9) becomes the usual matrix factorization formulation for factorizing the data matrix \( \mathbf{X} \in \mathbb{R}^{1 \times \cdot \times 1} \times \mathbb{R}^{1 \times \cdot \times 1} \) into the (matrix) product of a dictionary matrix \( \mathbf{D} \in \mathbb{R}^{1 \times \cdot \times 1} \times \mathbb{R}^{1 \times \cdot \times 1} \) and the code matrix \( \mathbf{A} \in \mathbb{R}^{1 \times \cdot \times 1} \). Aharon, 2006; Mairal et al., 2007; Peyré, 2009).

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

\[
\min_{\mathbf{D} \in \mathbb{R}^{1 \times \cdot \times 1} \times \mathbb{R}^{1 \times \cdot \times 1}, \mathbf{A} \in \mathbb{R}^{1 \times \cdot \times 1} \times \mathbb{R}^{1 \times \cdot \times 1}} \delta\left(\mathbf{X}, \mathbf{D} \times \cdot \times 1 \mathbf{A}\right),
\]

where \( \delta : (\mathbb{R}^{1 \times \cdot \times 1} \times \mathbb{R}^{1 \times \cdot \times 1})^2 \to [0, \infty) \) is a ‘dissimilarity function’ that maps a pair of tensors \( (\mathbf{X}, \mathbf{X}') \) to a nonnegative number \( \delta(\mathbf{X}, \mathbf{X}') \). This function is used to measure the difference between the data tensor \( \mathbf{X} \) and the ‘reconstruction’ \( \mathbf{D} \times \cdot \times 1 \mathbf{A} \). For \( d = 1 \), standard choices of \( \delta \) 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 \( \mathbf{M} \in \mathbb{R}^{1 \times \cdot \times 1} \times \mathbb{R}^{1 \times \cdot \times 1} \) for \( d \)-mode tensors to be the tensor defined by \( \mathbf{M}(J_1, J_2) = \| J_1 - J_2 \|_2 \) for any multi-indices \( J_1, J_2 \in [1] \times \cdot \times [1] \).

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

\[
U(p_1, p_2) := \left\{ T \in \Sigma_{m,n} \left| p_1(i) = \sum_{j=1}^{m} T(i, j), p_2(j) = \sum_{i=1}^{m} T(i, j) \right\} \right.,
\]

the set of all couplings between \( p_1 \) and \( p_2 \).

Now, we can define the Wasserstein distance. Fix a cost tensor \( \mathbf{M} \in \mathbb{R}^{1 \times \cdot \times 1} \times \mathbb{R}^{1 \times \cdot \times 1} \) and let \( \mathbf{M}^2 \in \mathbb{R}^{1 \cdot \cdot \cdot 1} \cdot \cdot \cdot 1 \) denote its matricization (see (Kolda & Bader, 2009)). Fix a parameter \( \gamma \geq 0 \). For \( \mathbf{A}, \mathbf{B} \in \mathbb{R}^{1 \times \cdot \times 1} \times \mathbb{R}^{1 \times \cdot \times 1} \), define

\[
W_\gamma(\mathbf{A}, \mathbf{B}) := W_\gamma(\text{vec}(\mathbf{A}), \text{vec}(\mathbf{B})) := \min_{T \in U(\text{vec}(\mathbf{A}), \text{vec}(\mathbf{B}))} \langle \mathbf{M}^2, T \rangle + \gamma \langle T, \log T \rangle,
\]

where \( \text{vec}(\mathbf{A}) \) and \( \text{vec}(\mathbf{B}) \) denote the vectorization of \( \mathbf{A} \) and \( \mathbf{B} \), respectively. When \( \gamma = 0 \), \( W_\gamma \) 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 \( \mathbf{X}_1, \ldots, \mathbf{X}_N \) describe \( d \)-dimensional probability mass functions. Namely, we denote

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

We can think of an element \( \mathbf{X} \) of \( \Sigma_{I_1, \ldots, I_d} \) as the joint probability mass function of \( d \) discrete random variables \( Y_1, \ldots, Y_d \) where each \( Y_i \) takes values from \( \{1, \ldots, I_i\} \). For this reason, we will call an element of \( \Sigma_{I_1, \ldots, I_d} \) simply as a ‘\( d \)-dimensional joint distribution’. We also denote by \( \Sigma_{I_1, \ldots, I_d}^N \) the \( N \)-fold product of \( \Sigma_{I_1, \ldots, I_d} \), which we identify as a subset of \( \mathbb{R}^{I_1 \times \cdot \times 1 \times I_d \times \cdot \times 1} \) 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 \( \delta \) in (10) should measure the dissimilarity between two tuples of \( d \)-dimensional joint distribution. By using the entropy-regularized Wasserstein distance \( W_\gamma \) (see (11)), we formulate the \( d \)-dimensional Wasserstein Dictionary Learning (dWDL) as (9), where the dictionary atoms \( \mathbf{D}_1, \ldots, \mathbf{D}_r \) are taken to be \( d \)-dimensional joint distributions (elements of \( \Sigma_{I_1, \ldots, I_d} \)) and the dissimilarity function \( \delta : \Sigma_{I_1, \ldots, I_d}^N \times \Sigma_{I_1, \ldots, I_d} \to [0, \infty) \) is

\[
\delta([X_1, \ldots, X_N], [X'_1, \ldots, X'_N]) := \sum_{i=1}^{N} W_\gamma(X_i, X'_i).
\]

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

\[
(d\text{WDL}) \quad \hat{D} = \{ \mathbf{D}_1, \ldots, \mathbf{D}_r \} \in \Sigma_{I_1, \ldots, I_d}^N \min_{\Lambda \in \Sigma_{I_1, \ldots, I_d}^N} f_W(\hat{D}, \Lambda),
\]

where \( f_W(\hat{D}, \Lambda) := \sum_{i=1}^{N} W_\gamma(\mathbf{X}_i, \mathbf{D} \times d+1 \Lambda[i, i]) \).

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 \((\Lambda_{n-1}, D_{n-1})\), we compute the updated estimate \((\Lambda_n, D_n)\) by solving convex sub-problems as follows:

\[
\Lambda_n \in \arg \min_{\Lambda \in \Sigma_r^N} f_W(D_{n-1}, \Lambda) + \frac{\tau_n}{2} \|\Lambda - \Lambda_{n-1}\|_F^2. \tag{13}
\]

\[
D_n \in \arg \min_{D \in \Sigma_I^{1 \times \cdots \times I_d}} f_W(D, \Lambda_n) + \frac{\tau_n}{2} \|D - D_{n-1}\|_F^2. \tag{14}
\]

For the standard nonnegative matrix factorization using the Frobenius norm instead of the Wasserstein norm, solving 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 \(W_\gamma\), 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: \(\theta_0 = (D_0, \Lambda_0) \in \Sigma_I^{1 \times \cdots \times I_d} \times \Sigma_r^{N}\) (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: \(\Lambda_n \leftarrow \text{Algorithm 2 with input } (D_{n-1}, \Lambda_{n-1})\) \hspace{1cm} (15)
4: \(D_n \leftarrow \text{Algorithm 3 with input } (D_{n-1}, \Lambda_n)\) \hspace{1cm} (16)
5: end for
6: output: \(\theta_N\)

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 \(X, Y \in \Sigma_I^{1 \times \cdots \times I_d}\) and for given \(\Lambda_0 \in \Sigma_r\),

\[
H_X(y) := W_\gamma(X, y) \quad \text{and} \quad F_{\Lambda_0}(\lambda) := \begin{cases} \frac{1}{2} \|\lambda - \Lambda_0\|_F^2 & \text{for } \lambda \in \Sigma_r, \\ +\infty & \text{otherwise}. \end{cases} \tag{17}
\]

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

\[
\min_{\Lambda \in \Sigma_r^N} \sum_{i=1}^N \left\{ H_{X_i}(D_{n-1} \times_{d+1} \Lambda[: i]) \right\} + \tau_n F_{\Lambda_{n-1}[: i]}(\Lambda[: i]). \tag{18}
\]

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 \(H_X\) above involves finding an optimal transport map \(T \in U(\text{vec}(A), \text{vec}(B))\). 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:

\[
\begin{align*}
\min_{G \in \mathbb{R}^{I_1 \times \cdots \times I_d \times N}} & \sum_{i=1}^N \left( H_{X_i}^\ast(-G[: i]) + \tau_n F_{\Lambda_{n-1}[: i]}^\ast(D_{n-1} \times_{d+1} G[: i] / \tau_n) \right).
\end{align*} \tag{19}
\]

Here, the conjugate \(f^\ast\) of \(f\) is defined as

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

This dual problem can be solved without having to deal with a matrix-solving problem, as in the primal one (see Cuturi & Peyré, 2014). We postpone further discussion about the conjugate functions \(H^\ast\) and \(F^\ast\) to the subsequent sections.

Algorithm 2 Solving for \(\Lambda\)

1: Input: \(\theta_{n-1} = (D_{n-1}, \Lambda_{n-1}) \in \Sigma_I^{1 \times \cdots \times I_d} \times \Sigma_r^{N}\) (current estimate); \((\tau_n)_{n \geq 1}\);
2: Update estimate \(\Lambda_{n-1}\) by
3: \(G_n^\circ \leftarrow \text{the minimizer of } (19)\)
4: \(\Lambda_n \leftarrow \left( \Lambda_{n-1} + \frac{D_{n-1} \times_{d+1} G_n^\circ}{\tau_n} - J^\circ \otimes c_n \right)_{+}\)

where \(c_n \in \mathbb{R}^{N \times 1}\) is chosen to satisfy \(\Lambda_n \in \Sigma_r^N\) and all entries of \(J^\circ \in \mathbb{R}^{r \times r}\) are one.
5: output: \(\theta_{n-\frac{1}{2}} = (D_{n-1}, \Lambda_n)\)

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

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

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

\[
\begin{align*}
\min_{G \in \mathbb{R}^{I_1 \times \cdots \times I_d \times N}} & \left\{ \sum_{i=1}^N H_{X_i}^\ast(-G[: i]) \right\} + \tau_n F_{\Lambda_{n-1}}^\ast(G \times_{d+1} \Lambda_n^T / \tau_n) \tag{22}
\end{align*}
\]
We prove that our computationally efficient algorithm, Algorithm 3, solves for Theorem 5.1. 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 $R : \mathcal{H} \to \mathcal{K}$,

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

Here, $\mathcal{H}$ and $\mathcal{K}$ are real Hilbert spaces with inner product $\langle \cdot, \cdot \rangle$, and $\mathcal{K}$ 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 and 3.4.

**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 $(\theta_n)_{n \geq 0}$ of Algorithm 1, the following hold:

(i) (Global convergence to stationary points) Every limit point of $(\theta_n)_{n \geq 0}$ is a stationary point of $f_W$ over $\Theta := \Sigma_{I_1 \times \cdots \times I_d} \times \Sigma^N$.

(ii) (Worst-case rate of convergence) There exists a constant $M$ independent of $\theta_0$ such that for $n \geq 1$, (8) in Theorem 3.4 holds.

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

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

**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 $H_X$ given in (17) has a closed form (Cuturi & Peyré 2016): for $g \in \mathbb{R}^{I_1 \times \cdots \times I_d}$ and given $X \in \Sigma_{I_1 \times \cdots \times I_d}$,

\[
H_X^*(g; \Sigma_{I_1 \times \cdots \times I_d}) := \sup_{y \in \Sigma_{I_1 \times \cdots \times I_d}} \langle g, y \rangle - H_X(y) = \gamma \left( \langle X, \log X \rangle + (\gamma \log (1/\gamma)) \right).
\]

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

\[
\nabla H_X^*(g) = \alpha \circ \left( K \frac{X}{K} \right) \in \Sigma_{I_1 \times \cdots \times I_d}. \tag{24}
\]
Therefore, Assumption 3.1 is satisfied. Furthermore, the constraint set $\Sigma_{l_1 \times \cdots \times l_d}$ and $\Sigma_N^I$ satisfy Assumption 3.2.

Next, we compute the per-iteration cost of Algorithms 2 and 3. The dual function of $F_{X_0}$ is given by for $g \in \mathbb{R}^r$

$$F_{X_0}^*(g) := \sup_{\lambda \in \Sigma_r} \langle g, \lambda \rangle - \frac{1}{2} \|\lambda - \lambda_0\|_F^2.$$  

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

$$\lambda^* = (g + \lambda_0 - cl_r)_+$$  

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

$$F_{X_0}^*(g) = \frac{1}{2}(g + \lambda_0 - cl_r)_+ (g + \lambda_0 + cl_r) - \frac{1}{2} \|\lambda_0\|_F^2.$$  

By the duality as in Lem. 7.15 in Santambrogio (2015), its gradient is given as the optimizer (25): \[ \nabla F_{X_0}^*(g) = \lambda = (g + \lambda_0 - cl_r)_+ \in \Sigma_r. \] Therefore, each gradient descent step to solve (19) or (22) requires $O((l_1 \ldots l_d)^2 N)$.

Lastly, (19) and (22) are convex problems, we conclude (iii). \[ \square \]

### 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 $D_1, \ldots, D_d$ that we allow. Namely, the CONDECOMP/PARAFAC (CP)-dictionary learning (Lyu et al., 2020) assumes that each $D_i$ 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 $X \in \mathbb{R}^{l_1 \times \cdots \times l_d}$ is given and fix an integer $r \geq 1$. In the CANDECOMP/PARAFAC (CP) decomposition of $X$ (Kolda & Bader, 2009), we would like to find $r$ loading matrices $U^{(i)} \in \mathbb{R}^{l_i \times r}$ for $i = 1, \ldots, d$ such that the sum of the outer products of their respective columns approximate $X$:

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

where $U^{(i)}[:, k]$ denotes the $k$th column of the $l_i \times r$ loading 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:

$$\arg\min_{U^{(1)} \in \Theta^{(1)}, \ldots, U^{(d)} \in \Theta^{(d)}} f_{\text{CP}}(U^{(1)}, \ldots, U^{(d)})$$

where

$$f_{\text{CP}}(U^{(1)}, \ldots, U^{(d)}) := \|X - [U^{(1)}, U^2, \ldots, U^{(d)}]\|_F^2$$

and $\Theta^{(i)} \subseteq \mathbb{R}^{l_i \times r}$ denotes a compact and convex constraint set and $\lambda_i \geq 0$ is an $\ell_1$-regularizer for the $i$th loading matrix $U^{(i)}$ for $i = 1, \ldots, d$. In particular, by taking $\lambda_i = 0$ and $\Theta^{(i)}$ to be the set of nonnegative $l_i \times r$ matrices with bounded norm for $i = 1, \ldots, d$, (26) reduces to the nonnegative CP decomposition (NCPD) (Shashua & Hazan, 2005; Zafeiriou, 2009).

Also, it is easy to see that $f_{\text{CP}}$ is equal to

$$\|X - \bigotimes_{i=1}^{d} U^{(i)}[:, 1], \bigotimes_{k=1}^{d} U^{(i)}[:, 2], \ldots, \bigotimes_{k=1}^{d} U^{(i)}[:, r]\|_F^2,$$

where $\bigotimes_{k=1}^{d} U^{(i)}[:, r]$ denotes the mode-$d$ product of the $d$-mode tensor $X_i$ as $I_d$ observations of $(d - 1)$-mode tensors, and the $R$ rank-1 tensors in $\bigotimes_{i=1}^{d} U^{(i)}$, 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 $X_i$ 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, $U^{(i)} \in \Sigma_{l_i}$ for $i = 1, \ldots, d$.

$$|X_1, \ldots, X_N| \approx \bigotimes_{i=1}^{d} U^{(1)}_i, \ldots, U^{(d)}_i \times d+1 \Lambda$$

for some code matrix $\Lambda \in \Sigma_{l_d}^N$ where $\bigotimes$ 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 (WCPL):
6.2. Algorithm (WCPLD)

We state our algorithm to solve Wasserstein CP-Dictionary Learning (30). Given the previous estimates $U^{(1)}_{n-1}, \ldots, U^{(d)}_{n-1}$ and $\Lambda_{n-1}$, we compute the updated estimates $U^{(1)}_n, \ldots, U^{(d)}_n$ and $\Lambda_n$ by solving convex sub-problems, iteratively, as follows.

First, let $D_{n-1} = \text{out}(U^{(1)}_{n-1}, \ldots, U^{(d)}_{n-1}) \in \Sigma_{I_1 \times I_2 \times \cdots \times I_d}$. For a given data tensor $X \in \Sigma_{I_1 \times I_2 \times \cdots \times I_d}$, $\tau_n > 0$, and the previous estimates above, the code matrix is updated as follows:

$$\Lambda_n = \arg \min_{\Lambda \in \Sigma_X} \left( \sum_{i=1}^{N} W_\gamma(X_i, (D_{n-1} \times d+1) \Lambda)[i, i] \right) + \frac{\tau_n}{2} \| \Lambda - \Lambda_{n-1} \|^2_F. \quad (31)$$

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

$$\text{out}(U^{(1)}_{n-1}, \ldots, U^{(k-1)}_{n-1}, U^{(k+1)}_{n-1}, \ldots, U^{(d)}_{n-1}, \Lambda^T_n)$$

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

$$U^{(k)}_n = \arg \min_{U \in \Sigma_X} \left( \sum_{i=1}^{N} W_\gamma(X_i, \overline{X}[i, i] \times_k U^T) \right) + \frac{\tau_n}{2} \| U^{(k)} - U^{(k)}_{n-1} \|^2_F. \quad (32)$$

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, $\Lambda \in \Sigma^N_1$ and thus all entries of $\Lambda$ are 1’s, which corresponds to the Wasserstein barycenter problem with equal weights: $\min_{D \in \Sigma_{I_1} \times \cdots \times I_d} \sum_{i=1}^{N} W_\gamma(X_i, D)$.

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](image1)

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
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.

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.

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


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


A. Block Coordinate Descent with proximal regularization

A.1. Proof of Theorem 3.4

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

\[ f_n^{(i)} : \theta \mapsto f(\theta_n^{(1)}, \ldots, \theta_n^{(i-1)}, \theta, \theta_n^{(i+1)}, \ldots, \theta_n^{(m)}), \]  

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 \(g_n^{(i)}(\theta) = f_n^{(i)}(\theta) + \frac{\tau_n^{(i)}}{2} \| \theta - \theta_n^{(i)} \|^2\) is \((\tau_n^{(i)} - L(i))\)-strongly convex. Also, denote

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

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) \(f(\theta_{n-1}) - f(\theta_n) \geq \tau_n^- \| \theta_{n-1} - \theta_n \|^2 - m \Delta_n;\)

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

Proof. Fix \(i \in \{1, \ldots, m\}\). Let \(\theta_n^{(i)}\) 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 \(g_n^{(i)}(\theta_n^{(i)}) \leq f_n^{(i)}(\theta_n^{(i-1)}) = g_n^{(i)}(\theta_n^{(i-1)})\), for \(n \geq 1\). Hence we deduce

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

It follows that

\[ f(\theta_{n-1}) - f(\theta_n) \geq \sum_{i=1}^{n} f([\theta_n^{(1)}, \ldots, \theta_n^{(i-1)}, \theta_n^{(i)}, \theta_n^{(i+1)}, \ldots, \theta_n^{(m)}]) - f([\theta_n^{(1)}, \ldots, \theta_n^{(i-1)}, \theta_n^{(i)}, \theta_n^{(i+1)}, \ldots, \theta_n^{(m)}]) \tag{37} \]

\[ = \sum_{i=1}^{n} f_n^{(i)}(\theta_n^{(i-1)}) - f_n^{(i)}(\theta_n^{(i)}) \tag{38} \]

\[ \geq \sum_{i=1}^{m} \left( \tau_n^- \| \theta_n^{(i)} - \theta_n^{(i-1)} \|^2 - \Delta_n \right) = \frac{\tau_n^-}{2} \| \theta_n^{(i)} - \theta_n^{(i-1)} \|^2 - \Delta_n, \tag{39} \]

This shows (i).

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

\[ \sum_{k=1}^{n} \frac{\tau_k^-}{2} \| \theta_k - \theta_k \|^2 \leq \left( \sum_{k=1}^{n} f(\theta_{k-1}) - f(\theta_k) \right) + m \sum_{n=1}^{\infty} \Delta_n = f(\theta_0) + m \sum_{n=1}^{\infty} \Delta_n < \infty, \tag{40} \]

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

\[ \sum_{n=1}^{\infty} |\nabla f(\theta_n + 1) \| \theta_n - \theta_n + 1) \| \leq \frac{L + 2}{2} \sup_{\theta \in \Theta} f(\theta) + 3m \sum_{n=1}^{\infty} \Delta_n < \infty. \]
Proof. According to Assumptions 3.1 and 3.2, it follows that $\nabla f$ over $\Theta$ is Lipschitz with Lipshitz constant $L$. Hence by Lemma D.3, for all $t \geq 1$,

$$|f(\theta_t) - f(\theta_{t+1}) - \langle \nabla f(\theta_{t+1}), \theta_t - \theta_{t+1} \rangle| \leq \frac{L}{2} \|\theta_t - \theta_{t+1}\|^2_F.$$  

Using Proposition A.1, it follows that

$$|f(\theta_{n-1}) - f(\theta_n)| \leq f(\theta_{n-1}) - f(\theta_n) + 2m\Delta_n.$$  

Hence this yields

$$|\langle \nabla f(\theta_{n+1}), \theta_n - \theta_{n+1} \rangle| \leq \frac{L}{2} \|\theta_n - \theta_{n+1}\|^2_F + |f(\theta_n) - f(\theta_{n+1})|$$

$$\leq \frac{L}{2} \|\theta_n - \theta_{n+1}\|^2_F + f(\theta_n) - f(\theta_{n+1}) + 2m\Delta_n$$  

for $n \geq 1$. Also note that $\sum_{t=1}^n f(\theta_t) - f(\theta_{t+1}) = f(\theta_1) - f(\theta_{n+1}) \leq f(\theta_1)$. Hence

$$\sum_{n=0}^{\infty} |\langle \nabla f(\theta_{n+1}), \theta_n - \theta_{n+1} \rangle| \leq \frac{L}{2} \left( \sum_{n=0}^{\infty} \|\theta_n - \theta_{n+1}\|^2_F \right) + f(\theta_0) + 2m \sum_{n=1}^{\infty} \Delta_n$$

$$\leq \frac{L}{2} \left( \sum_{n=0}^{\infty} \tau_n^- \|\theta_n - \theta_{n+1}\|^2_F \right) + f(\theta_0) + 2m \sum_{n=0}^{\infty} \Delta_n$$

$$\leq 2f(\theta_0) + 3m \sum_{n=1}^{\infty} \Delta_n < \infty,$$

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

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

$$\sup_{n \geq 1} f(\theta_n) \leq f(\theta_1) + T < \infty.$$  

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 $\theta_0 \in \Theta$ such that for all $n \geq 1$,

$$\langle \nabla f(\theta_{n+1}), \theta_{n+1} - \theta_n \rangle \leq b_{n+1} \inf_{\theta \in \Theta} \left\langle \nabla f(\theta_n), \frac{\theta - \theta_n}{\|\theta - \theta_n\|} \right\rangle + c_1 b_{n+1} \|\theta_n - \theta_{n+1}\|$$  

$$\quad + c_1 \|\theta_{n+1} - \theta_n\|^2 + c_2 (L + \tau_{n+1}) b_{n+1}^2 + \Delta_{n+1}.$$  

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

$$f^{(i)}_{n+1}(\theta^{(i)}_{n+1}) + \frac{\tau_{n+1}^{(i)}}{2} \|\theta^{(i)}_{n+1} - \theta^{(i)}_n\|^2 - \Delta_{n+1} \leq f^{(i)}_{n+1}(\theta^{(i)}_n) + \tau_{n+1}^{(i)} \|\theta^{(i)}_n - \theta^{(i)}_{n+1}\|^2$$

$$\leq f^{(i)}_{n+1}(\theta^{(i)} + a(\theta^{(i)} - \theta^{(i)}_n)) + \frac{\tau_{n+1}^{(i)} a^2}{2} \|\theta^{(i)} - \theta^{(i)}_n\|^2.$$  

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

$$\left\langle \nabla f^{(i)}_{n+1}(\theta^{(i)}_{n+1}), \theta^{(i)}_{n+1} - \theta^{(i)}_n \right\rangle \leq a \left\langle \nabla f^{(i)}_{n+1}(\theta^{(i)}_n), \theta^{(i)} - \theta^{(i)}_n \right\rangle$$

$$+ \frac{L^{(i)}}{2} \|\theta^{(i)}_{n+1} - \theta^{(i)}_n\|^2 + \frac{L^{(i)}}{2} \|\theta^{(i)} - \theta^{(i)}_n\|^2 + \frac{\tau^{(i)}_{n+1} a^2}{2} \|\theta^{(i)} - \theta^{(i)}_n\|^2 + \Delta_{n+1}.$$  

Adding up these inequalities for \( i = 1, \ldots, m \),
\[
\left\langle \nabla f_{n+1}(\theta_n^{(1)}), \ldots, \nabla f_{n+1}(\theta_n^{(m)}) \right\rangle, \theta_{n+1} - \theta_n \leq a \left\langle \left[ \nabla f_{n+1}(\theta_n^{(1)}), \ldots, \nabla f_{n+1}(\theta_n^{(m)}) \right], \theta - \theta_n \right\rangle + \frac{L}{2} \| \theta_{n+1} - \theta_n \|^2 + \frac{(L + \tau_{n+1}a^2)^2}{2} \| \theta - \theta_n \|^2 + \Delta_{n+1}. \tag{51}
\]
Since for each \( i = 1, \ldots, m \) \( \nabla f \) is \( L^i \)-Lipschitz in the \( i \)th block coordinate, we have
\[
\| \nabla f(\theta_n) - \nabla f(\theta_n^{(i)}) \| \leq L^i \| \theta_n - \theta_n^{(i)} \|. \tag{53}
\]
Hence there exists constants \( c_1, c_2 > 0 \) independent of \( \theta_0 \in \Theta \), such that
\[
\left\langle \nabla f(\theta_{n+1}), \theta_{n+1} - \theta_n \right\rangle \leq a \left\langle \nabla f(\theta_n), \theta - \theta_n \right\rangle + amL \| \theta_n - \theta_{n+1} \| \cdot \| \theta - \theta_n \| + c_1 \| \theta_{n+1} - \theta_n \|^2 + c_2(L + \tau_{n+1}a^2) \| \theta - \theta_n \|^2 + \Delta_{n+1}. \tag{54}
\]
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 \( a \left\langle \nabla f(\theta_n), \theta - \theta_n \right\rangle \), whose absolute value is bounded above by \( a \| \nabla f(\theta_n) \| \| \theta - \theta_n \| \). By Proposition A.3 and Assumption 3.3, \( \| \nabla f(\theta_n) \| \) is uniformly bounded, so this is bounded above by \( ac_3 \| \theta - \theta_n \| \) 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 \( a \geq 0 \). In particular, we can choose \( a = b_{n+1}/\| \theta - \theta_n \| \). This and using \( \| \theta - \theta_n \| \leq b_{n+1} \) yield
\[
\left\langle \nabla f(\theta_{n+1}), \theta_{n+1} - \theta_n \right\rangle \leq b_{n+1} \left\langle \nabla f(\theta_n), \frac{\theta - \theta_n}{\| \theta - \theta_n \|} \right\rangle + c_0 \| \theta_n - \theta_{n+1} \|b_{n+1} + c_1 \| \theta_{n+1} - \theta_n \|^2 + c_2(L + \tau_{n+1}a^2) \| \theta - \theta_n \|^2 + \Delta_{n+1}, \tag{56}
\]
where we wrote \( c_0 := mL \).

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

**Proposition A.5 (Optimality gap for iterates).** For each \( n \geq 1 \) and \( i \in \{1, \ldots, m\} \), let \( \theta_n^{(i)} \) be the exact minimizer of the \((\tau_n i - L^i)\)-strongly convex function \( \theta ightarrow g_n^{(i)}(\theta) \) in (3) over the convex set \( \Theta^{(i)} \). Then
\[
\frac{\tau_n}{2} - L^i \| \theta_n^{(i)} - \theta_n^{(i)} \|^2 \leq \Delta_n. \tag{58}
\]

**Proof.** The assertion follows from
\[
\frac{\tau_n}{2} - L^i \| \theta_n^{(i)} - \theta_n^{(i)} \|^2 \leq \frac{g_n^{(i)}(\theta_n^{(i)}) - g_n^{(i)}(\theta_n^{(i)})}{2} \leq \Delta_n \tag{59}
\]
for \( n \geq 1 \). Indeed, the first inequality follows from the second-order growth property (see Lemma D.4) since \( g_n^{(i)} \) is \((\tau_n i - L^i)\)-strongly convex minimized at \( \theta_n^{(i)} \), and the second inequality follows from the definition of optimality gap \( \Delta_n \) in (7). \( \Box \)

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 \( n \geq 1 \) for some \( \delta > 0 \). Also assume \( \tau_n i = O(1) \). We first show (i). Fix a convergent subsequence \( (\theta_{n_k})_{k \geq 1} \) of \( (\theta_n)_{n \geq 1} \). We wish to show that \( \theta_{\infty} = \lim_{k \to \infty} \theta_{n_k} \)
is a stationary point of \( f \) over \( \Theta \). To this end, for each \( i \in \{1, \ldots, m\} \), let \( \theta_n^{(i)} \) 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 \( \theta_n^{(i)} \), we have
\[
\left\langle \nabla g_n^{(i)}(\theta_n^{(i)}), \theta - \theta_n^{(i)} \right\rangle = \left\langle \nabla f_n^{(i)}(\theta_n^{(i)}), \tau_n^{(i)}(\theta_n^{(i)} - \theta_n^{(i)}) \right\rangle + \tau_n^{(i)}(\theta_n^{(i)} - \theta_n^{(i)}) \geq 0 \quad \forall \theta \in \Theta^{(i)}.
\]
(60)

Let \( T := m \sum_{k=1}^{\infty} 1 \), which is finite by Assumption 3.3. Recall that by Proposition A.1, we have
\[
\sup_{n \geq 1} f(\theta_n) \leq f(\theta_1) + T < \infty.
\]
(61)

Let \( K := \{ \theta : f(\theta) \leq f(\theta_1) + T \} \) and let \( K(T) := \{ \theta : \exists \theta' \in K \text{ s.t. } \|\theta - \theta'\| \leq T \} \) 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 \( i \)th block coordinate, \( \|\nabla g_n^{(i)}\| \) is uniformly bounded over \( \theta \in K(T) \) by some constant, say, \( L_K > 0 \). Now observe that
\[
\left\langle \nabla g_n^{(i)}(\theta_n^{(i)}), \theta - \theta_n^{(i)} \right\rangle \geq \left\langle \nabla g_n^{(i)}(\theta_n^{(i)}), \theta - \theta_n^{(i)} \right\rangle - \left| \left\langle \nabla g_n^{(i)}(\theta_n^{(i)}), \theta_n^{(i)} - \theta_n^{(i)} \right\rangle \right|
\]
(62)
\[
\geq -\|\nabla g_n^{(i)}(\theta_n^{(i)})\| \|\theta_n^{(i)} - \theta_n^{(i)}\|
\]
(63)
\[
\geq -L_K \Delta_n.
\]
(64)

Next, using \( L^{(i)} \)-Lipschitzness of \( f \) in the \( i \)th block coordinate and Proposition A.5, we have
\[
\left| \left\langle \nabla g_n^{(i)}(\theta_n^{(i)}), \theta - \theta_n^{(i)} \right\rangle - \left\langle \nabla g_n^{(i)}(\theta_n^{(i)}), \theta - \theta_n^{(i)} \right\rangle \right|
\]
(65)
\[
\leq \left| \left\langle \nabla f_n^{(i)}(\theta_n^{(i)}), \theta_n^{(i)} - \theta_n^{(i)} \right\rangle \right|
\]
(66)
\[
\leq \left( \|\nabla f_n^{(i)}(\theta_n^{(i)})\| + \tau_n^{(i)} \right) \|\theta_n^{(i)} - \theta_n^{(i)}\| \|\theta_n^{(i)} - \theta_n^{(i)}\|
\]
(67)
\[
\leq (L^{(i)} + \tau_n^{(i)}) \|\theta_n^{(i)} - \theta_n^{(i)}\| \|\nabla f_n^{(i)}(\theta_n^{(i)})\|
\]
(68)
\[
\leq (L^{(i)} + \tau_n^{(i)}) \|\theta_n^{(i)} - \theta_n^{(i)}\| \frac{2\Delta_n}{\tau_n^{(i)} - L^{(i)}} = \sqrt{\frac{8\tau_n^{(i)} \Delta_n}{1 - L^{(i)} / \tau_n^{(i)}}},
\]
(69)
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, \( \|\theta_n^{(i)} - \theta_n^{(i)}\| \) is uniformly bounded in \( k \) since \( \theta_n^{(i)} \) converges as \( k \to \infty \). Hence
\[
\liminf_{k \to \infty} \left\langle \nabla g_n^{(i)}(\theta_n^{(i)}), \theta - \theta_n^{(i)} \right\rangle \geq 0 \quad \forall \theta \in \Theta^{(i)}.
\]
(70)

Note that by Proposition A.1 (ii) and \( \tau_n^{(i)} = O(1) \), we get \( \tau_n^{(i)} \|\theta_n^{(i)} - \theta_n^{(i-1)}\| = o(1) \). So if we write \( \theta_{\infty} = [\theta_{\infty}^{(1)}, \ldots, \theta_{\infty}^{(m)}] \),

For each \( \theta \in \Theta^{(i)} \), by the hypothesis, we get
\[
\lim_{k \to \infty} \left| \left\langle \nabla f_n^{(i)}(\theta_n^{(i)}), \theta - \theta_n^{(i)} \right\rangle \right|
\]
(71)
\[
\leq \lim_{k \to \infty} 2\tau_n^{(i)} \|\theta_n^{(i)} - \theta_n^{(i)}\| \|\theta_n^{(i)} - \theta_n^{(i)}\|
\]
(72)
\[
= 2\|\theta - \theta_{\infty}\| \lim_{k \to \infty} \tau_n^{(i)} \|\theta_n^{(i)} - \theta_n^{(i)}\| = 0.
\]
(73)

It follows that, for each \( \theta \in \Theta^{(i)} \), using the continuity of \( \nabla f \) in Assumption 3.2,
\[
\left\langle \nabla f(\theta^{(1)}, \ldots, \theta_{\infty}^{(1)}, \theta_{\infty}^{(i)}, \theta_{\infty}^{(i+1)}, \ldots, \theta_{\infty}^{(m)}), \theta - \theta_{\infty}^{(i)} \right\rangle \geq 0.
\]
(74)

This holds for all \( i = 1, \ldots, m \). Therefore we verify \( \langle \nabla f(\theta_{\infty}), \theta - \theta_{\infty}\rangle \geq 0 \) for all \( \theta \in \Theta \), which means that \( \theta_{\infty} \) is a stationary point of \( f \) over \( \Theta \), as desired. This shows (i).
Next, we show (ii). Let \( b_n \) be any square-summable sequence of positive numbers. By Cauchy-Schwarz inequality,
\[
\sum_{k=1}^{n} b_k \| \theta_k - \theta_{k+1} \| \leq \left( \sum_{k=1}^{n} b_k^2 \right)^{1/2} \left( \sum_{k=1}^{n} \| \theta_k - \theta_{k+1} \|^2 \right)^{1/2}.
\]
(75)
Then by Proposition A.1, the right hand side is uniformly bounded in \( n \geq 1 \), so we see that the left hand side is also uniformly bounded in \( n \). Hence using Propositions A.1 and A.2,
\[
\sum_{n=1}^{\infty} b_{n+1} \left[ - \inf_{\theta \in \Theta} \left\langle \nabla f(\theta_n), \frac{\theta - \theta_n}{\| \theta - \theta_n \|} \right\rangle \right] \leq C \left( \sup_{\theta \in \Theta} f(\theta) + \sum_{n=1}^{\infty} \Delta_n(\theta_0) + \sum_{n=1}^{\infty} b_n^2 + \sum_{n=1}^{\infty} b_n \| \theta_n - \theta_{n+1} \| \right)
\]
for some constant \( C > 0 \) independent of \( \theta_0 \), and the right hand side is finite. Thus by taking \( b_n = 1/(\sqrt{n} \log n) \), using Lemma D.2, we deduce
\[
\min_{1 \leq k \leq n} \left[ - \inf_{\theta \in \Theta} \left\langle \nabla f(\theta_k), \frac{\theta - \theta_k}{\| \theta - \theta_k \|} \right\rangle \right] \leq \frac{M + c \sum_{n=1}^{\infty} \Delta_n(\theta_0)}{\sqrt{n} / \log n}
\]
(77)
for some constants \( M, c > 0 \) independent of \( \theta_0 \). This shows (ii).

Lastly, we show (iii). Assume \( \sup_{\theta \in \Theta} \sum_{n=1}^{\infty} \Delta_n(\theta_0) < \infty \). Then the above implies that for some constant \( M' > 0 \) independent of \( \theta_0 \),
\[
\min_{1 \leq k \leq n} \sup_{\theta \in \Theta} \left[ - \inf_{\theta \in \Theta} \left\langle \nabla f(\theta_k), \frac{\theta - \theta_k}{\| \theta - \theta_k \|} \right\rangle \right]^2 \leq \frac{M'(\log n)^2}{n}.
\]
(78)
Then one can conclude (iii) by using the fact that \( n \geq 2e^{-1}(\log e^{-1})^2 \) implies \( (\log n)^2/n \leq \varepsilon \) for all sufficiently small \( \varepsilon > 0 \). This completes the proof. \( \square \)

### 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 \( (D_{n-1}, \Lambda_{n-1}) \in \Sigma_{I_1 \times \ldots \times I_d} \times \Sigma_{N}^r \) and \( \tau_n > 0 \), let \( \Lambda_n \in \Sigma_{N}^r \) be a solution of (18). Suppose each fiber of \( D_{n-1} \) along the last mode is not identically zero. Then, \( \Lambda_n \) is uniquely determined by
\[
\Lambda_n = \left( \Lambda_{n-1} + \frac{D_{n-1} \times \ldots \times d^G_n}{\tau_n} - J^0 \otimes c_n^0 \right)_+. \tag{79}
\]
Here, \( G_n^0 \in \mathbb{R}^N \) is defined as the unique solution of the dual problem (19), \( c_n^0 \in \mathbb{R}^{N \times 1} \) is chosen to satisfy \( \Lambda_n \in \Sigma_{N}^r \) and all entries of \( J^0 \in \mathbb{R}^{r \times 1} \) are one.

As shown later in the proof, the assumption on \( D_{n-1} \) 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 \( (D_{n-1}, \Lambda_n) \in \Sigma_{I_1 \times \ldots \times I_d} \times \Sigma_{N}^r \), let \( D_n \in \Sigma_{I_1 \times \ldots \times I_d} \) be a solution of (14). Suppose each fiber of \( \Lambda_n \in \Sigma_{N}^r \) along the 2nd mode is not identically zero. Then \( D_n \) is uniquely determined by
\[
D_n = \left( D_{n-1} + \frac{G_n^1}{\tau_n} + d^A_T - J^1 \otimes c_n^1 \right)_+. \tag{80}
\]
Here, \( G_n^1 \) is defined as the unique solution of the dual problem (22), \( c_n^1 \in \mathbb{R}^{r \times 1} \) is chosen to satisfy \( D_n \in \Sigma_{I_1 \times \ldots \times I_d} \) and all entries of \( J^1 \in \mathbb{R}^{I_1 \times \ldots \times I_d \times 1} \) 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 $R : \mathcal{H} \rightarrow \mathcal{K}$ 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 $\mathcal{J} : \mathcal{H} \times \mathcal{K} \rightarrow (-\infty, +\infty]$ is a function given as
\[
\mathcal{L} : \mathcal{H} \times \mathcal{K} \rightarrow [-\infty, +\infty] : (x, v) \mapsto \inf_{y \in \mathcal{K}} (\mathcal{J}(x, y) + (y, v)).
\]
Moreover, $(x, v) \in \mathcal{H} \times \mathcal{K}$ is a saddle point of $\mathcal{L}$ if
\[
\mathcal{L}(x, v) = \sup_{x \in \mathcal{H}} \mathcal{L}(x, K) = \inf_{v \in \mathcal{K}} \mathcal{L}(\mathcal{H}, v).
\]

The primal problem and the dual problem of $\mathcal{J} : \mathcal{H} \times \mathcal{K} \rightarrow (-\infty, +\infty]$ are respectively given as
\[
\min_{x \in \mathcal{H}} \mathcal{J}(x, 0), \quad \text{and} \quad \min_{v \in \mathcal{K}} \mathcal{J}^*(0, v).
\]

We first observe that the primal problem of
\[
\mathcal{J} : \mathcal{H} \times \mathcal{K} \rightarrow (-\infty, +\infty] : (x, y) \mapsto \begin{cases} f(Rx - y) + h(x), & \text{if} \ Rx \in y + K, \\ +\infty, & \text{if} \ Rx \notin y + K, \end{cases}
\]
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 $R : \mathcal{H} \rightarrow \mathcal{K}$ be continuous, convex with respect to $K$ such that $K \cap R(\text{dom} h) \neq \emptyset$. For $\mathcal{J}$ given in (83), the following hold:

1. The dual problem of $\mathcal{J}$ is given as
\[
\min_{v \in \mathcal{K}} f^*(v; K) + h^*(R^*v)
\]
where $f^*(\cdot; K) = \sup_{z \in \mathcal{K}} \langle z, \cdot \rangle - f(z)$.

2. The Lagrangian $\mathcal{L} : \mathcal{H} \times \mathcal{K} \rightarrow [-\infty, +\infty]$ is given as
\[
\mathcal{L}(x, v) = \begin{cases} -\infty, & \text{if} \ x \in \text{dom} h \text{ and } v \notin \text{dom} f^*(\cdot; K); \\ -f^*(v; K) + h(x) + \langle Rx, v \rangle, & \text{if} \ x \in \text{dom} h \text{ and } v \in \text{dom} f^*(\cdot; K); \\ +\infty, & \text{if} \ x \notin \text{dom} h. \end{cases}
\]

3. Suppose that the optimal values $\mu$ and $\mu^*$ of the primal problem and the dual problem satisfy the strong duality $\mu = -\mu^*$. Then, $(x^o, v^o) \in \mathcal{H} \times \mathcal{K}$ is a saddle point of $\mathcal{L}$ if and only if
\[
x^o \in \text{dom} h, \quad Rx^o \in K, \quad -v^o \in \text{dom} f^*(\cdot; K), \quad R^*v^o \in \partial h(x^o) \text{ and } -v^o \in \partial f(Rx^o).
\]

**Proof.** (1): For any $v \in \mathcal{K}$, it holds that
\[
\mathcal{J}^*(0, v) = \sup_{(x, y) \in \mathcal{H} \times \mathcal{K}} \langle y, v \rangle - \mathcal{J}(x, y),
\]
\[
= \sup_{(x, y) \in \mathcal{H} \times \mathcal{K}} \langle y, v \rangle - h(x) - f(Rx - y),
\]
\[
= \sup_{(x, z) \in \mathcal{H} \times \mathcal{K}} \langle x, R^*v \rangle - h(x) + \langle z, -v \rangle - f(z),
\]
\[
= h^*(R^*v) + f^*(-v; K).
\]
From the definition of the dual problem, we conclude.

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

\[
L(x, v) = h(x) + \inf_{y \in K} f(Rx - y) + \langle y, v \rangle,
\]

\[
= h(x) + \langle Rx, v \rangle + \inf_{z \in K} f(z) - \langle z, v \rangle,
\]

\[
= h(x) + \langle Rx, v \rangle - \sup_{z \in K} \langle z, v \rangle - f(z),
\]

\[
= h(x) + \langle Rx, v \rangle - f^*(v; K).
\]

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

(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 \( J \in \Gamma_0(\mathcal{H} \times \mathcal{K}) \). Applying Corollary 19.17 in (Bauschke et al., 2011), we obtain that \( (x^0, -v^0) \) is a saddle point of \( L \) if and only if \( x^0 \) is a solution of the primal problem (23) and \( v^0 \) 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. \( \square \)

Now we are ready to prove Proposition B.1.

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

Let \( \Lambda_n \) 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 \( K = \Sigma_{i_1 \times \cdots \times I_d} \), let us consider

\[
J : \mathbb{R}^{r \times N} \times \mathbb{R}^{I_1 \times \cdots \times I_d \times N} \to (-\infty, +\infty] : (A, Y) \mapsto \sum_{i=1}^{N} \left\{ Hx_i (D_{n-1} \times I_d + \Lambda[i,:]) - Y[i,:]) + \tau_n F_{\Lambda_{n-1}[i,:]}(\Lambda[i,:]) \right\}
\]

if \( D_{n-1} \times I_d + \Lambda \in Y + K \),

if \( D_{n-1} \times I_d + \Lambda \notin Y + K \).

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

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

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

for all \( i = 1, 2, \ldots, r \), and \( j = 1, 2, \cdots, N \). Hence

\[
D_{n-1} \times I_d + G_n / \tau_n = \xi + J^o \otimes c_o
\]

for some \( \xi \) satisfying (86) and \( c_o \in \mathbb{R}^{N \times 1} \). Now combining (87) and (86) yields (79). Finally, since we must have \( \Lambda_n \in \text{dom}(F) \), \( c_o \in \mathbb{R}^{N \times 1} \) should be such that \( \Lambda_n \) in (79) satisfies \( \Lambda_n \in \Sigma_{r} \).

\( \square \)

**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 \( \Lambda \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_{k-1} \times r \times I_{k+1} \times \cdots \times I_d \times N} \) be obtained from

\[
\Omega \cup \{(U_n^{(1)}, \ldots, U_{n}^{(k-1)}, U_{n-1}^{(k+1)}, \ldots, U_{n}^{(d)}, \Lambda_n^T) \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_{k-1} \times r \times I_{k+1} \times \cdots \times I_d \times N \times r}
\]

(88)
by inserting the last mode into the kth mode. Let \( U_{n}^{(k)} \in \Sigma^{r}_{I_{k}} \) be a solution of (14). Suppose each fiber of \( \Lambda_{n} \in \Sigma^{N}_{1} \) along the kth mode is not identically zero. Then \( U_{n}^{(k)} \) is uniquely determined by

\[
U_{n}^{(k)} = \left( U_{n-1}^{(k)} + \frac{G_{n}^{T} \otimes k \Lambda \otimes \tau_{n}}{\tau_{n}} - J^{1} \otimes c_{n}^{1} \right) + .
\]

(89)

Here, \( G_{n}^{T} \) is defined as the unique solution of the dual problem (22), \( c_{n}^{1} \in \mathbb{R}^{r \times 1} \) is chosen to satisfy \( U_{n}^{(k)} \in \Sigma^{r}_{I_{k}} \) and all entries of \( J^{1} \in \mathbb{R}^{I_{k} \times 1} \) are one.

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

\[
\begin{aligned}
\min_{U \in \Sigma^{N}_{I_{k}}} & \left( \sum_{i=1}^{N} H_{X_{i}} (\Lambda[i], i \times k, U^{T}) \right) + \tau_{n} F_{U_{n-1}^{(k)}} (U) \\
= & \min_{U \in \Sigma^{N}_{I_{k}}} \max_{R \in \mathbb{R}^{I_{k} \times I_{k}}} \left\{ \sum_{i=1}^{N} H_{X_{i}} (Q[i], i \times k, U^{T}) \right\} + \tau_{n} F_{U_{n-1}^{(k)}} (U), \\
= & \max_{G \in \mathbb{R}^{I_{k} \times I_{k}}} \min_{U \in \Sigma^{N}_{I_{k}}} \left\{ \sum_{i=1}^{N} \left( H_{X_{i}} (Q[i], i \times k, U^{T}, G[i], i) \right) \right\} + \tau_{n} F_{U_{n-1}^{(k)}} (U), \\
= & \max_{G \in \mathbb{R}^{I_{k} \times I_{k}}} \left\{ \sum_{i=1}^{N} \left( H_{X_{i}} (G[i], i \times k, U^{T}) \right) \right\} + \tau_{n} F_{U_{n-1}^{(k)}} (U), \\
= & \frac{1}{\tau_{n}} \left\{ \sum_{i=1}^{N} \left( H_{X_{i}} (G[i], i \times k, U^{T}) \right) \right\} + \tau_{n} F_{U_{n-1}^{(k)}} (U).
\end{aligned}
\]

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

\[
(\Lambda \times k, U^{T}) = [\Lambda[1], \ldots, \Lambda[N] \times k, U^{T}] = [\Lambda[1], \ldots, \Lambda[N] \times k, U^{T}],
\]

and (c) follows from the identity \( (\Lambda \times k, U^{T}, G) = (U, G \times \Lambda) \), 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 \( K = \Sigma^{N}_{I_{1} \times \ldots \times I_{d}} \) and

\[
J : \mathbb{R}^{I_{k} \times r} \times \mathbb{R}^{I_{k} \times I_{k} \times I_{d} \times N} \rightarrow (-\infty, +\infty) : \\
(U, Y) \mapsto \left\{ \begin{array}{ll}
\sum_{i=1}^{N} H_{X_{i}} (\Lambda[i], i \times k, U^{T} - Y[i], i) + \tau_{n} F_{U_{n-1}^{(k)}} (U) & \text{if } \Lambda \times k, U^{T} \in Y + K, \\
+\infty & \text{if } \Lambda \times k, U^{T} \notin Y + K.
\end{array} \right.
\]

\[
D. \text{ Auxiliary lemmas}
\]

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

\[
\sup_{\lambda \in \Sigma_{r}} \langle g, \lambda \rangle - \frac{1}{2} \| \lambda - \lambda_{0} \|_{2}^{2}
\]

(91)
As \( h \in (91) \). For any \( \lambda \), \( \lambda \)

This yields that

\[ \text{Proof.} \quad \text{The assertion follows from noting that} \]

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

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

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

\[ 0 \geq h'(0) = (g - \lambda^* + \lambda_0, \lambda - \lambda^*). \]

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

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

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

\[ (g - \lambda^* + \lambda_0)[i] = c. \]

This yields that

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

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

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

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

**Proof.** The assertion follows from noting that

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

**Lemma D.3 (Convex Surrogate for Functions with Lipschitz Gradient).** Let \( f : \mathbb{R}^p \to \mathbb{R} \) be differentiable and \( \nabla f \) be \( L \)-Lipschitz continuous. Then for each \( \theta, \theta' \in \mathbb{R}^p \),

\[ |f(\theta') - f(\theta) - \langle \nabla f(\theta), \theta' - \theta \rangle| \leq \frac{L}{2} \|\theta - \theta'\|^2. \]

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

\[ f(\theta') - f(\theta) = \int_{0}^{1} \langle \nabla f(\theta + s(\theta' - \theta)), \theta' - \theta \rangle \, ds. \]
By Cauchy-Schwarz inequality and $L$-Lipschitzness of $\nabla f$,

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

(102)

$$
\leq \int_0^1 L s \| \theta' - \theta \|^2 \, ds
$$

(103)

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

(104)

Then the assertion follows.

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

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

**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 $\Theta \subseteq \mathbb{R}^p$ and $\rho > 0$. The following conditions are equivalent.

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

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

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

**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 $\nabla f$ is $L$-Lipschitz for some $L > 0$. Then $f$ is $L$-weakly convex, that is, $\theta \mapsto f(\theta) + \frac{L}{2} \| \theta \|^2$ is convex.

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