Accelerated Gradient-free Neural Network Training by Multi-convex Alternating Optimization

Junxiang Wang [1] **Hongyi Li** [2] **Yongchao Wang** [2] **Liang Zhao** [1]

Abstract

In recent years, even though Stochastic Gradient Descent (SGD) and its variants are well-known for training neural networks, it suffers from limitations such as the lack of theoretical guarantees, vanishing gradients, and excessive sensitivity to input. To overcome these drawbacks, alternating minimization methods have attracted fast-increasing attention recently. As an emerging and open domain, however, several new challenges need to be addressed, including 1) Convergence properties are sensitive to penalty parameters, and 2) Slow theoretical convergence rate. We, therefore, propose a novel Deep Learning Alternating Minimization (DLAM) algorithm, and a monotonous DLAM (mDLAM) algorithm to deal with these two challenges. Our innovative inequality-constrained formulation infinitely approximates the original problem with non-convex equality constraints, enabling our convergence proof of the DLAM algorithm and the mDLAM algorithm regardless of the choice of hyperparameters. Our mDLAM algorithm is shown to achieve a fast linear convergence by the Nesterov acceleration technique.

1. Introduction

Stochastic Gradient Descent (SGD) and its variants have become popular optimization methods for training deep neural networks. Many variants of SGD methods have been presented, including SGD momentum (Sutskever et al., 2013), AdaGrad (Duchi et al., 2011), RMSProp (Tieleman & Hinton, 2017), Adam (Kingma & Ba, 2015) and AMSGrad (Reddi et al., 2018).While many researchers have provided solid theoretical guarantees on the convergence of SGD (Kingma & Ba, 2015; Reddi et al., 2018; Sutskever et al., 2013), the assumptions of their proofs cannot be applied to problems involving deep neural networks, which are highly nonsmooth and nonconvex. Aside from the lack of theoretical guarantees, several additional drawbacks restrict the applications of SGD. It suffers from the gradient vanishing problem, meaning that the error signal diminishes as the gradient is backpropagated, which prevents the neural networks from utilizing further training (Taylor et al., 2016), and the gradient of the activation function is highly sensitive to the input (i.e. poor conditioning), so a small change in the input can lead to a dramatic change in the gradient.

To tackle these intrinsic drawbacks of gradient descent optimization methods, alternating minimization methods have started to attract attention as a potential way to solve deep learning problems. A neural network problem is reformulated as a nested function associated with multiple linear and nonlinear transformations across multi-layers. This nested structure is then decomposed into a series of linear and nonlinear equality constraints by introducing auxiliary variables and penalty hyperparameters. The linear and nonlinear equality constraints generate multiple subproblems, which can be minimized alternately. Some recent alternating minimization methods have focused on applying the Alternating Direction Method of Multipliers (ADMM) (Taylor et al., 2016; Wang et al., 2019), Block Coordinate Descent (BCD) (Zeng et al., 2019) and auxiliary coordinates (MAC) (Carreira-Perpinan & Wang, 2014) to replace a nested neural network with a constrained problem without nesting, with empirical evaluations demonstrating good scalability in terms of the number of layers and high accuracy on the test sets (Taylor et al., 2016; Wang et al., 2019). These methods also avoid gradient vanishing problems and allow for non-differentiable activation functions such as binarized neural networks (Courbariaux et al., 2015), as well as allowing for complex non-smooth regularization and the constraints that are increasingly important for deep neural architectures that are required to satisfy practical requirements such as interpretability, energy-efficiency, and cost awareness (Carreira-Perpinan & Wang, 2014).

However, as an emerging domain, alternating mini-

[1]Department of Computer Science and Informatics, Emory University, Georgia, United States. [2]The State Key Laboratory of Integrated Service Networks, Xidian University, Shaanxi, China. Correspondence to: Junxiang Wang <jwan936@emory.edu>.

Workshop on "Beyond first-order methods in ML systems" of the 38^{th} *International Conference on Machine Learning*.

Table 1. Notations used in this paper

Notations	Descriptions
L	Number of layers.
W_l	The weight vector in the l-th layer.
z_l	The output of the linear mapping in the l-th layer.
$h_l(z_l)$	The nonlinear activation function in the l-th layer.
a_l	The output of the l-th layer.
x	The input matrix of the neural network.
y	The predefined label vector.
$R(z_L; y)$	The loss function in the L-th layer.
$\Omega_l(W_l)$	The regularization term in the l-th layer.
ε	The tolerance of the nonlinear mapping.

mization for deep model optimization suffers from several unsolved challenges including: **1. Convergence properties are sensitive to penalty parameters.** One recent work by Wang et al. firstly proved the convergence guarantee of ADMM in the fully-connected neural network problem (Wang et al., 2019). However, such convergence guarantee is dependent on the choice of penalty hyperparameters: the convergence can not be guaranteed any more when penalty hyperparameters are small; **2. Slow convergence rate.** To the best of our knowledge, almost all existing alternating minimization methods can only achieve a sublinear convergence rate. For example, The convergence rate of the ADMM and the BCD is proven to be $O(1/k)$, where k is the number of iteration (Wang et al., 2019; Zeng et al., 2019). Therefore, there still lacks a theoretical framework that can achieve a faster convergence rate.

To simultaneously address these technical problems, we propose a new formulation of the neural network problem, along with a novel Deep Learning Alternating Minimization (DLAM) algorithm and a monotonous DLAM (mDLAM) algorithm. Specifically, we, for the first time, transform the original neural network optimization problem into an inequality-constrained problem that can infinitely approximate to the original one. Applying this innovation to an inequality-constraint based transformation not only ensures the convexity of all subproblems, and hence easily ensures global minima, but also prevents the output of a nonlinear function from changing much and reduces sensitivity to the input. Moreover, our proposed mDLAM algorithm can achieve a linear convergence rate theoretically, and the choice of hyperparameters does not affect the convergence of our DLAM algorithm and the mDLAM algorithm theoretically.

2. Model and Algorithms

2.1. Inequality Approximation for Deep Learning

Important notations used in this paper are shown in Table 1. A typical fully-connected neural network consists of L layers, each of which are defined by a linear mapping and a nonlinear activation function. A linear mapping is composed of a weight vector $W_l \in \mathbb{R}^{n_l \times n_{l-1}}$, where n_l is the number of neurons on the l-th layer; a nonlinear mapping is defined by a continuous activation function $h_l(\bullet)$. Given an input $a_{l-1} \in \mathbb{R}^{n_{l-1}}$ from the $(l-1)$-th layer, the l-th layer outputs $a_l = h_l(W_l a_{l-1})$. By introducing an auxiliary variable z_l as the output of the linear mapping, the neural network problem is formulated mathematically as follows:

Problem 1.

$$
\begin{aligned}
&\min\nolimits_{a_l, W_l, z_l} R(z_L; y) + \sum\nolimits_{l=1}^{L} \Omega_l(W_l) \\
&s.t.\ z_l = W_l a_{l-1} (l = 1, \cdots, L), \\
&\quad a_l = h_l(z_l) (l = 1, \cdots, L-1)
\end{aligned}
$$

where $a_0 = x \in \mathbb{R}^d$ is the input of the neural network, d is the number of feature dimensions, and y is a predefined label vector. $R(z_L; y) \geq 0$ is a continuous loss function for the L-th layer, which is convex and proper, and $\Omega_l(W_l) \geq 0$ is a regularization term on the l-th layer, which is also continuous, convex and proper.

The equality constraint $a_l = h_l(z_l)$ is the most challenging one to handle here, because common activation functions such as sigmoid are nonlinear. This makes them nonconvex constraints and hence it is difficult to obtain the optimal solution when solving the z_l-subproblem (Taylor et al., 2016). To deal with this challenge, the following assumption is required for problem transformation:

Assumption 1. $h_l(z_l)(l = 1, \ldots, n)$ *are quasilinear.*

The quaslinearity is defined in the appendix. Assumption 1 is so mild that most of the widely used nonlinear activation functions satisfy it, including tanh (Zamanlooy & Mirhassani, 2014), smooth sigmoid (Glorot & Bengio, 2010), and the rectified linear unit (ReLU) (Maas et al., 2013). Then we innovatively transform the original nonconvex constraints into convex inequality constraints, which can infinitely approximate to Problem 1. To do this, we introduce a tolerance $\varepsilon > 0$ and reformulate Problem 1 to reach the following form:

$$
\begin{aligned}
&\min\nolimits_{W_l, z_l, a_l} R(z_L; y) + \sum\nolimits_{l=1}^{L} \Omega_l(W_l) \\
&\quad + \sum\nolimits_{l=1}^{L-1} \mathbb{I}(h_l(z_l) - \varepsilon \leq a_l \leq h_l(z_l) + \varepsilon) \\
&\quad s.t. z_l = W_l a_{l-1} (l = 1, \cdots, L)
\end{aligned}
$$

$\mathbb{I}(h_l(z_l) - \varepsilon \leq a_l \leq h_l(z_l) + \varepsilon)$ is an indicator function such that the value is 0 if $h_l(z_l) - \varepsilon \leq a_l \leq h_l(z_l) + \varepsilon$ and ∞ otherwise. For the linear constraint $z_l = W_l a_{l-1}$, this can be transformed into a penalty term in the objective function to minimize the difference between z_l and $W_l a_{l-1}$. The formulation is shown as follows:

Problem 2.

$$\begin{aligned}
&\min_{W_l, z_l, a_l} F(\boldsymbol{W}, \boldsymbol{z}, \boldsymbol{a}) \\
&= R(z_L; y) + \sum\nolimits_{l=1}^{L} \Omega_l(W_l) + \sum\nolimits_{l=1}^{L} \phi(a_{l-1}, W_l, z_l) \\
&+ \sum\nolimits_{l=1}^{L-1} \mathbb{I}(h_l(z_l) - \varepsilon \leq a_l \leq h_l(z_l) + \varepsilon)
\end{aligned}$$

The penalty term is defined as $\phi(a_{l-1}, W_l, z_l) = \frac{\rho}{2}\|z_l - W_l a_{l-1}\|_2^2$, where $\rho > 0$ a penalty parameter. $\mathbf{W} = \{W_l\}_{l=1}^{L}$, $\mathbf{z} = \{z_l\}_{l=1}^{L}$,$\mathbf{a} = \{a_l\}_{l=1}^{L-1}$.

The introduction of ε is to project the nonconvex constraints to convex ε-balls, thus transforming the nonconvex Problem 1 into the multi-convex Problem 2, which is much easier to solve by alternating minimization (Xu & Yin, 2013). For example, Problem 2 is convex with regard to $\mathbf{z}$ when $\mathbf{W}$, and $\mathbf{a}$ are fixed. As $\rho \to \infty$ and $\varepsilon \to 0$, Problem 2 approaches Problem 1.

Algorithm 1 The DLAM Algorithm

Require: $y, a_0 = x$.
Ensure: $a_l, W_l, z_l (l = 1, \cdots, L)$.
1: Initialize ρ, $k = 0$. $s^0 = 0$.
2: **repeat**
3: $\quad s^{k+1} \leftarrow \frac{1+\sqrt{1+4(s^k)^2}}{2}$
4: $\quad \omega^k \leftarrow \frac{s^k - 1}{s^{k+1}}$
5: $\quad$ **for** $l = 1$ to L **do**
6: $\quad\quad \overline{W}_l^{k+1} \leftarrow W_l^k + (W_l^k - W_l^{k-1})\omega^k$ and update W_l^{k+1} in Equation (3).
7: $\quad\quad \overline{z}_l^{k+1} \leftarrow z_l^k + (z_l^k - z_l^{k-1})\omega^k$
8: $\quad\quad$ **if** $l = L$ **then**
9: $\quad\quad\quad$ Update z_L^{k+1} in Equation (5).
10: $\quad\quad$ **else**
11: $\quad\quad\quad$ Update z_l^{k+1} in Equation (4).
12: $\quad\quad\quad \overline{a}_l^{k+1} \leftarrow a_l^k + (a_l^k - a_l^{k-1})\omega^k$ and update a_l^{k+1} in Equation (6).
13: $\quad\quad$ **end if**
14: $\quad$ **end for**
15: $\quad k \leftarrow k + 1$.
16: **until** convergence.
17: Output a_l, W_l, z_l.

2.2. Alternating Optimization

We present the DLAM algorithm and the mDLAM algorithm to solve Problem 2, shown in Algorithm 1 and Algorithm 2, respectively. To simplify the notation, $\mathbf{W}_{\leq l}^{k+1} = \{\{W_i^{k+1}\}_{i=1}^{l}, \{W_i^k\}_{i=l+1}^{L}\}$, $\mathbf{z}_{\leq l}^{k+1} = \{\{z_i^{k+1}\}_{i=1}^{l}, \{z_i^k\}_{i=l+1}^{L}\}$ and $\mathbf{a}_{\leq l}^{k+1} = \{\{a_i^{k+1}\}_{i=1}^{l}, \{a_i^k\}_{i=l+1}^{L-1}\}$. In Algorithm 1, Lines 6, 7, and 12 apply the Nestrov acceleration technique and update W_l, z_l and a_l, respectively. The difference between the DLAM algorithm and the mDLAM algorithm is that the mDLAM algorithm guarantees the decrease of objective F: for example, if the updated W_l^{k+1} in Line 7 of Algorithm 2 increases the value of F, i.e. $F(\mathbf{W}_{\leq l}^{k+1}, \mathbf{z}_{\leq l-1}^{k+1}, \mathbf{a}_{\leq l-1}^{k+1}) \geq F(\mathbf{W}_{\leq l-1}^{k+1}, \mathbf{z}_{\leq l-1}^{k+1}, \mathbf{a}_{\leq l-1}^{k+1})$, then W_l^{k+1} is updated again by setting $\overline{W}_l^{k+1} = W_l^k$ in Line 8 of Algorithm 2, which ensures the decline of F.

Algorithm 2 The mDLAM Algorithm

Require: $y, a_0 = x$.
Ensure: $a_l, W_l, z_l (l = 1, \cdots, L)$.
1: Initialize ρ, $k = 0$. $s^0 = 0$.
2: **repeat**
3: $\quad s^{k+1} \leftarrow \frac{1+\sqrt{1+4(s^k)^2}}{2}$
4: $\quad \omega^k \leftarrow \frac{s^k - 1}{s^{k+1}}$
5: $\quad$ **for** $l = 1$ to L **do**
6: $\quad\quad \overline{W}_l^{k+1} \leftarrow W_l^k + (W_l^k - W_l^{k-1})\omega^k$ and update W_l^{k+1} in Equation (3).
7: $\quad\quad$ **if** W_l^{k+1} increases the objective F **then**
8: $\quad\quad\quad \overline{W}_l^{k+1} \leftarrow W_l^k$ and update W_l^{k+1} in Equation (3).
9: $\quad\quad$ **end if**
10: $\quad\quad \overline{z}_l^{k+1} \leftarrow z_l^k + (z_l^k - z_l^{k-1})\omega^k$
11: $\quad\quad$ **if** $l = L$ **then**
12: $\quad\quad\quad$ Update z_L^{k+1} in Equation (5).
13: $\quad\quad\quad$ **if** z_L^{k+1} increases the objective F **then**
14: $\quad\quad\quad\quad \overline{z}_L^{k+1} \leftarrow z_L^k$ and update z_L^{k+1} in Equation (5).
15: $\quad\quad\quad$ **end if**
16: $\quad\quad$ **else**
17: $\quad\quad\quad$ Update z_l^{k+1} in Equation (4).
18: $\quad\quad\quad$ **if** z_l^{k+1} increases the objective F **then**
19: $\quad\quad\quad\quad \overline{z}_l^{k+1} \leftarrow z_l^k$ and update z_l^{k+1} in Equation (4).
20: $\quad\quad\quad$ **end if**
21: $\quad\quad\quad \overline{a}_l^{k+1} \leftarrow a_l^k + (a_l^k - a_l^{k-1})\omega^k$ and update a_l^{k+1} in Equation (6).
22: $\quad\quad\quad$ **if** a_l^{k+1} increases the objective F **then**
23: $\quad\quad\quad\quad \overline{a}_l^{k+1} \leftarrow a_l^k$ and update a_l^{k+1} in Equation (6).
24: $\quad\quad\quad$ **end if**
25: $\quad\quad$ **end if**
26: $\quad$ **end for**
27: $\quad k \leftarrow k + 1$.
28: **until** convergence.
29: Output a_l, W_l, z_l.

The same procedure is applied in Lines 13-15, Lines 18-20, and Lines 22-24 in Algorithm 2, respectively.

Next, all subproblems are shown as follows:

1. Update W_l

The variables $W_l (l = 1, \cdots, L)$ are updated as follows:

$$W_l^{k+1} \leftarrow \arg\min_{W_l} \phi(a_{l-1}^{k+1}, W_l, z_l^k) + \Omega_l(W_l) \qquad (1)$$

Because W_l and a_{l-1} are coupled in $\phi(\bullet)$, solving W_l requires an inversion operation of a_{l-1}^{k+1}, which is computationally expensive. Motivated by the dlADMM algorithm (Wang et al., 2019), we define $P_l^{k+1}(W_l; \theta_l^{k+1})$ as a quadratic approximation of ϕ at W_l^k as follows:

$$\begin{aligned}
&P_l^{k+1}(W_l; \theta_l^{k+1}) = \phi(a_{l-1}^{k+1}, \overline{W}_l^{k+1}, z_l^k) \\
&+ (\nabla_{\overline{W}_l^{k+1}} \phi)^T (W_l - \overline{W}_l^{k+1}) + \frac{\theta_l^{k+1}}{2} \|W_l - \overline{W}_l^{k+1}\|_2^2
\end{aligned}$$

where $\theta_l^{k+1} > 0$ is a scalar parameter, which can be chosen by the backtracking algorithm (Wang et al., 2019) to meet the following condition

$$P_l^{k+1}(W_l^{k+1}; \theta_l^{k+1}) \geq \phi(a_{l-1}^{k+1}, W_l^{k+1}, z_l^k) \qquad (2)$$

Rather than minimizing Equation (1), we instead minimize the following:

$$W_l^{k+1} \leftarrow \arg\min_{W_l} P_l^{k+1}(W_l; \theta_l^{k+1}) + \Omega_l(W_l) \qquad (3)$$

For $\Omega_l(W_l)$, common regularization terms like ℓ_1 or ℓ_2 regularizations lead to closed-form solutions.

2. Update z_l

The variables $z_l(l=1,\cdots,L)$ are updated as follows:

$$z_l^{k+1} \leftarrow \arg\min_{z_l} \phi(a_{l-1}^{k+1}, W_l^{k+1}, z_l) + \mathbb{I}(h_l(z_l)-\varepsilon \le a_l^k \le h_l(z_l)+\varepsilon)(l<L)$$
$$z_L^{k+1} \leftarrow \arg\min_{z_L} \phi(a_{L-1}^{k+1}, W_L^{k+1}, z_L) + R(z_L;y)$$

Similar to updating W_l, we define $V_l^{k+1}(z_l)$ as follows:

$$V_l^{k+1}(z_l) = \phi(a_{l-1}^{k+1}, W_l^{k+1}, \overline{z}_l^{k+1}) + (\nabla_{\overline{z}_l^{k+1}}\phi)^T(z_l-\overline{z}_l^{k+1}) + \frac{\rho}{2}\|z_l-\overline{z}_l^{k+1}\|_2^2$$

Hence, we solve the following problems:

$$z_l^{k+1} \leftarrow \arg\min_{z_l} V_l^{k+1}(z_l) + \mathbb{I}(h_l(z_l)-\varepsilon \le a_l^k \le h_l(z_l)+\varepsilon)(l<L) \tag{4}$$
$$z_L^{k+1} \leftarrow \arg\min_{z_L} V_L^{k+1}(z_L) + R(z_L;y) \tag{5}$$

As for $z_l(l=1,\cdots,l-1)$, the solution is

$$z_l^{k+1} \leftarrow \min(\max(B_1^{k+1}, \overline{z}_l^{k+1} - \nabla\phi_{\overline{z}_l^{k+1}}/\rho), B_2^{k+1}).$$

where B_1^{k+1} and B_2^{k+1} represent the lower bound and the upper bound of the set $\{z_l|h_l(z_l)-\varepsilon \le a_l^k \le h_l(z_l)+\varepsilon\}$. Equation (5) is easy to solve using the Fast Iterative Soft Thresholding Algorithm (FISTA) (Beck & Teboulle, 2009).

3. Update a_l

The variables $a_l(l=1,\cdots,L-1)$ are updated as follows:

$$a_l^{k+1} \leftarrow \arg\min_{a_l} \phi(a_l, W_{l+1}^k, z_{l+1}^k) + \mathbb{I}(h_l(z_l^{k+1})-\varepsilon \le a_l \le h_l(z_l^{k+1})+\varepsilon)$$

Similar to updating W_l^{k+1}, $Q_l^{k+1}(a_l;\tau_l^{k+1})$ is defined as

$$Q_l^{k+1}(a_l;\tau_l^{k+1}) = \phi(\overline{a}_l^{k+1}, W_{l+1}^k, z_{l+1}^k) + (\nabla_{\overline{a}_l^{k+1}}\phi)^T(a_l-\overline{a}_l^{k+1}) + \frac{\tau_l^{k+1}}{2}\|a_l-\overline{a}_l^{k+1}\|_2^2$$

and this allows us to solve the following problem instead:

$$a_l^{k+1} \leftarrow \arg\min_{a_l} Q_l^{k+1}(a_l;\tau_l^{k+1}) + \mathbb{I}(h_l(z_l^{k+1})-\varepsilon \le a_l \le h_l(z_l^{k+1})+\varepsilon) \tag{6}$$

where $\tau_l^{k+1} > 0$ is a scalar parameter, which can be chosen by the backtracking algorithm (Wang et al., 2019) to meet the following condition:

$$Q_l^{k+1}(a_l^{k+1};\tau_l^{k+1}) \ge \phi(a_l^{k+1}, W_{l+1}^k, z_{l+1}^k)$$

The solution can be obtained by

$$a_l^{k+1} \leftarrow \min(\max(h_l(z_l^{k+1})-\varepsilon, \overline{a}_l^{k+1} - \nabla_{\overline{a}_l^{k+1}}\phi/\tau_l^{k+1}), h_l(z_l^{k+1})+\varepsilon)$$

3. Convergence Analysis

In this section, the convergence of two proposed algorithm is analyzed. Due to space limit, all proofs are detailed in the appendix. The following mild assumption is required for the convergence analysis of the proposed DLAM algorithm and the mDLAM algorithm:

Assumption 2. *$F(\boldsymbol{W},\boldsymbol{z},\boldsymbol{a})$ is coercive over the domain* $\{(\boldsymbol{W},\boldsymbol{z},\boldsymbol{a})|h_l(z_l)-\varepsilon \le a_l \le h_l(z_l)+\varepsilon\ (l=1,\cdots,L-1)\}$.

The coercivity is defined in the Appendix. Common loss functions such as the least square loss and the cross-entropy loss are coercive (Wang et al., 2019).

3.1. Convergence of the DLAM algorithm

We now summarize the convergence of the DLAM algorithm using two theorems, which ensure that the DLAM algorithm converges to a stationary point sublinearly, no matter what ρ and ε are chosen.

Theorem 1 (Convergence to a Stationary Point). *If $(\omega^k)^2 < \min(\frac{\theta_l^k}{\theta_l^{k+1}}, \frac{\tau_l^k}{\tau_l^{k+1}})(l=1,\cdots,L-1)$ and $(\omega^k)^2 < \frac{\theta_L^k}{\theta_L^{k+1}}$ in Algorithm 1, for $\boldsymbol{W}$ in Problem 2, starting from any $\boldsymbol{W}^0$, any limit point $\boldsymbol{W}^*$ is a stationary point of Problem 2. That is, $0 \in \partial_{\boldsymbol{W}^*}F$.*

Theorem 2 (Sublinear Convergence Rate). *In Algorithm 1, for a sequence $(\boldsymbol{W}^k,\boldsymbol{z}^k,\boldsymbol{a}^k)$, define $c_k = \min_{1\le i\le k}(\sum_{l=1}^L((\frac{\theta_l^i}{2}-\frac{\theta_l^{i+1}}{2}(\omega^i)^2)\|W_l^i-W_l^{i-1}\|_2^2+\frac{\rho}{2}(1-(\omega^i)^2)\|z_l^i-z_l^{i-1}\|_2^2)+\sum_{l=1}^{L-1}(\frac{\tau_l^i}{2}-\frac{\tau_l^{i+1}}{2}(\omega^i)^2)\|a_l^i-a_l^{i-1}\|_2^2)$, then the convergence rate of c_k is $o(\frac{1}{k})$.*

3.2. Convergence of the mDLAM algorithm

Next we discuss the convergence of the mDLAM algorithm. The following lemma guarantees the objective decrease, which is not satisfied for the DLAM algorithm.

Lemma 1 (Objective Decrease). *In Algorithm 2, it holds that for any $k \in \mathbb{N}$, $F(\boldsymbol{W}^k,\boldsymbol{z}^k,\boldsymbol{a}^k) \ge F(\boldsymbol{W}^{k+1},\boldsymbol{z}^{k+1},\boldsymbol{a}^{k+1})$. Moreover, F is convergent. That is, $F(\boldsymbol{W}^k,\boldsymbol{z}^k,\boldsymbol{a}^k) \to F^*$ as $k\to\infty$.*

The next two theorems guarantee that the mDLAM algorithm converges to a stationary point with a fast linear convergence rate.

Theorem 3 (Convergence to a Stationary Point). *In Algorithm 2, for $\boldsymbol{W}$ in Problem 2, starting from any $\boldsymbol{W}^0$, any limit point $\boldsymbol{W}^*$ is a stationary point of Problem 2. That is, $0 \in \partial_{\boldsymbol{W}^*}F$.*

Theorem 4 (Linear Convergence Rate). *In Algorithm 2, if F is locally strongly convex, then for any ρ, there exist $\varepsilon > 0$, $k_1 \in \mathbb{N}$ and $0 < C_1 < 1$ such that it holds for $k > k_1$*

that

$$F(\boldsymbol{W}^{k+1}, \boldsymbol{z}^{k+1}, \boldsymbol{a}^{k+1}) - F^* \\ \leq C_1(F(\boldsymbol{W}^{k-1}, \boldsymbol{z}^{k-1}, \boldsymbol{a}^{k-1}) - F^*)$$

where F^* is the convergent value of F.

References

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Carreira-Perpinan, M. and Wang, W. Distributed optimization of deeply nested systems. In *Artificial Intelligence and Statistics*, pp. 10–19, 2014.

Courbariaux, M., Bengio, Y., and David, J.-P. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems*, pp. 3123–3131, 2015.

Deng, W., Lai, M.-J., Peng, Z., and Yin, W. Parallel multi-block admm with o (1/k) convergence. *Journal of Scientific Computing*, 71(2):712–736, 2017.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul): 2121–2159, 2011.

Ghadimi, S. and Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*. Citeseer, 2013.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147, 2013.

Taylor, G., Burmeister, R., Xu, Z., Singh, B., Patel, A., and Goldstein, T. Training neural networks without gradients: A scalable admm approach. In *International Conference on Machine Learning*, pp. 2722–2731, 2016.

Tieleman, T. and Hinton, G. Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. Technical report, 2017.

Wang, J., Yu, F., Chen, X., and Zhao, L. Admm for efficient deep learning with global convergence. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 111–119, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330936. URL http://doi.acm.org/10.1145/3292500.3330936.

Wang, Y., Yin, W., and Zeng, J. Global convergence of admm in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, pp. 1–35, 2015.

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

Zamanlooy, B. and Mirhassani, M. Efficient vlsi implementation of neural networks with hyperbolic tangent activation function. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 22(1):39–48, 2014.

Zeng, J., Lau, T. T.-K., Lin, S., and Yao, Y. Global convergence of block coordinate descent in deep learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7313–7323, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/zeng19a.html.

Appendix

Several definitions are shown here for the propose of convergence analysis.

Definition 1 (Coercivity). *Any arbitrary function $G_2(x)$ is coercive over a nonempty set $dom(G_2)$ if as $\|x\| \to \infty$ and $x \in dom(G_2)$, we have $G_2(x) \to \infty$, where $dom(G_2)$ is a domain set of G_2.*

Definition 2 (Multi-convexity). *A function $f(x_1, x_2, \cdots, x_m)$ is a multi-convex function if f is convex with regard to $x_i (i = 1, \cdots, m)$ while fixing other variables.*

Definition 3 (Lipschitz Differentiability). *A function $f(x)$ is Lipschitz differentiable with Lipschitz coefficient $L > 0$ if for any $x_1, x_2 \in \mathbb{R}$, the following inequality holds:*

$$\|\nabla f(x_1) - \nabla f(x_2)\| \leq L\|x_1 - x_2\|$$

For Lipschitz differentiability, we have the following lemma (Lemma 2.1 in (Beck & Teboulle, 2009)):

Lemma 2. *If $f(x)$ is Lipschitz differentiable with $L > 0$, then for any $x_1, x_2 \in \mathbb{R}$*

$$f(x_1) \leq f(x_2) + \nabla f^T(x_2)(x_1 - x_2) + \frac{L}{2}\|x_1 - x_2\|^2$$

Definition 4 (Fréchet Subdifferential). *For each $x_1 \in dom(u_1)$, the Fréchet subdifferential of u_1 at x_1, which is denoted as $\hat{\partial} u_1(x_1)$, is the set of vectors v, which satisfy*

$$\lim_{x_2 \neq x_1} \inf_{x_2 \to x_1} (u_1(x_2) - u_1(x_1) - v^T(x_2 - x_1))/\|x_2 - x_1\| \geq 0.$$

The vector $v \in \hat{\partial} u_1(x_1)$ is a Fréchet subgradient.

Then the definition of the limiting subdifferential, which is based on Fréchet subdifferential, is given in the following (Rockafellar & Wets, 2009):

Definition 5 (Limiting Subdifferential). *For each $x \in dom(u_2)$, the limiting subdifferential (or subdifferential) of u_2 at x is*

$$\partial u_2(x) = \{v_1 | \exists\, x^k \to x, s.t.\ u_2(x^k) \to u_2(x), v^k \in \hat{\partial} u_2(x^k), v^k \to v\}$$

where x^k is a sequence whose limit is x and the limit of $u_2(x^k)$ is $u_2(x)$, v^k is a sequence, which is a Fréchet subgradient of u_2 at x^k and whose limit is v. The vector $v \in \partial u_2(x)$ is a limiting subgradient.

Specifically, when u_2 is convex, its limiting subdifferential is reduced to regular subdifferential (Rockafellar & Wets, 2009), which is defined as follows:

Definition 6 (Regular Subdifferential). *For each $x_1 \in dom(f)$, the regular subdifferential of a convex function f at x_1, which is denoted as $\partial f(x_1)$, is the set of vectors v, which satisfy*

$$f(x_2) \geq f(x_1) + v^T(x_2 - x_1)$$

The vector $v \in \partial f(x_1)$ is a regular subgradient.

Definition 7 (Quasilinearity). *A function $f(x)$ is quasiconvex if for any sublevel set $S_\nu(f) = \{x | f(x) \leq \nu\}$ is a convex set. Likewise, A function $f(x)$ is quasiconcave if for any superlevel set $S_\nu(f) = \{x | f(x) \geq \nu\}$ is a convex set. A function $f(x)$ is quasilinear if it is both quasiconvex and quasiconcave.*

Definition 8 (Locally Strong Convexity). *A function $f(x)$ is locally strongly convex within a bound set $\mathbb{D}$ with a constant μ if*

$$f(y) \geq f(x) + g^T(y - x) + \frac{\mu}{2}\|x - y\|_2^2\ \forall\, g \in \partial f(x) \text{ and } x, y \in \mathbb{D}$$

Simply speaking, a locally strongly convex function lies above a quadratic function within a bounded set.

Definition 9 (Kurdyka-Lojasiewicz (KL) Property). *A function $f(x)$ has the KL Property at $\overline{x} \in dom\, \partial f = \{x \in \mathbb{R} : \partial f(x) \neq \emptyset\}$ if there exists $\eta \in (0, +\infty]$, a neighborhood X of $\overline{x}$ and a function $\psi \in \Psi_\eta$, such that for all*

$$x \in X \cap \{x \in \mathbb{R} : f(\overline{x}) < f(x) < f(\overline{x}) + \eta\}$$

the following inequality holds

$$\psi^{'}(f(x) - f(\overline{x}))dist(0, \partial f(x)) \geq 1$$

where Ψ_η stands for a class of function $\psi : [0, \eta] \to \mathbb{R}^+$ satisfying: (1). ϕ is concave and $\psi^{'}(x)$ continuous on $(0, \eta)$; (2). ψ is continuous at 0, $\psi(0) = 0$; and (3). $\psi^{'}(x) > 0, \forall x \in (0, \eta)$.

The following lemma shows that a locally strongly convex function satisfies the KL Property:

Lemma 3 ((Xu & Yin, 2013)). *A locally strongly convex function $f(x)$ with a constant μ satisfies the KL Property at any $x \in \mathbb{D}$ with $\psi(x) = \frac{2}{\mu}\sqrt{x}$ and $X = \mathbb{D} \cap \{y : f(y) \geq f(x)\}$.*

Preliminary Results

In this section, we present preliminary lemmas of the DLAM algorithm and the mDLAM algorithm. The limiting subdifferential is used to prove the convergence of the DLAM algorithm and the mDLAM algorithm in the following convergence analysis. Without loss of generality, ∂R and $\partial \Omega_l (l = 1, \cdots, n)$ are assumed to be nonempty, and the limiting subdifferential of F defined in Problem 2 is (Xu & Yin, 2013):

$$\partial F(\mathbf{W}, \mathbf{z}, \mathbf{a}) = \partial_{\mathbf{W}} F \times \partial_{\mathbf{z}} F \times \partial_{\mathbf{a}} F$$

where $\times$ means the Cartesian product.

Lemma 4. *If Equation (3) holds, then there exists $p \in \partial \Omega_l(W_l^{k+1})$, the subgradient of $\Omega_l(W_l^{k+1})$ such that*

$$\nabla_{\overline{W}_l^{k+1}} \phi + \theta_l^{k+1}(W_l^{k+1} - \overline{W}_l^{k+1}) + p = 0$$

Likewise, if Equation (4) holds, then there exists $q \in \partial \mathbb{I}(h_l(z_l^{k+1}) - \varepsilon \leq a_l^k \leq h_l(z_l^{k+1}) + \varepsilon)$ such that

$$\nabla_{\overline{z}_l^{k+1}} \phi + \rho(z_l^{k+1} - \overline{z}_l^{k+1}) + q = 0$$

If Equation (5) holds, then there exists $u \in \partial R(z_L^{k+1}; y)$ such that

$$\nabla_{\overline{z}_L^{k+1}} \phi + \rho(z_L^{k+1} - \overline{z}_L^{k+1}) + u = 0$$

If Equation (6) holds, then there exists $v \in \partial \mathbb{I}(h_l(z_l^{k+1}) - \varepsilon \leq a_l^{k+1} \leq h_l(z_l^{k+1}) + \varepsilon)$ such that

$$\nabla_{\overline{a}_l^{k+1}} \phi + \tau_l^{k+1}(a_l^{k+1} - \overline{a}_l^{k+1}) + v = 0$$

Proof. These can be obtained by directly applying the optimality conditions of Equation (3), Equation (4), Equation (5) and Equation (6), respectively. ☐

Lemma 5. *For Equation (4) and Equation (5), the following inequalities hold:*

$$V_l^{k+1}(z_l^{k+1}) \geq \phi(a_{l-1}^{k+1}, W_l^{k+1}, z_l^{k+1}) \tag{7}$$

Proof. Because $\phi(a_{l-1}, W_l, z_l)$ is Lipschitz differentiable with respect to z_l with Lipschitz coefficient ρ, we directly apply Lemma 2 to ϕ to obtain Equation (7). ☐

Lemma 6. *In Algorithm 1, it holds for* $\forall k \in \mathbb{N}$, $W_l^k, z_l^k (l = 1, 2, \cdots, L)$, *and* $a_l^k (l = 1, 2, \cdots, L-1)$ *that*

$$F(\boldsymbol{W}_{\leq l-1}^{k+1}, \boldsymbol{z}_{\leq l-1}^{k+1}, \boldsymbol{a}_{\leq l-1}^{k+1}) - F(\boldsymbol{W}_{\leq l}^{k+1}, \boldsymbol{z}_{\leq l-1}^{k+1}, \boldsymbol{a}_{\leq l-1}^{k+1}) \geq \frac{\theta_l^{k+1}}{2}(\|W_l^{k+1} - W_l^k\|_2^2 - (\omega^k)^2\|W_l^k - W_l^{k-1}\|_2^2). \tag{8}$$

$$F(\boldsymbol{W}_{\leq l}^{k+1}, \boldsymbol{z}_{\leq l-1}^{k+1}, \boldsymbol{a}_{\leq l-1}^{k+1}) - F(\boldsymbol{W}_{\leq l}^{k+1}, \boldsymbol{z}_{\leq l}^{k+1}, \boldsymbol{a}_{\leq l-1}^{k+1}) \geq \frac{\rho}{2}(\|z_l^{k+1} - z_l^k\|_2^2 - (\omega^k)^2\|z_l^k - z_l^{k-1}\|_2^2). \tag{9}$$

$$F(\boldsymbol{W}_{\leq l}^{k+1}, \boldsymbol{z}_{\leq l}^{k+1}, \boldsymbol{a}_{\leq l-1}^{k+1}) - F(\boldsymbol{W}_{\leq l}^{k+1}, \boldsymbol{z}_{\leq l}^{k+1}, \boldsymbol{a}_{\leq l}^{k+1}) \geq \frac{\tau^{k+1}}{2}(\|a_l^{k+1} - a_l^k\|_2^2 - (\omega^k)^2\|a_l^k - a_l^{k-1}\|_2^2). \tag{10}$$

In Algorithm 2, there exist $\alpha_l^k, \gamma_l^k, \delta_l^k > 0$ *such that for* $\forall k \in \mathbb{N}$, $W_l^k, z_l^k (l = 1, 2, \cdots, L)$, *and* $a_l^k (l = 1, 2, \cdots, L-1)$ *it holds that*

$$F(\boldsymbol{W}_{\leq l-1}^{k+1}, \boldsymbol{z}_{\leq l-1}^{k+1}, \boldsymbol{a}_{\leq l-1}^{k+1}) - F(\boldsymbol{W}_{\leq l}^{k+1}, \boldsymbol{z}_{\leq l-1}^{k+1}, \boldsymbol{a}_{\leq l-1}^{k+1}) \geq \frac{\alpha_l^{k+1}}{2}\|W_l^{k+1} - W_l^k\|_2^2. \tag{11}$$

$$F(\boldsymbol{W}_{\leq l}^{k+1}, \boldsymbol{z}_{\leq l-1}^{k+1}, \boldsymbol{a}_{\leq l-1}^{k+1}) - F(\boldsymbol{W}_{\leq l}^{k+1}, \boldsymbol{z}_{\leq l}^{k+1}, \boldsymbol{a}_{\leq l-1}^{k+1}) \geq \frac{\gamma_l^{k+1}}{2}\|z_l^{k+1} - z_l^k\|_2^2. \tag{12}$$

$$F(\boldsymbol{W}_{\leq l}^{k+1}, \boldsymbol{z}_{\leq l}^{k+1}, \boldsymbol{a}_{\leq l-1}^{k+1}) - F(\boldsymbol{W}_{\leq l}^{k+1}, \boldsymbol{z}_{\leq l}^{k+1}, \boldsymbol{a}_{\leq l}^{k+1}) \geq \frac{\delta^{k+1}}{2}\|a_l^{k+1} - a_l^k\|_2^2. \tag{13}$$

Proof. In Algorithm 1, all inequalities can be obtained by applying optimality conditions of updating W_l^{k+1}, z_l^{k+1} and a_l^{k+1}, respectively. We only prove Equation (8) because Equation (9) and Equation (10) follow the same routine of Equation (8).
Because $\Omega_{W_l}(W_l)$ and $\phi(a_{l-1}, W_l, z_l)$ are convex with regard to W_l, according to the definition of regular subgradient, we have

$$\Omega_l(W_l^k) \geq \Omega_l(W_l^{k+1}) + p^T(W_l^k - W_l^{k+1}) \tag{14}$$

$$\phi(a_{l-1}^{k+1}, W_l^k, z_l^k) \geq \phi(a_{l-1}^{k+1}, \overline{W}_l^{k+1}, z_l^k) + \nabla_{\overline{W}_l^{k+1}}\phi^T(W_l^k - \overline{W}_l^{k+1}) \tag{15}$$

where p is defined in the premise of Lemma 4. Therefore, we have

$$\begin{aligned}
&F(\mathbf{W}_{\leq l-1}^{k+1}, \mathbf{z}_{\leq l-1}^{k+1}, \mathbf{a}_{\leq l-1}^{k+1}) - F(\mathbf{W}_{\leq l}^{k+1}, \mathbf{z}_{\leq l-1}^{k+1}, \mathbf{a}_{\leq l-1}^{k+1}) \\
&= \phi(a_{l-1}^{k+1}, W_l^k, z_l^k) + \Omega_l(W_l^k) - \phi(a_{l-1}^{k+1}, W_l^{k+1}, z_l^k) - \Omega_l(W_l^{k+1}) \text{ (Definition of } F \text{ in Problem 2)} \\
&\geq \Omega_l(W_l^k) - \Omega_l(W_l^{k+1}) - (\nabla_{\overline{W}_l^{k+1}}\phi)^T(W_l^{k+1} - \overline{W}_l^{k+1}) - \frac{\theta_l^{k+1}}{2}\|W_l^{k+1} - \overline{W}_l^{k+1}\|_2^2 - \phi(a_{l-1}^{k+1}, \overline{W}_l^{k+1}, z_l^k) \\
&+ \phi(a_{l-1}^{k+1}, W_l^k, z_l^k) \text{(Equation (2))} \\
&\geq p^T(W_l^k - W_l^{k+1}) - (\nabla_{\overline{W}_l^{k+1}}\phi)^T(W_l^{k+1} - W_l^k) - \frac{\theta_l^{k+1}}{2}\|W_l^{k+1} - \overline{W}_l^{k+1}\|_2^2 \text{ (Equation (14) and Equation (15))} \\
&= -(\nabla_{\overline{W}_l^{k+1}}\phi + \theta_l^{k+1}(W_l^{k+1} - \overline{W}_l^{k+1}))^T(W_l^k - W_l^{k+1}) - (\nabla_{\overline{W}_l^{k+1}}\phi)^T(W_l^{k+1} - W_l^k) - \frac{\theta_l^{k+1}}{2}\|W_l^{k+1} - \overline{W}_l^{k+1}\|_2^2 \text{(Lemma 4)} \\
&= \frac{\theta_l^{k+1}}{2}\|W_l^{k+1} - \overline{W}_l^{k+1}\|_2^2 + \theta_l^{k+1}(W_l^{k+1} - \overline{W}_l^{k+1})^T(\overline{W}_l^{k+1} - W_l^k) \\
&= \frac{\theta_l^{k+1}}{2}(\|W_l^{k+1} - W_l^k\|_2^2 - \|\overline{W}_l^{k+1} - W_l^k\|_2^2) \\
&= \frac{\theta_l^{k+1}}{2}(\|W_l^{k+1} - W_l^k\|_2^2 - (\omega^k)^2\|W_l^k - W_l^{k-1}\|_2^2) \text{ (Nesterov Acceleration)}
\end{aligned} \tag{16}$$

In Algorithm 2, we only show Equation (11) because Equation (12) and Equation (13) follow the same routine of Equation (11).

In Line 7 of Algorithm 2, if $F(\mathbf{W}_{\leq l}^{k+1}, \mathbf{z}_{\leq l-1}^{k+1}, \mathbf{a}_{\leq l-1}^{k+1}) < F(\mathbf{W}_{\leq l-1}^{k+1}, \mathbf{z}_{\leq l-1}^{k+1}, \mathbf{a}_{\leq l-1}^{k+1})$, then obviously there exists $\alpha_l^{k+1} >$

0 such that Equation (11) holds. Otherwise, according to Line 8 of Algorithm 2, we have

$$
\begin{aligned}
&F(\mathbf{W}^{k+1}_{\leq l-1},\mathbf{z}^{k+1}_{\leq l-1},\mathbf{a}^{k+1}_{\leq l-1})-F(\mathbf{W}^{k+1}_{\leq l},\mathbf{z}^{k+1}_{\leq l-1},\mathbf{a}^{k+1}_{\leq l-1})\\
&\geq \frac{\theta^{k+1}_l}{2}(\|W^{k+1}_l-W^k_l\|^2_2-\|\overline{W}^{k+1}_l-W^k_l\|^2_2)\ (\text{Equation } (16))\\
&=\frac{\theta^{k+1}_l}{2}\|W^{k+1}_l-W^k_l\|^2_2\ (\overline{W}^{k+1}_l=W^k_l)
\end{aligned}
$$

Let $\alpha^{k+1}_l=\theta^{k+1}_l$, then Equation (11) still holds. □

Lemma 7 (Convergent Sequence)**.** *In Algorithm 1, it holds that*
(a). If $(\omega^k)^2<\min(\frac{\theta^k_l}{\theta^{k+1}_l},\frac{\tau^k_l}{\tau^{k+1}_l})(l=1,\cdots,L-1)$ and $(\omega^k)^2<\frac{\theta^k_L}{\theta^{k+1}_L}$, then $\boldsymbol{F}(\boldsymbol{W}^k,\boldsymbol{z}^k,\boldsymbol{a}^k)$ is upper bounded. Moreover, $\lim_{k\to\infty}\boldsymbol{W}^{k+1}-\boldsymbol{W}^k=0$, $\lim_{k\to\infty}\boldsymbol{z}^{k+1}-\boldsymbol{z}^k=0$, and $\lim_{k\to\infty}\boldsymbol{a}^{k+1}-\boldsymbol{a}^k=0$.
(b). $(\boldsymbol{W}^k,\boldsymbol{z}^k,\boldsymbol{a}^k)$ is bounded. That is, there exist scalars $M_{\boldsymbol{W}}$, $M_{\boldsymbol{z}}$ and $M_{\boldsymbol{a}}$ such that $\|\boldsymbol{W}^k\|\leq M_{\boldsymbol{W}}$, $\|\boldsymbol{z}^k\|\leq M_{\boldsymbol{z}}$ and $\|\boldsymbol{a}^k\|\leq M_{\boldsymbol{a}}$.
In Algorithm 2, it holds that
(c). $\boldsymbol{F}(\boldsymbol{W}^k,\boldsymbol{z}^k,\boldsymbol{a}^k)$ is upper bounded. Moreover, $\lim_{k\to\infty}\boldsymbol{W}^{k+1}-\boldsymbol{W}^k=0$, $\lim_{k\to\infty}\boldsymbol{z}^{k+1}-\boldsymbol{z}^k=0$, and $\lim_{k\to\infty}\boldsymbol{a}^{k+1}-\boldsymbol{a}^k=0$.
(d). $(\boldsymbol{W}^k,\boldsymbol{z}^k,\boldsymbol{a}^k)$ is bounded. That is, there exist scalars $M_{\boldsymbol{W}}$,$M_{\boldsymbol{z}}$ and $M_{\boldsymbol{a}}$ such that $\|\boldsymbol{W}^k\|\leq M_{\boldsymbol{W}}$, $\|\boldsymbol{z}^k\|\leq M_{\boldsymbol{z}}$ and $\|\boldsymbol{a}^k\|\leq M_{\boldsymbol{a}}$.

Proof. In Algorithm 1:
(a). We sum Equation (8), Equation (9) and Equation (10) from $l=1$ to L and from $k=0$ to K to obtain

$$
\begin{aligned}
&F(\mathbf{W}^0,\mathbf{z}^0,\mathbf{a}^0)-F(\mathbf{W}^K,\mathbf{z}^K,\mathbf{a}^K)\\
&\geq \sum\nolimits_{k=0}^K\sum\nolimits_{l=1}^L\frac{\theta^{k+1}_l}{2}(\|W^{k+1}_l-W^k_l\|^2_2-(\omega^k)^2\|W^k_l-W^{k-1}_l\|^2_2)+\sum\nolimits_{k=0}^K\sum\nolimits_{l=1}^L\frac{\rho}{2}(\|z^{k+1}_l-z^k_l\|^2_2-(\omega^k)^2\|z^k_l-z^{k-1}_l\|^2_2)\\
&+\sum\nolimits_{k=0}^K\sum\nolimits_{l=1}^{L-1}\frac{\tau^{k+1}_l}{2}(\|a^{k+1}_l-a^k_l\|^2_2-(\omega^k)^2\|a^k_l-a^{k-1}_l\|^2_2)\\
&\geq \sum\nolimits_{k=1}^{K+1}\sum\nolimits_{l=1}^L(\frac{\theta^k_l}{2}-\frac{\theta^{k+1}_l}{2}(\omega^k)^2)\|W^k_l-W^{k-1}_l\|^2_2+\sum\nolimits_{k=1}^{K+1}\sum\nolimits_{l=1}^L\frac{\rho}{2}(1-(\omega^k)^2)\|z^k_l-z^{k-1}_l\|^2_2\\
&+\sum\nolimits_{k=1}^{K+1}\sum\nolimits_{l=1}^{L-1}(\frac{\tau^k_l}{2}-\frac{\tau^{k+1}_l}{2}(\omega^k)^2)\|a^k_l-a^{k-1}_l\|^2_2\ (\omega^0=0)
\end{aligned}
\tag{17}
$$

On one hand, it is easy to verify that $0<\omega^k<1$, so $1-(\omega^k)^2>0$. On the other hand, Because $(\omega^k)^2<\min(\frac{\theta^k_l}{\theta^{k+1}_l},\frac{\tau^k_l}{\tau^{k+1}_l})(l=1,\cdots,L-1)$ and $(\omega^k)^2<\frac{\theta^k_L}{\theta^{k+1}_L}$, so $\frac{\theta^k_l}{2}-\frac{\theta^{k+1}_l}{2}(\omega^k)^2>0(l=1,\cdots,L)$ and $\frac{\tau^k_l}{2}-\frac{\tau^{k+1}_l}{2}(\omega^k)^2>0(l=1,\cdots,L-1)$. So $F(\mathbf{W}^K,\mathbf{z}^K,\mathbf{a}^K)\leq F(\mathbf{W}^0,\mathbf{z}^0,\mathbf{a}^0)$. This proves the upper boundness of F. Let $K\to\infty$ in Equation (17), since $F>0$ is lower bounded, so

$$
\begin{aligned}
&\sum\nolimits_{k=1}^{\infty}\sum\nolimits_{l=1}^L(\frac{\theta^k_l}{2}-\frac{\theta^{k+1}_l}{2}(\omega^k)^2)\|W^k_l-W^{k-1}_l\|^2_2+\sum\nolimits_{k=1}^{\infty}\sum\nolimits_{l=1}^L\frac{\rho}{2}(1-(\omega^k)^2)\|z^k_l-z^{k-1}_l\|^2_2\\
&+\sum\nolimits_{k=1}^{\infty}\sum\nolimits_{l=1}^{L-1}(\frac{\tau^k_l}{2}-\frac{\tau^{k+1}_l}{2}(\omega^k)^2)\|a^k_l-a^{k-1}_l\|^2_2<\infty
\end{aligned}
\tag{18}
$$

Since the sum of this infinite series is finite, every term converges to 0. This means that $\lim_{k\to\infty}W^{k+1}_l-W^k_l=0$, $\lim_{k\to\infty}z^{k+1}_l-z^k_l=0$ and $\lim_{k\to\infty}a^{k+1}_l-a^k_l=0$. In other words, $\lim_{k\to\infty}\mathbf{W}^{k+1}-\mathbf{W}^k=0$, $\lim_{k\to\infty}\mathbf{z}^{k+1}-\mathbf{z}^k=0$, and $\lim_{k\to\infty}\mathbf{a}^{k+1}-\mathbf{a}^k=0$.
(b). Because $F(\mathbf{W}^k,\mathbf{z}^k,\mathbf{a}^k)$ is bounded, by the definition of coercivity and Assumption 2, $(\mathbf{W}^k,\mathbf{z}^k,\mathbf{a}^k)$ is bounded.
In Algorithm 2:
(c). We sum Equation (11), Equation (12) and Equation (13) from $l=1$ to L and from $k=0$ to K to obtain

$$
\begin{aligned}
&F(\mathbf{W}^0,\mathbf{z}^0,\mathbf{a}^0)-F(\mathbf{W}^K,\mathbf{z}^K,\mathbf{a}^K)\\
&\geq \sum\nolimits_{k=0}^K(\sum\nolimits_{l=1}^L(\frac{\alpha^{k+1}_l}{2}\|W^{k+1}_l-W^k_l\|^2_2+\frac{\gamma^{k+1}_l}{2}\|z^{k+1}_l-z^k_l\|^2_2)+\sum\nolimits_{l=1}^{L-1}\frac{\delta^{k+1}_l}{2}\|a^{k+1}_l-a^k_l\|^2_2)
\end{aligned}
\tag{19}
$$

So $F(\mathbf{W}^K,\mathbf{z}^K,\mathbf{a}^K)\leq F(\mathbf{W}^0,\mathbf{z}^0,\mathbf{a}^0)$. This proves the upper boundness of F. Let $K\rightarrow\infty$ in Equation (19), since $F>0$ is lower bounded, so

$$\sum\nolimits_{k=0}^{K}(\sum\nolimits_{l=1}^{L}(\frac{\alpha_l^{k+1}}{2}\|W_l^{k+1}-W_l^k\|_2^2+\frac{\gamma_l^{k+1}}{2}\|z_l^{k+1}-z_l^k\|_2^2)+\sum\nolimits_{l=1}^{L-1}\frac{\delta_l^{k+1}}{2}\|a_l^{k+1}-a_l^k\|_2^2)<\infty \tag{20}$$

Since the sum of this infinite series is finite, every term converges to 0. This means that $\lim_{k\rightarrow\infty}W_l^{k+1}-W_l^k=0$, $\lim_{k\rightarrow\infty}z_l^{k+1}-z_l^k=0$ and $\lim_{k\rightarrow\infty}a_l^{k+1}-a_l^k=0$. In other words, $\lim_{k\rightarrow\infty}\mathbf{W}^{k+1}-\mathbf{W}^k=0$, $\lim_{k\rightarrow\infty}\mathbf{z}^{k+1}-\mathbf{z}^k=0$, and $\lim_{k\rightarrow\infty}\mathbf{a}^{k+1}-\mathbf{a}^k=0$.
(d). This follows the same routine as the proof of (b) in Algorithm 1. $\square$

Lemma 8 (Subgradient Bound)**.** *In Algorithms 1 and 2, there exists* $C_2=\max(\rho M_{\boldsymbol{a}},\rho M_{\boldsymbol{a}}^2+\theta_1^{k+1},\rho M_{\boldsymbol{a}}^2+\theta_2^{k+1},\cdots,\rho M_{\boldsymbol{a}}^2+\theta_L^{k+1})$*, some* $g_1^{k+1}\in\partial_{\mathbf{W}^{k+1}}F$ *such that*

$$\|g_1^{k+1}\|\leq C_2(\|\boldsymbol{W}^{k+1}-\boldsymbol{W}^k\|+\|\boldsymbol{z}^{k+1}-\boldsymbol{z}^k\|+\|\boldsymbol{W}^k-\boldsymbol{W}^{k-1}\|)$$

Proof. As shown in Remark 2.2 in (Xu & Yin, 2013),

$$\partial_{\mathbf{W}^{k+1}}F=\{\partial_{W_1^{k+1}}F\}\times\{\partial_{W_2^{k+1}}F\}\times\cdots\times\{\partial_{W_L^{k+1}}F\}.$$

where $\times$ denotes Cartesian Product.

In Algorithm 1,

$$\begin{aligned}\partial_{W_l^{k+1}}F&=\partial\Omega_l(W_l^{k+1})+\nabla_{W_l^{k+1}}\phi(a_{l-1}^{k+1},W_l^{k+1},z_l^{k+1})(\text{Definition of } F \text{ in Problem 2})\\&=\nabla_{W_l^{k+1}}\phi(a_{l-1}^{k+1},W_l^{k+1},z_l^{k+1})-\nabla_{\overline{W}_l^{k+1}}\phi(a_{l-1}^{k+1},\overline{W}_l^{k+1},z_l^k)-\theta_l^{k+1}(W_l^{k+1}-\overline{W}_l^{k+1})+\partial\Omega_l(W_l^{k+1})\\&\quad+\nabla_{\overline{W}_l^{k+1}}\phi(a_{l-1}^{k+1},\overline{W}_l^{k+1},z_l^k)+\theta_l^{k+1}(W_l^{k+1}-\overline{W}_l^{k+1})\\&=\rho(W_l^{k+1}-\overline{W}_l^{k+1})a_{l-1}^{k+1}(a_{l-1}^{k+1})^T-\rho(z_l^{k+1}-z_l^k)(a_{l-1}^{k+1})^T-\theta_l^{k+1}(W_l^{k+1}-\overline{W}_l^{k+1})+\partial\Omega_l(W_l^{k+1})\\&\quad+\nabla_{\overline{W}_l^{k+1}}\phi(a_{l-1}^{k+1},\overline{W}_l^{k+1},z_l^k)+\theta_l^{k+1}(W_l^{k+1}-\overline{W}_l^{k+1})\end{aligned}\tag{21}$$

On one hand, we have

$$\begin{aligned}&\|\rho(W_l^{k+1}-\overline{W}_l^{k+1})a_{l-1}^{k+1}(a_{l-1}^{k+1})^T-\rho(z_l^{k+1}-z_l^k)(a_{l-1}^{k+1})^T-\theta_l^{k+1}(W_l^{k+1}-\overline{W}_l^{k+1})\|\\&\leq\rho\|(W_l^{k+1}-\overline{W}_l^{k+1})a_{l-1}^{k+1}(a_{l-1}^{k+1})^T\|+\rho\|(z_l^{k+1}-z_l^k)(a_{l-1}^{k+1})^T\|+\theta_l^{k+1}\|W_l^{k+1}-\overline{W}_l^{k+1}\|(\text{Triangle Inequality})\\&\leq\rho\|W_l^{k+1}-\overline{W}_l^{k+1}\|\|a_{l-1}^{k+1}\|\|a_{l-1}^{k+1}\|+\rho\|z_l^{k+1}-z_l^k\|\|a_{l-1}^{k+1}\|+\theta_l^{k+1}\|W_l^{k+1}-\overline{W}_l^{k+1}\|(\text{Cauchy-Schwarz Inequality})\\&\leq\rho M_{\mathbf{a}}\|z_l^{k+1}-z_l^k\|+(\rho M_{\mathbf{a}}^2+\theta_l^{k+1})\|W_l^{k+1}-\overline{W}_l^{k+1}\|\ (\text{Lemma 7})\\&\leq\rho M_{\mathbf{a}}\|z_l^{k+1}-z_l^k\|+(\rho M_{\mathbf{a}}^2+\theta_l^{k+1})\|W_l^{k+1}-(W_l^k+\omega^k(W_l^k-W_l^{k-1}))\|(\text{Nesterov Acceleration})\\&\leq\rho M_{\mathbf{a}}\|z_l^{k+1}-z_l^k\|+(\rho M_{\mathbf{a}}^2+\theta_l^{k+1})\|W_l^{k+1}-W_l^k\|+(\rho M_{\mathbf{a}}^2+\theta_l^{k+1})\|W_l^k-W_l^{k-1}\|(\text{Triangle Inequality and }\omega^k<1)\end{aligned}\tag{22}$$

On the other hand, the optimality condition of Equation (3) yields

$$0\in\partial\Omega_l(W_l^{k+1})+\nabla_{\overline{W}_l^{k+1}}\phi(a_{l-1}^{k+1},\overline{W}_l^{k+1},z_l^k)+\theta_l^{k+1}(W_l^{k+1}-\overline{W}_l^{k+1})$$

Therefore, there exists $g_{1,l}^{k+1}\in\partial_{W_l^{k+1}}F$ such that

$$\|g_{1,l}^{k+1}\|\leq\rho M_{\mathbf{a}}\|z_l^{k+1}-z_l^k\|+(\rho M_{\mathbf{a}}^2+\theta_l^{k+1})\|W_l^{k+1}-W_l^k\|+(\rho M_{\mathbf{a}}^2+\theta_l^{k+1})\|W_l^k-W_l^{k-1}\|$$

This shows that there exists $g_1^{k+1}=g_{1,1}^{k+1}\times g_{1,2}^{k+1}\times\cdots\times g_{1,L}^{k+1}\in\partial_{\mathbf{W}^{k+1}}F$ and $C_2=\max(\rho M_{\mathbf{a}},\rho M_{\mathbf{a}}^2+\theta_1^{k+1},\rho M_{\mathbf{a}}^2+\theta_2^{k+1},\cdots,\rho M_{\mathbf{a}}^2+\theta_L^{k+1})$ such that

$$\|g_l^{k+1}\|\leq C_2(\|\mathbf{W}^{k+1}-\mathbf{W}^k\|+\|\mathbf{z}^{k+1}-\mathbf{z}^k\|+\|\mathbf{W}^k-\mathbf{W}^{k-1}\|)$$

In Algorithm 2, for W_l^{k+1}, according to Line 7 of Algorithm 2, if $F(\mathbf{W}_{\leq l}^{k+1}, \mathbf{z}_{\leq l-1}^{k+1}, \mathbf{a}_{\leq l-1}^{k+1}) < F(\mathbf{W}_{\leq l-1}^{k+1}, \mathbf{z}_{\leq l-1}^{k+1}, \mathbf{a}_{\leq l-1}^{k+1})$, then as proven in the previous proof in Algorithm 1, there exists $g_{1,l}^{k+1} \in \partial_{W_l^{k+1}} F$ such that

$$\|g_{1,l}^{k+1}\| \leq \rho M_{\mathbf{a}} \|z_l^{k+1} - z_l^k\| + (\rho M_{\mathbf{a}}^2 + \theta_l^{k+1})\|W_l^{k+1} - W_l^k\| + (\rho M_{\mathbf{a}}^2 + \theta_l^{k+1})\|W_l^k - W_l^{k-1}\| \tag{23}$$

Otherwise, we have

$$\begin{aligned}
&\|\rho(W_l^{k+1} - \overline{W}_l^{k+1}) a_{l-1}^{k+1}(a_{l-1}^{k+1})^T - \rho(z_l^{k+1} - z_l^k)(a_{l-1}^{k+1})^T - \theta_l^{k+1}(W_l^{k+1} - \overline{W}_l^{k+1})\| \\
&\leq \rho M_{\mathbf{a}}\|z_l^{k+1} - z_l^k\| + (\rho M_{\mathbf{a}}^2 + \theta_l^{k+1})\|W_l^{k+1} - \overline{W}_l^{k+1}\| (\text{Equation (22)}) \\
&= \rho M_{\mathbf{a}}\|z_l^{k+1} - z_l^k\| + (\rho M_{\mathbf{a}}^2 + \theta_l^{k+1})\|W_l^{k+1} - W_l^k\| (\overline{W}_l^{k+1} = W_l^k)
\end{aligned}$$

The optimality condition of Equation (3) yields

$$0 \in \partial \Omega_l(W_l^{k+1}) + \nabla_{\overline{W}_l^{k+1}} \phi(a_{l-1}^{k+1}, \overline{W}_l^{k+1}, z_l^k) + \theta_l^{k+1}(W_l^{k+1} - \overline{W}_l^{k+1})$$

By Equation (21), we know that there exists $g_{1,l}^{k+1} \in \partial_{W_l^{k+1}} F$ such that

$$\|g_{1,l}^{k+1}\| \leq \rho M_{\mathbf{a}}\|z_l^{k+1} - z_l^k\| + (\rho M_{\mathbf{a}}^2 + \theta_l^{k+1})\|W_l^{k+1} - W_l^k\| \tag{24}$$

Combining Equation (23) with Equation (24), we show that there exists $g_1^{k+1} = g_{1,1}^{k+1} \times g_{1,2}^{k+1} \times \cdots \times g_{1,L}^{k+1} \in \partial_{\mathbf{W}^{k+1}} F$ and $C_2 = \max(\rho M_{\mathbf{a}}, \rho M_{\mathbf{a}}^2 + \theta_1^{k+1}, \rho M_{\mathbf{a}}^2 + \theta_2^{k+1}, \cdots, \rho M_{\mathbf{a}}^2 + \theta_L^{k+1})$ such that

$$\|g_l^{k+1}\| \leq C_2(\|\mathbf{W}^{k+1} - \mathbf{W}^k\| + \|\mathbf{z}^{k+1} - \mathbf{z}^k\| + \|\mathbf{W}^k - \mathbf{W}^{k-1}\|)$$

□

Proof of Theorem 1

Proof. By Lemma 7 (a), $\lim_{k\to\infty} \mathbf{W}^{k+1} - \mathbf{W}^k = 0$. By Lemma 7 (b), there exists a subsequence $\mathbf{W}^s$ such that $\mathbf{W}^s \to \mathbf{W}^*$, where W^* is a limit point. From Lemma 8, there exist $g_1^s \in \partial_{\mathbf{W}^s} F$ such that $\|g_1^s\| \to 0$ as $s \to \infty$. According to the definition of limiting subdifferential, we have $0 \in \partial_{\mathbf{W}^*} F$. In other words, $\mathbf{W}^*$ is a stationary point of F in Problem 2. □

Proof of Theorem 2

Proof. In Algorithm 1, 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 $o(\frac{1}{k})$ based on these two conditions. Specifically, first, we have

$$\begin{aligned}
c_k &= \min_{1\leq i\leq k}\Big(\sum\nolimits_{l=1}^L\Big(\Big(\frac{\theta_l^i}{2} - \frac{\theta_l^{i+1}}{2}(\omega^i)^2\Big)\|W_l^i - W_l^{i-1}\|_2^2 + \frac{\rho}{2}(1-(\omega^i)^2)\|z_l^i - z_l^{i-1}\|_2^2\Big) + \sum\nolimits_{l=1}^{L-1}\Big(\frac{\tau_l^i}{2} - \frac{\tau_l^{i+1}}{2}(\omega^i)^2\Big)\|a_l^i - a_l^{i-1}\|_2^2\Big) \\
&\geq \min_{1\leq i\leq k+1}\Big(\sum\nolimits_{l=1}^L\Big(\Big(\frac{\theta_l^i}{2} - \frac{\theta_l^{i+1}}{2}(\omega^i)^2\Big)\|W_l^i - W_l^{i-1}\|_2^2 + \frac{\rho}{2}(1-(\omega^i)^2)\|z_l^i - z_l^{i-1}\|_2^2\Big) + \sum\nolimits_{l=1}^{L-1}\Big(\frac{\tau_l^i}{2} - \frac{\tau_l^{i+1}}{2}(\omega^i)^2\Big)\|a_l^i - a_l^{i-1}\|_2^2\Big) \\
&= c_{k+1}
\end{aligned}$$

Therefore c_k satisfies the first condition. Second, $\sum_{k=0}^\infty c_k$ is bounded, which is obtained directly from Equation (18). Finally, it has been proved that the sufficient conditions of convergence rate $o(\frac{1}{k})$ are: (1) $c_k \geq c_{k+1}$, (2) $\sum_{k=0}^\infty c_k$ is bounded, and (3) $c_k \geq 0$ (Lemma 1.2 in (Deng et al., 2017)). Since we have proved the first two conditions and the third one $c_k \geq 0$ is obvious, the $o(\frac{1}{k})$ convergence rate of Algorithm 1 is proven. □

Proof of Lemma 1

Proof. We add Equation (11), Equation (12), and Equation (13) from $l=1$ to L to obtain

$$\begin{aligned}&F(\mathbf{W}^k,\mathbf{z}^k,\mathbf{a}^k)-F(\mathbf{W}^{k+1},\mathbf{z}^{k+1},\mathbf{a}^{k+1})\\&\geq\sum\nolimits_{l=1}^{L}(\frac{\alpha_l^{k+1}}{2}\|W_l^{k+1}-W_l^k\|_2^2+\frac{\gamma_l^{k+1}}{2}\|z_l^{k+1}-z_l^k\|_2^2)+\sum\nolimits_{l=1}^{L-1}\frac{\delta_l^{k+1}}{2}\|a_l^{k+1}-a_l^k\|_2^2\end{aligned}$$

Let $C_5=\min(\frac{\alpha_l^{k+1}}{2},\frac{\gamma_l^{k+1}}{2},\frac{\delta_l^{k+1}}{2})>0$, we have

$$\begin{aligned}&F(\mathbf{W}^k,\mathbf{z}^k,\mathbf{a}^k)-F(\mathbf{W}^{k+1},\mathbf{z}^{k+1},\mathbf{a}^{k+1})\\&\geq C_5(\sum\nolimits_{l=1}^{L}(\|W_l^{k+1}-W_l^k\|_2^2+\|z_l^{k+1}-z_l^k\|_2^2)+\sum\nolimits_{l=1}^{L-1}\|a_l^{k+1}-a_l^k\|_2^2)\\&=C_5(\|\mathbf{W}^{k+1}-\mathbf{W}^k\|_2^2+\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_2^2+\|\mathbf{a}^{k+1}-\mathbf{a}^k\|_2^2)\\&\geq 0.\end{aligned}\tag{25}$$

By Lemma 7(d) and a monotone sequence is convergent if it is bounded, then $F(\mathbf{W}^k,\mathbf{z}^k,\mathbf{a}^k)$ is convergent. □

Proof of Theorem 3

Proof. By Lemma 7 (c), $\lim_{k\to\infty}\mathbf{W}^{k+1}-\mathbf{W}^k=0$. By Lemma 7 (d), there exists a subsequence $\mathbf{W}^s$ such that $\mathbf{W}^s\to\mathbf{W}^*$, where W^* is a limit point. From Lemma 8, there exist $g_1^s\in\partial_{\mathbf{W}^s}F$ such that $\|g_1^s\|\to 0$ as $s\to\infty$. According to the definition of limiting subdifferential, we have $0\in\partial_{\mathbf{W}^*}F$. In other words, $\mathbf{W}^*$ is a stationary point of F in Problem 2. □

Proof of Theorem 4

Proof. In Algorithm 2, we prove this by the KL Property.
Firstly, we consider Equation (4) and Equation (6), by Lemma 7, $h_l(\overline{z}_l^{k+1}-\nabla\phi_{\overline{z}_l^{k+1}}/\rho)-a_l^k$ and $h_l(z^{k+1})-\overline{a}_l^{k+1}+\nabla_{\overline{a}_l^{k+1}}\phi/\tau_l^{k+1}$ are bounded, i.e. there exist constants D_1 and D_2 such that

$$\begin{aligned}&\mid h_l(\overline{z}_l^{k+1}-\nabla_{\overline{z}_l^{k+1}}\phi/\rho)-a_l^k\mid<D_1,\\&\mid h_l(z^{k+1})-\overline{a}_l^{k+1}+\nabla_{\overline{a}_l^{k+1}}\phi/\tau_l^{k+1}\mid<D_2\end{aligned}$$

Let $\varepsilon=\max(D_1,D_2)$, then the solutions to Equation (4) and Equation (6) are simplified as follows:

$$z_l^{k+1}\leftarrow\overline{z}_l^{k+1}-\nabla_{\overline{z}_l^{k+1}}\phi/\rho.\tag{26}$$

$$a_l^{k+1}\leftarrow\overline{a}_l^{k+1}-\nabla_{\overline{a}_l^{k+1}}\phi/\tau_l^{k+1}.\tag{27}$$

This is because $h_l(z_l^{k+1})-\varepsilon\leq a_l^k\leq h_l(z_l^{k+1})+\varepsilon$ and $h_l(z_l^{k+1})-\varepsilon\leq a_l^{k+1}\leq h_l(z_l^{k+1})+\varepsilon$ hold in Equation (4) and Equation (6), respectively.
Next, we prove that given $\varepsilon=\max(D_1,D_2)$, there exists $C_3=\max(\rho M_{\mathbf{W}}^2+\tau_1^{k+1},\rho M_{\mathbf{W}}^2+\tau_2^{k+1},\rho M_{\mathbf{W}}^2+\tau_3^{k+1},\cdots,\rho M_{\mathbf{W}}^2+\tau_{L-1}^{k+1},2\rho M_{\mathbf{W}}M_{\mathbf{a}}+\rho M_{\mathbf{z}})$, some $g_3^{k+1}\in\partial_{\mathbf{z}^{k+1}}F$ and $g_4^{k+1}\in\partial_{\mathbf{a}^{k+1}}F$ such that

$$\begin{aligned}&\|g_3^{k+1}\|=0,\\&\|g_4^{k+1}\|\leq C_3(\|\mathbf{a}^{k+1}-\mathbf{a}^k\|+\|\mathbf{a}^k-\mathbf{a}^{k-1}\|+\|\mathbf{W}^{k+1}-\mathbf{W}^k\|+\|\mathbf{z}^{k+1}-\mathbf{z}^k\|)\end{aligned}$$

As shown in (Wang et al., 2015; Xu & Yin, 2013),

$$\begin{aligned}&\partial_{\mathbf{z}^{k+1}}F=\partial_{z_1^{k+1}}F\times\partial_{z_2^{k+1}}F\times\cdots\times\partial_{z_L^{k+1}}F.\\&\nabla_{\mathbf{a}^{k+1}}F=\nabla_{a_1^{k+1}}F\times\nabla_{a_2^{k+1}}F\times\cdots\times\nabla_{a_{L-1}^{k+1}}F.\end{aligned}$$

where $\times$ denotes Cartesian Product.
For $z_l^{k+1}(l<L)$, according to Line 18 of Algorithm 2, no matter
$F(\mathbf{W}_{\leq l}^{k+1},\mathbf{z}_{\leq l}^{k+1},\mathbf{a}_{\leq l-1}^{k+1})\geq F(\mathbf{W}_{\leq l}^{k+1},\mathbf{z}_{\leq l-1}^{k+1},\mathbf{a}_{\leq l-1}^{k+1})$ or not, we have

$$\begin{aligned}\partial_{z_l^{k+1}}F&=\nabla_{z_l^{k+1}}\phi(a_{l-1}^{k+1},W_l^{k+1},z_l^{k+1})\\&=\nabla_{z_l^{k+1}}\phi(a_{l-1}^{k+1},W_l^{k+1},z_l^{k+1})-\nabla_{\overline{z}_l^{k+1}}\phi(a_{l-1}^{k+1},W_l^{k+1},\overline{z}_l^{k+1})-\rho(z_l^{k+1}-\overline{z}_l^{k+1})(\text{Equation (26)})\\&=0\end{aligned}$$

For z_L^{k+1}, according to Line 12 of Algorithm 2, no matter
$F(\mathbf{W}_{\leq L}^{k+1},\mathbf{z}_{\leq L}^{k+1},\mathbf{a}_{\leq L-1}^{k+1})\geq F(\mathbf{W}_{\leq L}^{k+1},\mathbf{z}_{\leq L-1}^{k+1},\mathbf{a}_{\leq L-1}^{k+1})$ or not, we have

$$\begin{aligned}\partial_{z_L^{k+1}}F&=\nabla_{z_L^{k+1}}\phi(a_{L-1}^{k+1},W_L^{k+1},z_L^{k+1})+\partial R(z_L^{k+1};y)\\&=\nabla_{z_L^{k+1}}\phi(a_{L-1}^{k+1},W_L^{k+1},z_L^{k+1})+\partial R(z_L^{k+1};y)+\nabla_{\overline{z}_L^{k+1}}\phi(a_{L-1}^{k+1},W_L^{k+1},\overline{z}_L^{k+1})\\&+\rho(z_L-\overline{z}_L^{k+1})-\nabla_{\overline{z}_L^{k+1}}\phi(a_{L-1}^{k+1},W_L^{k+1},\overline{z}_L^{k+1})-\rho(z_L^{k+1}-\overline{z}_L^{k+1})\\&=\nabla_{z_L^{k+1}}\phi(a_{L-1}^{k+1},W_L^{k+1},z_L^{k+1})-\nabla_{\overline{z}_L^{k+1}}\phi(a_{L-1}^{k+1},W_L^{k+1},\overline{z}_L^{k+1})-\rho(z_L^{k+1}-\overline{z}_L^{k+1})\\&(0\in\partial R(z_L^{k+1};y)+\nabla_{\overline{z}_L^{k+1}}\phi(a_{L-1}^{k+1},W_L^{k+1},\overline{z}_L^{k+1})+\rho(z_L^{k+1}-\overline{z}_L^{k+1})\text{by the optimality condition of Equation (5)})\\&=0\end{aligned}$$

Therefore, there exists $g_{3,l}^{k+1}=\nabla_{z_l^{k+1}}F$ such that

$$\|g_{3,l}^{k+1}\|=0$$

This shows that there exists $g_3^{k+1}=g_{3,1}^{k+1}\times g_{3,2}^{k+1}\times\cdots\times g_{3,L}^{k+1}=\nabla_{\mathbf{z}^{k+1}}F$ such that

$$\|g_3^{k+1}\|=0 \tag{28}$$

For a_l^{k+1}, we have

$$\begin{aligned}\partial_{a_l^{k+1}}F&=\nabla_{a_l^{k+1}}\phi(a_l^{k+1},W_{l+1}^{k},z_{l+1}^{k+1})\\&=\nabla_{a_l^{k+1}}\phi(a_l^{k+1},W_{l+1}^{k+1},z_{l+1}^{k+1})-\nabla_{\overline{a}_l^{k+1}}\phi(\overline{a}_l^{k+1},W_{l+1}^{k},z_{l+1}^{k})-\tau_l^{k+1}(a_l^{k+1}-\overline{a}_l^{k+1})(\text{Equation (27)})\\&=\rho(W_{l+1}^{k+1})^T(W_{l+1}^{k+1}a_l^{k+1}-z_{l+1}^{k+1})-\rho(W_{l+1}^{k})^T(W_{l+1}^{k}\overline{a}_l^{k+1}-z_{l+1}^{k})-\tau_l^{k+1}(a_l^{k+1}-\overline{a}_l^{k+1})\\&=\rho(W_{l+1}^{k+1})^TW_{l+1}^{k+1}(a_l^{k+1}-\overline{a}_l^{k+1})+\rho(W_{l+1}^{k+1})^T(W_{l+1}^{k+1}-W_{l+1}^{k})\overline{a}_l^{k+1}\\&+\rho(W_{l+1}^{k+1}-W_{l+1}^{k})^TW_{l+1}^{k}\overline{a}_l^{k+1}-\rho(W_{l+1}^{k+1})^T(z_{l+1}^{k+1}-z_{l+1}^{k})-\rho(W_{l+1}^{k+1}-W_{l+1}^{k})^Tz_{l+1}^{k}-\tau_l^{k+1}(a_l^{k+1}-\overline{a}_l^{k+1})\end{aligned}$$

Therefore

$$\begin{aligned}\|\partial_{a_l^{k+1}}F\|&\leq\rho\|W_{l+1}^{k+1}\|\|W_{l+1}^{k+1}\|\|a_l^{k+1}-\overline{a}_l^{k+1}\|+\rho\|W_{l+1}^{k+1}\|\|W_{l+1}^{k+1}-W_{l+1}^{k}\|\|\overline{a}_l^{k+1}\|\\&+\rho\|W_{l+1}^{k+1}-W_{l+1}^{k}\|\|W_{l+1}^{k}\|\|\overline{a}_l^{k+1}\|+\rho\|W_{l+1}^{k+1}\|\|z_{l+1}^{k+1}-z_{l+1}^{k}\|\\&+\rho\|W_{l+1}^{k+1}-W_{l+1}^{k}\|\|z_{l+1}^{k}\|+\tau_l^{k+1}\|a_l^{k+1}-\overline{a}_l^{k+1}\|\\&(\text{Triangle Inequality and Cauthy-Schwarz Inequality})\\&\leq\rho M_{\mathbf{W}}^2\|a_l^{k+1}-\overline{a}_l^{k+1}\|+\rho M_{\mathbf{W}}\|W_{l+1}^{k+1}-W_{l+1}^{k}\|M_{\mathbf{a}}+\rho\|W_{l+1}^{k+1}-W_{l+1}^{k}\|M_{\mathbf{W}}M_{\mathbf{a}}+\rho M_{\mathbf{W}}\|z_{l+1}^{k+1}-z_{l+1}^{k}\|\\&+\rho\|W_{l+1}^{k+1}-W_{l+1}^{k}\|M_{\mathbf{z}}+\tau_l^{k+1}\|a_l^{k+1}-\overline{a}_l^{k+1}\|\ (\text{Lemma 7})\\&=(\rho M_{\mathbf{W}}^2+\tau_l^{k+1})\|a_l^{k+1}-\overline{a}_l^{k+1}\|+(2\rho M_{\mathbf{W}}M_{\mathbf{a}}+\rho M_{\mathbf{z}})\|W_{l+1}^{k+1}-W_{l+1}^{k}\|+\rho M_{\mathbf{W}}\|z_{l+1}^{k+1}-z_{l+1}^{k}\|\end{aligned}$$

According to Line 22 of Algorithm 2, if

$F(\mathbf{W}_{\leq l}^{k+1},\mathbf{z}_{\leq l}^{k+1},\mathbf{a}_{\leq l}^{k+1})<F(\mathbf{W}_{\leq l}^{k+1},\mathbf{z}_{\leq l}^{k+1},\mathbf{a}_{\leq l-1}^{k+1})$, then we have

$$\begin{aligned}
\|\partial_{a_l^{k+1}}F\| &\leq (\rho M_{\mathbf{W}}^2+\tau_l^{k+1})\|a_l^{k+1}-a_l^k-(a_l^k-a_l^{k-1})\omega^k\|+(2\rho M_{\mathbf{W}}M_{\mathbf{a}}+\rho M_{\mathbf{z}})\|W_{l+1}^{k+1}-W_{l+1}^k\|+\rho M_{\mathbf{W}}\|z_{l+1}^{k+1}-z_{l+1}^k\| \\
&\text{(Nestrov Acceleration)} \\
&\leq (\rho M_{\mathbf{W}}^2+\tau_l^{k+1})\|a_l^{k+1}-a_l^k\|+(\rho M_{\mathbf{W}}^2+\tau_l^{k+1})\|a_l^k-a_l^{k-1}\|+(2\rho M_{\mathbf{W}}M_{\mathbf{a}}+\rho M_{\mathbf{z}})\|W_{l+1}^{k+1}-W_{l+1}^k\| \\
&+\rho M_{\mathbf{W}}\|z_{l+1}^{k+1}-z_{l+1}^k\| \text{ (Triangle Inequality and } \omega^k<1)
\end{aligned}$$

Therefore, there exists $g_{4,l}^{k+1}\in\partial_{a_l^{k+1}}F$ such that

$$\begin{aligned}
\|g_{4,l}^{k+1}\| &\leq (\rho M_{\mathbf{W}}^2+\tau_l^{k+1})\|a_l^{k+1}-a_l^k\|+(\rho M_{\mathbf{W}}^2+\tau_l^{k+1})\|a_l^k-a_l^{k-1}\|+(2\rho M_{\mathbf{W}}M_{\mathbf{a}}+\rho M_{\mathbf{z}})\|W_{l+1}^{k+1}-W_{l+1}^k\| \\
&+\rho M_{\mathbf{W}}\|z_{l+1}^{k+1}-z_{l+1}^k\|
\end{aligned} \tag{29}$$

Otherwise,

$$\|\partial_{a_l^{k+1}}F\|\leq(\rho M_{\mathbf{W}}^2+\tau_l^{k+1})\|a_l^{k+1}-a_l^k\|+(2\rho M_{\mathbf{W}}M_{\mathbf{a}}+\rho M_{\mathbf{z}})\|W_{l+1}^{k+1}-W_{l+1}^k\|+\rho M_{\mathbf{W}}\|z_{l+1}^{k+1}-z_{l+1}^k\|\ (\overline{a}_l^{k+1}=a_l^k)$$

Therefore, there exists $g_{4,l}^{k+1}\in\partial_{a_l^{k+1}}F$ such that

$$\|g_{4,l}^{k+1}\|\leq(\rho M_{\mathbf{W}}^2+\tau_l^{k+1})\|a_l^{k+1}-a_l^k\|+(2\rho M_{\mathbf{W}}M_{\mathbf{a}}+\rho M_{\mathbf{z}})\|W_{l+1}^{k+1}-W_{l+1}^k\|+\rho M_{\mathbf{W}}\|z_{l+1}^{k+1}-z_{l+1}^k\| \tag{30}$$

Combining Equation (29) and Equation (30), we show that there exists $g_4^{k+1}=g_{4,1}^{k+1}\times g_{4,2}^{k+1}\times\cdots\times g_{4,L}^{k+1}\in\partial_{\mathbf{a}^{k+1}}F$ and $C_3=\max(\rho M_{\mathbf{W}}^2+\tau_1^{k+1},\rho M_{\mathbf{W}}^2+\tau_2^{k+1},\rho M_{\mathbf{W}}^2+\tau_3^{k+1},\cdots,\rho M_{\mathbf{W}}^2+\tau_{L-1}^{k+1},2\rho M_{\mathbf{W}}M_{\mathbf{a}}+\rho M_{\mathbf{z}})$ such that

$$\|g_4^{k+1}\|\leq C_3(\|\mathbf{a}^{k+1}-\mathbf{a}^k\|+\|\mathbf{a}^k-\mathbf{a}^{k-1}\|+\|\mathbf{W}^{k+1}-\mathbf{W}^k\|+\|\mathbf{z}^{k+1}-\mathbf{z}^k\|) \tag{31}$$

Combining Lemma 8, Equation (28) and Equation (31), we prove that there exists $g^{k+1}\in\partial F(\mathbf{W}^{k+1},\mathbf{z}^{k+1},\mathbf{a}^{k+1})=\{\partial_{\mathbf{W}^{k+1}}F,\partial_{\mathbf{z}^{k+1}}F,\partial_{\mathbf{a}^{k+1}}F\}$ and $C_4=\max(C_2,C_3,\rho)$ such that

$$\|g^{k+1}\|\leq C_4(\|\mathbf{a}^{k+1}-\mathbf{a}^k\|+\|\mathbf{a}^k-\mathbf{a}^{k-1}\|+\|\mathbf{W}^{k+1}-\mathbf{W}^k\|+\|\mathbf{W}^k-\mathbf{W}^{k-1}\|+\|\mathbf{z}^{k+1}-\mathbf{z}^k\|) \tag{32}$$

Finally, we prove the linear convergence rate by the KL Property given Equation (32) and Equation (25). Because F is locally strongly convex with a constant μ, F satisfies the KL Property by Lemma 3. Let $F^*=F(\mathbf{W}^*,\mathbf{z}^*,\mathbf{a}^*)$ be the convergent value of F, by Lemma 1, $F(\mathbf{W}^k,\mathbf{z}^k,\mathbf{a}^k)\to F^*$, then for any $\eta_1>0$ there exists $k_2\in\mathbb{N}$ such that it holds for $k>k_2$ that $F^*<F(\mathbf{W}^k,\mathbf{z}^k,\mathbf{a}^k)<F^*+\eta_1$. Also by Lemma 7(c) and Equation (32), $g^{k+1}\to 0$ as $k\to\infty$, then for any $\eta_2>0$ there exists $k_3\in\mathbb{N}$, such that it holds for $k>k_3$ that $\|g^{k+1}\|<\eta_2$. Therefore, for any $k>k_1=\max(k_2,k_3)$, $(\mathbf{W}^k,\mathbf{z}^k,\mathbf{a}^k)\in\{(\mathbf{W},\mathbf{z},\mathbf{a}):|F^*<F(\mathbf{W},\mathbf{z},\mathbf{a})<F^*+\eta_1\cap\exists g\in F(\mathbf{W},\mathbf{z},\mathbf{a})\ s.t.\ \|g\|<\eta_2\}$. By the KL Property and Lemma 3, it holds that

$$\begin{aligned}
1 &\leq \|g^{k+1}\|/(\mu\sqrt{F(\mathbf{W}^{k+1},\mathbf{z}^{k+1},\mathbf{a}^{k+1})-F^*}) \\
&\leq C_4(\|\mathbf{a}^{k+1}-\mathbf{a}^k\|+\|\mathbf{a}^k-\mathbf{a}^{k-1}\|+\|\mathbf{W}^{k+1}-\mathbf{W}^k\|+\|\mathbf{W}^k-\mathbf{W}^{k-1}\|+\|\mathbf{z}^{k+1}-\mathbf{z}^k\|)/(\mu\sqrt{F(\mathbf{W}^{k+1},\mathbf{z}^{k+1},\mathbf{a}^{k+1})-F^*}) \\
&\text{(Equation (32))} \\
&\leq C_4^2(\|\mathbf{a}^{k+1}-\mathbf{a}^k\|+\|\mathbf{a}^k-\mathbf{a}^{k-1}\|+\|\mathbf{W}^{k+1}-\mathbf{W}^k\|+\|\mathbf{W}^k-\mathbf{W}^{k-1}\|+\|\mathbf{z}^{k+1}-\mathbf{z}^k\|)^2/(\mu^2(F(\mathbf{W}^{k+1},\mathbf{z}^{k+1},\mathbf{a}^{k+1})-F^*)) \\
&\leq (5C_4^2(\|\mathbf{a}^{k+1}-\mathbf{a}^k\|_2^2+\|\mathbf{a}^k-\mathbf{a}^{k-1}\|_2^2+\|\mathbf{W}^{k+1}-\mathbf{W}^k\|_2^2+\|\mathbf{W}^k-\mathbf{W}^{k-1}\|_2^2+\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_2^2))/(\mu^2(F(\mathbf{W}^{k+1},\mathbf{z}^{k+1},\mathbf{a}^{k+1})-F^*)) \\
&\text{(Mean Inequality)} \\
&\leq (5C_4^2(F(\mathbf{W}^{k-1},\mathbf{z}^{k-1},\mathbf{a}^{k-1})-F(\mathbf{W}^{k+1},\mathbf{z}^{k+1},\mathbf{a}^{k+1})))/(C_5\mu^2(F(\mathbf{W}^{k+1},\mathbf{z}^{k+1},\mathbf{a}^{k+1})-F^*))\text{(Equation (25))}
\end{aligned}$$

This indicates that

$$(C_5\mu^2+5C_4^2)(F(\mathbf{W}^{k+1},\mathbf{z}^{k+1},\mathbf{a}^{k+1})-F^*)\leq 5C_4^2(F(\mathbf{W}^{k-1},\mathbf{z}^{k-1},\mathbf{a}^{k-1})-F^*)$$

Let $0 < C_1 = \frac{5C_4^2}{C_5\mu^2+5C_4^2} < 1$, we have

$$F(\mathbf{W}^{k+1}, \mathbf{z}^{k+1}, \mathbf{a}^{k+1}) - F^* \leq C_1(F(\mathbf{W}^{k-1}, \mathbf{z}^{k-1}, \mathbf{a}^{k-1}) - F^*)$$

So in summary, for any ρ, there exist $\varepsilon = \max(D_1, D_2)$, $k_1 = \max(k_2, k_3)$, and $0 < C_1 = \frac{5C_4^2}{C_5\mu^2+5C_4^2} < 1$ such that

$$F(\mathbf{W}^{k+1}, \mathbf{z}^{k+1}, \mathbf{a}^{k+1}) - F^* \leq C_1(F(\mathbf{W}^{k-1}, \mathbf{z}^{k-1}, \mathbf{a}^{k-1}) - F^*)$$

for $k > k_1$. In other words, the linear convergence rate is proven. $\square$

Discussion

We discuss convergence conditions of the DLAM algorithm compared with SGD-type methods and the dlADMM method. The comparison demonstrates that our convergence conditions are more general than others.

1. DLAM versus SGD

One influential work by Ghadimi et al. (Ghadimi & Lan, 2016) guaranteed that the SGD converges to a stationary point, which is similar to our convergence results. While the SGD requires the objective function to be Lipschitz differentiable, bounded from below (Ghadimi & Lan, 2016), our DLAM allows for non-smooth functions such as ReLU. Therefore, our convergence conditions are milder than SGD.

2. DLAM versus dlADMM

Wang et al. (Wang et al., 2019) proposed an improved version of ADMM for deep learning models called dlADMM. They showed that the dlADMM is convergent to a stationary point. However, assumptions of our DLAM are milder than those of the dlADMM: the DLAM requires activation functions to be quasilinear, which includes sigmoid, tanh, ReLU and leaky ReLU, while the dlADMM assumes that activation functions make subproblems solvable, which only includes ReLU and leaky ReLU.