An Inexact Augmented Lagrangian Algorithm for Training Leaky ReLU Neural Network with Group Sparsity

Wei Liu liuwei175@lsec.cc.ac.cn

Xin Liu liuxin@lsec.cc.ac.cn

Institute of Computational Mathematics and Scientific/Engineering Computing
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing 100190, China

Xiaojun Chen xiaojun.chen@polyu.edu.hk

Department of Applied Mathematics
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong

Editor: Miguel Carreira-Perpiñán

Abstract

The leaky ReLU network with a group sparse regularization term has been widely used in the recent years. However, training such network yields a nonsmooth nonconvex optimization problem and there exists a lack of approaches to compute a stationary point deterministically. In this paper, we first resolve the multi-layer composite term in the original optimization problem by introducing auxiliary variables and additional constraints. We show the new model has a nonempty and bounded solution set and its feasible set satisfies the Mangasarian-Fromovitz constraint qualification. Moreover, we show the relationship between the new model and the original problem. Remarkably, we propose an inexact augmented Lagrangian algorithm for solving the new model, and show the convergence of the algorithm to a KKT point. Numerical experiments demonstrate that our algorithm is more efficient for training sparse leaky ReLU neural networks than some well-known algorithms.

Keywords: sparse neural network, leaky ReLU, group sparsity, penalty method, inexact augmented Lagrangian method

1. Introduction

In this paper, we focus on the parameter estimation problem of the leaky ReLU network (Maas et al., 2013) with the $l_{2,1}$ regularizer, which pursues the group sparsity. The problem can be formulated as

$$\min_{w,b} \frac{1}{N} \sum_{n=1}^{N} \|\sigma(W_L\sigma(\cdots\sigma(W_1x_n + b_1) + \cdots) + b_L) - y_n\|^2 + \mathcal{R}_1(w). \tag{1}$$

Here $\{x_n \in \mathbb{R}^{N_0}\}_{n=1}^N$ and $\{y_n \in \mathbb{R}^{N_L}\}_{n=1}^N$ are the given input and output data, respectively, σ stands for the component-wise activation function, variables $W_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and $b_\ell \in \mathbb{R}^{N_\ell}$ represent the weight matrices and the bias vectors for all $\ell \in [L]$, respectively, $\|\cdot\|$ refers to

the l_2 norm, and $\mathcal{R}_1 : \mathbb{R}^{\widetilde{N}} \to \mathbb{R}$ is the sparse regularizer of w. For convenience, we set

$$w = \left(\mathrm{vec}(W_1)^\top, \ldots, \mathrm{vec}(W_L)^\top\right)^\top \in \mathbb{R}^{\widetilde{N}},\ b = \left(b_1^\top, \ldots, b_L^\top\right)^\top \in \mathbb{R}^{\overline{N}},$$

where $\mathrm{vec}(W_\ell) \in \mathbb{R}^{N_{\ell-1}N_\ell}$ is the column-wise vectorization of W_ℓ, $\widetilde{N} := \sum_{\ell=1}^{L} N_\ell N_{\ell-1}$ $\overline{N} := \sum_{\ell=1}^{L} N_\ell$, and $[L]$ denotes $\{1, 2, \ldots, L\}$. In this paper, we set σ as the leaky ReLU activation function and $\mathcal{R}_1$ as the group lasso regularizer, i.e.,

$$\sigma(z) = \max\{z, \alpha z\},\ \mathcal{R}_1(w) := \lambda_w \sum_{\ell=1}^{L} \|W_\ell\|_{2,1} = \lambda_w \sum_{\ell=1}^{L} \sum_{j=1}^{N_{\ell-1}} \|(W_\ell)_{\cdot,j}\|,$$

where $\lambda_w > 0$, $0 < \alpha < 1$, $(W_\ell)_{\cdot,j}$ stands for the j-th column of W_ℓ, and $\max\{z, \alpha z\} = (\max\{z_1, \alpha z_1\}, \ldots, \max\{z_K, \alpha z_K\})^\top$ for any $z \in \mathbb{R}^K$.

It is worth noting that the activation functions ReLU and the leaky ReLU get more and more popular in recent applications, as they can alleviate the overfitting phenomenon and pursue the model (neuron) sparsity, e.g., almost half of the neurons in the ReLU network are zero (Jarrett et al., 2009; Nair and Hinton, 2010; Glorot et al., 2011; Maas et al., 2013; Dahl et al., 2013; He et al., 2015; Agarap, 2018). Moreover, the performance of the leaky ReLU network is reported to be slightly better than that of the ReLU network (Maas et al., 2013; Pedamonti, 2018). As we will show in Theorem 8, the leaky ReLU network with a regularization term has a nonempty and bounded solution set, while the ReLU network with a regularization term does not have the property (see a counterexample given by Liu et al. (2022)). For simplicity, we focus on the leaky ReLU network in this paper. Our new model, algorithm and theoretical analysis can be generalized to the ReLU network easily (see Remark 15).

In training a deep neural network (DNN, e.g., the leaky ReLU network), regularization techniques play an important role in reducing the generalization error (also called the test error) (Goodfellow et al., 2016). The l_2 regularizer (i.e., $\|\cdot\|^2$, also called the weight decay) has been widely used for training the DNN (Goodfellow et al., 2016). Recently, sparse regularizers, such as the lasso regularizer (Goodfellow et al., 2016) and the group sparse regularizer (Zhou et al., 2010; Wen et al., 2016; Feng and Simon, 2017; Yoon and Hwang, 2017; Scardapane et al., 2017), are superior to the l_2 regularizer in pursuing the parameter sparsity and lead to theoretical improvement in efficiency (Hoefler et al., 2021). Moreover, Wen et al. (2016) show that by using the gradient descent methods, less training time is required by DNN with a group sparse regularizer compared with that required by DNN with a lasso regularizer. The group sparse regularizer also appears in convolution neural network (Bui et al., 2021) and other machine learning problems (Meier et al., 2008; Jenatton et al., 2011; Simon et al., 2013), etc. Hence, we focus on training the leaky ReLU network with the $l_{2,1}$ regularizer for pursuing the group sparsity.

The stochastic gradient descent based methods, including the stochastic gradient descent methods (SGD), are widely used in training DNN including the leaky ReLU network with group sparsity, while they neglect the fact that the subdifferentials of the objective function at those nondifferentiable points are not available (Abadi et al., 2016; Paszke et al., 2019). Instead, they calculate the "gradient" via the "chain rule" brutally no matter the "chain rule" applies or not (Telgarsky, 2020; Bolte and Pauwels, 2021)). Therefore, gradient descent

based approaches can not deterministically yield Clarke stationary points (see Definition 1) and may encounter numerical troubles in extreme cases. Recently, Davis et al. (2020) prove that the stochastic subgradient (SSGD) method for training the nonsmooth network can obtain Clarke stationary points for the leaky ReLU network with probability 1. However, they have not explained how to calculate a subgradient practically. Moreover, even though a Clarke stationary point is obtained, it may be far away from any local minimizer (see Example 1).

Carreira-Perpiñán and Wang (2012, 2014) proposed a new methodology that introduces auxiliary variables and constraints to resolve the multi-layer nonsmoothness in neural networks. By using quadratic penalties to enforce equality constraints, their method enables efficient training of deep networks and allows distributed computing. This methodology has been adopted for training deep neural networks in (Taylor et al., 2016; Lau et al., 2018; Zeng et al., 2019; Evens et al., 2021). However, it is important to note that such approach may not always guarantee finding stationary points of the original problem (1). To address this issue, Cui et al. (2020) propose an l_1 penalty method, which yields a directional stationary point theoretically, for training the DNN with piecewise activation functions and an l_2 regularizer. Liu et al. (2022) propose a smoothing method that finds a Clarke stationary point for the two-layer ReLU network. To the best of our knowledge, algorithms with guaranteed global convergence to KKT points for a nonsmooth deep neural network with group sparsity have not been developed yet.

1.1 Motivation

In this paper, we aim to explore efficient approaches for solving problem (1) with guaranteed convergence. Hence, we pay our attention to the methods which introduce auxiliary variables and constraints to resolve the multi-layer nonsmoothness. To peel the complicated composite objective of (1) like bamboo shoot, we first introduce the following variables,

$$v := \left(v_{1,1}^\top, v_{2,1}^\top, \ldots, v_{1,L}^\top, v_{2,L}^\top, \ldots, v_{N,L}^\top\right)^\top \in \mathbb{R}^m, \tag{2}$$

where $m := N\overline{N}$, $v_{n,\ell} := \sigma(W_\ell\sigma(\cdots\sigma(W_1x_n + b)_+ + \cdots) + b_\ell)$, $v_{n,0} := x_n$ for all $\ell \in [L]$ and $n \in [N]$. Specifically, Liu et al. (2022) introduce a linearly constrained model for training a two-layer ReLU network with a regularization term. For solving the sparse leaky ReLU network with more than two layers, we introduce in this paper a regularization term $\mathcal{R}_2(v) : \mathbb{R}^m \mapsto \mathbb{R}$ by

$$\mathcal{R}_2(v) := \lambda_v\|v\|^2,$$

and a new group of variables

$$u = \left(u_{1,1}^\top, u_{2,1}^\top, \ldots, u_{1,L}^\top, u_{2,L}^\top, \ldots, u_{N,L}^\top\right)^\top \in \mathbb{R}^m, \tag{3}$$

where $\lambda_v > 0$, $u_{n,\ell} = W_\ell v_{n,\ell-1} + b_\ell$ for all $n \in [N]$ and $\ell \in [L]$. Then, we derive the following model

$$\begin{aligned} \min_{w,b,v,u}\ & \bar{\mathcal{O}}(w,v) := \frac{1}{N}\sum_{n=1}^{N}\|v_{n,L} - y_n\|^2 + \mathcal{R}_1(w) + \mathcal{R}_2(v) \\ \text{s.t. }\ & \sigma(u_{n,\ell}) - v_{n,\ell} = 0,\ u_{n,\ell} - (W_\ell v_{n,\ell-1} + b_\ell) = 0, \\ & n \in [N],\ \ell \in [L]. \end{aligned} \tag{P}$$

The regularization teams $\mathcal{R}_1$ and $\mathcal{R}_2$ lead to the level boundedness of the objective function $\bar{\mathcal{O}}$. Moreover, $\mathcal{R}_1$ imposes column-wise sparsity of the weight matrices W_ℓ for all $\ell \in [L]$. By defining the linear operator $\Psi(v) : \mathbb{R}^m \mapsto \mathbb{R}^{m\times\widetilde{N}}$ and the matrix $A \in \mathbb{R}^{m\times\overline{N}}$ as

$$\Psi(v) = \begin{bmatrix} X^\top \otimes I_{N_1} & \dots & \dots & 0 \\ 0 & V_1^\top \otimes I_{N_2} & \dots & 0 \\ 0 & \dots & \dots & 0 \\ 0 & \dots & \dots & V_{L-1}^\top \otimes I_{N_L} \end{bmatrix} \text{ and}$$

$$A = \begin{bmatrix} e_N \otimes I_{N_1} & \dots & \dots & 0 \\ 0 & e_N \otimes I_{N_2} & \dots & 0 \\ 0 & \dots & \dots & 0 \\ 0 & \dots & \dots & e_N \otimes I_{N_L} \end{bmatrix},$$

respectively, where $X := (x_1, x_2, \dots, x_n)$, $V_\ell := (v_{1,\ell}, v_{2,\ell}, \dots, v_{N,\ell}) \in \mathbb{R}^{N_\ell \times N}$ for all $\ell \in [L]$, $\otimes$ represents the Kronecker product and $e_K \in \mathbb{R}^K$ denotes the all-ones vector for $K \in \mathbb{N}_+$, the constraint set of problem (P) can be simply written as

$$v - \sigma(u) = 0, u = \Psi(v)w + Ab.$$

Due to the nonsmoothness of $\sigma(u)$, problem (P) does not satisfy a standard constraint qualification for mathematical programming. We consider to have $v \geq \sigma(u)$ as a constraint and add a penalty term $\beta^\top(v - \sigma(u))$ in the objective function, where

$$\beta = (\beta_1 e_{NN_1}^\top, \dots, \beta_L e_{NN_L}^\top)^\top \in \mathbb{R}^m$$

with constants $\beta_\ell > 0$ for all $\ell \in [L]$. Note that the inequality $v \geq \sigma(u)$ is equivalent to the inequalities $v - u \geq 0$ and $v - \alpha u \geq 0$, we then present the partial l_1 penalty model for problem (P) as follows

$$\begin{aligned} \min_{w,b,v,u} \ & \mathcal{O}(w, v, u) = \bar{\mathcal{O}}(w, v) + \beta^\top(v - \sigma(u)) \\ \text{s.t.} \ & v - u \geq 0, v - \alpha u \geq 0, u = \Psi(v)w + Ab. \end{aligned} \tag{PP}$$

For brevity, we denote the feasible sets of problems (P) and (PP) by Ω_1 and Ω_2, respectively.

It is worth noting that problem (PP) is a nonsmooth nonconvex mathematical programming, where the second term $\beta^\top(v - \sigma(u))$ is nonsmooth, $\mathcal{R}_1$ is a convex nonsmooth regularizer, the inequality constraints are linear, the equality constraints are nonconvex bilinear. Hence, both the objective function and the feasible region of problem (PP) are much more complicated than the optimization problem for two-layer network proposed by Liu et al. (2022), which is a linearly constrained programming. Hence, the approaches therein can not be straightforwardly extended to solve problem (PP).

1.2 Contributions

We consider a regularized minimization model (P) with auxiliary variables and nonsmooth constraints for training the leaky ReLU network with group sparsity. We investigate its partial l_1 penalty model (PP) and establish the relationships between these two models

with respect to global minimizers, local minimizers, and stationary points under some mild conditions. Moreover, we show that the solution set of problem (PP) is bounded and any feasible point of problem (PP) satisfies the Mangasarian-Fromovitz constraint qualification. Based on these results, we theoretically verify the equivalence between the KKT points and the limiting stationary points of (PP), and further prove that any KKT point of (PP) is an MPCC W-stationary point of problem (P).

By exploiting the structure of problem (PP), we propose an inexact augmented Lagrangian method, whose subproblem at each iteration is solved by an alternating minimization method (IALAM). Different from the existing inexact augmented Lagrangian methods for nonsmooth nonconvex optimization problems (Lu and Zhang, 2012; Chen et al., 2017), we design a new rule for updating the Lagrangian penalty parameter. We also prove that any iterate sequence generated by IALAM has accumulation points, any of which is a limiting stationary point (or equivalently KKT point) of problem (PP) without assuming the existence of accumulation points. Moreover, any limiting stationary point of problem (PP) is a Clarke stationary point of problem (PP).

The numerical experiments demonstrate that IALAM, equipped with prefixed algorithm parameters, outperforms the popular SGD-based methods (e.g., Adam, Adadelda, and vanilla SGD) and ProxSGD in solving problems arisen from both synthetic and MNIST data sets. More specifically, compared with SGD-based methods, IALAM achieves lower training error and test error, and obtains sparser solutions.

By applying IALAM to training both the ReLU and the leaky ReLU networks under the same settings, we find that the leaky ReLU network with a small positive α (e.g., $\alpha = 0.01$) often leads to slightly better performance than that of the ReLU network, which verifies the observations of Maas et al. (2013); Pedamonti (2018).

1.3 Organizations

The rest of this paper is organized as follows. In Section 2, we introduce some notations, preliminary definitions, lemmas, and results. The relationships between the models (P) and (PP) are illustrated in Section 3. In Section 4, we propose an augmented Lagrangian method with the alternating minimization for solving problem (PP) and establish the global convergence of the algorithm. In Section 5, we illustrate the performance of our proposed algorithm through extensive numerical experiments. Concluding remarks are drawn in the last section.

2. Notations and Preliminaries

In this section, we introduce some notations, preliminary definitions, examples, and lemmas.

The $m \times m$ identity matrix is denoted by I_m. We use $\mathbb{N}_+$ to represent the set of positive integers. Given a point $z \in \mathbb{R}^m$ and $\epsilon > 0, \mathcal{B}_\epsilon(z)$ denotes a closed ball centered at z with radius ϵ, $(\mathrm{sign}(z))_i$ denotes the sign function of z_i, and $\mathrm{diag}(z)$ denotes the diagonal matrix whose diagonal vector is z. $\mathrm{dist}(z^*, \Omega) = \min_{z\in\Omega} \|z - z^*\|$ represents the distance from a point z^* to a nonempty closed set Ω. We use $\mathrm{int}(\Omega)$, $\mathrm{co}(\Omega)$ to represent the interior and convex hull of Ω, respectively. The indicator function of a set Ω is denoted by δ_Ω. The Hadamard product is denoted by $\circ$. We let $\nabla_{(z_1,z_2)} f(z) = \nabla_{z_1} f(z) \times \nabla_{z_2} f(z)$ for a smooth function f with respect to $z = (z_1, z_2)$.

Let H be a symmetric positive definite matrix, $\Omega \subseteq \mathbb{R}^m$ be a convex set, and $\mathrm{Proj}_{\Omega}^{H}(z^*) = \arg\min\{\|z - z^*\|_H : z \in \Omega\}$ be the orthogonal projection of a vector $z^* \in \mathbb{R}^m$ onto Ω (Facchinei and Pang, 2003). If H is the identity matrix, we will use $\mathrm{Proj}_{\Omega}(z^*)$ instead. The proximal mapping $\mathrm{Prox}_f(\cdot)$ of a proper closed convex function f is defined as $\mathrm{Prox}_f(z^*) = \arg\min_{z\in\mathbb{R}^n}\left\{\frac{1}{2}\|z - z^*\|^2 + f(z)\right\}$.

2.1 Subdifferentials and Stationarity

Let $f : \Omega \to \mathbb{R}$ be a locally Lipschitz continuous and directionally differentiable function defined on an open set $\Omega \subseteq \mathbb{R}^n$. The directional derivative of f at z along the direction d is defined as

$$f'(z; d) = \lim_{t\downarrow 0} \frac{f(z + td) - f(z)}{t}.$$

It is worth mentioning that any piecewise smooth and Lipschitz continuous function is directionally differentiable (Mifflin, 1977).

Let $\bar{z} \in \Omega$ be given. The Clarke subdifferential (Clarke, 1990) of f at $\bar{z}$ is defined by

$$\partial^c f(\bar{z}) = \mathrm{co}\left\{\lim_{z\to\bar{z}} \nabla f(z) : f \text{ is smooth at } z\right\}.$$

According to (Rockafellar and Wets, 1998, Definition 8.3), the limiting subdifferential of f at $\bar{z}$ is defined by $\partial f(\bar{z}) :=$

$$\left\{v : \exists z^k \xrightarrow{f} \bar{z}, v^k \to v \text{ such that } \liminf_{z\to z^k} \frac{f(z) - f\left(z^k\right) - \left\langle v^k, z - z^k\right\rangle}{\|z - z^k\|} \geq 0,\ \forall k\right\},$$

where $z^k \xrightarrow{f} \bar{z}$ means that $z^k \to \bar{z}$ and $f\left(z^k\right) \to f(\bar{z})$. If f is convex, then ∂f coincides with $\partial^c f$. If f is furthermore smooth, it holds that $\partial f(z) = \partial^c f(z) = \{\nabla f(z)\}$. In general, one has $\mathrm{co}(\partial f(\bar{z})) = \partial^c f(\bar{z})$.

For $z \in \mathbb{R}^n$, we have

$$\partial\|z\| = \partial^c\|z\| = \begin{cases} \frac{z}{\|z\|}, & \text{if } \|z\| \neq 0, \\ \{r : r \in \mathbb{R}^n, \|r\| \leq 1\}, & \text{if } \|z\| = 0. \end{cases}$$

Let $\mathcal{T}_\Omega(\bar{z}) = \left\{d : d = \lim_{z\in\Omega, z\to\bar{z}, t\downarrow 0} \frac{z-\bar{z}}{t}\right\}$ be the tangent cone of a set Ω at $\bar{z} \in \Omega$ and $\mathcal{N}_\Omega(z)$ be the limiting normal cone at $z \in \Omega$. If Ω is a convex set, then $\mathcal{N}_\Omega(z)$ coincides with the classical (Clarke) normal cone in the convex analysis, where the Clarke normal cone $\mathcal{N}_\Omega^c(z)$ is defined by $\mathcal{N}_\Omega^c(z) = \mathrm{clco}\mathcal{N}_\Omega(z)$.

Definition 1 *Let $\mathcal{Z}$ be a closed set in Ω. We call $\bar{z} \in \mathcal{Z}$ a d(irectional)-stationary point of $\min_{z\in\mathcal{Z}} f(z)$ if $f'(\bar{z}; d) \geq 0$ for all $d \in \mathcal{T}_{\mathcal{Z}}(\bar{z})$. We say that a point $\bar{z} \in \mathcal{Z}$ is a limiting stationary point, a C(larke)-stationary point of $\min_{z\in\mathcal{Z}} f(z)$ if $0 \in \partial f(\bar{z}) + \mathcal{N}_{\mathcal{Z}}(\bar{z})$, $0 \in \partial^c f(\bar{z}) + \mathcal{N}_{\mathcal{Z}}^c(\bar{z})$, respectively.*

Based on Definition 1, we have the following relationships

$$\text{local minimizer} \ \Rightarrow\ \text{d-stationary} \ \Rightarrow\ \text{limiting stationary} \ \Rightarrow\ \text{C-stationary}. \tag{4}$$

Furthermore, $0 \in \partial^c f(\bar{z}) + \mathcal{N}^c_{\mathcal{Z}}(\bar{z})$ implies

$$f^\circ(\bar{z}; d) := \limsup_{z \to \bar{z}, t \downarrow 0} \frac{f(z + td) - f(z)}{t} \geq 0, \text{ for all } d \in \mathcal{T}_{\mathcal{Z}}(\bar{z}).$$

If a certain constraint qualification condition (see Subsection 2.2) holds at $\bar{z} \in \mathcal{Z}$, then $\bar{z}$ being a limiting stationary point is a necessary condition for $\bar{z}$ to be a local minimizer of f (see an example given by Chen et al. (2017)).

In general, a C-stationary point is not a good candidate for a local minimizer. We end this subsection with an example on the DNN to illustrate that a C-stationary point may not be a limiting stationary point, which further may not be a local minimizer.

Example 1 *Consider*

$$\min_{w_1 \in \mathbb{R}, w_2 \in \mathbb{R}, b_1 \in \mathbb{R}, b_2 \in \mathbb{R}} \left((w_2 \sigma\left(w_1 + b_1\right) + b_2) + 1\right)^2 + \left((w_2 \sigma\left(2w_1 + b_1\right) + b_2) - 1\right)^2. \quad (5)$$

Let $f(w_1, w_2, b_1, b_2)$ be the objective function of (5), $w_2^* = 1$, $b_1^* = 0$, $w_1^* = 0$, $b_2^* = 0$, we have

$$\partial^c f(w_1^*, w_2^*, b_1^*, b_2^*) = \left\{(t, 0, s, 0)^T : t \in [2\alpha - 4, 2 - 4\alpha], s \in [-2 + 2\alpha, 2 - 2\alpha]\right\},$$
$$\partial\left(f(w_1^*, w_2^*, b_1^*, b_2^*)\right)$$
$$= \left\{(-2\alpha, 0, 0, 0)^\top, (2\alpha - 4, 0, 2\alpha - 2, 0)^\top, (2 - 4\alpha, 0, 2 - 2\alpha, 0)^\top, (-2, 0, 0, 0)^\top\right\},$$
$$f(w_1^* + \epsilon, w_2^*, b_1^*, b_2^*) = 5\epsilon^2 - 2\epsilon + 2 < 2 = f(w_1^*, w_2^*, b_1^*, b_2^*) \text{ for some small positive number } \epsilon.$$

For some $0 < \alpha < \frac{1}{2}$, $(w_1^*, w_2^*, b_1^*, b_2^*)$ is a C-stationary point of (5), but it is neither a local minimizer nor a limiting stationary point of (5). Moreover, one can see that $(1, 2, -1, -1)$ is a global minimizer of (5), at which the function value is 0.

2.2 Necessary Optimality Conditions.

In this subsection, we provide first order necessary optimality conditions for local minimizers of problems (P) and (PP), respectively. Let

$$\mathcal{C}(v, u) := \begin{pmatrix} u - v \\ \alpha u - v \end{pmatrix}. \quad (6)$$

Definition 2 *We say that (w^*, b^*, v^*, u^*) is a KKT point of problem (PP) if there exist vectors $\mu \in \mathbb{R}^{2m}_+$ and $\xi \in \mathbb{R}^m$ such that*

$$0 = \nabla_w \bar{\mathcal{O}}(w^*, v^*) - \Psi(v^*)^\top \xi, \quad 0 = A^\top \xi, \quad (7)$$
$$0 = \nabla_v \bar{\mathcal{O}}(w^*, v^*) + \beta + \nabla_v \mu^\top \mathcal{C}(v^*, u^*) - \nabla_v \xi^\top \Psi(v^*) w^*, \quad (8)$$
$$0 \in \partial_u(-\beta^\top \sigma(u^*)) + \nabla_u \mu^\top \mathcal{C}(v^*, u^*) + \xi, \quad (9)$$
$$\mathcal{C}(v^*, u^*) \leq 0, \ \mu^\top \mathcal{C}(v^*, u^*) = 0, \ u^* - \Psi(v^*) w^* - A b^* = 0. \quad (10)$$

Since $v - \sigma(u) = 0$ can be written as the following complementarity problem

$$v - u \geq 0, (v - u)(v - \alpha u) = 0, v - \alpha u \geq 0,$$

we can define the Mathematical Programming with Complementarity Constraints (MPCC) W(eakly)-stationary point (Scheel and Scholtes, 2000; Guo and Chen, 2021) of problem (P) as follows.

Definition 3 *We say that $(w^*, b^*, v^*, u^*) \in \Omega_1$ is an MPCC W-stationary point of problem* (P) *if there exist vectors $\mu^1 \in \mathbb{R}^m$, $\mu^2 \in \mathbb{R}^m$ and $\xi \in \mathbb{R}^m$ such that*

$$0 = \nabla_w \bar{\mathcal{O}}(w^*, v^*) - \Psi(v^*)^\top \xi, \quad 0 = A^\top \xi, \tag{11}$$

$$0 = \nabla_v \bar{\mathcal{O}}(w^*, v^*) - \mu^1 - \mu^2 - \nabla_v \xi^\top \Psi(v^*) w^*, \tag{12}$$

$$0 = \mu^1 + \alpha \mu^2 + \xi, \tag{13}$$

$$\left(\mu^1\right)^\top (v^* - u^*) = 0, \ \left(\mu^2\right)^\top (v^* - \alpha u^*) = 0. \tag{14}$$

We say $(w^*, b^*, v^*, u^*) \in \Omega_1$ is an MPCC C(larke)-stationary point of problem (P)*, if it is an MPCC W-stationary point of problem* (P) *and $\mu_i^1 \mu_i^2 \geq 0$ for $u_i^* = v_i^* = 0$.*

Mangasarian-Fromovitz Constraint Qualification (MFCQ) (Mangasarian, 1994) and Linear Independence Constraint Qualification (LICQ) (Dimitri, 1997) are constraint qualifications that are widely used in mathematical programming to characterize the behavior of constraints at a particular point. Both of them are necessary conditions for optimality in constrained optimization problems. MFCQ is often considered more practical and widely used in optimization algorithms because it is less restrictive than LICQ. Notice that LICQ does not hold for problem (PP), as it requires all the gradients of the constraints to be linearly independent. Therefore, MFCQ is more appropriate in our cases.

To ensure that a local minimizer of problem (PP) is a KKT point, the following Lemma shows that the feasible set of problem (PP) satisfies the MFCQ.

Lemma 4 *The MFCQ holds at $(w^*, b^*, v^*, u^*) \in \Omega_2$ for problem* (PP)*, i.e., there exist no nonzero vectors $\xi \in \mathbb{R}^m, \mu \in \mathbb{R}_+^{2m}$ such that $\mu^\top \mathcal{C}(v^*, u^*) = 0$ and*

$$0 = \Psi(v^*)^\top \xi, \quad 0 = A^\top \xi, \tag{15}$$

$$0 = \nabla_v \mu^\top \mathcal{C}(v^*, u^*) - \nabla_v \xi^\top \Psi(v^*) w^*, \tag{16}$$

$$0 = \nabla_u \mu^\top \mathcal{C}(v^*, u^*) + \xi. \tag{17}$$

Proof We prove that the linear system (15)–(17) only has a zero solution.

Let $\mu = ((\mu_{1,1}^1)^\top, (\mu_{2,1}^1)^\top, \ldots, (\mu_{1,L}^1)^\top, \ldots, (\mu_{N,L}^1)^\top, (\mu_{1,1}^2)^\top, \ldots, (\mu_{N,L}^2)^\top)^\top$ and $\xi = (\xi_{1,1}^\top, \xi_{2,1}^\top, \ldots, \xi_{N,1}^\top, \xi_{1,2}^\top, \ldots, \xi_{N,L}^\top)^\top$, where $\mu_{n,\ell}^1, \mu_{n,\ell}^2 \in \mathbb{R}_+^{N_\ell}$, $\xi_{n,\ell} \in \mathbb{R}^{N_\ell}$ for all $n \in [N]$ and $\ell \in [L]$. Notice that $u^* = \Psi(v^*) w^* + A b^*$ is equivalent to $u_{n,\ell}^* - (W_\ell^* v_{n,\ell-1}^* + b_\ell^*) = 0$ for all $n \in [N]$ and $\ell \in [L]$, the equalities (16) and (17) yield

$$0 = \nabla_v \left(\sum_{n=1}^{N} \sum_{\ell=1}^{L} \left(\mu_{n,\ell}^1\right)^\top v_{n,\ell}^* + \left(\mu_{n,\ell}^2\right)^\top v_{n,\ell}^* + \xi_{n,\ell}^\top W_\ell^* v_{n,\ell-1}^* \right), \tag{18}$$

$$0 = \nabla_u \left(\sum_{n=1}^{N} \sum_{\ell=1}^{L} \left(\mu_{n,\ell}^1\right)^\top u_{n,\ell}^* + \alpha \left(\mu_{n,\ell}^2\right)^\top u_{n,\ell}^* + \xi_{n,\ell}^\top u_{n,\ell}^* \right). \tag{19}$$

We first consider the coefficient with respect to $v_{n,L}^*$ in (18) for all $n \in [N]$, which yields $0 = -(\mu_{n,L}^1 + \mu_{n,L}^2)$. Together with the inequalities $\mu_{n,L}^1 \geq 0$ and $\mu_{n,L}^2 \geq 0$, we obtain that $\mu_{n,L}^1 = \mu_{n,L}^2 = 0$ for all $n \in [N]$. We then consider the coefficient with respect to $u_{n,L}^*$ in (19) for all $n \in [N]$, which implies that $0 = \mu_{n,L}^1 + \alpha\mu_{n,L}^2 + \xi_{n,L}$. Hence, we have $\xi_{n,L} = 0$ for all $n \in [N]$. Substituting $\xi_{n,L} = \mu_{n,L}^1 = \mu_{n,L}^2 = 0$ for all $n \in [N]$ into (18) and (19), we obtain that

$$\begin{aligned}
0 &= \nabla_v \left(\sum_{n=1}^{N} \sum_{\ell=1}^{L-1} \left(\mu_{n,\ell}^1\right)^\top v_{n,\ell}^* + \left(\mu_{n,\ell}^2\right)^\top v_{n,\ell}^* + \xi_{n,\ell}^\top W_\ell^* v_{n,\ell-1}^* \right), \\
0 &= \nabla_u \left(\sum_{n=1}^{N} \sum_{\ell=1}^{L-1} \left(\mu_{n,\ell}^1\right)^\top u_{n,\ell}^* + \alpha \left(\mu_{n,\ell}^2\right)^\top u_{n,\ell}^* + \xi_{n,\ell}^\top u_{n,\ell}^* \right).
\end{aligned}$$

Then, we obtain that $\xi_{n,\ell} = \mu_{n,\ell}^1 = \mu_{n,\ell}^2 = 0$ for all $n \in [N]$ and $\ell \in [L]$ by mathematical induction. This completes the proof. ■

Under MFCQ, we can obtain the following equivalence between the limiting stationary points and KKT points of problem (PP).

Theorem 5 *(w^*, b^*, v^*, u^*) is a limiting stationary point of problem* (PP) *if and only if (w^*, b^*, v^*, u^*) is a KKT point of problem* (PP)*.*

Proof Since the MFCQ holds at $(w^*, b^*, v^*, u^*) \in \Omega_2$ for problem (PP), then (Rockafellar and Wets, 1998, Theorem 6.14) yields that $\mathcal{N}_{\Omega_2}(w^*, b^*, v^*, u^*)$ equals to

$$\left\{ \nabla \left(\mu^\top \mathcal{C}(v^*, u^*) + \xi^\top (u^* - \Psi(v^*)w^* - Ab^*) \right) : \ \mu^\top \mathcal{C}(v^*, u^*) = 0, \ \mu \in \mathbb{R}_+^{2m}, \ \xi \in \mathbb{R}^m \right\}.$$

If (w^*, b^*, v^*, u^*) is a limiting stationary point of problem (PP), then there exist vectors $\mu^1, \mu^2 \in \mathbb{R}_+^m$ and $\xi \in \mathbb{R}^m$ such that $[(\mu^1)^\top, (\mu^2)^\top]\mathcal{C}(v^*, u^*) = 0$ and

$$\begin{aligned}
&0 \in \partial \mathcal{O}(w^*, v^*, u^*) + \nabla \left(\left(\mu^1\right)^\top (u^* - v^*) + \left(\mu^2\right)^\top (\alpha u^* - v^*) + \xi^\top (u^* - \Psi(v^*)w^*) \right), \\
&0 = A^\top \xi.
\end{aligned} \tag{20}$$

Recall the definition of $\mathcal{O}$, the relationships (7)–(10) hold with $\mu = [(\mu^1)^\top, (\mu^2)^\top]^\top$. Hence (w^*, b^*, v^*, u^*) is a KKT point of problem (PP).

Conversely, if (w^*, b^*, v^*, u^*) is a KKT point of problem (PP), then there exist vectors $\mu \in \mathbb{R}_+^{2m}$ and $\xi \in \mathbb{R}^m$ such that (7)–(10) hold. Let $\mu = [(\mu^1)^\top, (\mu^2)^\top]^\top$ with $\mu^1, \mu^2 \in \mathbb{R}_+^m$. Since the MFCQ holds at $(w^*, b^*, v^*, u^*) \in \Omega_2$ for problem (PP), we then obtain (20) by the expression of $\mathcal{N}_{\Omega_2}(w^*, b^*, v^*, u^*)$. This completes the proof. ■

Since there are nonsmooth constraints in problem (P), then MFCQ does not hold for problem (P). To ensure that a local minimizer of problem (P) is also an MPCC C-stationary

point, the following lemma shows that the feasible set of problem (P) satisfies the No Nonzero Abnormal Multiplier Constraint Qualification (NNAMCQ) (Ye and Zhang, 2013). It is worth noting that the MPCC linear independent CQ (Scheel and Scholtes, 2000; Guo and Chen, 2021) does not hold for problem (P).

Lemma 6 *The NNAMCQ holds at* $(w^*, b^*, v^*, u^*) \in \Omega_1$ *for problem* (P)*, i.e., there exist no nonzero vectors* $\mu \in \mathbb{R}^m,\ \xi \in \mathbb{R}^m$ *such that*

$$0 \in \partial_{(v,u)} \mu^\top (\sigma(u^*) - v^*) + \nabla_{(v,u)} \xi^\top (u^* - \Psi(v^*) w^*),\ 0 = \Psi(v^*)^\top \xi,\ 0 = A^\top \xi. \tag{21}$$

Proof We prove that there exist no nonzero vectors $\xi \in \mathbb{R}^m, \mu \in \mathbb{R}^m$ such that

$$0 = -\mu + \nabla_v \xi^\top (u^* - \Psi(v^*) w^* - A b^*), \tag{22}$$

$$0 \in \partial_u \mu^\top \sigma(u^*) + \xi, \tag{23}$$

since it implies that there is no nonzero vectors $\mu \in \mathbb{R}^m,\ \xi \in \mathbb{R}^m$ satisfying (21).

Notice that $u^* = \Psi(v^*) w^* + A b^*$ is equivalent to $u^*_{n,\ell} - (W^*_\ell v^*_{n,\ell-1} + b^*_\ell) = 0$ for all $n \in [N]$ and $\ell \in [L]$, the equality (22) yields

$$0 = -\mu + \nabla_v \left(\sum_{n=1}^{N} \sum_{\ell=1}^{L} \xi_{n,\ell}^\top W^*_\ell v^*_{n,\ell-1} \right). \tag{24}$$

We first consider the coefficient with respect to $v^*_{n,L}$ in (24) for all $n \in [N]$, which yields $0 = -\mu_{n,L}$. We then consider the coefficient with respect to $u^*_{n,L}$ in (23) for all $n \in [N]$, which implies $0 \in \mu_{n,L}[\alpha, 1] + \xi_{n,L}$. Hence, we have $\xi_{n,L} = 0$ for all $n \in [N]$. Substituting $\xi_{n,L} = \mu_{n,L} = 0$ for all $n \in [N]$ into (24), we obtain that

$$0 = -\mu + \nabla_v \left(\sum_{n=1}^{N} \sum_{\ell=1}^{L-1} \xi_{n,\ell}^\top W^*_\ell v^*_{n,\ell-1} \right).$$

Then, we can derive $\xi_{n,\ell} = \mu_{n,\ell} = 0$ for all $n \in [N]$ and $\ell \in [L]$ by mathematical induction. This completes the proof. ■

We end this section with a lemma illustrating that the MPCC C-stationary point of problem (P) are necessary to be a local minimizer of problem (P).

Lemma 7 *If* (w^*, b^*, v^*, u^*) *is a local minimizer of problem* (P)*, then* (w^*, b^*, v^*, u^*) *is also an MPCC C-stationary point of problem* (P)*.*

Proof Since (w^*, b^*, v^*, u^*) is a local minimizer of problem (P), Lemma 6 yields that there exist vectors $\mu \in \mathbb{R}^m$ and $\xi \in \mathbb{R}^m$ satisfying

$$\begin{aligned} &0 \in \partial \left(\mu^\top (\sigma(u^*) - v^*) \right) + \nabla_{(v,u)} \left(\bar{\mathcal{O}}(w^*, v^*) + \xi^\top (u^* - \Psi(v^*) w^*) \right), \\ &0 = \nabla_w \bar{\mathcal{O}}(w^*, v^*) + \Psi(v^*)^\top \xi, 0 = A^\top \xi. \end{aligned} \tag{25}$$

Notice that $\mu^\top \sigma(u^*)$ is the only nonsmooth term in (25). We now analyze the following three cases for all $i \in [m]$.

Case (i): if $u_i^* > 0$, we have $v_i^* = u_i^*$ and $\partial\mu_i\sigma(u_i^*) = \{\mu_i\}$. Let $\mu_i^1 = \mu_i$ and $\mu_i^2 = 0$, then $\mu_i^1(v_i^* - u_i^*) = 0$ and $\mu_i^2(v_i^* - \alpha u_i^*) = 0$.

Case (ii): if $u_i^* < 0$, we have $v_i^* = \alpha u_i^*$ and $\partial\mu_i\sigma(u_i^*) = \{\alpha\mu_i\}$. Let $\mu_i^1 = 0$ and $\mu_i^2 = \mu_i$, then $\mu_i^1(v_i^* - u_i^*) = 0$ and $\mu_i^2(v_i^* - \alpha u_i^*) = 0$.

Case (iii): if $u_i^* = 0$, we have $v_i^* = 0$ and

$$\partial\mu_i\sigma(u_i^*) \subset [\alpha, 1]\mu_i = \left\{\mu_i^1 + \alpha\mu_i^2 : \mu_i^1 = t_i\mu_i, \mu_i^2 = (1 - t_i)\mu_i, t_i \in [0, 1]\right\}.$$

In this case, we have $\mu_i^1(v_i^* - u_i^*) = 0$, $\mu_i^2(v_i^* - \alpha u_i^*) = 0$ and $\mu_i^1\mu_i^2 \geq 0$.

Combining the above three cases, it holds that

$$\begin{aligned}\partial\mu^\top\sigma(u^*) \subset \Big\{&\mu^1 + \alpha\mu^2 : \mu^1 = t \circ \mu, \mu^2 = (e_m - t) \circ \mu, \\ &\left(\mu^1\right)^\top (v^* - u^*) = 0, \left(\mu^2\right)^\top (v^* - \alpha u^*) = 0, t \in \mathbb{R}_+^m, t \leq e_m\Big\}.\end{aligned} \tag{26}$$

Together with the inclusion $0 \in \partial\mu^\top\sigma(u^*) + \xi$, we have

$$\begin{aligned}0 \in \Big\{&\mu^1 + \alpha\mu^2 + \xi : \mu^1 = t \circ \mu, \mu^2 = (e_m - t) \circ \mu, \\ &\left(\mu^1\right)^\top (v^* - u^*) = 0, \left(\mu^2\right)^\top (v^* - \alpha u^*) = 0, t \in \mathbb{R}_+^m, t \leq e_m\Big\}.\end{aligned}$$

Hence there exist $\bar{\mu}^1$ and $\bar{\mu}^2$ such that

$$0 = \bar{\mu}^1 + \alpha\bar{\mu}^2 + \xi, \left(\bar{\mu}^1\right)^\top (v^* - u^*) = 0, \left(\bar{\mu}^2\right)^\top (v^* - \alpha u^*) = 0, \tag{27}$$

$$\bar{\mu}^1 = t \circ \mu, \bar{\mu}^2 = (e_m - t) \circ \mu, \bar{\mu}^1 \circ \bar{\mu}^2 \geq 0, t \in \mathbb{R}_+^m, t \leq e_m. \tag{28}$$

Combining (25), (27), (28) and $\sigma(u^*) - v^* = 0$, we obtain (11)–(14), and $\bar{\mu}^1 \circ \bar{\mu}^2 \geq 0$ with $\bar{\mu}^1, \bar{\mu}^2$ instead of μ^1, μ^2, respectively. This completes the proof. ■

3. Model Analysis

In this section, we aim to theoretically investigate the relationship between problems (P) and (PP).

3.1 The Existence and Boundedness of the Solution Set

In this subsection, we show that the solution set of problem (PP) is not empty and bounded. First, we define a level set Ω_θ of the objective function of problem (PP) by

$$\Omega_\theta = \{(w, b, v, u) \in \Omega_2 : \mathcal{O}(w, v, u) \leq \theta\} \text{ with } \theta > \frac{1}{N}\|Y\|_F^2, \tag{29}$$

where $Y = (y_1, y_2, \ldots, y_N)$ is the label matrix. Clearly $0 \in \Omega_\theta$. For all $(w, b, v, u) \in \Omega_\theta$, it holds that $\mathcal{R}_1(w) = \lambda_w \sum_{\ell=1}^L \|W_\ell\|_{2,1} \leq \theta$ and $\mathcal{R}_2(v) = \lambda_v\|v\|^2 \leq \theta$, which further implies that $\|w\|$ and $\|v\|$ are bounded. For brevity, we let $\theta_w := \frac{\theta}{\lambda_w}\sqrt{\overline{N} + N_0}$, $\theta_v := \sqrt{\frac{\theta}{\lambda_v}}$ be the upper bounds of $\|w\|$ and $\|v\|$ over the set Ω_θ, respectively.

Theorem 8 *The set Ω_θ is bounded. Furthermore, the solution set of problem* (PP) *is not empty and bounded.*

Proof Since $(w, b, v, u) \in \Omega_2$, it holds that $b_\ell \le v_{n,\ell} - W_\ell v_{n,\ell-1}$ for all $\ell \in [L]$ and $n \in [N]$. Together with the fact that θ_w, θ_v being the upper bounds of $\|w\|$ and $\|v\|$ over the set Ω_θ, respectively, we obtain that $\|b_+\|_\infty$ is bounded.

Since $\mathcal{O}(w, v, u) \le \theta$, $\mathcal{R}_1(w)$, $\mathcal{R}_2(v)$ are nonnegative, then for all $n \in [N]$, $\ell \in [L]$, $j \in [N_\ell]$, it holds that

$$(v_{n,\ell})_j - \sigma(b_{\ell,j} + W_{\ell,j} v_{n,\ell-1}) \le \frac{\theta}{\beta_\ell},$$

which further implies that either $(v_{n,\ell})_j - \frac{\theta}{\beta_\ell} \le b_{\ell,j} + W_{\ell,j} v_{n,\ell-1}$ or $(v_{n,\ell})_j - \frac{\theta}{\beta_\ell} \le \alpha(b_{\ell,j} + W_{\ell,j} v_{n,\ell-1})$. By $(w, b, v, u) \in \Omega_\theta$, the definition of θ_w and θ_v, we have $\frac{1}{\alpha}((v_{n,\ell})_j - \frac{\theta}{\beta_\ell}) - W_{\ell,j} v_{n,\ell-1} > -\frac{1}{\alpha}(\theta_v + \frac{\theta}{\beta_\ell}) - \overline{N}\theta_v\theta_w$ and $(v_{n,\ell})_j - \frac{\theta}{\beta_\ell} - W_{\ell,j} v_{n,\ell-1} > -\theta_v - \frac{\theta}{\beta_\ell} - \overline{N}\theta_v\theta_w$ for all $\ell \in [L]$ and $n \in [N]$. Since $0 < \alpha < 1$, it holds that

$$\begin{aligned} b_{\ell,j} &\ge \min\left\{\frac{1}{\alpha}\left((v_{n,\ell})_j - \frac{\theta}{\beta_\ell}\right) - W_{\ell,j} v_{n,\ell-1}, (v_{n,\ell})_j - \frac{\theta}{\beta_\ell} - W_{\ell,j} v_{n,\ell-1}\right\} \\ &> -\frac{1}{\alpha}\left(\theta_v + \frac{\theta}{\beta_\ell}\right) - \overline{N}\theta_v\theta_w. \end{aligned} \tag{30}$$

It follows from the boundedness of $\|b_+\|_\infty$ that $\|b\|$ is also bounded. Hence $\|u\| = \|\Psi(v)w + Ab\|$ is bounded, too. These facts imply that Ω_θ is a bounded set. Together with the inclusion $0 \in \Omega_\theta$ and the continuity of the objective function of problem (PP), we obtain that the solution set of problem (PP) is nonempty and bounded. ■

3.2 Exact Penalization

In this subsection, we consider problem (PP) with penalty parameter β satisfying

$$\beta_\ell > L L_{\bar{\mathcal{O}}} \max\{\theta_w, 1\}^L + 2 \sum_{j=\ell+1}^{L} \beta_j \theta_w \max\{\theta_w, 1\}^{j-\ell-1} \text{ for all } \ell \in [L], \tag{31}$$

and reveal the relationship between problems (PP) and (P), where $L_{\bar{\mathcal{O}}}$ is the Lipschitz constant of the function $\bar{\mathcal{O}}$ over Ω_θ. We first present a lemma, which shows that any limiting stationary point of problem (PP) is in the feasible set Ω_1 of problem (P).

Lemma 9 *Let the penalty parameter β satisfy* (31). *If $(w^*, b^*, v^*, u^*) \in \{(w, b, v, u) \in \Omega_2 : \mathcal{O}(w, v, u) < \theta\}$ is a limiting stationary point of problem* (PP), *then (w^*, b^*, v^*, u^*) is in the feasible set Ω_1 of problem* (P).

Proof If (w^*, b^*, v^*, u^*) is a limiting stationary point of problem (PP), then (w^*, b^*, v^*, u^*) is a C-stationary point of problem (PP), which implies that

$$\mathcal{O}^\circ(w^*, v^*, u^*; d_w, d_v, d_u) \ge 0, \text{ for all } (d_w, d_v, d_u, d_b) \in \mathcal{T}_{\Omega_2}(w^*, b^*, v^*, u^*). \tag{32}$$

We then prove $v^*_{n,\ell} = \sigma(u^*_{n,\ell})$ for all $n \in [N]$ and $\ell = L, L-1, \ldots, 1$ by mathematical induction.

Assume on contradiction that $(w^*, b^*, v^*, u^*) \notin \Omega_1$. When ℓ equals L, let $\mathcal{I}_L := \{n : v^*_{n,L} \geq \sigma(u^*_{n,L}), v^*_{n,L} \neq \sigma(u^*_{n,L})\}$. Without loss of generality, we assume that $\mathcal{I}_L$ is not an empty set. We set

$$d_v = \left((d_v)^\top_{1,1}, (d_v)^\top_{2,1}, \ldots, (d_v)^\top_{N,1}, \ldots, (d_v)^\top_{1,L}, (d_v)^\top_{2,L} \ldots, (d_v)^\top_{N,L}\right)^\top$$

for all $n \in [N]$, $\ell \in [L]$, $(d_v)_{n,\ell} \in \mathbb{R}^{N_\ell}$ and

$$(d_v)_{n,\ell} = \begin{cases} 0, & \text{if } \ell < L \text{ or } n \notin \mathcal{I}_L, \\ \sigma(u^*_{n,\ell}) - v^*_{n,\ell}, & \text{if } \ell = L \text{ and } n \in \mathcal{I}_L. \end{cases}$$

Clearly $d_v \leq 0$ and $(w^*, b^*, v^* + td_v, u^*) \in \Omega_2$ for all $0 \leq t < 1$. Hence $(0, 0, d_v, 0) \in \mathcal{T}_{\Omega_2}(w^*, b^*, v^*, u^*)$.

Since $\mathcal{O}(w^*, v^*, u^*) < \theta$ and $\mathcal{O}$ is locally Lipschitz continuous, there exists $\bar{\epsilon} \in (0, 1]$ such that $\mathcal{O}(w, v, u) < \theta$ for all $(w, v, u) \in \mathcal{B}_{\bar{\epsilon}}(w^*, v^*, u^*)$. Furthermore, for any $(w, v, u) \in \mathcal{B}_{\bar{\epsilon}}(w^*, v^*, u^*)$, there exists $\bar{t} \in (0, 1]$ such that $\mathcal{O}(w, v + td_v, u) < \theta$ for all $0 < t < \bar{t}$.

Together with the inequalities $d_v \leq 0$ and $\bar{\mathcal{O}}(w, v + td_v) < \mathcal{O}(w, v + td_v, u) < \theta$, it holds that

$$\begin{aligned} &\frac{1}{t}\left(\mathcal{O}(w, v + td_v, u) - \mathcal{O}(w, v, u)\right) \\ =&\frac{1}{t}\left(\bar{\mathcal{O}}(w, v + td_v) + \beta^\top(v + td_v - \sigma(u)) - \bar{\mathcal{O}}(w, v) - \beta^\top(v - \sigma(u))\right) \\ =&\frac{1}{t}\left(\bar{\mathcal{O}}(w, v + td_v) - \bar{\mathcal{O}}(w, v)\right) + \beta^\top d_v \leq (L_{\bar{\mathcal{O}}} - \beta_L) \sum_{n \in \mathcal{I}_L} \left\|v^*_{n,L} - \sigma(u^*_{n,L})\right\|_1. \end{aligned} \tag{33}$$

Hence we derive

$$\mathcal{O}^\circ(w^*, v^*, u^*; 0, d_v, 0) \leq \limsup_{(w,v,u) \to (w^*, v^*, u^*), t \downarrow 0} (L_{\bar{\mathcal{O}}} - \beta_L) \sum_{n \in \mathcal{I}_L} \left\|v^*_{n,L} - \sigma(u^*_{n,L})\right\|_1 < 0.$$

This leads to a contradiction. Hence, it holds that $v^*_{n,L} = \sigma(u^*_{n,L})$ for all $n \in [N]$.

Then, we suppose that $v^*_{n,\ell} = \sigma(u^*_{n,\ell})$ for all $n \in [N]$ and $\ell = L, L-1, \ldots, \bar{\ell}+1$. Let $\mathcal{I}_{\bar{\ell}} := \{n : v^*_{n,\bar{\ell}} \geq \sigma(u^*_{n,\bar{\ell}}), v^*_{n,\bar{\ell}} \neq \sigma(u^*_{n,\bar{\ell}})\}$. Without loss of generality, we suppose that $\mathcal{I}_{\bar{\ell}}$ is not an empty set. Then for all $n \in [N]$ and $\ell \in [L]$, we set

$$\begin{cases} \tilde{v}^\epsilon_{n,\ell} = v^*_{n,\ell},\ \tilde{u}^\epsilon_{n,\ell} = u^*_{n,\ell}, & \text{if } \ell < \bar{\ell} \text{ or } \ell = \bar{\ell}, n \notin \mathcal{I}_{\bar{\ell}}, \\ \tilde{v}^\epsilon_{n,\ell} = \epsilon\sigma(u^*_{n,\ell}) + (1-\epsilon)v^*_{n,\ell},\ \tilde{u}^\epsilon_{n,\ell} = u^*_{n,\ell}, & \text{if } \ell = \bar{\ell} \text{ and } n \in \mathcal{I}_{\bar{\ell}}, \\ \tilde{v}^\epsilon_{n,\ell} = \sigma(\tilde{u}^\epsilon_{n,\ell}),\ \tilde{u}^\epsilon_{n,\ell} = w^*_\ell \tilde{v}^\epsilon_{n,\ell-1} + b^*_\ell, & \text{if } \ell > \bar{\ell}. \end{cases} \tag{34}$$

Clearly $(w^*, b^*, \tilde{v}^\epsilon, \tilde{u}^\epsilon) \in \Omega_2$, $\lim_{\epsilon \downarrow 0}(\tilde{v}^\epsilon - v^*)/\epsilon$ and $\lim_{\epsilon \downarrow 0}(\tilde{u}^\epsilon - u^*)/\epsilon$ exist. Let

$$d_v = \lim_{\epsilon_k \downarrow 0} d_v^{(k)},\ d_v^{(k)} = \frac{\tilde{v}^{\epsilon_k} - v^*}{\epsilon_k},\ d_u = \lim_{\epsilon_k \downarrow 0} d_u^{(k)},\ d_u^{(k)} = \frac{\tilde{u}^{\epsilon_k} - u^*}{\epsilon_k}, \tag{35}$$

then we have $(0,0,d_v,d_u) \in \mathcal{T}_{\Omega_2}(w^*,b^*,v^*,u^*)$. Besides, it follows from (35), the Lipschitz continuity and directional differentiability of $\mathcal{O}$ that

$$\begin{aligned}
&\limsup_{(w,v,u)\to(w^*,v^*,u^*),t\downarrow 0} \frac{1}{t}\left(\mathcal{O}(w,v+td_v,u+td_u)-\mathcal{O}(w,v,u)\right)\\
=&\limsup_{(w,v,u)\to(w^*,v^*,u^*),\epsilon_k\downarrow 0} \frac{1}{\epsilon_k}\left(\mathcal{O}\left(w,v+\epsilon_k d_v,u+\epsilon_k d_u\right)-\mathcal{O}(w,v,u)\right)\\
=&\limsup_{(w,v,u)\to(w^*,v^*,u^*),\epsilon_k\downarrow 0} \frac{1}{\epsilon_k}\left(\mathcal{O}\left(w,v+\epsilon_k d_v^{(k)},u+\epsilon_k d_u^{(k)}\right)-\mathcal{O}(w,v,u)\right)\\
&\qquad -\frac{1}{\epsilon_k}\left(\mathcal{O}\left(w,v+\epsilon_k d_v^{(k)},u+\epsilon_k d_u^{(k)}\right)-\mathcal{O}(w,v+\epsilon_k d_v,u+\epsilon_k d_u)\right)\\
=&\limsup_{(w,v,u)\to(w^*,v^*,u^*),\epsilon_k\downarrow 0} \frac{1}{\epsilon_k}\left(\mathcal{O}\left(w,v+\epsilon_k d_v^{(k)},u+\epsilon_k d_u^{(k)}\right)-\mathcal{O}(w,v,u)\right).
\end{aligned} \tag{36}$$

Since $\mathcal{O}(w^*,v^*,u^*)<\theta$ and $\mathcal{O}$ is locally Lipschitz continuous, there exists $\bar{\epsilon}\in(0,1]$ such that $\mathcal{O}(w,v,u)<\theta$ for all $(w,v,u)\in\mathcal{B}_{\bar{\epsilon}}(w^*,v^*,u^*)$. Furthermore, for any $(w,v,u)\in\mathcal{B}_{\bar{\epsilon}}(w^*,v^*,u^*)$, there exists $\bar{t}\in(0,1]$ such that $\mathcal{O}(w,v+td_v,u+td_u)<\theta$ for all $0<t<\bar{t}$. Together with the equalities (35), there exists $\tilde{\epsilon}\in(0,1]$ such that $\mathcal{O}(w,v+\epsilon d_v^{(k)},u+\epsilon d_u^{(k)})<\theta$ for all $k\in\mathbb{N}$ and $0<\epsilon<\tilde{\epsilon}$. Without loss of generality, we assume that $\epsilon_k<\tilde{\epsilon}$ for all $k\in\mathbb{N}$.

Recall the definition of θ_w, d_u, d_v and $v^*_{n,\ell}=\sigma(W^*_\ell v^*_{n,\ell-1}+b^*_\ell)$ for all $\ell>\bar{\ell}$ and $n\in[N]$, we obtain that $\|(d_v^{(k)})_{n,\ell}\|_1\le\theta_w\|(d_v^{(k)})_{n,\ell-1}\|_1$ and $\|(d_u^{(k)})_{n,\ell}\|_1\le\theta_w\|(d_v^{(k)})_{n,\ell-1}\|_1$ for all $k\in\mathbb{N}$, $\ell>\bar{\ell}$ and $n\in[N]$. Hence for all $k\in\mathbb{N}$, we have

$$\begin{aligned}
\max\left\{\left\|d_v^{(k)}\right\|_1,\left\|d_u^{(k)}\right\|_1\right\}&\le L\max\{\theta_w,1\}^L\sum_{n=1}^N\left\|\left(d_v^{(k)}\right)_{n,\bar{\ell}}\right\|_1\\
&=L\max\{\theta_w,1\}^L\sum_{n\in\mathcal{I}_{\bar{\ell}}}\left\|v^*_{n,\bar{\ell}}-\sigma\left(u^*_{n,\bar{\ell}}\right)\right\|_1,
\end{aligned} \tag{37}$$

where the last equality comes from the definition (35). We also obtain that

$$\begin{aligned}
&\sum_{n=1}^N\sum_{\ell=1}^L\beta_\ell e_{N_\ell}^\top\left(d_v^{(k)}\right)_{n,\ell}-\frac{1}{\epsilon_k}\sum_{n=1}^N\sum_{\ell=1}^L\beta_\ell e_{N_\ell}^\top\left(\sigma\left(u_{n,\ell}+\epsilon_k\left(d_u^{(k)}\right)_{n,\ell}\right)-\sigma(u_{n,\ell})\right)\\
=&-\beta_{\bar{\ell}}\sum_{n\in\mathcal{I}_{\bar{\ell}}}\left\|v^*_{n,\bar{\ell}}-\sigma\left(u^*_{n,\bar{\ell}}\right)\right\|_1+\sum_{\ell=\bar{\ell}+1}^L\sum_{n=1}^N\beta_\ell\left\|\left(d_v^{(k)}\right)_{n,\ell}\right\|_1\\
&-\frac{1}{\epsilon_k}\sum_{n=1}^N\sum_{\ell=\bar{\ell}+1}^L\beta_\ell e_{N_\ell}^\top\left(\sigma\left(u_{n,\ell}+\epsilon_k\left(d_u^{(k)}\right)_{n,\ell}\right)-\sigma(u_{n,\ell})\right)\\
\le&-\beta_{\bar{\ell}}\sum_{n\in\mathcal{I}_{\bar{\ell}}}\left\|v^*_{n,\bar{\ell}}-\sigma\left(u^*_{n,\bar{\ell}}\right)\right\|_1+\sum_{\ell=\bar{\ell}+1}^L\sum_{n=1}^N\beta_\ell\left(\left\|\left(d_v^{(k)}\right)_{n,\ell}\right\|_1+\left\|\left(d_u^{(k)}\right)_{n,\ell}\right\|_1\right)\\
\le&\left(2\sum_{j=\bar{\ell}+1}^L\beta_j\theta_w\max\{\theta_w,1\}^{j-\bar{\ell}-1}-\beta_{\bar{\ell}}\right)\sum_{n\in\mathcal{I}_{\bar{\ell}}}\left\|v^*_{n,\bar{\ell}}-\sigma\left(u^*_{n,\bar{\ell}}\right)\right\|_1,
\end{aligned} \tag{38}$$

where the first equality comes from $-e_{N_{\bar{\ell}}}^\top (d_v^{(k)})_{n,\bar{\ell}} = \sum_{n\in\mathcal{I}_{\bar{\ell}}} \|v_{n,\bar{\ell}}^* - \sigma(u_{n,\bar{\ell}}^*)\|_1$, $(d_u^{(k)})_{n,\bar{\ell}} = 0$ and $(d_v^{(k)})_{n,\ell} = (d_u^{(k)})_{n,\ell} = 0$ for all $1 \le \ell < \bar{\ell}$ and $n \in [N]$ by definition (35), and the last inequality yields from $\|(d_v^{(k)})_{n,\ell}\|_1 \le \theta_w \|(d_v^{(k)})_{n,\ell-1}\|_1$ and $\|(d_u^{(k)})_{n,\ell}\|_1 \le \theta_w \|(d_v^{(k)})_{n,\ell-1}\|_1$ for all $k \in \mathbb{N}$, $\ell > \bar{\ell}$ and $n \in [N]$.

It then holds that

$$\begin{aligned}
&\frac{1}{\epsilon_k}\left(\mathcal{O}\left(w, v+\epsilon_k d_v^{(k)}, u+\epsilon_k d_u^{(k)}\right) - \mathcal{O}(w,v,u)\right)\\
=&\frac{1}{\epsilon_k}\left(\bar{\mathcal{O}}\left(w, v+\epsilon_k d_v^{(k)}\right) - \bar{\mathcal{O}}(w,v)\right) + \sum_{n=1}^{N}\sum_{\ell=1}^{L} \beta_\ell e_{N_\ell}^\top \left(d_v^{(k)}\right)_{n,\ell}\\
&- \frac{1}{\epsilon_k}\sum_{n=1}^{N}\sum_{\ell=1}^{L} \beta_\ell e_{N_\ell}^\top \left(\sigma\left(u_{n,\ell} + \epsilon_k \left(d_u^{(k)}\right)_{n,\ell}\right) - \sigma(u_{n,\ell})\right)
\end{aligned}$$

$$\le \left(L L_{\bar{\mathcal{O}}} \max\{\theta_w, 1\}^L + 2\sum_{j=\bar{\ell}+1}^{L} \beta_j \theta_w \max\{\theta_w,1\}^{j-\bar{\ell}-1} - \beta_{\bar{\ell}}\right) \sum_{n\in\mathcal{I}_{\bar{\ell}}} \left\|v_{n,\bar{\ell}}^* - \sigma\left(u_{n,\bar{\ell}}^*\right)\right\|_1 < 0,$$

where the equality comes from the definitions (35), and the first inequality yields from the inequalities (37), (38) and $\bar{\mathcal{O}}(w,v) < \mathcal{O}(w,v,u) < \theta$.

Together with the relationships (32) and (36), there is a contradiction. We then conclude that $v_{n,\bar{\ell}}^* = \sigma(u_{n,\bar{\ell}}^*)$ for all $n \in [N]$. The proof is completed by mathematical induction. ■

We next present the main theorem illustrating the fact that problems (PP) and (P) sharing the same global and local minimizers.

Theorem 10 *Let the penalty parameter β satisfy* (31)*. Then the following statements hold.*

(a) (w^*, b^*, v^*, u^*) is a global minimizer of problem (PP) *if and only if (w^*, b^*, v^*, u^*) is a global minimizer of problem* (P)*.*

(b) If $(w^*, b^*, v^*, u^*) \in \{(w,b,v,u) \in \Omega_2 : \mathcal{O}(w,v,u) < \theta\}$ is a local minimizer of problem (PP)*, then (w^*, b^*, v^*, u^*) is also a local minimizer of problem* (P)*.*

(c) $(w^*, b^*, v^*, u^*) \in \{(w,b,v,u) \in \Omega_1 : \mathcal{O}(w,v,u) < \theta\}$ is a local minimizer of problem (P) *if and only if (w^*, b^*, v^*, u^*) is a local minimizer of problem* (PP)*.*

Proof (a) If (w^*, b^*, v^*, u^*) is a global minimizer of problem (PP), we obtain that $\bar{\mathcal{O}}(w^*, v^*) < \theta$ by $\bar{\mathcal{O}}(0,0) = \frac{1}{N}\|Y\|_F^2 < \theta$. Lemma 4 yields that (w^*, b^*, v^*, u^*) is a limiting stationary point of problem (PP). From Lemma 9, we have $(w^*, b^*, v^*, u^*) \in \Omega_1$. Together with the inclusion $\Omega_1 \subset \Omega_2$, we know that $(w^*, b^*, v^*, u^*) \in \Omega_1$ must be a global minimizer of problem (P).

Conversely, suppose that $(\bar{w}, \bar{b}, \bar{v}, \bar{u})$ is a global minimizer of problem (PP). From what we have proved, it holds that $\bar{\mathcal{O}}(\bar{w}, \bar{v}) < \theta$, $(\bar{w}, \bar{b}, \bar{v}, \bar{u}) \in \Omega_1$ and

$$\min_{(w,b,v,u)\in\Omega_2} \mathcal{O}(w,v,u) = \mathcal{O}(\bar{w}, \bar{v}, \bar{u}) = \bar{\mathcal{O}}(\bar{w}, \bar{v}) = \min_{(w,b,v,u)\in\Omega_1} \bar{\mathcal{O}}(w,v). \tag{39}$$

Since $(w^*, b^*, v^*, u^*) \in \Omega_1$ is a global minimizer of problem (P), we have

$$\min_{(w,b,v,u)\in\Omega_1} \bar{\mathcal{O}}(w, v) = \bar{\mathcal{O}}(w^*, v^*) = \mathcal{O}(w^*, v^*, u^*). \tag{40}$$

Together with the facts (39) and $(w^*, b^*, v^*, u^*) \in \Omega_1 \subset \Omega_2$, we have (w^*, b^*, v^*, u^*) is a global minimizer of problem (PP).

(b) Lemma 4 yields that (w^*, b^*, v^*, u^*) is a limiting stationary point of problem (PP). From Lemma 9, it holds that $(w^*, b^*, v^*, u^*) \in \Omega_1$. Together with the inclusion $\Omega_1 \subset \Omega_2$, $(w^*, b^*, v^*, u^*) \in \Omega_1$ must be a local minimizer of problem (P).

(c) By using a similar method as that in the proof of (a) and (b), we complete the statement (c). ■

Remark 11 *Recall the definition of Ω_1 and Ω_2, it holds that $\Omega_1 \subsetneq \Omega_2$. Hence the set $\{(w, b, v, u) \in \Omega_2 : \mathcal{O}(w, v, u) < \theta\}$ is distinct from the set $\{(w, b, v, u) \in \Omega_1 : \mathcal{O}(w, v, u) < \theta\}$. Therefore, the inclusion (b) is proper.*

Since $\mathcal{R}_1$ and $\mathcal{R}_2$ are directionally differentiable, we can show that if $(\bar{w}, \bar{b}, \bar{v}, \bar{u}) \in \{(w, b, v, u) \in \Omega_2 : \mathcal{O}(w, b, v, u) < \theta\}$ is a d-stationary point of problem (PP), then $(\bar{w}, \bar{b}, \bar{v}, \bar{u})$ is also a d-stationary point of problem (P) by using Lemma 9 and ideas from (Cui et al., 2020, Theorem 2.1) and (Liu et al., 2022, Theorem 2.5). However, computing a d-stationary point is difficult, we will consider a limiting stationary point of problem (PP), and show that it is an MPCC W-stationary point of problem (P).

Theorem 12 *Let the penalty parameter β satisfy* (31). *If $(w^*, b^*, v^*, u^*) \in \{(w, b, v, u) \in \Omega_2 : \mathcal{O}(w, v, u) < \theta\}$ is a limiting stationary point of problem* (PP), *then (w^*, b^*, v^*, u^*) is an MPCC W-stationary point of problem* (P).

Proof Since it holds that $\mathcal{O}(w^*, v^*, u^*) < \theta$, Lemma 9 yields that $(w^*, b^*, v^*, u^*) \in \Omega_1$.

From Theorem 5 and (w^*, b^*, v^*, u^*) being a limiting stationary point of problem (PP), there exist vectors $\mu \in \mathbb{R}^{2m}_+$ and $\xi \in \mathbb{R}^m$ such that (7)– (10) hold.

Let $\mu = [\left(\mu^1\right)^\top, \left(\mu^2\right)^\top]^\top$ with $\mu^1, \mu^2 \in \mathbb{R}^m_+$. Recall the definition of $\mathcal{C}$, it holds that

$$\nabla_v \mu^\top \mathcal{C}(v^*, u^*) = -\left(\mu^1 + \mu^2\right), \nabla_u \mu^\top \mathcal{C}(v^*, u^*) = \mu^1 + \alpha\mu^2. \tag{41}$$

Then, we obtain from (8)–(10) that

$$\begin{aligned} &0 = \nabla_v \bar{\mathcal{O}}(w^*, v^*) + \beta - \left(\mu^1 + \mu^2\right) + \nabla_v \xi^\top (u^* - \Psi(v^*)w^*), \\ &0 \in \partial_u(-\beta^\top \sigma(u^*)) + \mu^1 + \alpha\mu^2 + \xi, \\ &\left(\mu^1\right)^\top (u^* - v^*) = 0, \left(\mu^2\right)^\top (\alpha u^* - v^*) = 0, \mu^1 \geq 0, \mu^2 \geq 0. \end{aligned} \tag{42}$$

Now, we prove that there exist $\bar{\mu}^1$ and $\bar{\mu}^2$ such that

$$0 = \bar{\mu}^1 + \alpha\bar{\mu}^2 + \xi, \bar{\mu}^1 + \bar{\mu}^2 = \mu^1 + \mu^2 - \beta, \left(\bar{\mu}^1\right)^\top (v^* - u^*) = 0, \left(\bar{\mu}^2\right)^\top (v^* - \alpha u^*) = 0 \tag{43}$$

by analyzing the following cases for all $i \in [m]$.

Case (i): if $u_i^* < 0$, the relation (42) together with the definition of σ yield $v_i^* = \alpha u_i^*$, $\mu_i^1 = 0$ and

$$\partial(-(\beta)_i\sigma(u_i^*)) + \mu_i^1 + \alpha\mu_i^2 = \left\{\mu_i^1 + \alpha\left(\mu_i^2 - (\beta)_i\right)\right\},$$

where $(\beta)_i$ denotes the i-th element of β. In this case, we have $\mu_i^1(v_i^* - u_i^*) = 0$ and $(\mu_i^2 - (\beta)_i)(v_i^* - \alpha u_i^*) = 0$. Let $\bar{\mu}_i^1 = \mu_i^1 = 0, \bar{\mu}_i^2 = \mu_i^2 - (\beta)_i$.

Case (ii): if $u_i^* > 0$, the relation (42) together with the definition of σ yield $v_i^* = u_i^*$, $\mu_i^2 = 0$ and

$$\partial(-(\beta)_i\sigma(u_i^*)) + \mu_i^1 + \alpha\mu_i^2 = \left\{\left(\mu_i^1 - (\beta)_i\right) + \alpha\mu_i^2\right\}.$$

In this case, we have $(\mu_i^1 - (\beta)_i)(v_i^* - u_i^*) = 0$ and $\mu_i^2(v_i^* - \alpha u_i^*) = 0$. Let $\bar{\mu}_i^1 = \mu_i^1 - (\beta)_i, \bar{\mu}_i^2 = \mu_i^2 = 0$.

Case (iii): if $u_i^* = 0$, we have $v_i^* = \sigma(u_i^*) = 0$ and

$$\partial(-(\beta)_i\sigma(u_i^*)) + \mu_i^1 + \alpha\mu_i^2 = \left\{\left(\mu_i^1 - (\beta)_i\right) + \alpha\mu_i^2, \mu_i^1 + \alpha\left(\mu_i^2 - (\beta)_i\right)\right\}. \tag{44}$$

In this case, we have $v_i^* - u_i^* = 0$, $v_i^* - \alpha u_i^* = 0$. It then holds that either $0 = (\mu_i^1 - (\beta)_i) + \alpha\mu_i^2 + \xi$ or $0 = \mu_i^1 + \alpha(\mu_i^2 - (\beta)_i) + \xi$ by (42). If $0 = \mu_i^1 + \alpha(\mu_i^2 - (\beta)_i) + \xi$, let $\bar{\mu}_i^1 = \mu_i^1 \geq 0, \bar{\mu}_i^2 = \mu_i^2 - (\beta)_i$. Otherwise, let $\bar{\mu}_i^2 = \mu_i^2 \geq 0, \bar{\mu}_i^1 = \mu_i^1 - (\beta)_i$.

Combining the above three cases, we obtain (43). Recall the relation (42), we then derive (11)– (14) with $\bar{\mu}^1$, $\bar{\mu}^2$ instead of μ^1 and μ^2, respectively. The proof is then completed. ■

Remark 13 *Notice that the conditions* (31) *for the penalty parameter β are recursively define from L to* 1. *This coincides the intuition that the inner layer should have larger penalty parameter than the outer one to avoid error accumulation.*

Remark 14 *By simple calculation, we have*

$$\bar{\mu}_i^1\bar{\mu}_i^2 = \begin{cases} \mu_i^1\left(\mu_i^2 - (\beta)_i\right) \text{ or } \mu_i^2\left(\mu_i^1 - (\beta)_i\right), & \text{if } u_i^* = 0, \\ 0, & \text{if } u_i^* \neq 0. \end{cases}$$

If $\mu_i^1 = \mu_i^2 = 0$ for all $i \in \{i : u_i^* = 0\}$, then (w^*, b^*, v^*, u^*) is an MPCC C-stationary point of problem (P).

Remark 15 *Our theoretical results can also be extended to ReLU network with $\alpha = 0$. Notice that the solution set of problem* (PP) *with $\alpha = 0$ is unbounded (see a counterexample given by Liu et al. (2022)), we introduce a constrained set Ω_b, and minimize the objective function of problem* (PP) *over $\Omega_2 \cap \{(w, b, v, u) : b \in \Omega_b\}$, where*

$$\Omega_b := \left\{b : b \geq -e_{\overline{N}}\overline{N}\theta_w\theta_v\right\}.$$

We call the resulted problem ($\mathrm{PP_b}$). *By using a similar method as that in the proof of Theorems 8 and 10, we can prove that the solution set of problem* ($\mathrm{PP_b}$) *is nonempty and bounded; the global (local) minimizer of problem* ($\mathrm{PP_b}$) *is a global (local) minimizer of problem* (P). *However, a limiting stationary point of problem* ($\mathrm{PP_b}$) *may not be a MPCC*

W-stationary point of (P) *with* $\alpha = 0$. *Specifically, if* (w^*, b^*, v^*, u^*) *is a limiting stationary point of problem* ($\mathrm{PP_b}$)*, then* (11)–(14) *hold with* $0 \le A^\top \xi$ *instead of* $0 = A^\top \xi$.

Furthermore, our numerical algorithm for solving problem (PP)*, which will be proposed in Section 4, can also be applied to solve problem* ($\mathrm{PP_b}$).

We end this section by summarizing our results for the relationship of problems (P) and (PP) in the following diagram, where the function value at the related points is less than θ.

(PP) :	global (local) minimizer		limiting stationary point $\Leftrightarrow$	KKT point
	$\Downarrow\Uparrow$	conditions in Remark 13 $\Downarrow$	$\Downarrow$	
(P) :	global (local) minimizer	MPCC C-stationary point	MPCC W-stationary point	

4. An Inexact Augmented Lagrangian Method with the Alternating Minimization (IALAM)

Problem (PP) is to minimize a nonsmooth nonconvex function subject to linear and bilinear constraints. By exploring the structure of problem (PP), we propose a variation of the inexact augmented Lagrangian (IALM) framework in Subsection 4.1. Then, we present an alternating minimization algorithm to solve the augmented Lagrangian subproblem in Subsection 4.2. Combining these two parts, we call our new algorithm IALAM. In Subsections 4.3 and 4.4, we prove that any iterate sequence generated by IALAM has at least one accumulation point and any accumulation point is a KKT point of problem (PP), which is a MPCC W-stationary point of problem (P) according to Theorem 12.

4.1 The Algorithm Framework

By penalizing the equality constraint of problem (PP), we can obtain its augmented Lagrangian (AL) function as follows

$$\mathcal{L}_\rho(w, b, v, u; \xi) := \mathcal{O}(w, v, u) + \langle \xi, u - \Psi(v)w - Ab\rangle + \frac{\rho}{2} \|u - \Psi(v)w - Ab\|^2, \tag{45}$$

where $\rho > 0$ is the penalty parameter and

$$\xi := \left(\xi_{1,1}^\top, \xi_{2,1}^\top, \ldots, \xi_{N,1}^\top, \xi_{1,2}^\top, \ldots, \xi_{N,L}^\top\right)^\top$$

is the Lagrangian multiplier associate with $u = \Psi(v)w - Ab$, $\xi_{n,\ell} \in \mathbb{R}^{N_\ell}$ for all $n \in [N]$ and $\ell \in [L]$. Recall the definition of $\Psi(v)$ and A, it holds that

$$\langle \xi, u - \Psi(v)w - Ab\rangle = \sum_{\ell=1}^{L} \sum_{n=1}^{N} \langle \xi_{n,\ell}, u_{n,\ell} - W_\ell v_{n,\ell-1} - b_\ell\rangle .$$

In the framework of any augmented Lagrangian based approach, it requires to solve the following subproblem with the dual variables fixed at each iteration to update the prime variables

$$\min_{(w,b,v,u)\in\Omega_3} \mathcal{L}_\rho(w, b, v, u; \xi), \tag{46}$$

where

$$\Omega_3 := \{(w, b, v, u) : w \in \mathbb{R}^{\widetilde{N}}, b \in \mathbb{R}^{\overline{N}}, \mathcal{C}(v, u) \leq 0\}$$

and $\mathcal{C}(v, u)$ is defined in (6). We denote $(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)}, \xi^{(k)})$ as the k-th iterate tuple. At the k-th iteration, we inexactly solve (46) with $\rho = \rho^{(k-1)}$ and $\xi = \xi^{(k-1)}$ to obtain an approximate solution $(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)}) \in \Omega_3$ satisfying the following two conditions,

$$\mathcal{L}_{\rho^{(k-1)}}\left(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)}; \xi^{(k-1)}\right) < \theta, \tag{47}$$

and

$$\operatorname{dist}\left(0, \partial \mathcal{L}_{\rho^{(k-1)}}\left(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)}; \xi^{(k-1)}\right) + \mathcal{N}_{\Omega_3}\left(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)}\right)\right) \leq \epsilon_k. \tag{48}$$

We describe the IALM framework with inexact criteria (47) and (48) in Algorithm 1. The definitions of θ and Ω_θ can be found in Subsection 3.1.

Algorithm 1 The inexact augmented Lagrangian method for solving problem (PP)

Input: initial point $(w^{(0)}, b^{(0)}, v^{(0)}, u^{(0)}) \in \Omega_\theta$, parameters $\rho^{(0)} > 0$, $\eta_1, \eta_2, \eta_4 \in (0, 1)$, $\eta_3 > 0$, $\xi^{(0)} \in \mathbb{R}^m$, $\gamma \in \mathbb{N}_+$, and $\epsilon_0 > 0$. Set $k := 1$.

while the stop criterion is not met **do**

Step 1: Solve (46) with $\rho = \rho^{(k-1)}$ and $\xi = \xi^{(k-1)}$ and obtain $(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)}) \in \Omega_3$ satisfying (47) and (48).

Step 2: Update the Lagrangian multipliers by

$$\xi^{(k)} = \xi^{(k-1)} + \rho^{(k-1)}\left(u^{(k)} - \Psi(v^{(k)})w^{(k)} - Ab^{(k)}\right). \tag{49}$$

Step 3: If $k \leq \gamma$, set $\rho^{(k)} = \rho^{(k-1)}$ and $\epsilon_k = \epsilon_{k-1}$. Else if $k > \gamma$, and

$$\left\|u^{(k)} - \Psi(v^{(k)})w^{(k)} - Ab^{(k)}\right\| \leq \eta_1 \max_{t=k-\gamma,\ldots,k-1} \left\|u^{(t)} - \Psi(v^{(t)})w^{(t)} - Ab^{(t)}\right\|, \tag{50}$$

then set $\rho^{(k)} = \rho^{(k-1)}$ and $\epsilon_k = \sqrt{\eta_1}\epsilon_{k-1}$. Otherwise, set

$$\rho^{(k)} = \max\left\{\rho^{(k-1)}/\eta_2, \left\|\xi^{(k)}\right\|^{1+\eta_3}\right\} \text{ and } \epsilon_k = \eta_4\epsilon_{k-1}. \tag{51}$$

Set $k := k + 1$.

end while

Output: $(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})$.

4.2 The Alternating Minimization Algorithm

Subproblem (46) is to minimize a nonsmooth nonconvex function subject to linear constraints. We utilize its block structure and propose an alternating minimization algorithm. Before we present the detailed algorithm framework, we introduce how to choose an initial point and update the two blocks. We assume to be at the k-th iteration of Algorithm 1.

Initialization. Let $(w_{\text{init}}^{(k)}, b_{\text{init}}^{(k)}, v_{\text{init}}^{(k)}, u_{\text{init}}^{(k)})$ be the initial point of the Algorithm 2, which is updated recursively as follows

$$\left(w_{\text{init}}^{(k)}, b_{\text{init}}^{(k)}, v_{\text{init}}^{(k)}, u_{\text{init}}^{(k)}\right) = \begin{cases} \left(w_{\text{init}}^{(k-1)}, b_{\text{init}}^{(k-1)}, v_{\text{init}}^{(k-1)}, u_{\text{init}}^{(k-1)}\right), \text{ if } k > 1 \text{ and} \\ \qquad\qquad \mathcal{L}_{\rho^{(k-1)}}\left(w^{(k-1)}, b^{(k-1)}, \bar{v}, \bar{u}; \xi^{(k-1)}\right) \geq \theta, \\ \left(w^{(k-1)}, b^{(k-1)}, \bar{v}, \bar{u}\right), \qquad \text{otherwise}, \end{cases} \tag{52}$$

where $\bar{v}_{n,0} = x_n$, $\bar{u}_{n,\ell} = W_\ell^{(k-1)} \bar{v}_{n,\ell-1} + b_\ell^{(k-1)}$, and $\bar{v}_{n,\ell} = \sigma\left(\bar{u}_{n,\ell}\right)$ for all $n \in [N]$ and $\ell \in [L]$. Clearly, $(w_{\text{init}}^{(k)}, b_{\text{init}}^{(k)}, v_{\text{init}}^{(k)}, u_{\text{init}}^{(k)})$ is a feasible point of problem (PP) for all $k \in \mathbb{N}_+$ by its definition.

The notations $(w^{(k,j)}, b^{(k,j)}, v^{(k,j)}, u^{(k,j)})$ stands for the j-th iterate of the alternating minimization algorithm and the k-th iterate of Algorithm 1. For brevity, we drop the superscript $(k-1)$ (and (k)) and abuse the notations $\rho, \xi, w^{(j)}, b^{(j)}, v^{(j)}, u^{(j)}$ to denote $\rho^{(k-1)}, \xi^{(k-1)}, w^{(k,j)}, b^{(k,j)}, v^{(k,j)}, u^{(k,j)}$, respectively. We assume to be at the j-th iterate of the alternating minimization algorithm.

Update of the (w, b) block. Once (v, u) block is fixed at $(v^{(j)}, u^{(j)})$. We compute $(w^{(j+1)}, b^{(j+1)})$ by solving the following convex problem $\min_{w,b} \mathcal{L}_\rho(w, b, v^{(j)}, u^{(j)}; \xi)$, i.e.,

$$\min_{w,b} \mathcal{R}_1(w) + \left\langle \xi, u^{(j)} - \Psi\left(v^{(j)}\right) w - Ab \right\rangle + \frac{\rho}{2} \left\| u^{(j)} - \Psi\left(v^{(j)}\right) w - Ab \right\|^2. \tag{53}$$

This can be solved by some existing methods, for example, the proximal gradient method.

Update of the (v, u) block: After obtaining $(w^{(j+1)}, b^{(j+1)})$, we calculate $(u^{(j+1)}, v^{(j+1)})$ in the following way. We define an proximal term $\mathcal{P}(u, v; u^{(j)}, v^{(j)}, \tau^{(j)})$ by

$$\mathcal{P}\left(u, v; u^{(j)}, v^{(j)}, \tau^{(j)}\right) := \frac{1}{2} \sum_{n=1}^{N} \sum_{\ell=2}^{L} \left\| \begin{pmatrix} v_{n,\ell-1} \\ u_{n,\ell} \end{pmatrix} - \begin{pmatrix} v_{n,\ell-1}^{(j)} \\ u_{n,\ell}^{(j)} \end{pmatrix} \right\|_{S_\ell^{(j)}}^2 + \frac{\tau_1}{2} \sum_{n=1}^{N} \left\| u_{n,1} - u_{n,1}^{(j)} \right\|^2, \tag{54}$$

where $\tau_1 > 0$ is a given parameter, $\tau^{(j)} := (\tau_2^{(j)}, \ldots, \tau_L^{(j)})^\top \in \mathbb{R}^{L-1}$, $\tau_\ell^{(j)}$ and matrix $S_\ell^{(j)}$ are defined by

$$\tau_\ell^{(j)} := \rho \left\| \left[-W_\ell^{(j+1)} \quad I_{N_\ell} \right] \right\|^2 + \tau_1, \tag{55}$$

$$S_\ell^{(j)} := \tau_\ell^{(j)} I_{N_\ell + N_{\ell-1}} - \rho \left[-W_\ell^{(j+1)} \quad I_{N_\ell} \right]^\top \left[-W_\ell^{(j+1)} \quad I_{N_\ell} \right], \tag{56}$$

respectively for all $\ell = 2, 3, \ldots, L$. Clearly, $S_\ell^{(j)} \succeq \tau_1 I_{N_\ell + N_{\ell-1}}$ is a symmetric positive definite matrix, since $\|[-W_\ell^{(j+1)} \quad I_{N_\ell}]\|^2$ is the maximal eigenvalue of $[-W_\ell^{(j+1)} \quad I_{N_\ell}]^\top [-W_\ell^{(j+1)} \quad I_{N_\ell}]$ for any $\ell = 2, 3, \ldots, L$. Then, we arrive at a linearly constrained problem

$$\begin{aligned} \arg\min_{v,u} \ & \mathcal{L}_\rho\left(w^{(j+1)}, b^{(j+1)}, v, u; \xi\right) + \mathcal{P}\left(u, v; u^{(j)}, v^{(j)}, \tau^{(j)}\right) \\ \text{s.t. } & v \geq u, v \geq \alpha u. \end{aligned} \tag{57}$$

We can calculate its unique solution $(v^{(j+1)}, u^{(j+1)})$ in the following way.

Notice that

$$\begin{aligned}
&\frac{\rho}{2}\left\|u-\Psi(v)w^{(\jmath+1)}-Ab^{(\jmath+1)}\right\|^2+\mathcal{P}\left(u,v;u^{(\jmath)},v^{(\jmath)},\tau^{(\jmath)}\right)\\
=&\frac{1}{2}\sum_{n=1}^{N}\sum_{\ell=2}^{L}\left(\rho\left\|u_{n,\ell}-\left(W_\ell^{(\jmath+1)}v_{n,\ell-1}+b_\ell^{(\jmath+1)}\right)\right\|^2+\left\|\begin{pmatrix}v_{n,\ell-1}\\u_{n,\ell}\end{pmatrix}-\begin{pmatrix}v_{n,\ell-1}^{(\jmath)}\\u_{n,\ell}^{(\jmath)}\end{pmatrix}\right\|_{S_\ell^{(\jmath)}}^2\right)\\
&+\frac{\rho}{2}\sum_{n=1}^{N}\left\|u_{n,1}-\left(W_1^{(\jmath+1)}x_n+b_1^{(\jmath+1)}\right)\right\|^2+\frac{\tau_1}{2}\sum_{n=1}^{N}\left\|u_{n,1}-u_{n,1}^{(\jmath)}\right\|^2,
\end{aligned}$$

where

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{\ell=2}^{L}\left(\rho\left\|u_{n,\ell}-\left(W_\ell^{(\jmath+1)}v_{n,\ell-1}+b_\ell^{(\jmath+1)}\right)\right\|^2+\left\|\begin{pmatrix}v_{n,\ell-1}\\u_{n,\ell}\end{pmatrix}-\begin{pmatrix}v_{n,\ell-1}^{(\jmath)}\\u_{n,\ell}^{(\jmath)}\end{pmatrix}\right\|_{S_\ell^{(\jmath)}}^2\right)\\
=&\sum_{n=1}^{N}\sum_{\ell=2}^{L}\left(\rho\left\|u_{n,\ell}^{(\jmath)}-\left(W_\ell^{(\jmath+1)}v_{n,\ell-1}^{(\jmath)}+b_\ell^{(\jmath+1)}\right)\right\|^2+\left\|\begin{pmatrix}v_{n,\ell-1}\\u_{n,\ell}\end{pmatrix}-\begin{pmatrix}v_{n,\ell-1}^{(\jmath)}\\u_{n,\ell}^{(\jmath)}\end{pmatrix}\right\|_{S_\ell^{(\jmath)}}^2\right.\\
&+2\rho\left(u_{n,\ell}^{(\jmath)}-\left(W_\ell^{(\jmath+1)}v_{n,\ell-1}^{(\jmath)}+b_\ell^{(\jmath+1)}\right)\right)^\top\left[-W_\ell^{(\jmath+1)}\quad I_{N_\ell}\right]\begin{pmatrix}v_{n,\ell-1}-v_{n,\ell-1}^{(\jmath)}\\u_{n,\ell}-u_{n,\ell}^{(\jmath)}\end{pmatrix}\\
&\left.+\rho\begin{pmatrix}v_{n,\ell-1}-v_{n,\ell-1}^{(\jmath)}\\u_{n,\ell}-u_{n,\ell}^{(\jmath)}\end{pmatrix}^\top\begin{bmatrix}(W_\ell^{(\jmath+1)})^\top W_\ell^{(\jmath+1)}&-(W_\ell^{(\jmath+1)})^\top\\-W_\ell^{(\jmath+1)}&I_{N_\ell}\end{bmatrix}\begin{pmatrix}v_{n,\ell-1}-v_{n,\ell-1}^{(\jmath)}\\u_{n,\ell}-u_{n,\ell}^{(\jmath)}\end{pmatrix}\right)\\
=&\sum_{n=1}^{N}\sum_{\ell=2}^{L}\left(\rho\left\|u_{n,\ell}^{(\jmath)}-\left(W_\ell^{(\jmath+1)}v_{n,\ell-1}^{(\jmath)}+b_\ell^{(\jmath+1)}\right)\right\|^2+\tau_\ell^{(\jmath)}\begin{pmatrix}v_{n,\ell-1}-v_{n,\ell-1}^{(\jmath)}\\u_{n,\ell}-u_{n,\ell}^{(\jmath)}\end{pmatrix}^\top\begin{pmatrix}v_{n,\ell-1}-v_{n,\ell-1}^{(\jmath)}\\u_{n,\ell}-u_{n,\ell}^{(\jmath)}\end{pmatrix}\right.\\
&\left.+2\rho\left(u_{n,\ell}^{(\jmath)}-\left(W_\ell^{(\jmath+1)}v_{n,\ell-1}^{(\jmath)}+b_\ell^{(\jmath+1)}\right)\right)^\top\left[-W_\ell^{(\jmath+1)}\quad I_{N_\ell}\right]\begin{pmatrix}v_{n,\ell-1}-v_{n,\ell-1}^{(\jmath)}\\u_{n,\ell}-u_{n,\ell}^{(\jmath)}\end{pmatrix}\right).
\end{aligned}$$

Then, the objective function of problem (57) can be simplified as

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\|v_{n,L}-y_n\|^2+\lambda_v\|v\|^2+\beta^\top(v-\sigma(u))+\sum_{\ell=1}^{L}\sum_{n=1}^{N}\left\langle\xi_{n,\ell},u_{n,\ell}-W_\ell^{(\jmath+1)}v_{n,\ell-1}\right\rangle\\
+&\frac{1}{2}\sum_{n=1}^{N}\sum_{\ell=2}^{L}\left(\tau_\ell^{(\jmath)}\left\|u_{n,\ell}-u_{n,\ell}^{(\jmath)}\right\|^2+2\rho\left(u_{n,\ell}^{(\jmath)}-\left(W_\ell^{(\jmath+1)}v_{n,\ell-1}^{(\jmath)}+b_\ell^{(\jmath+1)}\right)\right)^\top\left(u_{n,\ell}-u_{n,\ell}^{(\jmath)}\right)\right.\\
+&\left.\tau_\ell^{(\jmath)}\left\|v_{n,\ell-1}-v_{n,\ell-1}^{(\jmath)}\right\|^2-2\rho\left(u_{n,\ell}^{(\jmath)}-\left(W_\ell^{(\jmath+1)}v_{n,\ell-1}^{(\jmath)}+b_\ell^{(\jmath+1)}\right)\right)^\top W_\ell^{(\jmath+1)}\left(v_{n,\ell-1}-v_{n,\ell-1}^{(\jmath)}\right)\right)\\
+&\frac{\rho}{2}\sum_{n=1}^{N}\left(\left\|u_{n,1}-\left(W_1^{(\jmath+1)}x_n+b_1^{(\jmath+1)}\right)\right\|^2\right)+\frac{\tau_1}{2}\sum_{n=1}^{N}\left\|u_{n,1}-u_{n,1}^{(\jmath)}\right\|^2.
\end{aligned}\tag{58}$$

Hence, subproblem (57) can be separated into m independent subproblems of the following structure

$$\begin{aligned} \min_{r,s\in\mathbb{R}} \; & r - \max\{s, \alpha s\} + \frac{d_1}{2}\left(r - \frac{d_3}{d_1}\right)^2 + \frac{d_2}{2}\left(s - \frac{d_4}{d_2}\right)^2 \\ \text{s.t.} \quad & r \geq s, r \geq \alpha s, \end{aligned} \tag{59}$$

where the constants $d_1, d_2 > 0$ and $d_3, d_4 \in \mathbb{R}$ are dependent on the parameters of problem (57).

Restricting problem (59) to $\{(r,s) : s \geq 0\}$, we obtain $\min_{r\geq s, s\geq 0} \; r - s + d_1(r - d_3/d_1)^2/2 + d_2(s - d_4/d_2)^2/2$ and its closed-form solution (Facchinei and Pang, 2003, page 81)

$$(r_1^*,\ s_1^*) := \mathrm{Proj}^{\mathrm{diag}(d_1,d_2)}_{\{(r,\ s):r\geq s,\ s\geq 0\}}\left(\frac{d_3-1}{d_1},\ \frac{d_4+1}{d_2}\right). \tag{60}$$

On the other hand, restricting problem (59) to $\{(r,s) : s \leq 0\}$, we obtain $\min_{r\geq \alpha s,\ s\leq 0} \; r - \alpha s + d_1(r - d_3/d_1)^2/2 + d_2(s - d_4/d_2)^2/2$ and its closed-form solution

$$(r_2^*,\ s_2^*) = \mathrm{Proj}^{\mathrm{diag}(d_1,d_2)}_{\{(r,\ s):r\geq \alpha s,\ s\leq 0\}}\left(\frac{d_3-1}{d_1},\ \frac{d_4+\alpha}{d_2}\right). \tag{61}$$

By comparing the objective function values at $(r_1^*,\ s_1^*)$ and $(r_2^*,\ s_2^*)$, we obtain the unique solution of (59).

We next present the framework of the alternating minimization method for solving (46) as follows

Algorithm 2 An alternating minimization method for solving (46)

Input: matrix A, the vector ξ, the parameters $\rho > 0$ and $\tau_1 > 0$. Initialize $(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)}, u^{(\jmath)})$ by (52). Set $\jmath = 0$.
Step 1: Update $(w^{(\jmath+1)}, b^{(\jmath+1)})$ by solving problem (53).
Step 2: Update $(u^{(\jmath+1)}, v^{(\jmath+1)})$ by solving problem (57).
Step 3: Set $\jmath := \jmath + 1$. If the stop criterion is not met, return to Step 1.
Output: $(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)}, u^{(\jmath)})$.

Remark 16 *The reasons why we divide subproblem* (46) *into* (w, b) *and* (v, u) *blocks are two-fold. Firstly,* w, b *are the vectorization of weight matrices and bias vectors, respectively, meanwhile* v, u *are the auxiliary variables. Secondly, subproblem* (46) *restricted to both of these two blocks are easy to solve. More precisely, the* (w, b) *subproblem is strongly convex and has one unique solution, meanwhile the* (v, u) *subproblem has a closed-form unique solution.*

4.3 Convergence Analysis of Algorithm 1

In this subsection, we establish the convergence of Algorithm 1. The proof is given in Appendix A.

Theorem 17 *Let* $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ *be the sequence generated by Algorithm 1 with* $\eta_3 > 1$. *Then the following statements hold.*

(a) $\liminf_{k\to\infty} \|u^{(k)} - \Psi(v^{(k)})w^{(k)} - Ab^{(k)}\| = 0$ *and the sequence* $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ *has at least one accumulation point.*

(b) $\liminf_{k\to\infty} \mathrm{dist}((w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)}), \mathcal{Z}^*) = 0$*, where* $\mathcal{Z}^*$ *is the set of KKT points of problem* (PP).

(c) If in addition that $\gamma = 1$*, then* $\lim_{k\to\infty} \|u^{(k)} - \Psi(v^{(k)})w^{(k)} - Ab^{(k)}\| = 0$*. Furthermore, any accumulation point* (w^*, b^*, v^*, u^*) *of* $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ *is a KKT point of problem* (PP).

4.4 Convergence Analysis of Algorithm 2

In this subsection, we prove the global convergence of Algorithm 2 for solving subproblem (46) and show that the condition (47) always hold. The proof is given in Appendix B.

Theorem 18 *Let* $\{(w^{(j)}, b^{(j)}, v^{(j)}, u^{(j)})\}$ *be the sequence generated by Algorithm 2. Then we have the following statements.*

(a) It holds that

$$\begin{aligned} &\mathcal{L}_\rho\left(w^{(j+1)}, b^{(j+1)}, v^{(j+1)}, u^{(j+1)}; \xi\right) - \mathcal{L}_\rho\left(w^{(j)}, b^{(j)}, v^{(j)}, u^{(j)}; \xi\right) \\ \le &-\frac{\lambda_w}{2}\left\|w^{(j+1)} - w^{(j)}\right\|^2 - \frac{\tau_1}{2}\left\|u^{(j+1)} - u^{(j)}\right\|^2 - \frac{\tau_1}{2}\sum_{n=1}^{N}\sum_{\ell=1}^{L-1}\left\|v_{n,\ell}^{(j+1)} - v_{n,\ell}^{(j)}\right\|^2. \end{aligned} \tag{62}$$

(b) The sequence $\{\mathcal{L}_\rho(w^{(j)}, b^{(j)}, v^{(j)}, u^{(j)}; \xi)\}$ *is convergent.*

(c) The sequence $\{(w^{(j)}, b^{(j)}, v^{(j)}, u^{(j)})\}$ *is bounded.*

(d) It holds that

$$\lim_{j\to\infty} \left\|w^{(j+1)} - w^{(j)}\right\|^2 + \left\|b^{(j+1)} - b^{(j)}\right\|^2 + \left\|v^{(j+1)} - v^{(j)}\right\|^2 + \left\|u^{(j+1)} - u^{(j)}\right\|^2 = 0. \tag{63}$$

(e) The sequence $\{(w^{(j)}, b^{(j)}, v^{(j)}, u^{(j)})\}$ *has at least one accumulation point, and any accumulation point* (w^*, b^*, v^*, u^*) *of* $\{(w^{(j)}, b^{(j)}, v^{(j)}, u^{(j)})\}$ *is a KKT point of* (46).

Remark 19 *The statement (a) of Theorem 18 shows that Algorithm 2 yields a monotonic nonincreasing function value sequence* $\{\mathcal{L}_\rho\left(w^{(j)}, b^{(j)}, v^{(j)}, u^{(j)}; \xi\right)\}$ *for fixed* ρ *and* ξ*. Together with the selected initial guess, we can conclude that condition* (47) *always holds. Meanwhile, the statement (e) of Theorem 18 guarantees an inexact stationarity condition* (48) *can hold by certain iterate. Therefore, the inner iteration, Algorithm 2, is qualified to be Step 1 of the outer iteration, namely, Algorithm 1.*

5. Numerical Experiments

In this section, we evaluate the numerical performance of IALAM for training the sparse leaky ReLU network with group sparsity through comparing with some state-of-the-art SGD-based approaches. All the numerical experiments are conducted under MATLAB R2018b with windows 7 on a desktop with 3.4 GHz Inter Core i7-6700 CPU and 16 GB RAM.

5.1 Implementation Details

Algorithm parameters. For our IALAM, we set $\eta_1 = 0.99$, $\eta_2 = \frac{5}{6}$, $\eta_3 = 0.01$, $\eta_4 = \frac{2}{3}$, $\epsilon_0 = 0.1$, $\rho^{(0)} = 1/N$, $\xi^{(0)} = 0$, and $\gamma = 2L$. It is worthy of mentioning that although $\eta_3 = 0.01$ does not satisfy the requirement in Theorem 18 to guarantee the global convergence of Algorithm 1, this choice always yields better performance than $\eta_3 > 1$ in practice. In the inner iteration, subproblem (53) is solved by the proximal gradient method (Dai and Fletcher, 2005). We set the initial proximal parameter in subproblem (57) as $\tau_1 = \frac{1}{10N}$, and update $\tau_\ell^{(\jmath)}$, for all $\jmath = 2, 3, \ldots, L$, by formulation (54).

Stopping criterion and initial guess. Except for otherwise mentioned, we terminate our algorithm whenever $\epsilon_k < 10^{-6}$ or $\rho^{(k)} > 10^3\rho^{(0)}$. For all $\ell \in [L]$, the variables $W_\ell^{(0)}$ are randomly generated by $W_\ell^{(0)} = \mathrm{randn}(N_\ell, N_\ell - 1)/N$, where $\mathrm{randn}(n, p)$ stands for an $n \times p$ randomly generated matrix under the standard Gaussian distribution. Let $b^{(0)} = 0$, $v_{n,0}^{(0)} = x_n$, $u_{n,\ell}^{(0)} = W_\ell^{(0)} v_{n,\ell-1}^{(0)}$ and $v_{n,\ell}^{(0)} = \sigma(u_{n,\ell}^{(0)})$ for all $n \in [N]$ and $\ell \in [L]$.

Algorithms in Comparison. For comparison, we choose a few state-of-the-art SGD-based approaches, including the Adam (Kingma and Ba, 2014), the Adamax (Kingma and Ba, 2014), the Adadelata (Zeiler, 2012), the Adagrad (Duchi et al., 2011), the Adagrad-Decay (Duchi et al., 2011), and the Vanilla SGD (Cramir, 1946) with batch-size (Vanilla SGD (batch)). The MATLAB codes of these SGD-based approaches are downloaded from the SGD Library (Kasai, 2018). We also include ProxSGD (Yang et al., 2019). These approaches directly solve the unconstrained model, i.e., problem (1), neglecting the nonsmoothness. All of these algorithms are run under their defaulting settings. The batch-size of these methods is set to $\lceil\sqrt{N}\rceil$. We terminate these methods whenever the epoch (i.e., "Iteration×batch-size/N") reaches 1000 unless otherwise stated.

Model parameters (hyperparameters).

We introduce the model parameters of problem (PP) in the tests unless otherwise statement, which include $\alpha = 0.01$, $\lambda_w = \frac{1}{N}$ $\lambda_v = \frac{1}{100N}$, and $\beta = \frac{1}{N} e_m$. Specifically, results with various values of constant α and vector β are shown in Figures 2 and 6, and Table 3, respectively.

Test problems. The number of test samplings N_{test} is set to be $\lceil N/5 \rceil$.

There are three classes of test problems. The first class of test problems are generated randomly. We construct the training data sets and test data sets with a similar way as that proposed by Cui et al. (2020), i.e.,

$$y_n = \sigma(W_L \sigma(\cdots \sigma(W_1 x_n + b_1) + b_2 \cdots) + b_L) + \tilde{y}_n,$$

for all $n \in [N + N_{\text{test}}]$, where $x_n \sim \mathcal{N}(\zeta, \Sigma_0^T \Sigma_0)$, $\tilde{y}_n = \epsilon_y \mathrm{randn}(1, 1)$. Here, the parameter $\epsilon_y = 0.05$ is to control the noise level, $\zeta = \mathrm{randn}(N_0, 1)$, and $\Sigma_0 = \mathrm{randn}(N_0, 1)$.

The second class of test problems is the classification problem on the MNIST (LeCun, 1998) data set, consisting of 10-classes handwritten digits with the size 28×28, namely, $N_0 = 784$. In practice, we randomly pick up data entries from each class of MNIST under uniform distribution. The dataset consists of 60,000 training images and 10,000 test images. Since there are ten classes in the MNIST, we take $N_L = 10$.

The third class of test problems is the classification problem on the fashion MNIST (Han et al., 2017) data set, consisting of 10-classes images with the size 28×28, namely, $N_0 = 784$. Each image is labeled with a corresponding class label, ranging from 0 to 9,

which indicates the type of clothing item in the image (e.g., T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Ankle boot). In practice, we randomly pick up data entries from each class of fashion MNIST under uniform distribution. The dataset consists of 60,000 training images and 10,000 test images. Since there are ten classes in the fashion MNIST, we take $N_L = 10$.

Output evaluation. Finally, we introduce how to evaluate the performance of various approaches. We record the measurements including the training error, the test error, the first feasibility violation, the second feasibility violation, and the KKT violation, which are denoted by

$$\text{TrainErr} = \frac{1}{N}\sum_{n=1}^{N} \|\sigma(W_L\sigma(\cdots\sigma(W_1x_n + b_1) + b_2\cdots) + b_L) - y_n\|^2,$$

$$\text{TestErr} = \frac{1}{N}\sum_{n=N+1}^{N+N_{\text{test}}} \|\sigma(W_L\sigma(\cdots\sigma(W_1x_n + b_1) + b_2\cdots) + b_L) - y_n\|^2,$$

$$\text{FeasVi1} = \frac{1}{N}\sum_{n=1}^{N}\sum_{\ell=1}^{L} \|v_{n,\ell} - \sigma(u_{n,\ell})\|^2, \ \text{FeasVi2} \ = \frac{1}{N}\sum_{n=1}^{N}\sum_{\ell=1}^{L} \|u_{n,\ell} - (W_\ell v_{n,\ell-1} + b_\ell)\|^2,$$

and

$$\text{KKTVi} = \text{dist}(0, \partial\mathcal{L}_{\rho^{(k-1)}}(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)}; \xi^{(k-1)}) + \mathcal{N}_{\Omega_3}(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})) + \frac{1}{2}\text{FeasVi2},$$

respectively, and the average feasibility violation FeasVi=(FeasVi1+FeasVi2)/$\overline{N}$. Time is the CPU time in (minutes: seconds). For the classification task, we also record the classification accuracy for the training data, "Accuracy", and test data, "TestAcc", respectively.

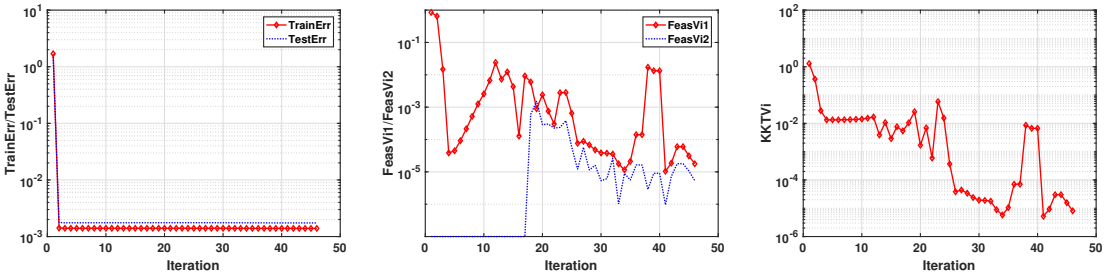
5.2 Numerical Performance of IALAM

In this subsection, we investigate the numerical performance of IALAM in solving problems with both randomly generated data sets and MNIST.

5.2.1 Solving Problems with Different Layers

We test IALAM in solving problem (PP) with different layers. Figure 1 shows the performance of IALAM in solving problem with synthetic data set, where we set $N = 500$, $L = 4$, $N_0 = 5$, $N_1 = 4$, $N_2 = 4$, $N_3 = 3$ and $N_4 = 1$. We can learn from Figure 1 that (i) the training error and the test error decrease in the same order; (ii) the feasibility violations and KKT violation reduce oscillatorily to zero.

Next, we demonstrate the numerical behavior of IALAM in solving problems with the MNIST data set, where $N = 60000$, $N_{\text{test}} = 10000$ and the number of hidden layers up to four (averaged over 100 simulations). We can learn from Table 1 and Table 2 that IALAM works well in dealing with classification data set MNIST with different layers.



(a) TrainErr/TestErr (b) FeasVi1/FeasVi2 (c) KKTVi

Figure 1: Algorithm performance of IALAM on the synthetic data set.

Table 1: Numerical Results of IALAM with up to four layers.

N_1	N_2	N_3	Iter	Time	FeasVi1	FeasVi2	TrainErr	Accuracy	TestAcc
500	–	–	138	47:06	2.00e-5	9.95e-9	(3.98±0.24)e-2	0.981±0.002	0.974±0.002
200	100	–	152	21:58	6.97e-5	4.06e-5	(3.35±0.41)e-2	0.980±0.004	0.976±0.004
500	200	–	125	21:33	1.07e-4	1.97e-5	(1.94±0.62)e-2	0.989±0.003	0.983±0.003
100	100	100	109	6:37	1.19e-4	7.49e-5	(5.23±0.86)e-2	0.947±0.006	0.943±0.005
200	400	200	112	20:10	8.68e-5	5.66e-5	(4.42±1.27)e-2	0.960±0.010	0.959±0.008
800	400	200	95	31:39	9.82e-5	3.74e-5	(5.21±1.31)e-2	0.960±0.012	0.958±0.008

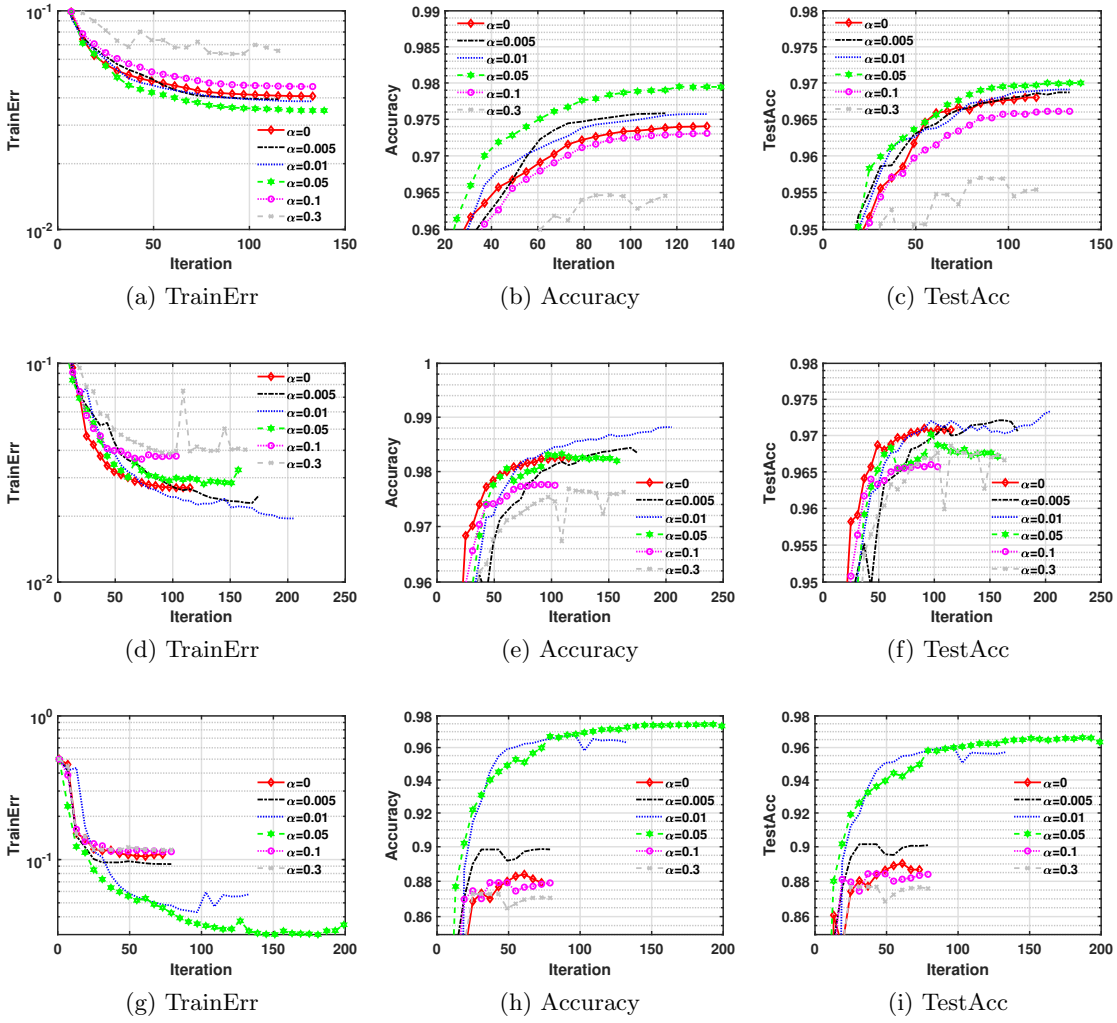
Table 2: Numerical Results of IALAM with up to four layers among 100 times.

	N_1 800	N_2 400	N_3 200	N_1 500	N_2 200	N_3 –	N_1 500	N_2 –	N_3 –
FeasVi1	[7.89e-7, 8.78e-4]			[1.67e-6, 3.09e-4]			[7.32e-7, 1.25e-4]		
FeasVi2	[1.06e-8, 4.29e-4]			[3.86e-9, 1.05e-4]			[7.75e-9, 1.87e-8]		
FeasVi	[6.67e-10, 9.33e-7]			[1.20e-9, 9.42e-5]			[5.30e-10, 8.90e-8]		
TrainErr	[1.94e-2, 6.63e-2]			[1.10e-2, 3.97e-2]			[3.60e-2, 4.61e-2]		
Accuracy	[0.946, 0.986]			[0.980, 0.996]			[0.976, 0.985]		
TestAcc	[0.943, 0.982]			[0.978, 0.990]			[0.969, 0.979]		

5.2.2 Investigating the Model Parameters in Sparse leaky ReLU Network

In this subsection, we first study the numerical performance of IALAM in solving problem (PP) with various leaky ReLU parameters α. Our test is based on the MNIST data set with $N = 60000$ and $N_{\text{test}} = 10000$ and a fixed initialization point. We can learn from Figure 2 that (i) IALAM can be extended to training the ReLU network, i.e. $\alpha = 0$; (ii) a small positive α often leads to better performance than $\alpha = 0$, but further increasing of α yields worse and worse performance.

Then, we study the numerical performance of IALAM in solving problem (PP) with different penalty parameters $\beta := \bar{\beta} e_m$, $\bar{\beta} > 0$ and a fixed initialization point. Our test



(a) TrainErr (b) Accuracy (c) TestAcc

(d) TrainErr (e) Accuracy (f) TestAcc

(g) TrainErr (h) Accuracy (i) TestAcc

Figure 2: Comparisons among IALAM with $N = 60000$, different networks, α and (a)–(c): $N_1 = 500, L = 2$; (d)–(f): $N_1 = 200, N_2 = 100, L = 3$; (g)–(i): $N_1 = 100$, $N_2 = 100$, $N_3 = 100$, $L = 4$.

is also based on the MNIST data set with $N = 60000$ and $N_{\text{test}} = 10000$. We can learn from Table 3 that (a) the bigger $\bar{\beta}$ always leads to slower convergence; (b) $\bar{\beta} = \frac{1}{N}$ performs the best among $\{1/N, 1/10N, 10/N, 100/N\}$. Hence, we choose $\frac{1}{N}e_m$ as the default value of penalty parameter β in IALAM.

Table 3: Comparisons among IALAM with vector $\beta = \bar{\beta}e_m$.

N_1	N_2	N_3	$\bar{\beta}$	Iter	FeasVi1	FeasVi2	FeasVi	TrainErr	Accuracy	TestAcc
500	–	–	1/10N	121	7.64e-6	1.51e-11	1.50e-8	3.66e-3	0.975	0.970
500	–	–	1/N	147	6.02e-9	8.89e-9	2.92e-11	3.88e-3	0.981	0.974
500	–	–	10/N	168	9.05e-9	4.80e-14	1.77e-11	3.71e-3	0.977	0.969
500	–	–	100/N	155	1.21e-8	0	2.37e-11	3.91e-3	0.975	0.967
200	100	–	1/10N	113	7.38e-5	6.52e-5	4.48e-7	3.43e-2	0.970	0.961
200	100	–	1/N	146	6.97e-5	4.06e-5	3.56e-7	2.76e-2	0.984	0.969
200	100	–	10/N	133	4.21e-6	5.99e-11	1.36e-8	2.98e-2	0.980	0.968
200	100	–	100/N	149	6.66e-6	1.61e-12	2.15e-8	2.80e-2	0.981	0.969
100	100	100	1/10N	88	9.53e-4	4.53e-4	4.54e-6	1.95e-1	0.772	0.776
100	100	100	1/N	135	1.73e-4	7.39e-5	7.96e-7	4.04e-2	0.959	0.955
100	100	100	10/N	87	3.94e-5	0	1.27e-7	6.63e-2	0.935	0.927
100	100	100	100/N	99	9.10e-6	0	2.94e-8	6.60e-2	0.934	0.932

5.3 Comparisons with the State-of-the-art Approaches

In this subsection, we compare IALAM with the existing SGD-based approaches including ProxSGD in solving problem (P) through different ways.

5.3.1 Testing on Synthetic Data Sets

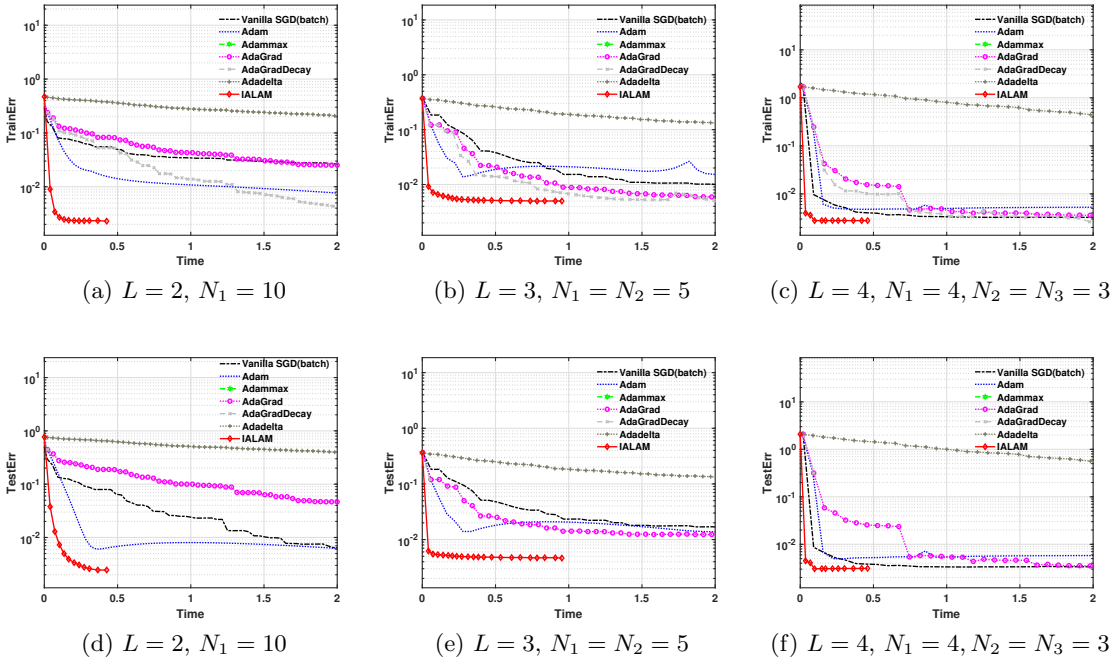
The synthetic data sets are generated with $N = 500$, $N_0 = 5$ and $N_L = 1$. We compare our IALAM with vanilla SGD, Adam, Adammax, AdaGrad, AdaGradDecay and Adadelta. We depict the "TrainErr" and the "TestErr" with the x-axis varying on CPU time. We can learn from Figure 3 that (i) IALAM converges faster than the other approaches; (ii) IALAM can always reach comparable TrainErr and TestErr with the other approaches.

5.3.2 Testing on MNIST Data Set

Now we consider the test on MNIST data set. We first investigate how the "TrainErr" and "Accuracy" with the x-axis varying on "Iteration×batch-size/N", which is equivalent to "iteration" for IALAM and "epoch" for the SGD-based approaches. We also display the "Column Sparsity Ratio" with the x-axis varying on "Tolerance". Here, let $(W_1, \ldots, W_L)$ be the derived weight matrix of solver s, we denote $t^s_{\ell,j} = ||(W_\ell)_{\cdot,j}||_2$ for all $\ell \in [L]$ and $j \in [N_{\ell-1}]$ and for a given tolerance ω, the "Column Sparsity Ratio" r^s_ω of solver s is defined by

$$r^s_\omega := \sum_{\ell=1}^{L} \sum_{j=1}^{N_{\ell-1}} \delta(t^s_{\ell,j} \le \omega) \Big/ \sum_{\ell=0}^{L-1} N_\ell \,,$$

where $\delta(\Gamma) = 1$ if the statement "Γ" is true, otherwise $\delta(\Gamma) = 0$.



(a) $L = 2,\ N_1 = 10$ (b) $L = 3,\ N_1 = N_2 = 5$ (c) $L = 4,\ N_1 = 4, N_2 = N_3 = 3$

(d) $L = 2,\ N_1 = 10$ (e) $L = 3,\ N_1 = N_2 = 5$ (f) $L = 4,\ N_1 = 4, N_2 = N_3 = 3$

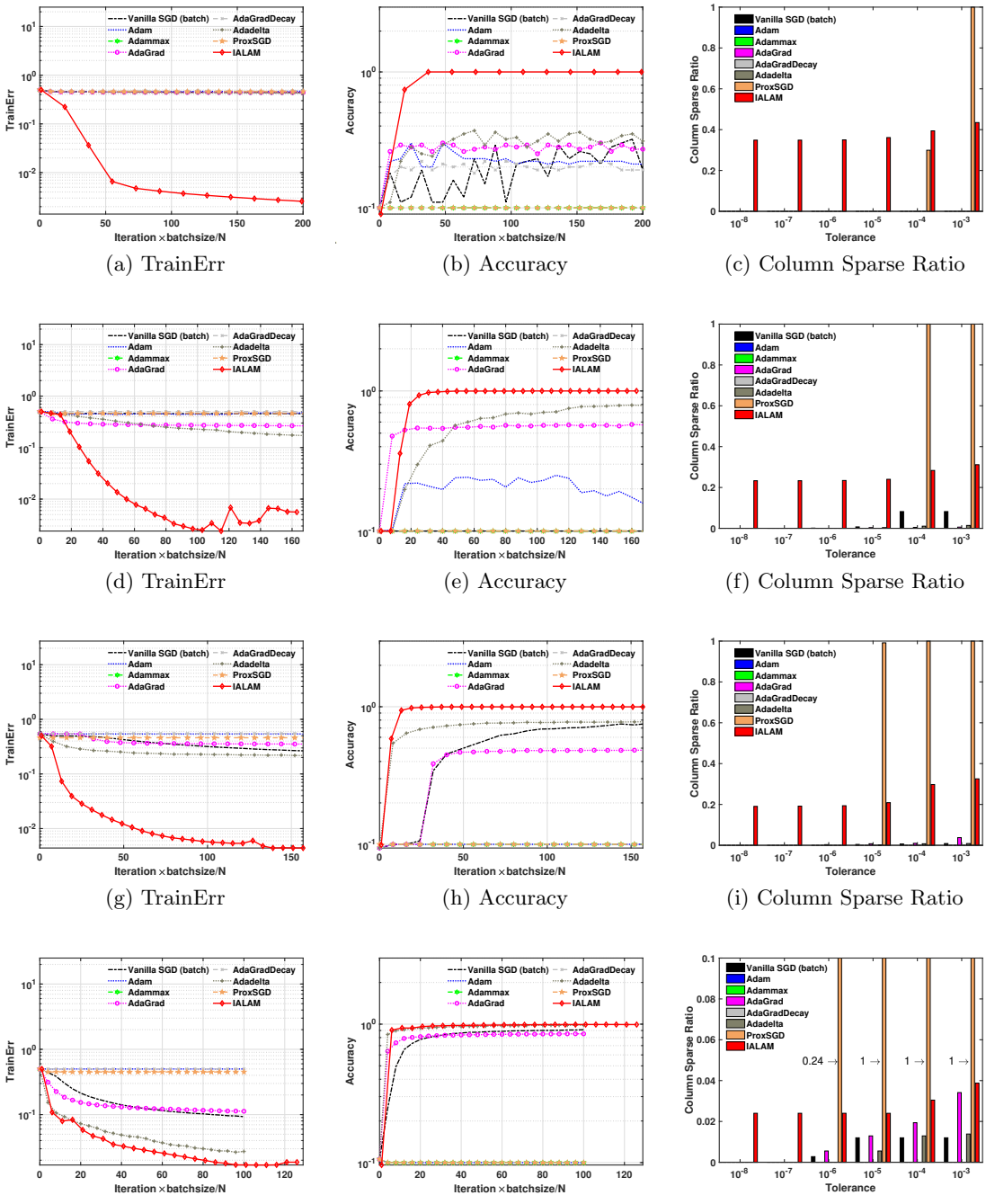
Figure 3: Comparisons among IALAM and SGD-based approaches on the synthetic data set.

We can conclude from Figures 4–5 that (i) IALAM can find sparser solution than the SGD-based approaches; (ii) IALAM can yield comparable TrainErr and Accuracy with other approaches, if not better; (iii) ProxSGD can find as sparse solutions as those of IALAM but much worse behavior on "TrainErr" and "Accuracy".

Finally, we select 720 test problems based on MNIST data set with different network parameter combinations $\{(L = 2, N_1 = 20), (L = 2, N_1 = 50), (L = 3, N_1 = 20, N_2 = 10), (L = 3, N_1 = 50, N_2 = 20), (L = 4, N_1 = 10, N_2 = 10, N_3 = 10), (L = 4, N_1 = 40, N_2 = 20, N_3 = 10)\}$, $\alpha \in \{0, 0.01, 0.05, 0.1\}$, $\lambda_w = i/10N$, $i \in \{1, 2, \ldots, 10\}$, $N \in \{100, 500, 2000\}$. We investigate the performance profiles (Dolan and Moré, 2002) of Vanilla SGD, Adam, Adamdelta and our IALAM through three measurements "TrainErr", and "TestErr".We terminate Valinna SGD, Adadelta and Adam whenever the epoch reaches 100. We describe how to plot the performance profiles. For problem p and solver s, we use t_p^s to represent the output meansurement ("TrainErr" or "TestErr"). Performance ratio is defined as $r_p^s := t_p^s / \min_s \{t_p^s\}$. If solver s fails to solve problem p, the ratio r_p^s is set to 10000. Finally, the overall performance of solver s is defined by

$$\pi_s(\omega) := \sum_{p=1}^{720} \delta(r_p^s \leq \omega) \Big/ 720.$$

Clearly, the closer π_s is to 1, the better performance the solver s has. The performance profiles with respect to "TrainErr" and "TestErr" are given in Figure 6. We can conclude that IALAM outperforms the others with respect to both "TrainErr" and "TestErr". It is



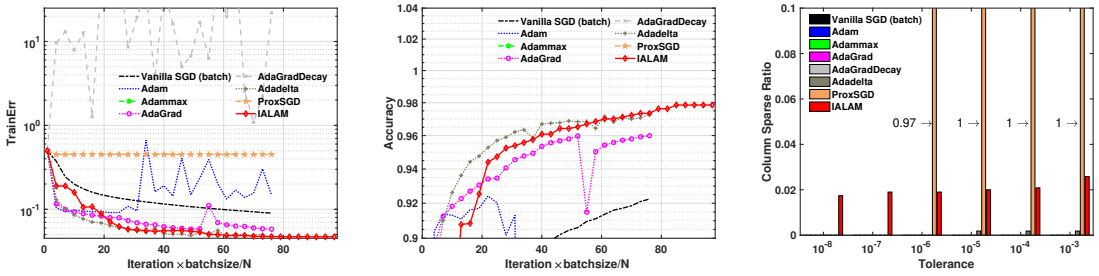
(a) TrainErr (b) Accuracy (c) Column Sparse Ratio

(d) TrainErr (e) Accuracy (f) Column Sparse Ratio

(g) TrainErr (h) Accuracy (i) Column Sparse Ratio

(j) TrainErr (k) Accuracy (l) Column Sparse Ratio

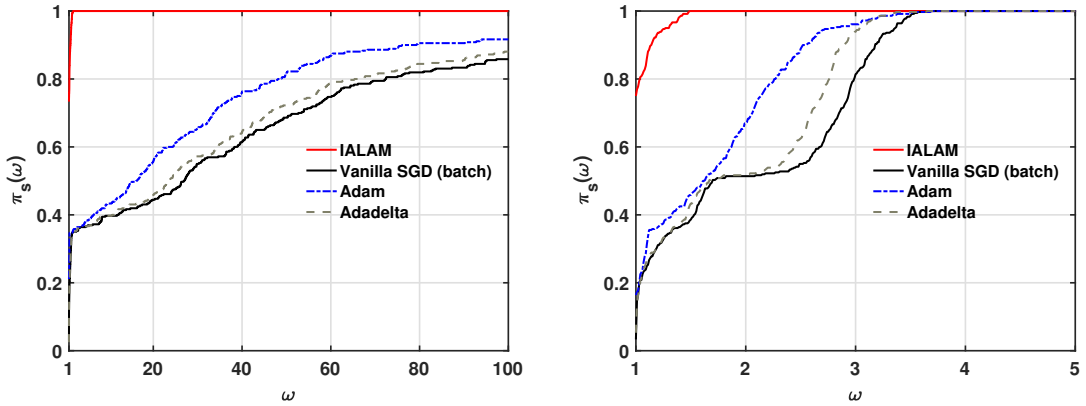
Figure 4: Comparisons among IALAM and SGD-based approaches on MNIST with (a)–(c): $N = 100,\ N_1 = 5,\ L = 2$; (d)–(f): $N = 500, N_1 = 50, N_2 = 20, L = 3$; (g)–(i): $N = 1000, N_1 = 100, N_2 = 50, L = 3$; (j)–(l) $N = 5000, N_1 = 200, N_2 = 100, L = 3$.



(a) TrainErr (b) Accuracy (c) Column Sparse Ratio

Figure 5: Comparisons among IALAM and SGD-based approaches on MNIST with $N = 60000, N_1 = 200, N_2 = 100, L = 3$.

worth noting that we use the same batch-size for Valinna SGD, Adadelta, and Adam, and stop them whenever the the number of epochs reaches 100.

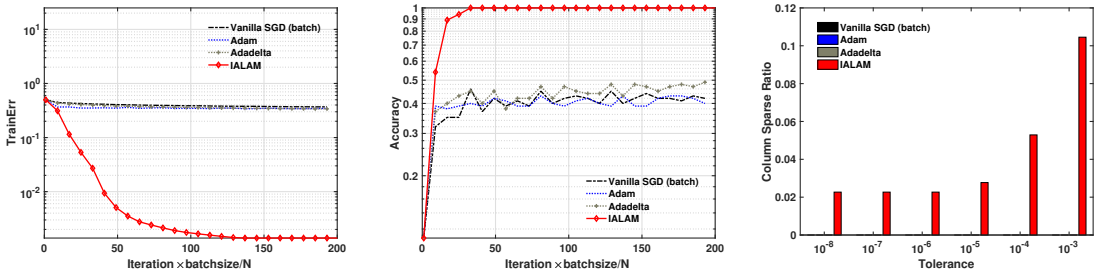


(a) TrainErr (b) TestErr

Figure 6: Performance profile for IALAM, Valinna SGD, Adadelta and Adam on TrainErr and TestErr.

5.3.3 Testing on fashion MNIST Data Set

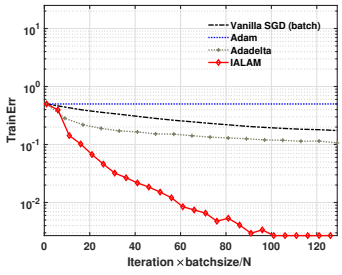
Last but not least, we consider the test on fashion MNIST data set. We investigate the "TrainErr" and "Accuracy" with the x-axis varying on "Iteration×batch-size/N" for Vanilla SGD, Adam, Adamdelta and our IALAM. We also display the "Column Sparsity Ratio" with the x-axis varying on "Tolerance". We can conclude from Figure 7 that (i) IALAM can find sparser solutions than the chosen SGD-based approaches; (ii) IALAM can yield better TrainErr and Accuracy than other approaches.



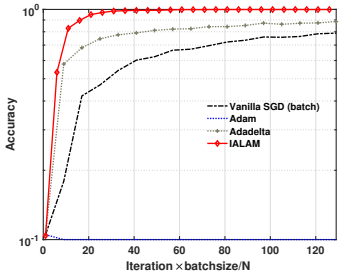
(a) TrainErr

(b) Accuracy

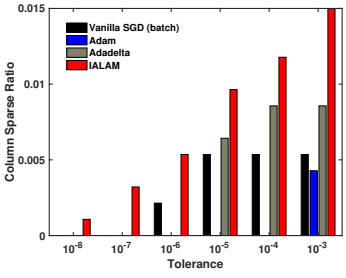
(c) Column Sparse Ratio



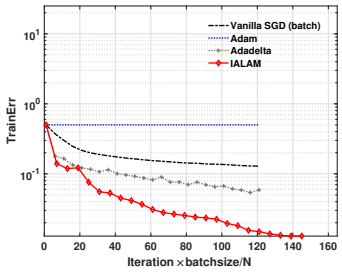
(d) TrainErr



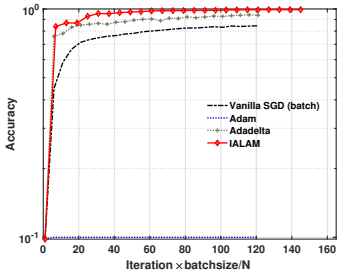
(e) Accuracy



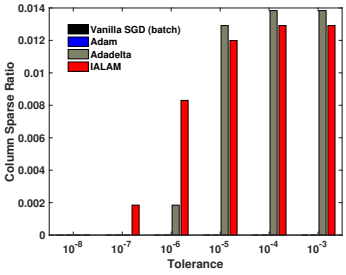
(f) Column Sparse Ratio



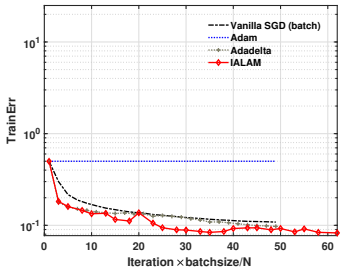
(g) TrainErr



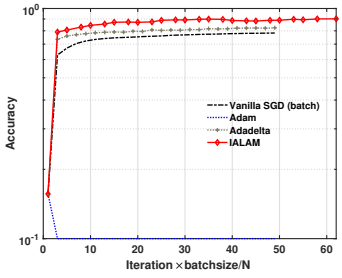
(h) Accuracy



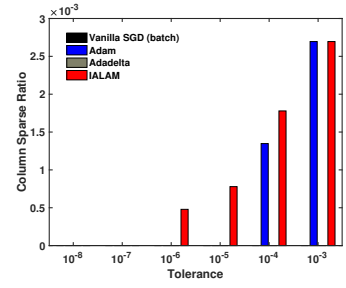
(i) Column Sparse Ratio



(j) TrainErr



(k) Accuracy



(l) Column Sparse Ratio

Figure 7: Comparisons among IALAM and SGD-based approaches on fashion MNIST with (a)–(c): $N = 100,\ N_1 = 10,\ L = 2$; (d)–(f) $N = 1000, N_1 = 100, N_2 = 50, L = 3$; (g)–(i) $N = 5000, N_1 = 200, N_2 = 100, L = 3$; (j)–(l) $N = 60000, N_1 = 500, N_2 = 200, L = 3$.

6. Conclusion

We focus on the regularized minimization model (P) for training leaky ReLU with group sparsity. We first present an l_1-norm penalty model (named PP) for problem (P) and then theoretically demonstrate that these two models share the same global minimizers, local minimizers and limiting stationary points under mild conditions. In addition, we prove that problem (PP) has a nonempty and bounded solution set and its feasible set satisfies the MFCQ, under which the KKT point of (PP) is also an MPCC W-stationary point of problem (P). We propose an inexact augmented Lagrangian algorithm with the alternating minimization (IALAM) to solve problem (PP). The global convergence to the KKT point has been established. Comprehensive numerical experiments have illustrated the efficiency of IALAM as well as its ability to seek sparse solution.

Acknowledgments

We would like to acknowledge support for this project from the National Natural Science Foundation of China (No. 12125108, 11971466, 12288201, 12021001 and 11991021), Hong Kong Research Grants Council grant PolyU15300021, Key Research Program of Frontier Sciences, Chinese Academy of Sciences (No. ZDBS-LY-7022) and the CAS AMSS-PolyU Joint Laboratory in Applied Mathematics. We would like to thank Professor Carreira-Perpiñán and two referees for their helpful comments which helped us to improve the content of this paper.

References

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

A. F. Agarap. Deep learning using rectified linear units (ReLU). *Preprint, arxiv:1803.08375*, 2018.

A. Beck. *First-order Methods in Optimization*. SIAM, 2017.

J. Bolte and E. Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188(1):19–51, 2021.

K. Bui, F. Park, S. Zhang, Y. Qi, and J. Xin. Structured sparsity of convolutional neural networks via nonconvex sparse group regularization. *Frontiers in Applied Mathematics and Statistics*, page 62, 2021.

M. Carreira-Perpiñán and W. Wang. Distributed optimization of deeply nested systems. *CoRR*, abs/1212.5921, 2012. URL `http://arxiv.org/abs/1212.5921`.

M. Carreira-Perpiñán and W. Wang. Distributed optimization of deeply nested systems. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pages 10–19, 2014.

X. Chen, L. Guo, Z. Lu, and J. J. Ye. An augmented lagrangian method for non-lipschitz nonconvex programming. *SIAM Journal on Numerical Analysis*, 55(1):168–193, 2017.

F. H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, Philadelphia, 1990.

H. Cramir. Mathematical methods of statistics. *Princeton University Press, Princeton*, page 500, 1946.

Y. Cui, Z. He, and J.-S. Pang. Multicomposite nonconvex optimization for training deep neural networks. *SIAM Journal on Optimization*, 30(2):1693–1723, 2020.

G. E. Dahl, T. N. Sainath, and G. E. Hinton. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8609–8613, 2013.

Y.-H. Dai and R. Fletcher. Projected barzilai-borwein methods for large-scale box-constrained quadratic programming. *Numerische Mathematik*, 100(1):21–47, 2005.

D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1):119–154, 2020.

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

E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, 2002.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011.

B. Evens, P. Latafat, A. Themelis, J. Suykens, and P. Patrinos. Neural network training as an optimal control problem: An augmented lagrangian approach. *Preprint, arXiv:2103.14343*, 2021.

F. Facchinei and J.-S. Pang. *Finite-dimensional Variational Inequalities and Complementarity Problems*, volume 1. Springer, 2003.

J. Feng and N. Simon. Sparse-input neural networks for high-dimensional nonparametric regression and classification. *Preprint, arXiv:1711.07592*, 2017.

X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep Learning*, volume 1. MIT press, Cambridge, 2016.

L. Guo and X. Chen. Mathematical programs with complementarity constraints and a non-lipschitz objective: optimality and approximation. *Mathematical Programming*, 185(1): 455–485, 2021.

X. Han, R. Kashif, and V. Roland. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

T. Hoefler, D. Alistarh, T. Ben-Nun, N. Dryden, and A. Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124, 2021.

K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pages 2146–2153, 2009.

R. Jenatton, J. Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.

H. Kasai. SGDLibrary: A MATLAB library for stochastic optimization algorithms. *Journal of Machine Learning Research*, 18(215):1–5, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *Preprint, arxiv:1412.6980*, 2014.

T. T.-K. Lau, J. Zeng, B. Wu, and Y. Yao. A proximal block coordinate descent algorithm for deep neural network training. *Preprint, arXiv:1803.09082*, 2018.

Y. LeCun. The mnist database of handwritten digits. 1998.

W. Liu, X. Liu, and X. Chen. Linearly-constrained nonsmooth optimization for training autoencoders. *SIAM Journal on Optimization*, 32: 1931-1957, 2022.

Z. Lu and Y. Zhang. An augmented lagrangian approach for sparse principal component analysis. *Mathematical Programming*, 135(1):149–193, 2012.

A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th International Conference on Machine Learning*, number 1, page 3, 2013.

O. L. Mangasarian. *Nonlinear Programming*. SIAM, 1994.

L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

R. Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization*, 15(6):959–972, 1977.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

D. Pedamonti. Comparison of non-linear activation functions for deep neural networks on mnist classification task. *Preprint, arXiv:1804.02763*, 2018.

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer Science & Business Media, 1998.

S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017.

H. Scheel and S. Scholtes. Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity. *Mathematics of Operations Research*, 25(1):1–22, 2000.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

G. Taylor, R. Burmeister, Z. Xu, B. Singh, A. Patel, and T. Goldstein. Training neural networks without gradients: A scalable admm approach. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2722–2731, 2016.

M. Telgarsky. Deep learning theory lecture notes, 2020.

W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li. Learning structured sparsity in deep neural networks. *Advances in Neural Information Processing Systems*, 29, 2016.

Y. Yang, Y. Yuan, A. Chatzimichailidis, R. G. van Sloun, L. Lei, and S. Chatzinotas. Proxsgd: Training structured neural networks under regularization and constraints. In *Proceedings of the International Conference on Learning Representations*, 2019.

J. Ye and J. Zhang. Enhanced karush–kuhn–tucker condition and weaker constraint qualifications. *Mathematical Programming*, 139, 2013.

J. Yoon and S. J. Hwang. Combined group and exclusive sparsity for deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3958–3966, 2017.

M. D. Zeiler. Adadelta: an adaptive learning rate method. *Preprint, arXiv:1212.5701*, 2012.

J. Zeng, T. T.-K. Lau, S. Lin, and Y. Yao. Global convergence of block coordinate descent in deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7313–7323, 2019.

Y. Zhou, R. Jin, and S. C.-H. Hoi. Exclusive lasso for multi-task feature selection. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 988–995, 2010.

A. Proof to Theorem 17

Proof (a) Case (i): the sequence $\{\rho^{(k)}\}$ is bounded. In this case, the formula (51) can be called at most finite times. That is for some $k_0 > \gamma$, it holds that $\rho^{(k)} = \rho^{(k_0)}$ and

$$\left\|u^{(k)} - \Psi\left(v^{(k)}\right) w^{(k)} - Ab^{(k)}\right\| \leq \eta_1 \max_{t=k-\gamma,\ldots,k-1} \left\|u^{(t)} - \Psi\left(v^{(t)}\right) w^{(t)} - Ab^{(t)}\right\|. \tag{64}$$

for all $k > k_0$. By simple calculation, it then follows that for all $k > k_0$,

$$\max_{t=k,k+1,\ldots,k+\gamma-1} \left\|u^{(t)} - \Psi\left(v^{(t)}\right) w^{(t)} - Ab^{(t)}\right\| \leq \eta_1 \max_{t=k-\gamma,\ldots,k-1} \left\|u^{(t)} - \Psi\left(v^{(t)}\right) w^{(t)} - Ab^{(t)}\right\|, \tag{65}$$

and hence $\lim_{k\to\infty} \max_{t=k-\gamma,\ldots,k-1} \|u^{(t)} - \Psi(v^{(t)})w^{(t)} - Ab^{(t)}\| = 0$. This yields

$$\liminf_{k\to\infty} \left\|u^{(k)} - \Psi\left(v^{(k)}\right) w^{(k)} - Ab^{(k)}\right\| = \lim_{k\to\infty} \left\|u^{(k)} - \Psi\left(v^{(k)}\right) w^{(k)} - Ab^{(k)}\right\| = 0. \tag{66}$$

We next show that for all $k > k_0$,

$$\left\|\xi^{(k)}\right\| \leq \left\|\xi^{(k_0)}\right\| + \rho^{(k_0)} \sum_{i=1}^{k-k_0} \left\|u^{(k_0+i)} - \Psi\left(v^{(k_0+i)}\right) w^{(k_0+i)} - Ab^{(k_0+i)}\right\|. \tag{67}$$

In view of $\rho^{(k)} = \rho^{(k_0)}$ for all $k > k_0$ and the updating rule (49) of $\xi^{(k)}$ that

$$\left\|\xi^{(k_0+i)}\right\| \leq \left\|\xi^{(k_0+i-1)}\right\| + \rho^{(k_0)} \left\|u^{(k_0+i)} - \Psi\left(v^{(k_0+i)}\right) w^{(k_0+i)} - Ab^{(k_0+i)}\right\|$$

for all $i \in \mathbb{N}_+$. Summing up the above inequalities for all $i \in \mathbb{N}_+$ yields (67).

Combining inequalities (64), (65) with (67), we obtain that

$$\begin{aligned}\left\|\xi^{(k)}\right\| &\leq \left\|\xi^{(k_0)}\right\| + \rho^{(k_0)} \sum_{i=1}^{k-k_0} \eta_1^{\lceil i/\gamma \rceil} \max_{t=k_0-\gamma+1,\ldots,k_0} \left\|u^{(t)} - \Psi\left(v^{(t)}\right) w^{(t)} - Ab^{(t)}\right\| \\ &\leq \left\|\xi^{(k_0)}\right\| + \frac{\gamma\eta_1\rho^{(k_0)}}{1-\eta_1} \max_{t=k_0-\gamma+1,\ldots,k_0} \left\|u^{(t)} - \Psi\left(v^{(t)}\right) w^{(t)} - Ab^{(t)}\right\|.\end{aligned}$$

Hence $\{\xi^{(k)}\}$ is bounded.

Furthermore, we obtain by the inequality (47) and the definition of $\mathcal{L}_\rho$ that for all $k \in \mathcal{K}$,

$$\begin{aligned}&\mathcal{O}\left(w^{(k+1)}, v^{(k+1)}, u^{(k+1)}\right) + \frac{\rho^{(k)}}{2} \left\|\frac{\xi^{(k)}}{\rho^{(k)}} + u^{(k+1)} - \Psi\left(v^{(k+1)}\right) w^{(k+1)} - Ab^{(k+1)}\right\|^2 \\ \leq &\theta + \frac{\left\|\xi^{(k)}\right\|^2}{2\rho^{(k)}}.\end{aligned} \tag{68}$$

It follows the inclusion $\left\{\left(w^{(k+1)}, b^{(k+1)}, v^{(k+1)}, u^{(k+1)}\right)\right\} \subseteq \Omega_3$ and the definition of $\mathcal{O}$ that $\left\{\mathcal{O}\left(w^{(k+1)}, v^{(k+1)}, u^{(k+1)}\right)\right\}$ is not less than 0. Using this, $\{\xi^{(k)}\}$ being bounded, $\rho^{(k)} = \rho^{(k_0)}$ for all $k > k_0$ and the relation (68) yield that $\left\{\mathcal{O}\left(w^{(k+1)}, v^{(k+1)}, u^{(k+1)}\right)\right\}$ is bounded. Using a similar method as that in the proof of Theorem 8, we obtain that the sequence

$\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ is bounded. Hence, the sequence $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ has at least one accumulation point.

Case (ii): the sequence $\{\rho^{(k)}\}$ is unbounded. In this case, the set

$$\mathcal{K} := \left\{k : \rho^{(k)} = \max\left\{\rho^{(k-1)}/\eta_2, \left\|\xi^{(k)}\right\|^{1+\eta_3}\right\}\right\} \tag{69}$$

is infinite. Together with the inclusion $\eta_2 \in (0,1)$, it follows that $\left\{\rho^{(k)}\right\} \to \infty$ as $k \to \infty, k \in \mathcal{K}$. For all $k \in \mathcal{K}$, we have $\|\xi^{(k)}\|^{1+\eta_3} \leq \rho^{(k)}$, which further yields $\frac{\left\|\xi^{(k)}\right\|}{\rho^{(k)}} \leq (\rho^{(k)})^{-\eta_3/(1+\eta_3)}$. Together with the fact $\left\{\rho^{(k)}\right\} \to \infty$ as $k \to \infty, k \in \mathcal{K}$, we derive

$$\lim_{k\to\infty, k\in\mathcal{K}} \frac{\left\|\xi^{(k)}\right\|}{\rho^{(k)}} = 0. \tag{70}$$

Similarly, we have $\frac{\left\|\xi^{(k)}\right\|^2}{\rho^{(k)}} \leq (\rho^{(k)})^{(1-\eta_3)/(1+\eta_3)}$ and

$$\lim_{k\to\infty, k\in\mathcal{K}} \frac{\left\|\xi^{(k)}\right\|^2}{\rho^{(k)}} = 0 \tag{71}$$

by $\eta_3 > 1$ and $\left\{\rho^{(k)}\right\} \to \infty$ as $k \to \infty, k \in \mathcal{K}$.

Similarly, we obtain (68) with $k \in \mathbb{N}_+$ replaced by $k \in \mathcal{K}$. Dividing both sides of the above inequality by $\rho^{(k)}/2$, we obtain that for all $k \in \mathcal{K}$,

$$\begin{aligned} &\left\|\frac{\xi^{(k)}}{\rho^{(k)}} + u^{(k+1)} - \Psi\left(v^{(k+1)}\right) w^{(k+1)} - Ab^{(k+1)}\right\|^2 \\ \leq &\frac{2}{\rho^{(k)}}\left(\theta - \mathcal{O}\left(w^{(k+1)}, v^{(k+1)}, u^{(k+1)}\right)\right) + \frac{\left\|\xi^{(k)}\right\|^2}{(\rho^{(k)})^2}. \end{aligned} \tag{72}$$

Using this, the relation (70), $\left\{\mathcal{O}\left(w^{(k+1)}, v^{(k+1)}, u^{(k+1)}\right)\right\}$ is not less than 0, and $\left\{\rho^{(k)}\right\} \to \infty$ as $k \to \infty$ and $k \in \mathcal{K}$, we then derive that

$$\lim_{k\to\infty, k\in\mathcal{K}} \left\|\frac{\xi^{(k)}}{\rho^{(k)}} + u^{(k+1)} - \Psi\left(v^{(k+1)}\right) w^{(k+1)} - Ab^{(k+1)}\right\| = 0, \tag{73}$$

which together with (70) yield

$$\begin{aligned} &\liminf_{k\to\infty} \left\|u^{(k)} - \Psi\left(v^{(k)}\right) w^{(k)} - Ab^{(k)}\right\| \\ = &\lim_{k\to\infty, k\in\mathcal{K}} \left\|u^{(k+1)} - \Psi\left(v^{(k+1)}\right) w^{(k+1)} - Ab^{(k+1)}\right\| = 0. \end{aligned} \tag{74}$$

Combining the relations (68), (71), $\rho^{(k)} > 0$ with the sequence $\left\{\mathcal{O}\left(w^{(k+1)}, v^{(k+1)}, u^{(k+1)}\right)\right\}$ being not less than 0, there exists $\bar{k} > 0$ such that $\mathcal{O}\left(w^{(k+1)}, v^{(k+1)}, u^{(k+1)}\right) < 2\theta$ for all $k > \bar{k}$, $k \in \mathcal{K}$. Using similar arguments as those in the proof of Theorem 8, we obtain that the set $\Omega_{2\theta} = \{(w, b, v, u) \in \Omega_2 : \mathcal{O}(w, v, u) \leq 2\theta\}$ is bounded. Hence the sequence

$\{\left(w^{(k+1)}, b^{(k+1)}, v^{(k+1)}, u^{(k+1)}\right)\}_{k\in\mathcal{K}}$ is bounded. Moreover, the sequence $\{(w^{(k+1)}, b^{(k+1)}, v^{(k+1)}, u^{(k+1)})\}_{k\in\mathcal{K}}$ has at least one accumulation point.

Combining the above two cases, the proof is completed.

(b) Let (w^*, b^*, v^*, u^*) be an accumulation point of $\{(w^{(k+1)}, b^{(k+1)}, v^{(k+1)}, u^{(k+1)})\}_{k\in\mathcal{K}}$, where $\mathcal{K}$ is defined in (69). By what we have proved in the statement (a), the closedness of Ω_2, we obtain that the point (w^*, b^*, v^*, u^*) belongs to the feasible set of problem (PP).

Since (w^*, b^*, v^*, u^*) is an accumulation point of $\left\{\left(w^{(k+1)}, b^{(k+1)}, v^{(k+1)}, u^{(k+1)}\right)\right\}_{k\in\mathcal{K}}$, there exists a subsequence $\left\{(w^{(j_k)}, b^{(j_k)}, v^{(j_k)}, u^{(j_k)})\right\}$ of $\left\{\left(w^{(k+1)}, b^{(k+1)}, v^{(k+1)}, u^{(k+1)}\right)\right\}_{k\in\mathcal{K}}$ such that

$$\lim_{k\to\infty}\left(w^{(j_k)}, b^{(j_k)}, v^{(j_k)}, u^{(j_k)}\right) = (w^*, b^*, v^*, u^*).$$

Using similar arguments as those in the proof of Lemma 4 and the fact $\Omega_3 := \{(w, b, v, u) : w \in \mathbb{R}^{\widetilde{N}}, b \in \mathbb{R}^{\overline{N}}, \mathcal{C}(v, u) \leq 0\}$, we obtain that the MFCQ holds at any feasible point (w, b, v, u) for problem (46). Together with (Rockafellar and Wets, 1998, Theorem 6.14), we obtain that $\mathcal{N}_{\Omega_3}(w, b, v, u) = \{\nabla_{(v,u)}(\mu^\top\mathcal{C}(v, u)) : \ \mu^\top\mathcal{C}(v, u) = 0, \ \mu \in \mathbb{R}^{2m}_+\}$. Together with the inequality (48), and the definition of $\mathcal{L}_\rho(w, b, v, u; \xi)$, there exist $\mu^{(j_k)} \geq 0$ and $\zeta^{(j_k)}$ satisfying $(\mu^{(j_k)})^\top\mathcal{C}(v^{(j_k)}, u^{(j_k)}) = 0$, $\|\zeta^{(j_k)}\| \leq \epsilon_{j_k}$ such that

$$\begin{aligned}\zeta^{(j_k)} \in \partial_{(w,b,v,u)}\mathcal{O}\left(w^{(j_k)}, v^{(j_k)}, u^{(j_k)}\right) + \nabla\Bigg(&\left(\xi^{(j_k-1)}\right)^\top\left(u^{(j_k)} - \Psi\left(v^{(j_k)}\right)w^{(j_k)} - Ab^{(j_k)}\right)\\ &+\frac{\rho^{(j_k-1)}}{2}\left\|u^{(j_k)} - \Psi\left(v^{(j_k)}\right)w^{(j_k)} - A^{(j_k)}b^{(j_k)}\right\|^2 + \left(\mu^{(j_k)}\right)^\top\mathcal{C}\left(v^{(j_k)}, u^{(j_k)}\right)\Bigg).\end{aligned} \tag{75}$$

In view of $\xi^{(j_k)} = \xi^{(j_k-1)} + \rho^{(j_k-1)}\left(u^{(j_k)} - \Psi(v^{(j_k)})w^{(j_k)} - Ab^{(j_k)}\right)$ and $\nabla_z\frac{1}{2}\|z\|^2 = z\nabla_z z$ for $z \in \mathbb{R}^m$, we then obtain that

$$\begin{aligned}\zeta^{(j_k)} \in&\partial_{(w,b,v,u)}\mathcal{O}\left(w^{(j_k)}, v^{(j_k)}, u^{(j_k)}\right)\\ &+\nabla\left(\left(\xi^{(j_k)}\right)^\top\left(u^{(j_k)} - \Psi\left(v^{(j_k)}\right)w^{(j_k)} - Ab^{(j_k)}\right) + \left(\mu^{(j_k)}\right)^\top\mathcal{C}\left(v^{(j_k)}, u^{(j_k)}\right)\right).\end{aligned} \tag{76}$$

It holds from the update rule in Algorithm 1 that $\epsilon_k \leq \max\{\sqrt{\eta_1}, \eta_4\}\epsilon_{k-1}$ for all $k > \gamma$. Together with the relationship $0 < \eta_1, \eta_4 < 1$, we derive $\lim_{k\to\infty}\epsilon_k = 0$. It then follows that $\zeta^{(j_k)} \to 0$ as $k \to \infty$.

Let $r_k := \max\left\{\|\xi^{(j_k)}\|_\infty, \|\mu^{(j_k)}\|_\infty\right\}$. Suppose that $\{r_k\}$ is unbounded. Without loss of generality, we assume that as $k \to \infty$, it holds that

$$\frac{\xi^{(j_k)}}{r_k} \to \xi^*, \text{ and } \frac{\mu^{(j_k)}}{r_k} \to \mu^*. \tag{77}$$

It then holds that $\max\{\|\xi^*\|_\infty, \|\mu^*\|_\infty\} = 1$ and $\mu^* \geq 0$, since $\mu^{(j_k)} \geq 0$ for all $k \geq 0$.

Dividing by r_k and taking the limit $k \to \infty$ on the both sides of (76), we obtain

$$0 = \nabla_{(w,b,v,u)}\left((\xi^*)^\top(u^* - \Psi(v^*)w^* - Ab^*) + (\mu^*)^\top\mathcal{C}(v^*, u^*)\right).$$

This together with Lemma 4 and the equality $\max\{\|\xi^*\|_\infty, \|\mu^*\|_\infty\} = 1$ lead to a contradiction. $\{r_k\}$ is hence bounded. Without loss of generality, we assume that as $k \to \infty$,

$$\xi^{(j_k)} \to \xi^*, \text{ and } \mu^{(j_k)} \to \mu^*.$$

Since $\mu^{(j_k)} \geq 0$ for all $k \geq 0$, we have $\mu^* \geq 0$.

Taking the limit $k \to \infty$ on both sides of (76), we obtain that

$$0 \in \partial_{(w,b,v,u)} \left(\mathcal{O}(w^*, v^*, u^*) + (\xi^*)^\top (u^* - \Psi(v^*)w^* - Ab^*) + (\mu^*)^\top \mathcal{C}(v^*, u^*) \right),$$

which together with the inequality $\mu^* \geq 0$ yield that (w^*, b^*, v^*, u^*) is a KKT point of problem (PP). Hence $\liminf_{k\to\infty} \operatorname{dist}((w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)}), \mathcal{Z}^*) = 0$ by the definition of (w^*, b^*, v^*, u^*), which completes the statement (b).

(c) Case (i): the set $\mathcal{K}$ defined in (69) is finite. By what we have proved in the statement (a), we have $\lim_{k\to\infty} \left\| u^{(k)} - \Psi\left(v^{(k)}\right) w^{(k)} - Ab^{(k)} \right\| = 0$ in this case. Using a similar method as that in the proof of statement (b), any accumulation point (w^*, b^*, v^*, u^*) of $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ is a KKT point of problem (PP).

Case (ii): the set $\mathcal{K}$ defined in (69) is infinite.

For any given $k \in \mathbb{N}_+$, let t_k be the largest element in $\mathcal{K}$ satisfying $t_k \leq k$. We then show that

$$\frac{\left\|\xi^{(k)}\right\|}{\rho^{(k)}} \leq \frac{\left\|\xi^{(t_k)}\right\|}{\rho^{(t_k)}} + \sum_{i=1}^{k-t_k} \left\| u^{(t_k+i)} - \Psi\left(v^{(t_k+i)}\right) w^{(t_k+i)} - Ab^{(t_k+i)} \right\|. \tag{78}$$

Clearly, the inequality (78) holds when $k = t_k$. We now suppose $k > t_k$. In view of the fact that $\rho^{(t_k+i)} = \rho^{(t_k)}$ for all $0 < i \leq k - t_k$ and the updating rule (49) of $\xi^{(k)}$, we have

$$\frac{\left\|\xi^{(t_k+i)}\right\|}{\rho^{(t_k+i)}} = \frac{\left\|\xi^{(t_k+i)}\right\|}{\rho^{(t_k+i-1)}} \leq \left\| \frac{\xi^{(t_k+i-1)}}{\rho^{(t_k+i-1)}} \right\| + \left\| u^{(t_k+i)} - \Psi\left(v^{(t_k+i)}\right) w^{(t_k+i)} - Ab^{(t_k+i)} \right\|$$

for all $i \in [k - t_k]$. Summing up the above inequalities for all $i \in [k - t_k]$ yields (78).

For all $i \in [k - t_k]$, we obtain from $\gamma = 1$, (50) and the definition of t_k that

$$\begin{aligned} &\left\| u^{(t_k+i)} - \Psi\left(v^{(t_k+i)}\right) w^{(t_k+i)} - Ab^{(t_k+i)} \right\| \\ \leq &\eta_1 \left\| u^{(t_k+i-1)} - \Psi\left(v^{(t_k+i-1)}\right) w^{(t_k+i-1)} - Ab^{(t_k+i-1)} \right\|. \end{aligned} \tag{79}$$

Together with the inequality (78), we derive

$$\begin{aligned} \frac{\left\|\xi^{(k)}\right\|}{\rho^{(k)}} &\leq \frac{\left\|\xi^{(t_k)}\right\|}{\rho^{(t_k)}} + \sum_{i=1}^{k-t_k} \eta^{i-1} \left\| u^{(t_k+1)} - \Psi\left(v^{(t_k+1)}\right) w^{(t_k+1)} - Ab^{(t_k+1)} \right\| \\ &\leq \frac{\left\|\xi^{(t_k)}\right\|}{\rho^{(t_k)}} + \frac{1}{1-\eta_1} \left\| u^{(t_k+1)} - \Psi\left(v^{(t_k+1)}\right) w^{(t_k+1)} - Ab^{(t_k+1)} \right\|. \end{aligned} \tag{80}$$

Together with the equalities (70), (74), $t_k \in \mathcal{K}$ and $\mathcal{K}$ being infinite, we can conclude that

$$\lim_{k\to\infty} \frac{\left\|\xi^{(k)}\right\|}{\rho^{(k)}} = 0. \tag{81}$$

Similarly, we obtain (72) with $k \in \mathcal{K}$ replaced by $k \in \mathbb{N}_+$. This together with the fact (81) and the lower boundedness of $\mathcal{O}\left(w^{(k+1)}, v^{(k+1)}, u^{(k+1)}\right)$ imply that

$$\lim_{k\to\infty} \left\| u^{(k)} - \Psi\left(v^{(k)}\right) w^{(k)} - Ab^{(k)} \right\| = \lim_{k\to\infty} \left\| u^{(k+1)} - \Psi\left(v^{(k+1)}\right) w^{(k+1)} - Ab^{(k+1)} \right\| = 0.$$

Using a similar method as that in the proof of statement (b), any accumulation point (w^*, b^*, v^*, u^*) of $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ is a KKT point of problem (PP).

This proof is then completed by summarizing the above two cases. ■

B. Proof to Theorem 18

Proof (a) Since we have $S_\ell^{(\jmath)} \succeq \tau_1 I_{N_\ell + N_{\ell-1}}$ for all $\ell = 2, 3, \ldots, L$, we obtain that

$$\begin{aligned}
&\mathcal{P}\left(u^{(\jmath+1)}, v^{(\jmath+1)}; u^{(\jmath)}, v^{(\jmath)}, \tau^{(\jmath)}\right) \\
=&\frac{1}{2}\sum_{n=1}^{N}\sum_{\ell=2}^{L}\left\|\begin{pmatrix} v_{n,\ell-1}^{(\jmath+1)} \\ u_{n,\ell}^{(\jmath+1)} \end{pmatrix} - \begin{pmatrix} v_{n,\ell-1}^{(\jmath)} \\ u_{n,\ell}^{(\jmath)} \end{pmatrix}\right\|_{S_\ell^{(\jmath)}}^2 + \frac{\tau_1}{2}\sum_{n=1}^{N}\left\|u_{n,1}^{(\jmath+1)} - u_{n,1}^{(\jmath)}\right\|^2 \\
\geq&\frac{\tau_1}{2}\left\|u^{(\jmath+1)} - u^{(\jmath)}\right\|^2 + \frac{\tau_1}{2}\sum_{n=1}^{N}\sum_{\ell=1}^{L-1}\left\|v_{n,\ell}^{(\jmath+1)} - v_{n,\ell}^{(\jmath)}\right\|^2.
\end{aligned}$$

Together with the fact that $(u^{(\jmath+1)}, v^{(\jmath+1)})$ being the global minimizer of (57), we obtain

$$\begin{aligned}
&\mathcal{L}_\rho\left(w^{(\jmath+1)}, b^{(\jmath+1)}, v^{(\jmath+1)}, u^{(\jmath+1)}; \xi\right) - \mathcal{L}_\rho\left(w^{(\jmath+1)}, b^{(\jmath+1)}, v^{(\jmath)}, u^{(\jmath)}; \xi\right) \\
\leq&-\frac{\tau_1}{2}\left\|u^{(\jmath+1)} - u^{(\jmath)}\right\|^2 - \frac{\tau_1}{2}\sum_{n=1}^{N}\sum_{\ell=1}^{L-1}\left\|v_{n,\ell}^{(\jmath+1)} - v_{n,\ell}^{(\jmath)}\right\|^2.
\end{aligned} \tag{82}$$

Since $\mathcal{R}_1$ is λ_w-strongly convex with respect to w (Beck, 2017), we derive from the definition of $\mathcal{L}_\rho$ and the updating rule (53) that

$$\mathcal{L}_\rho\left(w^{(\jmath+1)}, b^{(\jmath+1)}, v^{(\jmath)}, u^{(\jmath)}; \xi\right) - \mathcal{L}_\rho\left(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)}, u^{(\jmath)}; \xi\right) \leq -\frac{\lambda_w}{2}\left\|w^{(\jmath+1)} - w^{(\jmath)}\right\|^2,$$

which together with the inequality (82) complete the statement (a).

(b) The statement (a) together with the inequality $\mathcal{L}_\rho(w_{\text{init}}^{(k)}, b_{\text{init}}^{(k)}, v_{\text{init}}^{(k)}, u_{\text{init}}^{(k)}; \xi) \leq \theta$ and the definition of $\mathcal{L}_\rho$, we obtain that

$$-\frac{1}{2\rho}\|\xi\|^2 \leq \mathcal{L}_\rho\left(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)}, u^{(\jmath)}; \xi\right) \leq \theta, \quad \text{for all } \jmath \in \mathbb{N}. \tag{83}$$

Then, the non-increasing sequence $\{\mathcal{L}_\rho\left(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)}, u^{(\jmath)}; \xi\right)\}$ is bounded and hence convergent.

(c) Recall the definition of $\mathcal{L}_\rho$ and (83), we obtain that $\lambda_w\|w\|_{2,1} + \lambda_v\|v\|^2 \leq \frac{1}{2\rho}\|\xi\|^2 + \theta$, then the sequence $\{(w^{(\jmath)}, v^{(\jmath)})\}$ is bounded. Together with the definition of $\mathcal{L}_\rho$ and (83), it holds that $\{(b^{(\jmath)})\}$ is bounded. Since $\mathcal{O}$ is not less than 0, (83) also yields that

$$\frac{\rho}{2}\left\|\frac{\xi}{\rho} + u^{(\jmath)} - \Psi(v^{(\jmath)})w^{(\jmath)} - Ab^{(\jmath)}\right\|^2 \leq \theta + \frac{1}{2\rho}\|\xi\|^2$$

for all $\jmath \in \mathbb{N}_+$, which together with the boundedness of $\{(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)})\}$ imply that $\{u^{(\jmath)}\}$ is bounded. Hence, the sequence $\{(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)}, u^{(\jmath)})\}$ is bounded, which completes the statement (c).

(d) From the statements (a), (b), (c), and $\tau_1 > 0$, we obtain that

$$\lim_{\jmath\to\infty} \left\|w^{(\jmath+1)} - w^{(\jmath)}\right\|^2 = 0, \lim_{\jmath\to\infty} \left\|u^{(\jmath+1)} - u^{(\jmath)}\right\|^2 = 0, \lim_{\jmath\to\infty} \sum_{n=1}^{N}\sum_{\ell=1}^{L-1} \left\|v_{n,\ell}^{(\jmath+1)} - v_{n,\ell}^{(\jmath)}\right\|^2 = 0. \quad (84)$$

From the updating rule of the (w, b), we obtain from the KKT condition of (53) and the definition of A, Ψ that for all $\ell \in [L]$, it holds that

$$b_\ell^{(\jmath+1)} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\xi_{n,\ell}}{\rho} + u_{n,\ell}^{(\jmath)} - W_\ell v_{n,\ell-1}^{(\jmath)}\right).$$

This together with the statement (c) and (84) imply that

$$\lim_{\jmath\to\infty} \left\|b^{(\jmath+1)} - b^{(\jmath)}\right\|^2 = 0. \quad (85)$$

From the updating rule of the (v, u), (60) and (61), we obtain that $v_{n,L}^{(\jmath+1)}$ has a closed-form associated with $\xi_{n,L}$, ρ, $\tau_L^{(\jmath)}$, $u_{n,L}^{(\jmath)}$, $W_L^{(\jmath+1)}$, $v_{n,L-1}^{(\jmath)}$, and $b_L^{(\jmath+1)}$ for all $n \in [N]$. Together with the facts (55) and (84), it holds that

$$\lim_{\jmath\to\infty} \sum_{n=1}^{N} \left\|v_{n,L}^{(\jmath+1)} - v_{n,L}^{(\jmath)}\right\|^2 = 0.$$

Using this and relations (84), (85), we complete the statement (d).

(e) The statement (c) yields that the sequence $\{(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)}, u^{(\jmath)})\}$ has at least one accumulation point. Let $\mathcal{J}$ be a index set of $\{(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)}, u^{(\jmath)})\}$ such that

$$\lim_{\jmath\to\infty, \jmath\in\mathcal{J}} \left(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)}, u^{(\jmath)}\right) = (w^*, b^*, v^*, u^*).$$

From the first-order optimality conditions for the updating schemes in Steps 1-2 of Algorithm 2 and the constraint set $\mathcal{C}$ being separable with respect to (v, u) block and (w, b) block, there exists $\mu^{(\jmath)} \in \mathbb{R}_+^{2m}$ such that

$$\begin{cases} 0 \in \partial_{(w,b)}\mathcal{L}_\rho\left(w^{(\jmath+1)}, b^{(\jmath+1)}, v^{(\jmath)}, u^{(\jmath)}; \xi\right) \\ 0 \in \partial_{(v,u)}\mathcal{L}_\rho\left(w^{(\jmath+1)}, b^{(\jmath+1)}, v^{(\jmath+1)}, u^{(\jmath+1)}; \xi\right) \\ \qquad + \nabla_{(v,u)}\left(\mathcal{P}\left(u^{(\jmath+1)}, v^{(\jmath+1)}; u^{(\jmath)}, v^{(\jmath)}, \tau^{(\jmath)}\right) + \left(\mu^{(\jmath)}\right)^\top \mathcal{C}\left(v^{(\jmath+1)}, u^{(\jmath+1)}\right)\right) \\ \left(\mu^{(\jmath)}\right)^\top \mathcal{C}\left(v^{(\jmath+1)}, u^{(\jmath+1)}\right) = 0. \end{cases} \quad (86)$$

Let $r_\jmath := \|\mu^{(\jmath)}\|_\infty$. Suppose that $\{r_\jmath\}$ is unbounded. Without loss of generality, we assume that as $\jmath \to \infty, \jmath \in \mathcal{J}$, it holds that $\frac{\mu^{(\jmath)}}{r_\jmath} \to \bar{\mu}^*$. It then follows the fact $\mu^{(\jmath)} \geq 0$ for all $\jmath \in \mathbb{N}$ that $\|\bar{\mu}^*\|_\infty = 1$ and $\bar{\mu}^* \geq 0$.

Dividing by $r_\jmath$ and taking the limit $\jmath \to \infty, \jmath \in \mathcal{J}$ on both sides of (86), we obtain $0 = \nabla_{(v,u)}(\mu^*)^\top \mathcal{C}(v^*, u^*)$ which results from the locally Lipschitz continuity of $\mathcal{L}_\rho, \mathcal{C}, \mathcal{P}, \nabla\mathcal{C}$ and the statements (a)–(d). Together with the equality $\|\bar{\mu}^*\|_\infty = 1$ and Lemma 4, it leads to contradiction. Thus, $\{r_\jmath\}$ is bounded as desired. Without loss of generality, we assume that as $\jmath \to \infty, \jmath \in \mathcal{J}$, it holds that

$$\mu^{(\jmath)} \to \bar{\mu}^*. \tag{87}$$

Similarly, we can obtain that $\bar{\mu}^* \geq 0$, since $\mu^{(\jmath)} \geq 0$ for all $\jmath \in \mathbb{N}$.

Again, taking the limit $\jmath \to \infty, \jmath \in \mathcal{J}$ on both sides of (86), we finally arrive at $(\bar{\mu}^*)^\top \mathcal{C}(v^*, u^*) = 0$ and

$$0 \in \partial_{(w,b)} \mathcal{L}_\rho(w^*, b^*, v^*, u^*; \xi),\ 0 \in \partial_{(v,u)} \left(\mathcal{L}_\rho(w^*, b^*, v^*, u^*; \xi) + (\bar{\mu}^*)^\top \mathcal{C}(v^*, u^*) \right).$$

This completes the proof. ■