ADMM for Efficient Deep Learning with Global Convergence

Junxiang Wang, Fuxun Yu, Xiang Chen and Liang Zhao
George Mason University
{jwang40, fyu2, xchen26, lzhao9}@gmu.edu

ABSTRACT

Alternating Direction Method of Multipliers (ADMM) has been used successfully in many conventional machine learning applications and is considered to be a useful alternative to Stochastic Gradient Descent (SGD) as a deep learning optimizer. However, as an emerging domain, several challenges remain, including
1. The lack of global convergence guarantees,
2. Slow convergence towards solutions, and
3. Cubic time complexity with regard to feature dimensions.

In this paper, we propose a novel optimization framework for deep learning via ADMM (dADMM) to address these challenges simultaneously. The parameters in each layer are updated backward and then forward so that the parameter information in each layer is exchanged efficiently. The time complexity is reduced from cubic to quadratic in (latent) feature dimensions via a dedicated algorithm design for subproblems that enhances them utilizing iterative quadratic approximations and backtracking. Finally, we provide the first proof of global convergence for an ADMM-based method (dADMM) in a deep neural network problem under mild conditions. Experiments on benchmark datasets demonstrated that our proposed dADMM algorithm outperforms most of the comparison methods.

CCS CONCEPTS

• Theory of computation → Nonconvex optimization; • Computing methodologies → Neural networks.

KEYWORDS

Deep Learning, Global Convergence, Alternating Direction Method of Multipliers

ACM Reference Format:

Junxiang Wang, Fuxun Yu, Xiang Chen and Liang Zhao. 2019. ADMM for Efficient Deep Learning with Global Convergence. In The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3292500.3330936

1 INTRODUCTION

Deep learning has been a hot topic in the machine learning community for the last decade. While conventional machine learning techniques have limited capacity to process natural data in their raw form, deep learning methods are composed of non-linear modules that can learn multiple levels of representation automatically [12]. Since deep learning methods are usually applied in large-scale datasets, such approaches require efficient optimizers to obtain a feasible solution within realistic time limits.

Stochastic Gradient Descent (SGD) and many of its variants are popular state-of-the-art methods for training deep learning models due to their efficiency. However, SGD suffers from many limitations that prevent its more widespread use: for example, the error signal diminishes as the gradient is backpropagated (i.e. the gradient vanishes); and SGD is sensitive to poor conditioning, which means a small input can change the gradient dramatically. Recently, the use of the Alternating Direction Method of Multipliers (ADMM) has been proposed as an alternative to SGD. ADMM splits a problem into many subproblems and coordinates them globally to obtain the solution. It has been demonstrated successfully for many machine learning applications [3]. The advantages of ADMM are numerous: it exhibits linear scaling as data is processed in parallel across cores; it does not require gradient steps and hence avoids gradient vanishing problems; it is also immune to poor conditioning [20].

Even though the performance of the ADMM seems promising, there are still several challenges at must be overcome: 1. The lack of global convergence guarantees. Despite the fact that many empirical experiments have shown that ADMM converges in deep learning applications, the underlying theory governing this convergence behavior remains mysterious. This is because a typical deep learning problem consists of a combination of linear and nonlinear mappings, causing optimization problems to be highly nonconvex. This means that traditional proof techniques cannot be directly applied. 2. Slow convergence towards solutions. Although ADMM is a powerful optimization framework that can be applied to large-scale deep learning applications, it usually converges slowly to high accuracy, even for simple examples [3]. It is often the case that ADMM becomes trapped in a modest solution and hence performs worse than SGD, as the experiment described later in this paper in Section 5 demonstrates. 3. Cubic time complexity with regard to feature dimensions. The implementation of the ADMM is very time-consuming for real-world datasets. Experiments conducted by Taylor et al. found that ADMM required more than 7000 cores to train a neural network with just 300 neurons [20]. This computational bottleneck mainly originates from the matrix inversion required to update the weight parameters. Computing an inverse matrix needs further subiterations, and its time complexity is approximately \(O(n^3)\), where \(n\) is a feature dimension [3].

In order to deal with these difficulties simultaneously, in this paper we propose a novel optimization framework for a deep learning Alternating Direction Method of Multipliers (dADMM) algorithm. Specifically, our new dADMM algorithm updates parameters first in a backward direction and then forwards. This update approach propagates parameter information across the whole network and accelerates the convergence process. It also avoids the operation of
The rest of this paper is organized as follows. In Section 2, we present the new dladMM algorithm, quadratic approximation, and the backtracking techniques utilized. In Section 4, we introduce the main convergence results for the dladMM algorithm. The results of extensive experiments conducted to show the convergence, efficiency, and effectiveness of our proposed new dladMM algorithm are presented in in Section 5, and Section 6 concludes this paper by summarizing the research.

2 RELATED WORK

Previous literature related to this research includes optimization for deep learning models and ADMM for nonconvex problems.

**Optimization for deep learning models:** The SGD algorithm and its variants play a dominant role in the research conducted by deep learning optimization community. The famous back-propagation algorithm was firstly introduced by Rumelhart et al. to train the neural network effectively [18]. Since the superior performance exhibited by AlexNet [11] in 2012, deep learning has attracted a great deal of researchers’ attention and many new optimizers based on SGD have been proposed to accelerate the convergence process, including the use of Polyak momentum [15], as well as research on the Nesterov momentum and initialization by Sutskever et al. [19]. Adam is the most popular method because it is computationally efficient and requires little tuning [10]. Other well-known methods that incorporate adaptive learning rates include AdaGrad [7], RMSProp [21], and AMSGrad [16]. Recently, the Alternating Direction Method of Multipliers (ADMM) has become popular with researchers due to its excellent scalability [20]. However, even though these optimizers perform well in real-world applications, their convergence mechanisms remain mysterious. This is because convergence assumptions are not applicable to deep learning problems, which often require non-differentiable activation functions such as the Rectifier linear unit (ReLU).

**ADMM for nonconvex problems:** The good performance achieved by ADMM over a range wide of convex problems has attracted the attention of many researchers, who have now begun to investigate the behavior of ADMM on nonconvex problems and made significant advances. For example, Wang et al. proposed an ADMM to solve multi-convex problems with a convergence guarantee [23], while Wang et al. presented convergence conditions for a coupled objective function that is nonconvex and nonsmooth [24]. Chen et al. discussed the use of ADMM to solve problems with quadratic coupling terms [4] and Wang et al. studied the convergence behavior of the ADMM for problems with nonlinear equality constraints [22]. Even though ADMM has been proposed to solve deep learning applications [8, 20], there remains a lack theoretical convergence analysis for the application of ADMM to such problems.

3 THE DLADM ALGORITHM

We present our proposed dladMM algorithm in this section. Section 3.1 formulates the deep neural network problem, Section 3.2 introduces how the dladMM algorithm works, and the quadratic approximation and backtracking techniques used to solve the subproblems are presented in Section 3.3.

3.1 Problem Formulation

Table 1 lists the important notation utilized in this paper. Even though there are many variants of formulations for deep neural networks, a typical neural network is defined by multiple linear mappings and nonlinear activation functions. A linear mapping for the $l$-th layer is composed of a weight matrix $W_l \in \mathbb{R}^{n_l \times n_{l-1}}$ and an intercept vector $b_l \in \mathbb{R}^{n_l}$, where $n_l$ is the number of neurons for the $l$-th layer; a nonlinear mapping for the $l$-th layer is defined by a
continuous activation function $f_\ell(\bullet)$. Given an input $a_{l-1} \in \mathbb{R}^{n_{l-1}}$ from the $(l-1)$-th layer, the $l$-th layer outputs $a_l = f_\ell(W_l a_{l-1} + b_l)$. Obviously, $a_{l-1}$ is nested in $a_l = f_\ell(\bullet)$. By introducing an auxiliary variable $z_l$, the task of training a deep neural network problem is formulated mathematically as follows:

**Problem 1.**

$$
\begin{align*}
\min_{W_l, b_l, z_l} & R(z_L; y) + \sum_{l=1}^{L} \Omega_l(W_l) \\
\text{s.t.} & \quad z_l = W_l a_{l-1} + b_l (l = 1, \ldots, L), \quad a_l = f_\ell(z_l) (l = 1, \ldots, L-1)
\end{align*}
$$

In Problem 1, $a_0 = x \in \mathbb{R}^{n_0}$ is the input of the deep neural network where $n_0$ is the number of feature dimensions, and $y$ is a predefined label vector. $R(z_L; y)$ is a risk function for the $L$-th layer, which is convex, continuous, and proper, and $\Omega_l(W_l)$ is a regularization term for the $l$-th layer, which is also convex, continuous, and proper. Rather than solving Problem 1 directly, we can relax Problem 1 by adding an $\ell_2$ penalty to address Problem 2 as follows:

**Problem 2.**

$$
\begin{align*}
\min_{W_l, b_l, z_l, a} & f(W, b, z, a) = R(z_L; y) + \sum_{l=1}^{L} \Omega_l(W_l) + \nu \sum_{l=1}^{L} (\| z_l - W_l a_{l-1} - b_l \|^2 + \| a_l - f_\ell(z_l) \|^2)
\end{align*}
$$

where $W = \{W_l\}_{l=1}^{L}$, $b = \{b_l\}_{l=1}^{L}$, $z = \{z_l\}_{l=1}^{L}$, $a = \{a_l\}_{l=1}^{L-1}$ and $\nu > 0$ is a tuning parameter. Compared with Problem 1, Problem 2 has only a linear constraint $z_L = W_L a_{L-1} + b_L$ and hence is easier to solve. It is straightforward to show that as $\nu \to \infty$, the solution of Problem 2 approaches that of Problem 1.

### 3.2 The dlADMM Algorithm

We introduce the dlADMM algorithm to solve Problem 2 in this section. The traditional ADMM strategy for optimizing parameters is to start from the first layer and then update parameters in the following layer sequentially [20]. In this case, the parameters in the final layer are subject to the parameter update in the first layer. However, the parameters in the final layer contain important information that can be transmitted towards the previous layers to speed up convergence. To achieve this, we propose our novel dlADMM framework, as shown in Figure 1. Specifically, the dlADMM algorithm updates parameters in two steps. In the first, the dlADMM begins updating from the $L$-th (final) layer and moves backward toward the first layer. The update order of parameters in the same layer is $a_L \rightarrow z_L \rightarrow b_L \rightarrow W_L$. In the second, the dlADMM reverses the update direction, beginning at the first layer and moving forward toward the $L$-th (final) layer. The update order of the parameters in the same layer is $W_L \rightarrow b_L \rightarrow z_L \rightarrow a_L$. The parameter information for all layers can be exchanged completely by adopting this update approach.

Now we can present our dlADMM algorithm mathematically. The augmented Lagrangian function of Problem 2 is shown in the following form:

$$
L_p(W, b, z, a, u) = R(z_L; y) + \sum_{l=1}^{L} \Omega_l(W_l) + \phi(W, b, z, a, u) \tag{1}
$$

where $\phi(W, b, z, a, u) = (\nu/2) \sum_{l=1}^{L} (\| z_l - W_L a_{l-1} - b_l \|^2 + \| a_l - f_\ell(z_l) \|^2) + u^T (z_L - W_L a_{L-1} - b_L) + \| \rho / 2 \| z_L - W_L a_{L-1} - b_L \|^2$, $u$ is a dual variable and $\rho > 0$ is a hyperparameter of the dlADMM algorithm. We denote $W^L_l, b^L_l, z^{k+1}_l$ and $\tilde{a}^{k+1}_l$ as the backward update of the dlADMM for the $l$-th layer in the $(k+1)$-th iteration, while $W^{k+1}_L, b^{k+1}_L, z^{k+1}_L$ and $\tilde{a}^{k+1}_L$ are denoted as the forward update of the dlADMM for the $L$-th layer in the $(k+1)$-th iteration. Moreover, we denote $W_l^{k+1} = \{(W_l^{k+1})_{i=1}^L\}_{l=1}^{L}$, $b_l^{k+1} = \{(b_l^{k+1})_{i=1}^L\}_{l=1}^{L}$, $z_l^{k+1} = \{(z_l^{k+1})_{i=1}^{L-1}\}_{l=1}^{L}$, $a_l^{k+1} = \{(a_l^{k+1})_{i=1}^{L-1}\}_{l=1}^{L}$, $\tilde{a}_l^{k+1} = \{(\tilde{a}_l^{k+1})_{i=1}^{L-1}\}_{l=1}^{L}$ and $\tilde{a}_L^{k+1} = \{(\tilde{a}_L^{k+1})_{i=1}^{L-1}\}_{l=1}^{L}$. Then the dlADMM algorithm is shown in Algorithm 1. Specifically, Lines 5, 6, 10, 11, 14, 15, 17 and 18 solve eight subproblems, namely, $\tilde{a}_l^{k+1}$, $z_l^{k+1}$, $b_l^{k+1}$, $W_l^{k+1}$, $a_l^{k+1}$, $\tilde{a}_L^{k+1}$ and $\tilde{a}_l^{k+1}$, respectively. Lines 21 and 22 update the residual $r^{k+1}$ and the dual variable $u^{k+1}$, respectively.

### 3.3 The Quadratic Approximation and Backtracking

The eight subproblems in Algorithm 1 are discussed in detail in this section. Most can be solved by quadratic approximation and the backtracking techniques described above, so the operation of the matrix inversion can be avoided.

#### 1. Update $\tilde{a}_l^{k+1}$

The variables $\tilde{a}_l^{k+1} (l = 1, \cdots, L-1)$ are updated as follows:

$$
\tilde{a}_l^{k+1} = \arg \min_{a_l} L_p(W_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, \{z_{l+1}^{k+1}\}_{l+1}^{L}, u^{k+1})
$$

The subproblem is transformed into the following form by replacing Equation (1).

$$
\tilde{a}_l^{k+1} = \arg \min_{a_l} \phi(W_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, \{z_{l+1}^{k+1}\}_{l+1}^{L}, u^{k+1}) \tag{2}
$$

Because $a_l$ and $W_{l+1}$ are coupled in $\phi(\bullet)$, in order to solve this problem, we must compute the inverse matrix of $W_l^{k+1}$, which involves subiterations and is computationally expensive [20]. In order to handle this challenge, we define $\tilde{Q}_l(a_l; \tilde{a}_l^{k+1})$ as a quadratic
Algorithm 2 The Backtracking Algorithm to update $\tilde{x}_{k+1}^l$

Require: $W_{l+1}^k, \tilde{x}_{k+1}^l, z_{k+1}^l, u_k^l, \rho$, some constant $\eta > 1$.
Ensure: $\tilde{x}_{k+1}^l, z_{k+1}^l, \rho$

1: Pick up $T$ and $\beta = a_k^l - \nabla_{a_k^l} \phi(T) / T$
2: while $\phi(W_{l+1}^k, b_{l+1}^k, z_{l+1}^k, \tilde{x}_{k+1}^l, u_k^l) > \bar{G}(T, T)$ do
3: $T = T\eta$
4: $\beta = a_k^l - \nabla_{a_k^l} \phi(T)$
5: end while
6: Output $\tilde{x}_{k+1}^l = \bar{T}$
7: Output $z_{k+1}^l = T$
8: given a suitable $T_{k+1}^l$.

Algorithm 1 the dLADMM Algorithm to Solve Problem 2

Require: $y, a_l = x, \rho, \nu$
Ensure: $y_l = B_i l = 1, \ldots, L$, $W_l(i = 1, \ldots, L), b_l(i = 1, \ldots, L), z_l(i = 1, \ldots, L)$

1: Initialize $k = 0$
2: while $W_{k+1}^l, b_{k+1}^l, u_{k+1}^l$ not converged do
3: for $l = 1$ to $L$ do
4: $\tilde{x}_{k+1}^l = \arg \min_{\tilde{x}_{k+1}^l} L_p(W_{l+1}^k, b_{l+1}^k, z_{l+1}^k, \tilde{x}_{k+1}^l)^L_{j=l+1} z_{l+1}^k, u_{k+1}^l$
5: end for
6: $L_{k+1}^l = \text{L} \sigma_k^l$
7: end while

approximation of $\phi$ at $a_k^l$, which is mathematically reformulated as follows:

$$\tilde{G}_l(a_i; \bar{k}_{k+1}^l) = \phi(W_{l+1}^k, b_{l+1}^k, z_k^l, \bar{x}_{k+1}^l, u_k^l) + (\nabla_{a_k^l} \phi)^T (W_{l+1}^k, b_{l+1}^k, z_k^l, \bar{x}_{k+1}^l, u_k^l) (a_i - a_k^l)$$

where $\bar{k}_{k+1}^l = 0$ is a parameter vector, $\sigma$ denotes the Hadamard product (the elementary product), and $a^{\sigma b}$ denotes $a$ to the Hadamard power of $b$ and $\| \cdot \|$ is the $l_1$ norm. $\nabla_{a_k^l} \phi$ is the gradient of $\bar{a}_i$ at $a_k^l$. Obviously, $\tilde{G}_l(a_i; \bar{k}_{k+1}^l) = \phi(W_{l+1}^k, b_{l+1}^k, z_k^l, \bar{x}_{k+1}^l, u_k^l)$.

Because $\tilde{G}_l(a_i; \bar{k}_{k+1}^l)$ is a quadratic function with respect to $a_i$, the solution can be obtained by

$$\bar{a}_{k+1}^l = \nabla_{a_k^l} \phi(T_{k+1}^l)$$

Algorithm 2 shows the backtracking algorithm utilized to find a suitable $T_{k+1}^l$. Lines 2-4 implement a while loop until the condition $\phi(W_{l+1}^k, b_{l+1}^k, z_k^l, \bar{x}_{k+1}^l, u_k^l) \leq \bar{G}(T_{k+1}^l, T_{k+1}^l)$ is satisfied. As $T_{k+1}^l$ becomes larger and larger, $\tilde{x}_{k+1}^l$ is close to $\bar{a}_{k+1}^l$ and $a_{k+1}^l$ satisfies the loop condition, which precludes the possibility of the infinite loop. The time complexity of Algorithm 2 is $O(n^2)$, where $n$ is the number of features or neurons.

2. Update $z_{k+1}^l$

The variables $z_{k+1}^l$ $(l = 1, \ldots, L)$ are updated as follows:

$$z_{k+1}^l \leftarrow \arg \min_{z_{k+1}^l} L_p(W_{l+1}^k, b_{l+1}^k, z_{l+1}^k, \tilde{x}_{k+1}^l)^L_{j=l+1} z_{l+1}^k, u_{k+1}^l$$

which is equivalent to the following forms: for $z_{k+1}^l (l = 1, \ldots, L - 1)$,

$$z_{k+1}^l \leftarrow \arg \min_{z_{k+1}^l} \phi(W_{l+1}^k, b_{l+1}^k, z_{l+1}^k)_{j=l+1} z_{l+1}^k, \tilde{x}_{k+1}^l, u_{k+1}^l$$

and for $z_L^*$,

$$z_{k+1}^l \leftarrow \arg \min_{z_{k+1}^l} \phi(W_{l+1}^k, b_{l+1}^k, z_{l+1}^k)_{j=l+1} z_{l+1}^k, a_k^l, u_k^l) + R(z_{l}^k)$$

Equation (4) is highly nonconvex because the nonlinear activation function $f(z_l)$ is contained in $\phi(\bullet)$. For common activation functions such as the Rectified linear unit (ReLU) and leaky ReLU, Equation (4) has a closed-form solution; for other activation functions like sigmoid and hyperbolic tangent (tanh), a look-up table is recommended [20].

Equation (5) is a convex problem because $\phi(\bullet)$ and $R(\bullet)$ are convex with regard to $z_l$. Therefore, Equation (5) can be solved by Fast Iterative Soft-Thresholding Algorithm (FISTA) [1].

3. Update $b_{l+1}^k$

The variables $b_{l+1}^k$ $(l = 1, \ldots, L)$ are updated as follows:

$$b_{l+1}^k \leftarrow \arg \min_{b_{l+1}^k} L_p(W_{l+1}^k, b_{l+1}^k, z_{l+1}^k, \bar{a}_{k+1}^l)_{j=l+1} b_{l+1}^k, \tilde{x}_{k+1}^l, u_{k+1}^l$$

which is equivalent to the following form:

$$b_{l+1}^k \leftarrow \arg \min_{b_{l+1}^k} \phi(W_{l+1}^k, b_{l+1}^k, z_{l+1}^k, \bar{a}_{k+1}^l)_{j=l+1} b_{l+1}^k, \tilde{x}_{k+1}^l, u_{k+1}^l$$

Similarly to the update of $z_{k+1}^l$, we define $\bar{U}_l(b_{l}, \bar{B})$ as a quadratic approximation of $\phi(\bullet)$ at $a_{k+1}^l$, which is formulated mathematically as follows [1]:

$$\bar{U}_l(b_{l}, \bar{B}) = \phi(W_{l+1}^k, b_{l+1}^k, z_{l+1}^k, \bar{x}_{k+1}^l, u_{k+1}^l) + (\nabla_{b_k^l} \phi)^T (W_{l+1}^k, b_{l+1}^k, z_{l+1}^k, \bar{x}_{k+1}^l, u_{k+1}^l) (b_{l} - b_{k+1}^l)$$

where $\bar{B} > 0$ is a parameter. Here $\bar{B} \geq v$ for $l = 1, \ldots, L - 1$ and $\bar{B} \geq \rho$ for $l = L$ are required for the convergence analysis [1]. Without loss of generality, we set $\bar{B} = v$, and solve the subsequent subproblem as follows:

$$b_{l+1}^k \leftarrow \arg \min_{b_{l+1}^k} \bar{U}_l(b_{l}^k; v)(l = 1, \ldots, L - 1)$$

$$b_{L+1}^k \leftarrow \arg \min_{b_{L+1}^k} \bar{U}_L(b_{L}; v)$$
Equation (6) is a convex problem and has a closed-form solution as follows:
\[
\begin{align*}
\zeta_{l+1}^{k+1} & \leftarrow b_l^k - \nabla b_l^k \phi / \nu, \quad (l = 1, \cdots, L - 1) \\
\zeta_{L}^{k+1} & \leftarrow b_L^k - \nabla b_L^k \phi / \rho.
\end{align*}
\]

4. Update \( W_l^{k+1} \)

The variables \( W_l^{k+1} \) \((l = 1, \cdots, L)\) are updated as follows:
\[
\begin{align*}
\bar{W}_l^{k+1} & \leftarrow \arg\min_{W_l} L_p(W_l^{k+1} = \bar{W}_l^{k+1} + b_l^{k+1}, z_l^{k+1}, \bar{a}_l^{k+1}, u_l^k)
\end{align*}
\]
which is equivalent to the following form:
\[
\begin{align*}
\bar{W}_l^{k+1} & \leftarrow \arg\min_{W_l} \phi(W_l^{k+1}, \bar{b}_l^{k+1}, \bar{z}_l^{k+1}, \bar{a}_l^{k+1}, u_l^k) + \Omega(W_l)
\end{align*}
\]
(8)

Due to the same challenge in updating \( b_l^{k+1} \), we define \( \bar{P}_l(W_l^k; \bar{b}_l^{k+1}) \) as a quadratic approximation of \( \phi \) at \( W_l^k \). The quadratic approximation is mathematically reformulated as follows [1]:
\[
\bar{P}_l(W_l^k; \bar{b}_l^{k+1}) = \phi(\bar{W}_l^{k+1}, \bar{b}_l^{k+1}, \bar{z}_l^{k+1}, \bar{a}_l^{k+1}, u_l^k) + \nabla \left( \nabla \bar{P}_l(W_l^k; \bar{b}_l^{k+1}) \right) (W_l^{k+1} - W_l^k)
\]
and
\[
\bar{P}_l(W_l^k; \bar{b}_l^{k+1}) > 0 \quad \text{is a parameter vector, which is chosen by the Algorithm 3. Instead of minimizing the Equation (8), we minimize the following:}
\]
\[
\bar{W}_l^{k+1} \leftarrow \arg\min_{W_l} \bar{P}_l(W_l^k; \bar{b}_l^{k+1}) + \Omega(W_l)
\]
(9)

Equation (9) is convex and hence can be solved exactly. If \( \Omega_l \) is either an \( \ell_1 \) or an \( \ell_2 \) regularization term, Equation (9) has a closed-form solution.

**Algorithm 3 The Backtracking Algorithm to update \( W_l^{k+1} \)**

**Require:** \( \bar{W}_l^{k+1}, \bar{b}_l^{k+1}, \bar{z}_l^{k+1}, \bar{a}_l^{k+1}, u_l^k, \rho, \) some constant \( \tau > 1.\)

**Ensure:** \( W_l^{k+1}, b_l^k, z_l^k, a_l^k \)

1. Pick up \( \bar{W}_l^k \) and \( \bar{W}_l^k \) = \( \phi(\bar{W}_l^k, \bar{b}_l^k, \bar{z}_l^k, \bar{a}_l^k, u_l^k) \).
2. while \( \phi(\bar{W}_l^{k+1}, \bar{z}_l^{k+1}, \bar{a}_l^{k+1}, u_l^k) > \bar{W}_l^k(z_l^k, \bar{a}_l^k, u_l^k) \)
3. \( \tau = \frac{\tau}{\tau} \)
4. Solve \( \bar{Z} \) by Equation (9).
5. end while
6. Output \( \bar{W}_l^{k+1} \) = \( \tau \)
7. Output \( W_l^{k+1} = \tau \)

5. Update \( W_l^{k+1} \)

The variables \( W_l^{k+1} \) \((l = 1, \cdots, L)\) are updated as follows:
\[
\begin{align*}
W_l^{k+1} & \leftarrow \arg\min_{W_l} L_p(W_l^k, b_l^k, z_l^k, a_l^k, u_l^k) + \Omega(W_l)
\end{align*}
\]
which is equivalent to
\[
\begin{align*}
W_l^{k+1} & \leftarrow \arg\min_{W_l} \phi(W_l^k, b_l^k, z_l^k, a_l^k, u_l^k) + \Omega(W_l)
\end{align*}
\]
(10)

Similarly, we define \( P_l(W_l^k; \bar{b}_l^{k+1}) \) as a quadratic approximation of \( \phi \) at \( W_l^{k+1} \). The quadratic approximation is then mathematically reformulated as follows [1]:
\[
\begin{align*}
P_l(W_l^k; \bar{b}_l^{k+1}) & = \phi(W_l^{k+1}, b_l^{k+1}, z_l^k, a_l^k, u_l^k) + \nabla \left( \nabla \phi(W_l^{k+1}, b_l^{k+1}, z_l^k, a_l^k, u_l^k) \right) (W_l^{k+1} - W_l^k)
\end{align*}
\]
\[
+ \Omega(W_l)
\]
(11)

where \( \bar{b}_l^{k+1} > 0 \) is a parameter vector. Instead of minimizing the Equation (10), we minimize the following:
\[
W_l^{k+1} \leftarrow \arg\min_{W_l} P_l(W_l^k; \bar{b}_l^{k+1}) + \Omega_l(W_l)
\]
(12)

The choice of \( \bar{b}_l^{k+1} \) is discussed in the supplementary materials.

6. Update \( b_l^{k+1} \)

The variables \( b_l^{k+1} \) \((l = 1, \cdots, L)\) are updated as follows:
\[
\begin{align*}
b_l^{k+1} & \leftarrow \arg\min_{b_l^k} L_p(b_l^{k+1}, b_l^k, z_l^{k+1}, a_l^k, u_l^k)
\end{align*}
\]
which is equivalent to the following formulation:
\[
\begin{align*}
b_l^{k+1} & \leftarrow \arg\min_{b_l^k} \phi(b_l^{k+1}, b_l^k, z_l^{k+1}, a_l^k, u_l^k) + \nabla \left( \nabla \phi(b_l^{k+1}, b_l^k, z_l^{k+1}, a_l^k, u_l^k) \right) (b_l^{k+1} - b_l^k)
\end{align*}
\]
(13)

Where \( b_l^{k+1} > 0 \) is a parameter. We set \( B = \nu \) for \( l = 1, \cdots, L - 1 \) and \( B = \rho \) for \( l = L \), and solve the resulting subproblems as follows:
\[
\begin{align*}
b_l^{k+1} & \leftarrow \arg\min_{b_l^k} U_l(b_l^k; \nu) \quad (l = 1, \cdots, L - 1)
\end{align*}
\]
(14)

7. Update \( z_l^{k+1} \)

The variables \( z_l^{k+1} \) \((l = 1, \cdots, L)\) are updated as follows:
\[
\begin{align*}
z_l^{k+1} & \leftarrow \arg\min_{z_l^k} L_p(z_l^{k+1}, b_l^k, z_l^k, a_l^k, u_l^k)
\end{align*}
\]
which is equivalent to the following forms for \( z_l \) \((l = 1, \cdots, L - 1)\):
\[
\begin{align*}
z_l^{k+1} & \leftarrow \arg\min_{z_l^k} \phi(z_l^{k+1}, b_l^k, z_l^k, a_l^k, u_l^k) + \nabla \left( \nabla \phi(z_l^{k+1}, b_l^k, z_l^k, a_l^k, u_l^k) \right) (z_l^{k+1} - z_l^k)
\end{align*}
\]
(15)

Solving Equations (14) and (15) proceeds exactly the same as solving Equations (4) and (5), respectively.

8. Update \( a_l^{k+1} \)

The variables \( a_l^{k+1} \) \((l = 1, \cdots, L - 1)\) are updated as follows:
\[
\begin{align*}
a_l^{k+1} & \leftarrow \arg\min_{a_l^k} L_p(z_l^{k+1}, a_l^{k+1})
\end{align*}
\]
(16)

The supplementary materials are available at [http://mason.gmu.edu/~zhao9/materials/papers/diADMM_supp.pdf](http://mason.gmu.edu/~zhao9/materials/papers/diADMM_supp.pdf)
which is equivalent to the following form:

\[ a_i^{k+1} \leftarrow \arg \min_{a_i} Q_i(a_i; z_i^{k+1}) \]  

\[ Q_i(a_i; z_i^{k+1}) \text{ is defined as the quadratic approximation of } \phi \text{ at } a_i^{k+1} \]  

as follows:

\[ Q_i(a_i; z_i^{k+1}) = \phi(W_i^{k+1}, b_i^{k+1}, z_i^{k+1}, a_i^{k+1}, u_i^{k+1}) \]  

\[ + \langle \nabla \phi(a_i^{k+1}, z_i), (W_i^{k+1}, b_i^{k+1}, z_i^{k+1}, a_i^{k+1}, u_i^{k+1}) \rangle \]  

\[ + \| z_i^{k+1} - a_i^{k+1} \|^2 \]  

and we can solve the following problem instead:

\[ a_i^{k+1} \leftarrow \arg \min_{a_i} Q_i(a_i; z_i^{k+1}) \]  

(16)

where \( z_i^{k+1} > 0 \) is a parameter vector. The solution to Equation (16) can be obtained by

\[ a_i^{k+1} \leftarrow a_i^{k+1} - \nabla \phi(a_i^{k+1}, z_i^{k+1}) \]  

To choose of an appropriate \( z_i^{k+1} \) is shown in the supplementary materials [1].

4 CONVERGENCE ANALYSIS

In this section, the theoretical convergence of the proposed dlADMM algorithm is analyzed. Before we formally present the convergence results of the dlADMM algorithms, Section 4.1 presents necessary assumptions to guarantee the global convergence of dlADMM. In Section 4.2, we prove the global convergence of the dlADMM algorithm.

4.1 Assumptions

**Assumption 1 (Closed-form Solution).** There exist activation functions \( a_i = f_i(z_i) \) such that Equations (4) and (14) have closed form solutions \( z_i^{k+1} = h(W_i^{k+1}, b_i^{k+1}, a_i^{k+1}), \) respectively, where \( h(\bullet) \) and \( h(\bullet) \) are continuous functions.

This assumption can be satisfied by commonly used activation functions such as ReLU and leaky ReLU. For example, for the ReLU activation function \( a_i = \max(z_i, 0), \) Equation (14) has the following solution:

\[ z_i^{k+1} = \begin{cases} \min(W_i^{k+1} a_i^{k+1} + b_i^{k+1}, 0) & 0 \\ \max((W_i^{k+1} a_i^{k+1} + b_i^{k+1})/2, 0) & 0 \end{cases} \]  

**Assumption 2 (Objective Function).** \( F(W, b, z, a) \) is coercive over the nonempty set \( G = \{(W, b, z, a) : z_i - W_i a_i - b_i = 0\}. \) In other words, \( F(W, b, z, a) \to \infty \) if \( F(W, b, z, a) \) \to \infty. Moreover, \( R(z; y) \) is Lipschitz differentiable with Lipschitz constant \( H \geq 0. \)

The Assumption 2 is mild enough for most common loss functions to satisfy. For example, the cross-entropy and square loss are Lipschitz differentiable.

4.2 Key Properties

We present the main convergence result of the proposed dlADMM algorithm in this section. Specifically, as long as Assumptions 1-2 hold, then Properties 1-3 are satisfied, which are important to prove the global convergence of the proposed dlADMM algorithm. The proof details are included in the supplementary materials [1].

**Property 1 (Boundedness).** If \( \rho > 2H, \) then \( \{W_i, b_i, z_i, a_i, u_i\} \) is bounded, and \( L_p(W_i, b_i, z_i, a_i, u_i) \) is lower bounded.

Property 1 concludes that all variables and the value of \( L_p \) have lower bounds. It is proven under Assumptions 1 and 2, and its proof can be found in the supplementary materials [1].

**Property 2 (Sufficient Descent).** If \( \rho > 2H \) so that \( C_1 = \rho/2 - H/2 - H^2/\rho > 0, \) then there exists

\[ C_2 = \min(\rho/2, C_1, \tau_i, \tau_i, \tau_i, \tau_i, \tau_i, \tau_i, \tau_i, \tau_i, \tau_i) \]  

such that

\[ L_p(W_i, b_i, z_i, a_i, u_i) \leq C_2 \left( \sum_{i=1}^{L} \| W_i^{k+1} - W_i^k \|^2 + \| W_i^{k+1} - W_i^k \|^2 \right) + \| b_i^{k+1} - b_i^k \|^2 + \| a_i^{k+1} - a_i^k \|^2 \]  

\[ + \| z_i^{k+1} - z_i^k \|^2 + \| W_i^{k+1} - W_i^k \|^2 \]  

(17)

Property 2 depicts the monotonic decrease of the objective value during iterations. The proof of Property 2 is detailed in the supplementary materials [1].

**Property 3 (Subgradient Bound).** There exist a constant \( C > 0 \) and \( g \in \delta L_p(W_i^{k+1}, b_i^{k+1}, z_i^{k+1}, a_i^{k+1}) \) such that

\[ \| g \| \leq C(\| W_i^{k+1} - W_i^k \| + \| b_i^{k+1} - b_i^k \|) \]  

\[ + \| z_i^{k+1} - z_i^k \| + \| a_i^{k+1} - a_i^k \| + \| z_i^{k+1} - z_i^k \|) \]  

(18)

Property 3 ensures that the subgradient of the objective function is bounded by variables. The proof of Property 3 requires Property 1 and the proof is elaborated in the supplementary materials [1].

**Theorem 4.1.** For any \( \rho > 2H, \) if Assumptions 1 and 2 are satisfied, then Properties 1-3 hold.

**Proof.** This theorem can be concluded by the proofs in the supplementary materials [1].

The next theorem presents the global convergence of the dlADMM algorithm.

**Theorem 4.2 (Global Convergence).** If \( \rho > 2H, \) then for the variables \( (W, b, z, a, W) \) in Problem 2, starting from any \( (W_0, b_0, z_0, a_0, u_0) \), it has at least a limit point \( (W^*, b^*, z^*, a^*, u^*) \), and any limit point \( (W^*, b^*, z^*, a^*, u^*) \) is a critical point of Problem 2. That is, \( 0 \in \delta L_p(W^*, b^*, z^*, a^*, u^*) \) or equivalently,

\[ z_i^* = W_i^* a_i^* + b_i^* \]  

\[ 0 \in \partial \rho L_p(W^*, b^*, z^*, a^*, u^*) \]  

\[ 0 \in \partial \rho L_p(W^*, b^*, z^*, a^*, u^*) \]  

(19)

**Proof.** Because \( (W^k, b^k, z^k, a^k, u^k) \) is bounded, there exists a subsequence \( (W^k, b^k, z^k, a^k, u^k) \) such that \( (W^k, b^k, z^k, a^k, u^k) \to (W^*, b^*, z^*, a^*, u^*) \) where \( (W^*, b^*, z^*, a^*, u^*) \) is a limit point. By Properties 1 and 2, \( L_p(W^k, b^k, z^k, a^k, u^k) \) is non-increasing and lower bounded and hence converges. By Property 2, we prove that \( L_p(W^{k+1} - W^k) \to 0, \) so that \( z^{k+1} - z^k \to 0, \)
\[\|W^{k+1} - \tilde{W}^{k+1}\| \to 0, \|b^{k+1} - \tilde{b}^{k+1}\| \to 0,\] and \[\|a^{k+1} - \tilde{a}^{k+1}\| \to 0,\] as \(k \to \infty\). Therefore \(\|W^{k+1} - W^k\| \to 0, \|b^{k+1} - b^k\| \to 0,\) and \(\|a^{k+1} - a^k\| \to 0,\) as \(k \to \infty\). Moreover, from Assumption 1, we know that \(z^{k+1} \to z^k\) and \(z^{k+1} \to z^k\) as \(k \to \infty\). Therefore, \(z^{k+1} \to z^k\).

We infer there exists \(g^k \in \partial L_p(W^k, b^k, z^k, a^k, u^k)\) such that \(\|g^k\| \to 0\) as \(k \to \infty\) based on Property 3. Specifically, \(\|g^k\| \to 0\) as \(s \to \infty\). According to the definition of general subgradient (Definition 8.3 in [17]), we have \(0 \in \partial L_p(W^k, b^k, z^k, a^k, u^k)\). In other words, the limit point \((W^*, b^*, z^*, a^*, u^*)\) is a critical point of \(L_p\) defined in Equation (1).

Theorem 4.2 shows that our dLADMM algorithm converges globally for sufficiently large \(\rho\), which is consistent with previous literature [9, 24]. The next theorem shows that the dLADMM converges globally with a sublinear convergence rate \(o(1/k)\).

**Theorem 4.3 (Convergence Rate).** For a sequence \((W^k, b^k, z^k, a^k, u^k)\), define \(c_k = \min_{0 \leq i \leq L} \left(\sum_{l=1}^L \|\tilde{W}_l^{i+1} - W_l^i\|^2 + \|\tilde{b}_l^{i+1} - b_l^i\|^2 + \|\tilde{a}_l^{i+1} - a_l^i\|^2 + \|\tilde{z}_l^{i+1} - z_l^i\|^2\right)\), then the convergence rate of \(c_k\) is \(o(1/k)\).

**Proof.** The proof of this theorem is included in the supplementary materials.

# 5 EXPERIMENTS

In this section, we evaluate dLADMM algorithm using benchmark datasets. Effectiveness, efficiency and convergence properties of dLADMM are compared with state-of-the-art methods. All experiments were conducted on 64-bit Ubuntu16.04 LTS with Intel(R) Xeon processor and GTX1080Ti GPU.

## 5.1 Experiment Setup

**5.1.1 Dataset.** In this experiment, two benchmark datasets were used for performance evaluation: MNIST [13] and Fashion MNIST [25]. The MNIST dataset has ten classes of handwritten-digit images, which was firstly introduced by Lecun et al. in 1998 [13]. It contains 55,000 training samples and 10,000 test samples with 784 features each, which is provided by the Keras library [5]. Unlike the MNIST dataset, the Fashion MNIST dataset has ten classes of assortment images on the website of Zalando, which is Europe’s largest online fashion platform [25]. The Fashion-MNIST dataset consists of 60,000 training samples and 10,000 test samples with 784 features each.

**5.1.2 Experiment Settings.** We set up a network architecture which contained two hidden layers with 1,000 hidden units each. The Rectified linear unit (ReLU) was used for the activation function for both network structures. The loss function was set as the deterministic cross-entropy loss. \(\nu\) was set to \(10^{-6}\). \(\rho\) was initialized as \(10^{-6}\) and was multiplied by 10 every 100 iterations. The number of iteration was set to 200. In the experiment, one iteration means one epoch.

**5.1.3 Comparison Methods.** Since this paper focuses on fully-connected deep neural networks, SGD and its variants and ADMM are state-of-the-art methods and hence were served as comparison methods. For SGD-based methods, the full batch dataset is used for training models. All parameters were chosen by the accuracy of the training dataset. The baselines are described as follows:

1. **Stochastic Gradient Descent (SGD)** [2]. The SGD and its variants are the most popular deep learning optimizers, whose convergence has been studied extensively in the literature. The learning rate of SGD was set to \(10^{-6}\) for both the MNIST and Fashion MNIST datasets.

2. **Adaptive gradient algorithm (AdaGrad)** [7]. AdaGrad is an improved version of SGD: rather than fixing the learning rate during iteration, it adapts the learning rate to the hyperparameter. The learning rate of AdaGrad was set to \(10^{-3}\) for both the MNIST and Fashion MNIST datasets.

3. **Adaptive learning rate method (Adadelta)** [26]. As an improved version of the AdaGrad, the Adadelta is proposed to overcome the sensitivity to hyperparameter selection. The learning rate of Adadelta was set to 0.1 for both the MNIST and Fashion MNIST datasets.

4. **Adaptive momentum estimation (Adam)** [10]. Adam is the most popular optimization method for deep learning models. It estimates the first and second momentum in order to correct the biased gradient and thus makes convergence fast. The learning rate of Adam was set to \(10^{-3}\) for both the MNIST and Fashion MNIST datasets.

5. **Alternating Direction Method of Multipliers (ADMM)** [20]. ADMM is a powerful convex optimization method because it can split an objective function into a series of subproblems, which are coordinated to get global solutions. It is scalable to large-scale datasets and supports parallel computations. The \(\rho\) of ADMM was set to 1 for both the MNIST and Fashion MNIST datasets.

## 5.2 Experimental Results

In this section, experimental results of the dLADMM algorithm are analyzed against comparison methods.

![Figure 2: Convergence objective value](image)

(a) Objective value

![Figure 2: Convergence residual](image)

(b) Residual

Figure 2: Convergence curves of dLADMM algorithm for MNIST and Fashion MNIST datasets when \(p = 1\): dLADMM algorithm converged.

### 5.2.1 Convergence.

First, we show that our proposed dLADMM algorithm converges when \(\rho\) is sufficiently large and diverges when \(\rho\) is small for both the MNIST dataset and the Fashion MNIST dataset.

The convergence and divergence of dLADMM algorithm are shown in Figures 2 and 3 when \(p = 1\) and \(\rho = 10^{-6}\) respectively. In Figures 2(a) and 3(a), the X axis and Y axis denote the number of iterations and the logarithm of objective value, respectively. In Figures, 2(b) and 3(b), the X axis and Y axis denote the number of iterations and the logarithm of residual, respectively.
iterations and the logarithm of the residual, respectively. Figure 2, both the objective value and the residual decreased monotonically for the MNIST dataset and the Fashion-MNIST dataset, which validates our theoretical guarantees in Theorem 4.2. Moreover, Figure 3 illustrates that both the objective value and the residual diverge when \( \rho = 10^{-6} \). The curves fluctuated drastically on the objective value. Even though there was a decreasing trend for the residual, it still fluctuated irregularly and failed to converge.

Figure 3: Divergence curves of the diADMM algorithm for the MNIST and the Fashion MNIST datasets when \( \rho = 10^{-6} \); diADMM algorithm diverged.

Figure 4: Performance of all methods for the MNIST dataset: diADMM algorithm outperformed most of the comparison methods.

Figure 5: Performance of all methods for the Fashion MNIST dataset: diADMM algorithm outperformed most of the comparison methods.

5.2.2 Performance. Figure 4 and Figure 5 show the curves of the training accuracy and test accuracy of our proposed diADMM algorithm and baselines, respectively. Overall, both the training accuracy and the test accuracy of our proposed diADMM algorithm outperformed most baselines for both the MNIST dataset and the Fashion MNIST dataset. Specifically, the curves of our diADMM algorithm soared to 0.8 at the early stage, and then raised steadily towards 0.9 or more. The curves of the most SGD-related methods, SGD, Adadelta, and Adagrad, moved more slowly than our proposed diADMM algorithm. The curves of the ADMM also rocketed to around 0.8, but decreased slightly later on. Only the state-of-the-art Adam performed better than diADMM slightly.

5.2.3 Scalability Analysis. In this subsection, the relationship between running time per iteration of our proposed diADMM algorithm and three potential factors, namely, the value of \( \rho \), the size of training samples, and the number of neurons was explored. The running time was calculated by the average of 200 iterations.

Firstly, when the training size was fixed, the computational result demonstrated that diADMM algorithm diverged.

Table 2: The relationship between running time per iteration (in second) and the number of neurons for each layer as well as value of \( \rho \). For the training size was fixed: generally, the running time increased as the number of neurons and the value of \( \rho \) became larger.

Table 3: The relationship between running time per iteration (in second) and the size of training samples as well as value of \( \rho \) when the number of neurons is fixed: generally, the running time increased as the training sample and the value of \( \rho \) became larger.
to $10^{-2}$, with multiplying by 10 each time. Generally, the running time increased as the number of neurons and the value of $\rho$ became larger. However, there were a few exceptions: for example, when there were 200 neurons for the MNIST dataset, and $\rho$ increased from $10^{-5}$ to $10^{-4}$, the running time per iteration dropped from 2.8778 seconds to 2.2761 seconds.

Secondly, we fixed the number of neurons for each layer as 1, 000. The relationship between running time per iteration, the training size and the value of $\rho$ is shown in Table 3. The value of $\rho$ ranged from $10^{-6}$ to $10^{-2}$, with multiplying by 10 each time. The training size of the MNIST dataset ranged from 11,000 to 55,000, with an increase of 11,000 each time. The training size of the Fashion MNIST dataset ranged from 12,000 to 60,000, with an increase of 12,000 each time. Similar to Table 3, the running time increased generally as the training sample and the value of $\rho$ became larger and some exceptions exist.

6 CONCLUSION AND FUTURE WORK

Alternating Direction Method of Multipliers (ADMM) is a good alternative to Stochastic Gradient Descent (SGD) for deep learning problems. In this paper, we propose a novel deep learning Alternating Direction Method of Multipliers (dADMM) to address some previously mentioned challenges. Firstly, the dADMM updates parameters from backward to forward in order to transmit parameter information more efficiently. The time complexity is successfully reduced from $O(n^3)$ to $O(n^2)$ by iterative quadratic approximations and backtracking. Finally, the dADMM is guaranteed to converge to a critical solution under mild conditions. Experiments on benchmark datasets demonstrate that our proposed dADMM algorithm outperformed most of the comparison methods.

In the future, we may extend our dADMM from the fully-connected neural network to the famous Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN), because our convergence guarantee is also applied to them. We also consider other non-linear activation functions such as sigmoid and hyperbolic tangent function (tanh).

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation grant: #1755850.

REFERENCES

[1] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences, 2(1):183–202, 2009.
[2] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT’2010, pages 177–186. Springer, 2010.
[3] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning, 3(1):1–122, 2011.
[4] Cailin Zhou, Min Li, Xin Liu, and Yunyi Ye. Extended adm and adm for separable convex minimization models with quadratic coupling terms: convergence analysis and insights. Mathematical Programming, pages 1–41, 2015.
[5] Francois Chollet. Deep learning with python. Manning Publications Co., 2017.
[6] Wei Deng, Ming-Jun Lai, ZhiBin Peng, and Wotao Yin. Parallel multi-blockadm with o (1/k) convergence. Journal of Scientific Computing, 71(2):712–736, 2017.
[7] John Duch, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12(Dec):2121–2159, 2011.
[8] Yuyang Gao, Liang Zhao, Lingfei Wu, Yanfang Ye, Hui Xiong, and Chaowei Yang. Incomplete label multi-task deep learning for spatio-temporal event subtype forecasting. 2019.
[9] Farkhondeh Kharee, Christian Gagné, and Mahdieh Abbasi. Alternating direction method of multipliers for sparse convolutional neural networks. arXiv preprint arXiv:1611.01596, 2016.
[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.
[12] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature, 521(7553):436, 2015.
[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.
[14] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In Eleventh Annual Conference of the International Speech Communication Association, 2010.
[15] Boris T Polyak. Some methods of speeding up the convergence of iterations methods. USSR Computational Mathematics and Mathematical Physics, 45(1):1–17, 1964.
[16] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In International Conference on Learning Representations, 2018.
[17] R Tyrrell Rockafellar and Roger J-B Wets. Varational analysis, volume 317. Springer Science & Business Media, 2009.
[18] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. nature, 323(6088):533, 1986.
[19] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In International conference on machine learning, pages 1139–1147, 2013.
[20] Gavin Taylor, Ryan Burmeister, Zheng Xu, Bharat Singh, Ankit Patel, and Tom Goldstein. Training neural networks without gradients: A scalable admn approach. In International Conference on Machine Learning, pages 2722–2731, 2016.
[21] T Tieleman and G Hinton. Divide the gradient by a running average of its recent magnitude (accessed on 21 April 2017).
[22] Junxiang Wang and Liang Zhao. Nonconvex generalizations of admm for nonlinear equality constrained problems. arXiv preprint arXiv:1705.03412, 2017.
[23] Junxiang Wang, Liang Zhao, and Lingfei Wu. Multi-convex inequality-constrained alternating direction method of multipliers. arXiv preprint arXiv:1902.10882, 2019.
[24] Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of adm in nonconvex non-smooth optimization. Journal of Scientific Computing, pages 1–35, 2015.
[25] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017.
[26] Matthew D Zeiler. Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701, 2012.
A ALGORITHMS TO UPDATE $W^{k+1}_l$ AND $a_l^{k+1}$

The algorithms to update $W^{k+1}_l$ and $a_l^{k+1}$ are described in the Algorithms 4 and 5, respectively.

Algorithm 4 The Backtracking Algorithm to update $W^{k+1}_l$

Require: $W^{k+1}_{l-1}, b^{k+1}_l, z^{k+1}_l, a^{k+1}_l, a^k_l, u^k$, $\rho$, some constant $\gamma > 1$.
Ensure: $\theta^k_l, W_l^{k+1}$.

1. Pick up $\alpha$ and $\zeta = W^{k+1}_l - \nabla W^{k+1}_l \phi/\alpha$.
2. while $\phi(W^{k+1}_l, l_{i=1}^L b^{k+1}_l, z^{k+1}_l, a^{k+1}_l, a^k_l, u^k) > P_1(\zeta, \alpha)$ do
   3. $\alpha \leftarrow \alpha \gamma$.
   4. Solve $\xi$ by Equation (11).
5. end while
6. Output $\theta^k_l \leftarrow \alpha$.
7. Output $W_l^{k+1} \leftarrow \zeta$.

Algorithm 5 The Backtracking Algorithm to update $a_l^{k+1}$

Require: $W^{k+1}_l, b^{k+1}_l, z^{k+1}_l, a^{k+1}_l, a^k_l, u^k$, $\rho$, some constant $\eta > 1$.
Ensure: $\tau_l^{k+1}, a^{k+1}_l$.

1. Pick up $\tau$ and $\beta = W^{k+1}_l - \nabla W^{k+1}_l \phi/\tau$.
2. while $\phi(W^{k+1}_l, l_{i=1}^L b^{k+1}_l, z^{k+1}_l, a^{k+1}_l, a^k_l, u^k) > Q_1(\beta, \tau)$ do
   3. $\tau \leftarrow t^\eta$.
   4. $\beta \leftarrow \beta^\eta - \nabla W^{k+1}_l \phi/\tau$.
5. end while
6. Output $\tau^{k+1}_l \leftarrow \tau$.
7. Output $a^{k+1}_l \leftarrow \beta$.

B LEMMAS FOR THE PROOFS OF PROPERTIES

The following several lemmas are preliminary results.

**Lemma B.1.** Equation (9) holds if and only if there exists $\tilde{\zeta} \in \partial \Omega_1(\tilde{W}_l^{k+1})$, the subgradient of $\Omega_1(\tilde{W}_l^{k+1})$ such that

$$\nabla_{W_l^{k+1}} \phi(W^{k+1}_l, b^{k+1}_l, z^{k+1}_l, a^{k+1}_l, u^k) + \frac{\partial^{k+1}_l}{\partial (\tilde{W}_l^{k+1} - W_l^{k+1})} \tilde{\zeta} = 0$$

**Likewise, Equation (11) holds if and only if there exists $\tilde{\zeta} \in \partial \Omega_1(W^{k+1}_l)$, the subgradient of $\Omega_1(W^{k+1}_l)$ such that**

$$\nabla_{W_l^{k+1}} \phi(W^{k+1}_l, b^{k+1}_l, z^{k+1}_l, a^{k+1}_l, u^k) + \frac{\partial^{k+1}_l}{\partial (W^{k+1}_l - \tilde{W}_l^{k+1})} \tilde{\zeta} = 0$$

**Proof.** These can be obtained by directly applying the optimality conditions of Equation (9) and Equation (11), respectively. \(\square\)

**Lemma B.2.** $\nabla_{z^{k+1}_L} R(z^{k+1}_L, y) + u^k = 0$ for all $k \in \mathbb{N}$.

**Proof.** The optimality condition of $z_L^{k+1}$ in Equation (15) gives rise to

$$\nabla_{z^{k+1}_L} R(z^{k+1}_L, y) + \rho(z^{k+1}_L - W_L^{k+1} u_{L-1} - b^k_L) + u^{k-1} = 0$$

Because $u^k = u^k - \rho(z^{k+1}_L - W_L^{k+1} u_{L-1} - b^k_L)$, then we have $\nabla_{z^{k+1}_L} R(z^{k+1}_L, y) + u^k = 0$. \(\square\)

**Lemma B.3.** It holds that $\forall z_{L-1}, z_L \in \mathbb{R}^{n_L}$,

$$R(z_{L-1}, y) \leq R(z_{L-2}, y) + \nabla z_L R^T z_{L-2} (z_{L-2}, y) (z_{L-1} - z_L) + (H/2) ||z_{L-1} - z_L||^2$$

$$- R(z_{L-1}, y) \leq -R(z_{L-2}, y) - \nabla z_L R^T z_{L-2} (z_{L-2}, y) (z_{L-1} - z_L) + (H/2) ||z_{L-1} - z_L||^2$$

**Proof.** Because $R(z_L, y)$ is Lipschitz differentiable by Assumption 2, so is $-R(z_L, y)$. Therefore, this lemma is proven exactly as same as Lemma 2.1 in [1]. \(\square\)
LEMMA B.4. For Equations (6) and (12), if $\overline{B}, B \geq v$, then the following inequalities hold:
\begin{align}
U_l(\overline{b}_l^{k+1}; B) & \geq \phi(W_l^{k+1}, \overline{b}_l^{k+1}, \overline{z}_l^{k+1}, \overline{a}_l^{k+1}, u^k) \\
U_B(b_B^{k+1}; B) & \geq \phi(W_B^{k+1}, b_B^{k+1}, z_B^{k+1}, a_B^{k+1}, u^k)
\end{align}

Proof. Because $\phi(W, b, z, a, u)$ is Lipschitz differentiable with respect to $b$ with Lipschitz coefficient $v$ (the definition of Lipschitz differentiability can be found in [1]), we directly apply Lemma 2.1 in [1] to $\phi$ to obtain Equations (19) and (20), respectively. \qed

LEMMA B.5. It holds that for all $k \in \mathbb{N}$,
\begin{align}
& L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) - L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) \geq \|\overline{z}_l^{k+1} - b_l^{k+1}\|^2/2(l = 1, \ldots, L - 1) \\
& L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) \geq L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k)(l = 1, \ldots, L - 1) \\
& L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) \geq \rho(2/\|b_l^{k+1} - b_l^k\|^2)(l = 1, \ldots, L - 1) \\
& L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) - L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) \geq \|\overline{z}_l^{k+1} - b_l^{k+1}\|^2/2(l = 1, \ldots, L - 1) \\
& L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) \geq L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k)(l = 1, \ldots, L - 1) \\
& L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) \geq \rho(2/\|b_l^{k+1} - b_l^k\|^2)(l = 1, \ldots, L - 1) \\
& L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) - L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) \geq \|\overline{z}_l^{k+1} - b_l^{k+1}\|^2/2(l = 1, \ldots, L - 1) \\
& L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) \geq L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k)(l = 1, \ldots, L - 1) \\
& L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) \geq \rho(2/\|b_l^{k+1} - b_l^k\|^2)(l = 1, \ldots, L - 1) \\
& L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) - L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) \geq \|\overline{z}_l^{k+1} - b_l^{k+1}\|^2/2(l = 1, \ldots, L - 1)
\end{align}

Proof. Essentially, all inequalities can be obtained by applying optimality conditions of updating $\overline{a}_l^{k+1}, \overline{z}_l^{k+1}, \overline{b}_l^{k+1}, \overline{W}_l^{k+1}, z_l^{k+1}, W_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}$, and $a_l^{k+1}$, respectively. We only prove Inequality (23), (27), (28) and (30). This is because Inequalities (21), (26) and (31) follow the routine of Inequality (27), Inequalities (24), (25) and (29) follow the routine of Inequality (28), and Inequality (22) follows the routine of Inequality (30).

Firstly, we focus on Inequality (23).
\begin{align}
L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) & - L_p(\overline{W}_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) \\
& = R(z_l^{k+1}; y) + (u^k)^T(z_l^{k+1} - W_l^{k}a_l^{k+1} - b_l^{k+1}) + \rho(2/\|b_l^{k+1} - b_l^k\|^2) \|z_l^{k+1} - W_l^{k}a_l^{k+1} - b_l^{k+1}\|^2 \\
& - R(z_l^{k+1}; y) - (u^k)^T(z_l^{k+1} - W_l^{k}a_l^{k+1} - b_l^{k+1}) - \rho(2/\|b_l^{k+1} - b_l^k\|^2) \|z_l^{k+1} - W_l^{k}a_l^{k+1} - b_l^{k+1}\|^2 \\
& = R(z_l^{k+1}; y) - R(z_l^{k+1}; y) + (u^k)^T(z_l^{k+1} - W_l^{k}a_l^{k+1} - b_l^{k+1}) + \rho(2/\|b_l^{k+1} - b_l^k\|^2) \|z_l^{k+1} - z_l^{k+1}\|^2 \\
& + \rho(2/\|b_l^{k+1} - b_l^k\|^2) \|z_l^{k+1} - W_l^{k}a_l^{k+1} - b_l^{k+1}\|^2 (z_l^{k+1} - z_l^{k+1})
\end{align}

where the second equality follows from the cosine rule $\|z_l^{k+1} - W_l^{k}a_l^{k+1} - b_l^{k+1}\|^2 - \|z_l^{k+1} - W_l^{k}a_l^{k+1} - b_l^{k+1}\|^2 = \|z_l^{k+1} - z_l^{k+1}\|^2 + (z_l^{k+1} - W_l^{k}a_l^{k+1} - b_l^{k+1})^T(z_l^{k+1} - z_l^{k+1})$.

According to the optimality condition of Equation (5), we have $\nabla_{z_l^{k+1}} R(z_l^{k+1}; y) + u^k + \rho(z_l^{k+1} - W_l^{k}a_l^{k+1} - b_l^{k+1}) = 0$. Because $R(z_l^{k+1}; y)$ is convex and differentiable with regard to $z_l^{k+1}$, its subgradient is also its gradient. According to the definition of subgradient, we have
\begin{align}
R(z_l^{k+1}; y) & \geq R(z_l^{k+1}; y) + \nabla_{z_l^{k+1}} R(z_l^{k+1}; y) (z_l^{k+1} - z_l^{k+1}) \\
& = R(z_l^{k+1}; y) - (u^k + \rho(z_l^{k+1} - W_l^{k}a_l^{k+1} - b_l^{k+1}))^T(z_l^{k+1} - z_l^{k+1})
\end{align}

We introduce Equation (33) into Equation (32) to obtain Equation (23).

Secondly, we focus on Inequality (27). The stopping criterion of Algorithm 4 shows that
\begin{align}
\phi(W_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) & \leq P_l(W_l^{k+1}, z_l^{k+1}) \\
\phi(W_l^{k+1}, b_l^{k+1}, z_l^{k+1}, a_l^{k+1}, u^k) & \leq P_l(W_l^{k+1}, z_l^{k+1})
\end{align}

Because $\Omega W_l(W_l)$ is convex, according to the definition of subgradient, we have
\begin{align}
\Omega_l(W_l^{k+1}) & \geq \Omega_l(W_l^{k+1}) + s^T(W_l^{k+1} - W_l^{k+1})
\end{align}
where \( s \) is defined in the premise of Lemma B.1. Therefore, we have

\[
L_p(W^{k+1}_{l-1}, b^{k+1}_{l-1}, a^{k+1}_{l-1}, u^k) - L_p(W^{k+1}_{l-1}, b^{k+1}_{l-1}, a^{k+1}_{l-1}, u^k) = \phi(W^{k+1}_{l-1}, b^{k+1}_{l-1}, a^{k+1}_{l-1}, u^k) + \Omega_l(W^{k+1}_{l-1}) - \phi(W^{k+1}_{l-1}, b^{k+1}_{l-1}, a^{k+1}_{l-1}, u^k) - \Omega_l(W^{k+1}_{l-1}) \quad \text{(Definition of } L_p).
\]

\[
\geq \Omega_l(W^{k+1}_{l-1}) - \Omega_l(W^{k+1}_{l-1}) - \nabla \phi(W^{k+1}_{l-1}, a^{k+1}_{l-1}, u^k) ||\phi(W^{k+1}_{l-1}, b^{k+1}_{l-1}, a^{k+1}_{l-1}, u^k) - (W^{k+1}_{l-1} - W^{k+1}_{l-1})^2||/2 \quad \text{(Equation (34))}
\]

\[
\geq s^T(W^{k+1}_{l-1} - W^{k+1}_{l-1}) - \nabla \phi(W^{k+1}_{l-1}, a^{k+1}_{l-1}, u^k) ||\phi(W^{k+1}_{l-1}, b^{k+1}_{l-1}, a^{k+1}_{l-1}, u^k) - (W^{k+1}_{l-1} - W^{k+1}_{l-1})^2||/2 \quad \text{(Equation (35))}
\]

\[
= (\gamma^T + \nabla \phi(W^{k+1}_{l-1}, a^{k+1}_{l-1}, u^k)) ||\phi(W^{k+1}_{l-1}, b^{k+1}_{l-1}, a^{k+1}_{l-1}, u^k) - (W^{k+1}_{l-1} - W^{k+1}_{l-1})^2||/2 \quad \text{(Lemma B.1)}.
\]

Thirdly, we focus on Inequality (28).

\[
L_p(W^{k+1}_{l-1}, b^{k+1}_{l-1}, a^{k+1}_{l-1}, u^k) - L_p(W^{k+1}_{l-1}, b^{k+1}_{l-1}, a^{k+1}_{l-1}, u^k) = \phi(W^{k+1}_{l-1}, b^{k+1}_{l-1}, a^{k+1}_{l-1}, u^k) - \phi(W^{k+1}_{l-1}, b^{k+1}_{l-1}, a^{k+1}_{l-1}, u^k) \geq (\gamma/2) ||\phi(x^{k+1}_{l-1} - b^{k+1}_{l-1})^2||. \quad \text{(Lemma B.4)}
\]

Finally, we focus on Equation (30). This follows directly from the optimality of \( x^{k+1}_l \) in Equation (14).

\[ \square \]

**Lemma B.6.** If \( \rho > 2H \) so that \( C_1 = \rho/2 - H/2 - H^2/\rho > 0 \), then it holds that

\[
L_p(W^{k+1}, b^{k+1}, a^{k+1}, u^k) - L_p(W^{k+1}, b^{k+1}, a^{k+1}, u^k) \geq C_1 ||x^{k+1}_l - z^{k+1}_l - (H^2/\rho) ||x^{k+1}_l - z^{k+1}_l ||_2^2. \quad \text{(36)}
\]

**Proof.**

\[
L_p(W^{k+1}, b^{k+1}, a^{k+1}, u^k) - L_p(W^{k+1}, b^{k+1}, a^{k+1}, u^k) = R(\mathbb{E}^{k+1}_l; y) - R(z^{k+1}_l; y) + (u^{k+1}_l)^T(z^{k+1}_l - z^{k+1}_l) + (\rho/2) ||z^{k+1}_l - z^{k+1}_l||^2 - (1/\rho) ||u^{k+1}_l - u^k||^2
\]

\[
= R(\mathbb{E}^{k+1}_l; y) - R(z^{k+1}_l; y) + \nabla R(z^{k+1}_l; y)^T(z^{k+1}_l - z^{k+1}_l) + (\rho/2) ||z^{k+1}_l - z^{k+1}_l||^2 - (1/\rho) ||u^{k+1}_l - u^k||^2 \quad \text{(Lemma B.2)}
\]

\[
\geq (-H/2) ||z^{k+1}_l - z^{k+1}_l||^2 + (\rho/2) ||z^{k+1}_l - z^{k+1}_l||^2 - (1/\rho) ||\nabla z^{k+1}_l R(z^{k+1}_l; y) - \nabla z^{k+1}_l; y||^2
\]

\[
= -R(z^{k+1}_l; y) \quad \text{(Lipschitz differentiable, Lemmas B.2 and B.3)}
\]

\[
\geq (-H/2) ||z^{k+1}_l - z^{k+1}_l||^2 + (\rho/2) ||z^{k+1}_l - z^{k+1}_l||^2 - (H^2/\rho) ||z^{k+1}_l - z^{k+1}_l||^2 \quad \text{(Assumption 2)}
\]

\[
\geq (-H/2) ||z^{k+1}_l - z^{k+1}_l||^2 + (\rho/2) ||z^{k+1}_l - z^{k+1}_l||^2 - (H^2/\rho) ||z^{k+1}_l - z^{k+1}_l||^2 \quad \text{(triangle inequality)}
\]

\[
= C_1 ||x^{k+1}_l - z^{k+1}_l||^2 - (H^2/\rho) ||x^{k+1}_l - z^{k+1}_l||^2. \quad \text{(36)}
\]

We choose \( \rho > 2H \) to make \( C_1 > 0 \). \[ \square \]

**C PROOF OF THEOREM 4.1**

Proving Theorem 4.1 is equal to proving jointly Theorem C.1, C.2, and C.3, which are elaborated in the following.

**Theorem C.1.** Given that Assumptions 1 and 2 hold, the dADMM satisfies Property 1.

**Proof.** There exists \( z^{k+1}_l \) such that \( z^{k+1}_l - W^{k+1}_l a^{k+1}_{l-1} - b^{k+1}_l = 0 \). By Assumption 2, we have

\[
F(W^{k+1}, b^{k+1}, (z^{k+1}_l)_{l=1}^{l-1}, z^{k+1}_l, a^{k+1}) \geq \min S > -\infty
\]
where \( S = \{ F(\mathbf{W}, \mathbf{b}, \mathbf{z}, \mathbf{a}) : z_L = W_i a_{l-1} - b_L = 0 \} \). Then we have

\[
L_p(\mathbf{W}^k, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k) = F(\mathbf{W}, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k)^T (x_L - \mathbf{W}_L^k a_{l-1}^k - b_L^k) + (\rho/2) \| z_L^k - \mathbf{W}_L^k a_{l-1}^k - b_L^k \|_2^2 \\
= F(\mathbf{W}, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k)^T (x_L - \mathbf{W}_L^k a_{l-1}^k - b_L^k) + (\rho/2) \| z_L^k - \mathbf{W}_L^k a_{l-1}^k - b_L^k \|_2^2 \\
= F(\mathbf{W}, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k)^T (x_L - \mathbf{W}_L^k a_{l-1}^k - b_L^k) + (\rho/2) \| z_L^k - \mathbf{W}_L^k a_{l-1}^k - b_L^k \|_2^2 (\text{Lemma B.2}) \\
= \sum_{l=1}^L \Omega_l(W_l^k) + (\rho/2) \sum_{l=1}^L (\| z_L^k - \mathbf{W}_L^k a_{l-1}^k - b_L^k \|_2^2 + \| a_l^k - f(z_L^k) \|_2^2) + (\rho/2) \| z_L^k - \mathbf{W}_L^k a_{l-1}^k - b_L^k \|_2^2 \\
> -\infty
\]

It concludes from Lemma 5.6 and Lemma B.6 that \( L_p(\mathbf{W}^k, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k) \) is upper bounded by \( L_p(\mathbf{W}, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k) \) and so are \( \sum_{l=1}^L \Omega_l(W_l^k)^T + (\rho/2) \sum_{l=1}^L (\| z_L^k - \mathbf{W}_L^k a_{l-1}^k - b_L^k \|_2^2 + \| a_l^k - f(z_L^k) \|_2^2) \\
\) and \( \| z_L^k - \mathbf{W}_L^k a_{l-1}^k - b_L^k \|_2^2 \). By Assumption 2, \( \mathbf{W}^k, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k \) is bounded. By Lemma B.2, it is obvious that \( \mathbf{u}^k \) is bounded as well.

\textbf{Theorem C.2.} Given that Assumptions 1 and 2 hold, the dlADM satisfies Property 2.

\textbf{Proof.} This follows directly from Lemma 5.5 and Lemma B.6.

\textbf{Theorem C.3.} Given that Assumptions 1 and 2 hold, the dlADM satisfies Property 3.

\textbf{Proof.} We know that

\[
\partial L_p(\mathbf{W}^k, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k) = \partial \Omega_l(W_l^k) + \nabla \Omega_l(W_l^k) \phi(W_l^k, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k)
\]

where \( \partial \Omega_l(W_l^k) = (\partial \Omega_l(W_l^k) + \nabla \Omega_l(W_l^k) \phi(W_l^k, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k) \) and \( \partial \Omega_l(W_l^k) = (\nabla \partial \Omega_l(W_l^k) + \partial \nabla \Omega_l(W_l^k) \phi(W_l^k, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k) \). To prove Property 3, we need to give an upper bound of \( \partial \Omega_l(W_l^k) + \nabla \Omega_l(W_l^k) \phi(W_l^k, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k) \) by a linear combination of \( \| \mathbf{W}^k - \mathbf{W}^k \|, \| \mathbf{b}^k - \mathbf{b}^k \|, \| \mathbf{z}^k - \mathbf{z}^k \| \), \( \| \mathbf{a}^k - \mathbf{a}^k \| \) and \( \| u^k - u^k \| \).

For \( W_l^k \) \( (l < L) \),

\[
\partial W_l^k = \partial \Omega_l(W_l^k) + \nabla \Omega_l(W_l^k) \phi(W_l^k, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k)
\]

Because

\[
\| \partial \Omega_l(W_l^k) + \nabla \Omega_l(W_l^k) \phi(W_l^k, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k) \| \leq \| \partial \Omega_l(W_l^k) \| + \| \nabla \Omega_l(W_l^k) \phi(W_l^k, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k) \| \\
\leq \| \partial \Omega_l(W_l^k) \| + \| \mathbf{W}^k - \mathbf{W}^k \| \| \partial \Omega_l(W_l^k) \| + \| \mathbf{b}^k - \mathbf{b}^k \| \| \partial \Omega_l(W_l^k) \| + \| \mathbf{z}^k - \mathbf{z}^k \| \| \partial \Omega_l(W_l^k) \| + \| \mathbf{a}^k - \mathbf{a}^k \| \| \partial \Omega_l(W_l^k) \| \\
= \| \partial \Omega_l(W_l^k) \| + \| \mathbf{W}^k - \mathbf{W}^k \| \| \partial \Omega_l(W_l^k) \| + \| \mathbf{b}^k - \mathbf{b}^k \| \| \partial \Omega_l(W_l^k) \| + \| \mathbf{z}^k - \mathbf{z}^k \| \| \partial \Omega_l(W_l^k) \| + \| \mathbf{a}^k - \mathbf{a}^k \| \| \partial \Omega_l(W_l^k) \|
\]

and the optimality condition of Equation (11) yields

\[
0 \in \partial \Omega_l(W_l^k) + \nabla \Omega_l(W_l^k) \phi(W_l^k, \mathbf{b}^k, \mathbf{z}^k, \mathbf{a}^k, u^k) + \partial \Omega_l(W_l^k)
\]
Because $a_{k+1}^k$ is bounded by Property 1, $\|\partial W_{k+1}^k L_{\rho}\|$ can be upper bounded by a linear combination of $\|W_{k+1}^k - \overline{W}_L^k\|$, $\|b_{k+1}^k - \overline{b}_L^k\|$ and $\|z_{k+1}^k - \overline{z}_L^k\|$. For $W_{k+1}^k$,

$$\partial W_{k+1}^k L_{\rho} = \partial Q_L(W_{k+1}^k) + \nabla W_{k+1}^k \phi(W_{k+1}^k, b_{k+1}^k, z_{k+1}^k, a_{k+1}^k, u_{k+1}^k)$$

$$= \partial Q_L(W_{k+1}^k) + \nabla W_{k+1}^k \phi(W_{k+1}^k, b_{L_{k+1}}^k, z_{L_{k+1}}^k, a_{L_{k+1}}^k, u_{k}^k) + \partial \theta_k^k o (W_{k+1}^k - \overline{W}_L^k)$$

$$- \nabla W_{k+1}^k \phi(W_{k+1}^k, b_{L_{k+1}}^k, z_{L_{k+1}}^k, a_{L_{k+1}}^k, u_{k}^k) + \nabla W_{k+1}^k \phi(W_{k+1}^k, b_{k+1}^k, z_{k+1}^k, a_{k+1}^k, u_{k+1}^k)$$

$$= \partial Q_L(W_{k+1}^k) + \nabla W_{k+1}^k \phi(W_{k+1}^k, b_{k+1}^k, z_{k+1}^k, a_{k+1}^k, u_{k+1}^k) + \partial \theta_k^k o (W_{k+1}^k - \overline{W}_L^k) - \partial \theta_k^k o (W_{k+1}^k - \overline{W}_L^k)$$

$$+ \rho(W_{k+1}^k b_{L_{k+1}}^k - b_{k+1}^k - z_{k+1}^k - u_{k}^k / \rho)(a_{L_{k+1}}^k)^T - \rho(W_{k+1}^k a_{L_{k+1}}^k + b_{L_{k+1}}^k - z_{L_{k+1}}^k - u_{k}^k / \rho)(a_{L_{k+1}}^k)^T)$$

Because

$$\| - \partial \theta_k^k o (W_{k+1}^k - \overline{W}_L^k) + \rho(W_{k+1}^k, a_{k+1}^k + b_{k+1}^k - z_{k+1}^k - u_{k}^k / \rho)(a_{k+1}^k)^T - \rho(W_{k+1}^k a_{k+1}^k + b_{k+1}^k - z_{k+1}^k - u_{k}^k / \rho)(a_{k+1}^k)^T\|$$

$$= \| - \partial \theta_k^k o (W_{k+1}^k - \overline{W}_L^k) + \rho(W_{k+1}^k a_{k+1}^k + b_{k+1}^k - z_{k+1}^k - u_{k}^k / \rho)(a_{k+1}^k)^T - \rho(W_{k+1}^k a_{k+1}^k + b_{k+1}^k - z_{k+1}^k - u_{k}^k / \rho)(a_{k+1}^k)^T\|$$

$$\leq \| \partial \theta_k^k o (W_{k+1}^k - \overline{W}_L^k) \| + \rho \| W_{k+1}^k - \overline{W}_L^k \| a_{k+1}^k (a_{k+1}^k)^T \| + \rho \| W_{k+1}^k - \overline{W}_L^k \| (a_{k+1}^k)^T \| + \rho \| z_{k+1}^k - \overline{z}_L^k \| \| a_{k+1}^k \| + H \| z_{k+1}^k - \overline{z}_L^k \| \| a_{k+1}^k \|$$

(Cauchy-Schwarz inequality, Lemma 2.3, R(Z(k); y) is Lipschitz differentiable)

and the optimality condition of Equation (11) yields

$$0 \in \partial Q_L(W_{k+1}^k, b_{L_{k+1}}^k, z_{L_{k+1}}^k, a_{L_{k+1}}^k, u_{k}^k) + \partial \theta_k^k o (W_{k+1}^k - \overline{W}_L^k)$$

Because $a_{k+1}^k$ is bounded by Property 1, $\|\partial W_{k+1}^k L_{\rho}\|$ can be upper bounded by a linear combination of $\|W_{k+1}^k - \overline{W}_L^k\|$, $\|b_{k+1}^k - \overline{b}_L^k\|$, $\|z_{k+1}^k - \overline{z}_L^k\|$ and $\|x_{k+1}^k - \overline{x}_L^k\|$. For $b_{k+1}^k (k < L)$,

$$\nabla b_{k+1}^k L_{\rho} = \nabla b_{k+1}^k \phi(W_{k+1}^k, b_{k+1}^k, z_{k+1}^k, a_{k+1}^k, i_{k+1}^k)$$

$$= \nabla b_{k+1}^k \phi(W_{L}^k, b_{L}^k, z_{L}^k, a_{L}^k, i_{L})$$

$$= \nabla b_{k+1}^k \phi(W_{L}^k, b_{L}^k, z_{L}^k, a_{L}^k, i_{L}) + \nabla b_{L}^k \phi(W_{L}^k, b_{L}^k, z_{L}^k, a_{L}^k, i_{L})$$

$$= \nabla b_{L}^k \phi(W_{L}^k, b_{L}^k, z_{L}^k, a_{L}^k, i_{L}) + \nabla b_{L}^k \phi(W_{L}^k, b_{k+1}^k, z_{k+1}^k, a_{k+1}^k, i_{k+1}^k)$$

The optimality condition of Equation (12) yields

$$0 \in \nabla b_{k+1}^k \phi(W_{k+1}^k, b_{k+1}^k, z_{k+1}^k, a_{k+1}^k, i_{k+1}^k) + \nabla b_{k+1}^k \phi(W_{L}^k, b_{L}^k, z_{L}^k, a_{L}^k, i_{L})$$

Therefore, $\|\partial b_{k+1}^k L_{\rho}\|$ is linearly independent on $\|x_{k+1}^k - \overline{x}_L^k\|$. For $b_{k+1}^k$,
Therefore, $\|\nabla z^k - \Pi_{L^k} z^k\|$ is upper bounded by a combination of $\|z^{k+1} - z^k\|$ and $\|z^k - z^k\|$.

For $\partial x^k (t < L)$,

$$\partial x^k \in \triangledown L_{R} \phi(W^{k+1}, b^{k+1}, x^{k+1}, a^{k+1}, u^k)$$

because $x^k$ is bounded and $f_t(z^k)$ is continuous and hence $f_t(z^k)$ is bounded, and the optimality condition of Equation (14) yields

$$0 \in \triangledown L_{R} \phi(W^{k+1}, b^{k+1}, x^k, a^{k+1}, u^k)$$

and the optimality condition of Equation (15) yields

$$0 \in \triangledown R(W^{k+1}, b^{k+1}, x^{k+1}, a^{k+1}, u^k)$$

and

$$\|u^k - u^k\| = \|\triangledown L_{R} R(z^{k+1}; y) - \triangledown L_{R} R(z^k; y)\| (\text{Lemma 2.2}) \leq H(\|z^{k+1} - z^k\|)$$

(Cauchy-Schwarz inequality, $R(z; y)$ is Lipschitz differentiable)
Because
\[
||-\mathbf{t}^{k+1} \circ (\mathbf{d}^{k+1} - \mathbf{a}^{k+1} || - v(\mathbf{W}^{k+1})^T (\mathbf{W}^{k+1} \mathbf{a}^{k+1} + \mathbf{b}^{k+1} - \mathbf{x}^{k+1} - v(\mathbf{a}^{k+1} - f(\mathbf{x}^{k+1}))) \\
+ v(\mathbf{W}^{k+1})^T (\mathbf{W}^{k+1} \mathbf{a}_{l+1} - \mathbf{b}_{l+1} - z_{l+1} - v(\mathbf{a}_{l+1}) - f(z_{l+1}))) \\
\leq ||-\mathbf{t}^{k+1} \circ (\mathbf{a}^{k+1} - \mathbf{a}_{l+1}) || + ||\mathbf{a}_{l+1} - \mathbf{a}^{k+1} || + ||(\mathbf{W}^{k+1})^T \mathbf{a}^{k+1} - (\mathbf{W}^{k+1})^T \mathbf{a}_{l+1} || \\
+ ||((\mathbf{W}^{k+1})^T \mathbf{a}^{k+1} - (\mathbf{W}^{k+1})^T \mathbf{a}_{l+1} || + ||(\mathbf{W}^{k+1})^T \mathbf{a}^{k+1} - (\mathbf{W}^{k+1})^T \mathbf{a}_{l+1} || ||(\text{triangle inequality})
\]

Then we need to show that \(v((\mathbf{W}^{k+1})^T \mathbf{b}^{k+1} - (\mathbf{W}^{k+1})^T \mathbf{b}_{l+1} ||, ||\mathbf{b}^{k+1} - \mathbf{b}_{l+1} ||, ||\mathbf{z}^{k+1} - \mathbf{z}_{l+1} ||, ||\mathbf{a}^{k+1} - \mathbf{a}_{l+1} ||, ||\mathbf{k}^{k+1} - \mathbf{k}_{l+1} ||, \) are upper bounded by \(\|\mathbf{W}^{k+1} - \mathbf{W}_{l+1} \||, ||\mathbf{b}^{k+1} - \mathbf{b}_{l+1} ||, ||\mathbf{z}^{k+1} - \mathbf{z}_{l+1} ||, \) and \(||\mathbf{a}^{k+1} - \mathbf{a}_{l+1} ||\).

Because \(\|\mathbf{b}^{k+1} \||, \|\mathbf{W}_{l+1} \|| \) are upper bounded, \(v((\mathbf{W}^{k+1})^T \mathbf{b}^{k+1} - (\mathbf{W}^{k+1})^T \mathbf{b}_{l+1} \) is therefore upper bounded by a combination of \(\|\mathbf{W}^{k+1} - \mathbf{W}_{l+1} \||\) and \(\|\mathbf{b}^{k+1} - \mathbf{b}_{l+1} \||.

Similarly, \(v((\mathbf{W}^{k+1})^T \mathbf{z}^{k+1} - (\mathbf{W}^{k+1})^T \mathbf{z}_{l+1} \) is upper bounded by a combination of \(\|\mathbf{W}^{k+1} - \mathbf{W}_{l+1} \||\) and \(\|\mathbf{z}^{k+1} - \mathbf{z}_{l+1} \||.

Because \(\|\mathbf{W}^{k+1} \||, \|\mathbf{W}_{l+1} \|| \) and \(\|\mathbf{a}^{k+1} \|| \) are upper bounded, \(v((\mathbf{W}^{k+1})^T \mathbf{a}^{k+1} - (\mathbf{W}^{k+1})^T \mathbf{a}_{l+1} \) is therefore upper bounded by a combination of \(\|\mathbf{W}^{k+1} - \mathbf{W}_{l+1} \||\) and \(\|\mathbf{a}^{k+1} - \mathbf{a}_{l+1} \||.

For \(\mathbf{a}_{l+1}^{k+1}\),
\[
\nabla_{\mathbf{a}_{l+1}}^k \phi (\mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \mathbf{z}^{k+1}, \mathbf{a}^{k+1}, \mathbf{u}^{k+1}) \\
= \mathbf{t}^{k+1} \circ (\mathbf{a}^{k+1} - \mathbf{a}_{l+1}^k) + \nabla_{\mathbf{a}_{l+1}}^k \phi (\mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \mathbf{z}^{k+1}, \mathbf{a}^{k+1}, \mathbf{u}^{k+1}) - \mathbf{t}^{k+1} \circ (\mathbf{a}^{k+1} - \mathbf{a}_{l+1}^k) - \nabla_{\mathbf{a}_{l+1}}^k \phi (\mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \mathbf{z}^{k+1}, \mathbf{a}^{k+1}, \mathbf{u}^{k+1}) \\
+ \nabla_{\mathbf{a}_{l+1}}^k \phi (\mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \mathbf{z}^{k+1}, \mathbf{a}^{k+1}, \mathbf{u}^{k+1}) \\
= \mathbf{t}^{k+1} \circ (\mathbf{a}^{k+1} - \mathbf{a}_{l+1}^k) + \nabla_{\mathbf{a}_{l+1}}^k \phi (\mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \mathbf{z}^{k+1}, \mathbf{a}^{k+1}, \mathbf{u}^{k+1}) - \mathbf{t}^{k+1} \circ (\mathbf{a}^{k+1} - \mathbf{a}_{l+1}^k) - \rho (\mathbf{W}^{k+1})^T (\mathbf{W}^{k+1} \mathbf{a}_{l+1}^k + \mathbf{b}_{l+1}^k - \mathbf{x}^{k+1} - u^{k+1})/\rho \\
- \nabla_{\mathbf{a}_{l+1}}^k \phi (\mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \mathbf{z}^{k+1}, \mathbf{a}^{k+1}, \mathbf{u}^{k+1})
\]

Because
\[
||-\mathbf{t}^{k+1} \circ (\mathbf{a}^{k+1} - \mathbf{a}_{l+1}^k) || - \rho (\mathbf{W}^{k+1})^T (\mathbf{W}^{k+1} \mathbf{a}_{l+1}^k + \mathbf{b}_{l+1}^k - \mathbf{x}^{k+1} - u^{k+1})/\rho ||-\mathbf{t}^{k+1} \circ (\mathbf{a}^{k+1} - \mathbf{a}_{l+1}^k) \|
\]

Then we need to show that \(\rho ||((\mathbf{W}^{k+1})^T \mathbf{b}^{k+1} - (\mathbf{W}^{k+1})^T \mathbf{b}_{l+1} \), \(\rho ||((\mathbf{W}^{k+1})^T \mathbf{z}^{k+1} - (\mathbf{W}^{k+1})^T \mathbf{z}_{l+1} \), \(||((\mathbf{W}^{k+1})^T \mathbf{u}^{k+1} - (\mathbf{W}^{k+1})^T \mathbf{u}_{l+1} \), and \(\rho ||((\mathbf{W}^{k+1})^T \mathbf{a}^{k+1} - (\mathbf{W}^{k+1})^T \mathbf{a}_{l+1} \) are upper bounded by \(\|\mathbf{W}^{k+1} - \mathbf{W}_{l+1} \||, ||\mathbf{b}^{k+1} - \mathbf{b}_{l+1} \||, ||\mathbf{z}^{k+1} - \mathbf{z}_{l+1} \||, ||\mathbf{a}^{k+1} - \mathbf{a}_{l+1} \||, \) and \(||\mathbf{k}^{k+1} - \mathbf{k}_{l+1} ||, \)

triangle inequality, Cauchy-Schwarz inequality.
Because \( \|k^{k+1}\| \) and \( \|W_L^{k+1}\| \) are upper bounded, \( \rho \| (W_L^{k+1})^T b_L^{k+1} - (W_L^{k+1})^T b_L^{k+1} \| \) is therefore upper bounded by a combination of \( \|W_L^{k+1} - W_L^{k+1}\| \) and \( \|b_L^{k+1} - b_L^{k+1}\| \).

Similarly, \( \rho \| (W_L^{k+1})^T z_L^{k+1} - (W_L^{k+1})^T z_L^{k+1} \| \) is upper bounded by a combination of \( \|W_L^{k+1} - W_L^{k+1}\| \) and \( \|z_L^{k+1} - z_L^{k+1}\| \).

\[
\|W_L^{k+1}\| \|u^{k+1} - u^k\| + \|W_L^{k+1} - W_L^{k+1}\| \|u^k\| (\text{Cauchy-Schwarz inequality})
\]

\[
H \|W_L^{k+1}\| \|z_L^{k+1} - z_L^k\| + \|W_L^{k+1} - W_L^{k+1}\| \|u^k\| (R(z_L; y) \text{ is Lipschitz differentiable})
\]

Because \( \|W_L^{k+1}\| \), \( \|W_L^{k+1}\| \) and \( \|z_L^{k+1}\| \) are upper bounded, \( \rho \| (W_L^{k+1})^T W_L^{k+1} a_L^{k+1} - (W_L^{k+1})^T W_L^{k+1} a_L^{k+1} \| \) is therefore upper bounded by a combination of \( \|W_L^{k+1} - W_L^{k+1}\| \) and \( \|a_L^{k+1} - a_L^{k+1}\| \).

For \( u^{k+1} \),

\[
\nabla_{u^{k+1}} L_{\rho} = \nabla_{u^{k+1}} \phi(W_{L}^{k+1}, b_{L}^{k+1}, z_{L}^{k+1}, u^{k+1}) = z_{L}^{k+1} - W_{L}^{k+1} a_{L}^{k+1} - b_{L}^{k+1} = (1/\rho)(u^{k+1} - u^k)
\]

\[
(1/\rho) \nabla_{x_L} R(z_L; y) - \nabla_{x_L} R(z_L; y) (\text{Lemma B.2})
\]

Because

\[
\| (1/\rho) \nabla_{x_L} R(z_L; y) - \nabla_{x_L} R(z_L; y) \| \leq (H/\rho) \|z_L^{k+1} - z_L^k\| (R(z_L; y) \text{ is Lipschitz differentiable})
\]

Therefore, \( \| \nabla_{u^{k+1}} L_{\rho} \| \) is upper bounded by \( \|z_L^{k+1} - z_L^k\| \).

\[\square\]

### D PROOF OF THEOREM 4.3

**Proof.** To prove this theorem, we will first show that \( c_k \) satisfies two conditions: (1) \( c_k \geq c_{k+1} \); (2) \( \sum_{k=0}^{\infty} c_k \) is bounded. We then conclude the convergence rate of \( o(1/k) \) based on these two conditions. Specifically, first, we have

\[
c_k = \min_{0 \leq l \leq k} \left( \sum_{l=0}^{L} (\|W_L^{k+1} - W_L^{k+1}\|_2 + \|W_L^{k+1} - W_L^{k+1}\|_2 + \|b_L^{k+1} - b_L^{k+1}\|_2 + \|b_L^{k+1} - b_L^{k+1}\|_2 + \|z_L^{k+1} - z_L^{k+1}\|_2 + \|z_L^{k+1} - z_L^{k+1}\|_2)
\]

\[
+ \|W_L^{k+1} - z_L^{k+1}\|_2 + \|z_L^{k+1} - z_L^{k+1}\|_2)
\]

\[
\geq \min_{0 \leq l \leq k} \left( \sum_{l=0}^{L} (\|W_L^{k+1} - W_L^{k+1}\|_2 + \|W_L^{k+1} - W_L^{k+1}\|_2 + \|b_L^{k+1} - b_L^{k+1}\|_2 + \|b_L^{k+1} - b_L^{k+1}\|_2 + \|z_L^{k+1} - z_L^{k+1}\|_2 + \|z_L^{k+1} - z_L^{k+1}\|_2)
\]

\[
+ \|W_L^{k+1} - z_L^{k+1}\|_2 + \|z_L^{k+1} - z_L^{k+1}\|_2)
\]

\[
= c_{k+1}
\]
Therefore \( c_k \) satisfies the first condition. Second,
\[
\sum_{k=0}^{\infty} c_k
= \sum_{k=0}^{\infty} \min_{0 \leq j \leq k} \left( \sum_{l=0}^{L} \left( \| W_{j}^{k+1} - W_{j}^{k} \|_2^2 + \| W_{j}^{k+1} - W_{j}^{k} \|_2^2 + \| \bar{b}_{j}^{k+1} - b_{j}^{k} \|_2^2 + \| \bar{b}_{j}^{k+1} - b_{j}^{k} \|_2^2 \right) + \sum_{l=0}^{L-1} \left( \| p_{j}^{k+1} - a_{j}^{k} \|_2^2 + \| a_{j}^{k+1} - a_{j}^{k} \|_2^2 \right) \right)
+ \sum_{l=0}^{L} \left( \| W_{L}^{k+1} - W_{L}^{k} \|_2^2 + \| W_{L}^{k+1} - W_{L}^{k} \|_2^2 + \| \bar{b}_{L}^{k+1} - b_{L}^{k} \|_2^2 + \| \bar{b}_{L}^{k+1} - b_{L}^{k} \|_2^2 \right) + \sum_{l=0}^{L-1} \left( \| p_{L}^{k+1} - a_{L}^{k} \|_2^2 + \| a_{L}^{k+1} - a_{L}^{k} \|_2^2 \right)
\leq (L \rho (\mathbf{W}, \mathbf{b}, \mathbf{z}, \mathbf{a}, \mathbf{u}) - L \rho (\mathbf{W}', \mathbf{b}', \mathbf{z}', \mathbf{a}', \mathbf{u}))/C_3 \text{(Property 2)}
\]
So \( \sum_{k=0}^{\infty} c_k \) is bounded and \( c_k \) satisfies the second condition. Finally, it has been proved that the sufficient conditions of convergence rate \( o(1/k) \) are: (1) \( c_k \geq c_{k+1} \), and (2) \( \sum_{k=0}^{\infty} c_k \) is bounded, and (3) \( c_k \geq 0 \) (Lemma 1.2 in [6]). Since we have proved the first two conditions and the third one \( c_k \geq 0 \) is obvious, the convergence rate of \( o(1/k) \) is proven. \( \square \)