AN AUGMENTED LAGRANGIAN METHOD FOR TRAINING RECURRENT NEURAL NETWORKS*

YUE WANG†, CHAO ZHANG‡, AND XIAOJUN CHEN§

Abstract. Recurrent Neural Networks (RNNs) are widely used to model sequential data in a wide range of areas, such as natural language processing, speech recognition, machine translation, and time series analysis. In this paper, we model the training process of RNNs with the ReLU activation function as a constrained optimization problem with a smooth nonconvex objective function and piecewise smooth nonconvex constraints. We prove that any feasible point of the optimization problem satisfies the no nonzero abnormal multiplier constraint qualification (NNAMCQ), and any local minimizer is a Karush-Kuhn-Tucker (KKT) point of the problem. Moreover, we propose an augmented Lagrangian method (ALM) and design an efficient block coordinate descent (BCD) method to solve the subproblems of the ALM. The update of each block of the BCD method has a closed-form solution. The stop criterion for the inner loop is easy to check and can be stopped in finite steps. Moreover, we show that the BCD method can generate a directional stationary point of the subproblem. Furthermore, we establish the global convergence of the ALM to a KKT point of the constrained optimization problem. Compared with the state-of-the-art algorithms, numerical results demonstrate the efficiency and effectiveness of the ALM for training RNNs.

Key words. recurrent neural network, nonsmooth nonconvex optimization, augmented Lagrangian method, block coordinate descent

MSC codes. 65K05, 90B10, 90C26, 90C30

1. Introduction.

Recurrent Neural Networks (RNNs) have been applied in a wide range of areas, such as speech recognition [15, 27], natural language processing [22, 28] and nonlinear time series forecasting [1, 23]. In this paper, we focus on the Elman RNN architecture [13], one of the earliest and most fundamental RNNs, and use Elman RNNs to deal with the regression task with the least squares loss function.

Given input data $x_t \in \mathbb{R}^n$ and output data $y_t \in \mathbb{R}^m$, $t = 1, \dots, T$, a widely used minimization problem for training RNNs is represented as (see [14, pp. 381])

$$\min_{A,W,V,b,c} \frac{1}{T}\sum_{t=1}^{T}\left\| y_t - \left(A\sigma\Big(W\big(...\sigma(Vx_1+b)...\big)+Vx_t+b\Big)+c\right)\right\|^2, \tag{1.1}$$

where $W \in \mathbb{R}^{r\times r}$, $V \in \mathbb{R}^{r\times n}$ and $A \in \mathbb{R}^{m\times r}$ are unknown weight matrices, $b \in \mathbb{R}^r$ and $c \in \mathbb{R}^m$ are unknown bias vectors, and $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonsmooth activation function that is applied component-wise on vectors and transforms the previous information and the input data x_t into the hidden layer at time t. The training process by (1.1) can be interpreted as looking for proper weight matrices A, W, V, and bias vectors b, c in RNNs to minimize the difference between the true value y_t and the output from RNNs across all time steps. It is worth mentioning that the Elman RNNs in (1.1) shares the same weight matrices and bias vectors at different time steps [14, pp. 374].

*Submitted on 28 December 2023.

Funding: This work was supported in part by Hong Kong Research Grants Council C5036-21E and National Science Foundation of China (No. 12171027).

†Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, China (yueyue.wang@connect.polyu.hk).

‡School of Mathematics and Statistics, Beijing Jiaotong University, Beijing 100044, China (zc.njtu@163.com).

§Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, China (maxjchen@polyu.edu.hk).

When the traditional backpropagation through time (BPTT) method is used to train RNNs, the highly nonlinear and nonsmooth composition function presented in (1.1) poses significant challenges. Gradient descent methods (GDs), as well as stochastic gradient descent-based methods (SGDs), are widely used to train RNNs in practice [8, 30], but the "gradient" of the loss function associated with the weighted matrices via the "chain rule" is calculated even if the "chain rule" does not hold. The "gradients" might exponentially increase to a very large value or shrink to zero as time t increases, which makes RNNs training with large time length T very challenging [4]. To overcome this shortcoming, various techniques have been developed, such as gradient clipping [22], gradient descent with Nesterov momentum [3], initialization with small values [24], adding sparse regularization [2], and so on. Because the essence of the above methods is to restrict the initial values of weighted matrices or gradients, they are sensitive to the choice of initial values [18]. Moreover, GDs and SGDs for training RNNs lack rigorous convergence analysis.

The objective function in (1.1) is nonsmooth nonconvex and has a highly composite structure. In this paper, we equivalently reformulate (1.1) as a constrained optimization problem with a simple smooth objective function by utilizing auxiliary variables to represent the composition structures and treating these representations as constraints. Moreover, we propose an augmented Lagrangian method (ALM) for the constrained optimization problem with ℓ_2-norm regularization, and design a block coordinate descent (BCD) method to solve the subproblem of the ALM at every iteration. The solution of the subproblems of the BCD method is very easy to compute with a closed-form. Utilizing auxiliary variables to reformulate highly nonlinear composite structured problems as constrained optimization problems has been adopted for training Deep Neural Networks (DNNs) [7, 12, 19, 20, 31]. However, these algorithms for DNNs cannot be used for RNNs directly because of the difference between their architectures. In fact, RNNs share the same weighted matrices and bias vectors across different layers, whereas DNNs have distinct weighted matrices and bias vectors in different layers. In DNNs, the weighted matrices and bias vectors can be updated layer by layer, allowing for the separation of the gradient calculation across different layers. However, in RNNs, the weighted matrices and bias vectors need to be updated simultaneously. Therefore, it is necessary to establish effective algorithms tailored to the characteristics of RNNs. To the best of our knowledge, the proposed ALM in this paper is the first first-order optimization method for training RNNs with solid convergence results.

Recently, several augmented Lagrangian-based methods have been proposed for nonconvex nonsmooth problems with composite structures. In [9], Chen et al. proposed an ALM for non-Lipschitz nonconvex programming, which requires the constraints to be smooth. Hallak and Teboulle in [16] transformed a comprehensive class of optimization problems into constrained problems with smooth constraints and nonsmooth nonconvex objective functions, and proposed a novel adaptive augmented Lagrangian-based method to solve the constrained problem. The assumption on the smoothness of constraints in [9, 16] is not satisfied for the optimization problem arising in training RNNs with nonsmooth activation functions considered in this paper.

Our contributions are summarized as follows:

- We prove that the solution set of the constrained problem with ℓ_2 regularization is nonempty and compact. Furthermore, we prove that any feasible point of the constrained optimization problem satisfies the no nonzero abnormal multiplier constraint qualification (NNAMCQ), which immediately guarantees any local minimizer of the constrained problems is a Karush-Kuhn-Tucker (KKT) point.

- We show that any accumulation point of the sequence generated by the BCD method is a directional stationary point of the subproblem. Moreover, we show that in the k-th iteration of the ALM, the stopping criterion of the BCD method for solving the subproblem can be satisfied within $O\big(1/(\epsilon_{k-1})^2\big)$ finite steps for any $\epsilon_{k-1} > 0$.
- We show that there exists an accumulation point of the sequence generated by the ALM for solving the constrained optimization problem with regularization and any accumulation point of the sequence is a KKT point.
- We compare the performance of the ALM with several state-of-the-art methods for both synthetic and real datasets. The numerical results verify that our ALM outperforms other algorithms in terms of forecasting accuracy for both the training sets and the test sets.

The rest of the paper is organized as follows. In section 2, we equivalently reformulate problem (1.1) as a nonsmooth nonconvex constrained minimization problem with a simple smooth objective function. Then we show that the solution set of the constrained problem with regularization is nonempty and bounded, and give the first-order necessary optimality conditions for the constrained problem and the regularized problem. We propose the ALM for the constrained problem with regularization, as well as the BCD method for the subproblems of the ALM in section 3. We establish the convergence results of the BCD method, and the ALM in section 4. Finally, we conduct numerical experiments on both the synthetic and real data in section 5, which demonstrate the effectiveness and efficiency of the ALM for the reformulated optimization problem.

Notation and terminology. Let $\mathbb{N}_+$ denote the set of positive integers. For column vectors $\pi_1, \pi_2, \ldots, \pi_l$, let us denote by $\boldsymbol{\pi} := (\pi_1; \pi_2; \ldots; \pi_l) = (\pi_1^\top, \pi_2^\top, \ldots, \pi_l^\top)^\top$ a column vector. For a given matrix $D \in \mathbb{R}^{k\times l}$, we denote by $D_{\cdot j}$ the j-th column of D and use $\mathrm{vec}(D) = (D_{\cdot 1}; D_{\cdot 2}; \ldots; D_{\cdot l}) \in \mathbb{R}^{kl}$ to represent a column vector. For a given vector g, we use $\mathrm{diag}(g)$ to represent the diagonal matrix, whose (i,i)-entry is the i-th component g_i of g. We use $\boldsymbol{e}_l$ to represent the vector of all ones in $\mathbb{R}^l$. For $\nu \in \mathbb{R}$, $\lceil \nu \rceil$ refers to the smallest integer that is greater than ν. For a given $N \in \mathbb{N}_+$, we denote $[N] := \{1, 2, \ldots, N\}$. We use $\|\cdot\|$ and $\|\cdot\|_\infty$ to denote the ℓ_2-norm and infinity norm of a vector or a matrix, respectively. We denote by $\|\cdot\|_F$ the Frobenius norm of a matrix.

Let $f : \mathbb{R}^{n_1} \to \mathbb{R}$ be a proper lower semicontinuous function defined on $\mathbb{R}^{n_1}$. The notation $x^k \xrightarrow{f} \bar{x}$ means that $x^k \to \bar{x}$ and $f(x^k) \to f(\bar{x})$. The Fréchet subdifferential $\hat{\partial} f(x)$ and the limiting subdifferential $\partial f(x)$ of f at $\bar{x} \in \mathbb{R}^{n_1}$ are defined as

$$\hat{\partial} f(\bar{x}) := \left\{ g \in \mathbb{R}^{n_1} : \liminf_{x\to\bar{x}, x\neq\bar{x}} \frac{f(x) - f(\bar{x}) - \langle g, x - \bar{x}\rangle}{\|x - \bar{x}\|} \geq 0 \right\},$$

$$\partial f(\bar{x}) := \left\{ g \in \mathbb{R}^{n_1} : \exists x^k \xrightarrow{f} \bar{x}, g^k \to g \text{ with } g^k \in \hat{\partial} f(x^k),\ \forall k \right\},$$

by [17, Definition 1.1] and [26, Definition 8.3, pp. 301], respectively. A point $\bar{x}$ is said to be a Fréchet stationary point of $\min f(x)$ if $0 \in \hat{\partial} f(\bar{x})$, and $\bar{x}$ is said to be a limiting stationary point of $\min f(x)$ if $0 \in \partial f(\bar{x})$. By [11, pp. 30], the usual (one-side) directional derivative of f at x in the direction $d \in \mathbb{R}^{n_1}$ is

$$f'(x; d) := \lim_{\lambda\downarrow 0} \frac{f(x + \lambda d) - f(x)}{\lambda},$$

when the limit exists. According to [25, Definition 2.1], we say that a point $\bar{x} \in \mathbb{R}^{n_1}$ is a d(irectional)-stationary point of $\min f(x)$ if

$$f'(\bar{x}; d) \geq 0, \quad \forall d \in \mathbb{R}^{n_1}.$$

2. Problem reformulation and optimality conditions.

For simplicity, we focus on the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ as the ReLU function, i.e.,

$$\sigma(u) = \max\{u, 0\} = (u)_+. \tag{2.1}$$

Our model, algorithms and theoretical analysis developed in this paper can be generalized to the leaky ReLU and the ELU activation functions. Detailed analysis for the extensions will be given in section 4.3.

2.1. Problem reformulation.

We utilize auxiliary variables $\mathbf{h}$, $\mathbf{u}$ and denote vectors $\mathbf{w}, \mathbf{a}, \mathbf{z}, \mathbf{s}$ as

$$\begin{aligned} &\mathbf{h} = (h_1; h_2; ...; h_T) \in \mathbb{R}^{rT}, \quad \mathbf{u} = (u_1; u_2; ...; u_T) \in \mathbb{R}^{rT}, \\ &\mathbf{w} = (\mathrm{vec}(W); \mathrm{vec}(V); b) \in \mathbb{R}^{N_\mathbf{w}}, \quad \mathbf{a} = (\mathrm{vec}(A); c) \in \mathbb{R}^{N_\mathbf{a}}, \\ &\mathbf{z} = (\mathbf{w}; \mathbf{a}) \in \mathbb{R}^{N_\mathbf{w}+N_\mathbf{a}}, \qquad \mathbf{s} = (\mathbf{z}; \mathbf{h}; \mathbf{u}) \in \mathbb{R}^{N_\mathbf{w}+N_\mathbf{a}+2rT}, \end{aligned}$$

where $N_\mathbf{w} = r^2 + rn + r$ and $N_\mathbf{a} = mr + m$.

We reformulate problem (1.1) as the following constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{s}} \quad & \frac{1}{T}\sum_{t=1}^{T} \|y_t - (Ah_t + c)\|^2 \\ \text{s.t.} \quad & u_t = Wh_{t-1} + Vx_t + b, \\ & h_0 = 0,\ h_t = (u_t)_+,\ t = 1, 2, ..., T. \end{aligned} \tag{2.2}$$

Problems (1.1) and (2.2) are equivalent in the sense that if $(A^*, W^*, V^*, b^*, c^*)$ is a global solution of (1.1), then $\mathbf{s}^* = (\mathbf{z}^*; \mathbf{h}^*; \mathbf{u}^*)$ is a global solution of (2.2) where $\mathbf{z}^*$ is defined by $(A^*, W^*, V^*, b^*, c^*)$ and $\mathbf{h}^*, \mathbf{u}^*$ satisfy the constraints of (2.2) with W^*, V^*, b^*. Conversely, if $\mathbf{s}^*$ is a global solution of (2.2), then $\mathbf{z}^*$ is a global solution of (1.1).

Let us denote the mappings $\Phi : \mathbb{R}^r \mapsto \mathbb{R}^{m \times N_\mathbf{a}}$ and $\Psi : \mathbb{R}^{rT} \mapsto \mathbb{R}^{rT \times N_\mathbf{w}}$ as

$$\Phi(h_t) = \begin{bmatrix} h_t^\top \otimes I_m & I_m \end{bmatrix}, \quad \Psi(\mathbf{h}) = \begin{bmatrix} 0_r^\top \otimes I_r & x_1^\top \otimes I_r & I_r \\ h_1^\top \otimes I_r & x_2^\top \otimes I_r & I_r \\ \vdots & \vdots & \vdots \\ h_{T-1}^\top \otimes I_r & x_T^\top \otimes I_r & I_r \end{bmatrix}, \tag{2.3}$$

where $\otimes$ represents the Kronecker product, I_r and I_m are the identity matrices with dimensions r and m respectively, and 0_r is the zero vector with dimension r. Thus, the objective function and constraints in problem (2.2) can be represented as

$$\begin{aligned} &\ell(\mathbf{s}) := \frac{1}{T}\sum_{t=1}^{T} \|y_t - \Phi(h_t)\mathbf{a}\|^2, \\ &\mathcal{C}_1(\mathbf{s}) := \mathbf{u} - \Psi(\mathbf{h})\mathbf{w} = 0, \qquad \mathcal{C}_2(\mathbf{s}) := \mathbf{h} - (\mathbf{u})_+ = 0. \end{aligned} \tag{2.4}$$

To mitigate the overfitting, we further add a regularization term

$$P(\mathbf{s}) := \lambda_1\|A\|_F^2 + \lambda_2\|W\|_F^2 + \lambda_3\|V\|_F^2 + \lambda_4\|b\|^2 + \lambda_5\|c\|^2 + \lambda_6\|\mathbf{u}\|^2 \tag{2.5}$$

with $\lambda_i > 0, i = 1, 2, \ldots, 6$ in the objective of problem (2.2), and consider the following problem:

$$\begin{aligned} \min \quad & \mathcal{R}(\mathbf{s}) := \ell(\mathbf{s}) + P(\mathbf{s}) \\ \text{s.t.} \quad & \mathbf{s} \in \mathcal{F} := \{\mathbf{s} : \mathcal{C}_1(\mathbf{s}) = 0,\ \mathcal{C}_2(\mathbf{s}) = 0\}. \end{aligned} \tag{2.6}$$

2.2. Optimality conditions. Problem (2.2) and problem (2.6) have the same feasible set $\mathcal{F}$. The constraint function $\mathcal{C}_1$ is continuously differentiable, while the other constraint function $\mathcal{C}_2$ is linear in $\mathbf{h}$ and piecewise linear in $\mathbf{u}$. We denote by $J\mathcal{C}_1(\mathbf{s})$ the Jacobian matrix of the function $\mathcal{C}_1$ at $\mathbf{s}$, and by $J_{\mathbf{z}}\mathcal{C}_1(\mathbf{s})$, $J_{\mathbf{h}}\mathcal{C}_1(\mathbf{s})$, $J_{\mathbf{u}}\mathcal{C}_1(\mathbf{s})$ the Jacobian matrix of function $\mathcal{C}_1$ at $\mathbf{s}$ with respect to the block $\mathbf{z}$, $\mathbf{h}$ and $\mathbf{u}$, respectively. Similarly, we use $J_{\mathbf{h}}\mathcal{C}_2(\mathbf{s})$ to represent the Jacobian matrix of $\mathcal{C}_2$ at $\mathbf{s}$ with respect to $\mathbf{h}$. Moreover, for a fixed vector $\zeta \in \mathbb{R}^{rT}$, we use $\partial\big(\zeta^\top\mathcal{C}_2(\mathbf{s})\big)$ to denote the limiting subdifferential of $\zeta^\top\mathcal{C}_2$ at $\mathbf{s}$ and $\partial_{\mathbf{u}}\big(\zeta^\top\mathcal{C}_2(\mathbf{s})\big)$ to denote the limiting subdifferential of $\zeta^\top\mathcal{C}_2$ at $\mathbf{s}$ with respect to $\mathbf{u}$.

The following lemma shows that the NNAMCQ [29, Definition 4.2, pp. 1451] holds at any feasible point $\mathbf{s} \in \mathcal{F}$. The proofs of all lemmas are given in Appendix A.

LEMMA 2.1. *The NNAMCQ holds at any* $\mathbf{s} \in \mathcal{F}$*, i.e., there exist no nonzero vectors* $\xi = (\xi_1; \xi_2; ...; \xi_T) \in \mathbb{R}^{rT}$ *and* $\zeta = (\zeta_1; \zeta_2; ...; \zeta_T) \in \mathbb{R}^{rT}$ *such that*

$$0 \in J\mathcal{C}_1(\mathbf{s})^\top\xi + \partial\big(\zeta^\top\mathcal{C}_2(\mathbf{s})\big). \tag{2.7}$$

DEFINITION 2.2. *We say that* $\mathbf{s} \in \mathcal{F}$ *is a KKT point of problem* (2.2) *if there exist* $\xi \in \mathbb{R}^{rT}$ *and* $\zeta \in \mathbb{R}^{rT}$ *such that*

$$0 \in \nabla\ell(\mathbf{s}) + J\mathcal{C}_1(\mathbf{s})^\top\xi + \partial\big(\zeta^\top\mathcal{C}_2(\mathbf{s})\big).$$

We say that $\mathbf{s} \in \mathcal{F}$ *is a KKT point of problem* (2.6) *if there exist* $\xi \in \mathbb{R}^{rT}$ *and* $\zeta \in \mathbb{R}^{rT}$ *such that*

$$0 \in \nabla\mathcal{R}(\mathbf{s}) + J\mathcal{C}_1(\mathbf{s})^\top\xi + \partial\big(\zeta^\top\mathcal{C}_2(\mathbf{s})\big).$$

Now we can establish the first order necessary conditions for problem (2.2) and problem (2.6).

THEOREM 2.3. *(i) If* $\bar{\mathbf{s}}$ *is a local solution of problem* (2.2)*, then* $\bar{\mathbf{s}}$ *is a KKT point of problem* (2.2)*. (ii) If* $\bar{\mathbf{s}}$ *is a local solution of problem* (2.6)*, then* $\bar{\mathbf{s}}$ *is a KKT point of problem* (2.6)*.*

Proof. Note that the objective functions of problem (2.2) and problem (2.6) are continuously differentiable. The constraint functions $\mathcal{C}_1$ is continuously differentiable, and $\mathcal{C}_2$ is Lipschitz continuous at any feasible point $\mathbf{s} \in \mathcal{F}$. By Lemma 2.1, NNAMCQ holds at any $\bar{s} \in \mathcal{F}$. Therefore, the conclusions of this theorem hold according to [29, Remark 2 and Theorem 5.2]. ∎

2.3. Nonempty and compact solution set of (2.6). Let $\mathcal{S}_1$ be the solution set of problem (2.6), and denote the level set

$$\mathcal{D}_{\mathcal{R}}(\rho) := \{\mathbf{s} \in \mathcal{F} : \mathcal{R}(\mathbf{s}) \le \rho\} \tag{2.8}$$

with a nonnegative scalar ρ.

LEMMA 2.4. *For any* $\rho > \mathcal{R}(0)$*, the level set* $D_{\mathcal{R}}(\rho)$ *is nonempty and compact. Moreover, the solution set* $\mathcal{S}_1$ *of* (2.6) *is nonempty and compact.*

3. ALM with BCD method for (2.6). To solve the regularized constrained problem (2.6), we develop in this section an ALM. The subproblems of ALM are approximately solved by a BCD method whose update of each block owns a closed-form expression. This is not an easy task due to the nonsmooth nonconvex constraints. The framework of the ALM is given in Algorithm 3.1, in which the updating schemes for Lagrangian multipliers and penalty parameters are motivated by [9]. It is worth mentioning that in [9], the constraints are smooth. In problem (2.6), the constraints are nonsmooth nonconvex. For solving the subproblems in the ALM, we design the BCD method in Algorithm 3.2 and provide the closed-form expression for the update of each block in the BCD. Due to the nonsmooth nonconvex constraints in (2.6), the convergence analysis is complex, which will be given in section 4.

The augmented Lagrangian (AL) function of problem (2.6) is

$$\begin{aligned}(3.1)\qquad &\mathcal{L}(\mathbf{s},\xi,\zeta,\gamma)\\ &:= \mathcal{R}(\mathbf{s}) + \langle \xi, \mathbf{u} - \Psi(\mathbf{h})\mathbf{w}\rangle + \langle \zeta, \mathbf{h} - (\mathbf{u})_+\rangle + \tfrac{\gamma}{2}\left\|\mathbf{u} - \Psi(\mathbf{h})\mathbf{w}\right\|^2 + \tfrac{\gamma}{2}\left\|\mathbf{h} - (\mathbf{u})_+\right\|^2\\ &= \mathcal{R}(\mathbf{s}) + \frac{\gamma}{2}\left\|\mathbf{u} - \Psi(\mathbf{h})\mathbf{w} + \frac{\xi}{\gamma}\right\|^2 + \frac{\gamma}{2}\left\|\mathbf{h} - (\mathbf{u})_+ + \frac{\zeta}{\gamma}\right\|^2 - \frac{\|\xi\|^2}{2\gamma} - \frac{\|\zeta\|^2}{2\gamma},\end{aligned}$$

where $\xi = (\xi_1;\xi_2;...;\xi_T) \in \mathbb{R}^{rT}$ and $\zeta = (\zeta_1;\zeta_2;...;\zeta_T) \in \mathbb{R}^{rT}$ are the Lagrangian multipliers, and $\gamma > 0$ is the penalty parameter for the two quadratic penalty terms of constraints $\mathbf{u} = \Psi(\mathbf{h})\mathbf{w}$ and $\mathbf{h} = (\mathbf{u})_+$. For convenience, we will also write $\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u},\xi,\zeta,\gamma)$ to represent $\mathcal{L}(\mathbf{s},\xi,\zeta,\gamma)$ when the blocks of $\mathbf{s}$ are emphasized.

We develop some basic results in the following two lemmas relating to the AL function $\mathcal{L}$. The explicit formulas for the gradients of $\mathcal{L}$ with respect to $\mathbf{z}$ and $\mathbf{h}$ in Lemma 3.1 (iii) and (iv) will be used for obtaining the closed-form updates for the $\mathbf{z}$ and $\mathbf{h}$ blocks in the BCD method, respectively. The Lipschitz constants $L_1(\xi,\zeta,\gamma,\hat{r})$ and $L_2(\xi,\zeta,\gamma,\hat{r})$ in Lemma 3.2 are essential to design a practical stopping condition (3.17) of the BCD method in Algorithm 3.2. The results will also be used for the convergence results of the BCD method in Theorems 4.3 and 4.4.

LEMMA 3.1. *For any fixed γ,ξ and ζ, the following statements hold.*

(i) The AL function $\mathcal{L}$ is lower bounded that satisfies

$$\mathcal{L}(\mathbf{s},\xi,\zeta,\gamma) \geq -\frac{\|\xi\|^2}{2\gamma} - \frac{\|\zeta\|^2}{2\gamma} \quad \textit{for all } \mathbf{s}.$$

(ii) For any $\hat{\mathbf{s}}$ and $\hat{\Gamma} \geq \hat{r} := \mathcal{L}(\hat{\mathbf{s}},\xi,\zeta,\gamma)$, the level set

$$\Omega_{\mathcal{L}}(\hat{\Gamma}) := \{\mathbf{s}\ : \mathcal{L}(\mathbf{s},\xi,\zeta,\gamma) \leq \hat{\Gamma}\}$$

is nonempty and compact.

(iii) The AL function $\mathcal{L}$ is continuously differentiable with respect to $\mathbf{z}$, and the gradient with respect to $\mathbf{z}$ is

$$\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u},\xi,\zeta,\gamma) = \begin{bmatrix} \hat{Q}_1(\mathbf{s},\xi,\zeta,\gamma)\mathbf{w} + \hat{q}_1(\mathbf{s},\xi,\zeta,\gamma) \\ \hat{Q}_2(\mathbf{s},\xi,\zeta,\gamma)\mathbf{a} + \hat{q}_2(\mathbf{s},\xi,\zeta,\gamma) \end{bmatrix},$$

where

$$\hat{Q}_1(\mathbf{s},\xi,\zeta,\gamma) = \gamma\Psi(\mathbf{h})^\top\Psi(\mathbf{h}) + 2\Lambda_1, \quad \hat{q}_1(\mathbf{s},\xi,\zeta,\gamma) = -\Psi(\mathbf{h})^\top(\xi + \gamma\mathbf{u})$$

$$\hat{Q}_2(\mathbf{s},\xi,\zeta,\gamma) = \frac{2}{T}\sum_{t=1}^{T}\Phi(h_t)^\top\Phi(h_t) + 2\Lambda_2, \quad \hat{q}_2(\mathbf{s},\xi,\zeta,\gamma) = -\frac{2}{T}\sum_{t=1}^{T}\Phi(h_t)^\top y_t$$

$$\Lambda_1 = \mathrm{diag}\Big(\big(\lambda_2\boldsymbol{e}_{r^2};\lambda_3\boldsymbol{e}_{rn};\lambda_4\boldsymbol{e}_r\big)\Big), \quad \Lambda_2 = \mathrm{diag}\Big(\big(\lambda_1\boldsymbol{e}_{rm};\lambda_5\boldsymbol{e}_m\big)\Big).$$

(iv) The AL function $\mathcal{L}$ is continuously differentiable with respect to $\mathbf{h}$*, and the gradient with respect to* $\mathbf{h}$ *is*

$$\begin{aligned}&\nabla_{\mathbf{h}}\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u},\xi,\zeta,\gamma)\\ &= \big(\nabla_{h_1}\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u},\xi,\zeta,\gamma);\nabla_{h_2}\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u},\xi,\zeta,\gamma);\ldots;\nabla_{h_T}\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u},\xi,\zeta,\gamma)\big),\end{aligned}$$

where

$$\nabla_{h_t}\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u},\xi,\zeta,\gamma) = \begin{cases} D_1(\mathbf{s},\xi,\zeta,\gamma)h_t - d_{1t}(\mathbf{s},\xi,\zeta,\gamma), & \text{if } t\in[T-1],\\ D_2(\mathbf{s},\xi,\zeta,\gamma)h_T - d_{2T}(\mathbf{s},\xi,\zeta,\gamma), & \text{if } t=T,\end{cases}$$

$$D_1(\mathbf{s},\xi,\zeta,\gamma) = \gamma W^\top W + \tfrac{2}{T}A^\top A + \gamma I_r,$$

$$D_2(\mathbf{s},\xi,\zeta,\gamma) = \tfrac{2}{T}A^\top A + \gamma I_r,$$

$$d_{1t}(\mathbf{s},\xi,\zeta,\gamma) = W^\top\left(\xi_{t+1} + \gamma(u_{t+1} - Vx_{t+1} - b)\right) + \gamma(u_t)_+ - \zeta_t + \tfrac{2}{T}A^\top(y_t - c),$$

$$d_{2T}(\mathbf{s},\xi,\zeta,\gamma) = \gamma(u_T)_+ - \zeta_T + \tfrac{2}{T}A^\top(y_T - c).$$

LEMMA 3.2. *For any* $\mathbf{z},\mathbf{h},\mathbf{u},\mathbf{h}',\mathbf{u}'$ *in the level set* $\Omega_{\mathcal{L}}(\hat r)$*, we have*

$$\|\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{z},\mathbf{h}',\mathbf{u}',\xi,\zeta,\gamma) - \nabla_{\mathbf{z}}\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u},\xi,\zeta,\gamma)\| \le L_1(\xi,\zeta,\gamma,\hat r)\left\|\begin{matrix}\mathbf{h}'-\mathbf{h}\\ \mathbf{u}'-\mathbf{u}\end{matrix}\right\|, \tag{3.2}$$

$$\|\nabla_{\mathbf{h}}\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u}',\xi,\zeta,\gamma) - \nabla_{\mathbf{h}}\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u},\xi,\zeta,\gamma)\| \le L_2(\xi,\zeta,\gamma,\hat r)\,\|\mathbf{u}'-\mathbf{u}\|, \tag{3.3}$$

where

$$L_1(\xi,\zeta,\gamma,\hat r) = \sqrt{2}\max\{\gamma\delta_1,\delta_2+\delta_3+\delta_4\},\ L_2(\xi,\zeta,\gamma,\hat r) = \gamma\delta_5, \tag{3.4}$$

with $X := (x_1;x_2;...;x_T)\in\mathbb{R}^{nT}$,

$$\delta = \hat r + \frac{\|\xi\|^2}{2\gamma} + \frac{\|\zeta\|^2}{2\gamma},\ \delta_0 = \sqrt{\frac{2\delta}{\gamma}} + \sqrt{\frac{\delta}{\lambda_6}} + \frac{\|\zeta\|}{\gamma},\ \delta_1 = \sqrt{r(\delta^2 + \|X\|^2 + T)},$$

$$\delta_2 = 2\gamma\delta_1\sqrt{\frac{r\delta}{\min\{\lambda_2,\lambda_3,\lambda_4\}}},\ \delta_3 = \sqrt{r}\|\xi\| + \gamma\sqrt{\frac{r\delta}{\lambda_6}},$$

$$\delta_4 = \frac{2\sqrt{m}}{\sqrt{T}}\left(2\sqrt{m(\delta_0^2+1)}\sqrt{\frac{\delta}{\min\{\lambda_1,\lambda_5\}}} + \max_{1\le t\le T}\|y_t\|\right),\ \delta_5 = \sqrt{\frac{\delta(T-1)}{\lambda_2}} + \sqrt{T}.$$

3.1. ALM for the regularized RNNs. To solve the regularized constrained problem (2.6), we propose the ALM in Algorithm 3.1. The ALM first approximately solves (3.5) that aims to minimize the AL function with the fixed Lagrange multipliers ξ^{k-1} and ζ^{k-1}, and the fixed penalty parameter γ_{k-1} for the quadratic terms, until $\mathbf{s}^k$ satisfies the approximate first-order optimality necessary condition (3.6) with tolerance ϵ_{k-1}. Then the Lagrange multipliers are updated, and the tolerance ϵ_k is reduced so that in the next iteration the subproblem is solved more accurately. Moreover, the penalty parameter γ_k is unchanged if the feasibility of $\mathbf{s}^k$ is sufficiently improved compared to that of $\mathbf{s}^{k-1}$, otherwise, γ_k is increased.

Remark 3.3. The main operation of Algorithm 3.1 is to approximately solve the subproblem (3.5). Furthermore, to show that Algorithm 3.1 is well-defined requires that the algorithm for solving the subproblem (3.5) can be terminated within finite steps to meet the stopping condition in (3.6).

In section 3.2, we will design a BCD method to solve the subproblem (3.5). The update of each block of the BCD method owns a closed-form formula, which makes the BCD method efficient. Moreover, the stopping condition (3.6) can be replaced by a simpler condition (3.17) as will be shown in Theorem 4.3.

Algorithm 3.1 The augmented Lagrangian method (ALM) for (2.6)

1: Set an initial penalty parameter $\gamma_0 > 0$, parameters $\eta_1, \eta_2, \eta_4 \in (0,1)$ and $\eta_3 > 1$, an initial tolerance $\epsilon_0 > 0$, vectors of Lagrangian multipliers ξ^0, ζ^0, and a feasible initial point $\mathbf{s}^0 = (\mathbf{z}^0, \hat{\mathbf{h}}, \hat{\mathbf{u}})$ where $\hat{h}_0 = 0$, $\hat{u}_t = W\hat{h}_{t-1} + Vx_t + b$ and $\hat{h}_t = (\hat{u}_t)_+$ for $t \in [T]$.

2: Set $k := 1$.

3: **Step 1:** Solve

$$\min_{\mathbf{s}} \quad \mathcal{L}(\mathbf{s}, \xi^{k-1}, \zeta^{k-1}, \gamma_{k-1}) \tag{3.5}$$

to obtain $\mathbf{s}^k$ satisfying the following condition

$$\operatorname{dist}\big(0, \partial \mathcal{L}(\mathbf{s}^k, \xi^{k-1}, \zeta^{k-1}, \gamma_{k-1})\big) \le \epsilon_{k-1}. \tag{3.6}$$

4: **Step 2:** Update $\epsilon_k = \eta_4 \epsilon_{k-1}$, ξ^{k-1} and ζ^{k-1} as

$$\xi^k = \xi^{k-1} + \gamma_{k-1}\left(\mathbf{u}^k - \Psi(\mathbf{h}^k)\mathbf{w}^k\right), \quad \zeta^k = \zeta^{k-1} + \gamma_{k-1}\left(\mathbf{h}^k - (\mathbf{u}^k)_+\right). \tag{3.7}$$

5: **Step 3:** Set $\gamma_k = \gamma_{k-1}$, if the following condition is satisfied

$$\max\left\{\|\mathcal{C}_1(\mathbf{s}^k)\|, \|\mathcal{C}_2(\mathbf{s}^k)\|\right\} \le \eta_1 \max\left\{\|\mathcal{C}_1(\mathbf{s}^{k-1})\|, \|\mathcal{C}_2(\mathbf{s}^{k-1})\|\right\}. \tag{3.8}$$

6: Otherwise, set

$$\gamma_k = \max\left\{\gamma_{k-1}/\eta_2, \left\|\xi^k\right\|^{1+\eta_3}, \left\|\zeta^k\right\|^{1+\eta_3}\right\}. \tag{3.9}$$

7: Let $k-1 := k$ and go to **Step 1**.

3.2. BCD method for subproblem. To solve the nonsmooth nonconvex problem (3.5) in Step 1 of Algorithm 3.1, we propose a BCD method in Algorithm 3.2 to solve the subproblem at the k-th iteration in the ALM by alternatively updating the blocks in the order of $\mathbf{z}$, $\mathbf{h}$, and $\mathbf{u}$ in $\mathbf{s}$, respectively. Let us choose a constant Γ such that

$$\Gamma \ge \mathcal{L}\big(\mathbf{s}^0, \xi^0, \zeta^0, \gamma_0\big). \tag{3.10}$$

Because at the k-th iteration of the ALM, $\xi^{k-1}, \zeta^{k-1}, \gamma_{k-1}$ are fixed, we just write ξ, ζ, γ in the BCD method for brevity. Furthermore, for the BCD solving the subproblem appeared at the k-th iteration of the ALM, we define

$$\mathbf{s}_{\mathbf{z}}^{k-1,j} := (\mathbf{z}^{k-1,j}; \mathbf{h}^{k-1,j-1}; \mathbf{u}^{k-1,j-1}), \ \mathbf{s}_{\mathbf{h}}^{k-1,j} := (\mathbf{z}^{k-1,j}; \mathbf{h}^{k-1,j}; \mathbf{u}^{k-1,j-1}) \tag{3.11}$$

to denote the point obtained after updating the $\mathbf{z}$ block, and updating the $\mathbf{h}$ block at the j-th iteration of the BCD method, and we use

$$\mathbf{s}^{k-1,j} = (\mathbf{z}^{k-1,j}; \mathbf{h}^{k-1,j}; \mathbf{u}^{k-1,j}) \tag{3.12}$$

to represent the point obtained at the j-th iteration of the BCD method after updating the $\mathbf{u}$ block.

Algorithm 3.2 Block Coordinate Descent (BCD) method for (3.5)

1: Set the initial point of BCD algorithm as

$$\mathbf{s}^{k-1,0} = \begin{cases} \mathbf{s}^{k-1}, & \text{if } k > 1 \text{ and } \mathcal{L}(\mathbf{s}^{k-1}, \xi, \zeta, \gamma) \leq \Gamma, \\ \mathbf{s}^0, & \text{otherwise.} \end{cases} \tag{3.13}$$

Compute $\hat{r}_{k-1} = \mathcal{L}(\mathbf{s}^{k-1,0}, \xi, \zeta, \gamma)$, $L_{1,k-1} = L_1(\xi, \zeta, \gamma, \hat{r}_{k-1})$ and $L_{2,k-1} = L_2(\xi, \zeta, \gamma, \hat{r}_{k-1})$ by formula (3.4).

2: Set $j := 1$.

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

4: **Step 1:** Update blocks $\mathbf{z}^{k-1,j}$, $\mathbf{h}^{k-1,j}$ and $\mathbf{u}^{k-1,j}$ separately as

$$\mathbf{z}^{k-1,j} = \arg\min_{\mathbf{z}} \ \mathcal{L}\left(\mathbf{z}, \mathbf{h}^{k-1,j-1}, \mathbf{u}^{k-1,j-1}, \xi, \zeta, \gamma\right), \tag{3.14}$$

$$\mathbf{h}^{k-1,j} = \arg\min_{\mathbf{h}} \ \mathcal{L}\left(\mathbf{z}^{k-1,j}, \mathbf{h}, \mathbf{u}^{k-1,j-1}, \xi, \zeta, \gamma\right), \tag{3.15}$$

$$\mathbf{u}^{k-1,j} \in \arg\min_{\mathbf{u}} \ \mathcal{L}\left(\mathbf{z}^{k-1,j}, \mathbf{h}^{k-1,j}, \mathbf{u}, \xi, \zeta, \gamma\right) + \frac{\mu}{2}\left\|\mathbf{u} - \mathbf{u}^{k-1,j-1}\right\|^2. \tag{3.16}$$

Then set $\mathbf{s}^{k-1,j} = (\mathbf{z}^{k-1,j}; \mathbf{h}^{k-1,j}; \mathbf{u}^{k-1,j})$.

5: **Step 2:** If the stop criterion

$$\left\|\mathbf{s}^{k-1,j} - \mathbf{s}^{k-1,j-1}\right\| \leq \frac{\epsilon_{k-1}}{\max\{L_{1,k-1}, L_{2,k-1}, \mu\}}, \tag{3.17}$$

is not satisfied, then set $j := j + 1$ and go to **Step 1**.

6: **end while**

7: **return** $\mathbf{s}^k = \mathbf{s}^{k-1,j}$.

Condition (3.6) is satisfied when (3.17) holds, which will be proved in Theorem 4.3. The closed-form solutions of problems (3.14), (3.15) and (3.16) are provided below.

Update $\mathbf{z}^{k-1,j}$: Problem (3.14) is an unconstrained optimization problem with smooth and strongly convex objective function. By employing Lemma 3.1 (iii) and solving

$$\nabla_{\mathbf{z}} \mathcal{L}(\mathbf{s}_{\mathbf{z}}^{k-1,j}, \xi, \zeta, \gamma) = 0,$$

the unique global minimizer $\mathbf{z}^{k-1,j} = (\mathbf{w}^{k-1,j}; \mathbf{a}^{k-1,j})$ can be computed as

$$\mathbf{w}^{k-1,j} = -\hat{Q}_1(\mathbf{s}_{\mathbf{z}}^{k-1,j}, \xi, \zeta, \gamma)^{-1} \hat{q}_1(\mathbf{s}_{\mathbf{z}}^{k-1,j}; \xi, \zeta, \gamma),$$

$$\mathbf{a}^{k-1,j} = -\hat{Q}_2(\mathbf{s}_{\mathbf{z}}^{k-1,j}, \xi, \zeta, \gamma)^{-1} \hat{q}_2(\mathbf{s}_{\mathbf{z}}^{k-1,j}, \xi, \zeta, \gamma).$$

Update $\mathbf{h}^{k-1,j}$: The objective function of (3.15) is also strongly convex and smooth. By employing Lemma 3.1 (iv) and solving $\nabla_{\mathbf{h}} \mathcal{L}(\mathbf{s}_{\mathbf{h}}^{k-1,j}, \xi, \zeta, \gamma) = 0$, we get its unique global minimizer, given by

$$h_t^{k-1,j} = \begin{cases} D_1(\mathbf{s}_{\mathbf{h}}^{k-1,j}, \xi, \zeta, \gamma)^{-1} d_{1t}(\mathbf{s}_{\mathbf{h}}^{k-1,j}, \xi, \zeta, \gamma), & \text{if } t \in [T-1], \\ D_2(\mathbf{s}_{\mathbf{h}}^{k-1,j}, \xi, \zeta, \gamma)^{-1} d_{2T}(\mathbf{s}_{\mathbf{h}}^{k-1,j}, \xi, \zeta, \gamma), & \text{if } t = T. \end{cases} \tag{3.18}$$

Update $\mathbf{u}^{k-1,j}$: Although problem (3.16) is nonsmooth nonconvex, one of its global solutions is accessible, because the objective function of problem (3.16) can be separated into rT one-dimensional functions with the same structure. Thus, we aim to solve the following one-dimensional problem:

$$\min_{u\in\mathbb{R}} \varphi(u) := \tfrac{\gamma}{2}(u-\theta_1)^2 + \tfrac{\gamma}{2}(\theta_2-(u)_+)^2 + \tfrac{\mu}{2}(u-\theta_3)^2 + \lambda_6 u^2, \tag{3.19}$$

where $\theta_1, \theta_2, \theta_3 \in \mathbb{R}$ are known real numbers. Denote

$$u^+ := \arg\min_{u\in\mathbb{R}_+} \varphi(u) \quad \text{and} \quad u^- := \arg\min_{u\in\mathbb{R}_-} \varphi(u). \tag{3.20}$$

By direct computation,

$$u^+ = \begin{cases} \dfrac{\gamma\theta_1+\gamma\theta_2+\mu\theta_3}{2\gamma+2\lambda_6+\mu}, & \text{if } \gamma\theta_1+\gamma\theta_2+\mu\theta_3 > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{3.21}$$

and

$$u^- = \begin{cases} \dfrac{\gamma\theta_1+\mu\theta_3}{\gamma+2\lambda_6+\mu}, & \text{if } \gamma\theta_1+\mu\theta_3 < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.22}$$

Then a solution of (3.19) can be given as

$$u^* = \begin{cases} u^+, & \text{if } \varphi(u^+) \le \varphi(u^-), \\ u^-, & \text{otherwise.} \end{cases}$$

By setting

$$\theta_1 = (\Psi(\mathbf{h}^{k-1,j})\mathbf{w}^{k-1,j})_i - \frac{\xi_i}{\gamma}, \quad \theta_2 = \mathbf{h}_i^{k-1,j} + \frac{\zeta_i}{\gamma}, \quad \theta_3 = \mathbf{u}_i^{k-1,j-1},$$

$$\mathbf{u}_i^{k-1,j} = u^*, \quad \mathbf{u}_i{}^+ = u^+, \quad \mathbf{u}_i{}^- = u^-,$$

we obtain a closed-form solution of problem (3.16) as

$$\mathbf{u}_i^{k-1,j} = \begin{cases} \mathbf{u}_i^+, & \text{if } \varphi(\mathbf{u}_i^+) \le \varphi(\mathbf{u}_i^-), \\ \mathbf{u}_i^-, & \text{otherwise,} \end{cases} \quad i = 1, \dots, rT.$$

Remark 3.4. It is important to mention that the solution set of problem (3.16) may not be a singleton. To ensure the selected solution is unique, we set $\mathbf{u}_i^{k-1,j} = \mathbf{u}_i^+$ when $\varphi(\mathbf{u}_i^+) = \varphi(\mathbf{u}_i^-)$ for every $i \in [rT]$.

4. Convergence analysis.

In this section, we show the convergence results of both the BCD method for the subproblem of the ALM, as well as the ALM for (2.6).

4.1. Convergence analysis of Algorithm 3.2.

It is clear that

$$\mathcal{L}(\mathbf{s}, \xi, \zeta, \gamma) = g(\mathbf{s}, \xi, \gamma) + q(\mathbf{s}, \zeta, \gamma), \tag{4.1}$$

where

$$g(\mathbf{s}, \xi, \gamma) = \mathcal{R}(\mathbf{s}) + \frac{\gamma}{2}\left\|\mathbf{u} - \Psi(\mathbf{h})\mathbf{w} + \frac{\xi}{\gamma}\right\|^2 - \frac{\|\xi\|^2}{2\gamma}, \tag{4.2}$$

$$q(\mathbf{s}, \zeta, \gamma) = \frac{\gamma}{2}\left\|\mathbf{h} - (\mathbf{u})_+ + \frac{\zeta}{\gamma}\right\|^2 - \frac{\|\zeta\|^2}{2\gamma}. \tag{4.3}$$

The function g is smooth but nonconvex, because it contains the bilinear structure $\Psi(\mathbf{h})\mathbf{w}$. The function q is nonsmooth nonconvex.

For the convergence analysis below, we further use $\mathbf{s}_{\mathbf{z}}^{(j)}$ and $\mathbf{s}_{\mathbf{h}}^{(j)}$ to represent $\mathbf{s}_{\mathbf{z}}^{k-1,j}$ and $\mathbf{s}_{\mathbf{h}}^{k-1,j}$ in (3.11), and use $\mathbf{s}^{(j)}$ to represent $s^{k-1,j}$ in (3.12) for brevity. We emphasize that the point $\mathbf{s}^k$ is generated by the ALM in Algorithm 3.1, while the point $\mathbf{s}^{(j)}$ is generated by the BCD method in Algorithm 3.2 for solving the subproblem in the ALM at the k-th iteration.

The following two lemmas will be used in proving the convergence results of the BCD method.

LEMMA 4.1. *Let* $\{\mathbf{s}^{(j)}\}$ *represent the sequence generated by Algorithm* 3.2*. Then* $\{\mathbf{s}^{(j)}\}$ *belongs to the level set* $\Omega_{\mathcal{L}}(\Gamma)$*, which is compact.*

LEMMA 4.2. *The AL function* $\mathcal{L}$ *is locally Lipschitz continuous and directionally differentiable on* $\Omega_{\mathcal{L}}(\Gamma)$.

We can now show that the stop criterion (3.17) in Algorithm 3.2 can be stopped in finite steps, and condition (3.6) in Algorithm 3.1 is satisfied when (3.17) holds. These results guarantee that the ALM in Algorithm 3.1 is well-defined, when the subproblems are solved by the BCD method in Algorithm 3.2.

THEOREM 4.3. *At the k-th iteration of ALM in Algorithm* 3.1*, the BCD method in Algorithm* 3.2 *for the subproblem* (3.5) *can be stopped within finite steps to satisfy the stop criterion in* (3.17)*, which is of order* $O(1/(\epsilon_{k-1})^2)$*. Moreover, condition* (3.6) *of the ALM in Algorithm* 3.1 *is satisfied at the output* $\mathbf{s}^k$ *of Algorithm* 3.2*.*

Proof. Since $\mathcal{L}$ is strongly convex with respect to the blocks $\mathbf{z}$ and $\mathbf{h}$, respectively, from (3.14) and (3.15), we obtain

$$\mathcal{L}(\mathbf{s}^{(j-1)},\xi,\zeta,\gamma)-\mathcal{L}(\mathbf{s}_{\mathbf{z}}^{(j)},\xi,\zeta,\gamma)\geq\tfrac{\alpha_1}{2}\|\mathbf{z}^{(j-1)}-\mathbf{z}^{(j)}\|^2, \tag{4.4}$$

$$\mathcal{L}(\mathbf{s}_{\mathbf{z}}^{(j)},\xi,\zeta,\gamma)-\mathcal{L}(\mathbf{s}_{\mathbf{h}}^{(j)},\xi,\zeta,\gamma)\geq\tfrac{\alpha_2}{2}\|\mathbf{h}^{(j-1)}-\mathbf{h}^{(j)}\|^2, \tag{4.5}$$

where α_1 and α_2 are the minimum eigenvalues of the Hessian matrices $\nabla^2_{\mathbf{z}}\mathcal{L}(\mathbf{s},\xi,\zeta,\gamma)$ and $\nabla^2_{\mathbf{h}}\mathcal{L}(\mathbf{s},\xi,\zeta,\gamma)$ for all $\mathbf{s}$ in the compact set $\Omega_{\mathcal{L}}(\Gamma)$, respectively. Furthermore, by (3.16), we have

$$\mathcal{L}(\mathbf{s}_{\mathbf{h}}^{(j)},\xi,\zeta,\gamma)-\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma)\geq\tfrac{\mu}{2}\left\|\mathbf{u}^{(j)}-\mathbf{u}^{(j-1)}\right\|^2. \tag{4.6}$$

It follows that

$$\begin{aligned}
&\mathcal{L}(\mathbf{s}^{(j-1)},\xi,\zeta,\gamma)-\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma)\\
&=\big(\mathcal{L}(\mathbf{s}^{(j-1)},\xi,\zeta,\gamma)-\mathcal{L}(\mathbf{s}_{\mathbf{z}}^{(j)},\xi,\zeta,\gamma)\big)+\big(\mathcal{L}(\mathbf{s}_{\mathbf{z}}^{(j)},\xi,\zeta,\gamma)-\mathcal{L}(\mathbf{s}_{\mathbf{h}}^{(j)},\xi,\zeta,\gamma)\big)\\
&\quad+\big(\mathcal{L}(\mathbf{s}_{\mathbf{h}}^{(j)},\xi,\zeta,\gamma)-\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma)\big)\\
&\geq\tfrac{\alpha_1}{2}\|\mathbf{z}^{(j)}-\mathbf{z}^{(j-1)}\|^2+\tfrac{\alpha_2}{2}\|\mathbf{h}^{(j)}-\mathbf{h}^{(j-1)}\|^2+\tfrac{\mu}{2}\|\mathbf{u}^{(j)}-\mathbf{u}^{(j-1)}\|^2\\
&\geq\ \max\{\tfrac{\alpha_1}{2},\tfrac{\alpha_2}{2},\tfrac{\mu}{2}\}\|\mathbf{s}^{(j)}-\mathbf{s}^{(j-1)}\|^2.
\end{aligned}$$

Summing up $\mathcal{L}(\mathbf{s}^{(j-1)},\xi,\zeta,\gamma)-\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma)$ from $j=1$ to J, we have

$$\begin{aligned}
\mathcal{L}(\mathbf{s}^{(0)},\xi,\zeta,\gamma)-\mathcal{L}(\mathbf{s}^{(J)},\xi,\zeta,\gamma)&\geq\max\{\tfrac{\alpha_1}{2},\tfrac{\alpha_2}{2},\tfrac{\mu}{2}\}\sum_{j=1}^{J}\|\mathbf{s}^{(j)}-\mathbf{s}^{(j-1)}\|^2\\
&\geq J\max\{\tfrac{\alpha_1}{2},\tfrac{\alpha_2}{2},\tfrac{\mu}{2}\}\min_{j\in[J]}\{\|\mathbf{s}^{(j)}-\mathbf{s}^{(j-1)}\|^2\}.
\end{aligned} \tag{4.7}$$

This, together with Lemma 3.1 (i), yields that

$$\min_{j\in[J]}\{\|\mathbf{s}^{(j)}-\mathbf{s}^{(j-1)}\|^2\} \le \frac{\mathcal{L}(\mathbf{s}^{(0)},\xi,\zeta,\gamma)+\frac{\|\xi\|^2}{2\gamma}+\frac{\|\zeta\|^2}{2\gamma}}{J\max\{\frac{\alpha_1}{2},\frac{\alpha_2}{2},\frac{\mu}{2}\}}.$$

It follows that the stop criterion (3.17) holds, as long as

$$J \ge \hat{J} := \left\lceil \frac{\left(\mathcal{L}(\mathbf{s}^{(0)},\xi,\zeta,\gamma)+\frac{\|\xi\|^2}{2\gamma}+\frac{\|\zeta\|^2}{2\gamma}\right)(\max\{L_{1,k-1},L_{2,k-1},\mu\})^2}{\max\{\frac{\alpha_1}{2},\frac{\alpha_2}{2},\frac{\mu}{2}\}(\epsilon_{k-1})^2} \right\rceil. \tag{4.8}$$

Therefore, at the k-th iteration of the ALM in Algorithm 3.1, the BCD method in Algorithm 3.2 can be stopped in at most $\hat{J}$ iterations defined in (4.8) and output $\mathbf{s}^k$, which is of order $O(1/(\epsilon_{k-1})^2)$.

Once condition (3.17) is satisfied, condition (3.6) in Algorithm 3.1 also holds, which will be proved in the following. By Step 1 in Algorithm 3.2, the first order optimality condition of the three blocked subproblems (3.14), (3.15) and (3.16) are

$$0=\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{s}_{\mathbf{z}}^{(j)},\xi,\zeta,\gamma),\ 0=\nabla_{\mathbf{h}}\mathcal{L}(\mathbf{s}_{\mathbf{h}}^{(j)},\xi,\zeta,\gamma),$$
$$0\in\nabla_{\mathbf{u}}g(\mathbf{s}^{(j)},\xi,\gamma)+\partial_{\mathbf{u}}q(\mathbf{s}^{(j)},\zeta,\gamma)+\mu(\mathbf{u}^{(j)}-\mathbf{u}^{(j-1)}).$$

Furthermore, the limiting subdifferential of the function $\mathcal{L}$ at $\mathbf{s}^{(j)}$ can be written as

$$\partial\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma)=\left(\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma);\nabla_{\mathbf{h}}\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma);\nabla_{\mathbf{u}}g(\mathbf{s}^{(j)},\xi)+\partial_{\mathbf{u}}q(\mathbf{s}^{(j)},\zeta)\right).$$

Hence

$$\begin{bmatrix}\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma)-\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{s}_{\mathbf{z}}^{(j)},\xi,\zeta,\gamma)\\ \nabla_{\mathbf{h}}\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma)-\nabla_{\mathbf{h}}\mathcal{L}(\mathbf{s}_{\mathbf{h}}^{(j)},\xi,\zeta,\gamma)\\ -\mu(\mathbf{u}^{(j)}-\mathbf{u}^{(j-1)})\end{bmatrix}\in\partial\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma).$$

By Lemma 3.2, we obtain

$$\begin{aligned}\operatorname{dist}\big(0,\partial\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma)\big) &\le \left\|\begin{matrix}\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma)-\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{s}_{\mathbf{z}}^{(j)},\xi,\zeta,\gamma)\\ \nabla_{\mathbf{h}}\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma)-\nabla_{\mathbf{h}}\mathcal{L}(\mathbf{s}_{\mathbf{h}}^{(j)},\xi,\zeta,\gamma)\\ -\mu(\mathbf{u}^{(j)}-\mathbf{u}^{(j-1)})\end{matrix}\right\|\\ &\le \max\{L_{1,k-1},L_{2,k-1},\mu\}\|\mathbf{s}^{(j)}-\mathbf{s}^{(j-1)}\|.\end{aligned}$$

Thus condition (3.17) that $\|\mathbf{s}^{(j)}-\mathbf{s}^{(j-1)}\|\le\epsilon_{k-1}/\max\{L_{1,k-1},L_{2,k-1},\mu\}$, together with $\mathbf{s}^k=\mathbf{s}^{(j)}$, implies $\operatorname{dist}(0,\partial\mathcal{L}(\mathbf{s}^{(k)},\xi,\zeta,\gamma))\le\epsilon_{k-1}$ in condition (3.6). ∎

Theorem 4.3 above guarantees that the BCD method in Algorithm 3.2 terminates within finite steps to meet the stop criterion (3.17) for a fixed $\epsilon_{k-1}>0$. In the rest of this subsection, we discuss the convergence of Algorithm 3.2 for the case $\epsilon_{k-1}=0$, i.e., we replace the stop criterion (3.17) by

$$\left\|\mathbf{s}^{k-1,j}-\mathbf{s}^{k-1,j-1}\right\|=0. \tag{4.9}$$

We will show in Theorem 4.6 that the BCD method converges to a d-stationary point if $\epsilon_{k-1}=0$. For this purpose, we first show the following theorem that provides the convergence of the sequences of the function values $\mathcal{L}$ with respect to the three blocks, as well as the convergence of the subsequences of the iterative points with respect to the three blocks.

THEOREM 4.4. *Suppose that (3.17) is replaced by (4.9) in Algorithm 3.2. If there is $\bar{j}$ such that (4.9) holds, then*

$$\mathcal{L}(\mathbf{s}_{\mathbf{z}}^{(\bar{j})},\xi,\zeta,\gamma)=\mathcal{L}(\mathbf{s}_{\mathbf{h}}^{(\bar{j})},\xi,\zeta,\gamma)=\mathcal{L}(\mathbf{s}^{(\bar{j})},\xi,\zeta,\gamma) \quad \text{and} \quad \mathbf{s}_{\mathbf{z}}^{(\bar{j})}=\mathbf{s}_{\mathbf{h}}^{(\bar{j})}=\mathbf{s}^{(\bar{j})}. \tag{4.10}$$

Otherwise, Algorithm 3.2 generates infinite sequences $\{\mathbf{s}_{\mathbf{z}}^{(j)}\}$, $\{\mathbf{s}_{\mathbf{h}}^{(j)}\}$ and $\{\mathbf{s}^{(j)}\}$, and the following statements hold.

(i) The sequences $\{\mathcal{L}(\mathbf{s}_{\mathbf{z}}^{(j)},\xi,\zeta,\gamma)\}$, $\{\mathcal{L}(\mathbf{s}_{\mathbf{h}}^{(j)},\xi,\zeta,\gamma)\}$ and $\{\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma)\}$ all converge to a constant $\mathcal{L}^*$.

(ii) There exists a subsequence $\{j_i\}\subseteq\{j\}$ such that $\{\mathbf{s}_{\mathbf{z}}^{(j_i)}\}$, $\{\mathbf{s}_{\mathbf{h}}^{(j_i)}\}$ and $\{\mathbf{s}^{(j_i)}\}$ converging to the same point.

Proof. If there is $\bar{j}$ such that (4.9) holds, then (4.10) is derived directly from $\mathbf{s}^{k-1,\bar{j}}=\mathbf{s}^{k-1,\bar{j}-1}$ and (3.14)-(3.16).

If there is no $\bar{j}$ such that (4.9) holds, then Algorithm 3.2 generates infinite sequences $\{\mathbf{s}_{\mathbf{z}}^{(j)}\}$, $\{\mathbf{s}_{\mathbf{h}}^{(j)}\}$ and $\{\mathbf{s}^{(j)}\}$.

(i) By Lemma 4.1, there exists an infinite subsequence $\{j_i\}\subseteq\{j\}$ such that $\mathbf{s}^{(j_i)}\to\bar{\mathbf{s}}$ as $j_i\to\infty$. Let $\mathcal{L}^*=\mathcal{L}(\bar{\mathbf{s}})$. We can easily deduce that statement (i) holds, by the descent inequality (A.14) and the lower boundedness of $\{\mathcal{L}(\mathbf{s}^{(j)},\xi,\zeta,\gamma)\}$ according to Lemma 3.1 (i).

(ii) To further prove that $\{\mathbf{s}_{\mathbf{z}}^{(j_i)}\}$ and $\{\mathbf{s}_{\mathbf{h}}^{(j_i)}\}$ also converge to $\bar{\mathbf{s}}$, it is sufficient to prove

$$\lim_{i\to\infty}\|\mathbf{s}^{(j_i)}-\mathbf{s}_{\mathbf{z}}^{(j_i)}\|=0,\quad \lim_{i\to\infty}\|\mathbf{s}^{(j_i)}-\mathbf{s}_{\mathbf{h}}^{(j_i)}\|=0. \tag{4.11}$$

Letting J go to infinity and replacing (j) in (4.7) by (j_i), it is easy to have that $\sum_{i=1}^{\infty}\|\mathbf{s}^{(j_i)}-\mathbf{s}^{(j_i-1)}\|^2<\infty$. Hence,

$$\lim_{i\to\infty}\|\mathbf{s}^{(j_i)}-\mathbf{s}^{(j_i-1)}\|=0, \tag{4.12}$$

which together with

$$\|\mathbf{s}^{(j_i)}-\mathbf{s}_{\mathbf{z}}^{(j_i)}\|\le\|\mathbf{h}^{(j_i)}-\mathbf{h}^{(j_i-1)}\|+\|\mathbf{u}^{(j_i)}-\mathbf{u}^{(j_i-1)}\|,$$
$$\|\mathbf{s}^{(j_i)}-\mathbf{s}_{\mathbf{h}}^{(j_i)}\|\le\|\mathbf{u}^{(j_i)}-\mathbf{u}^{(j_i-1)}\|,$$

implies the validity of (4.11). ∎

Now we turn to show that Algorithm 3.2 generates a d-stationary point of problem (3.5). For convenience, when considering the directional derivative of a function with respect to a direction and we want to emphasize the blocks of the direction, we adopt a simple expression. For example, if $d=(d_{\mathbf{z}};d_h;d_{\mathbf{u}})$, we also write $\mathcal{L}'(\mathbf{s},\xi,\zeta,\gamma;d)=\mathcal{L}'(\mathbf{s},\xi,\zeta,\gamma;(d_{\mathbf{z}},d_{\mathbf{h}},d_{\mathbf{u}}))$ instead of $\mathcal{L}'(\mathbf{s},\xi,\zeta,\gamma;(d_{\mathbf{z}};d_{\mathbf{h}};d_{\mathbf{u}}))$.

LEMMA 4.5. *If the directional derivatives of $\mathcal{L}$ at $\bar{\mathbf{s}}\in\Omega_{\mathcal{L}}(\Gamma)$ satisfy*

$$\mathcal{L}'\big(\bar{\mathbf{s}},\xi,\zeta,\gamma;(d_{\mathbf{z}},0,0)\big)\ge0,\ \mathcal{L}'\big(\bar{\mathbf{s}},\xi,\zeta,\gamma;(0,d_{\mathbf{h}},0)\big)\ge0,\ \mathcal{L}'\big(\bar{\mathbf{s}},\xi,\zeta,\gamma;(0,0,d_{\mathbf{u}})\big)\ge0,$$

along any $d_{\mathbf{z}}\in\mathbb{R}^{N_{\mathbf{w}}+N_{\mathbf{a}}}$, $d_{\mathbf{h}}\in\mathbb{R}^{rT}$ and $d_{\mathbf{u}}\in\mathbb{R}^{rT}$, then

$$\mathcal{L}'(\bar{\mathbf{s}},\xi,\zeta,\gamma;d)\ge0,\quad \forall\ d\in\mathbb{R}^{N_{\mathbf{w}}+N_{\mathbf{a}}+2rT}.$$

As problem (3.5) is nonsmooth nonconvex, there are many kinds of stationary points for it, such as a Fréchet stationary point, a limiting stationary point, and a d-stationary point. It is known that a Fréchet stationary point is a limiting stationary point, and a d-stationary point is a limiting stationary point, but not vise versa [19]. The theorem below guarantees that either the BCD method terminates at a d-stationary point of problem (3.5) in finite steps, or any accumulation point of the sequence generated by the BCD method is a d-stationary point of problem (3.5).

THEOREM 4.6. *Suppose that (3.17) is replaced by (4.9) in Algorithm* 3.2*. If there is $\bar{j}$ such that (4.9) holds, then $\mathbf{s}^{(\bar{j})}$ is a d-stationary point of problem* (3.5)*. Otherwise, Algorithm* 3.2 *generates an infinite sequence $\{\mathbf{s}^{(j)}\}$ and any accumulation point of $\{\mathbf{s}^{(j)}\}$ is a d-stationary point of problem* (3.5)*.*

Proof. If there is $\bar{j}$ such that (4.9) holds, then $\mathbf{s}^{k-1,\bar{j}} = \mathbf{s}^{k-1,\bar{j}-1}$, i.e., $\mathbf{s}^{(\bar{j})} = \mathbf{s}^{(\bar{j}-1)}$. This, combined with (4.10) of Theorem 4.4, yields that $\mathbf{s}_{\mathbf{z}}^{(\bar{j})} = \mathbf{s}_{\mathbf{h}}^{(\bar{j})} = \mathbf{s}^{(\bar{j})} = \mathbf{s}^{(\bar{j}-1)}$. Thus by (3.14)-(3.16) in Algorithm 3.2, we have for any $\lambda > 0$ and any $d_{\mathbf{z}} \in \mathbb{R}^{N_{\mathbf{w}}+N_{\mathbf{a}}}$, $d_{\mathbf{h}} \in \mathbb{R}^{rT}$, $d_{\mathbf{u}} \in \mathbb{R}^{rT}$,

$$\mathcal{L}(\mathbf{s}^{(\bar{j})}, \xi, \zeta, \gamma) \leq \mathcal{L}\big(\mathbf{s}^{(\bar{j})} + \lambda(d_{\mathbf{z}}, 0, 0), \xi, \zeta, \gamma\big),$$
$$\mathcal{L}(\mathbf{s}^{(\bar{j})}, \xi, \zeta, \gamma) \leq \mathcal{L}\big(\mathbf{s}^{(\bar{j})} + \lambda(0, d_{\mathbf{h}}, 0), \xi, \zeta, \gamma\big),$$
$$\mathcal{L}(\mathbf{s}^{(\bar{j})}, \xi, \zeta, \gamma) \leq \mathcal{L}\big(\mathbf{s}^{(\bar{j})} + \lambda(0, 0, d_{\mathbf{u}}), \xi, \zeta, \gamma\big).$$

By Lemma 4.2 and the definition of the directional derivative, we get for any $d_{\mathbf{z}}$, $d_{\mathbf{h}}$, $d_{\mathbf{u}}$,

$$\mathcal{L}'(\mathbf{s}^{(\bar{j})}, \xi, \zeta, \gamma; (d_{\mathbf{z}}, 0, 0)) \geq 0, \;\; \mathcal{L}'(\mathbf{s}^{(\bar{j})}, \xi, \zeta, \gamma; (0, d_{\mathbf{h}}, 0)) \geq 0,$$
$$\mathcal{L}'(\mathbf{s}^{(\bar{j})}, \xi, \zeta, \gamma; (0, 0, d_{\mathbf{u}})) \geq 0.$$

The above inequalities, along with Lemma 4.5, yields that $\mathcal{L}'(\mathbf{s}^{(\bar{j})}, \xi, \zeta, \gamma; d) \geq 0$ for any $d \in \mathbb{R}^{N_{\mathbf{w}}+N_{\mathbf{a}}+2rT}$. Hence, $\mathbf{s}^{(\bar{j})}$ is a d-stationary point of problem (3.5).

If there is no $\bar{j}$ such that (4.9) holds, then Algorithm 3.2 generates an infinite sequence $\{\mathbf{s}^{(j)}\}$. By (3.16), we have

$$\mathcal{L}(\mathbf{s}^{(j)}, \xi, \zeta, \gamma) \leq \mathcal{L}(\mathbf{s}^{(j)}, \xi, \zeta, \gamma) + \frac{\mu}{2}\|\mathbf{u}^{(j)} - \mathbf{u}^{(j-1)}\|^2 \leq \mathcal{L}(\mathbf{s}_{\mathbf{h}}^{(j)}, \xi, \zeta, \gamma).$$

Letting $j \to \infty$ in the above inequalities and using Theorem 4.4 (i), we have

$$\lim_{j\to\infty} \|\mathbf{u}^{(j)} - \mathbf{u}^{(j-1)}\| = 0.$$

By Theorem 4.4 (ii), let $\{\mathbf{s}_z^{(j_i)}\}$, $\{\mathbf{s}_h^{(j_i)}\}$ and $\{\mathbf{s}^{(j_i)}\}$ be any convergent subsequences with limit $\bar{\mathbf{s}}$. Furthermore, by (3.14)-(3.16) in Algorithm 3.2, we have for any $\lambda > 0$ and any $d_{\mathbf{z}} \in \mathbb{R}^{N_{\mathbf{w}}+N_{\mathbf{a}}}$, $d_{\mathbf{h}} \in \mathbb{R}^{rT}$, $d_{\mathbf{u}} \in \mathbb{R}^{rT}$,

$$\mathcal{L}(\mathbf{s}_{\mathbf{z}}^{(j_i)}, \xi, \zeta, \gamma) \leq \mathcal{L}\big(\mathbf{s}_{\mathbf{z}}^{(j_i)} + \lambda(d_{\mathbf{z}}, 0, 0), \xi, \zeta, \gamma\big),$$
$$\mathcal{L}\big(\mathbf{s}_{\mathbf{h}}^{(j_i)}, \xi, \zeta, \gamma\big) \leq \mathcal{L}\big(\mathbf{s}_{\mathbf{h}}^{(j_i)} + \lambda(0, d_{\mathbf{h}}, 0), \xi, \zeta, \gamma\big),$$
$$\mathcal{L}(\mathbf{s}^{(j_i)}, \xi, \zeta, \gamma) \leq \mathcal{L}\big(\mathbf{s}^{(j_i)} + \lambda(0, 0, d_{\mathbf{u}}), \xi, \zeta, \gamma\big) + \tfrac{\mu}{2}\|\mathbf{u}^{(j_i)} + \lambda d_{\mathbf{u}} - \mathbf{u}^{(j_i-1)}\|^2.$$

As $i \to \infty$, the above equality and inequalities imply that for any $\lambda > 0$ and any $d_{\mathbf{z}}$, $d_{\mathbf{h}}$, $d_{\mathbf{u}}$,

$$\mathcal{L}(\bar{\mathbf{s}}, \xi, \zeta, \gamma) \leq \mathcal{L}\big(\bar{\mathbf{s}} + \lambda(d_{\mathbf{z}}, 0, 0), \xi, \zeta, \gamma\big), \; \mathcal{L}(\bar{\mathbf{s}}, \xi, \zeta, \gamma) \leq \mathcal{L}\big(\bar{\mathbf{s}} + \lambda(0, d_{\mathbf{h}}, 0), \xi, \zeta, \gamma\big),$$
$$\mathcal{L}(\bar{\mathbf{s}}, \xi, \zeta, \gamma) \leq \mathcal{L}\big(\bar{\mathbf{s}} + \lambda(0, 0, d_{\mathbf{u}}), \xi, \zeta, \gamma\big) + \tfrac{\mu}{2}\lambda^2\|d_{\mathbf{u}}\|^2.$$

By Lemma 4.2 and the definition of directional derivative, it follows that

$$\mathcal{L}'(\bar{\mathbf{s}},\xi,\zeta,\gamma;(d_{\mathbf{z}},0,0))\geq 0,\ \ \mathcal{L}'(\bar{\mathbf{s}},\xi,\zeta,\gamma;(0,d_{\mathbf{h}},0))\geq 0,\ \ \mathcal{L}'(\bar{\mathbf{s}},\xi,\zeta,\gamma;(0,0,d_{\mathbf{u}}))\geq 0,$$

for any $d_{\mathbf{z}}$, $d_{\mathbf{h}}$ and $d_{\mathbf{u}}$. The above inequalities, along with Lemma 4.5, yield that $\bar{\mathbf{s}}$ is a d-stationary point of problem (3.5). ∎

4.2. Convergence analysis of Algorithm 3.1. By Theorem 4.3, the ALM in Algorithm 3.1 is well-defined, since Step 1 can always be fulfilled in finite steps by the BCD method in Algorithm 3.2.

It is well known that the classical ALM may converge to an infeasible point. In contrast, the following theorem guarantees that any accumulation point of the ALM in Algorithm 3.1 is a feasible point. The delicate strategy for updating the penalty parameter γ_k in Step 3 of Algorithm 3.1 plays an important role in the proof of the theorem.

THEOREM 4.7. *Let* $\{\mathbf{s}^k\}$ *be the sequence generated by Algorithm* 3.1*. Then the following statements hold.*

(i) $\lim_{k\to\infty}\left\|\mathbf{u}^k-\Psi(\mathbf{h}^k)\mathbf{w}^k\right\|=0$ *and* $\lim_{k\to\infty}\left\|\mathbf{h}^k-(\mathbf{u}^k)_+\right\|=0$*.*

(ii) There exists at least one accumulation point of $\{\mathbf{s}^k\}$*, and any accumulation point is a feasible point of* (2.6)*.*

Proof. (i) Let the index set

$$\mathcal{K}:=\left\{k:\gamma_k=\max\{\gamma_{k-1}/\eta_2,\|\xi^k\|^{1+\eta_3},\|\zeta^k\|^{1+\eta_3}\}\right\}. \tag{4.13}$$

If $\mathcal{K}$ is a finite set, then there exists $K\in\mathbb{N}_+$, such that for all $k>K$,

$$\begin{aligned}\max\left\{\|\mathcal{C}_1(\mathbf{s}^k)\|,\|\mathcal{C}_2(\mathbf{s}^k)\|\right\}&\leq\eta_1\max\left\{\|\mathcal{C}_1(\mathbf{s}^{k-1})\|,\|\mathcal{C}_2(\mathbf{s}^{k-1})\|\right\}\\&\leq\eta_1^{k-K}\max\left\{\|\mathcal{C}_1(\mathbf{s}^{K})\|,\|\mathcal{C}_2(\mathbf{s}^{K})\|\right\}.\end{aligned} \tag{4.14}$$

Since $\eta_1\in(0,1)$, we get $\lim_{k\to\infty}\max\left\{\|\mathbf{u}^k-\Psi(\mathbf{h}^k)\mathbf{w}^k\|,\|\mathbf{h}^k-(\mathbf{u}^k)_+\|\right\}=0$. The statement (i) can thus be proved for this case.

Otherwise, $\mathcal{K}$ is an infinite set. Then for those $k-1\in\mathcal{K}$,

$$\max\left\{\frac{\|\xi^{k-1}\|}{\gamma_{k-1}},\frac{\|\zeta^{k-1}\|}{\gamma_{k-1}}\right\}\leq(\gamma_{k-1})^{\frac{-\eta_3}{1+\eta_3}},\ \max\left\{\frac{\|\xi^{k-1}\|^2}{\gamma_{k-1}},\frac{\|\zeta^{k-1}\|^2}{\gamma_{k-1}}\right\}\leq(\gamma_{k-1})^{\frac{1-\eta_3}{1+\eta_3}}.$$

The above inequalities, together with $\eta_3>1$ yields that

$$\lim_{k\to\infty,k-1\in\mathcal{K}}\max\left\{\frac{\|\xi^{k-1}\|}{\gamma_{k-1}},\frac{\|\zeta^{k-1}\|}{\gamma_{k-1}},\frac{\|\xi^{k-1}\|^2}{\gamma_{k-1}},\frac{\|\zeta^{k-1}\|^2}{\gamma_{k-1}}\right\}=0. \tag{4.15}$$

Recalling (3.1), and employing condition (A.15) and Step 1 of Algorithm 3.2, we have

$$\begin{aligned}0\leq&\left\|\mathbf{u}^k-\Psi(\mathbf{h}^k)\mathbf{w}^k+\tfrac{\xi^{k-1}}{\gamma_{k-1}}\right\|^2+\left\|\mathbf{h}^k-(\mathbf{u}^k)_++\tfrac{\zeta^{k-1}}{\gamma_{k-1}}\right\|^2\\\leq&\tfrac{2}{\gamma_{k-1}}\left(\Gamma-\mathcal{R}(\mathbf{s}^k)\right)+\left(\tfrac{\|\xi^{k-1}\|}{\gamma_{k-1}}\right)^2+\left(\tfrac{\|\zeta^{k-1}\|}{\gamma_{k-1}}\right)^2.\end{aligned} \tag{4.16}$$

Then by (4.15) and the lower boundedness of $\left\{\mathcal{R}(\mathbf{s}^k)\right\}$, we have

$$\lim_{k\to\infty,k-1\in\mathcal{K}}\|\mathbf{u}^k-\Psi(\mathbf{h}^k)\mathbf{w}^k\|=0\quad\text{and}\quad\lim_{k\to\infty,k-1\in\mathcal{K}}\|\mathbf{h}^k-(\mathbf{u}^k)_+\|=0. \tag{4.17}$$

To extend the results in (4.17) to any $k > K$, let l_k denote the largest element in $\mathcal{K}$ satisfying $l_k < k$. If $l_k = k-1$, the limitations are the same as (4.17). If $l_k < k-1$, let us define an index set $\mathcal{I}_k := \{i : l_k < i < k\}$. The updating rule for the penalty parameter, as stated in (3.9), implies that $\gamma_i = \gamma_{l_k}$. This, combined with the updating rules for the Lagrangian multipliers, yields that for all $i \in \mathcal{I}_k$, the following holds:

$$\frac{\|\xi^i\|}{\gamma_i} = \frac{\|\xi^i\|}{\gamma_{i-1}} \leq \frac{\|\xi^{i-1}\|}{\gamma_{i-1}} + \left\|\mathbf{u}^i - \Psi(\mathbf{h}^i)\mathbf{w}^i\right\|, \tag{4.18}$$

$$\frac{\|\zeta^i\|}{\gamma_i} = \frac{\|\zeta^i\|}{\gamma_{i-1}} \leq \frac{\|\zeta^{i-1}\|}{\gamma_{i-1}} + \left\|\mathbf{h}^i - (\mathbf{u}^i)_+\right\|. \tag{4.19}$$

Summing up inequalities (4.18) and (4.19) for every $i \in \mathcal{I}_k$, we have

$$\frac{\|\xi^{k-1}\|}{\gamma_{k-1}} \leq \frac{\|\xi^{l_k}\|}{\gamma_{l_k}} + \sum_{i=1}^{k-l_k-1} \left\|\mathbf{u}^{k-i} - \Psi(\mathbf{h}^{k-i})\mathbf{w}^{k-i}\right\|, \tag{4.20}$$

$$\frac{\|\zeta^{k-1}\|}{\gamma_{k-1}} \leq \frac{\|\zeta^{l_k}\|}{\gamma_{l_k}} + \sum_{i=1}^{k-l_k-1} \left\|\mathbf{h}^{k-i} - (\mathbf{u}^{k-i})_+\right\|. \tag{4.21}$$

By the updating rule of γ_k in (3.8), (4.20) and (4.21), we obtain

$$\frac{\|\xi^{k-1}\|}{\gamma_{k-1}} \leq \frac{\|\xi^{l_k}\|}{\gamma_{l_k}} + \frac{\eta_1}{1-\eta_1} \max\left\{\left\|\mathbf{u}^{l_k+1} - \Psi(\mathbf{h}^{l_k+1})\mathbf{w}^{l_k+1}\right\|, \left\|\mathbf{h}^{l_k+1} - (\mathbf{u}^{l_k+1})_+\right\|\right\},$$

$$\frac{\|\zeta^{k-1}\|}{\gamma_{k-1}} \leq \frac{\|\zeta^{l_k}\|}{\gamma_{l_k}} + \frac{\eta_1}{1-\eta_1} \max\left\{\left\|\mathbf{u}^{l_k+1} - \Psi(\mathbf{h}^{l_k+1})\mathbf{w}^{l_k+1}\right\|, \left\|\mathbf{h}^{l_k+1} - (\mathbf{u}^{l_k+1})_+\right\|\right\}.$$

This, together with (4.15), (4.17) and $\eta_1 \in (0,1)$, yields that

$$\lim_{k\to\infty} \frac{\left\|\xi^{k-1}\right\|}{\gamma_{k-1}} = 0, \quad \lim_{k\to\infty} \frac{\left\|\zeta^{k-1}\right\|}{\gamma_{k-1}} = 0. \tag{4.22}$$

By the inequality (4.16) and nondecreasing sequence $\{\gamma_k\}$, we conclude that

$$\lim_{k\to\infty} \|\mathbf{u}^k - \Psi(\mathbf{h}^k)\mathbf{w}^k\| = 0, \quad \lim_{k\to\infty} \|\mathbf{h}^k - (\mathbf{u}^k)_+\| = 0, \tag{4.23}$$

using the same manner for showing (4.17).

(ii) When $\mathcal{K}$ is finite, there exists a constant K such that $\gamma_{k-1} = \gamma_K$ for those $k > K$. Then, we turn to consider the boundedness of $\{\xi^{k-1}\}$ and $\{\zeta^{k-1}\}$. Summing up (3.7) for those $k > K$, and using (3.8), we find

$$\begin{aligned} &\max\{\{\|\xi^{k-1}\|, \|\zeta^{k-1}\|\} \\ &\leq \max\{\|\xi^K\|, \|\zeta^K\|\} + \frac{\eta_1\gamma_K}{1-\eta_1} \max\left\{\left\|\mathbf{u}^K - \Psi(\mathbf{h}^K)\mathbf{w}^K\right\|, \left\|\mathbf{h}^K - (\mathbf{u}^K)_+\right\|\right\}. \end{aligned}$$

From the above, the boundedness of $\{\xi^{k-1}\}$ and $\{\zeta^{k-1}\}$ are thus proved. Together with $\gamma_{k-1} = \gamma_K$ for those $k > K$, we can further deduce that $\|\xi^{k-1}\|^2/\gamma_{k-1}$ and $\|\zeta^{k-1}\|^2/\gamma_{k-1}$ are bounded for those $k \in \mathbb{N}_+$.

When the set $\mathcal{K}$ is infinite, by (4.15) we know that $\|\xi^{k-1}\|^2/\gamma_{k-1}$ and $\|\zeta^{k-1}\|^2/\gamma_{k-1}$ are bounded for $k-1 \in \mathcal{K}$. Therefore, no matter $\mathcal{K}$ is finite or infinite, $\|\xi^{k-1}\|^2/\gamma_{k-1}$ and $\|\zeta^{k-1}\|^2/\gamma_{k-1}$ are bounded for $k-1 \in \mathcal{K}$.

Moreover, we can deduce the following inequality according to the expression of $\mathcal{L}_{k-1}$, condition (A.15), and $\mathbf{s}^k = \mathbf{s}^{k-1,j}$:

$$\begin{aligned} &\mathcal{R}(\mathbf{s}^k) + \frac{\gamma_{k-1}}{2}\left\|\mathbf{u}^k - \Psi(\mathbf{h}^k)\mathbf{w}^k + \frac{\xi^{k-1}}{\gamma_{k-1}}\right\|^2 + \frac{\gamma_{k-1}}{2}\left\|\mathbf{h}^k - (\mathbf{u}^k)_+ + \frac{\zeta^{k-1}}{\gamma_{k-1}}\right\|^2 \\ &\le \Gamma + \frac{\|\xi^{k-1}\|^2}{2\gamma_{k-1}} + \frac{\|\zeta^{k-1}\|^2}{2\gamma_{k-1}}. \end{aligned} \tag{4.24}$$

The above inequality, along with the boundedness of $\left\{\|\xi^{k-1}\|^2/\gamma_{k-1}\right\}_{k-1\in\mathcal{K}}$ and $\left\{\|\zeta^{k-1}\|^2/\gamma_{k-1}\right\}_{k-1\in\mathcal{K}}$, yields the boundedness of $\{\mathbf{s}^k\}_{k-1\in\mathcal{K}}$ by the same manner in Lemma 3.1 (ii). Hence there exists at least one accumulation point of $\{\mathbf{s}^k\}$.

Any accumulation point is a feasible point of (2.6), which can be derived immediately by (i), because of the continuity of the functions in the constraints of (2.6). ∎

Below we show the main convergence result of the ALM.

THEOREM 4.8. *Every accumulation point of* $\{\mathbf{s}^k\}$ *generated by Algorithm* 3.1 *is a KKT point of problem* (2.6).

Proof. Let $\{\mathbf{s}^{k_i}\}$ be a subsequence of $\{\mathbf{s}^k\}$ converging to $\bar{\mathbf{s}}$. Then $\bar{s} \in \mathcal{F}$ by Theorem 4.7. We claim that

$$\begin{aligned} &\partial\mathcal{L}\big(\mathbf{s}^{k_i}, \xi^{k_i-1}, \zeta^{k_i-1}, \gamma_{k_i-1}\big) \\ &= \nabla\mathcal{R}(\mathbf{s}^{k_i}) + \nabla_{\mathbf{s}}\Big(\langle\xi^{k_i-1}, \mathbf{u}^{k_i} - \Psi(\mathbf{h}^{k_i})\mathbf{w}^{k_i}\rangle + \tfrac{\gamma_{k_i-1}}{2}\big\|\mathbf{u}^{k_i} - \Psi(\mathbf{h}^{k_i})\mathbf{w}^{k_i}\big\|^2\Big) \\ &\quad + \partial_{\mathbf{s}}\Big(\langle\zeta^{k_i-1}, \mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+\rangle + \tfrac{\gamma_{k_i-1}}{2}\big\|\mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+\big\|^2\Big) \\ &= \nabla\mathcal{R}(\mathbf{s}^{k_i}) + J\mathcal{C}_1(\mathbf{s}^{k_i})^\top\xi^{k_i} + \partial\Big((\zeta^{k_i})^\top\mathcal{C}_2(\mathbf{s}^{k_i})\Big), \end{aligned} \tag{4.25}$$

where $\mathcal{C}_1$ and $\mathcal{C}_2$ are defined in (2.4).

First, by employing (3.7) and by direct computation, we have

$$\begin{aligned} &\nabla_{\mathbf{s}}\big(\langle\xi^{k_i-1}, \mathbf{u}^{k_i} - \Psi(\mathbf{h}^{k_i})\mathbf{w}^{k_i}\rangle + \tfrac{\gamma_{k_i-1}}{2}\big\|\mathbf{u}^{k_i} - \Psi(\mathbf{h}^{k_i})\mathbf{w}^{k_i}\big\|^2\big) \\ &= J\mathcal{C}_1(\mathbf{s}^{k_i})^\top\big(\xi^{k_i-1} + \gamma_{k_i-1}(\mathbf{u}^{k_i} - \Psi(\mathbf{h}^{k_i})\mathbf{w}^{k_i})\big) = J\mathcal{C}_1(\mathbf{s}^{k_i})^\top\xi^{k_i}. \end{aligned} \tag{4.26}$$

Then, it remains to verify that

$$\partial_{\mathbf{s}}(\langle\zeta^{k_i-1}, \mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+\rangle + \tfrac{\gamma_{k_i-1}}{2}\|\mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+\|^2) = \partial\Big((\zeta^{k_i})^\top\mathcal{C}_2(\mathbf{s}^{k_i})\Big). \tag{4.27}$$

To verify (4.27), it can be divided into the subdifferential associated with $\mathbf{h}$ and $\mathbf{u}$. We first prove that (4.27) is satisfied associated with $\mathbf{h}$. By simple computation,

$$\begin{aligned} &\nabla_{\mathbf{h}}\big(\langle\zeta^{k_i-1}, \mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+\rangle + \tfrac{\gamma_{k_i-1}}{2}\big\|\mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+\big\|^2\big) \\ &= J_{\mathbf{h}}\mathcal{C}_2(\mathbf{z}^{k_i}, \mathbf{h}^{k_i}, \mathbf{u}^{k_i})^\top\big(\zeta^{k_i-1} + \gamma_{k_i-1}(\mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+)\big) \\ &= J_{\mathbf{h}}\mathcal{C}_2(\mathbf{z}^{k_i}, \mathbf{h}^{k_i}, \mathbf{u}^{k_i})^\top\zeta^{k_i} = \nabla_{\mathbf{h}}\big(\langle\zeta^{k_i}, \mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+\rangle\big). \end{aligned} \tag{4.28}$$

Then we prove that (4.27) is satisfied associated with $\mathbf{u}$, which can be replaced by proving rT one dimensional equations with the similar structure as follows:

$$\partial_{\mathbf{u}_j}\Big(\zeta_j^{k_i-1}(\mathbf{h}_j^{k_i} - (\mathbf{u}_j^{k_i})_+) + \tfrac{\gamma_{k_i-1}}{2}(\mathbf{h}_j^{k_i} - (\mathbf{u}_j^{k_i})_+)^2\Big) = \partial_{\mathbf{u}_j}\Big(\zeta_j^{k_i}(\mathbf{h}_j^{k_i} - (\mathbf{u}_j^{k_i})_+)\Big), \tag{4.29}$$

where $j = 1, 2, ..., rT$. When $\mathbf{u}_j^{k_i} \neq 0$, equation (4.29) can be easily deduced by the same proof method as in (4.28). When $\mathbf{u}_j^{k_i} = 0$, the validity of (4.29) can be proved as follows:

$$
\begin{aligned}
&\partial_{\mathbf{u}_j}\Big(\zeta_j^{k_i-1}(\mathbf{h}_j^{k_i} - (\mathbf{u}_j^{k_i})_+) + \tfrac{\gamma_{k_i-1}}{2}(\mathbf{h}_j^{k_i} - (\mathbf{u}_j^{k_i})_+)^2\Big)\\
&= \begin{cases} \{0, -\zeta_j^{k_i-1} - \gamma_{k_i-1}(\mathbf{h}_j^{k_i} - \mathbf{u}_j^{k_i})\}, & \text{if } \gamma_{k_i-1}\mathbf{h}_j^{k_i} + \zeta_j^{k_i-1} \geq 0,\\ \big[0, -\zeta_j^{k_i-1} - \gamma_{k_i-1}(\mathbf{h}_j^{k_i} - \mathbf{u}_j^{k_i})\big], & \text{if } \gamma_{k_i-1}\mathbf{h}_j^{k_i} + \zeta_j^{k_i-1} < 0,\end{cases}\\
&= \begin{cases} \{0, -\zeta_j^{k_i}\}, & \text{if } \zeta_j^{k_i} \geq 0,\\ \big[0, -\zeta_j^{k_i}\big], & \text{if } \zeta_j^{k_i} < 0,\end{cases}\\
&= \partial_{\mathbf{u}_j}\Big(\zeta_j^{k_i}(\mathbf{h}_j^{k_i} - (\mathbf{u}_j^{k_i})_+)\Big).
\end{aligned} \tag{4.30}
$$

Combining (4.26) and (4.27) yields the validity of (4.25).

Up to now, we have verified that equation (4.25) holds. Thus, there exists a sequence $\{\varsigma^{k_i}\}$ satisfying $\|\varsigma^{k_i}\| \leq \epsilon^{k_i}$ such that

$$
\varsigma^{k_i} \in \nabla\mathcal{R}(\mathbf{s}^{k_i}) + J\mathcal{C}_1(\mathbf{s}^{k_i})^\top \xi^{k_i} + \partial\Big((\zeta^{k_i})^\top \mathcal{C}_2(\mathbf{s}^{k_i})\Big). \tag{4.31}
$$

However, the boundedness of $\{\xi^{k_i}\}$ and $\{\zeta^{k_i}\}$ in (4.31) are still not sure. Define $\varrho^i = \max\{\|\xi^{k_i}\|_\infty, \|\zeta^{k_i}\|_\infty\}$ and assume that $\{\varrho^i\}$ is unbounded. It is trivial to have bounded sequences $\{\xi^{k_i}/\varrho^i\}$ and $\{\zeta^{k_i}/\varrho^i\}$ according to the definition of ϱ^i. Without loss of generality, we assume $\{\xi^{k_i}/\varrho^i\} \to \bar{\xi}$ and $\{\zeta^{k_i}/\varrho^i\} \to \bar{\zeta}$ as $k \to \infty$ and thus have

$$
\max\{\|\bar{\xi}\|_\infty, \|\bar{\zeta}\|_\infty\} = 1. \tag{4.32}
$$

Dividing by ϱ^i on both sides of (4.31) and taking $i \to \infty$, and using the facts that the limiting subdifferential is outer semicontinuous [26, Proposition 8.7], and $\varsigma^{k_i} \to 0$ as $i \to \infty$, we derive that

$$
0 \in J\mathcal{C}_1(\bar{\mathbf{s}})^\top \bar{\xi} + \partial\Big(\bar{\zeta}^\top \mathcal{C}_2(\bar{\mathbf{s}})\Big). \tag{4.33}
$$

Combining (4.33) and Lemma 2.1 yields that $\bar{\xi} = 0$ and $\bar{\zeta} = 0$, which contradicts (4.32). Therefore, $\{\xi^{k_i}\}$ and $\{\zeta^{k_i}\}$ are bounded. Without loss of generality, we assume $\{\xi^{k_i}\} \to \bar{\xi}$ and $\{\zeta^{k_i}\} \to \bar{\zeta}$ as $i \to \infty$. Letting $i \to \infty$ in (4.31), we obtain

$$
0 \in \nabla\mathcal{R}(\bar{\mathbf{s}}) + J\mathcal{C}_1(\bar{\mathbf{s}})^\top \bar{\xi} + \partial\Big(\bar{\zeta}^\top \mathcal{C}_2(\bar{\mathbf{s}})\Big).
$$

Therefore, $\bar{\mathbf{s}}$ is a KKT point of problem (2.6). ∎

4.3. Extensions to other activation functions. Now we discuss the possible extensions of our methods, algorithms and theoretical analysis, using other activation functions rather than the ReLU.

First, we claim that the activation functions are required to be locally Lipschitz continuous, because the locally Lipschitz continuity of the ReLU function is used in $L_2(\xi, \zeta, \gamma, \hat{r})$ of Lemma 3.2 that depends on the Lipschitz constant of the ReLU function on a compact set. Then we find that in the analysis above only the following two places make use of the special piecewise linear structure of the ReLU function:

P1. Explicit formula for $\mathbf{u}^{k-1,j}$ in (3.16) of the BCD method in Algorithm 3.2.

P2. Equations (4.30) for proving (4.29) in the proof of Theorem 4.8.

For P1, even if the activation function in (2.1) is replaced by others, the objective function in problem (3.16) can still be separated into rT one-dimensional functions, which is obtained by substituting the ReLU function $(u)_+$ in (3.19) by a more general activation function. For P2, if an arbitrary smooth activation function is considered, then (4.29) holds obviously because the limiting subdifferential reduces to the gradient. Below we illustrate in detail the leaky ReLU and the ELU activation functions as examples for extensions. It is clear that the expression of $L_2(\xi,\zeta,\gamma,\hat{r})$ in Lemma 3.2 remains unchanged for the two activation functions because they all have Lipschitz constant 1, the same as that of the ReLU.

Extension to the leaky ReLU. Let us replace the ReLU activation function $\sigma(u) = (u)_+$ with the leaky ReLU activation function defined by

$$\sigma_{\rm lRe}(u) := \max\{u, \varpi u\},$$

where $\varpi \in (0,1)$ is a fixed parameter. The leaky ReLU activation function has been widely used in recent years. With regard to P1, by direct computation, a closed-form global solution of

$$\min_{u\in\mathbb{R}} \varphi_{\rm lRe}(u) := \tfrac{\gamma}{2}(u-\theta_1)^2 + \tfrac{\gamma}{2}(\theta_2 - \sigma_{\rm lRe}(u))^2 + \tfrac{\mu}{2}(u-\theta_3)^2 + \lambda_6 u^2, \tag{4.34}$$

can be obtained similarly using the procedures for ReLU in (3.20)-(3.22), except that the expression u^- of (3.22) changes to

$$u^- = \begin{cases} \dfrac{\gamma\theta_1 + \gamma\varpi\theta_2 + \mu\theta_3}{\gamma + \gamma\varpi^2 + 2\lambda_6 + \mu}, & \text{if } \gamma\theta_1 + \mu\theta_3 < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4.35}$$

For P2, (4.30) is modified as follows: when $\mathbf{u}_j^{k_i} = 0$,

$$\begin{aligned} &\partial_{\mathbf{u}_j}\left(\zeta_j^{k_i-1}(\mathbf{h}_j^{k_i} - \sigma_{\rm lRe}(\mathbf{u}_j^{k_i})) + \tfrac{\gamma_{k_i-1}}{2}(\mathbf{h}_j^{k_i} - \sigma_{\rm lRe}(\mathbf{u}_j^{k_i}))^2\right) \\ =& \begin{cases} \{-\varpi\zeta_j^{k_i}, -\zeta_j^{k_i-1} - \gamma_{k_i-1}(\mathbf{h}_j^{k_i} - \mathbf{u}_j^{k_i})\}, & \text{if } \gamma_{k_i-1}\mathbf{h}_j^{k_i} + \zeta_j^{k_i-1} \ge 0, \\ \big[-\varpi\zeta_j^{k_i}, -\zeta_j^{k_i-1} - \gamma_{k_i-1}(\mathbf{h}_j^{k_i} - \mathbf{u}_j^{k_i})\big], & \text{if } \gamma_{k_i-1}\mathbf{h}_j^{k_i} + \zeta_j^{k_i-1} < 0, \end{cases} \\ =& \begin{cases} \{-\varpi\zeta_j^{k_i}, -\zeta_j^{k_i}\}, & \text{if } \zeta_j^{k_i} \ge 0, \\ \big[-\varpi\zeta_j^{k_i}, -\zeta_j^{k_i}\big], & \text{if } \zeta_j^{k_i} < 0, \end{cases} \\ =& \,\partial_{\mathbf{u}_j}\left(\zeta_j^{k_i}(\mathbf{h}_j^{k_i} - \sigma_{\rm lRe}(\mathbf{u}_j^{k_i}))\right). \end{aligned} \tag{4.36}$$

Extension to the ELU. Let us replace the ReLU activation function with the convex and smooth activation function ELU defined by

$$\sigma_{\rm ELU}(u) := \begin{cases} u & \text{if } u \ge 0, \\ e^u - 1 & \text{if } u < 0. \end{cases}$$

When $u \ge 0$, the ELU activation function is the same as the ReLU function. Thus for P1, the solution of (4.34) can be obtained similarly as the ReLU case, except that we do not have the explicit formula of u^-, which is a global solution of

$$\min_{u\in(-\infty,0]} \varphi_{\rm ELU}(u) = \tfrac{\gamma}{2}(u-\theta_1)^2 + \tfrac{\gamma}{2}(\theta_2 - (e^u - 1))^2 + \tfrac{\mu}{2}(u-\theta_3)^2 + \lambda_6 u^2, \tag{4.37}$$

due to the presence of the exponential function in the ELU activation function.

Now we illustrate that u^- can be obtained numerically through solving several one-dimensional minimization problems. First, using the formula of $\varphi_{\mathrm{ELU}}(u)$ and the fact that $\varphi_{\mathrm{ELU}}(u) \to +\infty$ as $u \to -\infty$, we can easily find a lower bound $\underline{u} < 0$ such that (4.37) is equivalent to

$$\min_{u\in[\underline{u},0]} \varphi_{\mathrm{ELU}}(u). \tag{4.38}$$

The objective function $\varphi_{\mathrm{ELU}}(u)$ is smooth on $(-\infty, 0]$. We thus calculate the second-order derivative of $\varphi_{\mathrm{ELU}}(u)$ as

$$\varphi''_{\mathrm{ELU}}(u) = 2\gamma e^{2u} - \gamma(\theta_2 + 1)e^u + \mu + \gamma + 2\lambda_6. \tag{4.39}$$

Let $z = e^u$. (4.39) can be represented as

$$\psi_{\mathrm{ELU}}(z) := 2\gamma z^2 - \gamma(\theta_2 + 1)z + \mu + \gamma + 2\lambda_6, \tag{4.40}$$

which is a quadratic function. Hence there are at most two distinct roots of

$$\psi_{\mathrm{ELU}}(z) = 0,$$

and consequently at most two distinct roots for $\varphi''(u) = 0$ on $[\underline{u}, 0]$. Hence the convexity and concavity can only be changed at most three times in $[\underline{u}, 0]$. That is, we can divide $[\underline{u}, 0]$ into at most three closed intervals, and in each interval φ_{ELU} is either convex or concave. We minimize the objective function φ_{ELU} in each of those intervals that φ_{ELU} is convex, and obtain a global solution in each interval numerically. Then, we select a point among those solutions, 0, and $\underline{u}$ that has the minimal objective value. This point is a global solution of (4.37).

5. Numerical experiments.

We employ a real world dataset, **Volatility of S&P index**, and synthetic datasets to evaluate the effectiveness of our reformulation (2.6) and Algorithm 3.1 with Algorithm 3.2. To be specific, we first use RNNs with unknown weighted matrices to model these sequential datasets, and then utilize the ALM with the BCD method to train RNNs. After the training process, we can predict future values of these sequential datasets using the trained RNNs.

The numerical experiments consist of two components. The first part involves assessing whether the outputs generated by the ALM adhere to the constraints in (2.6). The second part is to compare the training and forecasting performance of the ALM with state-of-the-art gradient descent-based algorithms (GDs). All the numerical experiments were conducted using Python 3.9.8. For the datasets, **Synthetic dataset ($T = 10$)** and **Volatility of S&P index**, experiments were carried out on a desktop (Windows 10 with 2.90 GHz Inter Core i7-10700 CPU and 32GB RAM). Additionally, experiments for **Synthetic dataset ($T = 500$)** were implemented on a server (2 Intel Xeon Gold 6248R CPUs and 768GB RAM) at the high-performance servers of the Department of Applied Mathematics, the Hong Kong Polytechnic University.

5.1. Datasets.

The process of generating synthetic datasets is as follows. We randomly generate the weighted matrices $\hat{A}$, $\hat{W}$, $\hat{V}$, the bias vectors $\hat{b}$, $\hat{c}$, and the noises $\tilde{e}_t$, $t = 1, 2, ..., T$, and the input data X with some distributions. Then we calculate the output data $Y = (y_1; \ldots; y_t)$ by $y_t = (\hat{A}(\hat{W}(...(\hat{V}x_1 + \hat{b})_+ ...) + \hat{V}x_t + \hat{b})_+ + \hat{c}) + \tilde{e}_t$ for $t \in [T]$. In the numerical experiments, we generate two synthetic datasets with $T = 10$ and $T = 500$. The detailed information of the two synthetic datasets is listed

Table 1: Synthetic datasets

T	n	m	r	Distributions		
				weight matrices	the noise	the input data
10	5	3	4	$\mathcal{N}(0, 0.8)$	$\mathcal{N}(0, 10^{-3})$	$\mathcal{U}(-1, 1)$
500	80	30	100	$\mathcal{N}(0, 0.05)$	$\mathcal{N}(0, 10^{-5})$	$\mathcal{U}(-1, 1)$

in Table 1. Moreover, the ratio of splitting for the training and test sets is about 9 : 1.

The dataset, **Volatility of S&P index**, consists of the monthly realized volatility of the S&P index and 11 corresponding exogenous variables from February 1973 to June 2009, totaling 437 time steps, i.e., $T = 437$, $n = 11$ and $m = 1$. The dataset was collected in strict adherence to the guidelines in [6] and contains no missing values. In the dataset, the monthly realized volatility of S&P index is appointed as the output variable, while 11 exogenous variables are input variables. For training the RNNs, we first standardize the dataset as zero mean and unit variance, and then allocate 90% of the dataset, consisting of 393 time steps, as the training set, while the remaining 44 time steps are the test set. Moreover, we have $r = 20$ for the real dataset.

5.2. Evaluations. We define **FeasVio** $:= \max\{\|\mathbf{u} - \Psi(\mathbf{h})\mathbf{w}\|, \|\mathbf{h} - (\mathbf{u})_+\|\}$ to evaluate the feasibility violation for constraints $\mathbf{u} = \Psi(\mathbf{h})\mathbf{w}$ and $\mathbf{h} = (\mathbf{u})_+$. Moreover, the training and test errors are used to evaluate the forecasting accuracy of RNNs in training and test sets denoted as

$$\mathbf{TrainErr} := \frac{1}{T_1}\sum_{t=1}^{T_1} \|y_t - (A(W(...(Vx_1 + b)_+ ...) + Vx_t + b)_+ + c\|^2,$$

$$\mathbf{TestErr} := \frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2} \|y_t - (A(W(...(Vx_1 + b)_+ ...) + Vx_t + b)_+ + c)\|^2,$$

where T_1 and T_2 are the time lengths of the training set and test set, and A, W, V, b and c are the output solutions from ALM.

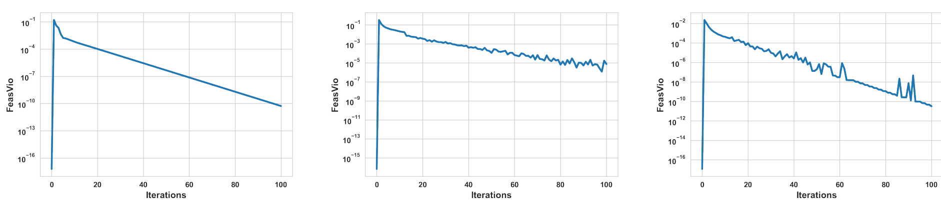
5.3. Investigating the feasibility. In this subsection, we aim to verify the outputs from the ALM satisfying the constraints of (2.2) through numerical experiments, while we have already proved the feasibility of any accumulation point of a sequence generated by the ALM in section 4. Initial values of weight matrices A^0, W^0, V^0 are randomly generated from the standard Gaussian distribution $\mathcal{N}(0, 0.1)$. Moreover, the bias b^0 and c^0 are set as 0. For all three datasets, we stop the outer loop (ALM) when it reaches 100 iterations, and the inner loop (BCD method) terminates at 500 iterations. Other parameters are listed in Table 2.

From Figure 1, we observe that the feasibility violation in each dataset is very small at the beginning, which implies that the selected initial point is feasible. As it turns to the first iteration, the feasibility violation goes to a large value. After that, the value goes to exhibit an oscillatory decrease and tends to zero. This indicates that the points generated by the ALM gradually satisfy the constraint conditions as the number of iterations increases.

5.4. Comparisons with state-of-the-art GDs. In this subsection, we compare the training and forecasting accuracy of RNNs using different methods. Specifi-

Table 2: Parameters of the ALM: the parameters for the given datasets are set as $\gamma^0 = 1$, $\xi^0 = \mathbf{0}$, $\zeta^0 = \mathbf{0}$, $\epsilon_0 = 0.1$, $\Gamma = 10^2$, $\mu = 10^{-5}$, $\lambda_1 = \tau/rm$, $\lambda_2 = \tau/r^2$, $\lambda_3 = \tau/rn$, $\lambda_4 = \tau/r$, $\lambda_5 = \tau/m$, $\lambda_6 = 10^{-8}$.

Datasets	Regularization parameters	Algorithm parameters
Synthetic dataset ($T = 10$)	$\tau = 1.2$	$\eta_1 = 0.99$, $\eta_2 = 5/6$, $\eta_3 = 0.01$, $\eta_4 = 5/6$.
Volatility of S&P index	$\tau = 1$	
Synthetic dataset ($T = 500$)	$\tau = 500$	$\eta_1 = 0.90$, $\eta_2 = 0.90$, $\eta_3 = 0.015$, $\eta_4 = 0.8$.



(a) Synthetic dataset ($T = 10$) (b) Synthetic dataset ($T = 500$) (c) Volatility of S&P index

Fig. 1: The feasibility violation of the ALM in different datasets

cally, we compare our ALM with the state-of-the-art GDs and SGDs with special techniques, i.e., gradient descent (GD), gradient descent with gradient clipping (GDC), gradient descent with Nesterov momentum (GDNM), Mini-batch SGD and Adam.

For the initial values of A^0, W^0, V^0, we use the following initialization strategies: random normal initialization [2] with zero mean and standard deviations of 10^{-3} and 10^{-1}, He initialization [32], Glorot initialization [33], and LeCun initialization [34]. Notably, the initial values of bias, b^0 and c^0, were both set to 0 according to [14, pp. 305].

We search the learning rates for GDs and SGDs over $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, as well as the clipping norm of GDC over $\{0.5, 1, 1.5, 2, 3, 4, 5, 6\}$. We employ the leave-P-out cross validation and repeated each method 30 trials with $P = 1$ in **Synthetic dataset** ($T = 10$), and $P = 10$ in **Volatility of S&P index** and **Synthetic dataset** ($T = 500$). We then select the learning rates and clipping norm with the best test error averaged over 30 trials, which are recorded in Table 4 of Appendix B. The batch size for SGDs is set to 2 for **Synthetic dataset** ($T = 10$), 50 for **Volatility of S&P index**, and 100 for **Synthetic dataset** ($T = 500$). We employ the Keras API [10] running on TensorFlow 2 to implement the GDs and SGDs. Additionally, the parameters for the ALM are listed in Table 2.

To evaluate the performance of different methods under various initialization strategies, we conducted the following experiments: each method was repeated 10 times under each initialization strategy. In each repetition, we recorded the final test error and the training error. We then calculated their means (**TrainErr** and **TestErr**) and the corresponding standard deviations, and listed them in Table 3. Each row records the results for a certain optimization method from different initialization strategies, with the best **TrainErr** or **TestErr** highlighted in bold. Each column provides the results of all the optimization methods with the same initial values, where the best **TrainErr** and **TestErr** are highlighted underline.

Table 3a and Table 3c demonstrate that for **Synthetic dataset** ($T = 10$) and **Synthetic dataset** ($T = 500$), no matter which initialization strategy is employed, our ALM method achieves the best **TrainErr** and **TestErr** among all the methods. Table 3b illustrates that our ALM achieves the best **TrainErr** under two types of initialization strategies, and obtains the best **TestErr** under three types of initialization strategies for **Volatility of S&P index**. For any of the three datasets, our ALM achieves the best **TrainErr** and **TestErr** among all combinations of optimization methods and initialization strategies, which we highlight in blue.

Table 3: Results of training Elman RNNs using different optimization methods and initialization strategies across multiple trials.

(a) **Synthetic dataset** ($T = 10$): For the ALM method, the maximum iteration for the outer loop is 50 and 10 for the inner loop. For GDs and SGDs, the number of epochs is set to 500.

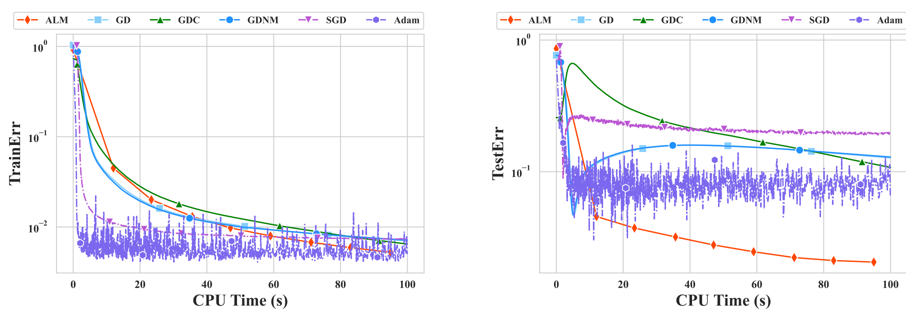
		He	$\mathcal{N}(0, 10^{-3})$	$\mathcal{N}(0, 10^{-1})$	Glorot	LeCun
ALM	**TrainErr**	0.345 ± 0.24	$\mathbf{0.113 \pm 0.03}$	0.143 ± 0.04	0.206 ± 0.10	0.279 ± 0.22
	TestErr	4.770 ± 1.25	$\mathbf{4.437 \pm 0.28}$	4.660 ± 0.35	4.628 ± 1.17	4.650 ± 0.62
GD	**TrainErr**	4.459 ± 0.77	$2.747 \pm 1.5\text{e-}6$	2.768 ± 0.01	1.814 ± 0.27	$\mathbf{1.604 \pm 0.17}$
	TestErr	6.432 ± 2.15	$5.311 \pm 9.3\text{e-}6$	5.057 ± 0.07	$\mathbf{4.696 \pm 0.90}$	5.056 ± 1.10
GDC	**TrainErr**	$\mathbf{1.479 \pm 0.32}$	$2.769 \pm 1.4\text{e-}6$	2.768 ± 0.01	1.684 ± 0.23	1.502 ± 0.26
	TestErr	5.376 ± 0.88	$5.079 \pm 1.0\text{e-}6$	5.057 ± 0.07	$\mathbf{4.922 \pm 1.20}$	5.266 ± 0.96
GDNM	**TrainErr**	2.689 ± 0.40	$2.769 \pm 1.4\text{e-}6$	2.768 ± 0.01	3.340 ± 0.54	$\mathbf{0.801 \pm 0.60}$
	TestErr	6.169 ± 2.06	$5.079 \pm 1.0\text{e-}6$	5.057 ± 0.07	7.469 ± 2.30	$\mathbf{4.844 \pm 0.64}$
SGD	**TrainErr**	$\mathbf{2.224 \pm 0.02}$	2.247 ± 0.02	2.232 ± 0.02	2.238 ± 0.02	2.225 ± 0.02
	TestErr	6.455 ± 0.23	$\mathbf{6.230 \pm 0.23}$	6.373 ± 0.18	6.543 ± 0.23	6.446 ± 0.18
Adam	**TrainErr**	2.283 ± 0.07	2.244 ± 0.02	2.237 ± 0.02	$\mathbf{2.231 \pm 0.01}$	2.239 ± 0.03
	TestErr	$\mathbf{6.335 \pm 0.61}$	6.432 ± 0.27	6.411 ± 0.25	6.508 ± 0.14	6.406 ± 0.20

(b) **Volatility of S&P index**: For the ALM method, the maximum iteration for the outer loop is 200 and 500 for the inner loop. For GDs and SGDs, the number of epochs is set to 5000.

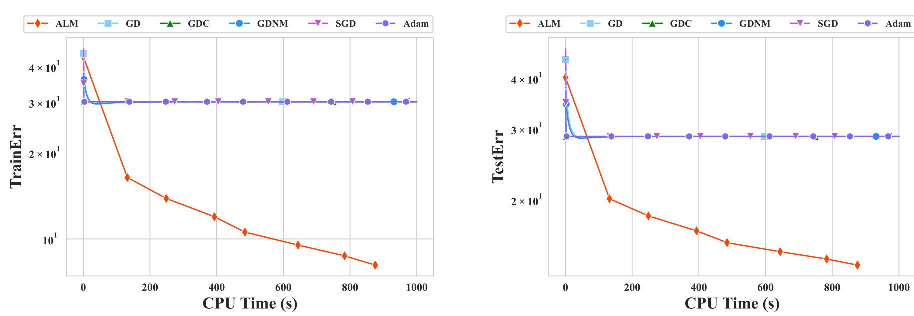
		He	$\mathcal{N}(0, 10^{-3})$	$\mathcal{N}(0, 10^{-1})$	Glorot	LeCun
ALM	**TrainErr**	0.058 ± 0.02	$0.004 \pm 3.6\text{e-}5$	$\mathbf{0.003 \pm 1.4}\text{e-}4$	0.009 ± 0.002	0.013 ± 0.002
	TestErr	0.229 ± 0.13	$0.041 \pm 4.7\text{e-}4$	$\mathbf{0.032 \pm 0.005}$	0.064 ± 0.04	0.053 ± 0.03
GD	**TrainErr**	$\mathbf{0.005 \pm 0.001}$	$0.015 \pm 1.8\text{e-}4$	$0.012 \pm 9.2\text{e-}4$	0.020 ± 0.003	0.025 ± 0.006
	TestErr	0.124 ± 0.10	0.077 ± 0.03	$\mathbf{0.0429 \pm 0.01}$	0.206 ± 0.20	0.307 ± 0.20
GDC	**TrainErr**	0.567 ± 0.47	$0.015 \pm 1.8\text{e-}4$	0.016 ± 0.009	$\mathbf{0.003 \pm 5.6}\text{e-}4$	0.011 ± 0.003
	TestErr	1.135 ± 0.55	0.077 ± 0.03	0.047 ± 0.02	0.107 ± 0.03	$\mathbf{0.041 \pm 0.01}$
GDNM	**TrainErr**	0.005 ± 0.001	$0.015 \pm 1.8\text{e-}4$	$0.012 \pm 9.2\text{e-}4$	$\mathbf{0.003 \pm 5.8}\text{e-}4$	$0.004 \pm 6.6\text{e-}4$
	TestErr	0.124 ± 0.10	0.077 ± 0.03	$\mathbf{0.043 \pm 0.01}$	0.097 ± 0.03	0.102 ± 0.02
SGD	**TrainErr**	$\mathbf{0.005 \pm 1.8}\text{e-}4$	0.006 ± 0.002	0.006 ± 0.002	0.006 ± 0.002	0.006 ± 0.002
	TestErr	$\mathbf{0.072 \pm 0.01}$	0.095 ± 0.02	0.086 ± 0.02	0.085 ± 0.01	0.096 ± 0.01
Adam	**TrainErr**	0.006 ± 0.001	$\mathbf{0.005 \pm 7.6}\text{e-}4$	0.006 ± 0.002	0.006 ± 0.001	$\mathbf{0.005 \pm 7.6}\text{e-}4$
	TestErr	0.079 ± 0.01	$\mathbf{0.074 \pm 0.01}$	0.084 ± 0.01	0.080 ± 0.02	0.080 ± 0.02

(c) **Synthetic dataset** ($T = 500$): For the ALM method, the maximum iteration for the outer loop is 100 and 500 for the inner loop. For GDs and SGDs, the number of epochs is set to 1000.

		He	$\mathcal{N}(0, 10^{-3})$	$\mathcal{N}(0, 10^{-1})$	Glorot	LeCun
ALM	**TrainErr**	4.639 ± 0.78	$\mathbf{3.461 \pm 0.06}$	3.472 ± 0.05	3.472 ± 0.06	3.475 ± 0.06
	TestErr	14.77 ± 0.93	12.418 ± 0.16	12.407 ± 0.27	$\mathbf{12.394 \pm 0.22}$	12.517 ± 0.16
GD	**TrainErr**	58.137 ± 2.42	30.010 ± 0.003	30.013 ± 0.008	30.000 ± 0.008	$\mathbf{29.985 \pm 0.007}$
	TestErr	58.314 ± 2.76	28.644 ± 0.006	28.641 ± 0.009	28.630 ± 0.006	$\mathbf{28.626 \pm 0.009}$
GDC	**TrainErr**	250.471 ± 399.70	$\mathbf{30.004 \pm 0.003}$	30.144 ± 0.001	$30.143 \pm 8.8\text{e-}4$	30.144 ± 0.001
	TestErr	119.007 ± 66.71	$\mathbf{28.640 \pm 0.007}$	28.723 ± 0.007	28.730 ± 0.006	28.725 ± 0.01
GDNM	**TrainErr**	58.137 ± 2.42	30.010 ± 0.003	30.013 ± 0.008	30.000 ± 0.008	$\mathbf{29.985 \pm 0.007}$
	TestErr	58.314 ± 2.76	28.644 ± 0.006	28.641 ± 0.009	28.730 ± 0.006	$\mathbf{28.626 \pm 0.009}$
SGD	**TrainErr**	$\mathbf{30.142 \pm 3.5}$e-$\mathbf{6}$	$30.142 \pm 4.7\text{e-}6$	$30.142 \pm 5.2\text{e-}6$	$30.142 \pm 4.4\text{e-}6$	$30.142 \pm 4.8\text{e-}6$
	TestErr	$\mathbf{28.725 \pm 3.2}$e-$\mathbf{5}$	$28.725 \pm 4.4\text{e-}5$	$28.725 \pm 4.7\text{e-}5$	$28.725 \pm 3.9\text{e-}5$	$28.725 \pm 4.1\text{e-}5$
Adam	**TrainErr**	$30.142 \pm 7.1\text{e-}5$	$30.142 \pm 6.5\text{e-}5$	$30.142 \pm 7.3\text{e-}5$	$\mathbf{30.142 \pm 5.1}$e-$\mathbf{5}$	$30.142 \pm 5.7\text{e-}5$
	TestErr	$28.726 \pm 6.1\text{e-}4$	$28.725 \pm 5.0\text{e-}4$	$28.726 \pm 5.9\text{e-}4$	$28.726 \pm 5.0\text{e-}4$	$\mathbf{28.725 \pm 4.8}$e-$\mathbf{4}$



(a) **Volatility of S&P index**



(b) **Synthetic dataset** ($T = 500$)

Fig. 2: Comparisons of the performance of the ALM, GDs and SGDs across different datasets.

We plot in Figure 2 the **TrainErr** and **TestErr** versus CPU time measured in seconds using **Volatility of S&P index** and **Synthetic dataset** ($T = 500$). Each line corresponds to a certain optimization method as indicated in the legend, with its most appropriate initialization strategy that leads to the final **TestErr** in bold as outlined in Table 3. For the real world dataset, **Volatility of S&P index**, the ALM achieves the smallest test error among all the methods. For the larger-scale **Synthetic dataset** ($T = 500$) with $N_{\mathbf{w}} = 1.81 \times 10^4$, $N_{\mathbf{a}} = 3.03 \times 10^3$ and $r = 500$, the ALM exhibits superior performance in terms of both training and test errors.

6. Conclusion.

In this paper, the minimization model (1.1) for training RNNs is equivalently reformulated as problem (2.2) by using auxiliary variables. We propose the ALM in Algorithm 3.1 with Algorithm 3.2 to solve the regularized problem (2.6). The BCD method in Algorithm 3.2 is efficient for solving the subproblems of the ALM, which has a closed-form solution for each block problem. We establish the solid convergence results of the ALM to a KKT point of problem (2.6), as well as the finite termination of the BCD method for the subproblem of the ALM at each iteration. The efficiency and effectiveness of the ALM for training RNNs are demonstrated by numerical results with real world datasets and synthetic data, and comparison with state-of-art algorithms. An interesting further study is to extend our algorithm to a stochastic algorithm that is potential to deal with problems of huge samples efficiently. We believe that it is possible to extend our method and its corresponding analysis to other more complex RNN architectures, such as LSTMs, and we will give rigorous analysis in the near future.

Acknowledgments. We are grateful to Prof. M. Mahoney and the anonymous referees for valuable comments.

Appendix A. Proofs of the lemmas.

A.1. Proof of Lemma 2.1.

Proof. By direct computation,

$$J\mathcal{C}_1(\mathbf{s})^\top \xi + \partial\big(\zeta^\top \mathcal{C}_2(\mathbf{s})\big) = \begin{bmatrix} J_{\mathbf{z}}\mathcal{C}_1(\mathbf{s})^\top \xi \\ J_{\mathbf{h}}\mathcal{C}_1(\mathbf{s})^\top \xi + J_{\mathbf{h}}\mathcal{C}_2(\mathbf{s})^\top \zeta \\ J_{\mathbf{u}}\mathcal{C}_1(\mathbf{s})^\top \xi + \partial_{\mathbf{u}}\left(\zeta^\top \mathcal{C}_2(\mathbf{s})\right) \end{bmatrix}, \tag{A.1}$$

where

$$J_{\mathbf{h}}\mathcal{C}_1(\mathbf{s})^\top \xi + J_{\mathbf{h}}\mathcal{C}_2(\mathbf{s})^\top \zeta = \left[-W^\top \xi_2 + \zeta_1; ...; -W^\top \xi_T + \zeta_{T-1}; \zeta_T\right], \tag{A.2}$$

$$J_{\mathbf{u}}\mathcal{C}_1(\mathbf{s})^\top \xi + \partial_{\mathbf{u}}\left(\zeta^\top \mathcal{C}_2(\mathbf{s})\right) = \xi + \partial_{\mathbf{u}}(-\zeta^\top (\mathbf{u})_+). \tag{A.3}$$

In order to achieve $0 \in J\mathcal{C}_1(\mathbf{s})^\top \xi + \partial\big(\zeta^\top \mathcal{C}_2(\mathbf{s})\big)$, it is necessary to require $\zeta_T = 0$, which is located in the last row of $J_{\mathbf{h}}\mathcal{C}_1(\mathbf{s})^\top \xi + J_{\mathbf{h}}\mathcal{C}_2(\mathbf{s})^\top \zeta$. Using $\zeta_T = 0$ and (A.3), we find $\xi_T = 0$. Substituting the results into (A.2) and (A.3) recursively and using (A.2) and (A.3) equal 0, we can derive that there exist no nonzero vectors ξ and ζ such that $0 \in J\mathcal{C}_1(\mathbf{s})^\top \xi + \partial\big(\zeta^\top \mathcal{C}_2(\mathbf{s})\big)$. ∎

A.2. Proof of Lemma 2.4.

Proof. It is clear that $0 \in \mathcal{D}_{\mathcal{R}}(\rho)$ and consequently $\mathcal{D}_{\mathcal{R}}(\rho)$ is nonempty. Moreover,

$$\begin{aligned} &\|A\|_F^2 \le \rho/\lambda_1, \|W\|_F^2 \le \rho/\lambda_2, \|V\|_F^2 \le \rho/\lambda_3, \\ &\|b\|^2 \le \rho/\lambda_4, \|c\|^2 \le \rho/\lambda_5, \|\mathbf{u}\|^{\mathbf{2}} \le \rho/\lambda_6, \end{aligned} \tag{A.4}$$

from $\mathcal{R}(\mathbf{s}) \le \rho$, $\ell(\mathbf{s}) \ge 0$ and $P(\mathbf{s}) \ge 0$. Hence for $\mathbf{s} = (\mathbf{z}; \mathbf{h}; \mathbf{u}) \in \mathcal{D}_{\mathcal{R}}(\rho)$, $\mathbf{z}$ and $\mathbf{u}$ are bounded, and consequently $\mathbf{h}$ is also bounded because $\mathbf{h} = (\mathbf{u})_+$.

Up to now, we have obtained the boundedness of $\mathcal{D}_{\mathcal{R}}(\rho)$. By the continuity of $\mathcal{R}(\mathbf{s})$, we can assert that $\mathcal{D}_{\mathcal{R}}(\rho)$ is closed according to [26, Theorem 1.6]. Thus we can claim that the level set $\mathcal{D}_{\mathcal{R}}(\rho)$ is nonempty and compact for any $\rho > \mathcal{R}(0)$. Then the solution set $\mathcal{S}_1$ is nonempty and compact according to [5, Proposition A.8]. ∎

A.3. Proof of Lemma 3.1.

Proof. Statement (i) can be easily obtained by the expression of $\mathcal{L}(\mathbf{s}, \xi, \zeta, \gamma)$ in (3.1) and the nonnegativity of $\mathcal{R}(\mathbf{s})$ in (2.6).

For statement (ii), the nonemptyness and closedness of the level set $\Omega_{\mathcal{L}}(\hat{\Gamma})$ are obvious. Moreover, we have $\mathcal{R}(\mathbf{s})$ and $\|\mathbf{h} - (\mathbf{u})_+ + \frac{\zeta}{\gamma}\|$ are upper bounded for all $\mathbf{s}$ in $\Omega_{\mathcal{L}}(\hat{\Gamma})$. The function $\mathcal{R}(\mathbf{s})$ is upper bounded implies that $\mathbf{w}, \mathbf{a}, \mathbf{u}$ are bounded. Then the boundedness of $\|\mathbf{h} - (\mathbf{u})_+ + \frac{\zeta}{\gamma}\|$ indicates that $\mathbf{h}$ is also bounded. Thus, $\mathbf{s}$ is bounded and statement (ii) holds.

Statements (iii) and (iv) can be obtained by direct computation. ∎

A.4. Proof of Lemma 3.2.

Proof. Using Lemma 3.1 (iii), we have

$$
\begin{aligned}
&\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{z}, \mathbf{h}', \mathbf{u}', \xi, \zeta, \gamma) - \nabla_{\mathbf{z}}\mathcal{L}(\mathbf{z}, \mathbf{h}, \mathbf{u}, \xi, \zeta, \gamma) \\
&= \begin{bmatrix} \gamma\Delta_1\mathbf{w} - (\Psi(\mathbf{h}') - \Psi(\mathbf{h}))^\top \xi - \gamma\Delta_3 \\ \frac{2}{T}\sum_{t=1}^T \Delta_{2,t}\mathbf{a} - \frac{2}{T}\sum_{t=1}^T (\Phi(h_t') - \Phi(h_t))^\top y_t \end{bmatrix},
\end{aligned} \tag{A.5}
$$

where $\Delta_1 = \Psi(\mathbf{h}')^\top\Psi(\mathbf{h}') - \Psi(\mathbf{h})^\top\Psi(\mathbf{h})$ and $\Delta_{2,t} = \Phi(h_t')^\top\Phi(h_t') - \Phi(h_t)^\top\Phi(h_t)$ and $\Delta_3 = \Psi(\mathbf{h}')\mathbf{u}' - \Psi(\mathbf{h})\mathbf{u}$. It is easy to see that

$$
\begin{aligned}
\|\Delta_1\| &= \|\Psi(\mathbf{h}')^\top\Psi(\mathbf{h}') - \Psi(\mathbf{h}')^\top\Psi(\mathbf{h}) + \Psi(\mathbf{h}')^\top\Psi(\mathbf{h}) - \Psi(\mathbf{h})^\top\Psi(\mathbf{h})\| \\
&\le (\|\Psi(\mathbf{h}')\| + \|\Psi(\mathbf{h})\|)\, \|\Psi(\mathbf{h}') - \Psi(\mathbf{h})\|.
\end{aligned} \tag{A.6}
$$

Similarly, we have

$$
\|\Delta_{2,t}\| \le (\|\Phi(h_t')\| + \|\Phi(h_t)\|)\, \|\Phi(h_t') - \Phi(h_t)\|, \ \forall t \in [T], \tag{A.7}
$$

$$
\|\Delta_3\| \le \|\Psi(\mathbf{h}')\|\|\mathbf{u}' - \mathbf{u}\| + \|\mathbf{u}\|\|\Psi(\mathbf{h}') - \Psi(\mathbf{h})\|. \tag{A.8}
$$

Since $\mathbf{s}, \mathbf{s}' \in \Omega_{\mathcal{L}}(\hat{\Gamma})$, we know that

$$
\ell(\mathbf{s}) + P(\mathbf{s}) + \frac{\gamma}{2}\left\|\mathbf{u} - \Psi(\mathbf{h})\mathbf{w} + \frac{\xi}{\gamma}\right\|^2 + \frac{\gamma}{2}\left\|\mathbf{h} - (\mathbf{u})_+ + \frac{\zeta}{\gamma}\right\|^2 \le \delta.
$$

This, together with the expressions of $\ell(\mathbf{s})$ in (2.6) and $P(\mathbf{s})$ in (2.5), yields

$$
\|W\|_F \le \sqrt{\frac{\delta}{\lambda_2}}, \ \|\mathbf{a}\| \le \sqrt{\frac{\delta}{\min\{\lambda_1, \lambda_5\}}}, \ \|\mathbf{w}\| \le \sqrt{\frac{\delta}{\min\{\lambda_2, \lambda_3, \lambda_4\}}}, \ \|\mathbf{u}\| \le \sqrt{\frac{\delta}{\lambda_6}}. \tag{A.9}
$$

Moreover, since $\|\mathbf{h}\| - \|(\mathbf{u})_+ - \frac{\zeta}{\gamma}\| \le \|\mathbf{h} - (\mathbf{u})_+ + \frac{\zeta}{\gamma}\| \le \sqrt{\frac{2\delta}{\gamma}}$, we find

$$
\|\mathbf{h}\| \le \delta_0. \tag{A.10}
$$

Using (2.3), we can easily obtain that

$$\text{(A.11)} \quad \|\Psi(\mathbf{h})-\Psi(\mathbf{h}')\| \le \sqrt{r}\|\mathbf{h}'-\mathbf{h}\|, \quad \|\Phi(h_t')-\Phi(h_t)\| \le \sqrt{m}\|h_t'-h_t\|,$$

$$\text{(A.12)} \quad \|\Psi(\mathbf{h})\| = \sqrt{r(\|\mathbf{h}\|^2+\|X\|^2+T)}, \quad \|\Phi(h_t)\| = \sqrt{m(\|h_t\|^2+1)}.$$

Using the facts that for any $\iota_1, \iota_1, \ldots, \iota_j \in \mathbb{R}$, any $g_1, g_2, \ldots, g_j \in \mathbb{R}^{n_r}$, and any matrices $B_1, B_2, \ldots, B_j \in \mathbb{R}^{n_c \times n_r}$, $\|B_1\| \le \|B_1\|_F$, and

$$\text{(A.13)} \quad \|\sum_{i=1}^{(j)} \iota_j B_j g_j\| \le \sum_{i=1}^{j} |\iota_j| \|B_j\| \|g_j\|, \quad \sum_{i=1}^{j} \|\iota_i g_i\| \le \max_{1\le i\le j}\{|\iota_i|\}\sqrt{j}\|(g_1;\ldots;g_j)\|,$$

taking the norm of both sides of (A.5), and employing (A.6)-(A.12), we can get (3.2) with the expression of $L_1(\xi,\zeta,\gamma,\hat{r})$ in (3.4) as desired.

Using Lemma 3.1 (iv), we have by direct computation

$$\begin{aligned}
&\nabla_{\mathbf{h}}\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u}',\xi,\zeta,\gamma) - \nabla_{\mathbf{h}}\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u},\xi,\zeta,\gamma)\\
&= \gamma W^T \sum_{t=1}^{T-1}(u_{t+1}-u_{t+1}') + \gamma\sum_{t=1}^{T}((u_t)_+ - (u_t')_+).
\end{aligned}$$

Taking the norm of both sides of the above system of equations, employing (A.9), (A.13), and the facts $\|(u_t)_+ - (u_t')_+\| \le \|u_t' - u_t\|$ for each t, we can get (3.3) with $L_2(\xi,\zeta,\gamma,\hat{r})$ in the form of (3.4) as desired. ∎

A.5. Proof of Lemma 4.1.

Proof. By (3.14), (3.15) and (3.16), we know that for any $j \in \mathbb{N}$:

$$\text{(A.14)} \quad \mathcal{L}\big(\mathbf{s}^{(j)},\xi,\zeta,\gamma\big) \le \mathcal{L}\big(\mathbf{s}_{\mathbf{h}}^{(j)},\xi,\zeta,\gamma\big) \le \mathcal{L}\big(\mathbf{s}_{\mathbf{z}}^{(j)},\xi,\zeta,\gamma\big) \le \mathcal{L}\big(\mathbf{s}^{(j-1)},\xi,\zeta,\gamma\big).$$

By the definition of Γ in Algorithm 3.2 and (A.14), we can deduce that

$$\text{(A.15)} \quad \mathcal{L}\big(\mathbf{s}^{(j)},\xi,\zeta,\gamma\big) \le \Gamma, \quad \forall j \in \mathbb{N}.$$

By the definition of $\Omega_{\mathcal{L}}(\Gamma)$ and Lemma 3.1 (ii), the proof is completed. ∎

A.6. Proof of Lemma 4.2.

Proof. It is clear that $\Omega_{\mathcal{L}}(\Gamma)$ is compact by Lemma 3.1 (ii). For the smooth part g in $\mathcal{L}$, its gradient for those $\mathbf{s} \in \Omega_{\mathcal{L}}(\Gamma)$ is upper bounded. Now, let us turn to consider the nonsmooth part q in $\mathcal{L}$. Let $\mathbf{s} = (\mathbf{z};\mathbf{h};\mathbf{u})$ and $\mathbf{s}' = (\mathbf{z}';\mathbf{h}';\mathbf{u}')$ be any two points in $\Omega_{\mathcal{L}}(\Gamma)$. We have

$$\begin{aligned}
&\big|q(\mathbf{s}',\zeta,\gamma) - q(\mathbf{s},\zeta,\gamma)\big|\\
&\le \tfrac{\gamma}{2}\Big|\big\|\mathbf{h}'-(\mathbf{u}')_+ + \tfrac{\zeta}{\gamma}\big\|^2 - \big\|\mathbf{h}-(\mathbf{u})_+ + \tfrac{\zeta}{\gamma}\big\|^2\Big|\\
&\le \tfrac{\gamma}{2}\big\|\mathbf{h}'-(\mathbf{u}')_+ - (\mathbf{h}-(\mathbf{u})_+)\big\|\big\|\mathbf{h}'-(\mathbf{u}')_+ + \mathbf{h}-(\mathbf{u})_+ + 2\tfrac{\zeta}{\gamma}\big\|\\
&\le \left(2\gamma\max_{\mathbf{s}\in\Omega_{\mathcal{L}}(\Gamma)}\{\|\mathbf{h}\|_\infty + \|\mathbf{u}\|_\infty\} + \|\zeta\|\right)(\|\mathbf{h}'-\mathbf{h}\| + \|\mathbf{u}'-\mathbf{u}\|).
\end{aligned}$$

Up to now, we have proved the Lipschitz continuity of g and q on $\Omega_{\mathcal{L}}(\Gamma)$, which implies that $\mathcal{L}$ is Lipschitz continuous on $\Omega_{\mathcal{L}}(\Gamma)$.

The above result, together with the piecewise smoothness of function $\mathcal{L}$, yields that $\mathcal{L}$ is directionally differentiable on $\Omega_{\mathcal{L}}(\Gamma)$ by [21]. ∎

A.7. Proof of Lemma 4.5.

Proof. By (4.1), the directional derivative of $\mathcal{L}$ at $\bar{\mathbf{s}}$ along $d \in \mathbb{R}^{N_{\mathbf{w}}+N_{\mathbf{a}}+2rT}$ refers to $\mathcal{L}'(\bar{\mathbf{s}}, \xi, \zeta, \gamma; d) = g'(\bar{\mathbf{s}}, \xi, \gamma; d) + q'(\bar{\mathbf{s}}, \zeta, \gamma; d)$. It is clear that

$$g'(\bar{\mathbf{s}}, \xi, \gamma; d) = \langle \nabla_{\mathbf{z}} g(\bar{\mathbf{s}}, \xi, \gamma), d_{\mathbf{z}} \rangle + \langle \nabla_{\mathbf{h}} g(\bar{\mathbf{s}}, \xi, \gamma), d_{\mathbf{h}} \rangle + \langle \nabla_{\mathbf{u}} g(\bar{\mathbf{s}}, \xi, \gamma), d_{\mathbf{u}} \rangle. \tag{A.16}$$

It remains to consider the directional derivative of nonsmooth part q. The function q can be separated into rT one dimensional functions with the same structure, i.e.,

$$\phi(\bar{h}, \bar{u}) = (\bar{h} - (\bar{u})_+ + \nu_1)^2 - \nu_1^2,$$

where $\bar{h}, \bar{u} \in \mathbb{R}$ are variables and $\nu_1 \in \mathbb{R}$ is a constant. The directional derivative of ϕ along the direction $(\bar{d}_1; \bar{d}_2) \in \mathbb{R}^2$ can be represented as the sum of the directional derivatives of ϕ along $(\bar{d}_1; 0)$ and $(0; \bar{d}_2)$ by the definition of directional derivative, i.e.,

$$\begin{aligned}
\phi'\big(\bar{h}, \bar{u}; (\bar{d}_1, \bar{d}_2)\big) &= \lim_{\lambda \downarrow 0} \frac{\Big(\bar{h} + \lambda \bar{d}_1 - \big(\bar{u} + \lambda \bar{d}_2\big)_+ + \nu_1\Big)^2 - \Big(\bar{h} - \big(\bar{u}\big)_+ + \nu_1\Big)^2}{\lambda} \\
&= \phi'\big(\bar{h}, \bar{u}; (\bar{d}_1, 0)\big) + \phi'\big(\bar{h}, \bar{u}; (0, \bar{d}_2)\big) - \lim_{\lambda \downarrow 0} \frac{2\lambda \bar{d}_1\big((u + \lambda \bar{d}_2)_+ - (u)_+\big)}{\lambda}
\end{aligned}$$

where

$$\begin{aligned}
\phi'\big(\bar{h}, \bar{u}; (\bar{d}_1, 0)\big) &= \lim_{\lambda \downarrow 0} \frac{\big(\bar{h} + \lambda \bar{d}_1 - (\bar{u})_+ + \nu_1\big)^2 - \big(\bar{h} - (\bar{u})_+ + \nu_1\big)^2}{\lambda} \\
&= \lim_{\lambda \downarrow 0} \frac{(\bar{h} + \lambda \bar{d}_1 + \nu_1)^2 - (\bar{h} + \nu_1)^2 - 2(\lambda \bar{d}_1)(\bar{u})_+}{\lambda},
\end{aligned}$$

$$\begin{aligned}
\phi'\big(\bar{h}, \bar{u}; (0, d_2)\big) &= \lim_{\lambda \downarrow 0} \frac{\big(\bar{h} + \nu_1 - (\bar{u} + \lambda \bar{d}_2)_+\big)^2 - \big(\bar{h} + \nu_1 - (\bar{u})_+\big)^2}{\lambda} \\
&= \lim_{\lambda \downarrow 0} \frac{(\bar{u} + \lambda \bar{d}_2)_+^2 - (\bar{u})_+^2 - 2(\bar{h} + \nu_1)\big((\bar{u} + \lambda \bar{d}_2)_+ - (\bar{u})_+\big)}{\lambda},
\end{aligned}$$

and $\lim_{\lambda \downarrow 0} \frac{2\lambda \bar{d}_1((u+\lambda \bar{d}_2)_+ - (u)_+)}{\lambda} = 0$. By setting $\bar{h} = \bar{\mathbf{h}}_i$, $\bar{u} = \bar{\mathbf{u}}_i$, $\bar{d}_1 = (d_{\mathbf{h}})_i$, $\bar{d}_2 = (d_{\mathbf{u}})_i$, $\nu_1 = \frac{\zeta_i}{\gamma}$, we have

$$\begin{aligned}
q'(\bar{\mathbf{s}}, \zeta, \gamma; \bar{d}) &= \frac{\gamma}{2} \sum_{i=1}^{rT} \phi'\big(\bar{\mathbf{h}}_i, \bar{\mathbf{u}}_i; ((d_{\mathbf{h}})_i, (d_{\mathbf{u}})_i)\big) \\
&= \frac{\gamma}{2} \sum_{i=1}^{rT} \phi'\big(\bar{\mathbf{h}}_i, \bar{\mathbf{u}}_i; ((d_{\mathbf{h}})_i, 0)\big) + \phi_i'\big(\bar{\mathbf{h}}_i, \bar{\mathbf{u}}_i; (0, (d_{\mathbf{u}})_i)\big) \\
&= q'\big(\bar{\mathbf{s}}, \zeta, \gamma; (0, d_{\mathbf{h}}, 0)\big) + q'\big(\bar{\mathbf{s}}, \zeta, \gamma; (0, 0, d_{\mathbf{u}})\big).
\end{aligned}$$

This, along with (A.16), yields that

$$\begin{aligned}
&\mathcal{L}'(\bar{\mathbf{s}}, \xi, \zeta, \gamma; d) \\
= &\langle \nabla_{\mathbf{z}} g(\bar{\mathbf{s}}, \xi, \gamma), d_{\mathbf{z}} \rangle + \langle \nabla_{\mathbf{h}} g(\bar{\mathbf{s}}, \xi, \gamma), d_{\mathbf{h}} \rangle + \langle \nabla_{\mathbf{u}} g(\bar{\mathbf{s}}, \xi, \gamma), d_{\mathbf{u}} \rangle \\
&+ q'\big(\bar{\mathbf{s}}, \zeta, \gamma; (0, d_{\mathbf{h}}, 0)\big) + q'\big(\bar{\mathbf{s}}, \zeta, \gamma; (0, 0, d_{\mathbf{u}})\big) \\
= &\mathcal{L}'(\bar{\mathbf{s}}, \xi, \zeta, \gamma; (d_{\mathbf{z}}, 0, 0)) + \mathcal{L}'(\bar{\mathbf{s}}, \xi, \zeta, \gamma; (0, d_{\mathbf{h}}, 0)) + \mathcal{L}'(\bar{\mathbf{s}}, \xi, \zeta, \gamma; (0, 0, d_{\mathbf{u}})).
\end{aligned}$$

Hence Lemma 4.5 holds. ∎

Appendix B. Parameters for numerical experiments in section 5.4. The final selected learning rates for GDs and SGDs, as well as the clipping norm for GDC, are listed in Table 4.

		He	$\mathcal{N}(0,10^{-3})$	$\mathcal{N}(0,10^{-1})$	Glorot	LeCun
GD	**Synthetic dataset** ($T=10$)	1e-4	1e-3	1e-4	1	1
	Volatility of S&P index	1e-4	0.01	0.01	0.01	0.01
	Synthetic dataset ($T=500$)	0.01	0.01	0.01	1e-3	1e-3
GDC	**Synthetic dataset** ($T=10$)	1 (6)	1e-4 (1)	1e-4 (1)	1 (6)	1 (6)
	Volatility of S&P index	1e-4 (3)	0.01 (1)	0.1 (1)	0.1 (4)	0.1 (1)
	Synthetic dataset ($T=500$)	1e-4 (1)	0.01 (1)	0.01 (4)	0.01 (1.5)	0.1 (0.5)
GDNM	**Synthetic dataset** ($T=10$)	1e-3	1e-4	1e-4	1e-4	0.1
	Volatility of S&P index	1e-4	0.01	0.01	0.01	0.01
	Synthetic dataset ($T=500$)	0.01	0.01	0.01	0.01	0.01
SGD	**Synthetic dataset** ($T=10$)	0.1	0.1	0.1	0.1	0.1
	Volatility of S&P index	0.01	0.01	0.01	0.01	0.01
	Synthetic dataset ($T=500$)	0.01	1e-3	0.01	0.01	0.01
Adam	**Synthetic dataset** ($T=10$)	0.1	0.01	0.01	0.01	0.01
	Volatility of S&P index	0.01	0.01	0.01	0.01	0.01
	Synthetic dataset ($T=500$)	0.01	0.01	0.01	0.01	0.01

Table 4: The learning rates for GDs and SGDs, and the clipping norm value for GDC (the second number in each cell for parameters) under different initialization strategies.

REFERENCES

[1] P. T. Arnerić Josip and A. Zdravka, *Garch based artificial neural networks in forecasting conditional variance of stock returns*, Croat. Oper. Res. Rev., 5 (2014), pp. 329–343, https://doi.org/10.17535/crorr.2014.0017.

[2] Y. Bengio, *Learning deep architectures for AI*, Found. Trends Mach. Learn., (2009), pp. 136, https://doi.org/10.1561/2200000006.

[3] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu, *Advances in optimizing recurrent networks*, in IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 8624–8628.

[4] Y. Bengio, P. Simard, and P. Frasconi, *Learning long-term dependencies with gradient descent is difficult*, IEEE Trans. Neural Netw. Learn. Syst., 5 (1994), pp. 157–166, https://doi.org/10.1109/72.279181.

[5] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Nashua, NH, 2nd ed., 1999.

[6] A. Bucci, *Realized volatility forecasting with neural networks*, J. Financ. Econ., 18 (2020), pp. 502–531, https://doi.org/10.1093/jjfinec/nbaa008.

[7] M. Carreira-Perpinan and W. Wang, *Distributed optimization of deeply nested systems*, in the 17th International Conference on Artificial Intelligence and Statistics, Reykjavic, Iceland, 2014, pp. 10–19.

[8] K. K. Chandriah and R. V. Naraganahalli, *RNN/LSTM with modified Adam optimizer in deep learning approach for automobile spare parts demand forecasting*, Multimed. Tools. Appl., 80 (2021), pp. 26145–26159, https://doi.org/10.1007/s11042-021-10913-0.

[9] X. Chen, L. Guo, Z. Lu, and J. J. Ye, *An augmented Lagrangian method for non-Lipschitz nonconvex programming*, SIAM J. Numer. Anal., 55 (2017), pp. 168–193, https://doi.org/10.1137/15M1052834.

[10] F. Chollet et al., *Keras*. https://keras.io, 2015.

[11] F. H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, Philadelphia, PA, 1990.

[12] Y. Cui, Z. He, and J.-S. Pang, *Multicomposite nonconvex optimization for training deep neural networks*, SIAM J. Optim., 30 (2020), pp. 1693–1723, https://doi.org/10.1137/

18M1231559.

[13] J. L. Elman, *Finding structure in time*, Cogn. Sci., 14 (1990), pp. 179–211, https://doi.org/10.1016/0364-0213(90)90002-E.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT Press, Cambridge, MA, 2016.

[15] A. Graves, A.-r. Mohamed, and G. Hinton, *Speech recognition with deep recurrent neural networks*, in the 38th IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, IEEE, pp. 6645–6649.

[16] N. Hallak and M. Teboulle, *An adaptive Lagrangian-based scheme for nonconvex composite optimization*, Math. Oper. Res., (2023), p. 136, https://doi.org/10.1287/moor.2022.1342.

[17] A. Y. Kruger, *On Fréchet subdifferentials*, J. Math. Sci., 116 (2003), pp. 3325–3358, https://doi.org/10.1023/A:1023673105317.

[18] Q. V. Le, N. Jaitly, and G. E. Hinton, *A simple way to initialize recurrent networks of rectified linear units*, preprint, https://arxiv.org/abs/1504.00941, 2015.

[19] W. Liu, X. Liu, and X. Chen, *Linearly constrained nonsmooth optimization for training autoencoders*, SIAM J. Optim., 32 (2022), pp. 1931–1957, https://doi.org/10.1137/21M1408713.

[20] W. Liu, X. Liu, and X. Chen, *An inexact augmented Lagrangian algorithm for training leaky ReLU neural network with group sparsity*, J. Mach. Learn. Res., 24 (2023), pp. 1–43, http://jmlr.org/papers/v24/22-0491.html.

[21] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972, https://doi.org/10.1137/0315061.

[22] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, *Recurrent neural network based language model*, in the 11th Annual Conference of the International Speech Communication Association, Chiba, 2010, ISCA, pp. 1045–1048.

[23] S. Mirmirani and H. C. Li, *A comparison of VAR and neural networks with genetic algorithm in forecasting price of oil*, in Applications of Artificial Intelligence in Finance and Economics, Emerald Publishing Limited, Leeds, 2004, pp. 203–223.

[24] R. Pascanu, T. Mikolov, and Y. Bengio, *On the difficulty of training recurrent neural networks*, in the 30th International Conference on Machine Learning, Atlanta GA, 2013, IMLS.

[25] D. Peng and X. Chen, *Computation of second-order directional stationary points for group sparse optimization*, Optim. Methods Softw., 35 (2020), pp. 348–376, https://doi.org/10.1080/10556788.2019.1684492.

[26] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 2009.

[27] H. Sak, A. Senior, and F. Beaufays, *Long short-term memory recurrent neural network architectures for large scale acoustic modeling*, in the 15th Annual Conference of the International Speech Communication Association, Singapore, 2014, ISCA, pp. 338–342.

[28] M. Sundermeyer, R. Schlüter, and H. Ney, *LSTM neural networks for language modeling*, in the 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, 2012, ISCA.

[29] J. J. Ye, *Multiplier rules under mixed assumptions of differentiability and Lipschitz continuity*, SIAM J. Control Optim., 39 (2000), pp. 1441–1460, https://doi.org/10.1137/S0363012999358476.

[30] X. Zhang, N. Gu, and H. Ye, *Multi-GPU based recurrent neural network language model training*, in the International Conference of Pioneering Computer Scientists, Engineers and Educators, Springer, 2016, pp. 484–493.

[31] Z. Zhang and M. Brand, *Convergent block coordinate descent for training Tikhonov regularized deep neural networks*, in the 31st International Conference on Neural Information Processing Systems, NY, 2017, pp. 1719–1728.

[32] K. He, X. Zhang, S. Ren, and J. Sun. *Delving deep into rectifiers: surpassing human-level performance on imagenet classification,* in the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1026–1034.

[33] X. Glorot and Y. Bengio. *Understanding the difficulty of training deep feedforward neural networks,* in the 13th International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 2010, pp. 249–256.

[34] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, *Self-normalizing neural networks,* in the 31st International Conference on Neural Information Processing Systems, California, USA, 2017, pp. 972–981.

[35] J. Bergstra and Y. Bengio, *Random search for hyper-parameter optimization*, J. Mach. Learn. Res., 13 (2012), pp. 281–305.