AN AUGMENTED LAGRANGIAN METHOD FOR TRAINING RECURRENT NEURAL NETWORKS WITH SAMPLE AVERAGE APPROXIMATION

YUE WANG

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Applied Mathematics

An Augmented Lagrangian Method for Training Recurrent Neural Networks with Sample Average Approximation

Yue Wang

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

December, 2024

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

Yue Wang (Name of student)

Dedicated to my family.

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 forecasting. The training process of RNNs with nonsmooth activation functions is formulated as an unconstrained optimization problem. The objective function of the problem is nonconvex, nonsmooth and has a highly composite structure, which poses significant challenges. State-of-the-art optimization methods, such as gradient descent-based methods (GDs) and stochastic gradient descent-based methods (SGDs), often lack a well-defined generalized gradient of the nonsmooth objective function and do not provide rigorous convergence analysis. In the thesis, we propose an augmented Lagrangian method (ALM) to solve the nonconvex, nonsmooth and highly composite optimization problem and provide a rigorous convergence analysis. Moreover, the aforementioned unconstrained problem arising from the RNN training process is typically a sample average approximation (SAA) of the original optimization problem whose objective function is formulated with an expectation. Therefore, it is necessary to prove that any accumulation point of minimizers and stationary points of the SAA problems is almost surely a minimizer and a stationary point of the original problem, respectively. The thesis is primarily divided into two parts.

In the first part of the thesis, we focus on the method to solve the nonconvex, nonsmooth and highly composite optimization problem. Specifically, we first re-

formulate the aforementioned unconstrained optimization problem equivalently 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. We prove the existence of global solutions and Karush-Kuhn-Tucker (KKT) points of the constrained problem. Moreover, we propose an 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 satisfied in finite steps. Moreover, we demonstrate that any accumulation point of the sequences generated by the BCD method is 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 state-of-the-art algorithms, numerical results demonstrate the efficiency and effectiveness of the ALM for training RNNs on synthetic datasets, the MNIST handwritten digit recognition task, the TIMIT audio denoising task, and the volatility of S&P index forecasting task.

In the second part of the thesis, we investigate the convergence of minimizers and stationary points of the SAA problems. Specifically, we first establish the existence of optimal solutions for both the original problem and the SAA problems. Next, we prove that any accumulation point of the sequences of minimizers and stationary points of the SAA problems is, respectively, a minimizer and a stationary point of the original problem with probability one, as the sample size goes to infinity. We also derive the uniform exponential rates of convergence of the objective functions of the SAA problems to those of the original problem.

This thesis contains research results of the following paper which has been published during the period of my Ph.D study at the Department of Applied Mathematics, The Hong Kong Polytechnic University:

- Y. WANG, C. ZHANG and X. CHEN, *An Augmented Lagrangian Method for Training Recurrent Neural Networks*. SIAM J. Sci. Comput., 47(1):C22-C51, 2025.

Acknowledgements

The journey of completing a PhD is one filled with challenges, growth, and the invaluable support of many individuals. I am deeply grateful to all those who have contributed to this academic and personal journey.

First and foremost, I wish to express my heartfelt gratitude to my chief supervisor, Prof. Xiaojun Chen, for her exceptional guidance, thoughtful mentorship, and constant encouragement throughout my PhD. Her intellectual insight and unwavering support have been instrumental not only in the completion of this research but also in shaping my academic career.

I would also like to sincerely thank my collaborator and senior, Prof. Chao Zhang, whose guidance and encouragement have been pivotal throughout my research. As both a mentor and a collaborator, she has provided me with invaluable advice, thoughtful discussions, and constant support. Her dedication to research and her willingness to share her experience have profoundly influenced my academic and personal growth.

A special thanks goes to my former co-supervisors, Prof. Heung Wong and Dr. Xin Guo, for their invaluable guidance and support during the early stages of my research. Their mentorship and insights laid the foundation for much of the progress I have made in this work.

I am deeply grateful to Prof. Xin Liu for providing me with the opportunity to visit the Academy of Mathematics and Systems Science of Chinese Academy of

Sciences. This visit was an eye-opening experience that allowed me to learn a great deal and significantly expand my academic knowledge. I truly appreciate his support and encouragement.

I extend my special thanks to my academic brothers and sisters: Prof. Chao Zhang, Prof. Wei Bian, Prof. Hailin Sun, Prof. Yanfang Zhang, Prof. Zaikun Zhang, Prof. Xiao Wang, Prof. Yang Zhou, Prof. Bo Wen, Prof. Jie Jiang, Prof. Lei Yang, Dr. Hong Wang, Dr. Chao Li, Dr. Fang He, Dr. Jianfeng Luo, Dr. Shisen Liu, Dr. Wei Liu, Dr. Xiaozhou Wang, Dr. Xiaoxia Liu, Dr. Fan Wu, Dr. Fang Fei, Dr. You Zhao, Dr. Yong Zhao, Dr. Zhihua Zhao, Dr. Lin Chen, Dr. Kaixin Gao, Dr. Ming Huang, Dr. Lei Wang, Mr. Zicheng Qiu, Mr. Shijie Yu, Mr. Yifan He, Mr. Xiao Zha, Mr. Zhouxing, Luo, Mr. Guang Wang, Mr. Wentao Ma, Miss Yixuan Zhang, Miss Xin Qu, Miss Lingzi Jin and Miss Yiyang Li. Especially, I would like to express my sincere gratitude to Dr. Xingbang Cui and Dr. Chao Li for carefully reviewing my thesis and providing valuable feedback. Their efforts and suggestions have greatly improved the quality of the work. I wish to thank my colleagues and friends: Miss Qian Li, Dr. Shu Ma, Miss Hui Yan, Dr. Yangzi Zheng, Mr. Chiyu Ma, Miss Xinyu Fan. Their camaraderie, thoughtful discussions, and collaboration have made this journey both enjoyable and intellectually stimulating.

In particular, I am deeply thankful to Dr. Yuqia Wu, not only my beloved partner in life, but also a true companion in my academic journey. His continuous companionship, generous help, and heartfelt encouragement throughout my Ph.D. journey meant the world to me. His presence during both challenging and joyful times gave me strength, comfort, and confidence, making this period of my life more colorful, meaningful, and unforgettable.

On a personal note, I am profoundly grateful to my family for their unconditional love and unwavering support. My parents and grandparents have been a constant source of strength, instilling in me the values of perseverance and dedication. This

accomplishment reflects their faith in me, their guidance, and the foundation they have built for my success.

Finally, I dedicate this thesis to all the friends and mentors who have inspired me to pursue this path. Thank you for believing in me.

Contents

List of Figures

List of Tables

List of Notation

$\mathbb{R}$	the set of real numbers
$\mathbb{R}^n$	the set of n-dimensional real vectors
$\mathbb{R}^{m\times n}$	the set of $m \times n$ real matrices
$\mathbb{R}_+$	the set of nonnegative real numbers
$\mathbb{R}_{++}$	the set of positive real numbers
$\mathbb{N}$	the set of natural numbers
$\mathbb{N}_+$	the set of positive integers
$[N]$	for a given $N \in \mathbb{N}_+$, $[N] := \{1, 2, \ldots, N\}$
$x^\top$	the transpose of a vector x
$\mathrm{diag}(x)$	the diagonal matrix of a given vector x, whose (i,i)-entry is the i-th component of x
$\boldsymbol{e}_l$	the vector of all ones in $\mathbb{R}^l$
A	a matrix in $\mathbb{R}^{m\times n}$, where $A_{\cdot j}$ represents the j-th column of A, $A_{i\cdot}$ is the i-th row of A, and a_{ij} denotes the (i,j)-entry of A
$A^\top$	the transpose of the matrix A
A^{-1}	the inverse of the nonsingular matrix A
I_n	identity matrix of order n
$\|x\|$	the ℓ_2-norm of a vector x

$\|x\|_\infty$	the infinity norm of a vector x
$\|A\|_F$	the Frobenius norm of the matrix $A \in \mathbb{R}^{m\times n}$, i.e., $\|A\|_F := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}$
$\mathrm{vec}(A)$	columnwise vectorization for the matrix A, i.e., $\mathrm{vec}(A) = (A_{.1}; A_{.2}; \ldots; A_{.l}) \in \mathbb{R}^{mn}$
$d(x, \mathcal{A})$	the distance from a vector x to the set $\mathcal{A}$, i.e., $d(x, \mathcal{A}) := \inf_{y\in\mathcal{A}} \|x - y\|$
$\mathbb{D}(\mathcal{A}, \mathcal{B})$	the deviation of the set $\mathcal{A}$ from the set $\mathcal{B}$, i.e., $\mathbb{D}(\mathcal{A}, \mathcal{B}) := \sup_{x\in\mathcal{A}} d(x, \mathcal{B})$
$\mathbb{H}(\mathcal{A}, \mathcal{B})$	the Hausdorff distance between the sets $\mathcal{A}$ and $\mathcal{B}$, i.e., $\mathbb{H}(\mathcal{A}, \mathcal{B}) := \max\{\mathbb{D}(\mathcal{A}, \mathcal{B}), \mathbb{D}(\mathcal{B}, \mathcal{A})\}$
$\min(x, y)$	the vector with its i-th component is $\min(x_i, y_i)$
$\nabla f(x)$	the gradient of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$
$J\mathcal{C}(x)$	the Jacobian matrix of a function $\mathcal{C} : \mathbb{R}^n \to \mathbb{R}^m$ $(m \geq 2)$ at $x \in \mathbb{R}^n$
$f'(x; d)$	the directional derivative of a function $f : \mathbb{R}^n \to \mathbb{R}$ along the direction $d \in \mathbb{R}^n$
$f^o(x; d)$	the generalized directional derivative of a function $f : \mathbb{R}^n \to \mathbb{R}$ along the direction $d \in \mathbb{R}^n$
$\hat{\partial} f(x)$	the Fréchet subdifferential of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$
$\partial f(x)$	the limiting subdifferential of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$
$\partial^c f(x)$	the Clarke subdifferential of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$

Chapter 1

Introduction

Recurrent neural networks (RNNs) are a class of neural networks to model sequential data, as they can caption temporal dynamic behavior. RNNs have been applied in a wide range of areas, such as speech recognition [22, 48], natural language processing [38, 54] and nonlinear time series forecasting [27, 39].

1.1 Problems of training recurrent neural networks (RNNs)

Once the model is selected, the training process can be described as follows: estimating the weight matrices and bias vectors in the RNNs such that the differences between the outputs from the model and the true values are minimized. The RNN training process can be represented by the following optimization problems.

1.1.1 Optimization problem in RNN training process for time series forecasting tasks

In this section, we present the optimization problem that arises from using RNNs to forecast time series. Given input data $x_t \in \mathbb{R}^n$ and output data $y_t \in \mathbb{R}^m$, $t = 1, \ldots, T$,

a widely used minimization problem is formulated as follows (see [21, p. 381]):

$$\min_{A,W,V,b,c} \frac{1}{T}\sum_{t=1}^{T}\left\|y_t - \left(A\sigma\Big(W\big(...\sigma(Vx_1+b)...\big)+Vx_t+b\Big)+c\right)\right\|^2, \tag{1.1.1}$$

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. Furthermore, $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonsmooth activation function that is applied component-wise for vectors. The training process by (1.1.1) 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. The composite structure in (1.1.1) represents the mathematical formulation of RNNs. It is worth mentioning that the RNNs share the same weight matrices and bias vectors at different time steps as shown in (1.1.1) [21, p. 374].

1.1.2 Optimization problem in RNN training process for general regression tasks

In section 1.1.1, the sample size of the input data x_t and the output data y_t at each time step t is equal to one, the scenario that primarily arises in time series forecasting tasks. In this section, we consider a more general case where the sample size at each time t equals N. This generalization leads to a new optimization problem that can represent general regression tasks for sequences, which is formulated as follows:

$$\min_{A,W,V,b,c} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|y_t^i - \left(A\sigma\Big(W\big(...\sigma\big(W\sigma(Vx_1^i+b)+Vx_2^i+b\big)...\big)+Vx_t^i+b\Big)+c\right)\right\|_2^2, \tag{1.1.2}$$

where x_t^i and y_t^i, $t \in [T]$, $i \in [N]$, denote the i-th sample of the known input and output data at time t, respectively.

Solving problems (1.1.1) and (1.1.2) are incredibly challenging due to the nonsmoothness and highly composite structures in their objective functions.

1.1.3 Optimization problem arising from the RNN training process with expectation

Denote $X^i = \left((x_1^i)^\top, (x_2^i)^\top, ..., (x_T^i)^\top\right)^\top$ and $Y^i = \left((y_1^i)^\top, (y_2^i)^\top, ..., (y_T^i)^\top\right)^\top$, $i = 1, 2, ..., N$, where $\{X^i\}$ and $\{Y^i\}$ can be viewed as independent and identically distributed (i.i.d.) samples of the random vectors $\mathbf{X} = \left((\mathbf{X}_1)^\top, \cdots, (\mathbf{X}_T)^\top\right)^\top$ and $\mathbf{Y} = \left((\mathbf{Y}_1)^\top, \cdots, (\mathbf{Y}_T)^\top\right)^\top$ respectively. Problem (1.1.2) can thus be identified as a sample average approximation (SAA) of the following problem [13]:

$$\min_{A,W,V,b,c} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|\mathbf{Y}_t - \left(A\sigma\Big(W\Big(...\sigma\big(W\sigma(V\mathbf{X}_1 + b) + V\mathbf{X}_2 + b\big)...\Big) + V\mathbf{X}_t + b\Big) + c\right)\right\|_2^2\right], \quad (1.1.3)$$

where the expectation is related to the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$.

Compared with (1.1.1) and (1.1.2), problem (1.1.3) involves an expectation whose closed-form expression is challenging to obtain [60]. The SAA method is commonly used to approximate the expectation [45, 51]. Nevertheless, an important theoretical problem needs to verify that any accumulation point of minimizers of SAA problem (1.1.2) is a minimizer of problem (1.1.3) as the sample size goes to infinity with probability one.

1.2 Literature reviews

1.2.1 Methods for solving nonconvex and nonsmooth optimization problems as training RNNs

The traditional backpropagation through time (BPTT) method with gradient descent methods (GDs) or stochastic gradient descent-based methods (SGDs) [12, 68] is commonly used to train RNNs. However, the "gradient" of the loss function associated with the weight matrices via the "chain rule" is calculated even if the "chain rule" does not hold. Furthermore, 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 proposed, including gradient clipping [38], gradient descent with Nesterov momentum [3], initialization with small values [42], the addition of sparse regularization [2], and so on. Because the essence of the above methods is to restrict the initial values of weight matrices or gradients, they are sensitive to the choice of initial values [32]. Moreover, GDs and SGDs for training RNNs lack rigorous convergence analysis.

The objective functions in (1.1.1) and (1.1.2) are nonsmooth nonconvex and have highly composite structures. In this paper, we equivalently reformulate (1.1.1) and (1.1.2) as constrained optimization problems with simple, smooth objective functions by employing auxiliary variables to represent the composition structures and treating these representations as constraints. Utilizing auxiliary variables to reformulate highly nonlinear composite structured problems as constrained optimization problems has been adopted for training Deep Neural Networks (DNNs) [11, 17, 36, 35, 69]. However, these algorithms for DNNs cannot be used for RNNs directly because of the difference between their architectures. In fact, RNNs share the same weight matrices and bias vectors across different layers, whereas DNNs have distinct weight matrices and bias vectors in different layers. In DNNs, the weight matrices and bias vectors can be updated layer by layer, whose gradients can be calculated separately. However, in RNNs, the weight matrices and bias vectors must 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.

After equivalently reformulating (1.1.1) and (1.1.2) as constrained optimization problems with the nonconvex smooth objective function and nonconvex nonsmooth constraints, we can solve them by the augmented Lagrangian method (ALM), which

is a type of classical methods to solve constrained optimization. The ALM extends the quadratic penalty method to reduce the possibility of ill conditioning by introducing explicit Lagrangian multiplier to the function [41, p. 514]. In recent years, with the rapid growth of data dimensions, the ALM has attracted increasing attention due to its ability to efficiently handle large-scale problems. The ALM was first proposed by Hestenes [25] and Powell [44] in 1969 to solve equality constrained problems. Subsequently, Bertsekas extended the method to address nonconvex problems [6, 5]. Furthermore, during that time, some works further expanded the ALM to handle convex problems with inequality constraints [46, 30].

Recently, several augmented Lagrangian-based methods have been proposed for nonconvex nonsmooth problems with composite structures. In [14], Chen et al. proposed an ALM for non-Lipschitz nonconvex programming, which requires the constraints to be smooth. Hallak and Teboulle in [23] 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. However, the assumption on the smoothness of constraints in [14, 23] is not satisfied for the optimization problem arising in training RNNs with nonsmooth activation functions considered in this paper. Several studies have also focused on extending the ALM to handle constrained problems with nonsmooth constraints. Specifically, Xu et al. [64] proposed a smoothing ALM to solve problems with nonsmooth and nonconvex constraints. However, it is tricky to adjust the smoothing parameters in practical computations. Furthermore, Kanzow et al. [28] applied the ALM for cardinality-constrained optimization problems. They equivalently reformulated the problem with non-continuous cardinality constraint into a continuous constrained problem with an orthogonality-type constraint, and then employed a safeguarded ALM to solve the reformulated constrained problem. The reformulation technique is tailored

for cardinality constraints. Therefore, this method cannot be extended to address our problems. Very recently, Xiao et al. [59] developed Lagrangian-based methods for nonsmooth constrained problems with expectation. Their work focused on linearized Lagrangian-based methods, where the primal variables are updated by a single proximal gradient step. They further embedded proximal SGD, proximal momentum SGD and proximal ADAM into Lagrangian-based methods. Moreover, they proved the global convergence of the method to the KKT points of the nonsmooth constrained problem in the sense of conservative Jacobians. In the thesis, we will investigate solving the constrained problem (1.1.1) and (1.1.2) with nonconvex and nonsmooth constraints by the Powell-Hestenes-Rockafellar (PHR) ALM, where the subproblems are required to be solved within controlled accuracy.

1.2.2 Properties of sample average approximation (SAA)

Stochastic programming focuses on those optimization problems involving uncertain parameters, which arise in almost all areas of science and engineering [50]. We focus on a class of problems whose objective function involves the expectation, where the closed-form expression of the expectation is unknown in general. A widely used method to solve the problem is SAA, which approximates the expectation by samples [49]. It is natural to ask whether the solutions of SAA problems converge to those of the original problem with expectation. That is, whether the solutions of SAA problems approximate the solutions of the original problem well. Many studies have already established a framework for analyzing such convergence.

For those problems whose objective functions are convex and lower semicontinuous (lsc), Rockafellar and Wets (1997) [47, Chapter 7] proved that the sequence of objective functions in SAA problems uniformly converges to the objective function of the original problem via epi-convergence. Thereby, the convergence of the

sequences of optimal values and optimal solutions of SAA problems has been established. Shapiro (2003) [49] extended the convergence results of the optimal value and the optimal solutions for SAA to problems with nonconvex objective functions and compact feasible sets.

However, optimal solutions for problems with nonconvex and nonsmooth objective functions are generally unavailable. Instead, numerical algorithms typically yield stationary points of such problems. Therefore, it is necessary to analyze the convergence of stationary points from the SAA problems to the original problem. In nonsmooth analysis, the stationary points have various definitions, such as the limiting (l-) stationary point and the Clarke (C-) stationary point. The detailed definitions of the above will be stated in section 2. When the set of stationary points of the SAA problem is bounded, it was established that any accumulation point of the sequence of weak C-stationary points of the SAA problem is a weak C-stationary point of the true problem with probability one (w.p.1) [61]. This result needs to consider the relationship between the Clarke subdifferential of expectation and the expectation of the Clarke subdifferential, which is also a challenging task.

1.3 Contribution of the thesis

The contributions of this thesis are summarized as follows.

- In the first part, we propose a method to solve problems (1.1.1) and (1.1.2), whose objective functions are nonconvex, nonsmooth, and highly composite.

  Specifically, we first reformulate (1.1.1) and (1.1.2) equivalently as constrained optimization problems with smooth objective functions. This is achieved by introducing auxiliary variables to represent the composition structures and treating these representations as constraints. We prove that the solution sets of the constrained problems with ℓ_2 regularization are nonempty and compact. Fur-

thermore, we establish that any feasible point of the constrained optimization problems satisfies the no nonzero abnormal multiplier constraint qualification (NNAMCQ), which immediately guarantees that any local minimizer of the constrained problems is a Karush-Kuhn-Tucker (KKT) point.

Moreover, we propose an augmented Lagrangian method (ALM) to solve the constrained optimization problems with ℓ_2-norm regularization, and design a block coordinate descent (BCD) method to address the subproblem of the ALM at each iteration. The solutions of the BCD subproblems are straightforward to be computed in closed-form. We prove that any accumulation point of the sequence generated by the BCD method is a directional stationary point of the subproblem. Furthermore, we establish that the stopping criterion of the BCD method for solving the subproblem of the k-th iteration of the ALM can be satisfied within $O\big(1/(\epsilon_{k-1})^2\big)$ finite steps for any $\epsilon_{k-1} > 0$. Additionally, we show that there exists at least an accumulation point of the sequence generated by the ALM, and any accumulation point of the sequence is a KKT point of the constrained problem with ℓ_2-norm regularization.

We compare the performance of the ALM with several state-of-the-art methods on synthetic datasets and real-world tasks, such as forecasting the volatility of the S&P index, denoising TIMIT audios and denoising MNIST images. The numerical results verify that our ALM outperforms other algorithms on both the training sets and the test sets.

- In the second part, we prove that any accumulation point of minimizers and stationary points of the SAA problems is a minimizer and a stationary point of the original problem, respectively, w.p.1 as the sample size goes to infinity.

  To be specific, we first explore the properties of the objective functions of (1.1.2) and (1.1.3). After that, the convergence of the optimal value and the optimal

solutions of SAA problems has been established through the uniform convergence of the objective functions. Moreover, we prove that any accumulation point of stationary points of SAA problem (1.1.2) is almost surely a stationary point of problem (1.1.2). Finally, numerical experiments are conducted to verify the theoretical results.

1.4 Organization of the thesis

The thesis is organized as follows.

- In Chapter 1, we introduce the optimization problem with the expectation from the RNN training process, and its corresponding SAA representation, which are main problems addressed in the thesis. We then summarize existing works that tackle these types of problems. In the last of the chapter, we summarize the main contributions of the thesis.

- In Chapter 2, we define basic notation and outline the primary knowledge required in the following chapters.

- In Chapter 3, we propose an ALM to solve the SAA problems (1.1.1) and (1.1.2) in section 3.1 and 3.2, respectively. The methods for two problems are similar, that is, we model the SAA form of RNNs training problem with the nonsmooth activation functions as constrained optimization problem with smooth nonconvex objective functions and piecewise smooth nonconvex constraints. Then, we propose an ALM and design an efficient BCD method to solve the subproblems of the ALM. 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.

- In Chapter 4, we analyze the convergence of the solutions and stationary points of the SAA problems. Numerical experiments are also provided to support these results.

- In Chapter 5, we conclude the main results of our work and give some possible further works.

Chapter 2

Basic Notation and Preliminaries

In this chapter, we introduce the basic notation and preliminary concepts that will be used throughout the thesis.

2.1 Basic notation

For column vectors $x_1, x_2, \ldots, x_l$, let

$$x := (x_1; x_2; \ldots; x_l) = (x_1^\top, x_2^\top, \ldots, x_l^\top)^\top.$$

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-wise vectorization for matrix D. 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 or equal to ν. Let $\mathbb{N}$ denote the set of natural numbers and $\mathbb{N}_+$ denote the set of positive integers. For a given $N \in \mathbb{N}_+$, we denote $[N] := \{1, 2, \ldots, N\}$. Let $\mathbb{R}_{++}$ represent the set of strictly positive real numbers. 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. For two functions $f : \mathcal{P} \to \mathcal{Q}$ and $g : \mathcal{Q} \to \mathcal{Z}$, the composite function $g(f(\cdot))$ is denoted by $(g \circ f)(\cdot)$.

For two sets $\mathcal{P}, \mathcal{Q} \subset \mathbb{R}^n$, $d(x, \mathcal{P}) := \inf_{x' \in \mathcal{P}} \|x - x'\|$ denotes the distance from any $x \in \mathbb{R}^n$ to $\mathcal{P}$, and $\mathbb{D}(\mathcal{P}, \mathcal{Q}) := \sup_{x \in \mathcal{P}} d(x, \mathcal{Q})$ denotes the deviation of $\mathcal{P}$ from the set $\mathcal{Q}$. Moreover, $\mathbb{H}(\mathcal{P}, \mathcal{Q}) := \max\{\mathbb{D}(\mathcal{P}, \mathcal{Q}), \mathbb{D}(\mathcal{Q}, \mathcal{P})\}$ represents the Hausdorff distance between the two sets.

2.2 Preliminaries

In this subsection, we introduce some fundamental concepts and definitions that will be used throughout the thesis. We first state the definition of the local Lipschitz continuity of a function.

In the following definitions in the section, let $\mathcal{U}$ represent an open subset of $\mathbb{R}^{n_1}$.

Definition 2.1. *(Local Lipschitz continuity) We say $f : \mathcal{U} \to \mathbb{R}$ is locally Lipschitz continuous on $\mathcal{U}$ if for any x_1 and x_2 in $\mathcal{U}$, $x_1 \neq x_2$, the following inequality is satisfied:*

$$|f(x_1) - f(x_2)| \leq L_f \|x_1 - x_2\|,$$

where $L_f > 0$ is the Lipschitz constant of f.

We then show the Lipschitz continuity for composite functions.

Lemma 2.1. *[18, Theorem 12.6] Let f_1 be Lipschitz continuous on a set $\mathcal{D}_1$ with Lipschitz constant a_1 and f_2 be Lipschitz continuous on $\mathcal{D}_2$ with Lipschitz constant a_2 such that $f_1(\mathcal{D}_1) \subset \mathcal{D}_2$. Then the composite function $f_2 \circ f_1$ is Lipschitz continuous on $\mathcal{D}_1$ with Lipschitz constant $a_1 a_2$.*

We now introduce relevant definitions related to random variables. Let $\xi : \Omega \to \Xi$ denote a random variable defined on the probability space $(\Omega, \mathcal{F}, P)$. The expectation of ξ is denoted by $\mathbb{E}[\xi]$.

Definition 2.2. *[50, p. 361] It is said that $\mathbb{E}[\xi]$ is well-defined if ξ is measurable and either $E[(\xi)_+] < +\infty$ or $E[(-\xi)_+] < +\infty$.*

Definition 2.3. *(Integrable) [50, p. 361] It is said that ξ is integrable if $\mathbb{E}[\xi]$ is well-defined and finite.*

2.3 Stationary points of the nonsmooth optimization

The problems (1.1.1), (1.1.2) and (1.1.3) are nonconvex and nonsmooth. Due to the nonsmoothness of the objective function in these problems, the gradients at nondifferential points are not well-defined. Therefore, the generalized gradients are proposed for nonsmooth functions [16]. In this situation, the generalized gradients at a point are not unique, where the collection of all generalized gradients at this point is named as the subdifferential of the point. Various definitions are proposed to represent subdifferentials for nonconvex functions. The following are the definitions for different subdifferentials.

Definition 2.4. *(Fréchet subdifferential and limiting subdifferential) [31, Definition 1.1][47, Definition 8.3, p. 301] Suppose that $f : \mathbb{R}^{n_1} \to \mathbb{R}$ is a lower semicontinuous (lsc) function defined on $\mathbb{R}^{n_1}$. The Fréchet (F-) subdifferential $\hat{\partial} f(\bar{x})$ and the limiting (l-) subdifferential $\partial f(\bar{x})$ of f at $\bar{x} \in \mathbb{R}^{n_1}$ are respectively defined as*

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

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

where $x^k \xrightarrow{f} \bar{x}$ denotes $x^k \to \bar{x}$ and $f(x^k) \to f(\bar{x})$. Furthermore, the l-subdifferential is also named the Mordukhovich subdifferential [40].

Definition 2.5. *(Clarke subdifferential) For an lsc function $f : \mathbb{R}^{n_1} \to \mathbb{R}$, the Clarke (C-) subdifferential of f at $\bar{x}$ is defined as follows:*

$$\partial^c f(\bar{x}) := conv\ \partial f(\bar{x}),$$

where conv represents the convex hull of the l-subdifferential of f at $\bar{x}$.

A point $\bar{x}$ is said to be a Fréchet stationary point of $\min f(x)$ if $0 \in \hat{\partial} f(\bar{x})$, $\bar{x}$ is said to be a limiting stationary point of $\min f(x)$ if $0 \in \partial f(\bar{x})$, and $\bar{x}$ is said to be a Clarke stationary point of $\min f(x)$ if $0 \in \partial^c f(\bar{x})$.

Then we give the definition of the directional (d-) stationary point, which corresponds to the directional derivative of functions, i.e.,

Definition 2.6. *(Directional derivative) [16, p. 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 [43, 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

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

The relationships among the above subdifferentials of f at $\bar{x}$ are as follows:

$$\hat{\partial} f(x) \subseteq \partial f(x) \subseteq \partial^c f(x), \tag{2.3.1}$$

that is, a F-stationary point is a l-stationary point, and a l-stationary point is a C-stationary point, but not vice versa [36, 34].

We now introduce a class of nonsmooth functions, known as Clarke regular functions, which possess several desirable properties related to the subdifferential.

Definition 2.7. *(Generalized directional derivative) [16, p. 10] The generalized directional derivative of f at x in the direction $d \in \mathbb{R}^{n_1}$ is defined as follows:*

$$f^o(x;d) := \limsup_{\substack{y \to x \\ \lambda \downarrow 0}} \frac{f(y+\lambda d) - f(y)}{\lambda}.$$

Chapter 3

An Augmented Lagrangian Method for Training Recurrent Neural Networks

This chapter outlines the core contributions of the thesis, introducing an augmented Lagrangian method (ALM) to solve the SAA problems (1.1.2) and (1.1.1) with non-convex and nonsmooth objective functions, which arise from training RNNs. We begin by introducing the method for problem (1.1.1) in section 3.1, which can be regarded as a special case of problem (1.1.2) when the sample size N is set to one. The practice significance of (1.1.1) lies in the application of time series forecasting using RNNs , where only a single sample point is available at each time step. For example, when forecasting the daily S&P index, we can only obtain one sample point per day. After presenting the details of the ALM and its corresponding convergence results, we extend the method to the more general problem (1.1.2) in section 3.2. The frameworks of the ALM in these two sections are the same, with only modifications to some expressions.

3.1 ALM for problem (1.1.1)

Recall problem (1.1.1):

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

In this section, we first equivalently reformulate problem (1.1.1) as a nonsmooth nonconvex constrained minimization problem with a simple smooth objective function, showing 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 in subsection 3.1.1. After that, we propose the ALM for the constrained problem with regularization, as well as the BCD method for the subproblems of the ALM in subsection 3.1.2. We establish the convergence results of the BCD method, and the ALM in subsection 3.1.3. Finally, we conduct numerical experiments on both the synthetic and real data in subsection 3.1.4, which demonstrate the effectiveness and efficiency of the ALM for the reformulated optimization problem.

3.1.1 Problem reformulation and optimality conditions

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

$$\sigma(u) = \max\{u, 0\} = (u)_+. \tag{3.1.1}$$

It is worth mentioning that the models, algorithms, and theoretical analysis can be generalized to the leaky ReLU and the ELU activation functions. Detailed analysis for the extensions will be given in section 3.1.3.

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 (1.1.1) as the following constrained optimization problem:

$$\begin{aligned}
\min_{\mathbf{s}} \quad & \frac{1}{T}\sum_{t=1}^{T} \|y_t - (Ah_t + c)\|^2 \\
\text{s.t.} \quad & u_t = Wh_{t-1} + Vx_t + b, \\
& h_0 = 0,\ h_t = (u_t)_+,\ t = 1, 2, ..., T.
\end{aligned} \tag{3.1.2}$$

Problems (1.1.1) and (3.1.2) are equivalent in the sense that if $(A^*, W^*, V^*, b^*, c^*)$ is a global solution of (1.1.1), then $\mathbf{s}^* = (\mathbf{z}^*; \mathbf{h}^*; \mathbf{u}^*)$ is a global solution of (3.1.2) where $\mathbf{z}^*$ is defined by $(A^*, W^*, V^*, b^*, c^*)$ and $\mathbf{h}^*, \mathbf{u}^*$ satisfy the constraints of (3.1.2) with W^*, V^*, b^*. Conversely, if $\mathbf{s}^*$ is a global solution of (3.1.2), then $\mathbf{z}^*$ is a global solution of (1.1.1).

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

$$\Phi(h_t) = \begin{bmatrix} h_t^\top \otimes I_m & I_m \end{bmatrix}, \quad \Psi(\mathbf{h}) = \begin{bmatrix} 0_r^\top \otimes I_r & x_1^\top \otimes I_r & I_r \\ h_1^\top \otimes I_r & x_2^\top \otimes I_r & I_r \\ \vdots & \vdots & \vdots \\ h_{T-1}^\top \otimes I_r & x_T^\top \otimes I_r & I_r \end{bmatrix}, \tag{3.1.3}$$

where $\otimes$ represents the Kronecker product, I_r and I_m are the identity matrices with dimensions $r \times r$ and $m \times m$ respectively, and 0_r is the zero vector with dimension r.

Thus, the objective function and constraints in problem (3.1.2) can be represented as

$$\begin{aligned} \ell(\mathbf{s}) &:= \frac{1}{T}\sum_{t=1}^{T} \|y_t - \Phi(h_t)\mathbf{a}\|^2, \\ \mathcal{C}_1(\mathbf{s}) := \mathbf{u} - \Psi(\mathbf{h})\mathbf{w} = 0, &\qquad \mathcal{C}_2(\mathbf{s}) := \mathbf{h} - (\mathbf{u})_+ = 0. \end{aligned} \tag{3.1.4}$$

To mitigate the overfitting, we further add a regularization term

$$p(\mathbf{s}) := \lambda_1\|A\|_F^2 + \lambda_2\|W\|_F^2 + \lambda_3\|V\|_F^2 + \lambda_4\|b\|^2 + \lambda_5\|c\|^2 + \lambda_6\|\mathbf{u}\|^2 \tag{3.1.5}$$

with $\lambda_i > 0, i = 1, 2, \ldots, 6$ in the objective of problem (3.1.2), and consider the following problem:

$$\begin{aligned} \min \quad & \mathcal{R}(\mathbf{s}) := \ell(\mathbf{s}) + p(\mathbf{s}) \\ \text{s.t.} \quad & \mathbf{s} \in \mathcal{F} := \{\mathbf{s} : \mathcal{C}_1(\mathbf{s}) = 0,\ \mathcal{C}_2(\mathbf{s}) = 0\}. \end{aligned} \tag{3.1.6}$$

Problem (3.1.2) and problem (3.1.6) 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 l-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 l-subdifferential of $\zeta^\top\mathcal{C}_2$ at $\mathbf{s}$ with respect to $\mathbf{u}$.

The following lemma shows that the NNAMCQ [67, Definition 4.2, p. 1451] holds at any feasible point $\mathbf{s} \in \mathcal{F}$.

Lemma 3.1. *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*

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

Proof. By direct computation,

$$J\mathcal{C}_1(\mathbf{s})^\top \xi + \partial\big(\zeta^\top \mathcal{C}_2(\mathbf{s})\big) = \begin{bmatrix} J_{\mathbf{z}}\mathcal{C}_1(\mathbf{s})^\top \xi \\ J_{\mathbf{h}}\mathcal{C}_1(\mathbf{s})^\top \xi + J_{\mathbf{h}}\mathcal{C}_2(\mathbf{s})^\top \zeta \\ J_{\mathbf{u}}\mathcal{C}_1(\mathbf{s})^\top \xi + \partial_{\mathbf{u}}\left(\zeta^\top \mathcal{C}_2(\mathbf{s})\right) \end{bmatrix}, \tag{3.1.7}$$

where

$$J_{\mathbf{h}}\mathcal{C}_1(\mathbf{s})^\top \xi + J_{\mathbf{h}}\mathcal{C}_2(\mathbf{s})^\top \zeta = \left[-W^\top \xi_2 + \zeta_1; ...; -W^\top \xi_T + \zeta_{T-1}; \zeta_T\right], \tag{3.1.8}$$

$$J_{\mathbf{u}}\mathcal{C}_1(\mathbf{s})^\top \xi + \partial_{\mathbf{u}}\left(\zeta^\top \mathcal{C}_2(\mathbf{s})\right) = \xi + \partial_{\mathbf{u}}(-\zeta^\top (\mathbf{u})_+). \tag{3.1.9}$$

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 $J_{\mathbf{h}}\mathcal{C}_1(\mathbf{s})^\top \xi + J_{\mathbf{h}}\mathcal{C}_2(\mathbf{s})^\top \zeta$. By $\zeta_T = 0$ and (3.1.9), we find $\xi_T = 0$. Substituting the results into (3.1.8) and (3.1.9) recursively and setting (3.1.8) and (3.1.9) equal to 0, we can derive that there exist no nonzero vectors ξ and ζ such that $0 \in J\mathcal{C}_1(\mathbf{s})^\top \xi + \partial\big(\zeta^\top \mathcal{C}_2(\mathbf{s})\big)$. □

Definition 3.1. *We say that $\mathbf{s} \in \mathcal{F}$ is a KKT point of problem* (3.1.2) *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 (3.1.6) *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 (3.1.2) and problem (3.1.6).

Theorem 3.1. *(i) If $\bar{\mathbf{s}}$ is a local solution of problem* (3.1.2)*, then $\bar{\mathbf{s}}$ is a KKT point of problem* (3.1.2)*. (ii) If $\bar{\mathbf{s}}$ is a local solution of problem* (3.1.6)*, then $\bar{\mathbf{s}}$ is a KKT point of problem* (3.1.6)*.*

Proof. Note that the objective functions of problem (3.1.2) and problem (3.1.6) 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 3.1, NNAMCQ holds at any $\bar{s} \in \mathcal{F}$. Therefore, the conclusions of this theorem hold according to [67, Remark 2 and Theorem 5.2]. □

Nonempty and compact solution set of (3.1.6)

Let $\mathcal{S}_1$ be the solution set of problem (3.1.6), and denote the level set

$$\mathcal{D}_{\mathcal{R}}(\rho) := \{\mathbf{s} \in \mathcal{F} : \mathcal{R}(\mathbf{s}) \leq \rho\} \tag{3.1.10}$$

with a nonnegative scalar ρ.

Lemma 3.2. *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 (3.1.6) is nonempty and compact.*

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

$$\begin{aligned} \|A\|_F^2 \leq \rho/\lambda_1, \|W\|_F^2 \leq \rho/\lambda_2, \|V\|_F^2 \leq \rho/\lambda_3, \\ \|b\|^2 \leq \rho/\lambda_4, \|c\|^2 \leq \rho/\lambda_5, \|\mathbf{u}\|^{\mathbf{2}} \leq \rho/\lambda_6, \end{aligned} \tag{3.1.11}$$

from $\mathcal{R}(\mathbf{s}) \leq \rho$, $\ell(\mathbf{s}) \geq 0$ and $p(\mathbf{s}) \geq 0$. 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 [47, 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 [7, Proposition A.8]. □

3.1.2 ALM with BCD method for (3.1.6)

To solve the regularized constrained problem (3.1.6), 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 1, in which the updating schemes for Lagrangian multipliers and penalty parameters are motivated by [14]. It is worth mentioning that in [14], the constraints are smooth. In problem (3.1.6), the constraints are nonsmooth nonconvex. For solving the subproblems in the ALM, we design the BCD method in Algorithm 2 and provide the closed-form expression for the update of each block in the BCD. Due to the nonsmooth nonconvex constraints in (3.1.6), the convergence analysis is complex, which will be given in subsection 3.1.3.

The augmented Lagrangian (AL) function of problem (3.1.6) is

$$\begin{aligned}\mathcal{L}(\mathbf{s},\xi,\zeta,\gamma) &:= \mathcal{R}(\mathbf{s}) + \langle \xi, \mathbf{u} - \Psi(\mathbf{h})\mathbf{w}\rangle + \langle \zeta, \mathbf{h} - (\mathbf{u})_+\rangle + \tfrac{\gamma}{2}\left\|\mathbf{u} - \Psi(\mathbf{h})\mathbf{w}\right\|^2 + \tfrac{\gamma}{2}\left\|\mathbf{h} - (\mathbf{u})_+\right\|^2 \\ &= \mathcal{R}(\mathbf{s}) + \frac{\gamma}{2}\left\|\mathbf{u} - \Psi(\mathbf{h})\mathbf{w} + \frac{\xi}{\gamma}\right\|^2 + \frac{\gamma}{2}\left\|\mathbf{h} - (\mathbf{u})_+ + \frac{\zeta}{\gamma}\right\|^2 - \frac{\|\xi\|^2}{2\gamma} - \frac{\|\zeta\|^2}{2\gamma}, \end{aligned} \tag{3.1.12}$$

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 3.3 (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 3.4 are essential to design a practical stopping condition (3.1.36) of the BCD method in Algorithm 2. The results will also be used for the convergence results of the BCD method in Theorems 3.2 and 3.3.

Lemma 3.3. *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

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

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

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

(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

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

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

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

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

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

Proof. Statement (i) can be easily obtained by the expression of $\mathcal{L}(\mathbf{s},\xi,\zeta,\gamma)$ in (3.1.12) and the nonnegativity of $\mathcal{R}(\mathbf{s})$ in (3.1.6).

For statement (ii), the nonemptiness and closeness of the level set $\Omega_{\mathcal{L}}(\hat{\Gamma})$ are obvious. Moreover, $\mathcal{R}(\mathbf{s})$ and $\|\mathbf{h}-(\mathbf{u})_+ +\frac{\zeta}{\gamma}\|$ are upper bounded for all $\mathbf{s}$ in $\Omega_{\mathcal{L}}(\hat{\Gamma})$. The fact that $\mathcal{R}(\mathbf{s})$ 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. □

Lemma 3.4. *For any* $\mathbf{z},\mathbf{h},\mathbf{u},\mathbf{h}',\mathbf{u}'$ *in the level set* $\Omega_{\mathcal{L}}(\hat{r})$*, we have*

$$\|\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{z},\mathbf{h}',\mathbf{u}',\xi,\zeta,\gamma)-\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u},\xi,\zeta,\gamma)\|\le L_1(\xi,\zeta,\gamma,\hat{r})\left\|\begin{array}{c}\mathbf{h}'-\mathbf{h}\\ \mathbf{u}'-\mathbf{u}\end{array}\right\|, \quad (3.1.13)$$

$$\|\nabla_{\mathbf{h}}\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u}',\xi,\zeta,\gamma)-\nabla_{\mathbf{h}}\mathcal{L}(\mathbf{z},\mathbf{h},\mathbf{u},\xi,\zeta,\gamma)\|\le L_2(\xi,\zeta,\gamma,\hat{r})\,\|\mathbf{u}'-\mathbf{u}\|, \quad (3.1.14)$$

where

$$L_1(\xi,\zeta,\gamma,\hat{r})=\sqrt{2}\max\{\gamma\delta_1,\delta_2+\delta_3+\delta_4\},\ L_2(\xi,\zeta,\gamma,\hat{r})=\gamma\delta_5, \quad (3.1.15)$$

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

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

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

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

Proof. Using Lemma 3.3 (iii), we have

$$\begin{aligned} &\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{z}, \mathbf{h}', \mathbf{u}', \xi, \zeta, \gamma) - \nabla_{\mathbf{z}}\mathcal{L}(\mathbf{z}, \mathbf{h}, \mathbf{u}, \xi, \zeta, \gamma) \qquad (3.1.16)\\ &= \begin{bmatrix} \gamma\Delta_1\mathbf{w} - (\Psi(\mathbf{h}') - \Psi(\mathbf{h}))^\top \xi - \gamma\Delta_3 \\ \frac{2}{T}\sum_{t=1}^T \Delta_{2,t}\mathbf{a} - \frac{2}{T}\sum_{t=1}^T (\Phi(h_t') - \Phi(h_t))^\top y_t \end{bmatrix}, \end{aligned}$$

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

$$\begin{aligned} \|\Delta_1\| &= \|\Psi(\mathbf{h}')^\top\Psi(\mathbf{h}') - \Psi(\mathbf{h}')^\top\Psi(\mathbf{h}) + \Psi(\mathbf{h}')^\top\Psi(\mathbf{h}) - \Psi(\mathbf{h})^\top\Psi(\mathbf{h})\| \\ &\le (\|\Psi(\mathbf{h}')\| + \|\Psi(\mathbf{h})\|)\,\|\Psi(\mathbf{h}') - \Psi(\mathbf{h})\|. \qquad (3.1.17) \end{aligned}$$

Similarly, we have

$$\|\Delta_{2,t}\| \le (\|\Phi(h_t')\| + \|\Phi(h_t)\|)\,\|\Phi(h_t') - \Phi(h_t)\|,\ \forall t \in [T], \qquad (3.1.18)$$

$$\|\Delta_3\| \le \|\Psi(\mathbf{h}')\|\|\mathbf{u}' - \mathbf{u}\| + \|\mathbf{u}\|\|\Psi(\mathbf{h}') - \Psi(\mathbf{h})\|. \qquad (3.1.19)$$

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 (3.1.6) and $p(\mathbf{s})$ in (3.1.5), yields

$$\|W\|_F \le \sqrt{\frac{\delta}{\lambda_2}},\ \|\mathbf{a}\| \le \sqrt{\frac{\delta}{\min\{\lambda_1, \lambda_5\}}},\ \|\mathbf{w}\| \le \sqrt{\frac{\delta}{\min\{\lambda_2, \lambda_3, \lambda_4\}}},\ \|\mathbf{u}\| \le \sqrt{\frac{\delta}{\lambda_6}}. \qquad (3.1.20)$$

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

$$\|\mathbf{h}\| \leq \delta_0. \tag{3.1.21}$$

Using (3.1.3), we can easily obtain that

$$\|\Psi(\mathbf{h}) - \Psi(\mathbf{h}')\| \leq \sqrt{r}\|\mathbf{h}' - \mathbf{h}\|, \quad \|\Phi(h_t') - \Phi(h_t)\| \leq \sqrt{m}\|h_t' - h_t\|, \tag{3.1.22}$$

$$\|\Psi(\mathbf{h})\| = \sqrt{r(\|\mathbf{h}\|^2 + \|X\|^2 + T)}, \quad \|\Phi(h_t)\| = \sqrt{m(\|h_t\|^2 + 1)}. \tag{3.1.23}$$

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 $B_1, B_2, \ldots, B_j \in \mathbb{R}^{n_c \times n_r}$, $\|B_1\| \leq \|B_1\|_F$, and

$$\begin{aligned} \left\| \sum_{i=1}^{(j)} \iota_j B_j g_j \right\| &\leq \sum_{i=1}^{j} |\iota_j| \|B_j\| \|g_j\|, \\ \sum_{i=1}^{j} \|\iota_i g_i\| &\leq \max_{1 \leq i \leq j} \{|\iota_i|\} \sqrt{j} \|(g_1; \ldots; g_j)\|, \end{aligned} \tag{3.1.24}$$

taking the norm of both sides of (3.1.16), and employing (3.1.17)-(3.1.23), we can get (3.1.13) with the expression of $L_1(\xi, \zeta, \gamma, \hat{r})$ in (3.1.15) as desired.

Using Lemma 3.3 (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 (3.1.20), (3.1.24), and the facts $\|(u_t)_+ - (u_t')_+\| \leq \|u_t' - u_t\|$ for each t, we can get (3.1.14) with $L_2(\xi, \zeta, \gamma, \hat{r})$ in the form of (3.1.15) as desired. □

ALM for the regularized RNNs

To solve the regularized constrained problem (3.1.6), we propose the ALM in Algorithm 1. The ALM first approximately solves (3.1.25) 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 (3.1.26) with tolerance ϵ_{k-1}. Then, the Lagrange multipliers are updated, and the tolerance ϵ_k is reduced so that the subproblem is solved more accurately in the next iteration. 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.

Algorithm 1 The augmented Lagrangian method (ALM) for (3.1.6)

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 $\hat{h}_0 = 0$, $\hat{u}_t = W^0\hat{h}_{t-1} + V^0 x_t + b^0$ and $\hat{h}_t = (\hat{u}_t)_+$ for $t \in [T]$.

2: Set $k := 1$.

3: **Step 1:** Solve

$$\min_{\mathbf{s}} \quad \mathcal{L}(\mathbf{s}, \xi^{k-1}, \zeta^{k-1}, \gamma_{k-1}) \tag{3.1.25}$$

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

$$\operatorname{dist}\big(0, \partial\mathcal{L}(\mathbf{s}^k, \xi^{k-1}, \zeta^{k-1}, \gamma_{k-1})\big) \le \epsilon_{k-1}. \tag{3.1.26}$$

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

$$\xi^k = \xi^{k-1} + \gamma_{k-1}\left(\mathbf{u}^k - \Psi(\mathbf{h}^k)\mathbf{w}^k\right), \quad \zeta^k = \zeta^{k-1} + \gamma_{k-1}\left(\mathbf{h}^k - (\mathbf{u}^k)_+\right). \tag{3.1.27}$$

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

$$\max\left\{\|\mathcal{C}_1(\mathbf{s}^k)\|, \|\mathcal{C}_2(\mathbf{s}^k)\|\right\} \le \eta_1 \max\left\{\|\mathcal{C}_1(\mathbf{s}^{k-1})\|, \|\mathcal{C}_2(\mathbf{s}^{k-1})\|\right\}. \tag{3.1.28}$$

Otherwise, set

$$\gamma_k = \max\left\{\gamma_{k-1}/\eta_2, \left\|\xi^k\right\|^{1+\eta_3}, \left\|\zeta^k\right\|^{1+\eta_3}\right\}. \tag{3.1.29}$$

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

Remark 3.1. *The main operation of Algorithm 1 is to approximately solve the sub-*

problem (3.1.25). Furthermore, to show that Algorithm 1 is well-defined requires that the algorithm for solving the subproblem (3.1.25) can be terminated within finite steps to meet the stopping condition in (3.1.26).

In the next subsection, we will design a BCD method to solve the subproblem (3.1.25). The update of each block of the BCD method owns a closed-form formula, which makes the BCD method efficient. Moreover, the stopping condition (3.1.26) can be replaced by a more straightforward condition (3.1.36) as will be shown in Theorem 3.2.

BCD method for subproblem

To solve the nonsmooth nonconvex problem (3.1.25) in Step 1 of Algorithm 1, we propose a BCD method in Algorithm 2 to solve the subproblem at the k-th iteration in the ALM. The main idea of the BCD method is to split the variables into several blocks and update each block by fixing the other components, so that a complex optimization problem can be solved by addressing several simpler subproblems. This method performs competitively in various applications, including several in computational statistics and machine learning [58]. The BCD method has made significant strides in recent years, with techniques such as randomization, acceleration, and parallel computing being successfully applied to enhance its performance and scalability. Meanwhile, the convergence of the BCD method has also been gradually established. In the early stages of using the BCD method to solve nonconvex and nonsmooth problems, Tseng [55] proposed that for a nondifferentiable and nonconvex problem with N block variables, if $N-1$ subproblems have at most one minimum, any accumulation point of the sequences is a coordinate-wise minimum point of the objective function. At this stage, each subproblem was required to be solved exactly and must yield a unique minimizer, which is difficult to achieve for many problems. Subsequently, the prox-linear update scheme was proposed to inex-

actly solve each subproblem, and the algorithm was shown to globally converge to a critical point [56, 8]. The new scheme makes the BCD method easier to compute and gives a better solution overall. Furthermore, Xu and Yin [65] used extrapolation to accelerate the convergence of the prox-linear update scheme, and they further established its global convergence (of the whole sequence) to a critical point under the Kurdyka–Łojasiewicz property of the objective function [66].

By observing our subproblems (3.1.25),

$$\begin{aligned}\mathcal{L}(\mathbf{s}, \xi^{k-1}, \zeta^{k-1}, \gamma_{k-1}) =& \mathcal{R}(\mathbf{s}) + \langle \xi^{k-1}, \mathbf{u} - \Psi(\mathbf{h})\mathbf{w} \rangle + \langle \zeta^{k-1}, \mathbf{h} - (\mathbf{u})_+ \rangle \\ &+ \tfrac{\gamma_{k-1}}{2} \left\| \mathbf{u} - \Psi(\mathbf{h})\mathbf{w} \right\|^2 + \tfrac{\gamma_{k-1}}{2} \left\| \mathbf{h} - (\mathbf{u})_+ \right\|^2,\end{aligned}$$

we find that the nonconvexity is mainly caused by the bilinear term $-\Psi(\mathbf{h})\mathbf{w}$, as well as the term $-(\mathbf{u})_+$. Therefore, we divide the variable $\mathbf{s}$ into three blocks $\mathbf{z}$, $\mathbf{h}$, and $\mathbf{u}$, and alternatively update each block. Furthermore, each subproblem admits a closed-form solution. The details of our method are presented below.

Let us choose a constant Γ such that

$$\Gamma \geq \mathcal{L}\big(\mathbf{s}^0, \xi^0, \zeta^0, \gamma_0\big). \tag{3.1.30}$$

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 appearing at the k-th iteration of the ALM, we define

$$\mathbf{s}_{\mathbf{z}}^{k-1,j} := (\mathbf{z}^{k-1,j}; \mathbf{h}^{k-1,j-1}; \mathbf{u}^{k-1,j-1}),\ \mathbf{s}_{\mathbf{h}}^{k-1,j} := (\mathbf{z}^{k-1,j}; \mathbf{h}^{k-1,j}; \mathbf{u}^{k-1,j-1}) \tag{3.1.31}$$

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

$$\mathbf{s}^{k-1,j} = (\mathbf{z}^{k-1,j}; \mathbf{h}^{k-1,j}; \mathbf{u}^{k-1,j}) \tag{3.1.32}$$

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

Algorithm 2 Block Coordinate Descent (BCD) method for (3.1.25)

Set the initial point of BCD algorithm as

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

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 (3.1.15).

Set $j := 1$.

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

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

$$\mathbf{z}^{k-1,j} = \arg\min_{\mathbf{z}} \mathcal{L}\left(\mathbf{z}, \mathbf{h}^{k-1,j-1}, \mathbf{u}^{k-1,j-1}, \xi, \zeta, \gamma\right), \tag{3.1.33}$$

$$\mathbf{h}^{k-1,j} = \arg\min_{\mathbf{h}} \mathcal{L}\left(\mathbf{z}^{k-1,j}, \mathbf{h}, \mathbf{u}^{k-1,j-1}, \xi, \zeta, \gamma\right), \tag{3.1.34}$$

$$\mathbf{u}^{k-1,j} \in \arg\min_{\mathbf{u}} \mathcal{L}\left(\mathbf{z}^{k-1,j}, \mathbf{h}^{k-1,j}, \mathbf{u}, \xi, \zeta, \gamma\right) + \tfrac{\beta}{2}\left\|\mathbf{u} - \mathbf{u}^{k-1,j-1}\right\|^2. \tag{3.1.35}$$

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

Step 2: If the stop criterion

$$\left\|\mathbf{s}^{k-1,j} - \mathbf{s}^{k-1,j-1}\right\| \leq \frac{\epsilon_{k-1}}{\max\{L_{1,k-1}, L_{2,k-1}, \beta\}} \tag{3.1.36}$$

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

end while

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

Condition (3.1.26) is satisfied when (3.1.36) holds, which will be proved in Theorem 3.2. The closed-form solutions of problems (3.1.33), (3.1.34) and (3.1.35) are provided below.

Update $\mathbf{z}^{k-1,j}$: Problem (3.1.33) is an unconstrained optimization problem with a smooth and strongly convex objective function. By employing Lemma 3.3 (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 (3.1.34) is also strongly convex and smooth. By employing Lemma 3.3 (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

$$h_t^{k-1,j} = \begin{cases} D_1(\mathbf{s}_{\mathbf{h}}^{k-1,j}, \xi, \zeta, \gamma)^{-1} d_{1t}(\mathbf{s}_{\mathbf{h}}^{k-1,j}, \xi, \zeta, \gamma), & \text{if } t \in [T-1], \\ D_2(\mathbf{s}_{\mathbf{h}}^{k-1,j}, \xi, \zeta, \gamma)^{-1} d_{2T}(\mathbf{s}_{\mathbf{h}}^{k-1,j}, \xi, \zeta, \gamma), & \text{if } t = T. \end{cases} \tag{3.1.37}$$

Update $\mathbf{u}^{k-1,j}$: Although problem (3.1.35) is nonsmooth nonconvex, one of its global solutions is accessible, because the objective function of problem (3.1.35) can be separated into rT one-dimensional functions with the same structure. Thus, we aim to solve the following one-dimensional problem:

$$\min_{u \in \mathbb{R}} \ \varphi(u) := \tfrac{\gamma}{2}(u - \theta_1)^2 + \tfrac{\gamma}{2}(\theta_2 - (u)_+)^2 + \tfrac{\beta}{2}(u - \theta_3)^2 + \lambda_6 u^2, \tag{3.1.38}$$

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

$$u^+ := \arg\min_{u \in \mathbb{R}_+} \varphi(u) \quad \text{and} \quad u^- := \arg\min_{u \in \mathbb{R}_-} \varphi(u). \tag{3.1.39}$$

By direct computation,

$$u^+ = \begin{cases} \dfrac{\gamma\theta_1 + \gamma\theta_2 + \beta\theta_3}{2\gamma + 2\lambda_6 + \beta}, & \text{if } \gamma\theta_1 + \gamma\theta_2 + \beta\theta_3 > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{3.1.40}$$

and

$$u^- = \begin{cases} \dfrac{\gamma\theta_1 + \beta\theta_3}{\gamma + 2\lambda_6 + \beta}, & \text{if } \gamma\theta_1 + \beta\theta_3 < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.1.41}$$

Then a solution of (3.1.38) can be given as

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

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 (3.1.35) as

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

Remark 3.2. *It is important to mention that the solution set of problem* (3.1.35) *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* $i \in [rT]$.

3.1.3 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 (3.1.6).

Convergence analysis of Algorithm 2

It is clear that

$$\mathcal{L}(\mathbf{s}, \xi, \zeta, \gamma) = g(\mathbf{s}, \xi, \gamma) + q(\mathbf{s}, \zeta, \gamma), \tag{3.1.42}$$

where

$$g(\mathbf{s}, \xi, \gamma) = \mathcal{R}(\mathbf{s}) + \frac{\gamma}{2}\left\|\mathbf{u} - \Psi(\mathbf{h})\mathbf{w} + \frac{\xi}{\gamma}\right\|^2 - \frac{\|\xi\|^2}{2\gamma}, \tag{3.1.43}$$

$$q(\mathbf{s}, \zeta, \gamma) = \frac{\gamma}{2}\left\|\mathbf{h} - (\mathbf{u})_+ + \frac{\zeta}{\gamma}\right\|^2 - \frac{\|\zeta\|^2}{2\gamma}. \tag{3.1.44}$$

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 (3.1.31), and use $\mathbf{s}^{(j)}$ to represent $s^{k-1,j}$ in (3.1.32) for brevity. We emphasize that the point $\mathbf{s}^k$ is generated by the ALM in Algorithm 1, while the point $\mathbf{s}^{(j)}$ is generated by the BCD method in Algorithm 2 for solving the subproblem (3.1.25) 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 3.5. *Let $\{\mathbf{s}^{(j)}\}$ represent the sequence generated by Algorithm 2. Then $\{\mathbf{s}^{(j)}\}$ belongs to the level set $\Omega_{\mathcal{L}}(\Gamma)$, which is compact.*

Proof. By (3.1.33), (3.1.34) and (3.1.35), we know that for any $j \in \mathbb{N}$,

$$\mathcal{L}\big(\mathbf{s}^{(j)},\xi,\zeta,\gamma\big) \le \mathcal{L}\big(\mathbf{s}_{\mathbf{h}}^{(j)},\xi,\zeta,\gamma\big) \le \mathcal{L}\big(\mathbf{s}_{\mathbf{z}}^{(j)},\xi,\zeta,\gamma\big) \le \mathcal{L}\big(\mathbf{s}^{(j-1)},\xi,\zeta,\gamma\big). \tag{3.1.45}$$

By the definition of Γ in Algorithm 2 and (3.1.45), we can deduce that

$$\mathcal{L}\big(\mathbf{s}^{(j)},\xi,\zeta,\gamma\big) \le \Gamma, \quad \forall j \in \mathbb{N}. \tag{3.1.46}$$

By the definition of $\Omega_{\mathcal{L}}(\Gamma)$ and Lemma 3.3 (ii), the proof is completed. □

Lemma 3.6. *The AL function $\mathcal{L}$ is locally Lipschitz continuous and directionally differentiable on $\Omega_{\mathcal{L}}(\Gamma)$.*

Proof. It is clear that $\Omega_{\mathcal{L}}(\Gamma)$ is compact by Lemma 3.3 (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

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

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 [37]. □

We can now show that the stop criterion (3.1.36) in Algorithm 2 can be satisfied in finite steps, and condition (3.1.26) in Algorithm 1 is satisfied when (3.1.36) holds. These results guarantee that the ALM in Algorithm 1 is well-defined, when the subproblems are solved by the BCD method in Algorithm 2.

Theorem 3.2. *At the k-th iteration of ALM in Algorithm 1, the BCD method in Algorithm 2 for the subproblem* (3.1.25) *can be stopped within finite steps to satisfy the stop criterion in* (3.1.36)*, which is of order* $O(1/(\epsilon_{k-1})^2)$*. Moreover, condition* (3.1.26) *of the ALM in Algorithm 1 is satisfied at the output* $\mathbf{s}^k$ *of Algorithm 2.*

Proof. Since $\mathcal{L}$ is strongly convex with respect to the blocks $\mathbf{z}$ and $\mathbf{h}$, respectively, from (3.1.33) and (3.1.34), we obtain

$$\mathcal{L}(\mathbf{s}^{(j-1)},\xi,\zeta,\gamma)-\mathcal{L}(\mathbf{s}_{\mathbf{z}}^{(j)},\xi,\zeta,\gamma)\geq\tfrac{\alpha_1}{2}\|\mathbf{z}^{(j-1)}-\mathbf{z}^{(j)}\|^2, \tag{3.1.47}$$

$$\mathcal{L}(\mathbf{s}_{\mathbf{z}}^{(j)},\xi,\zeta,\gamma)-\mathcal{L}(\mathbf{s}_{\mathbf{h}}^{(j)},\xi,\zeta,\gamma)\geq\tfrac{\alpha_2}{2}\|\mathbf{h}^{(j-1)}-\mathbf{h}^{(j)}\|^2, \tag{3.1.48}$$

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

(3.1.35), we have

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

It follows that

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

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

$$\begin{aligned}
\mathcal{L}(\mathbf{s}^{(0)}, \xi, \zeta, \gamma) - \mathcal{L}(\mathbf{s}^{(J)}, \xi, \zeta, \gamma) & \geq \max\{\tfrac{\alpha_1}{2}, \tfrac{\alpha_2}{2}, \tfrac{\beta}{2}\} \sum_{j=1}^{J} \|\mathbf{s}^{(j)} - \mathbf{s}^{(j-1)}\|^2 \quad (3.1.49) \\
& \geq J \max\{\tfrac{\alpha_1}{2}, \tfrac{\alpha_2}{2}, \tfrac{\beta}{2}\} \min_{j \in [J]} \{\|\mathbf{s}^{(j)} - \mathbf{s}^{(j-1)}\|^2\}.
\end{aligned}$$

This, together with Lemma 3.3 (i), yields that

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

It follows that the stop criterion (3.1.36) holds, as long as

$$J \geq \hat{J} := \left\lceil \frac{\left(\mathcal{L}(\mathbf{s}^{(0)}, \xi, \zeta, \gamma) + \frac{\|\xi\|^2}{2\gamma} + \frac{\|\zeta\|^2}{2\gamma} \right) (\max\{L_{1,k-1}, L_{2,k-1}, \beta\})^2}{\max\{\frac{\alpha_1}{2}, \frac{\alpha_2}{2}, \frac{\beta}{2}\} (\epsilon_{k-1})^2} \right\rceil . \quad (3.1.50)$$

Therefore, at the k-th iteration of the ALM in Algorithm 1, the BCD method in Algorithm 2 can be stopped in at most $\hat{J}$ iterations defined in (3.1.50) and output $\mathbf{s}^k$, which is of order $O(1/(\epsilon_{k-1})^2)$.

Once condition (3.1.36) is satisfied, condition (3.1.26) in Algorithm 1 also holds, which will be proved in the following. By Step 1 in Algorithm 2, the first order optimality condition of the three blocked subproblems (3.1.33), (3.1.34) and (3.1.35) are

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

$$0 \in \nabla_{\mathbf{u}} g(\mathbf{s}^{(j)}, \xi, \gamma) + \partial_{\mathbf{u}} q(\mathbf{s}^{(j)}, \zeta, \gamma) + \beta(\mathbf{u}^{(j)} - \mathbf{u}^{(j-1)}).$$

Furthermore, the l-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

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

By Lemma 3.4, we obtain

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

Thus condition (3.1.36) that $\|\mathbf{s}^{(j)} - \mathbf{s}^{(j-1)}\| \leq \epsilon_{k-1}/\max\{L_{1,k-1}, L_{2,k-1}, \beta\}$, together with $\mathbf{s}^k = \mathbf{s}^{(j)}$, implies $\operatorname{dist}(0, \partial\mathcal{L}(\mathbf{s}^{(k)}, \xi, \zeta, \gamma)) \leq \epsilon_{k-1}$ in condition (3.1.26). □

In Theorem 3.2, we have proved that a BCD method, with subproblems admitting closed-form solutions and employing a cyclic updating rule, achieves an iteration complexity of $O(1/\epsilon^2)$ for nonconvex and nonsmooth problems, while the BCD method achieves an iteration complexity of $O(1/\epsilon)$ for nonsmooth and convex problems [26], and $O(\log(1/\epsilon))$ for smooth and strongly convex problems [1].

Theorem 3.2 guarantees that the BCD method in Algorithm 2 terminates within finite steps to meet the stop criterion (3.1.36) for a fixed $\epsilon_{k-1} > 0$. In the rest of this subsection, we discuss the convergence of Algorithm 2 for the case $\epsilon_{k-1} = 0$, i.e., we replace the stop criterion (3.1.36) by

$$\left\|\mathbf{s}^{k-1,j} - \mathbf{s}^{k-1,j-1}\right\| = 0. \tag{3.1.51}$$

We will show in Theorem 3.4 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 3.3. *Suppose that (3.1.36) is replaced by (3.1.51) in Algorithm 2. If there is $\bar{j}$ such that (3.1.51) holds, then*

$$\mathcal{L}(\mathbf{s}_{\mathbf{z}}^{(\bar{j})}, \xi, \zeta, \gamma) = \mathcal{L}(\mathbf{s}_{\mathbf{h}}^{(\bar{j})}, \xi, \zeta, \gamma) = \mathcal{L}(\mathbf{s}^{(\bar{j})}, \xi, \zeta, \gamma) \quad \text{and} \quad \mathbf{s}_{\mathbf{z}}^{(\bar{j})} = \mathbf{s}_{\mathbf{h}}^{(\bar{j})} = \mathbf{s}^{(\bar{j})}. \tag{3.1.52}$$

Otherwise, Algorithm 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 $\{\mathbf{s}^{(j_i)}\}$ converge to the same point.

Proof. If there is $\bar{j}$ such that (3.1.51) holds, then (3.1.52) is derived directly from $\mathbf{s}^{k-1,\bar{j}} = \mathbf{s}^{k-1,\bar{j}-1}$ and (3.1.33)-(3.1.35).

If there is no $\bar{j}$ such that (3.1.51) holds, then Algorithm 2 generates infinite sequences $\{\mathbf{s}_{\mathbf{z}}^{(j)}\}$, $\{\mathbf{s}_{\mathbf{h}}^{(j)}\}$ and $\{\mathbf{s}^{(j)}\}$.

(i) By Lemma 3.5, 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 (3.1.45) and the lower boundedness of $\{\mathcal{L}(\mathbf{s}^{(j)}, \xi, \zeta, \gamma)\}$ according to Lemma 3.3 (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

$$\lim_{i\to\infty} \|\mathbf{s}^{(j_i)} - \mathbf{s}_{\mathbf{z}}^{(j_i)}\| = 0, \quad \lim_{i\to\infty} \|\mathbf{s}^{(j_i)} - \mathbf{s}_{\mathbf{h}}^{(j_i)}\| = 0. \tag{3.1.53}$$

Letting J go to infinity and replacing (j) in (3.1.49) 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,

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

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 (3.1.53).

□

Now, we show that Algorithm 2 generates a d-stationary point of problem (3.1.25). For convenience, we adopt a simple expression for the directional derivative of a function, emphasizing the blocks of the direction. 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 3.7. *If the directional derivatives of $\mathcal{L}$ at $\bar{\mathbf{s}} \in \Omega_{\mathcal{L}}(\Gamma)$ satisfy*

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

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

$$\mathcal{L}'(\bar{\mathbf{s}}, \xi, \zeta, \gamma; d) \ge 0, \quad \forall\ d \in \mathbb{R}^{N_{\mathbf{w}}+N_{\mathbf{a}}+2rT}.$$

Proof. By (3.1.42), 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

$$g'(\bar{\mathbf{s}}, \xi, \gamma; d) = \langle \nabla_{\mathbf{z}} g(\bar{\mathbf{s}}, \xi, \gamma), d_{\mathbf{z}} \rangle + \langle \nabla_{\mathbf{h}} g(\bar{\mathbf{s}}, \xi, \gamma), d_{\mathbf{h}} \rangle + \langle \nabla_{\mathbf{u}} g(\bar{\mathbf{s}}, \xi, \gamma), d_{\mathbf{u}} \rangle. \quad (3.1.54)$$

The directional derivative of nonsmooth part q remains to be considered. 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 (3.1.54), 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 3.7 holds. □

As problem (3.1.25) 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 limiting stationary point is a Fréchet stationary point, and a d-stationary point is a limiting stationary point, but not vice versa [36]. The theorem below guarantees that either the BCD method terminates at a d-stationary point of problem (3.1.25) in finite steps, or any accumulation point of the sequence generated by the BCD method is a d-stationary point of problem (3.1.25).

Theorem 3.4. *Suppose that (3.1.36) is replaced by (3.1.51) in Algorithm 2. If there is $\bar{j}$ such that (3.1.51) holds, then $\mathbf{s}^{(\bar{j})}$ is a d-stationary point of problem* (3.1.25). *Otherwise, Algorithm 2 generates an infinite sequence $\{\mathbf{s}^{(j)}\}$ and any accumulation point of $\{\mathbf{s}^{(j)}\}$ is a d-stationary point of problem* (3.1.25).

Proof. If there is $\bar{j}$ such that (3.1.51) 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 (3.1.52) of Theorem 3.3, 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 (3.1.33)-(3.1.35) in Algorithm 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 3.6 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 3.7, yield 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 (3.1.25).

If there is no $\bar{j}$ such that (3.1.51) holds, then Algorithm 2 generates an infinite sequence $\{\mathbf{s}^{(j)}\}$. By (3.1.35), we have

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

Letting $j \to \infty$ in the above inequalities and using Theorem 3.3 (i), we have

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

By Theorem 3.3 (ii), let $\{\mathbf{s}_z^{(j_i)}\}$, $\{\mathbf{s}_h^{(j_i)}\}$ and $\{\mathbf{s}^{(j_i)}\}$ be any convergent subsequence with limit $\bar{\mathbf{s}}$. Furthermore, by (3.1.33)-(3.1.35) in Algorithm 2, we have for any

$\lambda > 0$ and any $d_{\mathbf{z}} \in \mathbb{R}^{N_{\mathbf{w}}+N_{\mathbf{a}}}$, $d_{\mathbf{h}} \in \mathbb{R}^{rT}$, $d_{\mathbf{u}} \in \mathbb{R}^{rT}$,

$$\mathcal{L}(\mathbf{s}_{\mathbf{z}}^{(j_i)}, \xi, \zeta, \gamma) \le \mathcal{L}\big(\mathbf{s}_{\mathbf{z}}^{(j_i)} + \lambda(d_{\mathbf{z}}, 0, 0), \xi, \zeta, \gamma\big),$$

$$\mathcal{L}\big(\mathbf{s}_{\mathbf{h}}^{(j_i)}, \xi, \zeta, \gamma\big) \le \mathcal{L}\big(\mathbf{s}_{\mathbf{h}}^{(j_i)} + \lambda(0, d_{\mathbf{h}}, 0), \xi, \zeta, \gamma\big),$$

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

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

$$\mathcal{L}(\bar{\mathbf{s}}, \xi, \zeta, \gamma) \le \mathcal{L}\big(\bar{\mathbf{s}} + \lambda(d_{\mathbf{z}}, 0, 0), \xi, \zeta, \gamma\big),\ \mathcal{L}(\bar{\mathbf{s}}, \xi, \zeta, \gamma) \le \mathcal{L}\big(\bar{\mathbf{s}} + \lambda(0, d_{\mathbf{h}}, 0), \xi, \zeta, \gamma\big),$$

$$\mathcal{L}(\bar{\mathbf{s}}, \xi, \zeta, \gamma) \le \mathcal{L}\big(\bar{\mathbf{s}} + \lambda(0, 0, d_{\mathbf{u}}), \xi, \zeta, \gamma\big) + \tfrac{\beta}{2}\lambda^2\|d_{\mathbf{u}}\|^2.$$

By Lemma 3.6 and the definition of directional derivative, it follows that

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

for any $d_{\mathbf{z}}$, $d_{\mathbf{h}}$ and $d_{\mathbf{u}}$. The above inequalities, along with Lemma 3.7, yield that $\bar{\mathbf{s}}$ is a d-stationary point of problem (3.1.25). □

Convergence analysis of Algorithm 1

By Theorem 3.2, the ALM in Algorithm 1 is well-defined, since Step 1 can always be fulfilled in finite steps by the BCD method in Algorithm 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 1 is a feasible point. The delicate strategy for updating the penalty parameter γ_k in Step 3 of Algorithm 1 plays an important role in the proof of the theorem.

Theorem 3.5. *Let $\{\mathbf{s}^k\}$ be the sequence generated by Algorithm 1. Then the following statements hold.*

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

(ii) There exists at least one accumulation point of $\{\mathbf{s}^k\}$*, and any accumulation point is a feasible point of* (3.1.6).

Proof. (i) Let the index set

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

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

$$\begin{aligned}\max\left\{\|\mathcal{C}_1(\mathbf{s}^k)\|, \|\mathcal{C}_2(\mathbf{s}^k)\|\right\} &\leq \eta_1 \max\left\{\|\mathcal{C}_1(\mathbf{s}^{k-1})\|, \|\mathcal{C}_2(\mathbf{s}^{k-1})\|\right\} \\ &\leq \eta_1^{k-K} \max\left\{\|\mathcal{C}_1(\mathbf{s}^K)\|, \|\mathcal{C}_2(\mathbf{s}^K)\|\right\}. \quad (3.1.55)\end{aligned}$$

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 (3.1.29) and $\eta_3 > 1$ yield that

$$\lim_{k\to\infty, k-1\in\mathcal{K}} \max\left\{\frac{\|\xi^{k-1}\|}{\gamma_{k-1}}, \frac{\|\zeta^{k-1}\|}{\gamma_{k-1}}, \frac{\|\xi^{k-1}\|^2}{\gamma_{k-1}}, \frac{\|\zeta^{k-1}\|^2}{\gamma_{k-1}}\right\} = 0. \quad (3.1.56)$$

Recalling (3.1.12), and employing condition (3.1.46) and Step 1 of Algorithm 2, we have

$$\begin{aligned}0 \leq& \left\|\mathbf{u}^k - \Psi(\mathbf{h}^k)\mathbf{w}^k + \tfrac{\xi^{k-1}}{\gamma_{k-1}}\right\|^2 + \left\|\mathbf{h}^k - (\mathbf{u}^k)_+ + \tfrac{\zeta^{k-1}}{\gamma_{k-1}}\right\|^2 \\ \leq& \tfrac{2}{\gamma_{k-1}}\left(\Gamma - \mathcal{R}(\mathbf{s}^k)\right) + \left(\tfrac{\|\xi^{k-1}\|}{\gamma_{k-1}}\right)^2 + \left(\tfrac{\|\zeta^{k-1}\|}{\gamma_{k-1}}\right)^2.\end{aligned} \quad (3.1.57)$$

Then by (3.1.56) and the lower boundedness of $\left\{\mathcal{R}(\mathbf{s}^k)\right\}$, we have

$$\lim_{k\to\infty, k-1\in\mathcal{K}} \|\mathbf{u}^k - \Psi(\mathbf{h}^k)\mathbf{w}^k\| = 0 \quad \text{and} \quad \lim_{k\to\infty, k-1\in\mathcal{K}} \|\mathbf{h}^k - (\mathbf{u}^k)_+\| = 0. \quad (3.1.58)$$

To extend the results in (3.1.58) 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 (3.1.58). 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 (3.1.29), 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:

$$\frac{\|\xi^i\|}{\gamma_i} = \frac{\|\xi^i\|}{\gamma_{i-1}} \leq \frac{\|\xi^{i-1}\|}{\gamma_{i-1}} + \left\|\mathbf{u}^i - \Psi(\mathbf{h}^i)\mathbf{w}^i\right\|, \tag{3.1.59}$$

$$\frac{\|\zeta^i\|}{\gamma_i} = \frac{\|\zeta^i\|}{\gamma_{i-1}} \leq \frac{\|\zeta^{i-1}\|}{\gamma_{i-1}} + \left\|\mathbf{h}^i - (\mathbf{u}^i)_+\right\|. \tag{3.1.60}$$

Summing up inequalities (3.1.59) and (3.1.60) for every $i \in \mathcal{I}_k$, we have

$$\frac{\|\xi^{k-1}\|}{\gamma_{k-1}} \leq \frac{\|\xi^{l_k}\|}{\gamma_{l_k}} + \sum_{i=1}^{k-l_k-1} \left\|\mathbf{u}^{k-i} - \Psi(\mathbf{h}^{k-i})\mathbf{w}^{k-i}\right\|, \tag{3.1.61}$$

$$\frac{\|\zeta^{k-1}\|}{\gamma_{k-1}} \leq \frac{\|\zeta^{l_k}\|}{\gamma_{l_k}} + \sum_{i=1}^{k-l_k-1} \left\|\mathbf{h}^{k-i} - (\mathbf{u}^{k-i})_+\right\|. \tag{3.1.62}$$

By the updating rule of γ_k in (3.1.28), (3.1.61) and (3.1.62), 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 (3.1.56), (3.1.58) and $\eta_1 \in (0,1)$, yields that

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

By the inequality (3.1.57) and nondecreasing sequence $\{\gamma_k\}$, we conclude that

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

using the same manner for showing (3.1.58).

(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 (3.1.27) for those $k > K$, and using (3.1.28), 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 (3.1.56) 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 (3.1.46), and $\mathbf{s}^k = \mathbf{s}^{k-1,j}$:

$$\begin{aligned}&\mathcal{R}(\mathbf{s}^k) + \frac{\gamma_{k-1}}{2}\left\|\mathbf{u}^k - \Psi(\mathbf{h}^k)\mathbf{w}^k + \frac{\xi^{k-1}}{\gamma_{k-1}}\right\|^2 + \frac{\gamma_{k-1}}{2}\left\|\mathbf{h}^k - (\mathbf{u}^k)_+ + \frac{\zeta^{k-1}}{\gamma_{k-1}}\right\|^2\\&\leq \Gamma + \frac{\|\xi^{k-1}\|^2}{2\gamma_{k-1}} + \frac{\|\zeta^{k-1}\|^2}{2\gamma_{k-1}}.\end{aligned} \tag{3.1.63}$$

The above inequality, along with the boundedness of

$$\left\{\|\xi^{k-1}\|^2/\gamma_{k-1}\right\}_{k-1\in\mathcal{K}},\ \left\{\|\zeta^{k-1}\|^2/\gamma_{k-1}\right\}_{k-1\in\mathcal{K}},$$

yields the boundedness of $\{\mathbf{s}^k\}_{k-1\in\mathcal{K}}$ by the same manner in Lemma 3.3 (ii). Hence there exists at least one accumulation point of $\{\mathbf{s}^k\}$.

Any accumulation point is a feasible point of (3.1.6), which can be derived immediately by (i), because of the continuity of the functions in the constraints of (3.1.6). □

Below we show the main convergence result of the ALM.

Theorem 3.6. *Every accumulation point of* $\{\mathbf{s}^k\}$ *generated by Algorithm 1 is a KKT point of problem* (3.1.6).

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 (3.5) (ii). We claim that

$$\begin{aligned}
&\partial\mathcal{L}\big(\mathbf{s}^{k_i}, \xi^{k_i-1}, \zeta^{k_i-1}, \gamma_{k_i-1}\big) \\
&= \nabla\mathcal{R}(\mathbf{s}^{k_i}) + \nabla_{\mathbf{s}}\Big(\langle \xi^{k_i-1}, \mathbf{u}^{k_i} - \Psi(\mathbf{h}^{k_i})\mathbf{w}^{k_i}\rangle + \tfrac{\gamma_{k_i-1}}{2}\big\|\mathbf{u}^{k_i} - \Psi(\mathbf{h}^{k_i})\mathbf{w}^{k_i}\big\|^2\Big) \\
&\quad + \partial_{\mathbf{s}}\Big(\langle \zeta^{k_i-1}, \mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+\rangle + \tfrac{\gamma_{k_i-1}}{2}\big\|\mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+\big\|^2\Big) \\
&= \nabla\mathcal{R}(\mathbf{s}^{k_i}) + J\mathcal{C}_1(\mathbf{s}^{k_i})^\top \xi^{k_i} + \partial\Big((\zeta^{k_i})^\top \mathcal{C}_2(\mathbf{s}^{k_i})\Big),
\end{aligned} \tag{3.1.64}$$

where $\mathcal{C}_1$ and $\mathcal{C}_2$ are defined in (3.1.4).

First, by employing (3.1.27) and by direct computation, we have

$$\begin{aligned}
&\nabla_{\mathbf{s}}\big(\langle \xi^{k_i-1}, \mathbf{u}^{k_i} - \Psi(\mathbf{h}^{k_i})\mathbf{w}^{k_i}\rangle + \tfrac{\gamma_{k_i-1}}{2}\big\|\mathbf{u}^{k_i} - \Psi(\mathbf{h}^{k_i})\mathbf{w}^{k_i}\big\|^2\big) \\
&= J\mathcal{C}_1(\mathbf{s}^{k_i})^\top\big(\xi^{k_i-1} + \gamma_{k_i-1}(\mathbf{u}^{k_i} - \Psi(\mathbf{h}^{k_i})\mathbf{w}^{k_i})\big) = J\mathcal{C}_1(\mathbf{s}^{k_i})^\top \xi^{k_i}.
\end{aligned} \tag{3.1.65}$$

Then, it remains to verify that

$$\partial_{\mathbf{s}}(\langle \zeta^{k_i-1}, \mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+\rangle + \tfrac{\gamma_{k_i-1}}{2}\|\mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+\|^2) = \partial\Big((\zeta^{k_i})^\top \mathcal{C}_2(\mathbf{s}^{k_i})\Big). \tag{3.1.66}$$

To verify (3.1.66), it can be divided into the subdifferential associated with $\mathbf{h}$ and $\mathbf{u}$. We first prove that (3.1.66) is satisfied associated with $\mathbf{h}$. By simple computation,

$$\begin{aligned}
&\nabla_{\mathbf{h}}\big(\langle \zeta^{k_i-1}, \mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+\rangle + \tfrac{\gamma_{k_i-1}}{2}\big\|\mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+\big\|^2\big) \\
&= J_{\mathbf{h}}\mathcal{C}_2(\mathbf{z}^{k_i}, \mathbf{h}^{k_i}, \mathbf{u}^{k_i})^\top\big(\zeta^{k_i-1} + \gamma_{k_i-1}(\mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+)\big) \\
&= J_{\mathbf{h}}\mathcal{C}_2(\mathbf{z}^{k_i}, \mathbf{h}^{k_i}, \mathbf{u}^{k_i})^\top \zeta^{k_i} = \nabla_{\mathbf{h}}\big(\langle \zeta^{k_i}, \mathbf{h}^{k_i} - (\mathbf{u}^{k_i})_+\rangle\big).
\end{aligned} \tag{3.1.67}$$

Then we prove that (3.1.66) is satisfied associated with $\mathbf{u}$, which can be replaced by proving rT one dimensional equations with the similar structure as follows:

$$\partial_{\mathbf{u}_j}\Big(\zeta_j^{k_i-1}(\mathbf{h}_j^{k_i} - (\mathbf{u}_j^{k_i})_+) + \tfrac{\gamma_{k_i-1}}{2}(\mathbf{h}_j^{k_i} - (\mathbf{u}_j^{k_i})_+)^2\Big) = \partial_{\mathbf{u}_j}\Big(\zeta_j^{k_i}(\mathbf{h}_j^{k_i} - (\mathbf{u}_j^{k_i})_+)\Big), \tag{3.1.68}$$

where $j = 1, 2, ..., rT$. When $\mathbf{u}_j^{k_i} \neq 0$, equation (3.1.68) can be easily deduced by the same proof method as in (3.1.67). When $\mathbf{u}_j^{k_i} = 0$, the validity of (3.1.68) can be proved as follows:

$$\begin{aligned}
&\partial_{\mathbf{u}_j}\Big(\zeta_j^{k_i-1}(\mathbf{h}_j^{k_i} - (\mathbf{u}_j^{k_i})_+) + \tfrac{\gamma_{k_i-1}}{2}(\mathbf{h}_j^{k_i} - (\mathbf{u}_j^{k_i})_+)^2\Big) \\
&= \begin{cases} \{0, -\zeta_j^{k_i-1} - \gamma_{k_i-1}(\mathbf{h}_j^{k_i} - \mathbf{u}_j^{k_i})\}, & \text{if } \gamma_{k_i-1}\mathbf{h}_j^{k_i} + \zeta_j^{k_i-1} \geq 0, \\ \big[0, -\zeta_j^{k_i-1} - \gamma_{k_i-1}(\mathbf{h}_j^{k_i} - \mathbf{u}_j^{k_i})\big], & \text{if } \gamma_{k_i-1}\mathbf{h}_j^{k_i} + \zeta_j^{k_i-1} < 0, \end{cases} \\
&= \begin{cases} \{0, -\zeta_j^{k_i}\}, & \text{if } \zeta_j^{k_i} \geq 0, \\ \big[0, -\zeta_j^{k_i}\big], & \text{if } \zeta_j^{k_i} < 0, \end{cases} \\
&= \partial_{\mathbf{u}_j}\Big(\zeta_j^{k_i}(\mathbf{h}_j^{k_i} - (\mathbf{u}_j^{k_i})_+)\Big).
\end{aligned} \tag{3.1.69}$$

Combining (3.1.65) and (3.1.66) yields the validity of (3.1.64).

Up to now, we have verified that equation (3.1.64) holds. Thus, there exists a sequence $\{\varsigma^{k_i}\}$ satisfying $\|\varsigma^{k_i}\| \leq \epsilon^{k_i}$ such that

$$\varsigma^{k_i} \in \nabla\mathcal{R}(\mathbf{s}^{k_i}) + J\mathcal{C}_1(\mathbf{s}^{k_i})^\top \xi^{k_i} + \partial\Big((\zeta^{k_i})^\top \mathcal{C}_2(\mathbf{s}^{k_i})\Big). \tag{3.1.70}$$

However, the boundedness of $\{\xi^{k_i}\}$ and $\{\zeta^{k_i}\}$ in (3.1.70) 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

$$\max\{\|\bar{\xi}\|_\infty, \|\bar{\zeta}\|_\infty\} = 1. \tag{3.1.71}$$

Dividing by ϱ^i on both sides of (3.1.70) and taking $i \to \infty$, and using the facts that the l-subdifferential is outer semicontinuous [47, Proposition 8.7], and $\varsigma^{k_i} \to 0$ as $i \to \infty$, we derive that

$$0 \in J\mathcal{C}_1(\bar{\mathbf{s}})^\top \bar{\xi} + \partial\Big(\bar{\zeta}^\top \mathcal{C}_2(\bar{\mathbf{s}})\Big). \tag{3.1.72}$$

Combining (3.1.72) and Lemma 3.1 yields that $\bar{\xi} = 0$ and $\bar{\zeta} = 0$, which contradicts (3.1.71). 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 (3.1.70), 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 (3.1.6). □

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 local Lipschitz continuity of the ReLU function is used in $L_2(\xi, \zeta, \gamma, \hat{r})$ of Lemma 3.4 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 (3.1.35) of the BCD method in Algorithm 2.

P2. Equation (3.1.69) for proving (3.1.68) in the proof of Theorem 3.6.

For P1, even if the activation function in (3.1.1) is replaced by others, the objective function in problem (3.1.35) can still be separated into rT one-dimensional functions, which is obtained by substituting the ReLU function $(u)_+$ in (3.1.38) by a more general activation function. For P2, if an arbitrary smooth activation function is considered, then (4.29) holds obviously because the l-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 3.4 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

$$\min_{u\in\mathbb{R}} \ \varphi_{\rm lRe}(u) := \tfrac{\gamma}{2}(u-\theta_1)^2 + \tfrac{\gamma}{2}(\theta_2 - \sigma_{\rm lRe}(u))^2 + \tfrac{\beta}{2}(u-\theta_3)^2 + \lambda_6 u^2 \tag{3.1.73}$$

can be obtained similarly using the procedures for ReLU in (3.1.39)-(3.1.41), except that the expression u^- of (3.1.41) changes to

$$u^- = \begin{cases} \dfrac{\gamma\theta_1 + \gamma\varpi\theta_2 + \beta\theta_3}{\gamma + \gamma\varpi^2 + 2\lambda_6 + \beta}, & \text{if } \gamma\theta_1 + \beta\theta_3 < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.1.74}$$

For P2, (3.1.69) is modified as follows: when $\mathbf{u}_j^{k_i} = 0$,

$$\begin{aligned} &\partial_{\mathbf{u}_j}\Big(\zeta_j^{k_i-1}(\mathbf{h}_j^{k_i} - \sigma_{\rm lRe}(\mathbf{u}_j^{k_i})) + \tfrac{\gamma_{k_i-1}}{2}(\mathbf{h}_j^{k_i} - \sigma_{\rm lRe}(\mathbf{u}_j^{k_i}))^2\Big) \\ &= \begin{cases} \{-\varpi\zeta_j^{k_i}, -\zeta_j^{k_i-1} - \gamma_{k_i-1}(\mathbf{h}_j^{k_i} - \mathbf{u}_j^{k_i})\}, & \text{if } \gamma_{k_i-1}\mathbf{h}_j^{k_i} + \zeta_j^{k_i-1} \geq 0, \\ \big[-\varpi\zeta_j^{k_i}, -\zeta_j^{k_i-1} - \gamma_{k_i-1}(\mathbf{h}_j^{k_i} - \mathbf{u}_j^{k_i})\big], & \text{if } \gamma_{k_i-1}\mathbf{h}_j^{k_i} + \zeta_j^{k_i-1} < 0, \end{cases} \\ &= \begin{cases} \{-\varpi\zeta_j^{k_i}, -\zeta_j^{k_i}\}, & \text{if } \zeta_j^{k_i} \geq 0, \\ \big[-\varpi\zeta_j^{k_i}, -\zeta_j^{k_i}\big], & \text{if } \zeta_j^{k_i} < 0, \end{cases} \\ &= \partial_{\mathbf{u}_j}\Big(\zeta_j^{k_i}(\mathbf{h}_j^{k_i} - \sigma_{\rm lRe}(\mathbf{u}_j^{k_i}))\Big). \end{aligned} \tag{3.1.75}$$

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

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

When $u \geq 0$, the ELU activation function is the same as the ReLU function. Thus for P1, the solution of (3.1.73) 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

$$\min_{u\in(-\infty,0]} \ \varphi_{\mathrm{ELU}}(u) := \tfrac{\gamma}{2}(u-\theta_1)^2 + \tfrac{\gamma}{2}(\theta_2 - (e^u-1))^2 + \tfrac{\beta}{2}(u-\theta_3)^2 + \lambda_6 u^2, \quad (3.1.76)$$

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

Now we illustrate that u^- can be obtained numerically by 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 (3.1.76) is equivalent to

$$\min_{u\in[\underline{u},0]} \varphi_{\mathrm{ELU}}(u). \quad (3.1.77)$$

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

$$\varphi''_{\mathrm{ELU}}(u) = 2\gamma e^{2u} - \gamma(\theta_2+1)e^u + \beta + \gamma + 2\lambda_6. \quad (3.1.78)$$

Let $z = e^u$. (3.1.78) can be represented as

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

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 (3.1.76).

3.1.4 Numerical experiments

We employ a real-world dataset, **Volatility of S&P index**, and synthetic datasets to evaluate the effectiveness of our reformulation (3.1.6) and Algorithm 1 with Algorithm 2. To be specific, we first use RNNs with unknown weight 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 (3.1.6). 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.

Datasets

The process of generating synthetic datasets is as follows. We randomly generate the weight matrices $\hat{A}$, $\hat{W}$, $\hat{V}$, the bias vectors $\hat{b}$, $\hat{c}$, and the noise $\tilde{e}_t$, $t = 1, 2, ..., T$, 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 in Table

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

Table 3.1: 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	$\mathcal{N}(0, 0.05)$	$\mathcal{N}(0, 10^{-5})$	$\mathcal{U}(-1, 1)$

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 [9] 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.

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

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

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

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.

Investigating the feasibility

In this subsection, we aim to verify the outputs from the ALM satisfying the constraints of (3.1.2) 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 3.2.

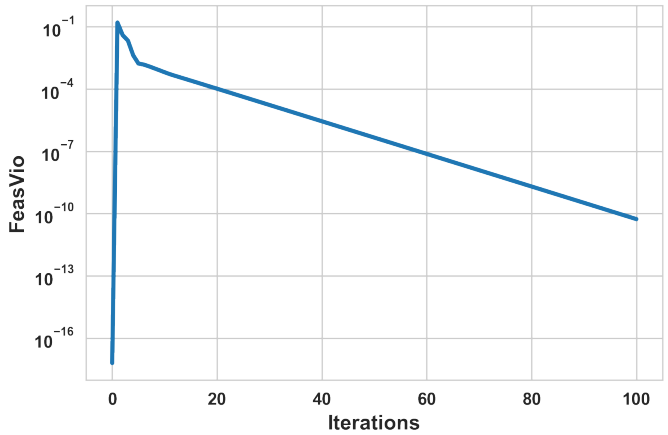
Table 3.2: Parameters of the ALM: the parameters for the given datasets are set as $\gamma^0 = 1$, $\xi^0 = \mathbf{0}$, $\zeta^0 = \mathbf{0}$, $\epsilon_0 = 0.1$, $\Gamma = 10^2$, $\beta = 10^{-5}$, $\lambda_1 = \tau/rm$, $\lambda_2 = \tau/r^2$, $\lambda_3 = \tau/rn$, $\lambda_4 = \tau/r$, $\lambda_5 = \tau/m$, $\lambda_6 = 10^{-8}$.

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 ($T = 500$)	$\tau = 500$	$\eta_1 = 0.90$, $\eta_2 = 0.90$, $\eta_3 = 0.015$, $\eta_4 = 0.8$.

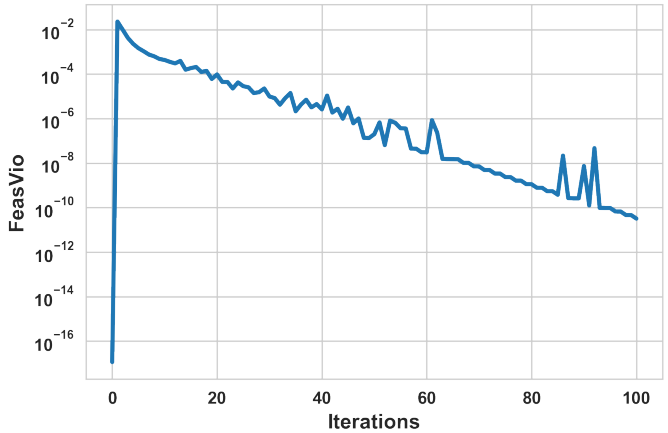
From Figure 3.1, we observe that the feasibility violation in each dataset is very small at the beginning, which implies that the selected initial point is feasible. As it turns to the first iteration, the feasibility violation goes to a large value. After that, the value goes to exhibit an oscillatory 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.

Figure 3.1: The feasibility violation of the ALM in different datasets

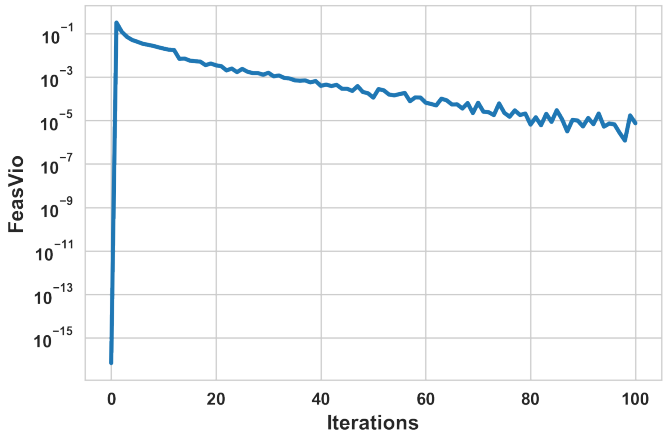
(a) Synthetic dataset ($T = 10$)



(b) Volatility of S&P index



(c) Synthetic dataset ($T = 500$)



Comparisons with state-of-the-art GDs

In this subsection, we compare the training and forecasting accuracy of RNNs using different methods. Specifically, 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 [24], Glorot initialization [20], and LeCun initialization [29]. Notably, the initial values of bias, b^0 and c^0, were both set to 0 according to [21, p. 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 3.3. 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 [15] running on TensorFlow 2 to implement the GDs and SGDs. Additionally, the parameters for the ALM are listed in Table 3.2.

Remark 3.3. *The values of regularization parameters λ_i, $i \in [6]$, influence the performance of the model. If these parameters are set too small, the regularization term fails to take effect, which may result in overfitting. Furthermore, small regularization parameters may result in large norms of the variables, making the solutions of subproblems in the BCD method more challenging. On the other hand, if the*

regularization parameters are set too large, the method focuses primarily on minimizing $p(\cdot)$, causing the norm of variables to become excessively small, which results in underfitting. Since the regularization parameters are primarily used to control the norms of variables, their selection strategy is closely related to the dimensions of the corresponding variables.

Table 3.3: 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.

		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

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 3.4.

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 3.4a and Table 3.4c 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 3.4b 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 3.4: Results of training 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	**0.113 ± 0.03**	0.143 ± 0.04	0.206 ± 0.10	0.279 ± 0.22
	TestErr	4.770 ± 1.25	**4.437 ± 0.28**	4.660 ± 0.35	4.628 ± 1.17	4.650 ± 0.62
GD	**TrainErr**	4.459 ± 0.77	$2.747 \pm$ 1.5e-6	2.768 ± 0.01	1.814 ± 0.27	**1.604 ± 0.17**
	TestErr	6.432 ± 2.15	$5.311 \pm$ 9.3e-6	5.057 ± 0.07	**4.696 ± 0.90**	5.056 ± 1.10
GDC	**TrainErr**	**1.479 ± 0.32**	$2.769 \pm$ 1.4e-6	2.768 ± 0.01	1.684 ± 0.23	1.502 ± 0.26
	TestErr	5.376 ± 0.88	$5.079 \pm$ 1.0e-6	5.057 ± 0.07	**4.922 ± 1.20**	5.266 ± 0.96
GDNM	**TrainErr**	2.689 ± 0.40	$2.769 \pm$ 1.4e-6	2.768 ± 0.01	3.340 ± 0.54	**0.801 ± 0.60**
	TestErr	6.169 ± 2.06	$5.079 \pm$ 1.0e-6	5.057 ± 0.07	7.469 ± 2.30	**4.844 ± 0.64**
SGD	**TrainErr**	**2.224 ± 0.02**	2.247 ± 0.02	2.232 ± 0.02	2.238 ± 0.02	2.225 ± 0.02
	TestErr	6.455 ± 0.23	**6.230 ± 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	**2.231 ± 0.01**	2.239 ± 0.03
	TestErr	**6.335 ± 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**	$30.142 \pm 3.5\text{e-}6$	$30.142 \pm 4.7\text{e-}6$	$\mathbf{30.142 \pm 5.2\text{e-}6}$	$30.142 \pm 4.4\text{e-}6$	$30.142 \pm 4.8\text{e-}6$
	TestErr	$\mathbf{28.725 \pm 3.2\text{e-}5}$	$28.725 \pm 4.4\text{e-}5$	$28.725 \pm 4.7\text{e-}5$	$28.725 \pm 3.9\text{e-}5$	$28.725 \pm 4.1\text{e-}5$
Adam	**TrainErr**	$\mathbf{30.142 \pm 7.1\text{e-}5}$	$30.142 \pm 6.5\text{e-}5$	$30.142 \pm 7.3\text{e-}5$	$30.142 \pm 5.1\text{e-}5$	$30.142 \pm 5.7\text{e-}5$
	TestErr	$28.726 \pm 6.1\text{e-}4$	$28.725 \pm 5.0\text{e-}4$	$28.726 \pm 5.9\text{e-}4$	$28.726 \pm 5.0\text{e-}4$	$\mathbf{28.725 \pm 4.8\text{e-}4}$

Figure 3.2: Comparisons of the performance of the ALM, GDs, and SGDs for **Volatility of S&P index**.

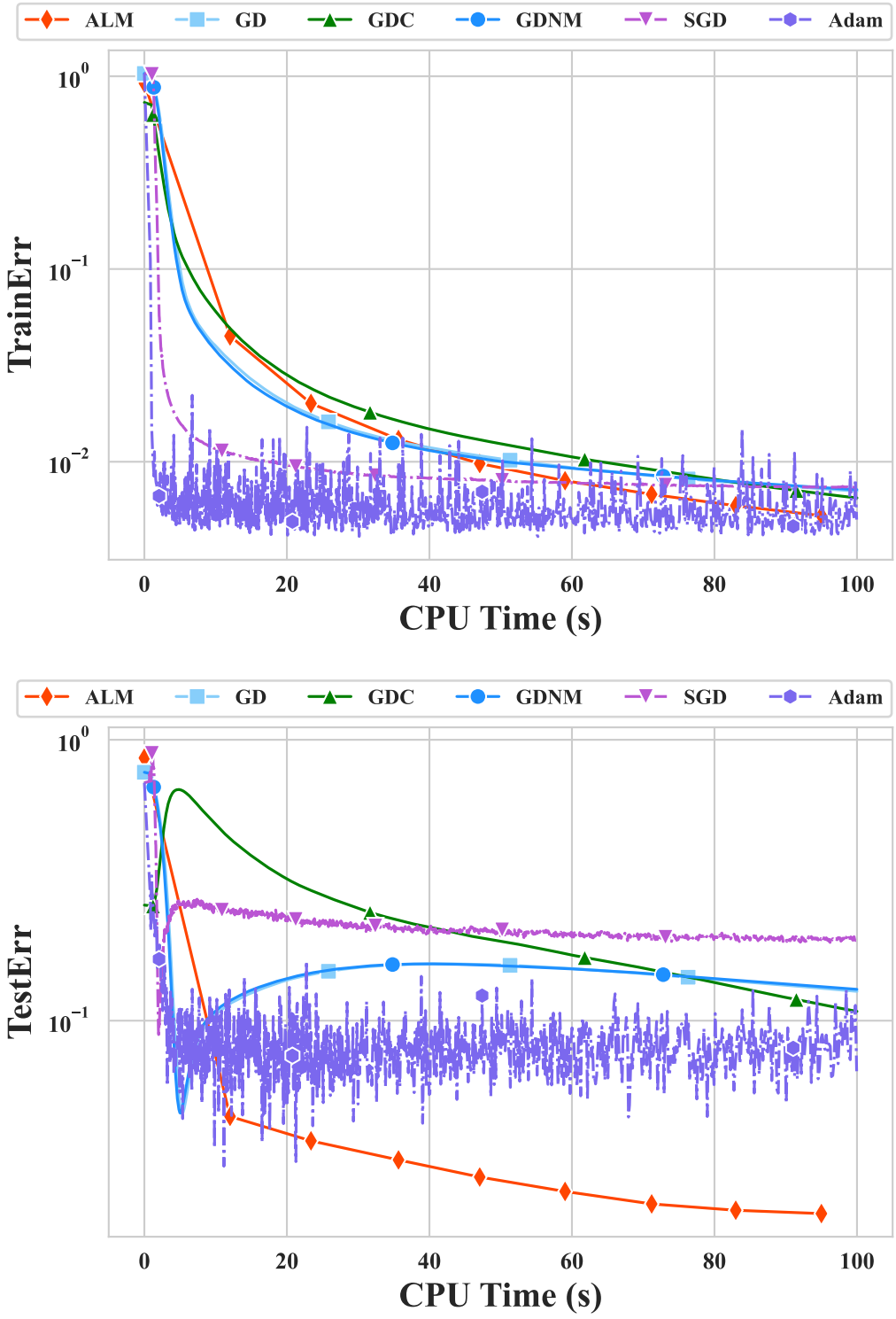
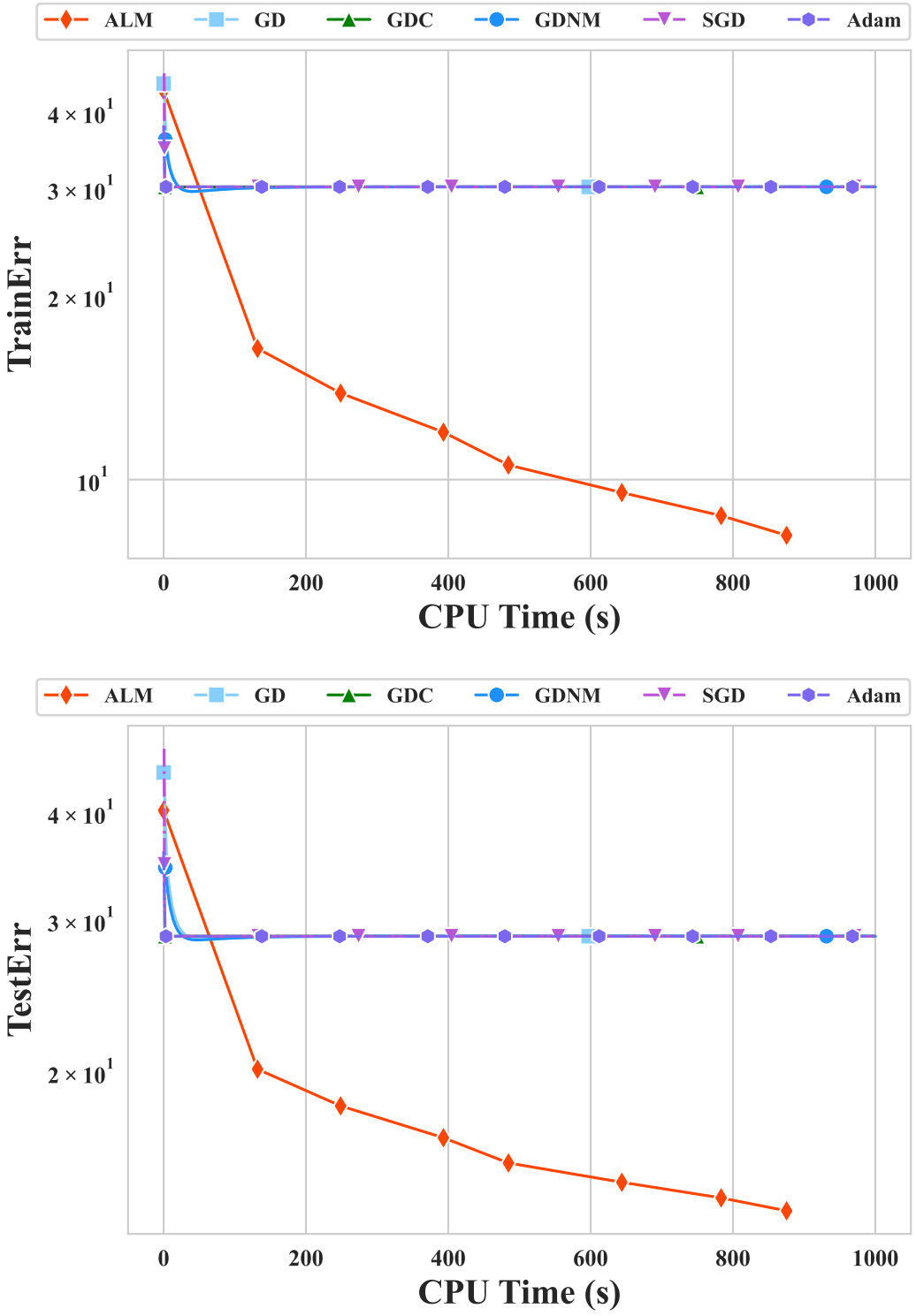


Figure 3.3: Comparisons of the performance of the ALM, GDs, and SGDs for **Synthetic dataset** ($T = 500$)



We plot in Figure 3.3 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 3.4. 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.

In this section, the minimization model (1.1.1) for training RNNs is equivalently reformulated as problem (3.1.2) by using auxiliary variables. We propose the ALM in Algorithm 1 with Algorithm 2 to solve the regularized problem (3.1.6). The BCD method in Algorithm 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 (3.1.6), 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.

3.2 ALM for problem (1.1.2)

In the previous section, we have already stated an ALM to solve problem (1.1.1) whose sample size equals one. Now, we extend the ALM to the more general case. Recall problem (1.1.2)

$$\min_{A,W,V,b,c} \ \frac{1}{NT}\sum_{t=1}^{N}\sum_{i=1}^{T}\left\| y_t^i - \Big(A\sigma\Big(W\Big(...\sigma\big(W\sigma(Vx_1^i + b) + Vx_2^i + b\big)...\Big) + Vx_t^i + b\Big) + c\Big)\right\|^2 .$$

We follow the similar procedure as in section 3.1 to deal with problem (1.1.1),

which involves utilizing auxiliary variables to represent the composite structure in the objective function and then setting them as constraints, adding a regularization term to the objective function, and employing the ALM to solve the modified constrained problem. The auxiliary variables for (1.1.2) are set as follows:

$$\mathbf{u} := (\mathbf{u}^1; ...; \mathbf{u}^i; ...; \mathbf{u}^N) \in \mathbb{R}^{rNT},\ \mathbf{u}^i := (\mathbf{u}_1^i; \mathbf{u}_t^i; ...; \mathbf{u}_T^i) \in \mathbb{R}^{rT},\ i = 1, 2, ..., N, \quad (3.2.1)$$

$$\mathbf{h} := (\mathbf{h}^1; ...; \mathbf{h}^i; ...; \mathbf{h}^N) \in \mathbb{R}^{rNT},\ \mathbf{h}^i := (\mathbf{h}_1^i; \mathbf{h}_t^i; ...; \mathbf{h}_T^i) \in \mathbb{R}^{rT},\ i = 1, 2, ..., N, \quad (3.2.2)$$

and A, W, V, b and c are vectorized by the same method as section 3.1, i.e.,

$$\mathbf{w} = (\mathrm{vec}(W); \mathrm{vec}(V); b),\ \mathbf{a} = (\mathrm{vec}(A); c),\ \mathbf{z} = (\mathbf{w}; \mathbf{a}), \quad (3.2.3)$$

$$\mathbf{s} = (\mathbf{z}; \mathbf{h}; \mathbf{u}) \in \mathbb{R}^{d+2rNT}. \quad (3.2.4)$$

To guarantee the existence of solutions, we still add the regularization term $p(\cdot)$ in (3.1.5) to the objective function. Now, the problem under consideration is the following:

$$\begin{aligned} \min_{\mathbf{s}} \quad & \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\| y_t^i - (A\mathbf{h}_t^i + c) \right\|^2 + p(\mathbf{z}) \\ \text{s.t.} \quad & \mathbf{u}_t^i = W\mathbf{h}_{t-1}^i + Vx_t^i + b, \\ & \mathbf{h}_0^i = 0,\ \mathbf{h}_t^i = (\mathbf{u}_t^i)_+,\ t \in [T],\ i \in [N]. \end{aligned} \quad (3.2.5)$$

To further simplification, the problem becomes

$$\begin{aligned} \min_{\mathbf{s}} \quad & \mathcal{R}(\mathbf{s}) := L(\mathbf{s}) + p(\mathbf{z}) \\ \text{s.t.} \quad & \mathcal{C}_1(\mathbf{z}) = 0,\ \mathcal{C}_2(\mathbf{z}) = 0, \end{aligned} \quad (3.2.6)$$

where

$$L(\mathbf{s}) := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|y_t^i - \Phi(\mathbf{h}_t^i)\mathbf{a}\right\|^2, \tag{3.2.7}$$

$$\mathcal{C}_1(\mathbf{z}) := \mathbf{u} - \Lambda(\mathbf{h})\mathbf{w},\ \mathcal{C}_2(\mathbf{z}) := \mathbf{h} - (\mathbf{u})_+, \tag{3.2.8}$$

$$\Phi(\mathbf{h}_t^i) = \left[(\mathbf{h}_t^i)^\top \otimes I_m \quad I_m\right],\ \Lambda(\mathbf{h}) = \Big(\Psi(\mathbf{h}^1); \Psi(\mathbf{h}^2); ...; \Psi(\mathbf{h}^N)\Big), \tag{3.2.9}$$

$$\Psi(\mathbf{h}^i) = \begin{bmatrix} 0_r^\top \otimes I_r & (x_1^i)^\top \otimes I_r & I_r \\ (\mathbf{h}_1^i)^\top \otimes I_r & (x_2^i)^\top \otimes I_r & I_r \\ \vdots & \vdots & \vdots \\ (\mathbf{h}_{T-1}^i)^\top \otimes I_r & (x_T^i)^\top \otimes I_r & I_r \end{bmatrix}. \tag{3.2.10}$$

3.2.1 ALM with BCD method for (3.2.6)

The augmented Lagrangian (AL) function of problem (3.2.6) is

$$\begin{aligned} &\mathcal{L}(\mathbf{s}, \varsigma, \zeta, \gamma) \\ :=\ &\mathcal{R}(\mathbf{s}) + \langle \varsigma, \mathbf{u} - \Lambda(\mathbf{h})\mathbf{w}\rangle + \langle \zeta, \mathbf{h} - (\mathbf{u})_+\rangle + \tfrac{\gamma}{2}\left\|\mathbf{u} - \Lambda(\mathbf{h})\mathbf{w}\right\|^2 + \tfrac{\gamma}{2}\left\|\mathbf{h} - (\mathbf{u})_+\right\|^2 \\ =\ &\mathcal{R}(\mathbf{s}) + \frac{\gamma}{2}\left\|\mathbf{u} - \Lambda(\mathbf{h})\mathbf{w} + \frac{\varsigma}{\gamma}\right\|^2 + \frac{\gamma}{2}\left\|\mathbf{h} - (\mathbf{u})_+ + \frac{\zeta}{\gamma}\right\|^2 - \frac{\|\varsigma\|^2}{2\gamma} - \frac{\|\zeta\|^2}{2\gamma}, \end{aligned} \tag{3.2.11}$$

where $\varsigma = (\varsigma^1; ...; \varsigma^i; ...; \varsigma^N) \in \mathbb{R}^{rNT}$, $\varsigma^i = (\varsigma_1^i; \varsigma_2^i; ...; \varsigma_T^i)$, $\zeta = (\zeta^1; ...; \zeta^i; ...; \zeta^T) \in \mathbb{R}^{rNT}$ and $\zeta^i = (\zeta_1^i; \zeta_2^i; ...; \zeta_T^i)$ are the Lagrangian multipliers, and $\gamma > 0$ is the penalty parameter for the two quadratic penalty terms of constraints $\mathbf{u} = \Lambda(\mathbf{h})\mathbf{w}$ and $\mathbf{u} = (\mathbf{h})_+$.

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

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

where

$$\hat{Q}_1(\mathbf{s},\varsigma,\zeta,\gamma)=\gamma\sum_{i=1}^{N}\Psi(\mathbf{h}^i)^\top\Psi(\mathbf{h}^i)+2\Lambda_1,\quad \hat{q}_1(\mathbf{s},\varsigma,\zeta,\gamma)=-\sum_{i=1}^{N}\Psi(\mathbf{h}^i)^\top(\varsigma^i+\gamma\mathbf{u}^i),$$

$$\hat{Q}_2(\mathbf{s},\varsigma,\zeta,\gamma)=\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Phi(\mathbf{h}_t^i)^\top\Phi(\mathbf{h}_t^i)+2\Lambda_2,\quad \hat{q}_2(\mathbf{s},\varsigma,\zeta,\gamma)=-\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Phi(\mathbf{h}_t^i)^\top y_t^i,$$

$$\Lambda_1=\mathrm{diag}\Big(\big(\lambda_2\boldsymbol{e}_{r2};\lambda_3\boldsymbol{e}_{rn};\lambda_4\boldsymbol{e}_r\big)\Big),\quad \Lambda_2=\mathrm{diag}\Big(\big(\lambda_1\boldsymbol{e}_{rm};\lambda_5\boldsymbol{e}_m\big)\Big).$$

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

$$\begin{aligned}&\nabla_{\mathbf{h}}\mathcal{L}(\mathbf{s},\varsigma,\zeta,\gamma)\\=\;&\big(\nabla_{\mathbf{h}^1}\mathcal{L}(\mathbf{s},\varsigma,\zeta,\gamma);\ldots;\nabla_{\mathbf{h}^i}\mathcal{L}(\mathbf{s},\varsigma,\zeta,\gamma);\ldots;\nabla_{\mathbf{h}^N}\mathcal{L}(\mathbf{s},\varsigma,\zeta,\gamma)\big),\end{aligned}$$

where

$$\begin{aligned}&\nabla_{\mathbf{h}^i}\mathcal{L}(\mathbf{s},\varsigma,\zeta,\gamma)\\=\;&\big(\nabla_{\mathbf{h}_1^i}\mathcal{L}(\mathbf{s},\varsigma^i,\zeta^i,\gamma);\nabla_{\mathbf{h}_2^i}\mathcal{L}(\mathbf{s},\varsigma^i,\zeta^i,\gamma);\ldots;\nabla_{\mathbf{h}_T^i}\mathcal{L}(\mathbf{s},\varsigma^i,\zeta^i,\gamma)\big),\end{aligned}$$

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

$$D_1(z,\mathbf{u}^i,\mathbf{h}^i,\varsigma^i,\zeta^i,\gamma)=\gamma W^\top W+\tfrac{2}{NT}A^\top A+\gamma I_r,$$

$$D_2(z,\mathbf{u}^i,\mathbf{h}^i,\varsigma^i,\zeta^i,\gamma)=\tfrac{2}{NT}A^\top A+\gamma I_r,$$

$$d_{1t}(z,\mathbf{u}^i,\mathbf{h}^i,\varsigma^i,\zeta^i,\gamma)=W^\top\big(\varsigma_t^i+\gamma(\mathbf{u}_{t+1}^i-Vx_{t+1}^i-b)\big)+\gamma(\mathbf{u}_t^i)_+-\zeta_t^i+\tfrac{2}{NT}A^\top(y_t^i-c),$$

$$d_{2T}(z,\mathbf{u}^i,\mathbf{h}^i,\varsigma^i,\zeta^i,\gamma)=\gamma(\mathbf{u}_T^i)_+-\zeta_T^i+\tfrac{2}{NT}A^\top(y_T^i-c).$$

The objective function of problem (3.2.6) can be separated into rNT one-dimensional functions with the same structure. Thus, we aim to solve the following one-dimensional problem:

$$\min_{\mathbf{u}\in\mathbb{R}}\ \varphi(\mathbf{u}):=\tfrac{\gamma}{2}(\mathbf{u}-\theta_1)^2+\tfrac{\gamma}{2}(\theta_2-(\mathbf{u})_+)^2+\tfrac{\beta}{2}(\mathbf{u}-\theta_3)^2+\lambda_6\mathbf{u}^2,\qquad(3.2.12)$$

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

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

By direct computation,

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

$$\mathbf{u}^- = \begin{cases} \dfrac{\gamma\theta_1 + \beta\theta_3}{\gamma + 2\lambda_6 + \beta}, & \text{if } \gamma_{k-1}\theta_1 + \beta\theta_3 < 0, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\theta_1 = [\Psi(h^i)\mathbf{w}]_k - \frac{[\zeta^i]_k}{\gamma}, \quad \theta_2 = [h^i]_k + \frac{[\zeta^i]_k}{\gamma}, \quad \theta_3 = [u^i]_k.$$

Then a solution of (3.2.12) can be given as

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

The detailed method to solve (3.2.6) is listed in Algorithm 3 and Algorithm 4.

Compared to problem (3.1.2), problem (3.2.6) has N times more constraints. However, the expressions of additional constraints are similar to $N = 1$. Therefore, Algorithm 3 and Algorithm 4 are not fundamentally different from Algorithm 1 and Algorithm 2. Hence, the convergence analysis for Algorithm 3 and Algorithm 4 follow analogous approaches to those of Algorithm 1 and Algorithm 2, and we omit the detailed discussion here.

Algorithm 3 The augmented Lagrangian method (ALM) for (3.2.6)

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 ς^0, ζ^0, and a feasible initial point $\mathbf{s}^0 = (z^0, \hat{\mathbf{u}}, \hat{\mathbf{h}})$ where $\hat{\mathbf{h}}_0 = 0$, $\hat{\mathbf{u}}_t = W^0\hat{\mathbf{h}}_{t-1} + V^0 x_t + b^0$ and $\hat{\mathbf{h}}_t = (\hat{\mathbf{u}}_t)_+$ for $t \in [T]$.

2: Set $k := 1$.

3: **Step 1:** Solve

$$\min_{\mathbf{s}} \quad \mathcal{L}(\mathbf{s}, \varsigma^{k-1}, \zeta^{k-1}, \gamma_{k-1}) \tag{3.2.13}$$

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

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

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

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

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

$$\max\left\{\|\mathcal{C}_1(\mathbf{s}^k)\|, \|\mathcal{C}_2(\mathbf{s}^k)\|\right\} \leq \eta_1 \max\left\{\|\mathcal{C}_1(\mathbf{s}^{k-1})\|, \|\mathcal{C}_2(\mathbf{s}^{k-1})\|\right\}. \tag{3.2.14}$$

Otherwise, set

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

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

3.2.2 Numerical experiments

In subsection 3.1.4, we have demonstrated that the ALM exhibits superior performance in solving optimization problems when applying RNNs to model time series forecasting tasks. Problem (3.2.5) is the optimization problem in using RNNs to model more general sequential regression tasks, such as the image denoising and the audio denoising.

Algorithm 4 Block Coordinate Descent (BCD) method for (3.2.13)

Set the initial point of the BCD algorithm as

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

Compute $\hat{r}_{k-1} = \mathcal{L}(\mathbf{s}^{k-1,0}, \varsigma, \zeta, \gamma)$, $L_{1,k-1} = L_1(\varsigma, \zeta, \gamma, \hat{r}_{k-1})$ and $L_{2,k-1} = L_2(\varsigma, \zeta, \gamma, \hat{r}_{k-1})$ by formula (3.4) in [57].

Set $j := 1$.

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

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

$$z^{k-1,j} = \arg\min_{z} \mathcal{L}\left(z, \mathbf{h}^{k-1,j-1}, \mathbf{u}^{k-1,j-1}, \varsigma, \zeta, \gamma\right), \tag{3.2.15}$$

$$\mathbf{h}^{k-1,j} = \arg\min_{\mathbf{h}} \mathcal{L}\left(z^{k-1,j}, \mathbf{h}, \mathbf{u}^{k-1,j-1}, \varsigma, \zeta, \gamma\right), \tag{3.2.16}$$

$$\mathbf{u}^{k-1,j} \in \arg\min_{\mathbf{u}} \mathcal{L}\left(z^{k-1,j}, \mathbf{h}^{k-1,j}, \mathbf{u}, \varsigma, \zeta, \gamma\right) + \tfrac{\beta}{2}\left\|\mathbf{u} - \mathbf{u}^{k-1,j-1}\right\|^2. \tag{3.2.17}$$

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

Step 2: If the stop criterion

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

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

end while

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

Datasets

Synthetic dataset: The synthetic dataset is generated by the following method. The weight matrices and bias $\hat{A}$, $\hat{W}$, $\hat{V}$, $\hat{b}$, $\hat{c}$ are randomly generated where each component of them follows the normal distribution $\mathcal{N}(0, 0.8)$. And for each sample of input data, $X^i = (x_1^i, ..., x_T^i) \in \mathbb{R}^{n \times T}$, $i = 1, 2, ..., N$, it is generated following the multivariate normal distribution $\mathcal{N}(\mathbf{0}_{nT}, \Sigma)$, where the covariance matrix $\Sigma \in \mathbb{R}^{nT \times nT}$ is a randomly generated positive-definite matrix. Then, corresponding output data $Y^i = (y_1^i, ..., y_T^i)$ is calculated by $y_t^i = (\hat{A}(\hat{W}(...(\hat{V}x_1^i + \hat{b})_+...) + \hat{V}x_t^i + \hat{b})_+ + \hat{c}) + \tilde{e}_t$ for $t \in \mathcal{T}$, where each component of $\tilde{e}_t$ follows the normal distribution $\mathcal{N}(0, 10^{-3})$.

Repeating the generating process of input and output data N times, we have the synthetic dataset. The N sample points are divided into training and test sets following the ratio $7:3$. The specific dimensions of the synthetic dataset are following: $n = 5$, $m = 3$, $r = 4$, $T = 10$ and $N = 100$.

Pixel-MNIST: The dataset, Pixel-MNIST, is a commonly used toy example for RNNs, which is modified from the general MNIST handwritten digit database [33]. To be specific, we flattened each 28×28 pixel matrix from left to right and top to bottom, resulting in a sequence of 784 pixels. We then normalized the data. After that, we added noise following a normal distribution $\mathcal{N}(0, 1)$ to the sequential data and used it as the input. Consequently, we employed a RNN model with $n = 1$, $m = 1$, $r = 5$ and $T = 784$ to perform the denoising task on the sequential Pixel-MNIST dataset. The total sample size is $N = 7 \times 10^4$, which is divided into training and test sets in a $6:1$ ratio.

TIMIT Audio: TIMIT [19] is a widely used audio dataset for speech recognition and denoising tasks. It contains 6,300 sentences, which are read by 630 speakers from 8 major dialects of American English and encoded at a sampling rate of 16 KHz. In this work, we use 90 of these sentences from 14 speakers as the original clean dataset and randomly add noise audio at a sampling rate of 16 kHz derived from airport, train, subway, babble, or drill environments to create a mixed noisy dataset. It is worth mentioning that we input frequency features of the audio samples to the RNNs, rather than the temporal waveforms directly, because frequency features can effectively reduce computational costs. For both the original clean dataset and the mixed noisy dataset, we extract the spectrograms of the audio samples using the Short-Time Fourier Transform (STFT) with an FFT size of 512, a hop size of 214, and Hann windows. The spectrograms serve as the frequency features of the audio. After the above process, we input the frequency features of the mixed noisy dataset into the RNNs, use the model to perform denoising, and approximate a clean

dataset with dimensions $n = 257$, $m = 257$, $r = 200$, and $T = 173$. The total sample size is $N = 140$, and we divide the data into training and test sets with a 7:3 ratio.

Comparisons with state-of-the-art GDs

This section aims to show that the ALM method is good at solving problem (1.1.2) compared with state-of-the-art methods, i.e., gradient descent with gradient clipping (GDC) and Adaptive Moment Estimation (Adam). All the numerical experiments were conducted using Python 3.9.8 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.

We employ the following **TrainErr** and **TestErr** to measure the performance of training sets and test sets in different methods:

$$\textbf{TrainErr} := \frac{1}{N_1 T}\sum_{i=1}^{N_1}\sum_{t=1}^{T}\left\|\boldsymbol{y}_t^i - \left(A\Big(W\Big(...\big(W(V\boldsymbol{x}_1^i+b)_+ + V\boldsymbol{x}_2^i+b\big)_+...\Big)+V\boldsymbol{x}_t^i+b\Big)_+ + c\right)\right\|^2,$$

$$\textbf{TestErr} := \frac{1}{N_2 T}\sum_{i=1}^{N_2}\sum_{t=1}^{T}\left\|\boldsymbol{y}_t^i - \left(A\Big(W\Big(...\big(W(V\boldsymbol{x}_1^i+b)_+ + V\boldsymbol{x}_2^i+b\big)_+...\Big)+V\boldsymbol{x}_t^i+b\Big)_+ + c\right)\right\|^2,$$

where N_1 and N_2 are the sample size of the training sets and test sets, and A, W, V, b and c are the output solutions from the ALM.

The initialization strategies of variables are the same as those in section 3.1. Furthermore, the methods for selecting hyper-parameters in GDs and SGDs are also consistent with those outlined in section 3.1 and the results are listed in Table 3.5. It is worth mentioning that the cross-validation values of He initialization, Glorot initialization and LeCun initialization may exceed 10^3 as selecting proper hyper-parameters. Therefore, these three initialization strategies are excluded in the following analysis. Moreover, The parameters for the ALM in Algorithm 2 and 3 are provided in Table 3.6.

Table 3.5: 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.

		$\mathcal{N}(0, 10^{-3})$	$\mathcal{N}(0, 10^{-2})$	$\mathcal{N}(0, 10^{-1})$
GD	**Synthetic dataset**	1e-3	1e-3	1e-3
	Pixel-MNIST	0.1	0.1	0.1
	TIMIT Audio	0.1	0.1	0.01
GDC	**Synthetic dataset**	1e-3 (1)	1e-3 (1)	0.1 (1)
	Pixel-MNIST	0.1(1)	0.1(1)	0.1(1)
	TIMIT Audio	1 (0.5)	0.1 (0.5)	1 (0.5)
GDNM	**Synthetic dataset**	0.1	0.1	0.1
	Pixel-MNIST	0.1	0.1	0.1
	TIMIT Audio	1	0.1	0.01
SGD	**Synthetic dataset**	0.1	0.1	0.1
	Pixel-MNIST	1e-3	1e-3	1e-3
	TIMIT Audio	0.1	0.1	0.1
Adam	**Synthetic dataset**	0.1	0.1	0.1
	Pixel-MNIST	1e-3	1e-3	1e-3
	TIMIT Audio	0.01	0.01	0.01

To assess the performance of various methods under different initialization strategies, we conducted a series of experiments, with each method being executed 5 times for each initialization strategy. In each trial, we recorded both the final test error and the training error. Subsequently, we computed their means along with the corresponding standard deviations, the results of which are presented in Table 3.7. The notation rules in the table are the same as those in Table 3.4.

Table 3.7 show that as the initialization strategy is selected as $\mathcal{N}(0, 10^{-1})$, ALM achieves the best **TrainErr** and **TestErr** among all combinations of optimization methods and initialization strategies for all three datasets, which we highlight in blue.

Table 3.6: Parameters for the ALM: the parameters for the given datasets are set as $\gamma^0 = 1$, $\varsigma^0 = \mathbf{0}$, $\zeta^0 = \mathbf{0}$, $\epsilon_0 = 0.1$, $\Gamma = 10^2$, $\beta = 10^{-5}$, $\lambda_1 = \tau/rm$, $\lambda_2 = \tau/r^2$, $\lambda_3 = \tau/rn$, $\lambda_4 = \tau/r$, $\lambda_5 = \tau/m$, $\lambda_6 = 10^{-8}$.

Datasets	Regularization parameters τ	Algorithm parameters
Synthetic dataset	5×10^{-2}	$\eta_1 = 0.99$, $\eta_2 = 5/6$, $\eta_3 = 0.01$, $\eta_4 = 5/6$.
Pixel-MNIST	0.25	$\eta_1 = 0.90$, $\eta_2 = 0.70$, $\eta_3 = 0.02$, $\eta_4 = 0.7$.
TIMIT Audio	5×10^{-4}	$\eta_1 = 0.90$, $\eta_2 = 0.70$, $\eta_3 = 0.02$, $\eta_4 = 0.7$.

Table 3.7: Results of training RNNs using different optimization methods and initialization strategies across multiple trials.

(a) **Synthetic dataset**: For the ALM method, the maximum iteration for the outer loop is 10 and 50 for the inner loop. For GDs and SGDs, the number of epochs is set to 2000. The batch size for SGDs is set to 20.

		$\mathcal{N}(0, 10^{-3})$	$\mathcal{N}(0, 10^{-2})$	$\mathcal{N}(0, 10^{-1})$
ALM	**TrainErr**	0.629 ± 0.05	0.827 ± 0.30	$\mathbf{0.502 \pm 0.08}$
	TestErr	0.685 ± 0.05	0.871 ± 0.29	$\mathbf{0.548 \pm 0.08}$
GD	**TrainErr**	$3.000 \pm$ 3.3e-6	$2.997 \pm$ 1.3e-3	$\mathbf{2.673 \pm 0.24}$
	TestErr	$3.000 \pm$ 3.3e-6	$2.997 \pm$ 1.3e-3	$\mathbf{2.698 \pm 0.21}$
GDC	**TrainErr**	$3.000 \pm$ 3.3e-6	$2.997 \pm$ 1.3e-3	$\mathbf{0.696 \pm 0.08}$
	TestErr	$3.000 \pm$ 3.3e-6	$2.997 \pm$ 1.3e-3	$\mathbf{0.767 \pm 0.09}$
GDNM	**TrainErr**	$\mathbf{0.657 \pm 0.01}$	$2.997 \pm$ 1.3e-3	0.685 ± 0.09
	TestErr	$\mathbf{0.726 \pm 0.01}$	$2.997 \pm$ 1.3e-3	0.754 ± 0.09
SGD	**TrainErr**	$\mathbf{0.617 \pm 0.01}$	2.027 ± 0.36	0.685 ± 0.10
	TestErr	$\mathbf{0.678 \pm 0.02}$	2.089 ± 0.37	0.751 ± 0.11
Adam	**TrainErr**	$\mathbf{0.621 \pm 0.01}$	1.206 ± 0.45	0.626 ± 0.01
	TestErr	$\mathbf{0.690 \pm 0.01}$	1.248 ± 0.43	0.692 ± 0.01

(b) **Pixel-MNIST**: For the ALM method, the maximum iteration for the outer loop is 100 and 200 for the inner loop. For GDs and SGDs, the number of epochs is set to 1000. The batch size for SGDs is set to 500.

		$\mathcal{N}(0, 10^{-3})$	$\mathcal{N}(0, 10^{-2})$	$\mathcal{N}(0, 10^{-1})$
ALM	**TrainErr**	0.092 ± 2.9e-3	0.093 ± 2.4e-3	**0.092 ± 1.1e-3**
	TestErr	0.100 ± 3.1e-3	0.100 ± 2.6e-3	**0.099 ± 1.2e-3**
GD	**TrainErr**	0.095 ± 1.1e-9	**0.095 ± 6.5e-10**	0.095 ± 8.9e-10
	TestErr	0.102 ± 1.6e-9	0.102 ± 2.3e-9	**0.102 ± 1.3e-9**
GDC	**TrainErr**	0.095 ± 1.2e-9	**0.095 ± 6.3e-10**	0.095 ± 8.9e-10
	TestErr	0.102 ± 2.2e-9	0.102 ± 3.2e-9	**0.102 ± 1.3e-9**
GDNM	**TrainErr**	0.095 ± 1.2e-9	**0.095 ± 8.7e-10**	0.095 ± 8.9e-10
	TestErr	0.102 ± 3.3e-9	0.102 ± 2.5e-9	**0.102 ± 1.3e-9**
SGD	**TrainErr**	0.095 ± 6.9e-4	**0.094 ± 8.9e-4**	0.095 ± 5.1e-4
	TestErr	0.102 ± 7.5e-4	**0.101 ± 9.7e-4**	0.102 ± 5.5e-4
Adam	**TrainErr**	0.095 ± 2.8e-8	**0.095 ± 1.2e-8**	0.095 ± 1.7e-8
	TestErr	0.102 ± 3.2e-8	**0.102 ± 1.4e-9**	0.102 ± 2.0e-8

(c) **TIMIT Audio**: For the ALM method, the maximum iteration for the outer loop is 10 and 50 for the inner loop. For GDs and SGDs, the number of epochs is set to 2000. The batch size for SGDs is set to 20.

		$\mathcal{N}(0, 10^{-3})$	$\mathcal{N}(0, 10^{-2})$	$\mathcal{N}(0, 10^{-1})$
ALM	**TrainErr**	4.788 ± 0.03	4.604 ± 5.2e-3	**4.412 ± 0.03**
	TestErr	4.387 ± 9.9e-3	4.121 ± 5.7e-3	**3.955 ± 0.03**
GD	**TrainErr**	6.742 ± 0.01	**5.006 ± 7.3e-3**	8.086 ± 0.08
	TestErr	6.021 ± 0.01	**4.472 ± 6.6e-3**	7.180 ± 0.06
GDC	**TrainErr**	5.106 ± 1.2e-4	5.006 ± 7.3e-3	**4.903 ± 0.18**
	TestErr	4.571 ± 1.3e-4	4.472 ± 6.6e-3	**4.377 ± 0.18**
GDNM	**TrainErr**	6.766 ± 0.05	**5.006 ± 7.3e-3**	8.086 ± 0.08
	TestErr	**6.057 ± 0.04**	**4.472 ± 6.6e-3**	7.180 ± 0.06
SGD	**TrainErr**	4.855 ± 8.0e-6	4.856 ± 4.1e-7	**4.855 ± 1.7e-6**
	TestErr	4.330 ± 7.6e-6	4.330 ± 3.5e-7	**4.330 ± 3.6e-6**
Adam	**TrainErr**	4.846 ± 2.1e-4	4.845 ± 8.1e-4	**4.845 ± 1.7e-3**
	TestErr	4.324 ± 1.9e-4	4.324 ± 5.2e-3	**4.324 ± 1.5e-3**

Chapter 4

SAA for Training Recurrent Neural Networks

In this chapter, we propose a method to solve (1.1.2) which is an approximation of (1.1.3) by the SAA method. Therefore, in this chapter, we aim to prove that any accumulation point of minimizers and stationary points of the SAA problems is a minimizer and a stationary point of the original problem respectively w.p.1 as the sample size goes to infinity. At the beginning of the chapter, we reformulate problems (1.1.2) and (1.1.3) to facilitate the subsequent analysis and investigate the properties of these two problems. We then establish the convergence of minimizers and stationary points of the SAA problem (1.1.2). Next, we discuss the uniform exponential rates of convergence of the objective function of the SAA problem (1.1.2) to those of the original problem (1.1.3).

4.1 Reformulations for (1.1.2) and (1.1.3), and properties of their objective functions

For simplicity of analysis, we define the nonsmooth activation function $\sigma : \mathbb{R} \to \mathbb{R}$ in (1.1.2) and (1.1.3) as the leaky ReLU function, i.e.,

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

where $\varpi \in (0, 1)$ is a fixed parameter.

4.1.1 Reformulations for (1.1.2) and (1.1.3)

We introduce the following notation:

$$\mathbf{z} = \big(\mathrm{vec}(W); \mathrm{vec}(V); b; \mathrm{vec}(A); c\big) \in \mathbb{R}^{N_{\mathbf{z}}},$$

$$\mathbf{X} = \big(\mathbf{X}_1; \mathbf{X}_2; ...; \mathbf{X}_T\big),\ \mathbf{Y} = \big(\mathbf{Y}_1; \mathbf{Y}_2; ...; \mathbf{Y}_T\big),\ \boldsymbol{\xi} = \big(\mathbf{X}; \mathbf{Y}\big), \tag{4.1.1}$$

$$m_t(\mathbf{z}, \boldsymbol{\xi}) := \sigma_{\mathrm{lRe}}\big(W m_{t-1}(\mathbf{z}, \boldsymbol{\xi}) + V\mathbf{X}_t + b\big),\ t \in [T],\ m_0(\mathbf{z}, \boldsymbol{\xi}) \equiv 0, \tag{4.1.2}$$

$$o_t(\mathbf{z}, \boldsymbol{\xi}) := A m_t(\mathbf{z}, \boldsymbol{\xi}) + c,\ \ell_t(\mathbf{z}, \boldsymbol{\xi}) := \|\mathbf{Y}_t - o_t(\mathbf{z}, \boldsymbol{\xi})\|^2,\ t \in [T], \tag{4.1.3}$$

$$\ell(\mathbf{z}, \boldsymbol{\xi}) = \frac{1}{T}\sum_{t=1}^{T} \ell_t(\mathbf{z}, \boldsymbol{\xi}), \text{‘} \tag{4.1.4}$$

where $\mathbb{R}^{N_{\mathbf{z}}} = mr + rn + r^2 + r + m$. Random vectors $\mathbf{X} : \Omega^x \to \mathcal{X} \subset \mathbb{R}^{nT}$, $\mathbf{Y} : \Omega^y \to \mathcal{Y} \subset \mathbb{R}^{mT}$ and $\boldsymbol{\xi} : \Omega^\xi \to \Xi \subset \mathbb{R}^{nT+mT}$ are defined on the probability space $(\Omega^x, \mathcal{F}^x, P_x)$, $(\Omega^y, \mathcal{F}^y, P_y)$ and $(\Omega^\xi, \mathcal{F}^\xi, P_\xi)$ respectively, where $\Omega^\xi = \Omega^x \times \Omega^y$. We assume that Ξ is compact. Moreover, $\ell : \mathbb{R}^{N_{\mathbf{z}}} \times \Xi \to \mathbb{R}$ is named as the loss function.

Then, problem (1.1.3) can be equivalently represented as

$$\min_{\mathbf{z}}\ f(\mathbf{z}) := \mathbb{E}\left[\ell(\mathbf{z}, \boldsymbol{\xi})\right], \tag{4.1.5}$$

and problem (1.1.2) is equivalently expressed as follows:

$$\min_{\mathbf{z}}\ \hat{f}_N(\mathbf{z}) := \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{z}, \xi^i). \tag{4.1.6}$$

To have bounded solution sets, we add the following regularized term:

$$p(\mathbf{z}) = \lambda_1\|A\|_F^2 + \lambda_2\|W\|_F^2 + \lambda_3\|V\|_F^2 + \lambda_4\|b\|^2 + \lambda_5\|c\|^2, \tag{4.1.7}$$

to the objective function of problems (4.1.5) and (4.1.6) with $\lambda_i > 0, i = 1, 2, \ldots, 5$. Then, the regularized problems of (4.1.5) and (4.1.6) respectively refer to

$$\min_{\mathbf{z}} \ \psi(\mathbf{z}) := f(\mathbf{z}) + p(\mathbf{z}), \tag{4.1.8}$$

and

$$\min_{\mathbf{z}} \ \hat{\psi}_N(\mathbf{z}) := \hat{f}_N(\mathbf{z}) + p(\mathbf{z}). \tag{4.1.9}$$

For the sake of analysis, we make the following assumption.

Assumption 4.1. *There exists a constant $\varrho \geq 0$ such that for any $\boldsymbol{\xi} \in \Xi$, $\|\boldsymbol{\xi}\| \leq \varrho$.*

It is worth mentioning that the assumption is assumed to be held throughout this chapter. The assumption ensures that each sample point drawn from the random variable $\boldsymbol{\xi}$ is almost surely bounded. This aligns with the fact that, in practice, data typically does not attain infinite value.

Remark 4.1. *Random variable $\boldsymbol{\xi}$ satisfying Assumption 4.1 implies that $\boldsymbol{\xi}$ is integrable.*

4.1.2 Properties of ψ and $\hat{\psi}_N$

To aid in the following analysis, we define a compact set as follows:

$$\mathcal{G}_\iota := \{\mathbf{z} \in \mathbb{R}^{N_\mathbf{z}} : \|\mathbf{z}\| \leq \iota\}, \tag{4.1.10}$$

with

$$\iota = \sqrt{\frac{\max\{\theta, \rho\}}{\min_{i\in[5]}\{\lambda_i\}}},$$

where θ and ρ are constants satisfying $\theta > \hat{\psi}_N(0)$ and $\rho > \psi(0)$.

By the expression of ψ and $\hat{\psi}_N$, we first explore properties of $\ell : \mathbb{R}^{N_{\mathbf{z}}} \times \Xi \to \mathbb{R}$ defined in (4.1.4) on $\mathcal{G}_\iota \times \Xi$, where $\mathcal{G}_\iota$ is a compact set defined in (4.1.10). We first show that for any fixed $\boldsymbol{\xi}$, $\ell(\cdot, \boldsymbol{\xi})$ is Lipschitz continuous on $\mathcal{G}_\iota$. For this purpose, we define the following two functions:

$$g_1(\mathbf{z}, \tilde{\mathbf{h}}) := \sigma_{\mathrm{lRe}}\big(W\tilde{\mathbf{h}} + Vx + b\big), \quad g_2(\mathbf{z}, \tilde{\mathbf{u}}) := A\tilde{\mathbf{u}} + c,$$

where $x \in \mathbb{R}^n$ is a fixed vector, the meaning of A, W, V, b, c, $\mathbf{z}$ are same as the above, $\tilde{\mathbf{h}}$ and $\tilde{\mathbf{u}}$ are in the bounded set $\mathcal{H} := \{\tilde{\mathbf{h}} : \|\tilde{\mathbf{h}}\| \le \tilde{a}_1\}$ and $\mathcal{U} := \{\tilde{\mathbf{u}} : \|\tilde{\mathbf{u}}\| \le \tilde{a}_2\}$ respectively. Then, we can deduce the following lemma.

Lemma 4.1. *For any $\mathbf{z}, \mathbf{z}' \in \mathcal{G}_\iota$, $\tilde{\mathbf{h}}, \tilde{\mathbf{h}}' \in \mathcal{H}$, $\tilde{\mathbf{u}}, \tilde{\mathbf{u}}' \in \mathcal{U}$, we have the following inequalities:*

$$\|g_1(\mathbf{z}, \tilde{\mathbf{h}}) - g_1(\mathbf{z}', \tilde{\mathbf{h}}')\| \le \iota\|\tilde{\mathbf{h}} - \tilde{\mathbf{h}}'\| + \max\{\tilde{a}_1, \|x\|, 1\}\|\mathbf{z} - \mathbf{z}'\|, \tag{4.1.11}$$

$$\|g_2(\mathbf{z}, \tilde{\mathbf{u}}) - g_2(\mathbf{z}', \tilde{\mathbf{u}}')\| \le \iota\|\tilde{\mathbf{u}} - \tilde{\mathbf{u}}'\| + \max\{\tilde{a}_2, 1\}\|\mathbf{z} - \mathbf{z}'\|. \tag{4.1.12}$$

Proof. It is known that the leaky ReLU function is Lipschitz continuous on compact sets with Lipschitz constant 1. Thus, by Lemma 2.1 and Lemma 4.3, we have that

$$\begin{aligned}
&\|g_1(\mathbf{z}, \tilde{\mathbf{h}}) - g_1(\mathbf{z}', \tilde{\mathbf{h}}')\| = \|\sigma_{\mathrm{lRe}}\big(Wh + Vx + b\big) - \sigma_{\mathrm{lRe}}\big(W'h' + V'x + b'\big)\| \\
&\le \|\big(W\tilde{\mathbf{h}} + Vx + b\big) - \big(W'\tilde{\mathbf{h}}' + V'x + b'\big)\| \\
&\le \|W\tilde{\mathbf{h}} - W'\tilde{\mathbf{h}} + W'\tilde{\mathbf{h}} - W'\tilde{\mathbf{h}}'\| + \|V' - V\|\|x\| + \|b' - b\| \\
&\le \|W - W'\|\|\tilde{\mathbf{h}}\| + \|W'\|\|\tilde{\mathbf{h}} - \tilde{\mathbf{h}}'\| + \|V' - V\|\|x\| + \|b' - b\| \\
&\le \iota\|\tilde{\mathbf{h}} - \tilde{\mathbf{h}}'\| + \max\{\tilde{a}_1, \|x\|, 1\}\|\mathbf{z} - \mathbf{z}'\|,
\end{aligned}$$

and

$$\begin{aligned}
&\|g_2(\mathbf{z}, \tilde{\mathbf{u}}) - g_2(\mathbf{z}', \tilde{\mathbf{u}}')\| = \|(A\tilde{\mathbf{u}} + c) - (A'\tilde{\mathbf{u}}' + c')\| \\
&= \|A\tilde{\mathbf{u}} - A'\tilde{\mathbf{u}} + A'\tilde{\mathbf{u}} - A'\tilde{\mathbf{u}}'\| + \|c - c'\| \\
&\le \|A - A'\|\|\tilde{\mathbf{u}}\| + \|A'\|\|\tilde{\mathbf{u}} - \tilde{\mathbf{u}}'\| + \|c - c'\| \\
&\le \iota\|\tilde{\mathbf{u}} - \tilde{\mathbf{u}}'\| + \max\{\tilde{a}_2, 1\}\|\mathbf{z} - \mathbf{z}'\|.
\end{aligned}$$

□

The following lemma shows the boundedness of $\|m_t(\mathbf{z}, \boldsymbol{\xi})\|$ for any $t \in [T]$.

Lemma 4.2. *For any $t \in [T]$ and any fixed $\boldsymbol{\xi} \in \Xi$, we have that*

$$\|m_t(\mathbf{z}, \boldsymbol{\xi})\| \leq \kappa_t(\boldsymbol{\xi}),$$

for any $\mathbf{z} \in \mathcal{G}_\iota$, where

$$\kappa_t(\boldsymbol{\xi}) = \sum_{i=1}^{t} \iota^i \|\mathbf{X}_{t-i+1}\| + \sum_{i=1}^{t} \iota^i, \ t \in [T], \ \kappa_0(\boldsymbol{\xi}_0) = 0.$$

Proof. By the expression of $m_t(\mathbf{z}, \boldsymbol{\xi})$ in (4.1.2) and the compactness of $\mathcal{G}_\iota$, it follows that

$$\|m_1(\mathbf{z}, \boldsymbol{\xi})\| = \|\sigma_{\mathrm{lRe}}(V\mathbf{X}_1 + b)\| \leq \|V\|_F \|\mathbf{X}_1\| + \|b\| \leq \iota(\|\mathbf{X}_1\| + 1) := \kappa_1(\boldsymbol{\xi}),$$

and

$$\begin{aligned} &\|m_2(\mathbf{z}, \boldsymbol{\xi})\| = \|\sigma_{\mathrm{lRe}}(W m_1(\mathbf{z}, \boldsymbol{\xi}) + V\mathbf{X}_2 + b)\| \\ &\leq \|W\|_F \|m_1(\mathbf{z}, \boldsymbol{\xi})\| + \|V\|_F \|\mathbf{X}_2\| + \|b\| \\ &\leq \iota^2 \|\mathbf{X}_1\| + \iota \|\mathbf{X}_2\| + \iota^2 + \iota := \kappa_2(\boldsymbol{\xi}). \end{aligned}$$

Using the above method recursively, we can deduce that

$$\|m_t(\mathbf{z}, \boldsymbol{\xi})\| \leq \sum_{i=1}^{t} \iota^i \|\mathbf{X}_{t-i+1}\| + \sum_{i=1}^{t} \iota^i := \kappa_t(\boldsymbol{\xi}).$$

□

By the above lemmas, we can thus deduce the local Lipschitz continuity of $\ell(\mathbf{z}, \boldsymbol{\xi})$ on $\mathcal{G}_\iota$.

Theorem 4.1. *If Assumption 4.1 holds, the following statements hold.*

(i) For any fixed $\boldsymbol{\xi} \in \Xi$, $\ell(\cdot, \boldsymbol{\xi})$ is Lipschitz continuous on $\mathcal{G}_\iota$ with the Lipschitz constant $L_\ell(\boldsymbol{\xi})$, where

$$L_\ell(\boldsymbol{\xi}) = \frac{2}{T}\sum_{t=1}^{T}\Big(\|\mathbf{Y}_t\| + \iota(\kappa_t(\boldsymbol{\xi}) + 1)\Big)\omega_t(\boldsymbol{\xi}), \tag{4.1.13}$$

$$\omega_t(\boldsymbol{\xi}) = \sum_{i=1}^{t}\iota^i \max\{\|\mathbf{X}_{t-i+1}\|, \kappa_{t-i}(\boldsymbol{\xi}), 1\} + \max\{\kappa_t(\boldsymbol{\xi}), 1\}. \tag{4.1.14}$$

(ii) There exists a nonnegative constant Θ such that for any $\boldsymbol{\xi} \in \Xi$, $|L_\ell(\boldsymbol{\xi})| \le \Theta$. Furthermore, $L_\ell(\cdot)$ is a nonnegative integrable function on Ξ.

Proof. (i) We first deduce that for any $t \in [T]$ and any fixed $\boldsymbol{\xi} \in \Xi$, $o_t(\mathbf{z}, \boldsymbol{\xi})$ is Lipschitz continuous on $\mathcal{G}_\iota$. The function $o_t(\mathbf{z}, \boldsymbol{\xi})$ can be represented by the compositions of g_1 and g_2 as follows:

$$\begin{aligned} o_t(\mathbf{z}, \boldsymbol{\xi}) &= g_2\big(\mathbf{z}, m_t(\mathbf{z}, \boldsymbol{\xi})\big) = \big(g_2 \circ g_1\big)\big(\mathbf{z}, m_{t-1}(\mathbf{z}, \boldsymbol{\xi})\big) \\ &= \big(g_2 \circ g_1 \circ ... \circ g_1\big)\big(\mathbf{z}, m_0(\mathbf{z}, \boldsymbol{\xi})\big). \end{aligned}$$

The above together with Lemma 2.1, Lemma 4.1, and Lemma 4.2 implies that for any $\mathbf{z}$, $\mathbf{z}' \in \mathcal{G}_\iota$,

$$\begin{aligned} &\|o_t(\mathbf{z}, \boldsymbol{\xi}) - o_t(\mathbf{z}', \boldsymbol{\xi})\| \\ \le&\iota\|m_t(\mathbf{z}, \boldsymbol{\xi}) - m_t(\mathbf{z}', \boldsymbol{\xi})\| + \max\{\kappa_t(\boldsymbol{\xi}), 1\}\|\mathbf{z} - \mathbf{z}'\| \\ \le&\big(\iota \max\{\|\mathbf{X}_t\|, \kappa_{t-1}(\boldsymbol{\xi}), 1\} + \max\{\kappa_t(\boldsymbol{\xi}), 1\}\big)\|\mathbf{z} - \mathbf{z}'\| \qquad (4.1.15) \\ &+ \iota^2\|m_{t-1}(\mathbf{z}, \boldsymbol{\xi}) - m_{t-1}(\mathbf{z}', \boldsymbol{\xi})\| \\ \le&\left(\sum_{i=1}^{t}\iota^i \max\{\|\mathbf{X}_{t-i+1}\|, \kappa_{t-i}(\boldsymbol{\xi}), 1\} + \max\{\kappa_t(\boldsymbol{\xi}), 1\}\right)\|\mathbf{z} - \mathbf{z}'\| \qquad (4.1.16) \\ =&\omega_t(\boldsymbol{\xi})\|\mathbf{z} - \mathbf{z}'\|, \end{aligned}$$

where Lemma 4.2 is repeatedly used from (4.1.15) to (4.1.16). Then, according to

Remark 4.2, we can further deduce that

$$\begin{aligned}\|\ell(\mathbf{z},\boldsymbol{\xi})-\ell(\mathbf{z}',\boldsymbol{\xi})\| &\leq \frac{1}{T}\sum_{t=1}^{T}\|\ell_t(\mathbf{z},\boldsymbol{\xi})-\ell_t(\mathbf{z}',\boldsymbol{\xi})\| \\ &\leq \frac{2}{T}\sum_{t=1}^{T}\Big(\|\mathbf{Y}_t\|+\|o_t(\mathbf{z},\boldsymbol{\xi})\|\Big)\|o_t(\mathbf{z},\boldsymbol{\xi})-o_t(\mathbf{z}',\boldsymbol{\xi})\| \\ &\leq \frac{2}{T}\sum_{t=1}^{T}\Big(\|\mathbf{Y}_t\|+\iota(\kappa_t(\xi)+1)\Big)\omega_t(\boldsymbol{\xi})\|\mathbf{z}-\mathbf{z}'\| \\ &\leq L_\ell(\boldsymbol{\xi})\|\mathbf{z}-\mathbf{z}'\|,\end{aligned}$$

where $L_\ell(\boldsymbol{\xi})=\frac{2}{T}\sum_{t=1}^{T}\Big(\|\mathbf{Y}_t\|+\iota(\kappa_t(\boldsymbol{\xi})+1)\Big)\omega_t(\boldsymbol{\xi})$.

(ii)Without loss of generality, we assume that T in the expression of $L_\ell(\boldsymbol{\xi})$ is finite. The above together with Assumption 4.1 implies that there exists a nonnegative constant Θ such that for any $\boldsymbol{\xi}\in\Xi$, $|L_\ell(\boldsymbol{\xi})|\leq\Theta$.

By the expression of $L_\ell(\cdot)$ deduced in (i) and the boundedness of $L_\ell(\cdot)$, it is easy to derive that $L_\ell(\boldsymbol{\xi})$ is nonnegative and $\mathbb{E}[L_\ell(\boldsymbol{\xi})]<\infty$ for any $\boldsymbol{\xi}\in\Xi$. □

According to Theorem 4.1, we can deduce the following conclusions.

Proposition 4.1. *For any $\mathbf{z}\in\mathcal{G}_\iota$, $\ell(\mathbf{z},\cdot)$ is dominated by a nonnegative integrable function $B_\ell(\cdot)$ where*

$$B_\ell(\boldsymbol{\xi}):=L_\ell(\boldsymbol{\xi})\iota+\varrho^2,$$

that is, there exists a nonnegative valued measurable function $B_\ell(\boldsymbol{\xi})$ such that $\mathbb{E}[B_\ell(\boldsymbol{\xi})]<+\infty$ and $|\ell(\mathbf{z},\boldsymbol{\xi})|\leq B_\ell(\boldsymbol{\xi})$ holds w.p.1, for those $\boldsymbol{\xi}\in\Xi$.

Proof. At first, we deduce the boundedness of $\ell(\cdot,\boldsymbol{\xi})$ for any fixed $\boldsymbol{\xi}\in\Xi$ based on Theorem 4.1. To be specific, according to the Lipschitz continuity of $\ell(\cdot,\boldsymbol{\xi})$ proved in Theorem 4.1 and the triangle inequality, it follows that for any $\mathbf{z}\in\mathcal{G}_\iota$,

$$|\ell(\mathbf{z},\boldsymbol{\xi})|-|\ell(0,\boldsymbol{\xi})|\leq L_\ell(\boldsymbol{\xi})\|\mathbf{z}\|.$$

It is easy to derive that $|\ell(0, \boldsymbol{\xi})| = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{Y}_t\|^2 \leq \varrho^2$ and $\|\mathbf{z}\| \leq \iota$. Hence, the conclusion can be established, i.e.,

$$|\ell(\mathbf{z}, \boldsymbol{\xi})| \leq L_\ell(\boldsymbol{\xi})\iota + \varrho^2 = B_\ell(\boldsymbol{\xi}).$$

It has been proved that $L_\ell(\boldsymbol{\xi})$ is nonnegative and integrable. Furthermore, ℓ defined in (4.1.10) and ϱ defined in Assumption 4.1 are both finite. Hence, we can claim that $\ell(\mathbf{z}, \cdot)$ is dominated by an integrable function $B_\ell(\cdot)$ for any $\mathbf{z} \in \mathcal{G}_\iota$. □

Proposition 4.2. *Under Assumption 4.1, ψ and $\hat{\psi}_N$ are both Lipschitz continuous on $\mathcal{G}_\iota$ with a Lipschitz constant $\Theta + 1$.*

Proof. We begin by establishing that ψ is Lipschitz continuous on $\mathcal{G}_\iota$. By the expression of ψ in (4.1.8) and the Lipschitz continuity of $\ell(\cdot, \boldsymbol{\xi})$, $\forall\ \boldsymbol{\xi} \in \Xi$, proved in Theorem 4.1, it follows that for any $\mathbf{z}_1$ and $\mathbf{z}_2$ in $\mathcal{G}_\iota$,

$$\begin{aligned}
|\psi(\mathbf{z}_1) - \psi(\mathbf{z}_2)| &\leq |\mathbb{E}[\ell(\mathbf{z}_1, \boldsymbol{\xi})] - \mathbb{E}[\ell(\mathbf{z}_2, \boldsymbol{\xi})]| + |p(\mathbf{z}_1) - p(\mathbf{z}_2)| \\
&\leq \mathbb{E}\left[|\ell(\mathbf{z}_1, \boldsymbol{\xi}) - \ell(\mathbf{z}_2, \boldsymbol{\xi})|\right] + \|\mathbf{z}_1 - \mathbf{z}_2\| \\
&\leq (\mathbb{E}[L_\ell(\boldsymbol{\xi})] + 1)\|\mathbf{z}_1 - \mathbf{z}_2\| \\
&\leq (\Theta + 1)\|\mathbf{z}_1 - \mathbf{z}_2\|.
\end{aligned}$$

Hence, ψ is Lipschitz continuous on $\mathcal{G}_\iota$ with Lipschitz constant $\Theta + 1$.

Moreover, according to Theorem 4.1, we can derive that for any $\mathbf{z}_1$, $\mathbf{z}_2 \in \mathcal{G}_\iota$,

$$|\hat{\psi}_N(\mathbf{z}_1) - \hat{\psi}_N(\mathbf{z}_2)| \leq \frac{1}{N}\sum_{i=1}^{N} \left|\ell(\mathbf{z}_1, \xi^i) - \ell(\mathbf{z}_2, \xi^i)\right| + |p(\mathbf{z}_1) - p(\mathbf{z}_2)| \tag{4.1.17}$$

$$\leq \frac{1}{N}\sum_{i=1}^{N} L_\ell(\xi^i)\|\mathbf{z}_1 - \mathbf{z}_2\| + \|\mathbf{z}_1 - \mathbf{z}_2\| \tag{4.1.18}$$

$$\leq (\Theta + 1)\|\mathbf{z}_1 - \mathbf{z}_2\|. \tag{4.1.19}$$

The statement is thus established. □

Having established properties for ℓ, we proceed to prove ψ is well-defined.

Theorem 4.2. *The expectation value function $\psi : \mathbb{R}^{N_{\mathbf{z}}} \to \mathbb{R}$ is well-defined and continuous on $\mathcal{G}_\iota$.*

Proof. By Theorem 4.1 and Remark 4.1, it is easy to derive that $\psi(\mathbf{z})$ is finite for any $\mathbf{z} \in \mathcal{G}_\iota$. Furthermore, it is known that $\ell(\mathbf{z}, \cdot)$ is measurable. Therefore, ψ is well-defined on $\mathcal{G}_\iota$. According to [50, Theorem 7.43], we can also derive that ψ is continuous on $\mathcal{G}_\iota$. □

We then establish the nonemptiness and compactness for the solution sets of problems (4.1.8) and (4.1.9), and prove that these solution sets are contained in the compact set $\mathcal{G}_\iota$.

Lemma 4.3. *The solution set of* (4.1.8)*, denoted by $\mathcal{S}^*$, is nonempty and compact. Moreover, $\mathcal{S}^*$ is contained within the compact set $\mathcal{G}_\iota$, i.e., $\mathcal{S}^* \subset \mathcal{G}_\iota$.*

Proof. We define the following level set:

$$\Gamma_\psi(\rho) := \{\mathbf{z} \in \mathbb{R}^{N_{\mathbf{z}}} : \psi(\mathbf{z}) \leq \rho\},$$

where ρ is a constant satisfying $\rho > \psi(0)$. Now, we establish the nonemptiness and compactness of the level set. As $\mathbf{z} = 0$, we have that

$$\psi(0) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \|\mathbf{Y}_t\|^2\right],$$

which together with Remark 4.1 implies that $\psi(0)$ is finite. The above together with $\rho > \psi(0)$ implies that $0 \in \Gamma_\psi(\rho)$, which guarantees the nonemptiness of the level set $\Gamma_\psi(\rho)$.

We then turn to consider the boundedness of $\Gamma_\psi(\rho)$. The expressions of f and p in (4.1.5) and (4.1.7) imply that $f(\mathbf{z}) \geq 0$ and $p(\mathbf{z}) \geq 0$ for any $\mathbf{z} \in \mathbb{R}^{N_{\mathbf{z}}}$. Then we can deduce that for those $\mathbf{z} \in \Gamma_\psi(\rho)$,

$$\|\mathbf{z}\| \leq \sqrt{\rho / \min_{i\in[5]}\{\lambda_i\}}. \tag{4.1.20}$$

The boundedness of $\Gamma_\psi(\rho)$ can thus be proved. The closeness of $\Gamma_\psi(\rho)$ can also be derived based on the continuity of function ψ according to [47, Theorem 1.6].

Therefore, we can deduce that $\mathcal{S}^*$ is nonempty, compact and $\mathcal{S}^* \subset \Gamma_\psi(\rho)$ according to the nonemptiness and compactness of the level set $\Gamma_\psi(\rho)$ based on [7, Proposition A.8].

Additionally, we can infer that $\Gamma_\psi(\rho) \subseteq \mathcal{G}_\iota$ by (4.1.10) and (4.1.20), which together with $\mathcal{S}^* \subset \Gamma_\psi(\rho)$ implies that $\mathcal{S}^* \subset \mathcal{G}_\iota$.

□

Remark 4.2. *From the definition of $\Gamma_\psi(\rho)$ and the expression of $f(\mathbf{z})$, we can also derive that $f(\mathbf{z}) = \mathbb{E}[\ell(\mathbf{z}, \boldsymbol{\xi})] \leq \rho$. The above, together with the definition of expectation, implies that for any fixed $\boldsymbol{\xi} \in \Xi$,*

$$\|\mathbf{Y}_t - o_t(\mathbf{z}, \boldsymbol{\xi})\| \leq \sqrt{\rho T},\ \forall t \in [T].$$

We proceed to prove that for any $N \in \mathbb{N}_+$, the solution set of the SAA problem (4.1.9) is nonempty, compact and contained within $\mathcal{G}_\iota$.

Lemma 4.4. *For any $N \in \mathbb{N}_+$, the solution set of problem (4.1.9), denoted by $\mathcal{S}_N$, is nonempty and compact w.p.1. Moreover, $\mathcal{S}_N$ is contained within the compact set $\mathcal{G}_\iota$ w.p.1, i.e., $\mathcal{S}_N \subset \mathcal{G}_\iota$.*

Proof. To prove the result, we define the following level set of $\hat{\psi}_N$ for any $N \in \mathbb{N}_+$:

$$\Gamma^N_{\hat{\psi}}(\theta) := \{\mathbf{z} \in \mathbb{R}^{N_\mathbf{z}} : \hat{\psi}_N(\mathbf{z}) \leq \theta\},$$

where θ is a constant satisfying $\theta > \hat{\psi}_N(0)$, $\forall N \in \mathbb{N}_+$. Specifically, by Assumption 4.1, it follows that

$$\hat{\psi}_N(0) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \|y_t^i\|^2 \leq \varrho^2, \tag{4.1.21}$$

where $\hat{\psi}_N(0)$ is bounded by a constant ϱ^2 for any $N \in \mathbb{N}_+$. The existence of θ can thus be proved. Hence, for any $N \in \mathbb{N}_+$, $0 \in \Gamma^N_{\hat{\psi}}(\theta)$ and thereby $\Gamma^N_{\hat{\psi}}(\theta)$ is nonempty w.p.1.

Furthermore, by the definition of level set $\Gamma^N_{\hat{\psi}}(\theta)$ and the expression of $\hat{\psi}_N$ in (4.1.6)-(4.1.9), we infer that

$$\|\mathbf{z}\| \leq \sqrt{\theta / \min_{i\in[5]}\{\lambda_i\}}, \tag{4.1.22}$$

which implies the boundedness of the level set $\Gamma^N_{\hat{\psi}}(\theta)$. By the similar proof method as in Lemma 4.3, we can deduce that $\mathcal{S}_N$ is bounded, compact and $\mathcal{S}_N \subset \Gamma^N_{\hat{\psi}}(\theta)$ for any $N \in \mathbb{N}_+$ w.p.1.

Additionally, according to (4.1.22), where θ is independent of the sample size N, we can deduce that $\Gamma^N_{\hat{\psi}}(\theta) \subset \mathcal{G}_\iota$. The above along with $\mathcal{S}_N \subset \Gamma^N_{\hat{\psi}}(\theta)$ implies that $\mathcal{S}_N \subset \mathcal{G}_\iota$ for any $N \in \mathbb{N}_+$. □

4.2 Convergence of SAA problems

In this section, we first show that any accumulation point of minimizers of the SAA problems is a minimizer of the original problem w.p.1 as $N \to \infty$. After that, we explore the convergence of the stationary points of SAA problems.

4.2.1 Convergence of the optimal value and optimal solutions of the SAA problem

Recall that $\mathcal{S}^*$ denotes the global solution set of problem (4.1.8), and $\mathcal{S}_N$ denotes the global solution set of problem (4.1.9) when the sample size is N. Furthermore, let ν^* and $\hat{\nu}_N$ denote the optimal value of problem (4.1.8) and (4.1.9) respectively. Before the formal statement of the convergence, we first show that function $\hat{\psi}_N$ uniformly converges to ψ w.p.1 on the compact set $\mathcal{G}_\iota$ as $N \to \infty$.

Lemma 4.5. *Under Assumption 4.1 and the i.i.d samples of* $\boldsymbol{\xi}$*,* $\hat{\psi}_N$ *uniformly converges to* ψ *on* $\mathcal{G}_\iota$ *w.p.1, as* $N \to \infty$*.*

Proof. Proposition 4.1 shows that the loss function $\ell(\mathbf{z}, \boldsymbol{\xi})$, $\mathbf{z} \in \mathcal{G}_\iota$, is dominated by an integrable function $B_\ell(\boldsymbol{\xi})$. Moreover, the sample $\{\xi^1, \xi^2, ..., \xi^N\}$ is i.i.d. Thus, we can claim that $\hat{\psi}_N(\mathbf{z})$ converges to $\psi(\mathbf{z})$ w.p.1, as $N \to \infty$, uniformly on $\mathcal{G}_\iota$ based on [49, Proposition 7]. □

Now, we present the main results of the convergence of SAA problems.

Theorem 4.3. *Under Assumption 4.1 and the i.i.d samples of* $\boldsymbol{\xi}$*,* $\hat{\nu}_N$ *converges to* ν^* *w.p.1 as* $N \to \infty$*. Furthermore, the solution set of SAA problem* (4.1.9)*, denoted by* $\mathcal{S}_N$*, converges to the solution set of problem* (4.1.8)*, denoted by* $\mathcal{S}^*$*, w.p.1 as* $N \to \infty$*. That is,* $\mathbb{D}(\mathcal{S}_N, \mathcal{S}^*) \to 0$ *w.p.1 as* $N \to \infty$*.*

Proof. The convergence of $\hat{\nu}_N$ to ν^* w.p.1 for N large enough can be equivalently represented as $|\hat{\nu}_N - \nu^*| \leq \epsilon$ w.p.1 as $N \to \infty$. This can be derived by Lemma 4.5, which implies that for any $\epsilon > 0$, $\sup_{\mathbf{z} \in \mathcal{G}_\iota} |\hat{\psi}_N(\mathbf{z}) - \psi(\mathbf{z})| \leq \epsilon$ as $N \to \infty$.

Now, we turn to prove the convergence of solution sets of SAA problems. We have already proved that the solution set $\mathcal{S}^*$ is nonempty in Lemma 4.3, and $\mathcal{S}^* \subset \mathcal{G}_\iota$ in Lemma 4.3. Furthermore, Proposition 4.2 shows that $\psi(\mathbf{z})$ is finite valued and continuous on $\mathcal{G}_\iota$. Moreover, It has been proved that $\hat{\psi}_N(\mathbf{z})$ converges to $\psi(\mathbf{z})$ w.p.1, as $N \to \infty$, uniformly on $\mathcal{G}_\iota$ in Lemma 4.5. At last, the condition where for N large enough the set $\mathcal{S}_N$ is nonempty and $\mathcal{S}_N \subset \mathcal{G}_\iota$ has been established in Lemma 4.4 and Lemma 4.4.

According to [49, Proposition 6], the assertion can thus be established. □

4.2.2 Convergence of stationary points of SAA problems

In the previous section, we utilized the SAA method to approximate problem (4.1.8) whose objective function involves expectation and discussed the convergence of the optimal solutions and the optimal value of the SAA problems (4.1.9). However, it is also challenging to derive the global or local optimal solutions for the SAA problems with nonconvex and nonsmooth objective functions. In general, stationary points of those nonconvex nonsmooth SAA problems can be obtained by numerical algorithms. Therefore, as an alternative, we focus on the convergence of stationary points of the SAA problems (4.1.9) to those of problem (4.1.8), which is related to the first-order optimality condition of the corresponding problem.

The first-order necessary conditions of problem (4.1.8) associated with l-subdifferential and C-subdifferential are as follows, respectively:

$$0 \in \mathcal{A}(\mathbf{z}) := \partial_{\mathbf{z}} \mathbb{E}[\ell(\mathbf{z}, \boldsymbol{\xi})] + \nabla p(\mathbf{z}), \tag{4.2.1}$$

$$0 \in \mathcal{A}^c(\mathbf{z}) := \partial^c_{\mathbf{z}} \mathbb{E}[\ell(\mathbf{z}, \boldsymbol{\xi})] + \nabla p(\mathbf{z}). \tag{4.2.2}$$

Furthermore, the first-order necessary condition of problem (4.1.9) associated with l-subdifferential is

$$0 \in \hat{\mathcal{A}}_N(\mathbf{z}) := \partial_{\mathbf{z}} \Big(\frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{z}, \xi^i) \Big) + \nabla p(\mathbf{z}). \tag{4.2.3}$$

Definition 4.1. *(Stationary points) We say a point $\mathbf{z} \in \mathbb{R}^{N_{\mathbf{z}}}$ is a limiting stationary point or Clarke stationary point of* (4.1.8) *if it satisfies* (4.2.1) *or* (4.2.2)*, respectively. Moreover, a point $\mathbf{z} \in \mathbb{R}^{N_{\mathbf{z}}}$ is a limiting stationary point* (4.1.9) *if it satisfies the generalized equation* (4.2.3)*.*

Now we define weak first-order necessary conditions of problems (4.1.8) and (4.1.9). The weak first-order necessary conditions of problem (4.1.8) associated with

l-subdifferential and C-subdifferential are as follows, respectively:

$$0 \in \mathcal{W}(\mathbf{z}) := \mathbb{E}[\partial_{\mathbf{z}} \ell(\mathbf{z}, \boldsymbol{\xi})] + \nabla p(\mathbf{z}), \tag{4.2.4}$$

$$0 \in \mathcal{W}^c(\mathbf{z}) := \mathbb{E}[\partial^c_{\mathbf{z}} \ell(\mathbf{z}, \boldsymbol{\xi})] + \nabla p(\mathbf{z}). \tag{4.2.5}$$

The weak first-order necessary condition of problem (4.1.9) associated with l-subdifferential is as follows:

$$0 \in \hat{\mathcal{W}}_N(\mathbf{z}) = \frac{1}{N} \sum_{i=1}^{N} \partial_{\mathbf{z}} \ell(\mathbf{z}, \xi^i) + \nabla p(\mathbf{z}). \tag{4.2.6}$$

Definition 4.2. *(Weak stationary points) It is said that a point $\mathbf{z} \in \mathbb{R}^{N_{\mathbf{z}}}$ is a weak limiting stationary point of* (4.1.8) *if it satisfies* (4.2.4) *and is a weak Clarke stationary point of* (4.1.8) *if it satisfies* (4.2.5)*. A point $\mathbf{z} \in \mathbb{R}^{N_{\mathbf{z}}}$ is a weak limiting stationary point of* (4.1.9)*, if it satisfies the generalized equation* (4.2.6)*.*

Remark 4.3. *We refer to* (4.2.4) *and* (4.2.5) *as the weak first-order necessary conditions of problem* (4.1.8) *due to the following relationship established in [10, p. 230]:*

$$\partial_{\mathbf{z}} \mathbb{E}[\ell(\mathbf{z}, \boldsymbol{\xi})] \subseteq \partial^c_{\mathbf{z}} \mathbb{E}[\ell(\mathbf{z}, \boldsymbol{\xi})] \subseteq \mathbb{E}[\partial^c_{\mathbf{z}} \ell(\mathbf{z}, \boldsymbol{\xi})].$$

Additionally, by the Lipschitz continuity of $\ell(\cdot, \boldsymbol{\xi})$ on the compact set $\mathcal{G}_\iota$ and the rule of the sum of l-subdifferential [47, Corollary 10.9], the following relationship can be established for any $\mathbf{z} \in int(\mathcal{G}_\iota)$:

$$\partial_{\mathbf{z}} \Big(\frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{z}, \xi^i) \Big) \subseteq \frac{1}{N} \sum_{i=1}^{N} \partial_{\mathbf{z}} \ell(\mathbf{z}, \xi^i). \tag{4.2.7}$$

The above implies that (4.2.6) *is a weak first-order necessary condition for problem* (4.1.9)*, in comparison to* (4.2.3)*.*

After that, we establish that the above set-valued mappings $\hat{\mathcal{A}}_N$, $\hat{\mathcal{W}}_N$, $\mathcal{A}$, $\mathcal{A}^c$, $\mathcal{W}$, and $\mathcal{W}^c$ are well-defined.

Lemma 4.6. *Under Assumption 4.1, the set-valued mappings $\hat{\mathcal{A}}_N$, $\hat{\mathcal{W}}_N$, $\mathcal{A}$, $\mathcal{A}^c$, $\mathcal{W}$, and $\mathcal{W}^c$ are well-defined on $\mathcal{G}_\iota$. That is, for every $\mathbf{z} \in \mathcal{G}_\iota$, the images of these mapping at $\mathbf{z}$ are nonempty.*

Proof. At first, we prove that $\hat{\mathcal{A}}_N$, $\hat{\mathcal{W}}_N$, $\mathcal{A}$ and $\mathcal{A}^c$ are well-defined on $\mathcal{G}_\iota$. By the Lipschitz continuity of $\ell(\cdot, \boldsymbol{\xi})$ proved in Theorem 4.1 (i) , and the Lipschitz continuity of $\hat{\psi}_N$ and ψ proved in Proposition 4.2, it is easy to derive that $\hat{\mathcal{A}}_N(\mathbf{z})$, $\hat{\mathcal{W}}_N(\mathbf{z})$ and $\mathcal{A}(\mathbf{z})$ are nonempty for any $\mathbf{z} \in \mathcal{G}_\iota$ based on [47, Theorem 9.13], and $\mathcal{A}^c(\mathbf{z})$ is nonempty according to [16, Proposition 2.1.2].

After that, we prove that $\mathcal{W}$ and $\mathcal{W}^c$ are well-defined. The Lipschitz continuity of $\ell(\cdot, \boldsymbol{\xi})$ on $\mathcal{G}_\iota$ also guarantees the boundedness of $\partial_{\mathbf{z}} \ell(\cdot, \boldsymbol{\xi})$ and $\partial^c_{\mathbf{z}} \ell(\cdot, \boldsymbol{\xi})$ as well according to [47, Theorem 9.13] and [16, Proposition 2.1.2]. Thus, we derive that $\mathbb{E}[\mathbb{H}(0, \partial_{\mathbf{z}} \ell(\mathbf{z}, \boldsymbol{\xi}))] < \infty$ and $\mathbb{E}[\mathbb{H}(0, \partial^c_{\mathbf{z}} \ell(\mathbf{z}, \boldsymbol{\xi}))] < \infty$ for any $\mathbf{z} \in \mathcal{G}_\iota$. The above implies that $\mathcal{W}$ and $\mathcal{W}^c$ are well-defined according to [62, p. 291]. □

Now we can establish the convergence of the stationary points of the SAA problems according to [61, Theorem 4.2].

Theorem 4.4. *Let $\mathbf{z}_N$ be a solution of* (4.2.3)*. Suppose that the sequence $\{\mathbf{z}_N\}$ is contained in the compact set $\mathcal{G}_\iota$. Then any accumulation point of $\{\mathbf{z}_N\}$ satisfies* (4.2.5) *w.p.1 as $N \to \infty$.*

Proof. Due to $\mathbf{z}_N$ is a solution of (4.2.3), we have that w.p.1

$$0 \in \hat{\mathcal{A}}_N(\mathbf{z}_N), \ \forall \ N \in \mathbb{N}_+.$$

Let $\{\mathbf{z}_{N_k}\}$ denote a subsequence of $\{\mathbf{z}_N\}$ converging to $\mathbf{z}^*$.

Now we prove that w.p.1

$$\mathbb{D}\left(\hat{\mathcal{A}}_{N_k}(\mathbf{z}_{N_k}), \mathcal{W}^c(\mathbf{z}^*)\right) \to 0, \text{ as } N_k \to \infty. \tag{4.2.8}$$

According to (4.2.7), it is easy to derive that

$$\mathbb{D}\left(\hat{\mathcal{A}}_{N_k}(\mathbf{z}_{N_k}), \mathcal{W}^c(\mathbf{z}^*)\right) \leq \mathbb{D}\left(\hat{\mathcal{W}}_{N_k}(\mathbf{z}_{N_k}), \mathcal{W}^c(\mathbf{z}^*)\right). \tag{4.2.9}$$

Moreover, by the Lipschitz continuity of $\partial\ell(\cdot, \boldsymbol{\xi})$ proved in Theorem 4.1, we infer that $\partial\ell(\cdot, \boldsymbol{\xi})$ is closed and locally bounded on compact set $\mathcal{G}_\iota$ according to [47, Theorem 9.13]. Therefore, $\partial\ell(\cdot, \boldsymbol{\xi})$ is upper semicontinuous on $\mathcal{G}_\iota$ based on [47, Theorem 5.19]. From the above, we can also deduce that

$$\|\partial\ell(\mathbf{z}, \boldsymbol{\xi})\| \leq L_\ell(\boldsymbol{\xi}),\ \forall(\mathbf{z}, \boldsymbol{\xi}) \in \mathcal{G}_\iota \times \Xi.$$

By applying [53, Theorem 2] and [63, Theorem 4.3], we can infer that

$$\mathbb{D}\left(\hat{\mathcal{W}}_{N_k}(\mathbf{z}_{N_k}), \mathcal{W}^c(\mathbf{z}^*)\right) \to 0,\ \text{as } N_k \to \infty.$$

The above together with (4.2.9) implies that

$$0 \in \mathcal{W}^c(\mathbf{z}^*).$$

The conclusion can thus be proved. □

4.3 Exponential rates of convergence

In the previous section, we proved that the solution sets of the SAA problems converge to the solution set of the original problem as the sample size becomes sufficiently large. In this subsection, we show the uniform exponential rate of convergence of $\hat{\psi}_N$ to $\psi(\mathbf{z})$ on $\mathcal{G}_\iota$.

A core concept to deduce the exponential convergence is the Cramér's Large Deviation (LD) theorem, which provides an exponential upper bound for the probabilities of large deviations. The formal statement of Cramér's LD theorem is presented as follows [49, p. 418]:

Lemma 4.7. *(Cramér's LD Theorem) Let* $\{r^1, r^2, ..., r^N\}$ *be an i.i.d sequence of replications of the random variable* $\mathsf{r} : \Omega^r \to \mathcal{R} \subset \mathbb{R}$. *Furthermore, let* $\bar{r}_N :=$

$N^{-1}\sum_{i=1}^{N} r^i$ *denote the corresponding sample average. Suppose that* r *has finite mean* $\mu_r := \mathbb{E}[\boldsymbol{\xi}]$. *For any real number* ϵ *satisfying* $\epsilon \geq \mu_r$, *we obtain that*

$$\mathbb{P}(\bar{r}_N \geq \epsilon) \leq e^{-NI(\epsilon)},$$

where

$$I(\epsilon) := \sup_{\tau \in \mathbb{R}} \{\tau\epsilon - \ln M(\tau)\}$$

is called the rate function of r, *and* $M(\tau) := \mathbb{E}[e^{\tau \mathsf{r}}]$ *is the moment generating function (MGF) of* r.

Let function $D : \mathbb{R}^{N_{\mathbf{z}}} \times \Xi \to \mathbb{R}$ have the following representation:

$$D(\mathbf{z}, \boldsymbol{\xi}) := \ell(\mathbf{z}, \boldsymbol{\xi}) - f(\mathbf{z}),$$

and its corresponding MGF and rate function are

$$M_D(\tau) := \mathbb{E}\left[e^{\tau D(\mathbf{z}, \boldsymbol{\xi})}\right], \; I_D(\epsilon) := \sup_{\tau \in \mathbb{R}} \{\tau\epsilon - \ln M_D(\tau)\}.$$

Furthermore, denote the MGF and rate function of $L_\ell(\boldsymbol{\xi})$ as follows:

$$M_{L_\ell}(\tau) := \mathbb{E}\left[e^{\tau L_\ell(\boldsymbol{\xi})}\right], \; I_{L_\ell}(\epsilon) := \sup_{\tau \in \mathbb{R}} \{\tau\epsilon - \ln M_{L_\ell}(\tau)\}.$$

Lemma 4.8. *Under Assumption 4.1,* $M_{L_\ell}(\tau)$ *is finite valued for any* $\tau \in \mathbb{B}_\epsilon(0) := \{\tau : \|\tau\| \leq \epsilon\}$.

Proof. According to Theorem 4.1 (ii), we can deduce the following:

$$M_{L_\ell}(\tau) = \mathbb{E}\left[e^{\tau L_\ell(\boldsymbol{\xi})}\right] = \int e^{\tau L_\ell(\boldsymbol{\xi})} dP_\xi \leq \int e^{\tau \Theta} dP_\xi = e^{\tau \Theta},$$

where Θ is a nonnegative constant. Hence, we prove the statement. □

Lemma 4.9. *Under Assumption 4.1,* $M_D(\tau)$ *is finite valued for any* $\mathbf{z} \in \mathcal{G}_\iota$ *and any* $\tau \in \mathbb{B}_\epsilon(0) := \{\tau : |\tau| \leq \epsilon\}$.

Proof. By definition, the moment generating function (MGF) of D is given by

$$M_D(\tau) = \mathbb{E}\left[e^{\tau D(\mathbf{z},\boldsymbol{\xi})}\right] = \int e^{\tau(\ell(\mathbf{z},\boldsymbol{\xi})-f(\mathbf{z}))}\, dP_\xi = e^{-\tau f(\mathbf{z})} \int e^{\tau \ell(\mathbf{z},\boldsymbol{\xi})}\, dP_\xi. \tag{4.3.1}$$

We first bound the integral $\int e^{\tau\ell(\mathbf{z},\boldsymbol{\xi})}\, dP_\xi$ for any fixed $\mathbf{z} \in \mathcal{G}_\iota$. Using Proposition 4.1 and Theorem 4.1 (ii), we have

$$\int e^{\tau\ell(\mathbf{z},\boldsymbol{\xi})}\, dP_\xi \le \int e^{\tau B(\boldsymbol{\xi})}\, dP_\xi \le \int e^{\tau(\Theta_\iota+\varrho^2)}\, dP_\xi = e^{\tau\left(\Theta_\iota+\varrho^2\right)}. \tag{4.3.2}$$

Additionally, since $\ell(\cdot,\boldsymbol{\xi})$ is a nonnegative integrable function, we conclude that

$$f(\mathbf{z}) = \mathbb{E}\left[\ell(\mathbf{z},\boldsymbol{\xi})\right] \ge 0, \quad \forall \mathbf{z} \in \mathcal{G}_\iota. \tag{4.3.3}$$

We now consider two cases based on the sign of τ:

Case 1: $\tau \ge 0$**.** In this case, since $f(\mathbf{z}) \ge 0$, the factor $e^{-\tau f(\mathbf{z})} \le 1$. Thus, substituting (4.3.2) and (4.3.3) into (4.3.1) we obtain

$$M_D(\tau) = e^{-\tau f(\mathbf{z})} \int e^{\tau\ell(\mathbf{z},\boldsymbol{\xi})}\, dP_\xi \le e^{\tau\left(\Theta_\iota+\varrho^2\right)}.$$

Since the right-hand side is finite for all τ with $|\tau| \le \epsilon$, $M_D(\tau)$ is finite for $\tau \ge 0$.

Case 2: $\tau < 0$**.** For negative τ, note that $\ell(\mathbf{z},\boldsymbol{\xi}) \ge 0$ implies

$$e^{\tau\ell(\mathbf{z},\boldsymbol{\xi})} \le 1.$$

Thus, we have

$$\int e^{\tau\ell(\mathbf{z},\boldsymbol{\xi})}\, dP_\xi \le \int 1\, dP_\xi = 1.$$

It follows that

$$M_D(\tau) = e^{-\tau f(\mathbf{z})} \int e^{\tau\ell(\mathbf{z},\boldsymbol{\xi})}\, dP_\xi \le e^{-\tau f(\mathbf{z})}.$$

Since $\tau \in \mathbb{B}_\epsilon(0)$, we have $|\tau| \le \epsilon$. In particular, for $\tau < 0$,

$$-\tau \le \epsilon,$$

so that

$$e^{-\tau f(\mathbf{z})} \leq e^{\epsilon f(\mathbf{z})}.$$

Because $f(\mathbf{z})$ is finite (due to the integrability of $\ell(\mathbf{z}, \boldsymbol{\xi})$), it follows that $M_D(\tau)$ is finite for $\tau < 0$ as well.

Combining the two cases, we conclude that for any $\mathbf{z} \in \mathcal{G}_\iota$ and any $\tau \in \mathbb{B}_\epsilon(0)$ the moment generating function $M_D(\tau)$ is finite.

□

Now, we consider the exponential convergence for SAA problems.

Theorem 4.5. *If Assumption 4.1 holds, then for any $\epsilon > 0$ there exist positive constants C_ϵ and χ_ϵ, independent of N, such that*

$$\mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{G}_\iota} |\hat{\psi}_N(\mathbf{z}) - \psi(\mathbf{z})| \geq \epsilon\right) \leq C_\epsilon e^{-N\chi_\epsilon}. \tag{4.3.4}$$

Proof. The proof is primarily based on the proof of [52, Theorem 5.1].

Due to the finiteness of $\mathcal{G}_\iota$ is unknown, we can use ν-net to approximate $\mathcal{G}_\iota$ as a finite set. Specifically, for a $\nu > 0$, let $\bar{\mathbf{z}}_1$, $\bar{\mathbf{z}}_2$, ..., $\bar{\mathbf{z}}_U \in \mathcal{G}_\iota$ be such that for any $\mathbf{z} \in \mathcal{G}_\iota$ there exists $\bar{\mathbf{z}}_i$, $i \in [U]$, such that $\|\mathbf{z} - \bar{\mathbf{z}}_i\| \leq \nu$. $\{\bar{\mathbf{z}}_1, \bar{\mathbf{z}}_2, ..., \bar{\mathbf{z}}_U\}$ is a ν-net in $\mathcal{G}_\iota$. The ν-net can be chosen by $U \leq \lceil O(1)d/\nu \rceil^{N_{\mathbf{z}}}$, where $d := \sup_{\mathbf{z},\mathbf{z}'\in\mathcal{G}_\iota} \|\mathbf{z} - \mathbf{z}'\|$ is the diameter of $\mathcal{G}_\iota$ and O(1) represents the generic constant. Let $i(\mathbf{z}) \in \arg\min_{i\in[U]} \|\mathbf{z} - \bar{\mathbf{z}}_i\|$, where $\mathbf{z} \in \mathcal{G}_\iota$. We can thus deduce the following inequality:

$$\begin{aligned} &\sup_{\mathbf{z}\in\mathcal{G}_\iota} |\hat{\psi}_N(\mathbf{z}) - \psi(\mathbf{z})| \\ \leq &\sup_{\mathbf{z}\in\mathcal{G}_\iota} |\hat{\psi}_N(\mathbf{z}) - \hat{\psi}_N(\bar{\mathbf{z}}_{i(\mathbf{z})})| + \max_{i\in[U]} |\hat{\psi}_N(\bar{\mathbf{z}}_{i(\mathbf{z})}) - \psi(\bar{\mathbf{z}}_{i(\mathbf{z})})| + \sup_{\mathbf{z}\in\mathcal{G}_\iota} |\psi(\bar{\mathbf{z}}_{i(\mathbf{z})}) - \psi(\mathbf{z})|. \end{aligned} \tag{4.3.5}$$

By Proposition 4.2, the first term and the last term in (4.3.5) can be deduced as

follows:

$$\sup_{\mathbf{z}\in\mathcal{G}_\iota} |\hat{\psi}_N(\mathbf{z}) - \hat{\psi}_N(\bar{\mathbf{z}}_{i(\mathbf{z})})| \leq (\Theta+1)\nu, \tag{4.3.6}$$

$$\sup_{\mathbf{z}\in\mathcal{G}_\iota} |\psi(\bar{\mathbf{z}}_{i(\mathbf{z})}) - \psi(\mathbf{z})| \leq (\Theta+1)\nu. \tag{4.3.7}$$

We turn to consider the second term in (4.3.5). The event $\{\max_{i\in[U]} |\hat{\psi}_N(\bar{\mathbf{z}}_{i(\mathbf{z})}) - \psi(\bar{\mathbf{z}}_{i(\mathbf{z})})| \geq \epsilon\}$ is equal to the union of the event $\{|\hat{\psi}_N(\bar{\mathbf{z}}_{i(\mathbf{z})}) - \psi(\bar{\mathbf{z}}_{i(\mathbf{z})})| \geq \epsilon\}$, $i \in [U]$. The above, together with Cramér's LD theorem, implies that

$$\begin{aligned}\mathbb{P}\left(\max_{i\in[U]} |\hat{\psi}_N(\bar{\mathbf{z}}_{i(\mathbf{z})}) - \psi(\bar{\mathbf{z}}_{i(\mathbf{z})})| \geq \epsilon_2\right) &\leq \sum_{i=1}^{U} \mathbb{P}\left(|\hat{\psi}_N(\bar{\mathbf{z}}_{i(\mathbf{z})}) - \psi(\bar{\mathbf{z}}_{i(\mathbf{z})})| \geq \epsilon_2\right) \\ &\leq 2\sum_{i=1}^{U} e^{-N\min\{I_D(\epsilon_2), I_D(-\epsilon_2)\}}.\end{aligned} \tag{4.3.8}$$

Substituting (4.3.6)-(4.3.8) into (4.3.5) and denote $\epsilon_2 = \epsilon - 2(\Theta+1)\nu$, we can deduce that

$$\begin{aligned}\mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{G}_\iota} |\hat{\psi}_N(\mathbf{z}) - \psi(\mathbf{z})| \geq \epsilon\right) &\leq \mathbb{P}\left(\max_{i\in[U]} |\hat{\psi}_N(\bar{\mathbf{z}}_{i(\mathbf{z})}) - \psi(\bar{\mathbf{z}}_{i(\mathbf{z})})| \geq \epsilon_2\right) \\ &\leq 2\sum_{i=1}^{U} e^{-N\min\{I_D(\epsilon_2), I_D(-\epsilon_2)\}}.\end{aligned} \tag{4.3.9}$$

By Lemma 4.9, we can derive that $I_D(\epsilon_2)$ and $I_D(-\epsilon_2)$ are positive based on [50, pp. 388-389]. Furthermore, the choice of ν-net does not depend on the sample. We thus obtain (4.3.4). □

4.4 Numerical experiments

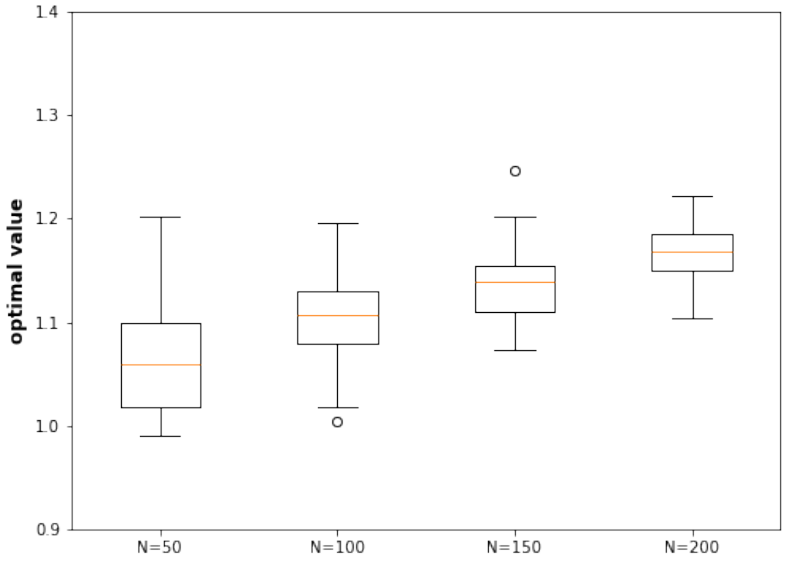
In this subsection, we aim to prove that the function value of stationary points of SAA problem (4.1.9) converges to the original problem with expectation in (4.1.8) w.p.1 as the sample size N goes to infinity.

We employ **Synthetic dataset** and **Pixel-MNIST** datasets described in section 3.2.2 for the numerical experiment in this section. The process of the experiment can be described as follows: we repeatedly employ the ALM to solve problem (3.2.6) and record the function values of the final output solutions 30 times with the sample size $N = 50$, 100, 150, 200 for **Synthetic dataset** and $N = 50$, 100, 200, 500 for **Pixel-MNIST** stated in section 3.2.2 respectively. It is worth mentioning that the parameters for all datasets are the same as those listed in Table 3.6. We then draw boxplots of the 30 times function values for each sample size respectively.

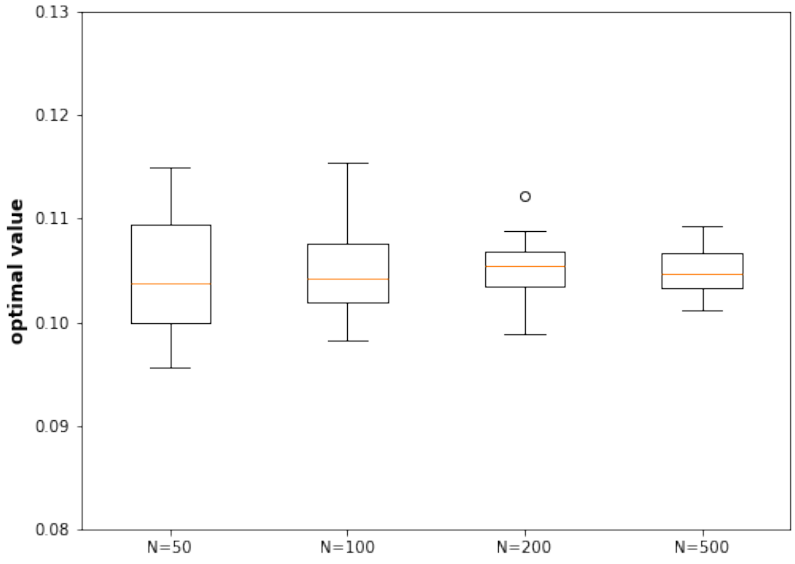
Figure 4.1 illustrates the convergence trend of the function value of stationary points as the sample size N increases.

Figure 4.1: Box plots of optimal values under different sample size N.

(a) Synthetic dataset



(b) Pixel-MNIST



Chapter 5

Conclusions and Future Work

This chapter draws conclusions on the thesis and points out some possible research directions related to the work done in this thesis.

5.1 Conclusions

In this thesis, we study optimization problems arising from the training process of RNNs, which are widely used in natural language processing, speech recognition, and time series forecasting.

- We propose an augmented Lagrangian-based method to solve SAA problems (1.1.2) and (1.1.1) arising from training RNNs, which are both nonconvex nonsmooth and highly composite. The method first reformulates (1.1.2) and (1.1.1) as (3.1.4) and (3.2.5), respectively using auxiliary variables. After adding a regularization term, we focus on solving the regularized problem (3.1.6) and (3.2.6). For (3.1.6), we present the ALM in Algorithm 1 along with BCD method in Algorithm 2. The BCD method in Algorithm 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 (3.1.6), as well as the finite termination of the BCD method for the subproblem of the ALM at each iteration. Similar algorithms are pro-

posed for problem (3.2.6). The efficiency and effectiveness of the ALM for training RNNs are demonstrated by numerical results with real-world datasets and synthetic data, and then comparison with state-of-art algorithms.

- We analyze the convergence of the optimal value, the solution set, and limiting stationary points of SAA problem (1.1.2) to those of the original problem (1.1.3) as the sample size goes to infinity. To achieve this, we first reformulate (1.1.2) and (1.1.3) as (4.1.8) and (4.1.9), respectively, and add regularization terms for them. We then investigate properties of the objective functions in problem (4.1.8) and (4.1.9). Based on the analysis, we established the convergence of the optimal value, solutions, and limiting stationary points of the SAA problems. Finally, we conduct numerical experiments to verify the theoretical convergence results.

5.2 Future work

In the future, we plan to enhance our algorithm by extending it into a stochastic framework, which holds great potential for efficiently addressing problems involving massive datasets. By incorporating stochastic techniques, we aim to significantly reduce computational costs while maintaining high levels of accuracy and scalability, enabling our method to handle large-scale problems that are common in real-world applications. This extension will open up new possibilities for deploying our approach in scenarios where deterministic algorithms may struggle due to the sheer size or complexity of the data.

Moreover, our proposed method, along with its theoretical analysis, has the flexibility to be adapted to more sophisticated and widely-used RNN architectures, such as Long Short-Term Memory networks (LSTMs) and Gated Recurrent Units (GRUs). These advanced architectures are particularly well-suited for capturing long-term de-

pendencies and handling sequential data with greater complexity. By extending our framework to these models, we aim to broaden the applicability of our method and provide a more comprehensive solution for complex time-series analysis, natural language processing, and other sequence modeling tasks. This future work will not only solidify the robustness of our approach but also contribute to the advancement of efficient learning algorithms in the domain of deep learning.

Bibliography

[1] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM J. Optim.*, 23 (2013), pp. 2037-2060.

[2] Y. Bengio. Learning deep architectures for AI. *Found. Trends Mach. Learn.*, 2 (2009), pp. 1-127.

[3] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu. Advances in optimizing recurrent networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1973, pp. 260-264.

[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.

[5] D. P. Bertsekas. On penalty and multiplier methods for constrained minimization. *SIAM J. Control Optim.*, 14 (1976), pp. 216-235.

[6] D. P. Bertsekas. Convergence rate of penalty and multiplier methods. In *Proceedings of the IEEE Conference on Decision and Control*, 1973, pp. 260-264.

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

[8] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146 (2014), pp. 459-494.

[9] A. Bucci. Realized volatility forecasting with neural networks. *J. Financ. Econ.*, 18 (2020), pp. 502-531.

[10] J. V. Burke, X. Chen, and H. Sun. The subdifferential of measurable composite max integrands and smoothing approximation. *Math. Program.*, 181 (2020), pp. 229-264.

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

[12] 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.

[13] M. Chen, X. Li, and T. Zhao. On generalization bounds of a family of recurrent neural networks. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020, pp. 1233-1243.

[14] 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.

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

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

[17] Y. Cui, Z. He, and J.-S. Pang. Multicomposite nonconvex optimization for training deep neural networks. *SIAM J. Optim.*, 30 (2020), pp. 1693-1723.

[18] K. Eriksson, D. Estep, and C. Johnson. *Applied Mathematics: Body and Soul (Volume 1)*. Springer, 1st edition, 2004.

[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and D. N. L. Darpa TIMIT acoustic-phonetic continous speech corpus CD-ROM. *NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD*, 1993.

[20] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249-256.

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

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

[23] N. Hallak and M. Teboulle. An adaptive Lagrangian-based scheme for nonconvex composite optimization. *Math. Oper. Res.*, 48 (2023), pp. 2337-2352.

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

[25] M. R. Hestenes. Multiplier and gradient methods. *J. Optim. Theory Appl.*, 4 (1969), pp. 303-320.

[26] M. Hong, X. Wang, M. Razaviyayn, and Z.-Q. Luo. Iteration complexity analysis of block coordinate descent methods. *Math. Program.*, 163 (2017), pp. 85-114.

[27] P. T. A. 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.

[28] C. Kanzow, A. B. Raharja, and A. Schwartz. An augmented lagrangian method for cardinality-constrained optimization problems. *J. Optimiz. Theory. App.*, 189 (2021), pp. 793-813.

[29] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 972-981.

[30] B. W. Kort and D. P. Bertsekas. Combined primal–dual and penalty methods for convex programming. *SIAM J. Control Optim.*, 14 (1976), pp. 268-294.

[31] A. Y. Kruger. On Fréchet subdifferentials. *J. Math. Sci.*, 116 (2003), pp. 3325-3358.

[32] Q. V. Le, N. Jaitly, and G. E. Hinton. A simple way to initialize recurrent networks of rectified linear units. preprint, 2015.

[33] Y. LeCun, C. Cortes, and C. J. Burges. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[34] J. Li, A. M.-C. So, and W.-K. Ma. Understanding notions of stationarity in nonsmooth optimization: A guided tour of various constructions of subdifferential for nonsmooth functions. *IEEE. Signal. Proc. Mag.*, 37 (2020), pp. 18-31.

[35] 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.

[36] W. Liu, X. Liu, and X. Chen. Linearly constrained nonsmooth optimization for training autoencoders. *SIAM J. Optim.*, 32 (2022), pp. 1931-1957.

[37] R. Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM J. Control Optim.*, 15 (1977), pp. 959-972.

[38] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication*, 2010, pp. 1045-1048.

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

[40] B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation I.* Springer, Berlin, 2006.

[41] J. Nocedal and S. J. Wright. *Numerical optimization.* Springer, NY, 2006.

[42] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *the 30th International Conference on Machine Learning*, 2013, pp. 1310-1318.

[43] D. Peng and X. Chen. Computation of second-order directional stationary points for group sparse optimization. *Optim. Methods Softw.*, 35 (2020), pp. 348-376.

[44] M. J. Powell. A method for nonlinear constraints in minimization problems. *Optimization*, 1969, pp. 283-298.

[45] S. M. Robinson. Analysis of sample-path optimization. *Math. Oper. Res*, 21 (1996), pp. 513-528.

[46] R. T. Rockafellar. The multiplier method of Hestenes and Powell applied to convex programming. *J. Optimiz. Theory. App.*, 12 (1973), pp. 555-562.

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

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

[49] A. Shapiro. *Monte Carlo sampling methods.* in Stochastic Programming, Handbooks in Operations Research and Management Science (Volume 10), Elsevier Science, 2003.

[50] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory.* SIAM, PA, US, 3rd edition, 2021.

[51] A. Shapiro, T. Homem-de Mello, and J. Kim. Conditioning of convex piecewise linear stochastic programs. *Math. Program.*, 94 (2002), pp. 1-19.

[52] A. Shapiro and H. Xu. Stochastic mathematical programs with equilibrium constraints, modelling and sample average approximation. *Optimization*, 57 (2008), pp. 395-418.

[53] A. Shapiro and H. Xu. Uniform laws of large numbers for set-valued mappings and subdifferentials of random functions. *J. Math. Anal. Appl.*, 325 (2007), pp. 1390-1399.

[54] M. Sundermeyer, R. Schlüter, and H. Ney. LSTM neural networks for language modeling. In *Proceedings of the 13th Annual Conference of the International Speech Communication*, 2012, pp. 194-197.

[55] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optimiz. Theory. App.*, 109 (2001), pp. 475-494.

[56] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program.*, 117 (2009), pp. 387-423.

[57] Y. Wang, C. Zhang, and X. Chen. An augmented Lagrangian method for training recurrent neural networks. *SIAM J. Sci. Comput.*, 47 (2025), pp. C22-C51.

[58] S. J. Wright. Coordinate descent algorithms. *Math. Program.*, 151 (2015), pp. 3-34.

[59] N. Xiao, K. Ding, X. Hu, and K.-C. Toh. Developing lagrangian-based methods for nonsmooth nonconvex optimization. *arXiv preprint arXiv:2404.09438*, 2024.

[60] H. Xu. Sample average approximation methods for a class of stochastic variational inequality problems. *ASIA. Pac. J. Oper. Res.*, 27 (2010), pp. 103-119.

[61] H. Xu. Uniform exponential convergence of sample average random functions under general sampling with applications in stochastic programming. *J. Math. Anal. Appl.*, 368 (2010), pp. 692-710.

[62] H. Xu and J. J. Ye. Approximating stationary points of stochastic mathematical programs with equilibrium constraints via sample averaging. *Set-valued Anal.*, 19 (2011), pp. 283-309.

[63] H. Xu and D. Zhang. Smooth sample average approximation of stationary points in nonsmooth stochastic optimization and applications. *Math. Program.*, 119 (2009), pp. 371-401.

[64] M. Xu, J. J. Ye, and L. Zhang. Smoothing augmented lagrangian method for nonsmooth constrained optimization problems. *J. Global. Optim.*, 62 (2015), pp. 675-694.

[65] Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM. J. Imaging Sci.*, 6 (2013), pp. 1758-1789.

[66] Y. Xu and W. Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *J. Sci. Comput.*, 72 (2017), pp. 700-734.

[67] J. J. Ye. Multiplier rules under mixed assumptions of differentiability and Lipschitz continuity. *SIAM J. Control Optim.*, 39 (2000), pp. 1441-1460.

[68] X. Zhang, N. Gu, and H. Ye. Multi-GPU based recurrent neural network language model training. In *Proceedings of the International Conference of Pioneering Computer Scientists*, 2016, pp. 484-493.

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