On Convergence Analysis of Stochastic and Distributed Gradient-Based Algorithms

by

Mengyao Zhang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Applied Mathematics

Waterloo, Ontario, Canada, 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Yongqiang Wang Associate Professor, Dept. of Electrical and Computer Engineering, Clemson University
Supervisor(s):	Jun Liu Associate Professor, Dept. of Applied Mathematics, University of Waterloo Xinzhi Liu Professor, Dept. of Applied Mathematics, University of Waterloo
Internal Member:	Sue Ann Campbell Professor, Dept. of Applied Mathematics, University of Waterloo Lilia Krivodonova Professor, Dept. of Applied Mathematics, University of Waterloo
Internal-External Member:	Sherman Shen Professor, Dept. of Electrical and Computer Engineering, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Optimization is a fundamental mathematical discipline focused on finding the best solution from a set of feasible choices. It is vital in various applications, including engineering, economics, data science, and beyond. Stochastic optimization and distributed optimization are crucial paradigms in the optimization field. Stochastic optimization deals with uncertainty and variability in problem parameters, providing a framework for decision-making under probabilistic conditions. On the other hand, distributed optimization tackles large-scale problems by harnessing the collective power of multiple agents or nodes, each with local information and local communication capabilities. This thesis aims to modify and analyze the existing stochastic methods and develop the algorithms and theory to solve the unconstrained distributed optimization problem.

For stochastic adaptive gradient-based methods, including Root Mean Square Propagation (RMSprop), Adaptive Delta (Adadelta), Adaptive Moment Estimation (Adam), Adaptive Gradient (AdaGrad), Nesterov-Accelerated Adaptive Moment Estimation (Nadam), and Accumulate-Squared Moving Average Gradient (AMSgrad), which are popular stochastic optimization methods commonly used in machine learning and deep learning, the Chapter 2 provides a concise and rigorous proof of the almost sure convergence guarantees towards a critical point in the context of smooth and non-convex objective functions. To the best of our knowledge, this work offers the first almost sure convergence rates for these stochastic adaptive gradient-based methods. For non-convex objective functions, we show that a weighted average of the squared gradient norms in each aforementioned method achieves a unified convergence rate of $o(1/t^{\frac{1}{2}-\theta})$ for all $\theta \in \left(0, \frac{1}{2}\right)$. Moreover, for strongly convex objective functions, the convergence rates for RMSprop and Adadelta can be further improved to $o(1/t^{1-\theta})$ for all $\theta \in \left(0, \frac{1}{2}\right)$. These rates are arbitrarily close to their optimal convergence rates possible.

As a locking-free parallel stochastic gradient descent algorithm, Hogwild! algorithm is commonly used for training large-scale machine learning models. In Chapter 3, we will provide an almost sure convergence rates analysis for Hogwild! algorithm under different assumptions on the loss function. We first prove its almost sure convergence rate on strongly convex function, which matches the optimal convergence rate of the classic stochastic gradient descent (SGD) method to an arbitrarily small factor. For non-convex loss function, a weighted average of the squared gradient, as well as the last iterations of the algorithm converges to zero almost surely. We also provide a last-iterate almost sure convergence rate analysis for this method on general convex smooth functions.

Another aspect of the research addresses the convergence rate analysis of the gradient-based distributed optimization algorithms, which have been shown to achieve computational efficiency

and rapid convergence while requiring weaker assumptions. We first propose a novel gradient-based proportional-integral (PI) algorithm in Chapter 4, and prove that its convergence rate matches that of the centralized gradient descent method under the strong convexity assumption. We then relax this assumption and discuss the local linear convergence of its virtual state for strictly convex cost functions. In Chapter 5, we propose the powered proportional-integral (PI) algorithm and prove its convergence in finite time under the assumption of strict convexity. Then, we discuss the fixed-time convergence of its virtual state for strongly convex cost functions. Finally, we demonstrate the practicality of the distributed algorithms proposed in this thesis through simulation results.

Acknowledgements

I am immensely grateful to my esteemed supervisors, Prof. Jun Liu and Prof. Xinzhi Liu, for providing me with invaluable opportunities to pursue a Ph.D.. With their assistance, I have acquired a wealth of research knowledge. I am particularly thankful to Prof. Jun Liu for sponsoring additional scholarships, for revising my papers word by word, and for his concern.

I would also like to thank my examining committee members Prof. Sue Ann Campbell and Prof. Lilia Krivodonova, my internal-external examiner Prof. Sherman Shen, and my external examiner Prof. Yongqiang Wang for taking their time reading my thesis and providing the valuable comments.

I am also grateful to my family and my friends for their unconditional support and assistance.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Adadelta Adaptive Delta iv, x, 5, 6, 10, 12, 15–17, 22, 29–32, 36, 42–44, 96

AdaGrad Adaptive Gradient iv, 5, 6

Adam Adaptive Moment Estimation iv, x, 5, 6, 10, 12–15, 17, 19, 28, 36, 42, 44, 52, 96, 97

AMSgrad Accumulate-Squared Moving Average Gradient iv, x, 5, 6, 10, 12, 17, 18, 22, 28, 29, 40–42, 96, 97

ASG Adaptive Stochastic Gradient 1, 12, 20, 22, 33, 42, 44, 46, 96, 97

Nadam Nesterov-Accelerated Adaptive Moment Estimation iv, x, 6, 10, 12, 15–17, 22, 23, 40–42, 96, 97

ODE Ordinary Differntial Equation 8, 12, 21, 23, 28, 30, 31, 41, 97, 108, 111

PI Proportional-Integral x, 2, 3, 8, 10, 11, 52, 53, 72, 74–78, 93, 95, 96

RMSprop Root Mean Square Propagation iv, x, 5, 6, 10, 12, 14, 15, 17, 22, 29, 30, 34, 42, 43, 96

SGD Stochastic Gradient Descent 1, 4–6, 44, 51, 96, 97

List of Symbols

To ensure clarity, we have compiled a list of frequently used notations along with their respective meanings below.

- $\mathbb{R}^n$: the set of all n-dimensional real vectors.
- $\mathbb{R}_+$: the set of all positive real numbers.
- $\mathbb{C}^n$: the set of all n-dimensional complex vectors.
- $\mathbb{R}^{n\times n}$: the set of all $n \times n$ real matrices.
- $\mathbb{O}^{n\times n}$: the set of all $n \times n$ orthogonal matrices.
- x^*: the minimizer of the overall objective function f, i.e.

$$x^* = \text{argmin} x \in \mathbb{R}^m f(x).$$

- $\mathbf{1}$: the all-ones vector in $\mathbb{R}^n$.
- $\mathbf{0}$: the all-zeros vector in $\mathbb{R}^n$.
- $\|\cdot\|$: the Euclidean norm for vectors and spectral norm for matrices.
- $\nabla^2 g$: the Hessian matrix of a twice differentible function g.
- I_n : the n-dimensional identity matrix.
- $\otimes$: the Kronecker product.
- λ_n, $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$: the n-th largest, smallest and largest eigenvalues of a matrix.

- $C^2(\mathbb{R}^m, \mathbb{R})$: the collection of twice differentible functions, where $\mathbb{R}^m$ is the domain and $\mathbb{R}$ is the codomain.
- $C[\alpha_1, \alpha_2]$: the collection of continuous real-valued functions defined on $[\alpha_1, \alpha_2] \subset \mathbb{R}$.
- $B(x, r)$: the set $\{y \in \mathbb{R}^{\dim x} \,|\, \|x - y\| \leq r\}$, where $\dim x$ is the dimension of x.
- $\mathbb{R}^n$: the set of all n-dimensional real vectors.
- $\mathbb{R}^n_+$: the set of all n-dimensional real vectors with nonnegative real elements.
- $z_i \in \mathbb{R}$ $(z \in \mathbb{R}^N)$: the i-th element of z if $z \in \mathbb{R}^N$.
- $z^i \in \mathbb{R}^m$ $(z \in \mathbb{R}^{nm}$ or $\mathbb{R}^{2nm})$: the j-th element of z^i is the $(im - m + j)$-th elements of z, i.e., $z = ((z^1)^T, ..., (z^n)^T)^T$ or $((z^1)^T, ..., (z^{2n})^T)^T$.
- $\|z\|_p$ $(z \in \mathbb{R}^N)$ $(p \geq 1)$: the p-norm of vector z is

$$\|z\|_p = \left(\sum_{i=1}^{N} |z_i|^p \right)^{\frac{1}{p}}.$$

Chapter 1

Introduction

1.1 Background

This thesis aims to formulate and theoretically analyze algorithms designed for solving unconstrained optimization problems. The overarching goal is to optimally minimize the objective function through distributed or stochastic approaches. Emphasis is placed on the enhancement of existing algorithms and a rigorous analysis of their convergence behaviors for objective functions under various non-convex and convex conditions.

One of the main challenges in solving machine learning problems is the involvement of large datasets. Stochastic gradient-based methods, by their nature, update the model parameters using only a small subset (or batch) of the entire dataset at a time. These approaches are far more efficient than traditional methods that require the entire dataset for each update. The Stochastic Gradient Descent (SGD) method is the simplest stochastic method for searching for a minimizer. Stochastic adaptive gradient-based methods and Hogwild! algorithm are both variants of SGD. The former can dynamically adjust the learning rate during training, which leads to more efficient training and faster convergence. The latter is not only highly scalable, but also allows all processors access to shared memory and implement SGD without any locking.

Another challenge in machine learning area is the sparseness of dataset. In many real-world datasets, especially in fields like natural language processing, the data can be highly sparse. Stochastic gradient-based methods, for example Adaptive Stochastic Gradient (ASG) methods and Hogwild! algorithm, are particularly effective when dealing with sparse data.

Discussing distributed optimization in machine learning is crucial due to its relevance in handling large-scale data, improving computational efficiency, ensuring scalability, and addressing

challenges in privacy and data security. On the one hand, datasets are often too large to be processed on a single machine in fields like image and video processing, natural language processing, and large-scale simulations. On the other hand, training complex models, such as deep neural networks, requires significant computational resources. Distributed systems allows for handling such large datasets by dividing the workload across multiple machines, and it provides the necessary computational power by harnessing the capabilities of multiple machines, leading to faster training times and more efficient use of resources. Distributed approaches are also inherently scalable. As the size of the data or the complexity of the model increases, more nodes can be added to the system to maintain or improve the performance. In real-world, data might be collected and processed in distributed environments. Training models on diverse datasets distributed across multiple nodes can potentially lead to more robust and generalized models, as the model is exposed to a wider variety of data samples during training. In some cases, data cannot be centralized due to privacy concerns or regulatory restrictions. Distributed learning allows data to remain on local nodes while still contributing to the learning of a global model.

Remark 1. *As shown in Figure 1.1, Hogwild! and the PI algorithm are distributed algorithms that cater to different system models. The Hogwild! model resembles centralized systems, featuring a server (or master) with a shared memory x, with each processor acting like a computer (or slave). The key difference lies in the fact that processors in Hogwild! bypass the need for a central synchronization point, allowing them to work independently and without waiting for each other. On the other hand, the PI algorithm is used in a system without shared memory or a central server (master). This setup involves different nodes sharing hardware, software, or data and communicating through a shared network without relying on a single point of control for synchronization.*

1.2 Motivation Example

A motivation example is to solve a deep learning problem for image classification with convolutional Neural Networks (CNNs) Scenario. The task is to develop a CNN for classifying images into various categories, such as distinguishing between different types of animals in a dataset, and it is usually converted to an unconstrained minimization problem. This is a common problem in computer vision with applications in areas like automated image tagging, surveillance, and medical imaging.

Consider the simplest image classification: binary image classification involving single data point $(x', y(x'))$, where x' is a single feature input of the image and $y(x')$ is a binary true label vector. The c-th element of $y(x')$, $y_c(x')$, denoted by 1 if the class label c is the correct classification, and 0 otherwise. In other words, $y(x')$ is $[0,1]$ if the data point belongs to class 1, and $y(x')$

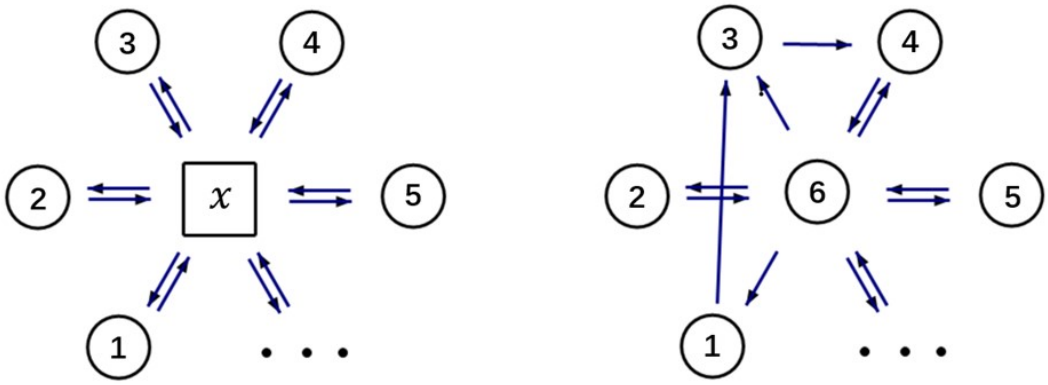

Figure 1.1: The left one is the distributed model for Hogwild! algorithm where each circle represents a processor, and all of them have access to the shared memory x. The right one is the distributed model for PI algorithm where each circle represents a processor, and they are allowed to communicate with their neighborhood.

is $[1,0]$ if it belongs to class 2. Let $\hat{y}(x',x)$ be the predicted probability distribution vector across the classes, where x is the parameter to be determined, the c-th element of $\hat{y}$, denoted by $\hat{y}_c(x,x')$ is the predicted probability of class c.

Our objective to find x such that the difference between the predicted probability vector and true probability vector are minimized, which can be mathematically formulated by

$$\min_x L(y,\hat{y}), \tag{1.1}$$

where L represents some measurement quantifying the difference between y and $\hat{y}$. A commonly used measurement is the cross-entropy loss [7,26,68] of the distribution $\hat{y}$ relative to a distribution y, and our goal is to optimize

$$\min_x L(y(x),\hat{y}(x)) \stackrel{\Delta}{=} -y_1(x)\log(\hat{y}_1(x)) - y_2(x).\log(\hat{y}_2(x)) \tag{1.2}$$

In practice, image datasets are typically vast in scale. For instance, ImageNet comprises over 14 million images, while ObjectNet contains upwards of 50 thousand images. We now consider binary image classification involving multiple data points, and extend this problem to a

distributed optimization problem, where each machine (node or processor) has access to a small batch of the large dataset, and only local communication and local computation are allowed. Each node computes gradients based on its subset of data, and these gradients are then combined to update the global model parameters. The objective function in a distributed setting is typically a summation of the local loss functions computed on each node, with an additional term to ensure regularization or consensus among the nodes. In the context of image classification using cross-entropy loss, the distributed optimization problem can be formulated as:

$$\min_x L_{\text{distributed}}(x) \triangleq \sum_{k=1}^{K} \frac{n_k}{K} L_k(x), \tag{1.3}$$

where $L_{\text{distributed}}$ is the global loss function dependent on the model parameter x, K is the total number of nodes in the distributed system, n_k is the number of samples in the k-th node, L_k is the local loss function for the k-th node, which is the cross-entropy loss function (1.2) for its subset of data [16, 49].

1.3 Stochastic Gradient-based Optimization Methods

Consider the unconstrained minimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{1.4}$$

where $f : \mathbb{R}^n \to \mathbb{R}$. Here, f may arise from optimizing an expected risk of the form $f(x) = \mathbb{E}(f(x;\xi))$, where ξ is a source of randomness indicating a sample (or a set of samples), or an empirical risk of the form $f(x) = \frac{1}{n}\sum_{i=1}^{n} f(x;\xi_i)$, where $\{\xi_i\}_{i=1}^{n}$ are realizations of ξ. For all $x \in \mathbb{R}^n$, we assume that $\nabla f(x;\xi)$ is an unbiased estimator of the actual gradient $\nabla f(x)$, i.e.,

$$\nabla f(x) = \mathbb{E}(\nabla f(x;\xi)). \tag{1.5}$$

The SGD is the simplest stochastic method for searching for a minimizer.

As a variant of stochastic gradient descent (SGD), Hogwild! combines the simplicity of SGD with accelerated learning on multi-core processors. Hogwild! (Higher-Order, Gradient-Wise learning for the Wild) is designed as a parallelize stochastic gradient descent method for training machine learning models. Introduced by Niu et al. [63], this distributed and parallel optimization algorithm removes the constraints of traditional locking mechanisms in SGD, enabling multiple CPU cores to update model parameters asynchronously, and scaling up model

training in data-intensive applications efficiently. In the next few years, numerous variations and extensions of Hogwild! algorithm has been proposed. As a variant of Hogwild! algorithm, Hogwild++ [94] extends the original algorithm by introducing decentralized averaging mechanisms to improve convergence while retaining the asynchrony. It addresses challenges related to communication overhead as well as non-convex optimization tasks. SPIDER [21] is an extension of HOGWILD! for non-convex optimization tasks. It combines stochastic gradient methods with variance-reduction techniques for improved efficiency. BUCKWILD! [15] is designed and analyzed as an asynchronous SGD algorithm, that uses lower-precision arithmetic. It has been demonstrated experimentally that BUCKWILD! can achieve speedups of up to 2.3 times over HOGWILD!-based algorithms for logistic regression.

Another variant of SGD, stochastic adaptive gradient-based methods can adaptively tunes learning rates for each parameter based on historical gradients. They help accelerate convergence, handle sparse data more effectively, and often lead to better optimization performance for complex models like deep neural networks and large-scale data analytics [38].

The first adaptive optimization algorithm, AdaGrad, was introduced by Duchi et al. in 2011 [19]. This algorithm proves to be particularly effective for handling sparse data. Over the next two years, additional methods such as RMSprop [27] and Adadelta [93] were proposed. RMSprop is an algorithm that rescales the step size using a weighted moving average of the squared gradient. On the other hand, Adadelta is an extension of RMSprop, designed to tackle the diminishing learning rate problem by utilizing a running average of past updates. Furthermore, Adadelta removes the necessity of setting an initial learning rate. In 2014, the Adam method and Adamax [39] were introduced, combining the strengths of both AdaGrad and RMSprop. Adam quickly gained popularity as a preferred choice for optimizing deep-learning models. Its widespread adoption can be attributed to its robustness, rapid convergence, and user-friendly default parameter settings. The same paper also proposed 'Adamax,' a variant of Adam that relies on the infinity norm. In [64], however, a rigorous proof was presented, demonstrating that there exists a stochastic optimization problem for which Adam fails to converge to the optimal solution. To address this limitation, AMSgrad was introduced in [64]. A study on the Nesterov-accelerated Adaptive Moment Estimation (Nadam) method is documented in [18], where the first momentum of Adam is replaced with the momentum used in Nesterov's Accelerated Gradient method (NAG).

For various adaptive optimization methods, some papers have proven the convergence to critical points in different optimization settings. The first convergence guarantees for adaptive optimization methods might be presented in [14]. This work provides convergence rates for the gradients of deterministic RMSprop and Adam algorithms, where the full gradient is calculated in each iteration. Additionally, the paper offers a convergence rate for stochastic RMSprop, assuming that all elements in gradients share the same sign. For large-scale non-convex stochastic op-

timization, [99] demonstrated the global decays of the gradients in stochastic Adam, RMSprop, and weighted AdaGrad with exponential moving average momentum (weighted AdaEMA) with a probability less than one. This paper also provides the convergence rate. In the non-convex setting, [30] established that the averages of the gradient squares in SHB, Adam, AMSgrad, AdaGrad, Adaform, and Adabound converge in expectation, and it provides the convergence rate by analyzing the Jacobian matrices. Conducting a locally exponential convergence analysis in batch mode for a deterministic fixed training set, the paper [9] presents insights into the performance of Adam methods. Moreover, the papers [42] and [85] consider the AdaGrad method as a variant of SGD with adaptive step size and provide a convergence guarantee for the minimum history gradient norm. It is proved in [17] that the squared norms of the gradients of Adam and AdaGrad share the same convergence rate $O(ln(N)/\sqrt{N})$ in expectation.

The research paper [4] introduces some modifications to the Adam optimization method. These enhancements incorporate the Robbins-Monro Algorithm (RMA), ensuring the modified Adam method's almost sure convergence to the critical point of the non-convex objective function. Addtionally, some methods of proving the almost convergence rates of SGD are proposed in [45].

Motivated by the above methods, we will apply RMA in proving the last-iterate almost sure convergences of stochastic adaptive gradient-based methods, including RMSprop, Adadelta, Adam, Nadam, and AMSgrad, with a more concise and accurate proof, whereas the proof in [4] has some mistakes.

To the best of our knowledge, the work to be presented in this thesis gives the first almost sure convergence rates for these stochastic adaptive gradient-based methods and Hogwild! algorithm. Using the results from [45], we show that a weighted average of the squared gradient norms for non-convex objective function achieves a unified $o(1/t^{\frac{1}{2}-\theta})$ convergence rate for all $\theta \in (0, \frac{1}{2})$ For strongly convex objective functions, the convergence rates of RMSprop, Adadelta and Hogwild! can be improved to $o(1/t^{1-\theta})$ for all $\theta \in (0, \frac{1}{2})$, which are arbitrarily close to their optimal convergence rates possible.

1.4 Distributed Gradient-based Optimization Methods

The overall objective is to find the value of x that minimizes the average of a collection of objective functions f_i through local communication with neighbors and local computation. Mathe-

matically, we express this problem as follows:

$$\min_{x\in\mathbb{R}^m} f(x) \triangleq \sum_{i=1}^{n} f_i(x), \tag{1.6}$$

Here, the local communication is defined by an undirected communication graph. The coordination of multi-agent dynamic systems in networks has gained significant attention recently, finding applications in various fields such as flocking of social insects, formation control, robotics, control engineering, economics, transportation, and social networks [57, 58, 79, 89].

Consider a network consisting of n agents $V = \{v_1, ..., v_n\}$, each of which has a convex objective function $f_i : \mathbb{R}^m \to \mathbb{R}$ to optimize. Denote by $X = \{1, 2, ..., n\}$ the set of indices of the agents. A weight matrix W is a symmetric matrix such that $w_{ij} \geq 0$ for all $i, j \in X$, and $w_{ii} = 0$ for $i \in X$. We say that a set (v_i, v_j) is an edge of the graph if and only if $w_{ij} > 0$, and we denote by $E = \{(v_i, v_j) \,|\, w_{ij} > 0\}$ the collection of all edges. The agents communicate through the corresponding weighted undirected communication graph $G = (V, E, W)$.

Definition 1. *Given a weighted undirected communication graph $G = (V, E, W)$, define the graph Laplacian matrix $L = [l_{ij}]$ of G by*

$$l_{ij} = \begin{cases} \sum_{k=1, k\neq i}^{n} w_{ik}, & j = i, \\ -w_{ij}, & j \neq i. \end{cases} \tag{1.7}$$

Definition 2. *An undirected graph is connected if there is a path between every pair of agents, where a path is a sequence of consecutive edges.*

Definition 3. *A network is said to reach consensus if all agents have the same value.*

The overall network objective is to find the x^* which satisfies

$$f(x^*) = \min_{x\in\mathbb{R}^m} f(x) \tag{1.8}$$

in a distributed manner, that is to say, an agent $v_i \in V$ is allowed to communicate with its neighborhood $N_i = \{v_j \in V | (v_i, v_j) \in E\}$ only.

Continuous-time optimization is important because real-world processes, such as physical systems, economic processes, and financial markets, evolve continuously over time. These processes are often described by differential equations that cannot be easily solved analytically. Continuous-time optimization provides a powerful tool for finding optimal solutions in these complex systems. Furthermore, continuous-time distributed optimization algorithms can offer a

dynamic perspective and provide physical insights that complement their discrete-time counterparts [53, 74]. In [95, 97], the authors accelerate the Heavy-ball method and Nesterov's gradient method by directly discretizing an Ordinary Differntial Equation (ODE) related to their continuous limits. This approach allows for leveraging the benefits of continuous-time optimization in the context of discrete-time algorithms.

Among the distributed optimization algorithms, continuous-time ones are often easier to understand and analyze, and many of them exhibit asymptotic behavior or linear convergence. The continuous-time distributed gradient descent method [98] is perhaps the simplest continuous-time distributed optimization algorithm, which is motivated by the discrete-time version and shares the same convergence rate with the discrete-time distributed and centralized gradient descent method. Motivated by the feedback control mechanism, Wang and Elia designed the gradient-based PI control strategy in [78]. Its asymptotic convergence for convex local cost functions [25] and linear convergence for strongly convex local cost functions [37] were discussed. Building on Kia et al.'s PI framework [37], Guo et al. investigated the distributed optimization problem of double-integrator multi-agent systems with unmatched constant disturbances and time-triggered communication [28]. Lu and Tang proposed the zero-gradient-sum (ZGS) algorithm, where the sum of the local gradients is always zero, and they proved its exponential stability for strongly convex cases [47]. Over the past few years, it has been shown that the ZGS algorithms can also solve the unconstrained optimization problem subject to switching topology and time-varying communication delays [29, 44]. However, these ZGS algorithms rely on the inverse of local Hessian matrices and are resource-intensive [29, 44, 47]. Despite the convergence guarantees of all these algorithms, there is no assurance that they can solve the distributed optimization problem in finite time. In order to improve traditional processing methods, stochastic and event-triggered algorithms have been investigated to achieve communication and computation efficiency [35, 48, 72, 90].

Various finite-time and fixed-time continuous-time distributed algorithms have been proposed [22, 34, 59, 81, 83, 84]. Finite-time algorithms depend on the initial value, while fixed-time algorithms have a predetermined settling time. Some distributed finite-time optimization algorithms have been developed based on the assumption that all local cost functions are strongly convex quadratic functions [22, 59, 81, 84]. These algorithms can handle mismatched disturbances or uncertain information. Feng and Hu proposed a finite-time distributed strongly convex optimization algorithm that incorporates disturbances using the inverse of Hessian matrices of the objective functions [23]. Wang et al. presented a distributed optimal signal generator that solves finite-time optimization problems when the local cost functions are quadratic-like and the overall cost function is strongly convex [83]. Importantly, this algorithm only relies on the gradients of local objective functions and does not require convexity of the global cost functions. Garg et al. developed two fixed-time distributed algorithms for time-varying topologies: one

Methods	Convergence	Rate	Global cost function	Local cost functions	Computation
[78]	asymptotic		compact solution set &convex	convex	gradients
[98]	exponential	μ	μ-strongly convex	μ-strongly convex	Hessian matrices
[47]	exponential	$< \mu$	μ-strongly convex	μ-strongly convex	the inverse of Hessian matrices
[37]	exponential	$< \mu$	μ-strongly convex	μ-strongly convex	gradients
Th 9	exponential	μ	**μ-strongly convex**	**μ-strongly convex**	**gradients**
Th 10	exponential		**strictly convex**	**convex**	**gradients**

Table 1.1: Theoretical comparisons with existing results in the literature

based on third-order derivatives for strictly convex global cost functions, and another based on Hessians for strongly convex global cost functions [24]. Song and Chen combined the ZSG algorithm from [47] and the powerball method from [92] to propose a finite-time ZSG algorithm for strongly convex objective functions [73]. Wu et al. investigated a distributed algorithm for finite-time and fixed-time optimization problems based on the ZGS framework [87]. Assuming all local cost functions are strongly convex, Shi et al. proposed a finite-time convergent distributed approach for time-varying distributed optimization [70]. In [46], a predefined-time multi-agent algorithm for solving multi-objective optimization is presented, where predefined-time optimization is a method of optimization capable of reaching a state very near to an optimal solution within a specified time frame. In summary, these algorithms can be computationally intensive or rely on strong assumptions.

In the Table 1.1, 1.2, we summarize the aforementioned results on finite-time and fixed-time continuous-time distributed optimization algorithms. It is observed that the algorithms presented in this thesis have lower computational requirements and are based on more lenient assumptions. The computational complexity of these algorithms primarily relies on the gradients of the objective functions. We rigorously prove the fixed-time convergence of the first proposed algorithm under the assumption of strong convexity. Additionally, we develop a decentralized control framework to solve the finite-time optimization problem under the assumption of weak convexity.

1.5 Organization of the Thesis

The primary goal of this thesis is to contribute to theory and algorithms in the fields of stochastic and distributed optimization problems. The rest of the thesis is organized as follows.

Methods	Convergence	Global cost function	Local cost functions	Computation
[59]	finite time	strongly convex & quadratic	strongly convex & quadratic	gradients
[73]	finite time	strongly convex	strongly convex	the inverse of Hessian matrices
[34]	finite time	strictly convex	strictly convex	the inverse of Hessian matrices& positive Hessians
[82, 83]	finite time	strongly convex	quadratic-like	gradients
[24]	fixed time	strictly convex	convex	third-order derivatives of cost functions
	fixed time	strongly convex	convex	Hessian matrices
[84]	finite time	strongly convex quadratic	strongly convex quadratic	Hessian matrices
[87]	finite time	local strongly convex	local strongly convex	the inverse of Hessian matrices
	fixed time	strongly convex & quadratic	strongly convex & quadratic	the inverse of Hessian matrices
Th 11	fixed time	**strongly convex**	**strongly convex**	**gradients**
Th 12	finite time	**strictly convex**	**convex**	**gradients**

Table 1.2: Comparison of main results with other algorithms

Chapter 2 presents a detailed proof of the almost sure convergence for popular stochastic adaptive gradient methods like RMSprop, Adadelta, Adam, Nadam, Adamax,and AMSgrad in machine and deep learning. This chapter marks the first to offer almost sure convergence rates for these methods, demonstrating a unified convergence rate of $o(1/t^{\frac{1}{2}-\theta})$ for non-convex objectives and improved rates for strongly convex objectives. These rates are arbitrarily close to their optimal convergence rates possible.

In Chapter 3, the Hogwild! algorithm, a parallel stochastic gradient descent method, is analyzed for its almost sure convergence rates under various loss functions. We show its optimal convergence rate for strongly convex functions and convergence rate to zero for non-convex loss functions, including a detailed analysis for general convex functions.

Chapter 4 introduces a novel gradient-based PI algorithm, showing its convergence rate under strong convexity matches centralized gradient descent. We also explore its local linear convergence for strictly convex functions.

Chapter 5 discusses the powered PI algorithm, proving finite-time convergence for strictly convex functions and fixed-time convergence for strongly convex functions, supported by practical simulation results.

The final Chapter summarizes the main contributions and bring up some related future research directions.

Chapter 2

On Almost Sure Convergence Rates of Adaptive Stochastic Gradient Methods

We have seen the merits and developments of ASG methods in Chapter 1, and there are some open questions: (1) Adam with fixed learning rate has been proved to fail to converge to the optimal solution [64]. How could it be modified? Will this flaw influences other ASG methods? (2) the last-iterate almost sure convergence analysis on Adam has been given [4]. Do the last-iterates of other ASG methods converge as well? (3) what is the almost sure convergence rate of these methods?

This chapter points out the existing flaws on the widely-used stochastic adaptive gradient techniques such as RMSprop, Adadelta, Adam, Nadam, and AMSgrad, and it is confirmed by simulations. A unified modification to these methods are given Motivated by the almost sure convergence analysis on Adam [4], we construct the limiting ODE of the above-mentioned methods, analyze their stability, and present last-iterate almost sure convergence analysis on non-convex smooth functions. Using the conclusions from [45], we provide a unified almost sure convergence rate $o(1/t^{\frac{1}{2}-\varepsilon})$ $(\varepsilon \in (0,1/2))$ for non-convex cost functions, and offer enhanced rates $o(1/t^{1-\varepsilon})$ $\varepsilon \in (0,1)$ for strongly convex objective functions.

2.1 Nonconvergence of Adaptive Methods

This section provides an illustrative example where some ASG methods with constant parameters do not converge toward the critical point. To address this issue, we propose a unified modification to these methods, resulting in convergence towards the critical point in this case.

2.1.1 Motivation Example

Consider the stochastic optimization problem in [64, Theorem 3], the Adam method defined in Algorithm 1 with constant α_t and β_t fails to converge to the optimal solution. In this case, the dramatic vibration of v_t, m_t causes them flawed estimations on the first and second raw moments, and $\mathbb{E}(x_{t+1}-x_t)$ might have the same direction as the expected gradient $\mathbb{E}(-\nabla f(x_t;\xi)) = -\nabla f(x_t)$, preventing x_t from approaching the optimal solution. As there is no critical point in this example, we propose a different stochastic optimization problem such that some adaptive methods with constant parameters, including Adam method, fail to converge towards the critical point if α_t and β_t are constant.

Example 1. Construct a strongly convex optimization problem

$$f(x;\xi) = \begin{cases} 5x^2+9x, & \text{with probability } 0.1, \\ -\frac{1}{2}x^2 - x, & \text{with probability } 0.9, \end{cases} \tag{2.1}$$

and the corresponding stochastic gradient is

$$\nabla f(x;\xi) = \begin{cases} 10x+9, & \text{with probability } 0.1, \\ -x-1, & \text{with probability } 0.9. \end{cases} \tag{2.2}$$

The feasible set is set to be $\mathscr{F} = [-1,1]$, i.e., if the x_t derived in the iteration exceeds the interval, we set $x_t \in \mathscr{F}$ to be the value closest to the iteration result. Noting that $f(x) = \mathbb{E}(f(x;\xi)) = 0.05x^2$, we know that the optimal solution occurs at the origin.

2.1.2 Modification on Adaptive Methods

Decreasing learning rate in stochastic algorithms is commonly employed. It allows the convergence to an optimal or near-optimal solution, enhances the stability of the optimization process, gradually balances the trade-off between exploration and exploitation, and so on. Except for the requirement of a decreasing learning rate, the research paper by Barakat et al. [4] also requires the increasing hyper-parameters α_t and β_t, and it is proven that the sequence of x_t in this context of Adam converges to the optimal solution with probability 1.

Motivated by the above arguments, we apply a similar modification as [4] on other adaptive methods, i.e., we assume the following holds

Assumption 1. *The following requirements hold:*

Algorithm 1: Adam$(\{\gamma_t\}_t, \{\alpha_t\}_t, \{\beta_t\}_t, \varepsilon)$

Data: $m_0 = v_0 = 0$

for $t = 1, 2, \ldots$ **do**

$g_t = \nabla f(x_{t-1}; \xi_t)$;

$m_t = \alpha_t m_{t-1} + (1-\alpha_t) g_t$;

$v_t = \beta_t v_{t-1} + (1-\beta_t) g_t^{\odot 2}$;

$\hat{m}_t = \frac{m_t}{1-\Pi_{i=1}^t \alpha_i}$ (bias correction to m_t);

$\hat{v}_t = \frac{v_t}{1-\Pi_{i=1}^t \beta_i}$ (bias correction to v_t);

$x_t = x_{t-1} - \gamma_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon \mathbf{1}}}$;

end

Algorithm 2: RMSprop$(\{\gamma_t\}_t, \{\beta_t\}_t, \varepsilon)$

Data: $v_0 = 0$

for $t = 1, 2, \ldots$ **do**

$g_t = \nabla f(x_{t-1}; \xi_t)$;

$v_t = \beta_t v_{t-1} + (1-\beta_t) g_t^{\odot 2}$;

$x_t = x_{t-1} - \gamma_t \frac{g_t}{\sqrt{\hat{v}_t + \varepsilon \mathbf{1}}}$;

end

1. $\sum_t \gamma_t = +\infty$ *and* $\sum_t \gamma_t^p < +\infty$ *for some* $p \geq 2$.
2. *There exist* $a, b > 0$ *such that* $b(1-\varepsilon) \leq 4a$, $\lim_{t\to\infty} \frac{1-\alpha_t}{\gamma_t} = a$ *and* $\lim_{t\to\infty} \frac{1-\beta_t}{\gamma_t} = b$.

or

Assumption 2. *The following requirements hold:*

1. $\sum_t \gamma_t = +\infty$ *and* $\sum_t \gamma_t^p < +\infty$ *for some* $p \geq 2$.
2. *There exists* $b > 0$ *such that* $\lim_{t\to\infty} \frac{1-\beta_t}{\gamma_t} = b$.

For adaptive methods involving parameters α_t and β_t, such as the Adam method, we assume the validity of Assumption 1. In cases where only β_t is present, such as RMSprop method, we consider Assumption

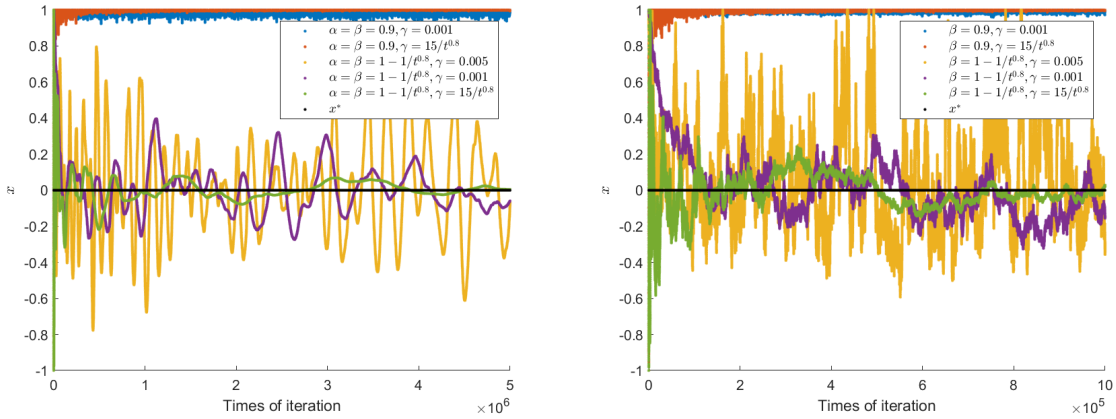

Figure 2.1: Simulations of Adam (left) and RMSprop (right) with different combinations of constant or time-varying parameters.

Adam and RMSprop, of which the iterations are given in Algorithms 1 and 2, are conducted with respect to different settings of the hyper-parameters α_t, β_t, and γ_t, and the simulation results are presented in Figure 2.1. When both α_t and β_t are kept constant, the values of x_t remain around the 'worst' solution within the feasible set $\mathscr{F}$ regardless of which step size is employed. When α_t and β_t increase and the learning rate γ_t maintains a constant, the variable x_t exhibits a damping effect, settling around the vicinity of the origin. It appears that using a smaller constant step size results in smaller oscillations and lower frequency of vibrations. The green lines meet Assumption 1. This combination demonstrates a more rapid and smoother convergence towards the optimal solution, and the almost sure convergence of the Adam method agrees with the result in [4]. In Sections 2.3 and 2.4, we will give the almost sure convergences guarantee and the almost sure convergence rates for both the Adam and RMSprop methods.

Nadam method [18], of which the iterations are given in Algorithm 3, follows a comparable argument. Figure 2.2 depicts a simulation of Example 1 conducted using the Nadam method. When α_t and β_t are constant, the trajectory converges to a neighbourhood of 1. When comparing Figure 2.2 with the upper figure in Figrue 2.1, it is evident that the yellow and purple lines exhibit less frequent damping. The green line, similar to the Adam method, will also be demonstrated to converge almost surely in Sections 2.3 and 2.4.

The constant value β_t also causes the non-convergence issue of the Adadelta method [93], defined in Algorithm 4. Setting ε to be 10^{-6}, we conduct Adadelta method on Example 1. It is observed from Figure 2.3 that x_t remains around 1 for a constant β_t. Increasing β_t and $\gamma_t = 1$ result in heavy isolation of x_t, and there is no tendency for the amplitude to decrease.

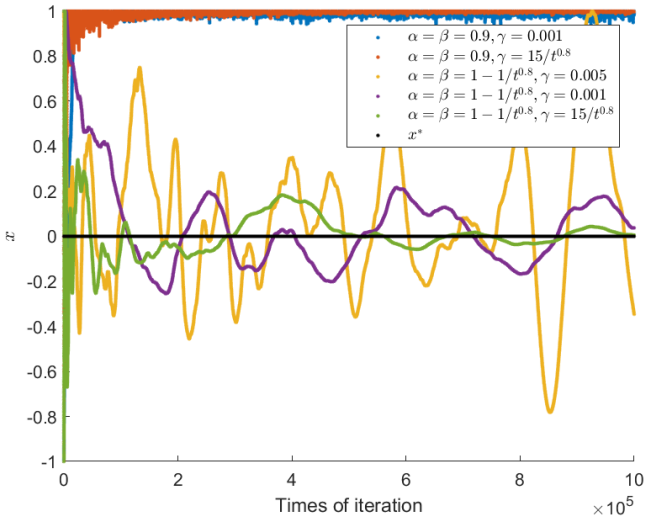

Figure 2.2: Simulations of Nadam with different combinations of constant or time-varying parameters.

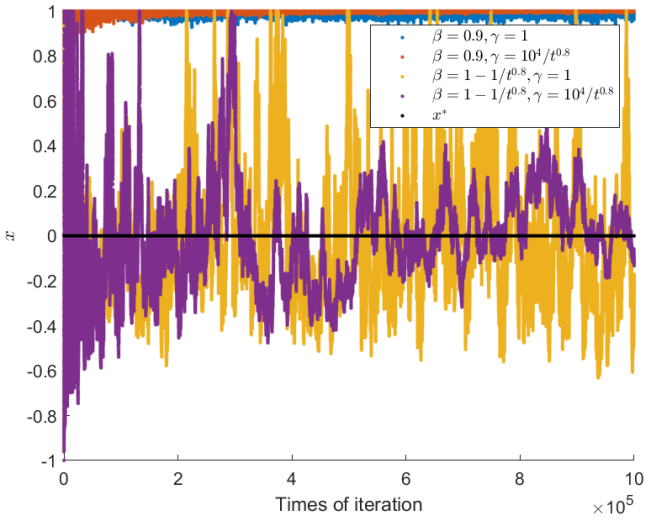

Figure 2.3: Simulations of Adadelta with different combinations of constant or time-varying parameters.

Algorithm 3: Nadam $(\{\gamma_t\}_t, \{\alpha_t\}_t, \{\beta_t\}_t, \varepsilon)$

Data: $m_0 = v_0 = 0$

for $t = 1, 2, \ldots$ **do**

$g_t = \nabla f(x_{t-1}; \xi_t)$;

$m_t = \alpha_t m_{t-1} + (1-\alpha_t) g_t$;

$v_t = \beta_t v_{t-1} + (1-\beta_t) g_t^{\odot 2}$;

$\hat{m}_t = \frac{\alpha_{t+1} m_t}{1-\Pi_{i=1}^{t+1}\alpha_i} + \frac{(1-\alpha_t) g_t}{1-\Pi_{i=1}^{t}\alpha_i}$ (bias correction to m_t);

$\hat{v}_t = \frac{v_t}{1-\Pi_{i=1}^{t}\beta_i}$ (bias correction to v_t);

$x_t = x_{t-1} - \gamma_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon \mathbf{1}}}$

end

Algorithm 4: Adadelta$(\{\gamma_t\}_t, \{\beta_t\}_t, \varepsilon)$

Data: $u_0 = v_0 = 0$

for $t = 1, 2, \ldots$ **do**

$g_t = \nabla f(x_{t-1}; \xi_t)$;

$v_t = \beta_t v_{t-1} + (1-\beta_t) g_t^{\odot 2}$;

$\Delta x_t = \frac{\sqrt{u_{t-1} + \varepsilon \mathbf{1}}}{\sqrt{v_t + \varepsilon \mathbf{1}}} g_t$;

$u_t = \beta_t u_{t-1} + \Delta x_t^2 (1-\beta_t)$;

$x_t = x_{t-1} - \gamma_t \Delta x_t$

end

By decreasing the learning rate and increasing β_t, the purple line in Figure 2.3 converged to the optimal solution with decreasing amplitude. This line's almost sure convergence behaviour will also be further explained in Sections 2.3 and 2.4.

Remark 2. *Unlike other adaptive methods, the Adadelta and RMSprop optimization methods display less smoothness and a considerable number of peaks in their plots. This disparity arises from using the term g_t in the iteration step of x_t instead of $\hat{m}_t$, which represents a weighted average of historical stochastic gradients. Consequently, the former methods exhibit more stochasticity, lacking the smoothing effect provided by historical gradients.*

The AMSgrad method, which iterates according to Algorithm 5, is a modification of the Adam method introduced in the paper [64]. In Figure 2.4, the red line appears smoother with smaller amplitudes compared with the red line. Additionally, by modifying α_t, β_t, and γ_t, the

Algorithm 5: AMSgrad $(\{\gamma_t\}_t, \{\alpha_t\}_t, \{\beta_t\}_t, \varepsilon)$

Data: $m_0 = v_0 = 0$

for $t = 1, 2, \ldots$ **do**

$g_t = \nabla f(x_{t-1}; \xi_t)$;

$m_t = \alpha_t m_{t-1} + (1 - \alpha_t) g_t$;

$v_t = \beta_t v_{t-1} + (1 - \beta_t) g_t^{\odot 2}$;

$\hat{m}_t = \frac{m_t}{1 - \Pi_{i=1}^{t} \alpha_i}$ (bias correction to m_t);

$\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$;

$x_t = x_{t-1} - \gamma_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon \mathbf{1}}}$

end

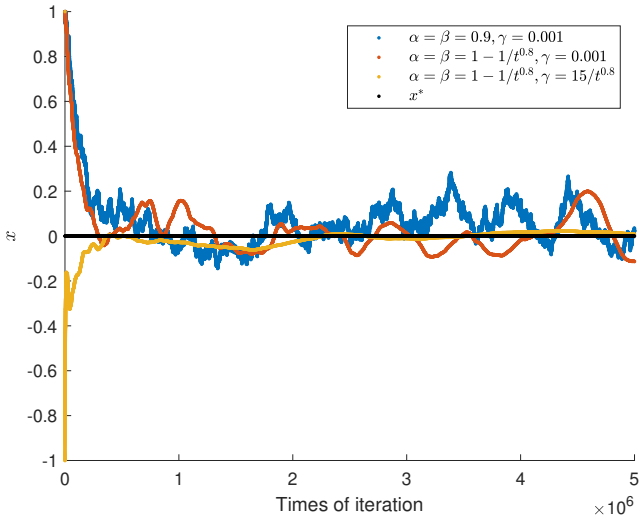

Figure 2.4: Simulations of AMSgrad with different combinations of constant or time-varying parameters.

variable x_t converges to the origin more rapidly and consistently. Its almost sure convergence will be discussed in Sections 2.3 and 2.4 as well.

Algorithm 6: Adamax $\{\gamma_t\}_t, \{\alpha_t\}_t, \{\beta_t\}_t, \varepsilon)$

Data: $m_0 = v_0 = 0$

for $t = 1, 2, \ldots$ **do**

$g_t = \nabla f(x_{t-1}; \xi_t)$;

$m_t = \alpha_t m_{t-1} + (1 - \alpha_t) g_t$;

$u_t = \max(\beta_t u_{t-1}, |g_t| + \varepsilon \mathbf{1})$;

$x_t = x_{t-1} - \gamma_t \frac{m_t}{(1 - \Pi_{i=1}^{t} \alpha_i) u_t}$

end

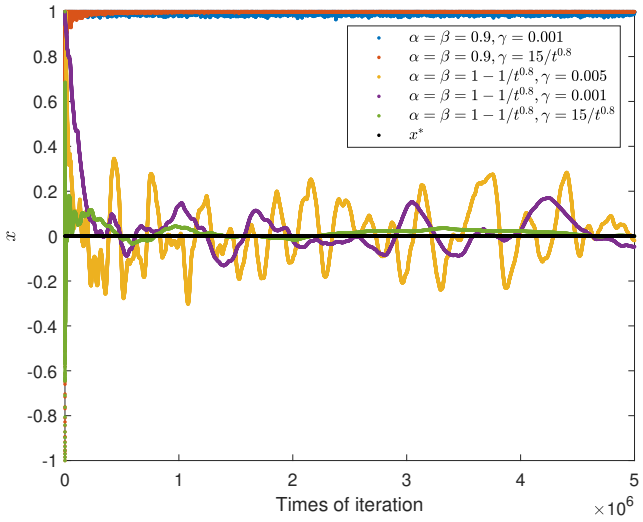

Figure 2.5: Simulations of Adamax with different combinations of constant or time-varying parameters.

Figure 2.5 for the Adamax method [39], which updates according to Algorithm 6, exhibits some similarity to those of the Adam method. The constant β_t here also leads to fluctuations in the variable u_t and can hinder the convergence process. Though the convergence of the green line will not be proven, this example illustrates that the modified parameters can enhance the performance of the Adamax method.

2.2 Assumptions

To analyze the almost sure convergence of the ASG methods, we make the following assumptions on the objective function.

Assumption 3. *The continuously differentiable cost function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the following:*

1. *f is coercive, i.e.,*
$$f(x) \to \infty \text{ as } \|x\| \to \infty. \tag{2.3}$$
2. *f has a global minimum value $f^* = f(x^*) = \min_{x\in\mathbb{R}^n} f(x)$, where x^* is the optimal solution.*
3. *∇f is locally Lipschitz, i.e., for any $x_0 \in \mathbb{R}^n$, there exists $\delta_0, L_0 > 0$ such that $\|x - x_0\| \le \delta_0$ implies*
$$\|\nabla f(x) - \nabla f(x_0)\| \le L_0 \|x - x_0\|. \tag{2.4}$$

Assuming that Assumption 3 is satisfied, and C is a compact set, we can use the Heine-Borel theorem to demonstrate that ∇f is Lipschitz continuous with some Lipschitz constant $L_C > 0$ over C, i.e.,
$$\|\nabla f(x) - \nabla f(y)\| \le L_C \|x - y\|, \quad x, y \in C. \tag{2.5}$$
A useful consequence of L_C-Lipschitzness of ∇f is the following inequality from [54, Lemma 1.2.3]:
$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_C}{2}\|y - x\|^2, \quad x, y \in C. \tag{2.6}$$
If we further assume that f is convex on C, i.e.,
$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle, \quad x, y \in C, \tag{2.7}$$
then it can be concluded from [54, Theorem 2.1.5] that
$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L_C}\|\nabla f(x) - \nabla f(y)\|^2, \quad x, y \in C. \tag{2.8}$$
In Assumption 3, the minimum point x^* is one of the critical points of f. In the following assumption, we assume that x^* is the unique critical point of f.

Assumption 4. *The set $\mathscr{E}\mathscr{P} = \{x \in \mathbb{R}^n | \nabla f(x) = 0\} = \{x^*\}$ is singleton. There exists some $\mu > 0$ such that the cost function is locally μ-strongly convex around the optimal solution x^*, i.e.,*
$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|x - y\|^2, \; x, y \in N^*, \tag{2.9}$$

where N^* is some neighborhood of x^*.

Combining the fact that $\nabla f(x^*) = 0$ and [54, Theorem 2.1.10], we know that (2.9) implies

$$f(x) - f^* \leq \frac{1}{2\mu}\|\nabla f(x)\|^2, \quad x \in N^*. \tag{2.10}$$

Assumption 5. *For all $x \in \mathbb{R}^n$, $S(x) = \mathbb{E}(\nabla f(x;\xi)^{\odot 2})$ is elementwise positive, and locally Lipschitz.*

This assumption, equivalent to $\mathbb{P}(\nabla f(x;\xi)) = 0$ is valid for all x, is often considered a weak hypothesis in practice.

2.3 Last-iterate Convergence Analysis

Before presenting the primary results, we will provide some contextual information concerning asymptotic pseudo-trajectories, the Robbins-Monro algorithm, and the Invariance Principle.

A semiflow Φ on the metric space (E,d) is a continuous function from $[0,\infty)\times E$ to E defined by $(t,x) \mapsto \Phi(t,x) = \Phi_t(x)$ such that Φ_0 is the identity mapping and $\Phi_{t+s} = \Phi_t \circ \Phi_s$ for all $(t,s) \in [0,\infty)^2$.

Definition 4. *A continuous function $z : \mathbb{R}_+ \to M$ is an asymptotic pseudo-trajectory (APT) for a semiflow Φ if*

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} d(z(t+h), \Phi_h(z(t))) = 0, \tag{2.11}$$

for any $T \in \mathbb{R}$

The Robbins-Monro algorithm is a stochastic approximation method [5]. By analyzing the related ODE, we are able to gain insights into the asymptotic tendencies of a stochastic process. Consider a discrete time stochastic process $\{z_t\}_t$ in $\mathbb{R}^m$ ($m \in \mathbb{N}$), which updates according to

$$z_t = z_{t-1} + \gamma_{t-1}(F(z_{t-1}) + U_t + b_t), \tag{2.12}$$

where $\{\gamma_t\}$ is a given sequence of deterministic nonnegative numbers satisfying $\sum_t \gamma_t = \infty$ and $\lim_{t\to\infty}\gamma_t = 0$, the martingale difference noise $\{U_t\}$ is measurable with respect to $\mathscr{F}_t$ for all $t \geq 1$ and satisfies $\mathbb{E}(U_t|\mathscr{F}_{t-1}) = 0$, and $\{b_t\}_t$ converges to zero almost surely.

The following conclusion is derived from [5, section 4.2].

Proposition 1. *Let* $\{z_t\}$ *given by* (2.12) *be a Robbin-Monro algorithm. Suppose that for some* $q \geq 2$

$$\sup_t \mathbb{E}\|U_t\|^q < \infty \text{ almost surely} \tag{2.13}$$

and

$$\sum_t \gamma_t^{1+q/2} < \infty. \tag{2.14}$$

If $\sup_t \|z_t\| < \infty$, *then the interpolated process* Z *defined by*

$$Z(\tau_t + s) = z_t + \frac{s}{\tau_{t+1} - \tau_t}(z_{t+1} - z_t)\,(0 \leq s < \gamma_{t+1}, t \in \mathbb{N}, \tau_0 = 0, \tau_t = \sum_{i=1}^{t} \gamma_i) \tag{2.15}$$

is an asymptotic pseudotrajectory of the flow induced by

$$\dot{z}(t) = F(z(t)) \tag{2.16}$$

with probability 1.

For clarity, we define z_t by the state derived from each iteration in (2.12), $z(t)$ by the continuous-time trajectory induced by (2.16), and $Z(t)$ by the linear interpolation of z_t.

This subsequent results establish the almost sure convergence of ASG methods towards the critical point. To the best of our knowledge, the following theorems provide the first last-iterate almost sure convergence analysis of x_t for Nadam, RMSprop, Adadelta, and AMSgrad in non-convex settings.

Theorem 1. *Suppose that Assumptions 1, 3, and 5 hold. Consider Nadam method described in Algorithm 3 with almost surely bounded* $\{x_t\}$ *and* $\{g_t\}$.

1. *The sequences of* $\{m_t\}$ *and* $\{v_t\}$ *are bounded almost surely.*
2. *Nadam converges almost surely:*

$$\lim_{t\to\infty} \begin{pmatrix} v_t \\ m_t \\ x_t \end{pmatrix} = \begin{pmatrix} S(x^c) \\ \boldsymbol{0} \\ x^c \end{pmatrix} \tag{2.17}$$

for some $x^c \in \mathscr{E}\mathscr{P} = \{x \in \mathbb{R}^n | \nabla f(x) = 0\}$.

Proof. 1. By our assumption, we know that there exist $X, G_1 > 0$ such that

$$\|x_t\| \leq X \text{ and } \|g_t\| \leq G_1 \text{ almost surely} \tag{2.18}$$

hold for $t \geq 1$. By the definition of $g_t^{\odot 2}$, there exists a constant $G_2 > 0$ such that

$$\|g_t^{\odot 2}\| \leq G_2 \text{ almost surely} \tag{2.19}$$

hold for $t \geq 1$. As a result, we can prove by induction that the following hold for all t:

$$\|m_t\| \leq G_1 \text{ almost surely} \tag{2.20}$$

and

$$\|v_t\| \leq G_2 \text{ almost surely.} \tag{2.21}$$

2. For every $t \geq 1$, the sequences $\{m_t\}, \{v_t\}$ and $\{x_t\}$ generated by Nadam method update according to

$$\begin{aligned} \begin{pmatrix} m_t \\ v_t \\ x_t \end{pmatrix} &= \begin{pmatrix} m_{t-1} \\ v_{t-1} \\ x_{t-1} \end{pmatrix} + \gamma_{t-1} \begin{pmatrix} \frac{1-\alpha_t}{\gamma_t}(g_t - m_{t-1}) \\ \frac{1-\beta_t}{\gamma_t}(g_t^{\odot 2} - v_{t-1}) \\ \left(\frac{\alpha_{t+1} m_t}{1-\Pi_{i=1}^{t+1}\alpha_i} + \frac{(1-\alpha_t)g_t}{1-\Pi_{i=1}^{t}\alpha_i}\right)\left(\frac{v_t}{1-\Pi_{i=1}^{t}\beta_i} + \varepsilon \mathbf{1}\right)^{-0.5} \end{pmatrix} \\ &= \begin{pmatrix} m_{t-1} \\ v_{t-1} \\ x_{t-1} \end{pmatrix} + \gamma_{t-1} F_t. \end{aligned} \tag{2.22}$$

To gain a deeper understanding of the Nadam method, it is convenient to analyze the following limit ODE first. As t goes to infinity, the terms $1 - \Pi_{i=1}^{t}\beta_i$ and $1 - \Pi_{i=1}^{t}\alpha_i$ converge to 1 according to [4, Lemma 9.1]. If we further ignore the stochasticity on gradient, the Nadam method can be considered to be a perturbed version of a time-varying step-size Cauchy-Euler approximation scheme for numerically solving the initial value problem

$$\begin{pmatrix} \dot{m} \\ \dot{v} \\ \dot{x} \end{pmatrix} = \begin{pmatrix} a(\nabla f(x) - m) \\ b(S(x) - v) \\ -\frac{m}{\sqrt{v+\varepsilon \mathbf{1}}} \end{pmatrix} = F(m, v, x), \quad \begin{pmatrix} m(0) \\ v(0) \\ x(0) \end{pmatrix} = \begin{pmatrix} m_0 \\ v_0 \\ x_0 \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R}^n_+ \times \mathbb{R}^n. \tag{2.23}$$

The solutions to m and v can be represented as

$$m(t) = exp(-at)\int_0^t exp(at)\nabla f(x(t))dt + m_0 exp(-at), \tag{2.24}$$

and

$$v(t) = exp(-bt)\int_0^t exp(bt)S(x(t))dt + v_0 exp(-bt). \tag{2.25}$$

As each element of $S(x)$ and v_0 is positive, every entry of $v(t)$ is positive for $t \in \mathbb{R}_+$, thus ensuring that $\frac{1}{\sqrt{v(t)+\varepsilon\mathbf{1}}}$ is well-defined for $t \in \mathbb{R}_+$.

We begin by proving the existence and uniqueness of the solution to (2.23) for $t \in [0,\infty)$. To do so, we consider the Lyapunov function $V : \mathbb{R}^n \times \mathbb{R}^n_+ \times \mathbb{R}^n \to \mathbb{R}$ derived from [4], defined as follows:

$$V(m,v,x) = f(x) + \frac{1}{2}\frac{\|m\|_2^2}{a\sqrt{v+\varepsilon\mathbf{1}}}. \tag{2.26}$$

Taking derivative on V yields

$$\begin{aligned}
\dot{V} &= -\nabla f(x)^T \tfrac{m}{\sqrt{v+\varepsilon\mathbf{1}}} + \tfrac{am^T(\nabla f(x)-m)}{a\sqrt{v+\varepsilon\mathbf{1}}} - \tfrac{1}{4}\tfrac{\|m\|_2^2}{a(v+\varepsilon\mathbf{1})^{1.5}} b(S(x)-v) \\
&= -\tfrac{\|m\|_2^2}{\sqrt{v+\varepsilon\mathbf{1}}} + \tfrac{b}{4a}\tfrac{\|m\|_2^2}{\sqrt{v+\varepsilon\mathbf{1}}} - \tfrac{b}{4a}\tfrac{\|m\|_2^2}{(v+\varepsilon\mathbf{1})^{1.5}}\varepsilon\mathbf{1} - \tfrac{b}{4a}\tfrac{\|m\|_2^2}{(v+\varepsilon\mathbf{1})^{1.5}}S(x) \\
&= -\tfrac{b}{4a}\tfrac{\|m\|_2^2}{(v+\varepsilon\mathbf{1})^{1.5}}S(x) - (1-\tfrac{b}{4a}(1-\varepsilon\mathbf{1}))\tfrac{\|m\|_2^2}{\sqrt{v+\varepsilon\mathbf{1}}} \\
&\overset{(a)}{\leq} -\tfrac{b}{4a}\tfrac{\|m\|_2^2}{(v+\varepsilon\mathbf{1})^{1.5}}S(x),
\end{aligned} \tag{2.27}$$

where (a) is due to Assumption 1.

Existence of the solution to (2.23). According to (2.27), the trajectory of $x(t)$ for $t \geq 0$ resides within the compact set

$$\Omega_x = \{x | f(x) \leq f(x_0) + \frac{1}{2}\frac{\|m_0\|_2^2}{a\sqrt{v_0+\varepsilon\mathbf{1}}}\}. \tag{2.28}$$

Within this set, both $S(x)$ and $\nabla f(x)$ are continuous and bounded, therefore, the solutions of m and v given by (2.24) and (2.25) are also bounded.

By the maximal interval of existence theorem [50], the initial value problem admits at least one solution on $t \in [0,\infty)$.

Uniqueness of the solution to (2.23). As both $S(x)$ and $\nabla f(x)$ are locally Lipschitz, F inherits local Lipschitz continuity. Let $(v(t^*)^T, m(t^*)^T, x(t^*)^T)^T$ be an arbitrary point on the solution. According to Picard's Existence and Uniqueness Theorem, there exists a unique solution on the

interval $[t^*-c, t^*+c]$ for some constant $c>0$. Since this solution exists for all $t \geq 0$ and the choice of $t^* \geq 0$ is arbitrary, the solution to the system (2.23) is unique.

Therefore, there exists a unique global solution to (2.23) starting from any point in $\mathbb{R}^n_+ \times \mathbb{R}^{2n}$, which admits a semiflow as mentioned at the beginning of this section.

As x^* is a critical point of f, the set $\mathscr{E}\mathscr{P} = \{x \in \mathbb{R}^n | \nabla f(x) = 0\}$ is nonempty. For all $x^c \in \mathscr{E}\mathscr{P}$, we have that

$$F(\mathbf{0}, S(x^c), x^c) = (\mathbf{0}^T, \mathbf{0}^T, \mathbf{0}^T)^T. \tag{2.29}$$

By Assumption 3, V is bounded below by f^*.

According to Barbashin-Krasovski-LaSalle Theorem, the trajectory will tend to the largest invariant set in $\{m=0\}$. Since m remains zero, we have $\dot{m} = a(\nabla f(x) - m) = a\nabla f(x) = 0$ and $\dot{x} = -\frac{m}{\sqrt{v+\varepsilon \mathbf{1}}} = 0$. It can be deduced that x stays at some vector $x^c \in \mathscr{E}\mathscr{P}$. That is to say, we have that

$$\lim_{t\to\infty} x = x^c, \text{ and } \lim_{t\to\infty} m = \mathbf{0}. \tag{2.30}$$

Considering the solution to v in (2.25) and the fact that x approaches x^c, we conclude that

$$\lim_{t\to\infty} v = S(x^c). \tag{2.31}$$

In the next step, we will check the fulfillment of conditions in Proposition 1.

Denote by

$$U_t = F_t - \mathbb{E}(F_t | \mathscr{F}_{t-1}) \tag{2.32}$$

and

$$b_t = \mathbb{E}(F_t | \mathscr{F}_{t-1}) - F(m_{t-1}, v_{t-1}, x_{t-1}). \tag{2.33}$$

Then, the condition $\mathbb{E}(U_t | \mathscr{F}_{t-1}) = 0$ follows directly.

Due to the almost sure boundedness of $\{x_t\}$ and the continuity of ∇f and S, there exists some $N>0$ such that

$$\|\nabla f(x_t)\| \leq N \text{ and } \|S(x_t)\| \leq N \text{ almost surely} \tag{2.34}$$

for all $t \geq 1$.

Consider the the first n components in (2.33):

$$\begin{aligned} & \|\mathbb{E}(\tfrac{1-\alpha_t}{\gamma_t}(g_t(x_{t-1})-m_{t-1})|\mathscr{F}_{t-1})-a(\nabla f(x_{t-1})-m_{t-1})\| \\ = \;& \|\tfrac{1-\alpha_t}{\gamma_t}(\nabla f(x_{t-1})-m_{t-1})-a(\nabla f(x_{t-1})-m_{t-1})\| \\ \leq \;& \left|\tfrac{1-\alpha_t}{\gamma_t}-a\right|\|m_{t-1}\|+\left|\tfrac{1-\alpha_t}{\gamma_t}-a\right|\|\nabla f(x_{t-1})\| \\ \leq \;& \left|\tfrac{1-\alpha_t}{\gamma_t}-a\right|G_1+\left|\tfrac{1-\alpha_t}{\gamma_t}-a\right|N \\ \to \;& 0 \quad \text{almost surely,} \end{aligned} \tag{2.35}$$

By a similar argument, we can derive that the second n component converge to zero almost surely as well, i.e.,

$$\lim_{t\to\infty}\mathbb{E}\left(\frac{1-\beta_t}{\gamma_t}(g_t^{\odot 2}(x_{t-1})-v_{t-1})\middle|\mathscr{F}_{t-1}\right)-b(S(x_{t-1})-v_{t-1})=0 \quad \text{almost surely.} \tag{2.36}$$

As for the last n components in (2.33), we will start with discussing the one-dimensional

case, where m_t, v_t, and x_t are scalar. By taking the norm of the difference, we obtain

$$
\begin{aligned}
& \left\| \mathbb{E}\left(\left(\frac{\alpha_{t+1} m_t}{1-\Pi_{i=1}^{t+1}\alpha_i} + \frac{(1-\alpha_t) g_t}{1-\Pi_{i=1}^{t}\alpha_i} \right) \left(\frac{v_t}{1-\Pi_{i=1}^{t}\beta_i} + \varepsilon \right)^{-0.5} \middle| \mathscr{F}_{t-1} \right) - \frac{m_{t-1}}{\sqrt{v_{t-1}+\varepsilon}} \right\| \\
\leq & \left\| \frac{\alpha_{t+1}}{1-\Pi_{i=1}^{t+1}\alpha_i} \mathbb{E}\left(m_t \left(\frac{v_t}{1-\Pi_{i=1}^{t}\beta_i} + \varepsilon \right)^{-0.5} \middle| \mathscr{F}_{t-1} \right) - \frac{m_{t-1}}{\sqrt{v_{t-1}+\varepsilon}} \right\| \\
& + (1-\alpha_t) \frac{1}{(1-\Pi_{i=1}^{t}\alpha_i)} \left\| \mathbb{E}\left(g_t \left(\frac{v_t}{1-\Pi_{i=1}^{t}\beta_i} + \varepsilon \right)^{-0.5} \middle| \mathscr{F}_{t-1} \right) \right\| \\
\leq & \left\| \frac{\alpha_{t+1}}{1-\Pi_{i=1}^{t+1}\alpha_i} \mathbb{E}\left(m_t \left(\frac{v_t}{1-\Pi_{i=1}^{t}\beta_i} + \varepsilon \right)^{-0.5} \middle| \mathscr{F}_{t-1} \right) - \mathbb{E}\left(m_t \left(\frac{v_t}{1-\Pi_{i=1}^{t}\beta_i} + \varepsilon \right)^{-0.5} \middle| \mathscr{F}_{t-1} \right) \right\| \\
& + \left\| \mathbb{E}\left(m_t \left(\frac{v_t}{1-\Pi_{i=1}^{t}\beta_i} + \varepsilon \right)^{-0.5} \middle| \mathscr{F}_{t-1} \right) - \frac{m_{t-1}}{\sqrt{v_{t-1}+\varepsilon}} \right\| + \frac{1-\alpha_t}{(1-\Pi_{i=1}^{t}\alpha_i)} \frac{G_1}{\sqrt{\varepsilon}} \\
\leq & \left| \frac{\alpha_{t+1}}{1-\Pi_{i=1}^{t+1}\alpha_i} - 1 \right| \frac{G_1}{\sqrt{\varepsilon}} + \left\| \mathbb{E}\left(m_t \left(\frac{v_t}{1-\Pi_{i=1}^{t}\beta_i} + \varepsilon \right)^{-0.5} \middle| \mathscr{F}_{t-1} \right) - \frac{m_{t-1}}{\sqrt{v_{t-1}+\varepsilon}} \right\| + \frac{1-\alpha_t}{(1-\Pi_{i=1}^{t}\alpha_i)} \frac{G_1}{\sqrt{\varepsilon}} \\
= & \left| \frac{\alpha_{t+1}}{1-\Pi_{i=1}^{t+1}\alpha_i} - 1 \right| \frac{G_1}{\sqrt{\varepsilon}} + \frac{1-\alpha_t}{(1-\Pi_{i=1}^{t}\alpha_i)} \frac{G_1}{\sqrt{\varepsilon}} \\
& + \left\| \mathbb{E}\left((\alpha_t m_{t-1} + (1-\alpha_t) g_t) \right) \left(\frac{\beta_t v_{t-1} + (1-\beta_t) g_t^{\odot 2}}{1-\Pi_{i=1}^{t}\beta_i} + \varepsilon \right)^{-0.5} \middle| \mathscr{F}_{t-1} \right) - \frac{m_{t-1}}{\sqrt{v_{t-1}+\varepsilon}} \right\| \\
\leq & \left| \frac{\alpha_{t+1}}{1-\Pi_{i=1}^{t+1}\alpha_i} - 1 \right| \frac{G_1}{\sqrt{\varepsilon}} + \frac{1-\alpha_t}{(1-\Pi_{i=1}^{t}\alpha_i)} \frac{G_1}{\sqrt{\varepsilon}} \\
& + \left\| \mathbb{E}\left(m_{t-1} \left(\frac{\beta_t v_{t-1} + (1-\beta_t) g_t^{\odot 2}}{1-\Pi_{i=1}^{t}\beta_i} + \varepsilon \right)^{-0.5} \middle| \mathscr{F}_{t-1} \right) - \frac{m_{t-1}}{\sqrt{v_{t-1}+\varepsilon}} \right\| + (1-\alpha_t) \frac{G_1+G_1}{\sqrt{\varepsilon}} \\
\overset{(b)}{\leq} & \left| \frac{\alpha_{t+1}}{1-\Pi_{i=1}^{t+1}\alpha_i} - 1 \right| \frac{G_1}{\sqrt{\varepsilon}} + \frac{1-\alpha_t}{(1-\Pi_{i=1}^{t}\alpha_i)} \frac{G_1}{\sqrt{\varepsilon}} + (1-\alpha_t) \frac{G_1+G_1}{\sqrt{\varepsilon}} \\
& + \left\| \mathbb{E}\left(m_{t-1} \left(v_{t-1} + \varepsilon \right)^{-0.5} \middle| \mathscr{F}_{t-1} \right) - \frac{m_{t-1}}{\sqrt{v_{t-1}+\varepsilon}} \right\| \\
& + \left\| \mathbb{E}\left(\frac{1}{2} m_{t-1} \frac{\frac{\beta_t v_{t-1} + (1-\beta_t) g_t^{\odot 2}}{1-\Pi_{i=1}^{t}\beta_i} - v_{t-1}}{\left(c_t v_{t-1} + (1-c_t) \frac{\beta_t v_{t-1} + (1-\beta_t) g_t^{\odot 2}}{1-\Pi_{i=1}^{t}\beta_i} + \varepsilon \right)^{1.5}} \middle| \mathscr{F}_{t-1} \right) \right\| \\
\leq & \left| \frac{\alpha_{t+1}}{1-\Pi_{i=1}^{t+1}\alpha_i} - 1 \right| \frac{G_1}{\sqrt{\varepsilon}} + \frac{1-\alpha_t}{(1-\Pi_{i=1}^{t}\alpha_i)} \frac{G_1}{\sqrt{\varepsilon}} + (1-\alpha_t) \frac{G_1+G_1}{\sqrt{\varepsilon}} \\
& + \frac{1}{2} \frac{G_1}{\varepsilon^{1.5}} \left(\left| \frac{\beta_t}{1-\Pi_{i=1}^{t}\beta_i} - 1 \right| G_2 + \frac{1-\beta_t}{1-\Pi_{i=1}^{t}\beta_i} G_2 \right) \\
\overset{(c)}{\to} & \ 0 \quad \text{almost surely,}
\end{aligned}
\tag{2.37}
$$

where the inequality (b) follows the mean value theorem and $c_t \in (0,1)$, and (c) is due to $\lim_{t\to\infty} 1-\Pi_{i=1}^{t}\beta_i = \lim_{t\to\infty} 1-\Pi_{i=1}^{t}\alpha_i = 1$ from [4, Lemma 9.1] and Assumption 1.

When $n > 1$, the discussion for each component is the same as the scalar case, and we omit

the detailed proof here. Therefore, we can conclude that

$$b_t \to 0 \text{ as } t \to \infty \quad \text{almost surely.} \tag{2.38}$$

On the other hand, all terms in U_t are bounded almost surely, thus, condition (2.13) should hold for all $q \geq 2$. According to Proposition 1, the linear interpolation of the stochastic process $\{(v_t^T, m_t^T, x_t^T)^T\}_t$, is almost surely an APT of the semiflow Φ induced by the system (2.23).

Combining the convergence behaviors in (2.30), (2.31) and the definition of APT, we know that

$$\lim_{t\to\infty} \begin{pmatrix} m_t \\ v_t \\ x_t \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ S(x^c) \\ x^c \end{pmatrix} \tag{2.39}$$

holds for some $x^c \in \mathscr{E}\mathscr{P}$. □

Corollary 1. *Adam method described in Algorithm 1 has the same conclusions as Theorem 1*

Proof. The proof is similar to that for 1. □

Corollary 2. *AMSgrad method described in Algorithm 5 has the same conclusions as Theorem 1*

Proof. 1. The discussion on the boundednesses of $\{m_t\}$ and $\{v_t\}$ is similar to that for Theorem 1.

2. We again utilize X, G_1, G_2 and N as almost sure upper bounds mentioned in (2.18), (2.19), (2.20), (2.21) and (2.34). Since $\{\hat{v}_t\}$ is non-decreasing and almost surely bounded, it converges to $V_u = \sup\{\hat{v}_t\}$ almost surely.

We derive a corresponding limiting ODE for the AMSgrad method:

$$\begin{pmatrix} \dot{m} \\ \dot{v} \\ \dot{x} \end{pmatrix} = \begin{pmatrix} a(\nabla f(x) - m) \\ b(S(x) - v) \\ -\frac{m}{\sqrt{V_u + \varepsilon \mathbf{1}}} \end{pmatrix} = F(m, v, x), \quad \begin{pmatrix} m(0) \\ v(0) \\ x(0) \end{pmatrix} = \begin{pmatrix} m_0 \\ v_0 \\ x_0 \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R}^n_+ \times \mathbb{R}^n. \tag{2.40}$$

We consider the Lyapunov function

$$V = f(x) + \frac{1}{2} \frac{\|m\|^2}{a\sqrt{V_u + \varepsilon \mathbf{1}}}. \tag{2.41}$$

Taking the derivative of V, we obtain

$$\dot{V} = -\frac{\|m\|^2}{\sqrt{V_u + \varepsilon \mathbf{1}}} \tag{2.42}$$

Similar to the arguments in Section 1, we can use Barbashin-Krasovski-LaSalle Theorem to derive

$$\lim_{t\to\infty} m = 0, \quad \lim_{t\to\infty} v = S(x^c), \quad \text{and} \quad \lim_{t\to\infty} x = x^c, \tag{2.43}$$

for some $x^c \in \mathscr{E}\mathscr{P} = \{x \in \mathbb{R}^n | \nabla f(x) = 0\}$.

Define F_t for the AMSgrad method in a similar way to (2.22). Similar to the proof of Section 1, we also need to check

$$\mathbb{E}(F_t|\mathscr{F}_{t-1}) - F(m_{t-1}, v_{t-1}, x_{t-1}) \to 0 \text{ as } t \to \infty \quad \text{almost surely.} \tag{2.44}$$

We begin with considering the scalar case $(n = 1)$. The last n components of (2.44) decays to zero:

$$\begin{aligned}
& \left\| \mathbb{E}\left(\frac{m_t}{(1-\Pi_{i=1}^t \alpha_i)\sqrt{\hat{v}_t+\varepsilon}} \middle| \mathscr{F}_{t-1} \right) - \frac{m_{t-1}}{\sqrt{V_u+\varepsilon}} \right\| \\
= & \left\| \mathbb{E}\left(\frac{\beta_t m_{t-1} + (1-\beta_t) g_t}{(1-\Pi_{i=1}^t \alpha_i)\sqrt{\hat{v}_t+\varepsilon}} \middle| \mathscr{F}_{t-1} \right) - \frac{m_{t-1}}{\sqrt{V_u+\varepsilon}} \right\| \\
= & \left\| \mathbb{E}\left(\left(\frac{m_{t-1}}{(1-\Pi_{i=1}^t \alpha_i)\sqrt{\hat{v}_t+\varepsilon}} - \frac{m_{t-1}}{\sqrt{\hat{v}_t+\varepsilon}} \right) + \frac{m_{t-1}}{\sqrt{\hat{v}_t+\varepsilon}} + \frac{(1-\beta_t)(g_t - m_{t-1})}{(1-\Pi_{i=1}^t \alpha_i)\sqrt{\hat{v}_t+\varepsilon}} \middle| \mathscr{F}_{t-1} \right) - \frac{m_{t-1}}{\sqrt{V_u+\varepsilon}} \right\| \\
\leq & \left\| \mathbb{E}\left(\frac{m_{t-1}}{\sqrt{\hat{v}_t+\varepsilon}} - \frac{m_{t-1}}{\sqrt{V_u+\varepsilon}} \middle| \mathscr{F}_{t-1} \right) \right\| + \left| \frac{1}{1-\Pi_{i=1}^t \alpha_i} - 1 \right| \frac{M}{\sqrt{\varepsilon}} + \left| \frac{1-\beta_t}{1-\Pi_{i=1}^t \alpha_i} \right| \frac{G+M}{\sqrt{\varepsilon}} \\
\overset{(a)}{\leq} & \left\| \mathbb{E}\left((V_u - \hat{v}_t) \frac{m_{t-1}}{2(\bar{v}_t+\varepsilon)^{3/2}} \middle| \mathscr{F}_{t-1} \right) \right\| + \left| \frac{1}{1-\Pi_{i=1}^t \alpha_i} - 1 \right| \frac{M}{\sqrt{\varepsilon}} + \left| \frac{1-\beta_t}{1-\Pi_{i=1}^t \alpha_i} \right| \frac{G+M}{\sqrt{\varepsilon}} \\
\leq & (V_u - \hat{v}_t) \frac{M}{2\varepsilon^{3/2}} + \left| \frac{1}{1-\Pi_{i=1}^t \alpha_i} - 1 \right| \frac{M}{\sqrt{\varepsilon}} + \left| \frac{1-\beta_t}{1-\Pi_{i=1}^t \alpha_i} \right| \frac{G+M}{\sqrt{\varepsilon}},
\end{aligned} \tag{2.45}$$

where (a) holds for some $\bar{v}_t \in [\hat{v}_t, V_u]$ due to the mean value theorem. When $n > 1$, the discussion for each component is the same as the scalar case, and we omit the detailed proof here. Combining this result and (2.35), (2.36), we can conclude (2.44)

The rest of the proof proceeds in the same way as that of 1. □

Since the RMSprop and Adadelta methods do not utilize the first moment estimate m_t, the parameter α_t is not involved. Hence, we will adopt Assumption 2 instead of Assumption 1. In line with the preceding statements, the following theorems demonstrate the almost sure convergence to a critical point of the non-convex objective function.

Theorem 2. *Suppose that Assumptions 2, 3, and 5 hold. Consider RMSprop method described in Algorithm 2 with almost surely bounded $\{x_t\}$ and $\{g_t\}$.*

1. *The sequence of $\{v_t\}$ is bounded almost surely.*
2. *RMSprop converges almost surely:*

$$\lim_{t\to\infty}\begin{pmatrix} v_t \\ x_t \end{pmatrix} = \begin{pmatrix} S(x^c) \\ x^c \end{pmatrix} \tag{2.46}$$

for some $x^c \in \mathscr{E}\mathscr{P} = \{x \in \mathbb{R}^n | \nabla f(x) = 0\}$.

Proof. 1. The discussion on the boundedness of $\{v_t\}$ is similar to that for 1.

2. Similar to the proof for 1, we derive a corresponding limiting ODE for the RMSprop method:

$$\begin{pmatrix} \dot{v} \\ \dot{x} \end{pmatrix} = \begin{pmatrix} b(S(x)-v) \\ -\frac{\nabla f(x)}{\sqrt{v+\varepsilon \mathbf{1}}} \end{pmatrix}, \quad \begin{pmatrix} v(0) \\ x(0) \end{pmatrix} = \begin{pmatrix} v_0 \\ x_0 \end{pmatrix} \in \mathbb{R}_+ \times \mathbb{R}. \tag{2.47}$$

Choose the Lyapunov function to be $V = f(x)$. Taking its derivative, we obtain

$$\frac{dF(x)}{dt} = -\frac{\|\nabla f(x)\|^2}{\sqrt{v+\varepsilon \mathbf{1}}} \tag{2.48}$$

According to the Invariance principle stated in Barbashin-Krasovski-LaSalle Theorem, the virtual state will converges to the maximal positive invariant set in $\mathscr{E}\mathscr{P} = \{x \in \mathbb{R}^n | \nabla f(x) = 0\}$. Within this set, x is a constant vector $x^c \in \mathscr{E}\mathscr{P}$ as $\dot{x} = -\frac{\nabla f(x)}{\sqrt{v+\varepsilon \mathbf{1}}} = 0$, resulting in the asymptotic behavior

$$\lim_{t\to\infty} x = x^c \tag{2.49}$$

From the solution (2.25), we know that

$$\lim_{t\to\infty} v = S(x^c). \tag{2.50}$$

The rest of the proof proceeds in the same way as Theorem 1. □

Theorem 3. *Suppose that Assumptions 2, 3, and 5 hold. Consider Adadelta method described in Algorithm 4 with almost surely bounded $\{u_t\}$, $\{x_t\}$ and $\{g_t\}$.*

1. *The sequence of $\{v_t\}$ is bounded almost surely.*

2. *Adadelta converges almost surely:*

$$\lim_{t\to\infty}\begin{pmatrix}u_t\\ v_t\\ x_t\end{pmatrix}=\begin{pmatrix}\frac{S(x^c)}{(S(x^c)+\varepsilon \mathbf{1})^2}\\ S(x^c)\\ x^c\end{pmatrix} \tag{2.51}$$

for some $x^c\in\mathscr{E}\mathscr{P}=\{x\in\mathbb{R}^n|\nabla f(x)=0\}$.

Proof. 1. The discussion on the boundedness of $\{v_t\}$ is similar to that for Theorem 1.

2. The proof is similar to that for Theorem 1. We first derive a corresponding limiting ODE for the Adadelta method:

$$\begin{pmatrix}\dot{v}\\ \dot{u}\\ \dot{x}\end{pmatrix}=\begin{pmatrix}b(S(x)-v)\\ b(\frac{u+\varepsilon\mathbf{1}}{v+\varepsilon\mathbf{1}}S(x)-u)\\ -\frac{\sqrt{u+\varepsilon\mathbf{1}}}{\sqrt{v+\varepsilon\mathbf{1}}}\nabla f(x)\end{pmatrix},\quad \begin{pmatrix}v(0)\\ u(0)\\ x(0)\end{pmatrix}=\begin{pmatrix}v_0\\ u_0\\ x_0\end{pmatrix}\in\mathbb{R}^{2n}_+\times\mathbb{R}. \tag{2.52}$$

Positivity of v and u. In the Adadelta method, the positivity of v_t and u_t ($t\geq 1$) can be proven by induction, ensuring that $\frac{\sqrt{u_t+\varepsilon\mathbf{1}}}{\sqrt{v_t+\varepsilon\mathbf{1}}}$ is always well-defined. We have already demonstrated the positivity of v in the discussion on (2.25). Define $p_1(t)=-\frac{bS(x(t))}{v(t)+\varepsilon\mathbf{1}}+b$ and $p_2(t)=\frac{\varepsilon bS(x(t))}{v(t)+\varepsilon\mathbf{1}}$. Observing that the second equation in (2.52) can be rewritten as

$$\dot{u}+p_1(t)u=p_2(t), \tag{2.53}$$

we know that the solution to u is given by

$$u(t)=u_0exp(-\int_0^t p_1(t)dt)+exp(-\int_0^t p_1(t)dt)\int_0^t exp(\int_0^t p_1(t)dt)p_2(t)dt,\quad (t>0). \tag{2.54}$$

As $p_2(t)$ is positive elementment-wise, $u(t)$ is also positive elementment-wise.

We consider the Lyapunov function $V_1=f(x)$, which is lower bounded by f^*. Taking the derivative of V_1, we obtain

$$\dot{V}_1=-\frac{\sqrt{u+\varepsilon\mathbf{1}}}{\sqrt{v+\varepsilon\mathbf{1}}}\|\nabla f(x)\|^2. \tag{2.55}$$

Similar to the argument in Section 2.3, we know that

$$\lim_{t\to\infty}x=x^c \text{ and } \lim_{t\to\infty}v=S(x^c) \tag{2.56}$$

for some $x^c \in \mathscr{E}\mathscr{P} = \{x \in \mathbb{R}^n | \nabla f(x) = 0\}$. The continuity of S implies that p_1 approaches $p_{1\infty} = \frac{b\varepsilon \mathbf{1}}{S(x^c)+\varepsilon \mathbf{1}}$ and p_2 approaches $p_{2\infty} = \frac{\varepsilon b S(x^c)}{S(x^c)+\varepsilon \mathbf{1}}$ as $t \to \infty$, therefore, eventually $p_1(t) \geq \frac{1}{2} p_{1\infty}$.

Let $V_2 = \frac{1}{2}\|u - \frac{p_{2\infty}}{p_{1\infty}}\|^2$. Taking its time derivative, we have

$$\begin{aligned}
\dot{V}_2 &= \textstyle\sum_i (u - \frac{p_{2\infty}}{p_{1\infty}})_i (-p_1(t)u + p_2(t))_i \\
&= -\textstyle\sum_i (p_1(t))_i (u - \frac{p_{2\infty}}{p_{1\infty}})_i (u - \frac{p_2}{p_1})_i \\
&\leq \textstyle\sum_i -(p_1(t))_i (u - \frac{p_{2\infty}}{p_{1\infty}})_i^2 + \left|(p_1(t))_i (u - \frac{p_{2\infty}}{p_{1\infty}})_i (\frac{p_{2\infty}}{p_{1\infty}} - \frac{p_2}{p_1})_i\right| \\
&\leq \textstyle\sum_i -(p_1(t))_i (u - \frac{p_{2\infty}}{p_{1\infty}})_i^2 + \frac{1}{2}(p_1(t))_i ((u - \frac{p_{2\infty}}{p_{1\infty}})_i^2 + (\frac{p_{2\infty}}{p_{1\infty}} - \frac{p_2}{p_1})_i^2) \\
&\leq \textstyle\frac{1}{2}\sum_i -(p_1(t))_i (u - \frac{p_{2\infty}}{p_{1\infty}})_i^2 + (p_1(t))_i (\frac{p_{2\infty}}{p_{1\infty}} - \frac{p_2}{p_1})_i^2 \\
&\leq -\min_i((p_{1\infty})_i) V_2 + \frac{b}{2}\max_i((p_{1\infty})_i) \left\| \frac{p_{2\infty}}{p_{1\infty}} - \frac{p_2}{p_1} \right\|^2
\end{aligned} \tag{2.57}$$

for sufficiently large t. As $\left\| \frac{p_{2\infty}}{p_{1\infty}} - \frac{p_2}{p_1} \right\|^2$ decays to zero, V also converges to zero, that is to say,

$$\lim_{t\to\infty} u = \frac{p_{2\infty}}{p_{1\infty}} = S(x^c). \tag{2.58}$$

Define F_t for the Adadelta method in a similar way to (2.22). Similar to the proof of Theorem 1, we need to check

$$\mathbb{E}(F_t | \mathscr{F}_{t-1}) - F(m_{t-1}, v_{t-1}, x_{t-1}) \to 0 \text{ as } t \to \infty \quad \text{almost surely.} \tag{2.59}$$

The arguments for the first n components are similar to (2.35). We again utilize X, G_1, G_2 and N as almost sure upper bounds mentioned in (2.18), (2.19), (2.21) and (2.34). Let U to be an almost sure upper bound for $\{u_t\}$. We begin with considering the scalar case $(n = 1)$. The last $2n$ components of (2.59) decays to zero:

$$\begin{aligned}
& \left\| b(\tfrac{u_{t-1}+\varepsilon}{v_{t-1}+\varepsilon} S(x_{t-1}) - u_{t-1}) - \tfrac{1-\beta_t}{\gamma_t}(\mathbb{E}(\tfrac{u_{t-1}+\varepsilon}{v_t+\varepsilon} g_t^{\odot 2} | \mathscr{F}_{t-1}) - u_{t-1})) \right\| \\
\leq & \left\| \tfrac{b}{v_{t-1}+\varepsilon} S(x_{t-1}) - \tfrac{1-\beta_t}{\gamma_t} \left(\mathbb{E}\left(\tfrac{1}{v_{t-1}+\varepsilon} g_t^{\odot 2} | \mathscr{F}_{t-1} \right) - \gamma_t \mathbb{E}\left(\tfrac{g_t^{\odot 2} - v_{t-1}}{(1-c_2\gamma_t)v_{t-1} + c_2\gamma_t g_t^{\odot 2} + \varepsilon)^2} | \mathscr{F}_t \right) \right) \right\| \\
& \|u_{t-1} + \varepsilon\| \\
\leq & \|u_{t-1} + \varepsilon\| \left| b - \tfrac{1-\beta_t}{\gamma_t} \right| \left\| \tfrac{S(x_{t-1})}{v_{t-1}+\varepsilon} \right\| + (1-\beta_t)\|u_{t-1} + \varepsilon\| \left\| \mathbb{E}\left(\tfrac{g_t^{\odot 2} - v_{t-1}}{(1-c_2\gamma_t)v_{t-1} + c_2\gamma_t g_t^{\odot 2} + \varepsilon)^2} | \mathscr{F}_t \right) \right\| \\
\leq & \left| b - \tfrac{1-\beta_t}{\gamma_t} \right| (U+\varepsilon)\tfrac{N}{\varepsilon} + |1-\beta_t|(U+\varepsilon)\tfrac{2G_2}{\varepsilon^2} \\
\to & \ 0 \quad \text{almost surely,}
\end{aligned} \tag{2.60}$$

and

$$
\begin{aligned}
& \left\| \mathbb{E}\left(\frac{\sqrt{u_{t-1}+\varepsilon}}{\sqrt{v_t+\varepsilon}} g_t | \mathscr{F}_{t-1} \right) - \frac{\sqrt{u_{t-1}+\varepsilon}}{\sqrt{v_{t-1}+\varepsilon}} \nabla f(x_{t-1}) \right\| \\
\leq & \left\| \mathbb{E}\left(\frac{\sqrt{u_{t-1}+\varepsilon}}{\sqrt{v_{t-1}+\varepsilon}} g_t - \gamma_t \frac{(g_t^{\odot 2} - v_{t-1})\sqrt{u_{t-1}+\varepsilon}}{2(1-c_1\gamma_t)v_{t-1} + c_1\gamma_t g_t^{\odot 2} + \varepsilon)^{3/2}} g_t | \mathscr{F}_{t-1} \right) - \frac{\sqrt{u_{t-1}+\varepsilon}}{\sqrt{v_{t-1}+\varepsilon}} \nabla f(x_{t-1}) \right\| \\
\leq & \left\| \mathbb{E}\left(\frac{\sqrt{u_{t-1}+\varepsilon}}{\sqrt{v_{t-1}+\varepsilon}} g_t | \mathscr{F}_{t-1} \right) - \frac{\sqrt{u_{t-1}+\varepsilon}}{\sqrt{v_{t-1}+\varepsilon}} \nabla f(x_{t-1}) \right\| \\
& + \gamma_t \left\| \mathbb{E}\left(\frac{(g_t^{\odot 2} - v_{t-1})\sqrt{u_{t-1}+\varepsilon}}{2(1-c_1\gamma_t)v_{t-1} + c_1\gamma_t g_t^{\odot 2} + \varepsilon)^{3/2}} g_t | \mathscr{F}_{t-1} \right) \right\| \\
= & \gamma_t \left\| \mathbb{E}\left(\frac{(g_t^{\odot 2} - v_{t-1})\sqrt{u_{t-1}+\varepsilon}}{2(1-c_1\gamma_t)v_{t-1} + c_1\gamma_t g_t^{\odot 2} + \varepsilon)^{3/2}} g_t | \mathscr{F}_{t-1} \right) \right\| \\
= & \gamma_t \frac{2G_2\sqrt{U+\varepsilon}}{\varepsilon^{1.5}} \\
\to & \ 0 \quad \text{almost surely,}
\end{aligned}
\tag{2.61}
$$

When $n > 1$, the discussion for each component is the same as the scalar case, and we omit the detailed proof here.

The rest of the proof proceeds in the same way as Theorem 1. □

Remark 3. *For the above-mentioned ASG methods, there are multiple choices of γ_t, such as $\gamma_t = O\left(\frac{1}{t}\right)$, $\gamma_t = O\left(\frac{1}{t^{3/4}}\right)$, and $\gamma_t = O\left(\frac{1}{t^{3/5}}\right)$, ensure the almost sure convergence of x_t to the critical point.*

2.4 Almost Sure Convergence Rate Analysis

It is usually more challenging to find the convergence rates of the ASG methods. A nice and short discussion for the almost sure convergence rates of stochastic gradient descent methods was made in [45], which relies on the results from Appendix B, i.e. the classical supermartingale convergence theorem from [67] and its corollaries derived by [45]. Motivated by [45], these two results will also be utilized in analyzing the almost sure convergence rates for ASG methods. These rates match the lower bounds for stochastic gradient-based algorithm, $O\left(\frac{1}{t}\right)$ for strongly convex loss function, and $O\left(\frac{1}{t^{0.5}}\right)$ for nonconvex loss function [2], to an ε-factor.

Proposition 2 (Supermartingale Convergence Theorem)**.** *Let $\{X_t\}$, $\{Y_t\}$, and $\{Z_t\}$ be three nonnegative sequences of random variables that are adapted to a filtration $\{\mathscr{F}_t\}$. Let $\{\gamma_t\}$ be a sequence of nonnegative real numbers such that $\Pi_{t=1}^{\infty}(1+\gamma_t) < \infty$. Suppose that the following conditions hold:*

1. $\mathbb{E}(Y_{t+1}|\mathscr{F}_t) \leq (1+\gamma_t)Y_t - X_t + Z_t$ *for all* $t > 1$.

2. $\sum_{t=1}^{\infty} Z_t < \infty$ *holds almost surely.*

Then, $\sum_{t=1}^{\infty} X_t < \infty$ *almost surely and* Y_t *converges almost surely.*

Corollary 3. *Suppose that* $\{Y_t\}$ *is a sequence of nonnegative random variables that are adapted to a filtration* $\{\mathscr{F}_t\}$.

1. *If* $\{Y_t\}$ *satisfies*

$$\mathbb{E}(Y_{t+1}|\mathscr{F}_t) \leq (1 - c_1\gamma_t)Y_t + c_2\gamma_t^2, \tag{2.62}$$

for all $t \geq 1$, *where* $\gamma_t = O\left(\frac{1}{t^{1-\theta_1}}\right)$ *for some* $\theta_1 \in (0, \frac{1}{2})$, *and* c_1, c_2 *are positive constants. Then, for any* $\theta_2 \in (2\theta_1, 1)$,

$$Y_t = o\left(\frac{1}{t^{1-\theta_2}}\right) \text{ almost surely.} \tag{2.63}$$

2. *Let* $\{\gamma_t\}$ *be a sequence of positive real numbers such that the following holds:*

$$\sum_{t=1}^{\infty} \gamma_t Y_t < \infty \text{ almost surely,} \tag{2.64}$$

$$\sum_i \gamma_i^2 < \infty \text{ and } \sum_{t=1}^{\infty} \frac{\gamma_t}{\sum_{i=1}^{t-1} \alpha_i} = \infty \tag{2.65}$$

Then, we know that

$$\min_{1 \leq i \leq t} Y_i = o\left(\frac{1}{\sum_{i=1}^{t-1} \gamma_i}\right) \text{ almost surely.} \tag{2.66}$$

To the best of our knowledge, the following provides the first almost sure convergence rates for adaptive methods on strong convexity, non-convexity and general convexity assumptions.

Theorem 4. *Suppose that Assumptions 2, 3, and 5 hold with* $p = 2$. *Consider the RMSprop method described in Algorithm 2 with almost surely bounded* $\{x_t\}$ *and* $\{g_t\}$. *Then, the following hold:*

1. *If Assumption 4 also hold and* $\gamma_t = O\left(\frac{1}{t^{1-\theta_1}}\right)$ *for* $\theta_1 \in (0, \frac{1}{2})$, *then it follows that*

$$f(x_t) - f^* = o\left(\frac{1}{t^{1-\theta_2}}\right) \text{ almost surely} \tag{2.67}$$

for any $\theta_2 \in (2\theta_1, 1)$.

2. *If* γ_t *satisfies* $\sum_{t=1}^{\infty} \frac{\gamma_t}{\sum_{i=1}^{t-1} \alpha_i} = \infty$, *then*

$$\min_{1\le i\le t} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{\sum_{i=1}^{t} \gamma_i}\right) \text{ almost surely.} \tag{2.68}$$

For example, if it is chosen that $\gamma_t = O\left(\frac{1}{t^{\frac{1}{2}+\theta_3}}\right)$ *for* $\theta_3 \in (0, \frac{1}{2})$, *then*

$$\min_{1\le i\le t} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{t^{\frac{1}{2}-\theta_3}}\right) \text{ almost surely.} \tag{2.69}$$

Proof. Given the almost sure boundedness of the trajectory $\{x_t\}$, there exists a compact set $C \subset \mathbb{R}^n$ containing the entire trajectory $\{x_t\}$. The locally Lipschitz condition of the gradient∇f implies that ∇f is L-Lipschitz with some constant $L > 0$ on C. We again utilize X, G_1, G_2 and N as almost sure upper bounds mentioned in (2.18), (2.19), (2.21) and (2.34). Using the inequality (2.6), we have

$$\begin{aligned}
f(x_t) &\le f(x_{t-1}) - \gamma_t \langle \nabla f(x_{t-1}), \tfrac{g_t}{\sqrt{v_t+\varepsilon \mathbf{1}}} \rangle + \tfrac{L}{2} (\tfrac{\gamma_t}{\sqrt{v_t+\varepsilon \mathbf{1}}})^2 \|g_t\|^2 \\
&= f(x_{t-1}) - \gamma_t \langle \nabla f(x_{t-1}), \tfrac{g_t}{\sqrt{\beta_t v_{t-1} + (1-\beta_t) g_t^{\odot 2} + \varepsilon \mathbf{1}}} \rangle + O(\gamma_t^2) \\
&= f(x_{t-1}) - \gamma_t \textstyle\sum_i \nabla f(x_{t-1})_i \frac{(g_t)_i}{\sqrt{\beta_t (v_{t-1})_i + (1-\beta_t)(g_t)_i^2 + \varepsilon}} + O(\gamma_t^2) \\
&\overset{(a)}{=} f(x_{t-1}) - \gamma_t \textstyle\sum_i \nabla f(x_{t-1})_i \frac{1}{\sqrt{(v_{t-1})_i + \varepsilon}} (g_t)_i \\
&\quad - \gamma_t (1-\beta_t) \tfrac{1}{2} \textstyle\sum_i \nabla f(x_{t-1})_i \frac{(v_{t-1})_i - (g_t)_i^2}{((1-c_{ti}\gamma_t)(v_{t-1})_i + c_{ti}\gamma_t (g_t)_i^2 + \varepsilon)^{3/2}} (g_t)_i + O(\gamma_t^2) \\
&= f(x_{t-1}) - \gamma_t \textstyle\sum_i \nabla f(x_{t-1})_i \frac{(g_t)_i}{\sqrt{(v_{t-1})_i + \varepsilon}} + O(\gamma_t^2)
\end{aligned} \tag{2.70}$$

where (a) holds for some $c_{ti} \in (0,1)$ according to the mean value theorem.

Taking the conditional expectation on (2.70), the following inequality holds with probability one:

$$\begin{aligned}
\mathbb{E}(f(x_t) - f^* | \mathscr{F}_{t-1}) &\le f(x_{t-1}) - f^* - \gamma_t \textstyle\sum_i \nabla f(x_{t-1})_i \frac{\nabla f(x_{t-1})_i}{\sqrt{(v_{t-1})_i + \varepsilon}} + O(\gamma_t^2) \\
&\le f(x_{t-1}) - f^* - \tfrac{\gamma_t}{\sqrt{G_1 + \varepsilon \mathbf{1}}} \|\nabla f(x_{t-1})\|^2 + O(\gamma_t^2).
\end{aligned} \tag{2.71}$$

1. When f is μ-strongly convex, (2) implies that x_t eventually enter a neighborhood of $x^c = x^*$

after finite-time iterations. We can use (2.71) and (2.10) to derive

$$\mathbb{E}(f(x_t)-f^*|\mathscr{F}_{t-1}) \leq (1-\gamma_t \frac{2\mu}{\sqrt{G_1+\varepsilon \mathbf{1}}})(f(x_{t-1})-f^*)+O(\gamma_t^2) \tag{2.72}$$

for sufficiently large t. By applying the first conclusion in Corollary 3, we have that $f(x_t)-f^* = o(\frac{1}{t^{1-\theta_2}})$ for any $\theta_2 \in (2\theta_1, 1)$.

2. Combining Proposition 2 and (2.71), we know that $\sum_{t=1}^{\infty} \gamma_t \|\nabla f(x_{t-1})\|^2$ is almost surely convergent, and the conclusion follows from Corollary 3. □

Corollary 4. *Suppose that Assumptions 2, 3, and 5 hold with $p=2$. Consider the Adadelta method described in Algorithm 4 with almost surely bounded $\{u_t\}$, $\{x_t\}$ and $\{g_t\}$. Then, it has the same conclusions as Theorem 4.*

Proof. We again utilize X, G_1, G_2 and N as almost sure upper bounds mentioned in (2.18), (2.19), (2.21) and (2.34).

Following a similar approach as in (2.70) to (2.71), we apply the mean value theorem and conditional expectation with respect to $\mathscr{F}_{t-1}$ on the definition of L-smoothness (2.6) and it gives

$$\begin{aligned} \mathbb{E}(f(x_t)-f^*|\mathscr{F}_{t-1}) &\leq f(x_{t-1})-f^*-\gamma_t \mathbb{E}(\langle \nabla f(x_{t-1}), \frac{\sqrt{u_{t-1}+\varepsilon \mathbf{1}}}{\sqrt{v_t+\varepsilon \mathbf{1}}} g_t \rangle + O(\gamma_t^2)\|\nabla f(x_{t-1})\|\|g_t\| \\ &\quad + \frac{L}{2}(\frac{\gamma_t}{\sqrt{\varepsilon \mathbf{1}}})^2 \|g_t\|^2 | \mathscr{F}_{t-1}) \\ &\leq f(x_{t-1})-f^*-\frac{\gamma_t \sqrt{\varepsilon \mathbf{1}}}{\sqrt{G_1+\varepsilon \mathbf{1}}}\|\nabla f(x_{t-1})\|^2 + O(\gamma_t^2). \end{aligned} \tag{2.73}$$

The rest of the proof is similar to that of Theorem 4. □

Theorem 5. *Suppose that Assumptions 1, 3, and 5 hold with $p=2$. Consider the Adam method described in Algorithm 1 with almost surely bounded $\{x_t\}$ and $\{g_t\}$. It is chosen that $\alpha_t = 1-a\gamma_t$, $\beta_t = 1-b\gamma_t$ and γ_t. Then, the following hold:*

1. *If Assumption 4 also hold,*

$$\|m_t\| = O(\|\nabla f(x_t)\|) \text{ almost surely,} \tag{2.74}$$

and $\gamma_t = O\left(\frac{1}{t^{1-\theta_1}}\right)$ for $\theta_1 \in (0, \frac{1}{2})$, then it follows that

$$f(x_t)-f^* = o\left(\frac{1}{t^{1-\theta_2}}\right) \tag{2.75}$$

for any $\theta_2 \in (2\theta_1, 1)$.

2. *If* γ_t *satisfies*

$$\sum_{t=1}^{\infty} \frac{\gamma_t}{\sum_{i=1}^{t-1} \gamma_i} = \infty \text{ and } \sum_t \gamma_t exp\left(-\sum_{i=1}^{t} \gamma_i\right) < \infty, \tag{2.76}$$

then

$$\min_{1 \le i \le t} \|m_i\|^2 = o\left(\frac{1}{\sum_{i=1}^{t} \gamma_i}\right) \text{ almost surely}. \tag{2.77}$$

If we further assume that

$$\sum_{i=1}^{t} \gamma_{i+1} \|\nabla f(x_i)\|^2 = O\left(\sum_{i=1}^{t} \gamma_{i+1} \|m_i\|^2\right) \text{ almost surely}, \tag{2.78}$$

then we have that

$$\min_{1 \le i \le t} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{\sum_{i=1}^{t} \gamma_i}\right) \text{ almost surely.} \tag{2.79}$$

For example, if we choose $\gamma_t = \frac{1}{t^{\frac{1}{2}+\theta_3}}$ *for* $\theta_3 \in (0, \frac{1}{2})$*, then* (2.77) *and* (2.79) *become*

$$\min_{1 \le i \le t} \|m_i\|^2 = o\left(\frac{1}{t^{\frac{1}{2}-\theta_3}}\right) \text{ and } \min_{1 \le i \le t} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{t^{\frac{1}{2}-\theta_3}}\right) \text{ almost surely}. \tag{2.80}$$

Proof. We begin by discussing the scalar case ($n = 1$). By applying the mean value theorem, we derive

$$\begin{aligned}
& \frac{\gamma_t}{\sqrt{\hat{v}_t + \varepsilon}} \nabla f(x_{t-1}) \hat{m}_t \\
\overset{(a)}{=} & \frac{1}{1 - \Pi_{i=1}^{t}(1 - a\gamma_i)} \frac{\gamma_t}{\sqrt{\hat{v}_t + \varepsilon}} \nabla f(x_{t-1}) m_t \\
= & \left(1 + \frac{\Pi_{i=1}^{t}(1 - a\gamma_i)}{(1 - c_t \Pi_{i=1}^{t}(1 - a\gamma_i))^2}\right) \frac{\gamma_t}{\sqrt{\hat{v}_t + \varepsilon}} \nabla f(x_{t-1}) m_t \\
= & \frac{\gamma_t}{\sqrt{\hat{v}_t + \varepsilon}} \nabla f(x_{t-1}) m_t + \gamma_t O(\Pi_{i=1}^{t}(1 - a\gamma_i)) \\
\le & \frac{\gamma_t}{\sqrt{\hat{v}_t + \varepsilon}} \nabla f(x_{t-1}) m_t + \gamma_t O(\Pi_{i=1}^{t} exp(-a\gamma_i)) \\
= & \frac{\gamma_t}{\sqrt{\hat{v}_t + \varepsilon}} \nabla f(x_{t-1}) m_t + \gamma_t O(exp(-\textstyle\sum_{i=1}^{t} \gamma_i)),
\end{aligned} \tag{2.81}$$

where (a) follows the mean value theorem and $c_{2t} \in (0, 1)$. In a similar way, we could further derive that

$$\frac{\gamma_t}{\sqrt{\hat{v}_t + \varepsilon}} \nabla f(x_{t-1}) \hat{m}_t = \frac{\gamma_t}{\sqrt{v_t + \varepsilon}} \nabla f(x_{t-1}) m_t + \gamma_t O(exp(-\textstyle\sum_{i=1}^{t} \gamma_i)). \tag{2.82}$$

Given the almost sure boundedness of the trajectory $\{x_t\}$, there exists a compact set $C \subset \mathbb{R}^n$ containing the entire trajectory $\{x_t\}$. The locally Lipschitz condition of the gradient ∇f implies that ∇f is L-Lipschitz with some constant $L > 0$ on C. We again utilize X, G_1, G_2 and N as almost sure upper bounds mentioned in (2.18), (2.19), (2.20), (2.21) and (2.34). It is implied by (2.6) that

$$
\begin{aligned}
& f(x_t) + \tfrac{1}{2a\sqrt{v_t+\varepsilon}}|m_t|^2 \\
\leq\ & f(x_{t-1}) - \tfrac{\gamma_t}{\sqrt{\hat{v}_t+\varepsilon}}\nabla f(x_{t-1})\hat{m}_t + \tfrac{L}{2}(\tfrac{\gamma_t}{\sqrt{\hat{v}_t+\varepsilon}})^2|\hat{m}_t|^2 + \tfrac{1}{2a\sqrt{v_t+\varepsilon}}|m_t|^2 \\
\overset{(2.82)}{\leq}\ & f(x_{t-1}) - \tfrac{\gamma_t}{\sqrt{v_t+\varepsilon}}\nabla f(x_{t-1})m_t + \gamma_t O(exp(-\textstyle\sum_{i=1}^t \gamma_i)) \\
& + \tfrac{L}{2}(\tfrac{\gamma_t}{\sqrt{\hat{v}_t+\varepsilon}})^2|\hat{m}_t|^2 + \tfrac{1}{2a\sqrt{v_t+\varepsilon}}|m_t|^2 \\
\leq\ & f(x_{t-1}) - (\tfrac{\gamma_t}{\sqrt{v_{t-1}+\varepsilon}} - b\gamma_t^2 \tfrac{|g_t^{\odot 2} - v_{t-1}|}{2(1-c_{2t}\gamma_t)v_{t-1}+c_1\gamma_t g_t^{\odot 2}+\varepsilon)^{3/2}})\nabla f(x_{t-1})m_t \\
& + \gamma_t^2 \tfrac{L}{2(\hat{v}_t+\varepsilon)}|\hat{m}_t|^2 + (\tfrac{1}{2a\sqrt{v_{t-1}+\varepsilon}} + b\gamma_t \tfrac{v_{t-1}-g_t^{\odot 2}}{4a((1-c_{2t}\gamma_t)v_{t-1}+c_1\gamma_t g_t^{\odot 2}+\varepsilon)^{3/2}})|m_t|^2 \\
\leq\ & f(x_{t-1}) - \tfrac{\gamma_t}{\sqrt{v_{t-1}+\varepsilon}}\nabla f(x_{t-1})m_t + \tfrac{1}{2a\sqrt{v_{t-1}+\varepsilon}}|m_t|^2 \\
& + b\gamma_t \tfrac{v_{t-1}-g_t^{\odot 2}}{4a((1-c_{2t}\gamma_t)v_{t-1}+c_1\gamma_t g_t^{\odot 2}+\varepsilon)^{3/2}}|m_t|^2 + O(\gamma_t^2) + \gamma_t O(exp(-\textstyle\sum_{i=1}^t \gamma_i)) \\
\leq\ & f(x_{t-1}) - \tfrac{\gamma_t}{\sqrt{v_{t-1}+\varepsilon}}\nabla f(x_{t-1})((1-a\gamma_t)m_{t-1} + a\gamma_t g_t) \\
& + \tfrac{1}{2a\sqrt{v_{t-1}+\varepsilon}}((1-2a\gamma_t+a^2\gamma_t^2)|m_{t-1}|^2 + a^2\gamma_t^2|g_t|^2 + 2a\gamma_t(1-a\gamma_t))\langle g_t, m_{t-1}\rangle) \\
& + b\gamma_t \tfrac{v_{t-1}-g_t^{\odot 2}}{4a((1-c_{2t}\gamma_t)v_{t-1}+c_1\gamma_t g_t^{\odot 2}+\varepsilon)^{3/2}}(|(1-a\gamma_t)m_{t-1} + a\gamma_t g_t|^2) \\
& + O(\gamma_t^2) + \gamma_t O(exp(-\textstyle\sum_{i=1}^t \gamma_i)) \\
\leq\ & f(x_{t-1}) + \tfrac{1}{2a\sqrt{v_{t-1}+\varepsilon}}|m_{t-1}|^2 - \tfrac{\gamma_t}{\sqrt{v_{t-1}+\varepsilon}}(\nabla f(x_{t-1}) - g_t)m_{t-1} \\
& + \gamma_t\left(\tfrac{b(v_{t-1}-g_t^{\odot 2})}{4a((1-c_{2t}\gamma_t)v_{t-1}+c_1\gamma_t g_t^{\odot 2}+\varepsilon)^{3/2}} - \tfrac{1}{\sqrt{v_{t-1}+\varepsilon}}\right)|m_{t-1}|^2 \\
& + O(\gamma_t^2) + \gamma_t O(exp(-\textstyle\sum_{i=1}^t \gamma_i))
\end{aligned}
\tag{2.83}
$$

Taking conditional expectation on the above inequality, we obtain

$$
\begin{aligned}
& \mathbb{E}(f(x_t) + \tfrac{1}{2a\sqrt{v_t+\varepsilon}}|m_t|^2|\mathscr{F}_{t-1}) \\
\leq\ & f(x_{t-1}) + \tfrac{1}{2a\sqrt{v_{t-1}+\varepsilon}}|m_{t-1}|^2 + + O(\gamma_t^2) + \gamma_t O(exp(-\textstyle\sum_{i=1}^t \gamma_i)) \\
& + \gamma_t \mathbb{E}\left(\tfrac{b(v_{t-1}-g_t^{\odot 2})}{4a((1-c_{2t}\gamma_t)v_{t-1}+c_1\gamma_t g_t^{\odot 2}+\varepsilon)^{3/2}} - \tfrac{1}{\sqrt{v_{t-1}+\varepsilon}}\middle|\mathscr{F}_{t-1}\right)|m_{t-1}|^2
\end{aligned}
\tag{2.84}
$$

From the result in Theorem 1, the term $\mathbb{E}\left(\frac{b(v_{t-1}-g_t^{\odot 2})}{4a((1-c_{2t}\gamma_t)v_{t-1}+c_1\gamma_t g_t^{\odot 2}+\varepsilon)^{3/2}}\middle|\mathscr{F}_{t-1}\right)$ eventually

converges to 0, and $\mathbb{E}\left(-\frac{1}{\sqrt{v_{t-1}+\varepsilon}}\middle|\mathscr{F}_{t-1}\right)$ eventually converges to $-\frac{1}{\sqrt{S(x^c)+\varepsilon}}<0$ almost surely. That is to say, we have

$$\mathbb{E}\left(\frac{b(v_{t-1}-g_t^{\odot 2})}{4a((1-c_{2t}\gamma_t)v_{t-1}+c_1\gamma_t g_t^{\odot 2}+\varepsilon)^{3/2}}-\frac{1}{\sqrt{v_{t-1}+\varepsilon}}\middle|\mathscr{F}_{t-1}\right)\leq -\frac{3}{4\sqrt{v_{t-1}+\varepsilon}}\leq -\frac{1}{4\sqrt{S(x^c)+\varepsilon}}-\frac{1}{4\sqrt{v_{t-1}+\varepsilon}} \tag{2.85}$$

holds almost surely after a finite time of iterations.

1. By assumption (2.74), (2.84) implies that

$$\begin{aligned}
&\mathbb{E}(f(x_t)-f^*+\tfrac{1}{2a\sqrt{v_t+\varepsilon}}|m_t|^2|\mathscr{F}_{t-1})\\
\leq\ & f(x_{t-1})-f^*+\tfrac{1}{2a\sqrt{v_{t-1}+\varepsilon}}|m_{t-1}|^2+O(\gamma_t^2)+\gamma_t O(exp(-\textstyle\sum_{i=1}^t\gamma_i))\\
&+\gamma_t\left(-\frac{1}{4\sqrt{S(x^*)+\varepsilon}}-\frac{1}{4\sqrt{v_{t-1}+\varepsilon}}\right)|m_{t-1}|^2\\
\leq\ & f(x_{t-1})-f^*+(1-O(\gamma))\tfrac{1}{2a\sqrt{v_{t-1}+\varepsilon}}|m_{t-1}|^2+O(\gamma_t^2)+\gamma_t O(exp(-\textstyle\sum_{i=1}^t\gamma_i))\\
&-\gamma_t\tfrac{1}{4\sqrt{S(x^*)+\varepsilon}}|\nabla f(x_{t-1})|^2\\
\leq\ & f(x_{t-1})+(1-O(\gamma))\tfrac{1}{2a\sqrt{v_{t-1}+\varepsilon}}|m_{t-1}|^2+O(\gamma_t^2)+\gamma_t O(exp(-\textstyle\sum_{i=1}^t\gamma_i))\\
&-\gamma_t\tfrac{1}{4\sqrt{S(x^*)+\varepsilon}}\tfrac{2}{\mu}(f(x_{t-1})-f^*)\\
\leq\ & (1-O(\gamma))\left(f(x_{t-1})+\tfrac{1}{2a\sqrt{v_{t-1}+\varepsilon}}|m_{t-1}|^2\right)+O(\gamma_t^2)+\gamma_t O(exp(-\textstyle\sum_{i=1}^t\gamma_i)).
\end{aligned} \tag{2.86}$$

The conclusion follows from Corollary 3.

2. We also obtain

$$\begin{aligned}
\mathbb{E}(f(x_t)+\tfrac{1}{2a\sqrt{v_t+\varepsilon}}|m_t|^2|\mathscr{F}_{t-1}) \leq\ & f(x_{t-1})+\tfrac{1}{2a\sqrt{v_{t-1}+\varepsilon}}|m_{t-1}|^2-\gamma_t\tfrac{1}{4\sqrt{S(x^c)+\varepsilon}}|m_{t-1}|^2\\
&+O(\gamma_t^2)+\gamma_t O(exp(-\textstyle\sum_{i=1}^t\gamma_i))
\end{aligned} \tag{2.87}$$

from (2.84) and (2.85).

The approach for the higher-dimensional case proceeds in a similar manner as the scalar case:

$$\begin{aligned}
&\mathbb{E}(f(x_t)+\textstyle\sum_i\frac{1}{2a\sqrt{(v_t)_i+\varepsilon}}|(m_t)_i|^2|\mathscr{F}_{t-1})\\
\leq\ & f(x_{t-1})+\textstyle\sum_i\frac{1}{2a\sqrt{(v_{t-1})_i+\varepsilon}}|(m_{t-1})_i|^2-\gamma_t\sum_i\frac{1}{2\sqrt{(S(x^c))_i+\varepsilon}}|(m_{t-1})_i|^2\\
&+O(\gamma_t^2)+\gamma_t O(exp(-\textstyle\sum_{i=1}^t\gamma_i))\\
\leq\ & f(x_{t-1})+\textstyle\sum_i\frac{1}{2a\sqrt{(v_{t-1})_i+\varepsilon}}|(m_{t-1})_i|^2-\gamma_t\frac{1}{2\sqrt{(\max_i S(x^c))_i+\varepsilon}}\|(m_{t-1})\|^2\\
&+O(\gamma_t^2)+\gamma_t O(exp(-\textstyle\sum_{i=1}^t\gamma_i)).
\end{aligned} \tag{2.88}$$

From Proposition 2 and Corollary 3, we have that

$$\sum_{i=1}^{\infty} \gamma_{i+1} \|m_i\|^2 < \infty \text{ almost surely,} \tag{2.89}$$

and the convergence rate for the estimated gradient is given by

$$\min_{1\le i\le t} \|m_i\|^2 = o\left(\frac{1}{\sum_{i=1}^{t} \gamma_i}\right) \text{ almost surely.} \tag{2.90}$$

If we further assume that

$$\sum_{i=1}^{t} \gamma_{i+1} \|\nabla f(x_i)\|^2 = O(\sum_{i=1}^{t} \gamma_{i+1} \|m_i\|^2) \text{ almost surely,} \tag{2.91}$$

we can conclude

$$\min_{1\le i\le t} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{\sum_{i=1}^{t} \gamma_i}\right) \text{ almost surely} \tag{2.92}$$

from Corollary 3. □

Corollary 5. *Nadam method described in Algorithm 3 has the same conclusions as Theorem 5*

Proof. The proof on Nadam method is similar to that of Theorem 5. We omitted the details here. □

Corollary 6. *AMSgrad method described in Algorithm 5 has the same conclusions as Theorem 5*

Proof. We again utilize X, G_1, G_2 and N as almost sure upper bounds mentioned in (2.18), (2.19), (2.20), (2.21) and (2.34). Considering the iteration of $\hat{v}_t$, we have

$$0 < \hat{v}_{t-1} \le \hat{v}_t = \max(v_{t-1} + b\gamma_t(g_t^{\odot 2} - v_{t-1}), \hat{v}_{t-1}) \le \hat{v}_{t-1} + b\gamma_t(G_2 + G_1) \text{ a.s.,} \tag{2.93}$$

and

$$\frac{1}{\sqrt{(\hat{v}_{t-1})_i + \varepsilon}} \ge \frac{1}{\sqrt{(\hat{v}_t)_i + \varepsilon}} \ge \frac{1}{\sqrt{(\hat{v}_{t-1})_i + b\gamma_t(G_2 + G_1) + \varepsilon}} \overset{(a)}{=} \frac{1}{\sqrt{(\hat{v}_{t-1})_i + \varepsilon}} + O(\gamma_t) \text{ a.s.} \tag{2.94}$$

where (a) follows from the mean value theorem.

We start with discussing the scalar case ($n = 1$). Applying (2.6) on $f(x_t)$, we have

$$
\begin{aligned}
& f(x_t) + \tfrac{1}{2a\sqrt{\hat{v}_t+\varepsilon}}|m_t|^2 \\
\leq\ & f(x_{t-1}) - \gamma_t \tfrac{1}{\sqrt{\hat{v}_t+\varepsilon}}\nabla f(x_{t-1})\hat{m}_t + \gamma_t^2 \tfrac{L^2}{\hat{v}_t+\varepsilon}|\hat{m}_t|^2 + \tfrac{1}{2a\sqrt{\hat{v}_t+\varepsilon}}|m_t|^2 \\
\overset{(2.81)}{\leq}\ & f(x_{t-1}) - \gamma_t \tfrac{1}{\sqrt{\hat{v}_t+\varepsilon}}\nabla f(x_{t-1})m_t + \tfrac{1}{2a\sqrt{\hat{v}_t+\varepsilon}}|m_t|^2 + O(\gamma_t exp(-\textstyle\sum_{i=1}^t \gamma_i)) + O(\gamma_t^2) \\
\leq\ & f(x_{t-1}) - \gamma_t \tfrac{1}{\sqrt{\hat{v}_t+\varepsilon}}\nabla f(x_{t-1})((1-a\gamma_t)m_{t-1} + a\gamma_t g_t) \\
& + \tfrac{1}{2a\sqrt{\hat{v}_t+\varepsilon}}((1-a\gamma_t)m_{t-1} + a\gamma_t g_t)^2 + O(\gamma_t exp(-\textstyle\sum_{i=1}^t \gamma_i)) + O(\gamma_t^2) \\
\leq\ & f(x_{t-1}) - \gamma_t \tfrac{1}{\sqrt{\hat{v}_t+\varepsilon}}\nabla f(x_{t-1})m_{t-1} + \tfrac{1}{2a\sqrt{\hat{v}_t+\varepsilon}}((1-2a\gamma_t)|m_{t-1}|^2 + 2a\gamma_t m_{t-1}g_t) \\
& + O(\gamma_t exp(-\textstyle\sum_{i=1}^t \gamma_i)) + O(\gamma_t^2) \\
\overset{(2.94)}{\leq}\ & f(x_{t-1}) + \tfrac{1}{2a\sqrt{\hat{v}_t+\varepsilon}}|m_{t-1}|^2 - \gamma_t \tfrac{1}{\sqrt{\hat{v}_{t-1}+\varepsilon}}\nabla f(x_{t-1})m_{t-1} + \\
& \gamma_t \tfrac{1}{\sqrt{\hat{v}_{t-1}+\varepsilon}}(-|m_{t-1}|^2 + m_{t-1}g_t) + O(\gamma_t exp(-\textstyle\sum_{i=1}^t \gamma_i)) + O(\gamma_t^2) \\
\overset{(2.94)}{\leq}\ & f(x_{t-1}) + \tfrac{1}{2a\sqrt{\hat{v}_{t-1}+\varepsilon}}|m_{t-1}|^2 - \gamma_t \tfrac{1}{\sqrt{\hat{v}_{t-1}+\varepsilon}}(\nabla f(x_{t-1}) - g_t)m_{t-1} - \gamma_t \tfrac{1}{\sqrt{\hat{v}_{t-1}+\varepsilon}}|m_{t-1}|^2 \\
& + O(\gamma_t exp(-\textstyle\sum_{i=1}^t \gamma_i)) + O(\gamma_t^2).
\end{aligned}
\tag{2.95}
$$

Taking the conditional expectation on both sides, we obtain

$$
\begin{aligned}
\mathbb{E}(f(x_t) + \tfrac{1}{2a\sqrt{\hat{v}_t+\varepsilon}}|m_t|^2|\mathscr{F}_{t-1}) \leq\ & f(x_{t-1}) + \tfrac{1}{2a\sqrt{\hat{v}_{t-1}+\varepsilon}}|m_{t-1}|^2 - \gamma_t \tfrac{1}{\sqrt{\hat{v}_{t-1}+\varepsilon}}|m_{t-1}|^2 \\
& + O(\gamma_t exp(-\textstyle\sum_{i=1}^t \gamma_i)) + O(\gamma_t^2).
\end{aligned}
\tag{2.96}
$$

The rest of the proof is similar to that of Theorem 5. □

Remark 4. *(Comments on (2.74) and (2.78)) We proved that the linearly interpolated processes of $(m_t^T, v_t^T, x_t^T)^T$ are asymptotic pseudo-trajectories for $(m^T, v^T, x^T)^T$ in the context of the Adam, Nadam, and AMSgrad optimization methods. It allows us to consider $(m^T, v^T, x^T)^T$ as an approximation of the linearly interpolated processes of $(m_t^T, v_t^T, x_t^T)^T$. Notably, when we treat v as a constant, the limiting ODE (2.23), common to the Adam, Nadam, and AMSgrad methods, takes the form*

$$
\begin{aligned}
\dot{m} &= a(\nabla f(x) - m), \\
\dot{x} &= -\frac{m}{\sqrt{v+\varepsilon \boldsymbol{I}}}.
\end{aligned}
\tag{2.97}
$$

This system describes a damped dynamical system with position x, velocity $\dot{x}$, and acceleration $\ddot{x}$:

$$
\ddot{x} + a\dot{x} + \frac{a}{\sqrt{v+\varepsilon \boldsymbol{I}}}\nabla f(x) = 0. \tag{2.98}
$$

In the context of an underdamped system, the kinetic energy $O(|\dot{x}|^2)$ and the potential energy $f(x)$ steadily diminish to zero while seamlessly transitioning between one another. The damping term facilitates the energy exchange, leading to an overall decrease in total energy. For functions adhering to the condition $f(x) = O(\|\nabla f(x)\|^2)$, the relationship

$$O\left(\int_{t=0}^{\infty} \|\nabla f(x)\|^2 dt\right) = O\left(\int_{t=0}^{\infty} f(x) dt\right) = \int_{t=0}^{\infty} \|m\|^2 dt \tag{2.99}$$

holds. This result agrees with the assumptions stated in (2.78). The underdamped phenomenon aligns with the illustrated figures in Section 2.1.2. In cases where the system exhibits overdamping or critical damping, the relationship

$$|x| = O(|\dot{x}|) = O(|m|) \tag{2.100}$$

holds. This result agrees with the assumption stated in (2.74), and further implies (2.78).

On the other hand, it is mentioned in the first paper on Adam method [39] that m_t is the estimate of the gradient; therefore, Assumptions 2.74 and 2.78 seem to be reasonable assumptions.

Remark 5. *(General convex functions) Consider Adam, Nadam, RMSprop, Adadelta, and AMSgrad methods. Suppose that the learning rate is chosen to be $\gamma_t = O\left(\frac{1}{t^{\frac{2}{3}+\varepsilon}}\right)$ for any $\varepsilon \in \left(0, \frac{1}{3}\right)$, and f is generally convex, then we have that*

$$f(x_t) - f^* = O\left(\frac{1}{t^{\frac{1}{3}-\varepsilon}}\right) \tag{2.101}$$

and $x \to x^*$. The proof can be conducted using the mean-value and boundedness techniques in this section and [45, Lemma 4].

2.5 Summary

The first part of this chapter highlights a significant issue with these methods when parameters are constant: they fail to converge towards the critical point. To tackle this, we propose a unified modification for these methods, successfully achieving convergence towards the crucial point in the given example. This modification addresses a fundamental limitation in current ASG methodologies.

Building on the ideas presented in [4], this chapter provides a clear and rigorous proof of almost sure convergence towards a critical point for smooth and non-convex objective functions,

while also correcting some errors found in [4]. This represents the first instance of establishing almost sure convergence rates for these methods. For non-convex objective functions, our findings, influenced by [45], demonstrate that a weighted average of squared gradient norms in these methods converges at a rate of $o(1/t^{\frac{1}{2}-\theta})$ for all $\theta \in (0, \frac{1}{2})$. Additionally, for strongly convex functions, we show that the convergence rates for RMSprop and Adadelta can be improved to $o(1/t^{1-\theta})$ for all $\theta \in (0, \frac{1}{2})$.

Overall, this chapter makes significant contributions to adaptive gradient-based optimization methods, particularly in addressing convergence issues and establishing convergence rates for both non-convex and strongly convex objective functions.

Chapter 3

On Almost Sure Convergence of Hogwild! Algorithm

In the last chapter, we have seen the almost sure convergence of ASG method, for example Adam and Adadelta. This chapter explores the guaranteed convergence rates of Hogwild! algorithm for different types of loss functions. Hogwild! algorithm is a lock-free approach to parallelizing SGD method, and its development can be found in Chapter 1. Considering the strongly convex overall loss function and convex local loss functions, Nguyen et al. show that the virtual state converges to the optimal solution with probability one [56]. Up to now, there are some unknown problems: Is it possible to relex the conditions? What is the almost sure convergence rate?

To answer the posed questions, we will regard Hogwild! as a delayed SGD, and leverage the findings from [45] to analyze the almost sure convergence. We first explore the Hogwild! algorithm's convergence rates for different loss functions. We prove its fast convergence on strongly convex functions, matching the best rates of classic SGD methods with minimal error. For non-convex functions, we show that both a weighted average of squared gradients and the algorithm's later iterations converge to zero. Additionally, we analyze the last-iterate convergence for general convex smooth functions, providing insights into its efficiency across various settings.

3.1 The Hogwild! Algorithm

The Hogwild! algorithm 7 is a parallel SGD method introduced by Feng et al in the paper [63]. The iteration can be rewritten as

$$(x_{t+1})_{u_t} = (x_t)_{u_t} - \eta_t d_{\xi_t}(\nabla f(\hat{x}_t;\xi_t))_{u_t}. \tag{3.1}$$

Algorithm 7: Hogwild! algorithm

Data: $x_0 \in \mathbb{R}^n$
for $t = 0, 1, 2, \ldots$ *in parallel* **do**
  read current shared memory $\hat{x}_t$;
  generate a random variable ξ_t and evaluate the stochastic gradient $\nabla f(\hat{x}_t; \xi_t)$;
  for *Sample position* u_t *uniformly from the set*
   $E_\xi = \{$*positions where* $\nabla f(x_t; \xi_t)$*is nonzero*$\} \subset \{1, 2, \ldots, n\}$ **do**
    $x_{t+1} = x_t - \eta_t d_{\xi_t} S_{\xi_t, u_t} \nabla f(\hat{x}_t; \xi_t)$
  end
end

Consider a fixed ξ. The scalar d_ξ represents the number of nonzero entries in $\nabla f(\cdot; \xi)$, and $S_{\xi,u}$ is a diagonal matrix equal to 1 on the u-th diagonal and zeros elsewhere. Therefore, $d_\xi \leq n$ for all ξ. The matrix $S_{\xi,u}$ filters which positions of $\nabla f(\cdot; \xi)$ is nonzero and contributes to the iteration. For a given ξ, we take the following expectation over u

$$d_\xi \mathbb{E}[S_{\xi,u} | \xi] = D_\xi, \tag{3.2}$$

and obtain that D_ξ is a diagonal $0/1$ matrix whose 1-entries corresponds to the non-zero positions in $\nabla f(w; \xi)$. In other words, for a given ξ, the i-th entry on D_ξ's diagonal is 1 if and only if the i-th position of $\nabla f(\cdot; \xi)$ is not a zero function. More details are shown in [56].

3.2 Convergence with Probability One

For the analysis of the almost sure convergence of the Hogwild! methods, the following assumptions are made.

Assumption 6. *[μ-strongly convex] The objective function $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex for some $\mu > 0$, i.e.,*

$$f(x) \geq f(x') + \langle \nabla f(x'), x - x'; \rangle + \frac{\mu}{2} \|x - x'\|^2 \tag{3.3}$$

for all $x, x' \in \mathbb{R}^n$.

Combining the fact that $\nabla f(x^*) = 0$ and [55, Theorem 2.1.10], we know that μ-strong convexity implies

$$f(x) - f^* \leq \frac{1}{2\mu} \|\nabla f(x)\|^2. \tag{3.4}$$

Assumption 7. *[L-smoothness] $\nabla f(w;\xi)$ is L-Lipschitz for every realization of ξ,i.e., there exists $L>0$ such that*

$$\|\nabla f(x;\xi)-\nabla f(x';\xi)\| \le L\|x-x'\| \tag{3.5}$$

for all $x,x' \in \mathbb{R}^n$.

The L-smoothness of a function f can be implied by equation 7. According to [55], we have that

$$f(x) \le f(x') + \langle \nabla f(x'), x-x' \rangle + \frac{L}{2}\|x-x'\|^2 \tag{3.6}$$

Noting that $\hat{x}_t$ is usually a lag state read many clock cycles earlier, i.e., $\{\hat{x}_t = x_{k_t}\}_t$ is a subsequence of $\{x_t\}_t$. We assume that there exists a constant delay satisfying the following assumption.

Assumption 8. *The shared memory is consistent with delay τ for all t, that is to say, $k_t - t$ is always less than or equal to τ.*

We assume that there exists a constant delay satisfying this assumption. That is to say, $\hat{x}_t$ might equals to some state in $\{x_{t-\tau}, x_{t-\tau+1}, ..., x_{t-1}\}$, and it includes some of the update during the $(t-\tau)$-th to $(t-1)$-th iterations. We define $\delta_{t,j}$ $(t-\tau \le j \le t-1)$ to be 1 if the j-th iteration is included in $\hat{x}_t$, i.e., $j \le k_t$ and $\delta_{t,j}$ is 0 elsewise. That is to say,

$$\hat{x}_t = x_{t-\tau} - \sum_{j=t-\tau}^{t-1} \delta_{t,j}\eta_j d_{\xi_j} S_{\xi_j,u_j} \nabla f(\hat{x}_j;\xi_j) \tag{3.7}$$

A nice and short discussion for the almost sure convergence rates of stochastic gradient descent methods was made in [45], which relies on the results in Appendix B, i.e., the classical supermartingale convergence theorem from [67] and its corollaries derived by [45]. Motivated by [45], these two results will also be utilized in analyzing the almost sure convergence rates for ASG methods. The rates in (3.13) and (3.24) match the lower bounds for stochastic gradient-based algorithm, $O\left(\frac{1}{t}\right)$ for strongly convex loss function, and $O\left(\frac{1}{t^{0.5}}\right)$ for nonconvex loss function [2], to an ε-factor.

Lemma 1. *[56, Lemma 6]*

$$\mathbb{E}[\|d_{\xi_t} S_{\xi_t,u_t} \nabla f(\hat{x}_t;\xi_t)\|^2 | \mathscr{F}_t, \xi_t] \le D\|\nabla f(\hat{x}_t;\xi_t)\|^2 \tag{3.8}$$

$$\mathbb{E}[d_{\xi_t} S_{\xi_t,u_t} \nabla f(\hat{x}_t;\xi_t) | \mathscr{F}_t] = \nabla F(\hat{x}_t) \tag{3.9}$$

Theorem 6. *Suppose that Assumptions 7 and 8 hold. Consider the Hogwild! method described in Algorithm 7. Then, the following hold:*

1. *If $M_0 = \max_\xi \|\nabla f(w_0;\xi)\|$ is finite almost surely, and we choose $\eta_t = \frac{1}{Ln(2+\beta)t}$ for some $\beta > 0$, then it follows that*

$$\min_{1\le i\le t} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{lnt}\right) \text{ almost surely.} \tag{3.10}$$

2. *If $\{\|\nabla f(\hat{x}_t,\xi_t)\|\}$ is almost surely bounded and $\{\eta_t\}$ is a decreasing sequence of positive real numbers satisfying*

$$\sum_{t=1}^{\infty} \frac{\eta_t}{\sum_{i=1}^{t-1}\eta_i} = \infty \text{ and } \sum_t \eta_t^2 < \infty, \tag{3.11}$$

then it follows that

$$\min_{1\le i\le t} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{\sum_{i=1}^{t}\eta_i}\right) \text{ almost surely.} \tag{3.12}$$

In particularly, if we choose $\eta_t = O\left(\frac{1}{t^{1/2+\rho}}\right)$ for some $0 < \rho < \frac{1}{2}$, then

$$\min_{1\le i\le t} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{t^{1/2-\rho}}\right) \text{ almost surely.} \tag{3.13}$$

Proof. According to (3.6), we have that

$$f(x_{t+1}) \le f(x_t) - \eta_t \langle \nabla f(x_t), d_{\xi_t} S_{\xi_t,u_t} \nabla f(\hat{x}_t;\xi_t)\rangle + \frac{L\eta_t^2}{2}\|d_{\xi_t} S_{\xi_t,u_t}\nabla f(\hat{x}_t;\xi_t)\|^2. \tag{3.14}$$

Taking the conditional expectation on (3.14), the following inequality holds with probability

one:

$$
\begin{aligned}
& \mathbb{E}[f(x_{t+1})|\mathscr{F}_t] \\
\leq\ & f(x_t) - \eta_t \langle \nabla f(x_t), \mathbb{E}[d_{\xi_t} S_{\xi_t,u_t} \nabla f(\hat{x}_t;\xi_t)|\mathscr{F}_t]\rangle + \tfrac{LD^2\eta_t^2}{2}\mathbb{E}[\|\nabla f(\hat{x}_t;\xi_t)\|^2|\mathscr{F}_t] \\
\leq\ & f(x_t) - \eta_t \langle \nabla f(x_t), \nabla f(\hat{x}_t)\rangle + \tfrac{LD^2\eta_t^2}{2}\mathbb{E}_{\xi_t}[\mathbb{E}[\|d_{\xi_t} S_{u_t}^{\xi_t} \nabla f(\hat{x}_t;\xi_t)\|^2|\mathscr{F}_t,\xi_t]] \\
\leq\ & f(x_t) - \eta_t \|\nabla f(x_t)\|^2 + \eta_t L \|\nabla f(x_t)\| \|\textstyle\sum_{j=t-\tau}^{t-1} \delta_{t,j}\eta_j d_{\xi_t} S_{\xi_t,u_t} \nabla f(\hat{x}_j;\xi_j)\| \\
& + \tfrac{LD^2\eta_t^2}{2}\mathbb{E}_{\xi_t}[\mathbb{E}[\|d_{\xi_t} S_{u_t}^{\xi_t} \nabla f(\hat{x}_t;\xi_t)\|^2|\mathscr{F}_t,\xi_t]] \\
\leq\ & f(x_t) - \eta_t \|\nabla f(x_t)\|^2 + \eta_t L \|\nabla f(x_t)\| \|\textstyle\sum_{j=t-\tau}^{t-1} (1-\delta t,j)\eta_j d_{\xi_j} S_{u_j}^{\xi_j} \nabla f(\hat{x}_j;\xi_j)\| \\
& + \tfrac{LD^2\eta_t^2}{2} n \mathbb{E}_{\xi_t}[\|\nabla f(\hat{x}_t;\xi_t)\|^2] \\
\leq\ & f(x_t) - \eta_t \|\nabla f(x_t)\|^2 + \eta_t L \|\nabla f(x_t)\| \textstyle\sum_{j=t-\tau}^{t-1} \eta_j d_{\xi_j} \|(1-\delta_{t,j}) S_{\xi_j,u_j} \nabla f(\hat{x}_j;\xi_j)\| \\
& + \tfrac{LD^2\eta_t^2}{2} n (\max_{\xi,0\leq j\leq t} \|\nabla f(w_j;\xi)\|)^2 \\
\leq\ & f(x_t) - \eta_t \|\nabla f(x_t)\|^2 + \eta_t L \|\nabla f(x_t)\| n \textstyle\sum_{j=t-\tau}^{t-1} \eta_j \|\nabla f(\hat{x}_j;\xi_j)\| \\
& + \tfrac{LD^2\eta_t^2}{2} n (\max_{\xi,0\leq j\leq t} \|\nabla f(w_j;\xi)\|)^2 \\
\leq\ & f(x_t) - \eta_t \|\nabla f(x_t)\|^2 + \eta_t L \|\nabla f(x_t)\| n\tau\eta_{t-\tau} (\max_{\xi,0\leq j\leq t-1} \|\nabla f(w_j;\xi)\|) \\
& + \tfrac{LD^2\eta_t^2}{2} n (\max_{\xi,0\leq j\leq t} \|\nabla f(w_j;\xi)\|)^2 \\
\leq\ & f(x_t) - \eta_t \|\nabla f(x_t)\|^2 + \tfrac{1}{2}\eta_t^2 L \|\nabla f(x_t)\| n\tau (\max_{\xi,0\leq j\leq t} \|\nabla f(w_j;\xi)\|) \\
& + \tfrac{LD^2\eta_t^2}{2} n (\max_{\xi,0\leq j\leq t} \|\nabla f(w_j;\xi)\|)^2 \\
\leq\ & f(x_t) - (\eta_t - O(\eta_t^2))\|\nabla f(x_t)\|^2 + O(\eta_t^2 (\max_{\xi,0\leq j\leq t} \|\nabla f(w_j;\xi)\|)^2)
\end{aligned}
\tag{3.15}
$$

for sufficiently large t.

1. It is concluded by [56, Lemma 13] that

$$
\max_{\xi,0\leq j\leq t} \|\nabla f(w_j;\xi)\| \leq M_0 exp\left(Ln \sum_{i=0}^{t-1} \eta_i\right) =: M_t \text{ almost surely.} \tag{3.16}
$$

Based on this conclusion, (3.15) becomes that

$$
\mathbb{E}[f(x_{t+1})|\mathscr{F}_t] \leq f(x_t) - (\eta_t - O(\eta_t^2))\|\nabla f(x_t)\|^2 + O(\eta_t^2 M_t^2). \tag{3.17}
$$

According to (3.16) and the proof of [56, Theorem 5], we have

$$
\eta_t M_t \leq \frac{M_0 e^{\frac{1}{2+\beta}}}{2+\beta} t^{-\frac{1+\beta}{2+\beta}} = \frac{M_0 e^{\frac{1}{2+\beta}}}{2+\beta} t^{-\frac{1}{2}-\rho} \text{ almost surely} \tag{3.18}
$$

and

$$\sum_{i=0}^{\infty} \eta_i^2 M_i^2 = \sum_{i=0}^{\infty} O(i^{-1-2\rho}) < \infty \text{ almost surely.} \tag{3.19}$$

From the Proposition 2 and the Corollary 3, we have that

$$\sum \eta_t \|\nabla f(x_t)\|^2 < \infty \text{ almost surely,} \tag{3.20}$$

and the convergence rate for the gradient is given by

$$\min_{1\le i\le t} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{\sum_{i=1}^{t} \eta_i}\right) = o\left(\frac{1}{lnt}\right) \text{ almost surely.} \tag{3.21}$$

2. Assume that $\|\nabla f(\hat{x}_t, \xi_t)\|$ is almost surely bounded by $M > 0$. Then, (3.15) gives

$$\mathbb{E}[f(x_{t+1})|\mathscr{F}_t] \le f(x_t) - (\eta_t - O(\eta_t^2))\|\nabla f(x_t)\|^2 + O(\eta_t^2). \tag{3.22}$$

Considering the fact $\sum_{t=0}^{\infty} O(\eta_t^2) < \infty$ and the Proposition 2, we have that

$$\sum \eta_t \|\nabla f(x_t)\|^2 < \infty \text{ almost surely.} \tag{3.23}$$

The Corollary 3 gives the convergence rate for the gradient. □

Remark 6. *(Remarks on bounded $\{\nabla f(x_t;\xi_t)\}$) A sufficient condition for bounded $\{\nabla f(x_t;\xi_t)\}$ is that If we assume that the collection of ξ_t is a finite set and $\{x_t\}$ is bounded*

Theorem 7. *Suppose that Assumptions 6 to 8 hold. Consider the Hogwild! method described in Algorithm 7. If we choose $\eta_t = O\left(\frac{1}{t^{1-\theta_1}}\right)$ for $\theta_1 \in (0, \frac{1}{2})$ and $\{\|\nabla f(\hat{x}_t, \xi_t)\|\}$ is almost surely bounded, then it follows that*

$$f(x_t) - f^* = o\left(\frac{1}{t^{1-\theta_2}}\right) \tag{3.24}$$

for any $\theta_2 \in (2\theta_1, 1)$.

Proof. Assume that $\|\nabla f(\hat{x}_t, \xi_t)\|$ is almost surely bounded by $M > 0$. Then, (3.15) becomes

$$\begin{aligned} \mathbb{E}[f(x_{t+1}) - f^*|\mathscr{F}_t] &\le f(x_t) - f^* - (\eta_t - O(\eta_t^2))\|\nabla f(x_t)\|^2 + O(\eta_t^2 M^2) \\ &\le (1 - O(\eta_t))(f(x_t) - f^*) + O(\eta_t^2) \end{aligned} \tag{3.25}$$

By applying the first conclusion of 3, we have that $f(x_t) - f^* = o\left(\frac{1}{t^{1-\theta_2}}\right)$ for any $\theta_2 \in (2\theta_1, 1)$. □

Lemma 2. *Let $\{b_t\}$, $\{\eta_t\}$ be two non-negative sequences and $\{a_t\}$ a sequence of vectors in a vector space X. Let $p \geq 1$ and assume $\sum_{t=1}^{\infty} \eta_t b_t^p < \infty$ and $\sum_{t=1}^{\infty} \eta_t = \infty$. Assume also that there exists $L \geq 0$ such that*

$$|b_{t+\tau} - b_t| \leq L \left(\sum_{i=t}^{t+\tau-1} \eta_i b_i + \| \sum_{i=t}^{t+\tau-1} \eta_i a_i \| \right), \tag{3.26}$$

where a_t is such that $\| \sum_{i=1}^{\infty} \eta_t a_t \| < \infty$. Then, b_t converges to 0.

The following proof proceed in a similar way as [45]

Theorem 8. *(Last iteration) Suppose that Assumption 7 and 8 hold. Suppose that $\{\|\nabla f(\hat{x}_t, \xi_t)\|\}$ is almost surely bounded. Consider the Hogwild! method described in Algorithm 7. Then, the following hold:*

1. *If we choose $\eta_t = \frac{1}{Ln(2+\beta)t}$ for some $\beta > 0$, then the last time iteration converges almost surely.*
2. *If we choose $\eta_t = O\left(\frac{1}{t^{1/2+\rho}}\right)$ for some $0 < \rho < \frac{1}{2}$, then the last time iteration converges almost surely.*

Proof. By (3.20), we know that $\sum_{i=0}^{\infty} \eta_i^2 M_i^2 = \sum_{i=0}^{\infty} O(i^{-1-2\rho}) < \infty$ almost surely. For any t', it can be implied by the L-smoothness of f that

$$\begin{aligned}
& |\|\nabla f(x_{t+t'})\| - \|\nabla f(x_t)\|| \\
\leq\ & L\|x_{t+t'} - x_t\| \\
\leq\ & L\left\|\sum_{i=t}^{t+t'-1} \eta_i d_{\xi_i} S_{\xi_i,u_i} \nabla f(\hat{x}_i, \xi_i)\right\| \\
\leq\ & L\left\|\sum_{i=t}^{t+t'-1} \eta_i d_{\xi_i} S_{\xi_i,u_i} (\nabla f(x_i) + \nabla f(\hat{x}_i, \xi_i) - \nabla f(x_i))\right\| \\
\leq\ & L\sum_{i=t}^{t+t'-1} \eta_i \|d_{\xi_i} S_{\xi_i,u_i}\| \|\nabla f(x_i)\| + L\left\|\sum_{i=t}^{t+t'-1} \eta_i d_{\xi_i} S_{\xi_i,u_i} (\nabla f(\hat{x}_i, \xi_i) - \nabla f(x_i))\right\| \\
\leq\ & Ln\sum_{i=t}^{t+t'-1} \eta_i \|\nabla f(x_i)\| + L\left\|\sum_{i=t}^{t+t'-1} \eta_i d_{\xi_i} S_{\xi_i,u_i} (\nabla f(\hat{x}_i, \xi_i) - \nabla f(x_i))\right\|
\end{aligned} \tag{3.27}$$

1. Let $m_t := \sum_{i=1}^{t} \eta_i d_{\xi_i} S_{\xi_i,u_i} (\nabla f(\hat{x}_i, \xi_i) - \nabla f(x_t))$. We can verify that it is a martingale by

definition. Combining (3.16), (3.20) and the triangle inequality, we have that

$$\begin{aligned} & \sum_{t=1}^{\infty} \mathbb{E}[\|m_t - m_{t-1}\|^2] \\ = \; & \sum_{t=1}^{\infty} \eta_i^2 \mathbb{E}[\|d_{\xi_i} S_{\xi_i,u_i}(\nabla f(\hat{x}_i,\xi_i) - \nabla f(x_t))\|^2] \\ \leq \; & \sum_{t=1}^{\infty} \eta_i^2 d_{\xi_i}^2 \mathbb{E}[\|S_{\xi_i,u_i}\|^2 (\|\nabla f(\hat{x}_i,\xi_i)\| + \|\nabla f(x_t)\|)^2] \\ \leq \; & 4n^2 \sum_{t=1}^{\infty} \eta_i^2 M_t^2 \\ < \; & \infty \text{ almost surely.} \end{aligned} \tag{3.28}$$

According to [86], the conclusion (3.28) implies that m_t is $\mathscr{L}^2$ bounded and hence converges almost surely. We conclude that $\nabla f(x_t)$ converges to 0 almost surely in view of Lemma 2.

2. Assume that $\|\nabla f(\hat{x}_t,\xi_t)\|$ is almost surely bounded by $M > 0$. Replacing M_t in part 1 by the constant M, we know that (3.28) becomes

$$\sum_{t=1}^{\infty} \mathbb{E}[\|m_t - m_{t-1}\|^2] \leq 4n^2 \sum_{t=1}^{\infty} \eta_i^2 M^2 \leq \sum_{t=1}^{\infty} O\left(\frac{1}{t^{1+2\rho}}\right). \tag{3.29}$$

The rest of the proof proceeds in the same way as part 1. □

3.3 Summary

Using the results from [45], we present a comprehensive analysis of the Hogwild! algorithm, a non-locking, parallelized form of SGD, widely used in training large-scale machine learning models. The focus of this study is on the algorithm's almost sure convergence rates under various conditions related to the loss function.

The chapter begins by establishing the almost sure convergence rate of the Hogwild! algorithm when applied to strongly convex functions. It is demonstrated that this rate matches the optimal convergence rate of the classic SGD method, achieving convergence to a negligibly small error margin. Further, this chapter explores the behavior of the Hogwild! algorithm when dealing with non-convex loss functions. In these cases, it is shown that both a weighted average of the squared gradient and the outcomes of the algorithm's final iterations converge to zero with high certainty.

Chapter 4

Continuous-time Distributed Convex Optimization via a Gradient-based Algorithm

Chapter 1 delineates the significance of continuous-time optimization and the benefits inherent in distributed optimization strategies. Different from chapters 2 and 3 concerning discrete-time stochastic optimization, for example, Adam and Hogwild algorithms, the subsequent discussions pivot towards exploring continuous-time distributed optimization algorithms.

This chapter proposes a new PI algorithm. As in the previous chapters, we are also interested in its convergence guarantee under different cost functions. Given the constraints of local communication, the PI algorithm's convergence trajectory is anticipated to lag behind that of the centralized gradient descent (GD) method. By analyzing the derivative of the PI algorithm, we demonstrate that its convergence rate aligns with that of centralized gradient descent for strongly convex functions. Additionally, the chapter investigates its local linear convergence when applied to strictly convex functions.

4.1 Description of the Algorithm

Denote by $x^i \in \mathbb{R}^m$ an estimate of the optimal solution x^* by agent $i \in X$, and $v^i \in \mathbb{R}^m$ an adapter for neutralizing the influence caused by the difference of $\nabla f_i(x^*)$ $(i \in X)$. Let $x = ((x^1)^T, \ldots, (x^n)^T)^T \in \mathbb{R}^{nm}$, $v = ((v^1)^T, \ldots, (v^n)^T)^T \in \mathbb{R}^{nm}$, L be the Laplacian matrix of G, and β be some positive constant.

The earliest PI algorithm published in [78] updates according to

$$\begin{aligned} \dot{x} &= -\bar{L}x - \bar{L}v - \nabla(x), \\ \dot{v} &= \bar{L}x, \end{aligned} \tag{4.1}$$

of which the initial values $x(0), v(0) \in \mathbb{R}^{nm}$ are arbitrary, $\bar{L} = L \otimes I_m$, and $\nabla(x) = \begin{pmatrix} \nabla f_1(x_1) \\ \dots \\ \nabla f_n(x_n) \end{pmatrix} \in \mathbb{R}^{nm}$.

Intuitively, the convergence rate of (4.1) should be restricted by the consensus coefficient $\bar{L}$ and the convergence rate of the centralized gradient descent (GD) method, where the rate of consensus is determined by the second largest eigenvalue of L [57]. When the eigenvalue is large enough, the latter limitation dominates, and the convergence rate of the new algorithm is expected to match that of the continuous-time GD method. By changing $\bar{L}$ into $\beta\bar{L}$, we gain more flexibility in tuning the eigenvalue and improving the rate of consensus. It is checked by Definition 1 that βL is also a Laplacian matrix of the given graph G

The novel PI algorithm begins with arbitrary $x(0), v(0) \in \mathbb{R}^{nm}$, and has the following update rule:

$$\begin{aligned} \dot{v} &= \beta\bar{L}x, \\ \dot{x} &= -\beta\bar{L}v - \beta\bar{L}x - \nabla(x). \end{aligned} \tag{4.2}$$

4.2 Assumptions

Here are some assumptions that will be utilized to study the behavior of the proposed algorithms.

Assumption 9. *All $f_i \in C^2(\mathbb{R}^m, \mathbb{R})$ $(i \in X)$ are μ-strongly convex and M-smooth for some $\mu, M > 0$, i.e.*

$$\mu I_m \leq \nabla^2 f_i(z) \leq M I_m \tag{4.3}$$

for all $z \in \mathbb{R}^m$.

Assumption 10. *The graph $G = (V, E, W)$ is connected.*

Assumption 11. *All objective functions $f_i \in C^2(\mathbb{R}^m, \mathbb{R})$ $(i \in X)$ are convex, and $f(x)$ is strictly convex on $\mathbb{R}^m$.*

Remark 7. *Assumption 11 is weaker than Assumption 9. First, $\nabla^2 f$ in Assumption 11 does not necessarily have a positive lower bound, whereas $\nabla^2 f \geq \mu I_m$ in Assumption 9 holds for some $\mu > 0$ on the entire space $\mathbb{R}^m$. Moreover, $\nabla^2 f_i$ in Assumption 11 might vanish for some i, whereas $\nabla^2 f_i \geq \mu I_m$ in Assumption 9 is always true.*

4.3 Main Convergence Results

Before discussing the system (4.2), we define $l = \dot{v} = \beta\bar{L}x$ and $q = \dot{x} = -\beta\bar{L}v - \beta\bar{L}x - \nabla(x)$. Differentiating q and l, we obtain a variation of the algorithm (4.2):

$$\begin{pmatrix} \dot{q} \\ \dot{l} \end{pmatrix} = \begin{pmatrix} -\beta\bar{L} - \nabla^2(x) & -\beta\bar{L} \\ \beta\bar{L} & 0 \end{pmatrix} \begin{pmatrix} q \\ l \end{pmatrix} = -A(x) \begin{pmatrix} q \\ l \end{pmatrix}, \tag{4.4}$$

where $\nabla^2(x) = \mathrm{diag}(\nabla^2 f_1(x_1), \ldots, \nabla^2 f_n(x_n))$ and $A(x) = \begin{pmatrix} \beta\bar{L} + \nabla^2(x) & \beta\bar{L} \\ -\beta\bar{L} & 0 \end{pmatrix}$ are matrix-valued functions.

4.3.1 Convergence Analysis under Strongly Convex Cost Function

We will analyze the auxiliary system (4.4), construct the relationship between (4.4) and (4.2), and conclude that the system (4.2) achieves the convergence rate $O(e^{-\mu t})$.

From the result of [57], we can obtain the following lemma.

Lemma 3. *Let Assumption (10) hold. Then,*

(1) all eigenvalues of L are nonnegative real numbers,

(2) the zero eigenvalue of L is simple, and $Null(L) = Null(L^T) = span(\mathbf{1})$,

(3) $Null(\bar{L}) = Null(\bar{L}^T) = \{\mathbf{1} \otimes \alpha : \alpha \in \mathbb{R}^m\}$, and

(4) L and $\bar{L}$ share the same spectrum.

Proof. (1) Since L is symmetric, all of its eigenvalues are real. According to [57, Theorem 10], they must be nonnegative real numbers.

(2) From [57, Theorem 9], the zero eigenvalue of L is simple, and thus the right and left null spaces are one-dimensional. Observing that the column sums and row sums of L are all zeros, we found that $\mathrm{Null}(L) = \mathrm{Null}(L^T) = \mathrm{span}(\mathbf{1})$.

(3) Since $\bar{L}(\mathbf{1} \otimes \alpha) = L\mathbf{1} \otimes \alpha = 0$ and $\bar{L}^T(\mathbf{1} \otimes \alpha) = \bar{L}(\mathbf{1} \otimes \alpha) = 0$, we know that

$$S \subset \mathrm{Null}(\bar{L}) = \mathrm{Null}(\bar{L}^T), \tag{4.5}$$

where $S = \{\mathbf{1} \otimes \alpha : \alpha \in \mathbb{R}^m\}$, and the equation holds as $\bar{L}$ is symmetric. Noting that

$$\dim(S) = m, \tag{4.6}$$

we know that $\dim(\text{Null}(\bar{L})) = \dim(\text{Null}(\bar{L}^T)) \geq m$.

By [40, Theorem 9], we have that

$$|\bar{L}| = |L|^m |I_m|^n = |L|^m. \tag{4.7}$$

By Lemma 3 (2), the algebraic multiplicity of $\bar{L}$ associated with zero eigenvalue is m, which implies that its geometric multiplicity is no greater than m, i.e.,

$$\dim(\text{Null}(\bar{L})) = \dim(\text{Null}(\bar{L}^T)) \leq m. \tag{4.8}$$

Therefore, we know that

$$\dim(\text{Null}(\bar{L})) = \dim(\text{Null}(\bar{L}^T)) = \dim(S), \tag{4.9}$$

and

$$\text{Null}(\bar{L}) = \text{Null}(\bar{L}^T) = S. \tag{4.10}$$

(4) Again, we have that

$$|\bar{L} - \lambda I_{nm}| = |L - \lambda I_n|^m |I_m|^n = |L - \lambda I_n|^m \tag{4.11}$$

by [40, Theorem 9]. □ □

Under Assumptions 9 and 10, the following lemma shows that (4.2)) and (4.4)) have the same convergence rate.

Lemma 4. *Assume Assumptions 9 and 10 hold. If the origin of* (4.4) *is exponentially stable, i.e.,*

$$\left\|(q^T, l^T)^T\right\| \leq c_1 \left\|(q^T(0), l^T(0))^T\right\| e^{-\mu t}, \tag{4.12}$$

for some $c_1 \geq 1$, all $t \geq 0$, and all $(q^T(0), l^T(0))^T \in \mathbb{R}^{2nm}$, then the virtual state x in algorithm (4.2) *converges to $\mathbf{1} \otimes x^*$ globally exponentially with rate no less than μ, i.e.,*

$$\|x - \mathbf{1} \otimes x^*\| \leq c_2 \left\|(q^T(0), l^T(0))^T\right\| e^{-\mu t} \tag{4.13}$$

for some $c_2 \geq 0$ and all $t \geq 0$, where c_2 relies on the initial value of (4.2)*, and x^* is the optimal solution of problem* (1.6).

Proof. Denote $Null^{\perp}(\bar{L})$ by the orthogonal complement of $Null(\bar{L})$. For any $z \in \mathbb{R}^{nm}$, there exist

unique vectors $z_p \in Null^{\perp}(\bar{L})$ and $z_o \in Null(\bar{L})$ such that $z = z_p + z_o$ as $Null^{\perp}(\bar{L}) \oplus Null(\bar{L}) = \mathbb{R}^{nm}$.

Denote β_2 by $\lambda_2(L)$, which equals to $\lambda_{m+1}(\bar{L})$ by Lemma 3. It is also the smallest positive eigenvalue of L and $\bar{L}$. Thus, $(\beta_2)^2 = \lambda_2(L^2) = \lambda_{m+1}(\bar{L}^2)$ as L and $\bar{L}$ are symmetric and positive semi-definite. By Courant–Fischer Theorem from [33], we know that

$$\begin{aligned}
\beta_2^2 &= \max_{\{S\subset R^{nm}:\dim S=mn-m\}} \min_{\{z:0\neq z\in S\}} \frac{z^T\bar{L}^2 z}{z^T z} \\
&= \max_{\{S\subset R^{nm}:\dim S=mn-m\}} \min_{\{z:0\neq z\in S\}} \frac{(z_p+z_o)^T\bar{L}^2(z_p+z_o)}{(z_p+z_o)^T(z_p+z_o)} \\
&\leq \max_{\{S\subset R^{nm}:\dim S=mn-m\}} \min_{\{z:0\neq z\in S\}} \frac{(z_p)^T\bar{L}^2 z_p}{(z_p)^T z_p} \\
&= \min_{\{z:0\neq z\in Null^{\perp}(\bar{L})\}} \frac{z^T\bar{L}^2 z}{z^T z} = \min_{\{z:0\neq z\in Null^{\perp}(\bar{L})\}} \frac{\|\bar{L}z\|^2}{\|z\|^2}.
\end{aligned} \tag{4.14}$$

We can conclude that x approaches the optimal solution exponentially:

$$\begin{aligned}
\|x-\mathbf{1}\otimes x^*\| &\leq \|\delta_x\| + \|\mathbf{1}\otimes(\bar{x}-x^*)\| \\
&\overset{(4.14)}{\leq} \tfrac{1}{\beta\beta_2}\|l\| + \sqrt{n\textstyle\sum_i(\bar{x}-x^*)_i^2} \\
&\leq \tfrac{1}{\beta\beta_2}\|l\| + \sqrt{n}\|\bar{x}-x^*\| \\
&\leq \tfrac{1}{\beta\beta_2}\|l\| + \tfrac{\sqrt{n}}{\mu}\|\nabla f(\bar{x})\| \\
&\leq \tfrac{1}{\beta\beta_2}\|l\| + \tfrac{\sqrt{n}}{\mu}\left(\left\|\nabla f(\bar{x}) - \tfrac{1}{n}\textstyle\sum_i \nabla f_i(x^i)\right\| + \left\|\tfrac{1}{n}\textstyle\sum_i \nabla f_i(x^i)\right\|\right) \\
&\overset{(a)}{\leq} \tfrac{1}{\beta\beta_2}\|l\| + \tfrac{\sqrt{n}M}{\mu}\|\delta_x\| + \tfrac{1}{n}\left\|(\mathbf{1}\otimes I_m)^T q\right\| \\
&\leq \tfrac{1}{\beta\beta_2}\|l\| + \tfrac{\sqrt{n}M}{\mu\beta\beta_2}\|l\| + \tfrac{1}{\sqrt{n}}\|q\| \\
&\leq c_2\left\|(q^T(0), l^T(0))^T\right\| e^{-\mu t}\ (t\geq 0),,
\end{aligned} \tag{4.15}$$

where $c_2 = c_1(\frac{1}{\beta\beta_2} + \frac{\sqrt{n}M}{\mu\beta\beta_2} + \frac{1}{\sqrt{n}})$, $\bar{x} = \frac{1}{n}\sum_i x^i$, $\delta_x = x - \mathbf{1}\otimes\bar{x} \in \mathbb{R}^{mn}$ is the disagreement vector of x, and (a) is implied by

$$\left\|\textstyle\sum_i \nabla f_i(x^i)\right\| = \left\|(\mathbf{1}\otimes I_m)^T\nabla(x)\right\| = \left\|(\mathbf{1}\otimes I_m)^T q\right\|. \tag{4.16}$$

□

The following is devoted to proving our first main result.

Theorem 9. *Suppose that Assumptions 9 and 10 hold, β satisfies*

$$\beta\beta_2 > \frac{\mu+M}{2}, \tag{4.17}$$

and there exist $D_i > 0$ such that

$$D_iB_i + (B_i)^T D_i - I - \frac{(M-\mu)^2}{4}(D_i)^2 \geq 0,\ 2 \leq i \leq n, \tag{4.18}$$

where $B_i = \begin{pmatrix} \beta\beta_i + \frac{M-\mu}{2} & \beta\beta_i \\ -\beta\beta_i & -\mu \end{pmatrix}$ and β_i is the i-th largest eigenvalue of L. Let x^* be the optimal solution of the problem (1.6). Then, the virtual state x in algorithm (4.2) converges to $\mathbf{1}\otimes x^*$ globally exponentially with rate no less than μ, i.e.,

$$\|x - \mathbf{1}\otimes x^*\| \leq c\left\|(q^T(0), l^T(0))^T\right\| e^{-\mu t}, \tag{4.19}$$

for all $t \geq 0$, $(x^T(0), v^T(0))^T \in \mathbb{R}^{2nm}$, where $l = \beta\bar{L}x$, $q = -\beta\bar{L}v - \beta\bar{L}x - \nabla(x)$, and $c > 0$ is a constant.

Proof. **Step 1.** Before analyzing the convergence behavior of system (4.4), which is a nonconstant coefficient linear dynamical system, we will first consider the following linear dynamical comparison system with constant coefficient:

$$\begin{pmatrix} \dot{q} \\ \dot{l} \end{pmatrix} = \begin{pmatrix} -\beta\bar{L} - \frac{\mu+M}{2} I_{mn} & -\beta\bar{L} \\ \beta\bar{L} & 0 \end{pmatrix} \begin{pmatrix} q \\ l \end{pmatrix} = -B \otimes I_m \begin{pmatrix} q \\ l \end{pmatrix}, \tag{4.20}$$

where $B = \begin{pmatrix} \beta L + \frac{\mu+M}{2} I_n & \beta L \\ -\beta L & 0 \end{pmatrix}$.

We shall determine the locations of the eigenvalues of $B\otimes I_m$ by analyzing the determinants of $B - \lambda I_m$ and $B\otimes I_m - \lambda I_n \otimes I_m$. Let

$$\begin{aligned} 0 &= \begin{vmatrix} \beta L + \frac{\mu+M}{2} I_n - \lambda I_n & \beta L \\ -\beta L & -\lambda I_n \end{vmatrix} \\ &\overset{(a)}{=} \left|\lambda^2 I_n - \lambda(\tfrac{\mu+M}{2} I_n + \beta L) + \beta^2 L^2\right| \\ &= \left|(S_L)^T(\lambda^2 I_n - \lambda(\tfrac{\mu+M}{2} I_n + \beta D_L) + \beta^2 (D_L)^2) S_L\right| \\ &= \left|(S_L)^T\right| \left|\lambda^2 - \lambda(\tfrac{\mu+M}{2} + \beta D_L) + \beta^2 (D_L)^2)\right| \left|S_L\right| \\ &= \Pi_{i=1}^n (\lambda^2 - \lambda(\tfrac{\mu+M}{2} + \beta\beta_i) + \beta^2(\beta_i)^2) \end{aligned} \tag{4.21}$$

where β_i $(1 \leq i \leq n)$ is the i-th largest eigenvalues of L, and the eigen-decomposition $L = (S_L)^T D_L S_L$ holds for some diagonal matrix D_L and orthogonal matrix S_L. The equation (a) follows the property

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |AD - BC| \text{ if } CD = DC. \tag{4.22}$$

from [71].

According Lemma 3, it holds that $0 = \beta_1 < \beta_2 \leq ... \leq \beta_n$. Assuming that β satisfies the conditions (4.17), We can derive that the eigenvalues of B are 0, $\frac{\mu+M}{2}$, and $\frac{\mu+M}{4} + \frac{\beta\beta_i}{2} + i\sqrt{(\beta\beta_i)^2 - (\frac{\mu+M}{4} + \frac{\beta\beta_i}{2})^2}$ $(i = 2, ..., n)$.

According to the results in [40, Theorem 9], we can obtain

$$\begin{aligned} 0 &= \left|B \otimes I_m - \lambda I_n \otimes I_m\right| \\ &= \left|(B - \lambda I_n) \otimes I_m\right| \\ &= \left|B - \lambda I_n\right|^m \left|I_m\right|^{2mn} \\ &= \left|B - \lambda I_n\right|^m. \end{aligned} \tag{4.23}$$

Therefore, the eigenvalues of $B \otimes I$ are also 0, $\frac{\mu+M}{2}$, and $\frac{\mu+M}{4} + \frac{\beta\beta_i}{2} + i\sqrt{(\beta\beta_i)^2 - (\frac{\mu+M}{4} + \frac{\beta\beta_i}{2})^2}$ $(i = 2, ..., n)$, and the algebraic multiplicities of eigenvalues 0 and $\frac{\mu+M}{2}$ are m.

In the following, we will rewrite the system (4.20) as an equivalent linear system with Hurwitz coefficient, of which the eigenvalues are the same as $-B \otimes I_m$ except for 0.

Define the matrix $P \in \mathbb{O}^{n \times n}$ and $S = \text{diag}\left(P, P\right) \in \mathbb{O}^{2n \times 2n}$, which is subject to

$$P^T L P = \text{diag}\begin{pmatrix} 0 & \beta_2 & \cdots & \beta_n \end{pmatrix}. \tag{4.24}$$

Therefore, the matrix B can be decoupled into

$$J^* = S^T B S = \left(\begin{array}{ccc|ccc} \frac{\mu+M}{2} & & & 0 & & \\ & \ddots & & & \ddots & \\ & & \beta\beta_n + \frac{\mu+M}{2} & & & \beta\beta_n \\ \hline & 0 & & & & \\ & & \ddots & & & \\ & & -\beta\beta_n & & & \end{array}\right) \tag{4.25}$$

Denote by a new variable

$$r^* = (S \otimes I) \begin{pmatrix} q \\ l \end{pmatrix}, \tag{4.26}$$

then the system (4.20) is equivalent to

$$\dot{r}^* = (J^* \otimes I)\, r^*. \tag{4.27}$$

By Lemma 3 (2), we know that the first column of P is $\frac{1}{\sqrt{n}}\mathbf{1}$, therefore, we know that

$$(r^*)^{n+1} = \left(\mathbf{1}^T \otimes I_m\right) l = \beta(\mathbf{1}^T L \otimes I_m)x = 0. \tag{4.28}$$

Equations (4.25) and (4.28) implies that $(r^*)^{n+1}$ and the $(n+1)$-th row and column of J^* are trivial. Discarding these trivial entries, we obtain a reduced system of (4.20)

$$\dot{r} = -(J_1 \otimes I_m)r. \tag{4.29}$$

In other words, J_1 is obtained by deleting the $(n+1)$-th column and row from J^*, and r is obtained by deleting $(r^*)^{n+1}$ from r^*.

Considering that

$$\begin{aligned} \text{Null}(B \otimes I) &\subset \text{Null}(A(x)), \\ \text{Null}((B \otimes I)^T) &\subset \text{Null}(A(x)^T), \end{aligned} \tag{4.30}$$

for any $x \in \mathbb{R}^{nm}$, the $(mn+1)$-th to $(mn+m)$-th columns and rows of the block matrix $(S \otimes I)^T A(x)(S \otimes I)$ are all trivial. By conducting the same transformation as we did on (4.20), the system (4.4) can be reduced to

$$\dot{r} = -J_2(x)r. \tag{4.31}$$

Step 2. Now we will prove that there exists a constant positive definite matrix P_1 such that

$$0 = P_1(J_1 - \mu I) + (J_1 - \mu I)^T P_1 - I - \frac{(M-\mu)^2}{4}(P_1)^2, \tag{4.32}$$

where $I = I_{m(n-1)}$ for short.

Noticing that $-(J_1 - \mu I)$ is a block Hurwitz matrix, the solution to the continuous-time Lyapunov equation $Q_1(J_1 - \mu I) + (J_1 - \mu I)^T Q_1 = I$ is a unique positive definite block matrix

$Q_1 = \text{diag}(\frac{1}{M-\mu}I, Q_{11})$, which admits

$$Q_1(J_1 - \mu I) + (J_1 - \mu I)Q_1 - I - \frac{(M-\mu)^2}{4}(Q_1)^2 \leq 0. \tag{4.33}$$

Let $D_i = \begin{pmatrix} (D_i)_{11} & (D_i)_{12} \\ (D_i)_{21} & (D_i)_2 2 \end{pmatrix}$ and

$$Q_2 = \left(\begin{array}{c|ccc|ccc} \frac{2}{M-\mu} & & & & & & \\ \hline & (D_2)_{11} & & & (D_2)_{12} & & \\ & & \ddots & & & \ddots & \\ & & & (D_n)_{11} & & & (D_n)_{12} \\ \hline & (D_2)_{21} & & & (D_2)_{22} & & \\ & & \ddots & & & \ddots & \\ & & & (D_n)_{21} & & & (D_n)_{22} \end{array}\right). \tag{4.34}$$

For any $h, g \in \mathbb{R}^{m(n-1)}$ and any $f \in \mathbb{R}$, we have that

$$\begin{array}{rl} & \begin{pmatrix} f & h^T & g^T \end{pmatrix} Q_2 \begin{pmatrix} f \\ h \\ g \end{pmatrix} \\ = & \frac{2}{M-\mu} f^2 + \sum_{i=1}^{n-1} \begin{pmatrix} h_i & g_i \end{pmatrix} D_{i+1} \begin{pmatrix} h_i \\ g_i \end{pmatrix} \geq 0, \end{array} \tag{4.35}$$

where the equality occurs only if f, h, g are trivial. Therefore, Q_2 is positive definite.

For any $h, g \in \mathbb{R}^{m(n-1)}$ and any $f \in \mathbb{R}$, define $fhg = \begin{pmatrix} f & h^T & g^T \end{pmatrix}^T$, and we have that

$$\begin{array}{rl} & fhg^T \left(Q_2(J_1 - \mu I) + (J_1 - \mu I)Q_2 - I - \frac{(M-\mu)^2}{4}(Q_2)^2 \right) fhg \\ = & \sum_{i=2}^{n} \begin{pmatrix} h_i & g_i \end{pmatrix} \left(D_i B_i + (B_i)^T D_i - I - \frac{(M-\mu)^2}{4}(D_i)^2 \right) \begin{pmatrix} h_i \\ g_i \end{pmatrix} \\ \geq & 0, \end{array} \tag{4.36}$$

which implies that

$$Q_2(J_1 - \mu I) + (J_1 - \mu I)Q_2 - I - \frac{(M-\mu)^2}{4}(Q_2)^2 \geq 0 \tag{4.37}$$

According to [1, Theorem 4.1.14], equation (4.32) has at least one positive definite Hermitian solution.

Step 3. This step will focus on analyzing the convergence behavior of system (4.31), which will be compared with (4.29).

Take a Lyapunov function $V = r^T P_2(t) r$, where $P_2(0) = P_1$ and P_2 updates according to

$$\dot{P}_2 = \quad P_2(J_2 - \mu I) + (J_2 - \mu I)^T P_2, \tag{4.38}$$

then, the derivative of the function V along system (4.4) is given as

$$\begin{aligned} \dot{V} &= r^T(-P_2 J_2 - J_2^T P_2 + \dot{P}_2) r \\ &= -2\mu V, \end{aligned} \tag{4.39}$$

which gives $V(t) = V(0)e^{-2\mu t}$.

In view of the facts that

$$\frac{\mu - M}{2} I_{2mn} \leq A(x) - B \otimes I_m \leq \frac{M - \mu}{2} I_{2mn}, \tag{4.40}$$

it can be recognized that

$$0 \leq ((S \otimes I_m)^T (A(x) - B \otimes I_m)(S \otimes I_m))^2 \leq \frac{(M-\mu)^2}{4} I_{2mn}. \tag{4.41}$$

Define $\Delta J = J_2 - J_1 \otimes I_m$. By step 1, it can be checked that ΔJ and $(\Delta J)^2$ are symmetric diagonal block matrices, and they are minors of $(S \otimes I_m)^T (A(x) - B \otimes I_m)(S \otimes I_m)$ and $(S \otimes I_m)^T (A(x) - B \otimes I_m)^2 (S \otimes I_m)$. Equation (4.41) gives that

$$(\Delta J)^2 \leq \frac{(M-\mu)^2}{4} I_{m(2n-1)}, \tag{4.42}$$

and we can derive that

$$\begin{aligned} &\begin{pmatrix} I & \Delta J \\ \Delta J & \frac{(M-\mu)^2}{4} I \end{pmatrix} = \begin{pmatrix} I & \frac{4}{(M-\mu)^2} \Delta J \\ 0 & I \end{pmatrix} \\ &\begin{pmatrix} I - \frac{4}{(M-\mu)^2} (\Delta J)^2 & 0 \\ 0 & \frac{(M-\mu)^2}{4} I \end{pmatrix} \begin{pmatrix} I & \frac{4}{(M-\mu)^2} \Delta J \\ 0 & I \end{pmatrix}^T \geq 0. \end{aligned} \tag{4.43}$$

According to [1, Theorem 4.1.4], the inequality

$$\begin{aligned} (JH)_1 &= \begin{pmatrix} I & -(J_1-\mu I)^T \\ -(J_1-\mu I) & \frac{(M-\mu)^2}{4} I \end{pmatrix} \\ \geq \quad (JH)_2 &= \begin{pmatrix} 0 & -(J_2-\mu I)^T \\ -(J_2-\mu I) & 0 \end{pmatrix}, \end{aligned} \tag{4.44}$$

which is directly derived from (4.43), implies that $P_2(t) \geq P_1$ on $[0,\infty)$. Therefore, we have that

$$\begin{aligned} & \lambda_{\min}(P_1)\|r\|^2 \leq r^T P_1 r \\ \leq \quad & r^T P_2(t) r = V(t) \leq V(0)e^{-2\mu t} \\ \leq \quad & \lambda_{\max}(P_1)\|r(0)\|^2 e^{-2\mu t} = \lambda_{\max}(P_1)\|r^*(0)\|^2 e^{-2\mu t} \\ \leq \quad & \lambda_{\max}(P_1)\|(S\otimes I)\|^2 \left\|(q^T(0), l^T(0))^T\right\|^2 e^{-2\mu t} \\ \leq \quad & \lambda_{\max}(P_1)\left\|(q^T(0), l^T(0))^T\right\|^2 e^{-2\mu t}, \end{aligned} \tag{4.45}$$

and thus

$$\begin{aligned} & \left\|(q^T\, l^T)^T\right\| \leq \|S\otimes I\|\|r^*\| \\ = \quad & \|S\otimes I\|\|r\| = \|r\| \\ \leq \quad & C_1\left\|(q^T(0), l^T(0))^T\right\| e^{-\mu t}, \end{aligned} \tag{4.46}$$

for all $t \geq 0$, where $C_1 = \sqrt{\lambda_{\max}(P_1)/\lambda_{\min}(P_1)}$.

Combining the above inequality with Lemma 2, we complete the proof. □

Remark 8. *A sufficient condition for* (4.18) *is that* β *is sufficiently large. Since we assume that* $\beta\beta_i > \frac{M+\mu}{2}$, $2 \leq i \leq n$, $-D_i$ *is a Hurwitz matrix. For each* $2 \leq i \leq n$, *the continuous-time Lyapunov equation* $d_i B_i + (B_i)^T d_i = I$ *has a unique solution. Since* $\frac{M}{4} - \frac{3\mu}{4} + \frac{\beta\beta_i}{2} > 0$ $(i = 2,\ldots,n)$ *are the real parts of the eigenvalues of* B_i, *every entry of* $e^{-B_i t}$ *is bounded by a linear combination of* $t^k e^{-(\frac{M}{4}-\frac{3\mu}{4}+\frac{\beta\beta_i}{2})t}$, *where* k *is some nonnegative integer. Due to the fact that* $\int_0^\infty t^k e^{-at}dt = \frac{k!}{a^{k+1}}$ *for all nonnegative integer* k *and* $a > 0$, *the positive definite matrix*

$$d_i = \int_0^\infty e^{-(B_i)^T t} e^{-B_i t} dt \tag{4.47}$$

and its eigenvalues would vanish as β *goes to infinity. When* $\beta > 0$ *is sufficiently large such that*

$$\lambda_{\max}(d_i) \leq \frac{1}{M-\mu}, \tag{4.48}$$

we can find a number $c = \frac{2}{(M-\mu)^2\lambda_{\max}^2(D_i)} > 0$ *such that*

$$-c^2\frac{(M-\mu)^2}{4}\lambda_{\max}^2(d_i) + c - 1 \geq 0. \tag{4.49}$$

Letting $D_i = cd_i$*, we know that*

$$\begin{array}{rl} & D_iB_i + (B_i)^T D_i - I_2 - \frac{(M-\mu)^2}{4}(D_i)^2 \\ \geq & -c^2\frac{(M-\mu)^2}{4}\lambda_{\max}^2(d_i)I_2 + cI_2 - I_2 \geq 0. \end{array} \tag{4.50}$$

Remark 9. *In [98], a distributed continuous-time algorithm that achieves the same convergence rate of the centralized gradient descent method was proposed. The iteration of the algorithm begins with an arbitrary* $x(0)$ *and* $s_i = \nabla f_i(x^i(0))$ *and has the following update rule*

$$\begin{array}{l} \dot{x} = -\beta\bar{L}x - s, \\ \dot{s} = -\beta\bar{L}s + \nabla^2(x)\dot{x} = -\beta\bar{L}s + \nabla^2(x)(-\beta\bar{L}x - s), \end{array} \tag{4.51}$$

where $\nabla^2(x) = diag(\nabla^2 f_1(x_1), ..., \nabla^2 f_n(x_n))$ *is a term involving the Hessian matrices of the local objective functions. The proposed algorithm improves upon the algorithm in [98] by removing the dependence of these Hessian matrices. It is interesting to note that the proof technique in this section can also be used to analyze algorithm (4.51) above. To see this, define* $l = \beta\bar{L}x$*,* $q = \dot{x}$*. Then the above system can be reduced to*

$$\begin{pmatrix} \dot{q} \\ \dot{l} \end{pmatrix} = \begin{pmatrix} -2\beta\bar{L} - \nabla^2(x) & -\beta\bar{L} \\ \beta\bar{L} & 0 \end{pmatrix} \begin{pmatrix} q \\ l \end{pmatrix}. \tag{4.52}$$

Similar to the proof of Theorem 9, it can be verified that the state x in (4.51) converges exponentially with rate no less than μ *if* β *is sufficiently large.*

4.3.2 Convergence Analysis under Strictly Convex Cost Function

With a weaker assumption, we provide a locally exponential convergence result for the algorithm (4.2). As in the previous section, we will first construct the relationship between (4.4) and (4.2), analyze the auxiliary system (4.4), and conclude the local exponential convergence of the system (4.2). Additionally, we will discuss the global asymptotic convergence of the system (4.2) as a separate part.

The following lemma shows that the system (4.2) converges as (4.4) converges and that they have the same local convergence behavior, similar to Lemma 4.

Proposition 3. *Assume that Assumptions 10 and 11 hold. The states q and l were defined in the last section. Consider a trajectory of (4.2) starting from $(x^T(0), v^T(0))^T$. Let x^* be the optimal solution to (1.6).*

(1) The state $(q^T, l^T)^T$ in (4.4) tends to zero if and only if $(x^T, v^T)^T$ in (4.2) tends to $((\mathbf{1}\otimes x^*)^T, (v^*)^T)^T$, i.e.,

$$\lim_{\substack{x\to \mathbf{1}\otimes x^* \\ v\to v^*}} \begin{pmatrix} q \\ l \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{4.53}$$

and

$$\lim_{\substack{q\to 0 \\ l\to 0}} \begin{pmatrix} x \\ v \end{pmatrix} = \begin{pmatrix} \mathbf{1}\otimes x^* \\ v^* \end{pmatrix}, \tag{4.54}$$

where v^* is uniquely determined by $\beta\bar{L}v^* = -\nabla(\mathbf{1}\otimes x^*)$ and $\sum_i (v^*)_i = \sum_i v^i(0)$.

(2) Suppose that the origin of (4.4) is locally exponentially stable, i.e., there exists $r_1 > 0$ such that

$$\left\|(q^T, l^T)^T\right\| \le c_1 \left\|(q^T(0), l^T(0))^T\right\| e^{-\mu t}, \tag{4.55}$$

for some $c_1 \ge 1$, all $t \ge 0$, and all $(q^T(0), l^T(0))^T \in B(0, r_1)$. If there exists $r_4 > 0$ such that $(x^T(0), v^T(0))^T$ lies in $B(((\mathbf{1}\otimes x^*)^T, (v^*)^T)^T, r_4)$, then,

$$\begin{array}{rl} & \left\|((x-\mathbf{1}\otimes x^*)^T, (v-v^*)^T)^T\right\| \\ \le & c_3 \left\|(q^T(0), l^T(0))^T\right\| e^{-\mu t}, \end{array} \tag{4.56}$$

for all $t \ge 0$, where $c_3 > 0$ relies on $x(0)$ and $v(0)$.

Proof. (1) ($\Longrightarrow$) Assume that $q, l \to 0$ holds throughout the $\Longrightarrow$ part.

Step 1. Recall that

$$\begin{array}{l} q = -v - \beta\bar{L}x - \nabla(x), \\ l = \beta\bar{L}x. \end{array} \tag{4.57}$$

It immediately follows from (4.15) and (4.16) that

$$\textstyle\sum_i \nabla f_i(x^i) \to 0 \text{ and } x \to \mathbf{1}\otimes\bar{x}, \tag{4.58}$$

which imply that $\nabla f_i(x^i) \to \nabla f_i(\bar{x})$ since ∇f_i ($i \in X$) are continuous globally. That is to say, it can be derived that

$$\textstyle\nabla f(\bar{x}) = \sum_i \nabla f_i(\bar{x}) \to 0 \text{ and } x \to \mathbf{1}\otimes\bar{x}. \tag{4.59}$$

In view of the continuity of ∇f and the strict convexity of f, we know that $\bar{x} \to x^*$ and $x \to \mathbf{1}\otimes x^*$.

Step 2. Next, it will be proved that there exists a unique v^* satisfying

$$\beta\bar{L}v^* = -\nabla(\mathbf{1}\otimes x^*) \text{ and } \sum_i (v^*)_i = \sum_i v^i(0). \tag{4.60}$$

The complete solution to $\beta\bar{L}v^* = -\nabla(\mathbf{1}\otimes x^*)$ is

$$v^* = v_s + v_g, \tag{4.61}$$

where v_s is the special solution and $v_g \in \text{Null}(\bar{L})$. Take $v_g = \mathbf{1}\otimes(\frac{1}{n}\sum_i v^i(0) - \frac{1}{n}\sum_i (v_s)^i) \in \text{Null}(\bar{L})$, then

$$\textstyle\sum_i (v^*)^i = \sum_i (v_s)^i + (\sum_i v^i(0) - \sum_i (v_s)^i) = \sum_i v^i(0). \tag{4.62}$$

We will prove the uniqueness of v^*. Assume that there are $v^* \neq w^*$ that satisfy condition (4.60). Define $\delta_v^* = v^* - w^*$, which satisfies $\sum_i (\delta_v^*)_i = 0$. We can derive that

$$\|\delta_v^*\| \overset{(4.14)}{\leq} \frac{1}{\beta\beta_2}\|\beta\bar{L}\delta_v^*\| = \frac{1}{\beta\beta_2}\|\beta\bar{L}v^* - \beta\bar{L}w^*\| = 0, \tag{4.63}$$

which contradicts our assumption.

Step 3. Define $\delta_v = v - v^*$, which satisfies $\sum_i (\delta_v)_i = 0$. We know that

$$\begin{aligned} & \|\delta_v\| \overset{(4.14)}{\leq} \tfrac{1}{\beta\beta_2}\|\beta\bar{L}\delta_v\| \\ = \; & \tfrac{1}{\beta\beta_2}\|\beta\bar{L}v + \nabla(\mathbf{1}\otimes x^*)\| \\ \leq \; & \tfrac{1}{\beta\beta_2}\|q+l\| + \tfrac{1}{\beta\beta_2}\|\nabla(x) - \nabla(\mathbf{1}\otimes x^*)\| \overset{(a)}{\to} 0, \end{aligned} \tag{4.64}$$

where (a) is due to the continuity of $\nabla(\cdot)$, $q, l \to 0$, and $x \to \mathbf{1}\otimes x^*$,

($\Longleftarrow$) Given that $v \to v^*$ as well as $x \to \mathbf{1}\otimes x^*$.

First, it can be checked that

$$\begin{aligned} & \|l\| = \|\beta\bar{L}x\| \\ = \; & \|\beta\bar{L}(x - \mathbf{1}\otimes x^*)\| \\ \leq \; & |\beta|\|\bar{L}\|\|x - \mathbf{1}\otimes x^*\| \to 0. \end{aligned} \tag{4.65}$$

By the continuity of ∇, we know that

$$\begin{array}{ll} & \|q+l\| = \|\beta\bar{L}v+\nabla(x)\| \\ \leq & \|\nabla(x)-\nabla(\mathbf{1}\otimes x^*)\|+\|\beta\bar{L}v-\beta\bar{L}v^*\| \\ \leq & \|\nabla(x)-\nabla(\mathbf{1}\otimes x^*)\|+\beta\|\bar{L}\|\|v-v^*\| \\ \to & 0+0=0. \end{array} \tag{4.66}$$

Thus,

$$\|q\| \leq \|q+l\|+\|l\| \to 0 \tag{4.67}$$

as $(x^T, v^T)^T \to ((\mathbf{1}\otimes x^*)^T, (v^*)^T)^T$.

(2) According to the part (1), there exists some $r_2 > 0$ such that $\|x-\mathbf{1}\otimes x^*\|^2+\|v-v^*\|^2 \leq (r_2)^2$ implies $\|q\|^2+\|l\|^2 \leq (r_1)^2$. There exists some $r_3 > 0$ such that $\|q\|^2+\|l\|^2 \leq (r_3)^2$ implies $\|x-\mathbf{1}\otimes x^*\|^2+\|v-v^*\|^2 \leq (r_2)^2$. There exists some $r_4 > 0$ such that $\|x-\mathbf{1}\otimes x^*\|^2+\|v-v^*\|^2 \leq (r_4)^2$ implies $\|q\|^2+\|l\|^2 \leq (r_3/c_1)^2$.

By our assumption, all positive trajectories of (4.4) starting from $(q^T(0), l^T(0)) \in B(0, r_3/c_1)$ lie in $B(0, r_3)$. Thus, all positive trajectories of (4.2) starting from $(x^T(0), v^T(0))^T \in B(((\mathbf{1}\otimes x^*)^T, (v^*)^T)^T, r_4)$ stay in $B(((\mathbf{1}\otimes x^*)^T, v^T)^T, r_2)$.

For all $\|x-\mathbf{1}\otimes x^*\|^2+\|v-v^*\|^2 \leq (r_2)^2$, the average $\bar{x} = \frac{1}{n}\sum_i x^i$ lies in the compact set $B(x^*, r_2)$. By Assumption 9, $\nabla^2 f > 0$ on the compact set $B(x^*, r_2)$. Therefore, there exists $\mu' > 0$ such that $\nabla^2 f(\bar{x}) > \mu' I$ for all $\bar{x} \in B(x^*, r_2)$. For all x in the compact set $B(\mathbf{1}\otimes x^*, r_2)$, there exists $M' > 0$ such that $\nabla^2(x) < M' I$, which implies that all gradients of the local objective functions ∇f_i ($i \in X$) are M'-Lipschitz on $B(\mathbf{1}\otimes x^*, r_2)$.

If c_2 in (4.15) is replaced by $c_2 = c_1(\frac{1}{\beta\beta_2}+\frac{\sqrt{n}M'}{\mu'\beta\beta_2}+\frac{1}{\sqrt{n}})$, (4.15) would also hold for $(x^T, v^T)^T \in B(((\mathbf{1}\otimes x^*)^T, (v^*)^T)^T, r_2)$ here.

Denote by the disagreement vector $\delta_v = v - v^*$. Since

$$\begin{array}{ll} \sum_i \dot{v}^i & = (\mathbf{1}\otimes I_m)^T\dot{v} \\ & = (\mathbf{1}\otimes I_m)^T(\beta L\otimes I_m)x \\ & = \beta(\mathbf{1}^T L\otimes I_m)x = 0, \end{array} \tag{4.68}$$

we can derive that $\sum_i (\delta_v)^i = 0$. It can be derived that

$$\begin{aligned} & \|\delta_v\| \overset{(4.14)}{\leq} \tfrac{1}{\beta_2}\|\bar{L}\delta_v\| \\ = \; & \tfrac{1}{\beta\beta_2}\|\beta\bar{L}v + \nabla(\mathbf{1}\otimes x^*)\| \\ \leq \; & \tfrac{1}{\beta\beta_2}\|\beta\bar{L}v + \nabla(x)\| + \tfrac{1}{\beta\beta_2}\|\nabla(x) - \nabla(\mathbf{1}\otimes x^*)\| \\ \leq \; & \tfrac{1}{\beta\beta_2}\|q + l\| + \tfrac{M'}{\beta\beta_2}\|x - \mathbf{1}\otimes x^*\| \\ \leq \; & \tfrac{1}{\beta\beta_2}\|q\| + \tfrac{1}{\beta\beta_2}\|l\| + \tfrac{M'}{\beta\beta_2}c_2\big\|(q^T(0), l^T(0))^T\big\|e^{-\mu t} \\ \leq \; & (\tfrac{2c_1}{\beta\beta_2} + \tfrac{M'}{\beta\beta_2}c_2)\big\|(q^T(0), l^T(0))^T\big\|e^{-\mu t} \; (t \geq 0). \end{aligned} \tag{4.69}$$

In conclusion, it is verified that

$$\begin{aligned} & \big\|((x - \mathbf{1}\otimes x^*)^T, (v - v^*)^T)^T\big\| \\ \leq \; & \sqrt{(c_2)^2 + (\tfrac{2c_1}{\beta\beta_2} + \tfrac{M'}{\beta\beta_2}c_2)^2}\big\|(q^T(0), l^T(0))^T\big\|e^{-\mu t} \; (t \geq 0) \end{aligned} \tag{4.70}$$

holds for all trajectories starting from $B(((\mathbf{1}\otimes x^*)^T, (v^*)^T)^T, r_4)$. □

Lemma 5. *Assume that Assumptions 10 and 11 hold. For all $z_i \in \mathbb{R}^m$ $(i \in X)$ and $\beta > 0$, the symmetric matrix $\hat{L} = \beta\bar{L} + \nabla^2(z)$ is positive definite if $\nabla^2(z) \neq 0$, where $z = (z_1^T, ..., z_n^T)^T$.*

Proof. Let $v = (v_1^T, ..., v_n^T)^T$ be a real vector in $\mathbb{R}^{mn}$, where $v^i \in \mathbb{R}^m$ $(i = 1, ..., n)$. Since

$$v^T\hat{L}v = v^T\beta\bar{L}v + v^T\nabla^2(z)v \geq 0 \tag{4.71}$$

holds for any $z \in \mathbb{R}^m$, the matrix $\hat{L}$ is positive semi-definite.

Assuming that $v^T\hat{L}v = 0$, we have

$$\begin{aligned} & v^T\beta\bar{L}v = 0 \\ & (v^i)^T\nabla^2 f_i(z_i)v^i = 0, \; (i = 1, ..., n). \end{aligned} \tag{4.72}$$

Assumption 9 implies that there exists some $j \in X$ such that $\nabla^2 f_j(z_j) > 0$, therefore, the second line of (4.72) indicates that $v_j = 0$. From Lemma 3, the first line of (4.72) indicates that all v^i are the same vector. We can conclude that $v^T\hat{L}v = 0$ occurs only when $v = 0$. □

Lemma 6. *Let Assumption 2 holds. If a variable is of the form $z_1 = \bar{L}z_2 \in \mathbb{R}^{nm}$, where $z_2 \in \mathbb{R}^{nm}$, then the following are equivalent:*

(1) z_1 reaches consensus.

(2) z_2 reaches consensus.

(3) z_1 is zero.

Proof. (1) $\Longrightarrow$ (2): Let z_1 reach consensus. Then, there exists $z_3 \in \mathbb{R}^m$ such that $z_1 = \bar{L}z_2 = \mathbf{1} \otimes z_3$. Therefore, we obtain

$$\begin{aligned} 0 &= (\mathbf{1}^T \otimes I_m)(L \otimes I_m)z_2 \\ &= (\mathbf{1}^T \otimes I_m)(\mathbf{1} \otimes z_3) \\ &= nz_3, \end{aligned} \tag{4.73}$$

which implies $z_1 = \bar{L}z_2 = 0$, and thus z_2 reaches consensus by Lemma 3 (3).

(2) $\Longrightarrow$ (3) $\Longrightarrow$ (1): Assume that z_2 reaches consensus. By Lemma 3 (3), we know that $z_1 = 0$, and it is trivial that z_1 reaches consensus as well. □

The following is devoted to proving our second main result.

Theorem 10. *Assume that Assumptions 10 and 11 hold. Consider a trajectory of (4.2) starting from $(x^T(0), v^T(0))^T \in \mathbb{R}^{2mn}$. Let x^* be the optimal solution to (1.6).*

(1) The trajectory converges to some equilibrium $((\mathbf{1} \otimes x^*)^T, (v^*)^T)^T$ globally asymptotically, i.e.,

$$\lim_{t\to\infty} \begin{pmatrix} x \\ v \end{pmatrix} = \begin{pmatrix} \mathbf{1} \otimes x^* \\ v^* \end{pmatrix}. \tag{4.74}$$

(2) The trajectory converges to some equilibrium $((\mathbf{1} \otimes x^*)^T, (v^*)^T)^T$ locally exponentially with rate no less than $\mu = J - \varepsilon$, where J is the smallest positive real part in the spectrum of

$$A(\mathbf{1} \otimes x^*) = \begin{pmatrix} \beta\bar{L} + \nabla^2(\mathbf{1} \otimes x^*) & -\beta\bar{L} \\ \beta\bar{L} & 0 \end{pmatrix}, \tag{4.75}$$

ε is an arbitrarily constant in $(0, J)$, and β is an arbitrary positive constant. That is to say, there exists some positive constant r such that

$$\left\|((x - \mathbf{1} \otimes x^*)^T, (v - v^*)^T)^T\right\| \le c\left\|(q^T(0), l^T(0))^T\right\| e^{-\mu t}, \tag{4.76}$$

for all $t \ge 0$, $(x^T(0), v^T(0))^T \in B(((\mathbf{1} \otimes x^*)^T, (v^*)^T)^T, r)$, where $l = \beta\bar{L}x$, $q = -\beta\bar{L}v - \beta\bar{L}x - \nabla(x)$, and $c > 0$ is a constant.

Proof. (1) Denote by $V = \frac{1}{2}\|q\|^2 + \frac{1}{2}\|l\|^2$ a Lyapunov function. After taking derivative, we obtain

that

$$\begin{aligned}\dot{V} &= q^T\dot{q} + l^T\dot{l} \\ &= \begin{pmatrix} q \\ l \end{pmatrix}^T \begin{pmatrix} -\nabla^2(x) - \beta\bar{L} & -\beta\bar{L} \\ \beta\bar{L} & 0 \end{pmatrix} \begin{pmatrix} q \\ l \end{pmatrix} \\ &= -q^T(\nabla^2(x) + \beta\bar{L})q \leq 0\end{aligned} \tag{4.77}$$

Now we will prove that $\nabla^2(x) \neq 0$ in $E = \{x : \dot{V} = 0\} \subset \{x : q \text{ reaches consensus}\}$. Consider a trajectory in E reaches a point x where $\nabla^2(x) = 0$. By Lemma 6, we know that the related q and $\dot{q} = -\beta\bar{L}l - \beta\bar{L}q = -\beta\bar{L}l$ reach consensus, which means that l has to reach consensus and $\dot{q} = 0$. Again, using Lemma 6, we know that $l = \beta\bar{L}x$ has to be zero and x reaches consensus, which contradicts with our assumption that $\nabla^2(x) = 0$.

By LaSalle's Invariance Principle, the solution will asymptotically approach the largest invariant set in $E = \{x : \dot{V} = 0\}$. Lemma 5 and Assumption 9 imply that $E = \{q = 0\}$. For any trajectory in E, $\dot{q} = -\beta\bar{L}l = 0$, which implies that l reaches consensus by Lemma 6. Again, using Lemma 6, we know that $l = \beta\bar{L}x$ is zero. Therefore, we can conclude that $E = \{q = l = 0\}$.

According to Proposition 3 (1), the state $(x^T, v^T)^T$ in the system (4.2) converges to $((\mathbf{1} \otimes x^*)^T, (v^*)^T)^T$ globally asymptotically.

(2) We now prove exponential convergence analysis for (4.2), which completes this proof.

Step 1. Similar to the above proof, we will first discuss the following comparison system:

$$\begin{pmatrix} \dot{q} \\ \dot{l} \end{pmatrix} = -\begin{pmatrix} \beta\bar{L} + \nabla^2(\mathbf{1} \otimes x^*) & -\beta\bar{L} \\ \beta\bar{L} & 0 \end{pmatrix} \begin{pmatrix} q \\ l \end{pmatrix} = -A(\mathbf{1} \otimes x^*) \begin{pmatrix} q \\ l \end{pmatrix}, \tag{4.78}$$

where x^* is the optimal solution of problem (1.6), and $A(\mathbf{1} \otimes x^*) = \begin{pmatrix} \beta\bar{L} + \nabla^2(\mathbf{1} \otimes x^*) & -\beta\bar{L} \\ \beta\bar{L} & 0 \end{pmatrix}$.

Part 1. In this part, we shall discuss the locations of the eigenvalues of $A(\mathbf{1} \otimes x^*)$.

Set a Lyapunov function $V = \frac{1}{2}(\|q\|^2 + \|l\|^2)$, then $\dot{V}$ along system (4.78) gives

$$\begin{aligned}\dot{V}(q,l) &= \dot{q}^T q + \dot{l}^T l \\ &= q^T(-\beta\bar{L} - \nabla^2(\mathbf{1} \otimes x^*))q \\ &\leq 0.\end{aligned} \tag{4.79}$$

Also, we know that the solution to (4.78) is

$$\begin{pmatrix} q \\ l \end{pmatrix} = e^{-A(\mathbf{1} \otimes x^*)t} \begin{pmatrix} q(0) \\ l(0) \end{pmatrix}, \tag{4.80}$$

and

$$V = \frac{1}{2}\left\| e^{-A(\mathbf{1}\otimes x^*)t}(q^T(0), l^T(0))^T \right\|^2 \leq \frac{1}{2}\left\|(q^T(0), l^T(0))^T\right\|^2. \tag{4.81}$$

Therefore, the fundamental matrix to (4.78) $e^{-A(\mathbf{1}\otimes x^*)t}$ has to be bounded on $[0,\infty)$, which further implies that all eigenvalues of $A(\mathbf{1}\otimes x^*)$ have non-negative real parts, and the algebraic multiplicities of the purely imaginary eigenvalues have to equal to their geometric multiplicities.

Next, we will investigate the purely imaginary eigenvalues of $A(\mathbf{1}\otimes x^*)$. Let variables $v_{11}, v_{12}, v_{21}, v_{22} \in \mathbb{R}^{nm}$, $v_1 = (v_{11}^T, v_{12}^T)^T$, $v_2 = (v_{21}^T, v_{22}^T)^T$, and $b \in \mathbb{R}$ such that

$$A(\mathbf{1}\otimes x^*)(v_1 + iv_2) = ib(v_1 + iv_2), \tag{4.82}$$

which implies

$$\begin{cases} A(\mathbf{1}\otimes x^*)v_1 = -bv_2, \\ A(\mathbf{1}\otimes x^*)v_2 = bv_1. \end{cases} \tag{4.83}$$

After pre-multiplying the first line by $(v_1)^T$, the second line by $(v_2)^T$, and adding them up, it can be derived that

$$\begin{aligned} 0 &= -b(v_1)^T v_2 + b(v_1)^T v_2 \\ &= (v_{11})^T(\beta\bar{L} + \nabla^2(\mathbf{1}\otimes x^*))v_{11} \\ &\qquad + (v_{21})^T(\beta\bar{L} + \nabla^2(\mathbf{1}\otimes x^*))v_{21} \end{aligned} \tag{4.84}$$

Combining Lemma 5 with the above equation, the only possibility is that $v_{11} = v_{21} = 0$. Then, the equation (4.83)) gives

$$\begin{pmatrix} -\beta\bar{L}v_{12} \\ 0 \end{pmatrix} = -b\begin{pmatrix} 0 \\ v_{22} \end{pmatrix} \text{ and } \begin{pmatrix} -\beta\bar{L}v_{22} \\ 0 \end{pmatrix} = b\begin{pmatrix} 0 \\ v_{12} \end{pmatrix}. \tag{4.85}$$

Assuming that $b \neq 0$, we know that $v_{22} = v_{12} = 0$ as well. The associated eigenvector $v_1 + iv_2 = 0$, however, makes non-sense.

Assuming that $b = 0$, we know that

$$\begin{pmatrix} \beta\bar{L} + \nabla^2(\mathbf{1}\otimes x^*) & -\beta\bar{L} \\ \beta\bar{L} & 0 \end{pmatrix}\begin{pmatrix} q \\ l \end{pmatrix} = 0. \tag{4.86}$$

Pre-multiplied by (q^T, l^T), the equation (4.86) reduces into $q^T(\beta\bar{L} + \nabla^2(\mathbf{1}\otimes x^*))q = 0$, which is equivalent to $q = 0$ by Lemma 5. Substituting $q = 0$ into equation (4.86), we can derive that l

reaches consensus. Lemma 3 (3) implies that the null space of $A(\mathbf{1}\otimes x^*)$ is

$$\text{Null}(A(\mathbf{1}\otimes x^*)) = \text{span}\left\{ \begin{pmatrix} \mathbf{0} \\ w\otimes e_1 \end{pmatrix}, ..., \begin{pmatrix} \mathbf{0} \\ w\otimes e_m \end{pmatrix} \right\}, \tag{4.87}$$

where $\{e_1, ..., e_m\}$ is an orthonormal basis for $\mathbb{R}^m$, and $w = \frac{1}{\sqrt{n}}\mathbf{1} \in \mathbb{R}^n$ is a unit vector. By a similar process, it can be checked that

$$\text{Null}(A(\mathbf{1}\otimes x^*)^T) = \text{Null}(A(\mathbf{1}\otimes x^*)). \tag{4.88}$$

We have proved that zero eigenvalue of $A(\mathbf{1}\otimes x^*)$ have the same number of algebraic multiplicity and geometric multiplicity, and we have shown that its null space is m-dimensional. Thus, zero eigenvalue of $A(\mathbf{1}\otimes x^*)$ has algebraic multiplicity and geometric multiplicity m. Similarly, it can be verified that zero eigenvalue of $A(\mathbf{1}\otimes x^*)^T$ has algebraic multiplicity and geometric multiplicity m as well.

Part 2. In the following, the system (4.78) will be rewritten as an equivalent linear system with Hurwitz coefficient.

Considering the construction of S, we know that the $(nm+1)$-th to $(nm+m)$-th columns and rows of the block matrix $(S\otimes I)A(\mathbf{1}\otimes x^*)(S\otimes I)^T$ are all zero vectors. By conducting the same transformation as we did in the proof of Theorem 9, the system (4.78) can be reduced into

$$\dot{r} = -J_3 r, \tag{4.89}$$

where r is defined in the proof of Theorem 9, and the sepctrum of J_3 is composed of the eigenvalues of $A(\mathbf{1}\otimes x^*)$ with positive real parts .

Step 2. This step is devoted to discussing the convergence behavior of system (4.4).

Noticing that

$$\begin{cases} \text{Null}(A(x)) \subset \text{Null}(A(\mathbf{1}\otimes x^*)), \\ \text{Null}(A(x)^T) \subset \text{Null}(A(\mathbf{1}\otimes x^*)^T), \end{cases} \tag{4.90}$$

we can apply the same transformation as step 1 on $A(x)$, and obtain that

$$\dot{r} = -J_4(x) r, \tag{4.91}$$

where $J_4(\mathbf{1}\otimes x^*) = J_3$ According to Proposition 3, we know that $\|r\| \to 0$ implies that $x \to \mathbf{1}\otimes x^*$, which further implies that $J_4(x) \to J_3$, that is to say, for every $\varepsilon \in (0,1]$, there exists some $\delta > 0$

such that

$$\|r\| \leq \delta \implies \|J_4(x) - J_3\| \leq \frac{\varepsilon}{2\lambda_{\max}(P_3)}. \tag{4.92}$$

Denote by J the smallest real part in the spectrum of J_3, which is also the smallest positive real part in the spectrum of $A(\mathbf{1} \otimes x^*)$. Let $\mu = J - \varepsilon$. Since $-J_3 + \mu I_{(n-1)m}$ is Hurwitz, we know that there exists a positive definite matrix P_3 which is the solution to the Lyapunov equation

$$P_3(J_3 - \mu I_{(n-1)m}) + (J_3 - \mu I_{(n-1)m})^T P_3 = I_{m(n-1)}. \tag{4.93}$$

Let

$$V(r) = r^T P_3 r. \tag{4.94}$$

The derivative of V along trajectories of (4.91) is given by

$$\begin{aligned} \dot{V}(r) &= -r^T(P_3 J_4 + J_4^T P_3)r \\ &= -r^T(P_3 J_3 + J_3^T P_3)r - r^T(P_3(J_4 - J_3) \\ &\qquad + (J_4 - J_3)^T P_3)r \\ &\leq -(1 - 2\lambda_{\max}(P_3)\tfrac{\varepsilon}{2\lambda_{\max}(P_3)})\|r\|^2 - 2\mu r^T P_3 r \\ &= -(1-\varepsilon)\|r\|^2 - 2\mu V \\ &\leq -2\mu V \end{aligned} \tag{4.95}$$

provided that $\|r\| \leq \delta$.

By the Grönwall's inequality, we have that

$$V(r) \leq V(r(0))e^{-2\mu t}, \tag{4.96}$$

and thus

$$\begin{aligned} & \left\|(q^T\, l^T)^T\right\| \leq \|S \otimes I\|\|r^*\| \\ = \;& \|S \otimes I\|\|r\| = \|r\| \\ \leq \;& C_2\left\|(q^T(0), l^T(0))^T\right\| e^{-\mu t}, \end{aligned} \tag{4.97}$$

for all $t \geq 0$, where $C_2 = \sqrt{\lambda_{\max}(P_3)/\lambda_{\min}(P_3)}$.

Combining the above inequality with Proposition 3 (2), we complete the proof. □

□

Remark 10. *The value of v^* is uniquely determined by $\beta \bar{L} v^* = -\nabla(\mathbf{1} \otimes x^*)$ and $\sum_i (v^*)^i = \sum_i v^i(0)$, which will be mentioned in Proposition 3.*

Remark 11. *Noting that the algorithm (4.2) with $\beta = 1$ is reduced to the classical PI algorithm*

(4.1), *we can prove that the system* (4.1) *also converges globally asymptotically and locally exponentially under the same assumptions as the above theorem.*

Remark 12. *Since Assumption 9 allows* $\nabla^2 f_i(x^*) = 0$ *for some i, Theorem 10 is more than a localization of Theorem 9.*

Assume that all objective functions are quadratic. As all Hessian matrices are constant, a straightforward corollary follows.

Corollary 7. *Assume that Assumptions 10 and 11 hold, and the local objective functions* f_i $(i \in X)$ *are quadratic functions. Consider a trajectory of* (4.2) *starting from* $(x^T(0), v^T(0))^T \in \mathbb{R}^{2mn}$. *Let* x^* *be the optimal solution to* (1.6).

(1) The trajectory converges to some equilibrium $((\mathbf{1}\otimes x^*)^T, (v^*)^T)^T$ *globally asymptotically, i.e.,*

$$\lim_{t\to\infty} \begin{pmatrix} x \\ v \end{pmatrix} = \begin{pmatrix} \mathbf{1}\otimes x^* \\ v^* \end{pmatrix}. \tag{4.98}$$

(2) The trajectory converges to some equilibrium $((\mathbf{1}\otimes x^*)^T, (v^*)^T)^T$ *globally exponentially with rate no less than* μ, *where* μ *is the smallest positive real part in the spectrum of the constant matrix*

$$A(x) = \begin{pmatrix} \beta\bar{L} + \nabla^2(x) & -\beta\bar{L} \\ \beta\bar{L} & 0 \end{pmatrix}, \tag{4.99}$$

and β *is an arbitrary positive constant. That is to say, there exists some positive constant r such that*

$$\left\| ((x - \mathbf{1}\otimes x^*)^T, (v - v^*)^T)^T \right\| \le ce^{-\mu t} \ (t \ge 0) \tag{4.100}$$

for all $(x^T(0), v^T(0))^T \in B(((\mathbf{1}\otimes x^*)^T, (v^*)^T)^T, r)$, *where* $c > 0$ *relies on* $x(0)$ *and* $v(0)$.

4.4 Simulations

In this section, we conduct numerical simulations to illustrate convergence behaviors of the proposed algorithm. For consistency, we discretize the algorithm (4.2) using the Euler's method with the same step size for each example and apply it on two objective functions to illustrate the convergence behavior established in Theorems 1 and 2. The step size is chosen to be 10^{-3} in the first example and 10^{-2} in the second example.

4.4.1 Strongly Convex Case

Define some three-dimensional μ-strongly convex and M-smooth functions as

$$f_i(x) = \sum_{j=1}^{3} |x_j - a_{ij}|^2, \quad i = 1, 2, \ldots, 20, \tag{4.101}$$

where $\mu = M = 2$, all entries of a_{ij}, and the initial value $x(0) \in \mathbb{R}^{60}$ are drawn from independent standard normal distributions. The first graph, G_1, is generated by the Erdős–Rényi model [20] with connectivity probability 0.75. The second graph, G_2, is a 20-cycle graph, which means that 20 agents are arranged into a cycle, and each agent is connected to its left and right agents. The third graph, G_3, is a 4×5 2-D grid.

Figure 4.4.1 illustrates the time history of the distances between x and x^* in (4.2), which agrees with the analysis in Section III. As expected in Section III, the simulation results show that increasing β improves the convergence behavior of the virtual state x in the classical PI algorithm (4.1) [78]. For a relatively smaller value of β (e.g., $\beta = 1$), the corresponding plots are almost straight lines after a few thousand iterations, which implies asymptotic and locally exponential convergence, and is consistent with the theoretical expectation in Theorem 10. When β becomes larger (e.g., $\beta = 10$), the simulation results are almost parallel to the dashed line, which implies that $\|x - x^*\| = O(e^{-\mu t})$ for all the first twenty thousand iterations. This is consistent with the conclusion of Theorem 9.

4.4.2 Strictly Convex Case

Define some three-dimensional convex functions

$$f_i(x) = \frac{1}{4} \sum_{j=1}^{3} b_{ij} |x_j - a_{ij}|^4, \, i = 1, 2, \ldots, 5, \tag{4.102}$$

where all a_{ij} and all entries of the initial values $x(0) \in \mathbb{R}^{60}$ are drawn from independent standard normal distributions, b_{ij} are uniformly distributed random numbers from $[0, 1]$, and Assumption 9 is satisfied. The undirected graphs G_1 to G_3 are the same as before.

Figure 4.4.2 shows the evolution of agents in system (4.2) for different parameters and topologies G_1, G_2, and G_3, which agrees with the analysis in Section III. As shown in Fig. 2, the virtual agents x reach a neighborhood of the minimizer after thousands of iterations. Afterwards, the lines in Figure 4.4.2 decrease exponentially, which agrees with Theorem 10.

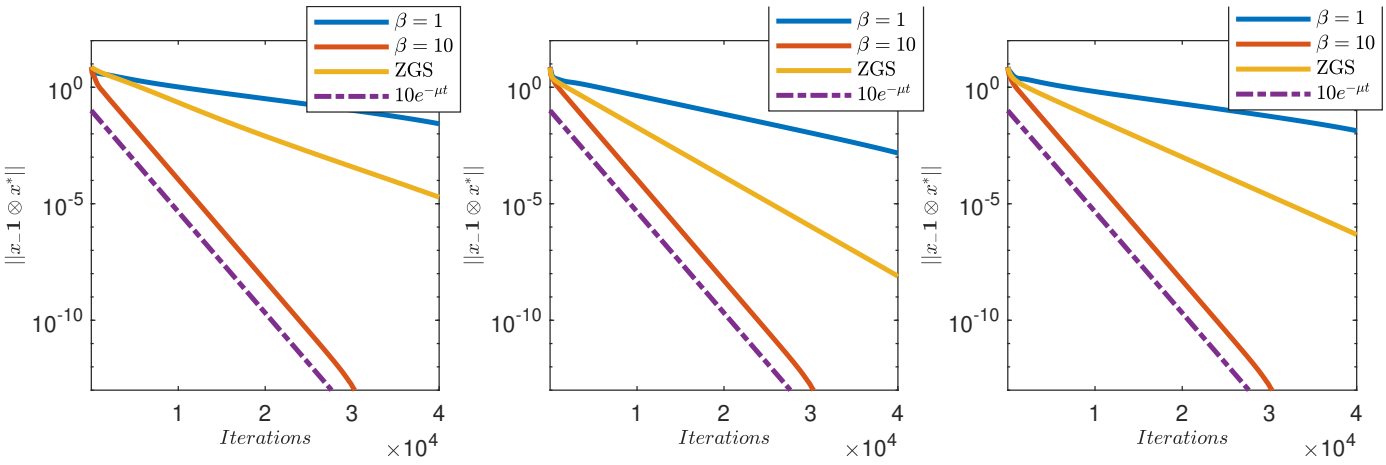

Figure 4.1: For μ-strongly convex case, the algorithm (4.2) with large β outperforms the classical PI algorithm (4.1) [78] and the ZGS algorithm [47]. Left: G_1 with $\beta_2 = 0.3502$. Middle: G_2 with $\beta_2 = 0.4894$. Right: G_3 with $\beta_2 = 0.3820$.

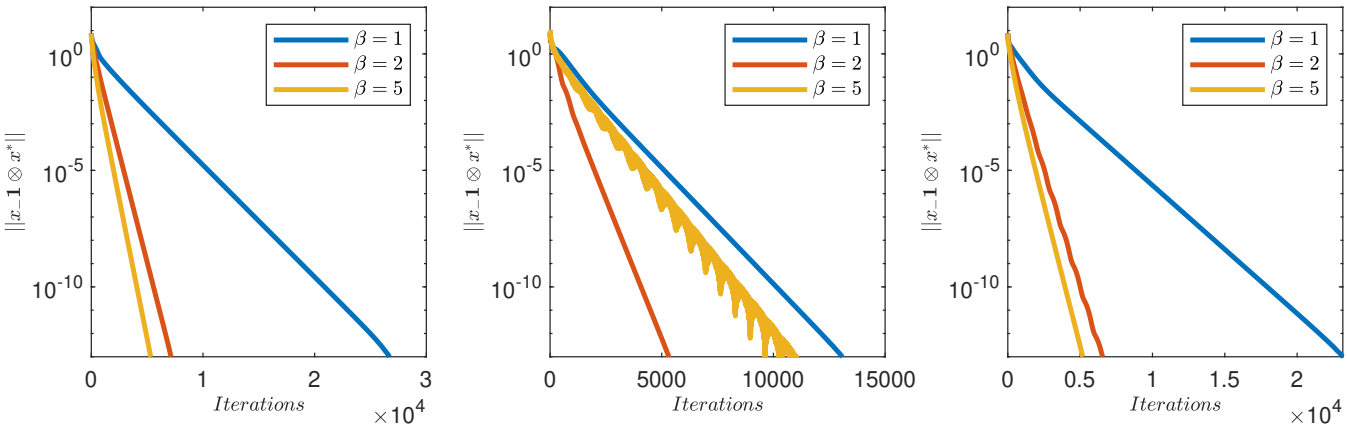

Figure 4.2: Under the strictly convex case, the algorithms (4.2) converge to some neighborhood of the optimal solution asymptotically, and converges to the optimal solution afterwards. Increasing β improves the algorithm (4.2) under different topologies. Left: G_1 with $\beta_2 = 0.3502$. Middle: G_2 with $\beta_2 = 0.4894$. Right: G_3 with $\beta_2 = 0.3820$.

Figure 4.4.2 also reveals unexpected behavior that surpasses our expectations. Under strictly convex cases, we did not analyze the influence of β on the convergence rate and settling time. Fig. 2, however, suggests that increasing β does improve the behavior of algorithm (4.2).

4.5 Summary

This chapter introduces an innovative gradient-based PI algorithm. Rather than directly examining the behavior of the PI algorithm (4.2), we first construct an auxiliary system (4.4) by differentiating the original system. By establishing a connection between (4.4) and (4.2), we can derive conclusions about the convergence rates of the PI algorithm.

Our findings are significant: We demonstrate that the convergence rate of this novel PI algorithm aligns with that of centralized gradient descent, particularly in scenarios involving strongly convex functions. Furthermore, the chapter delves into exploring the algorithm's local linear convergence characteristics when applied to strictly convex functions. This dual analysis provides a deeper understanding of the algorithm's performance and potential applications.

Chapter 5

Powered Algorithms for Finite-time and Fixed-time Distributed Optimization

In practical settings, achieving control objectives within a finite time frame is often essential. Furthermore, finite-time and fixed-time stable systems converge more rapidly and exhibit superior resistance to disturbances compared to those that are merely asymptotically stable [81]. As discussed in the introductory chapter, studying finite-time and fixed-time distributed convex optimization problems for continuous-time multi-agent systems has evolved significantly.

The exponential convergence of a novel proportional-integral (PI) algorithm is studied in the previous chapter. Inspired by the Powerball method and some finite-time or fixed-time algorithms [12,24,41,91,92], we have further developed the finite-time and fixed-time PI algorithms. This chapter confirms the algorithm's finite-time convergence for strictly convex functions, and establishes its fixed-time convergence for strongly convex functions. These theories are further supported by empirical evidence from simulations.

5.1 Finite-time PI Algorithm

The same as chapter 4, denote by $x_i \in \mathbb{R}^m$ an estimate of the optimal solution x^* by agent $i \in X$, and $v_i \in \mathbb{R}^n$ an adapter for neutralizing the influence caused by the difference of $\nabla f_i(x^*)$ $(i \in X)$. Let $x = ((x_1)^T, ..., (x_n)^T)^T$, $v = ((v_1)^T, ..., (v_n)^T)^T$, L be the Laplacian matrix of G, and β be some positive constant.

It was introduced in [92] and [91] that powerball function is of the form $\sigma^\gamma(z) = sign(z)|z|^\gamma$ $(z \in \mathbb{R}, \gamma \in (0,1))$, and $\sigma^\gamma(z) = (\sigma^\gamma(z_1), ..., \sigma^\gamma(z_n))^T$ $(z = (z_1, ..., z_n)^T \in \mathbb{R}^n, \gamma \in (0,1))$. Given

the intuition of [92] and [91] and the asymptotic convergence of the algorithm (4.1) [78], we propose a finite-time PI method of the form

$$\begin{aligned} \dot{v} &= \sigma^{\gamma}(\beta \bar{L} x), \\ \dot{x} &= -\sigma^{\gamma}(\beta \bar{L} v + \alpha \bar{L} x + \nabla(x)), \end{aligned} \tag{5.1}$$

where α and $\beta > 0$ and the other notations are the same as before.

5.1.1 Convergence Analysis

Similar to the analysis in chapter 4, we will start with analyzing the convergence of $\dot{x}$ and $\dot{v}$. Let $q = \nabla(x) + \beta \bar{L} v + \alpha \bar{L} x$ and $l = -\beta \bar{L} x$. Differentiation on q and l according to (5.1) gives that

$$\begin{pmatrix} \dot{q} \\ \dot{l} \end{pmatrix} = \begin{pmatrix} -\nabla^2(x) - \alpha \bar{L} & -\beta \bar{L} \\ \beta \bar{L} & 0 \end{pmatrix} \begin{pmatrix} \sigma^{\gamma}(q) \\ \sigma^{\gamma}(l) \end{pmatrix}, \tag{5.2}$$

where $\nabla^2(x) = \operatorname{diag}\begin{pmatrix} \nabla^2 f_1(x_1) & \cdots & \nabla^2 f_n(x_n) \end{pmatrix} \in \mathbb{R}^{nm}$, and Assumption 11 implies that $\mu I_{nm} \leq \nabla^2(x) \leq M I_{nm}$ for all $x \in \mathbb{R}^{nm}$.

The following proposition presents that the algorithm (5.2) has global asymptotic stability. The convergence behavior of (5.1), however, is not discussed here since we care about x only.

Proposition 4. *Under Assumptions 9 and 10, the origin of* (5.2) *is globally asymptotically stable.*

Proof. Denote by $V = \frac{1}{\gamma+1}\|q\|^{\gamma+1} + \frac{1}{\gamma+1}\|l\|^{\gamma+1}$ a Lyapunov function. After taking derivative, we obtain that

$$\begin{aligned} \dot{V} &= \sigma(q)^T \dot{q} + \sigma(l)^T \dot{l} \\ &= \begin{pmatrix} \sigma(q) \\ \sigma(l) \end{pmatrix}^T \begin{pmatrix} -\nabla^2(x) - \alpha \bar{L} & \beta \bar{L} \\ -\beta \bar{L}^T & 0 \end{pmatrix} \begin{pmatrix} \sigma(q) \\ \sigma(l) \end{pmatrix} \\ &\leq -\sigma(q)^T (\nabla^2(x) + \alpha \bar{L}) \sigma(q) \leq 0 \end{aligned} \tag{5.3}$$

Now we will prove that $\nabla^2(x) \neq 0$ in $E = \{x : \dot{V} = 0\} \subset \{x : q \text{ reaches consensus}\}$. Consider a trajectory in E reaches a point x where $\nabla^2(x) = 0$. Then, the related q and $\dot{q} = \beta \bar{L} \sigma(l)$ have to reach consensus, which means that $\sigma(l)$ has to reach consensus and $\dot{q} = 0$ due to Lemma 6. Since $\sigma(\cdot)$ is strictly monotone increasing, l also reaches consensus. Recalling that $l = \beta \bar{L} x$, we know that l has to be zero and x reaches consensus, which contradicts our assumption that $\nabla^2(x) = 0$.

By LaSalle's Invariance Principle, the solution will asymptotically approach the largest invariant set in $E = \{x : \dot{V} = 0\}$. Lemma 5 and Assumption 10 imply that $E = \{q = 0\}$. For

any trajectory in E, $\dot{q} = -\beta\bar{L}\sigma(l) = 0$, which implies that $\sigma(l)$ reaches consensus by Lemma 6. Since $\sigma(\cdot)$ is strictly monotone increasing, l also reaches consensus. Again, using Lemma 6, we know that $l = \beta\bar{L}x$ is zero. Therefore, we can conclude that $E = \{q = l = 0\}$. □

Before the convergence analysis on algorithm (5.1), we need to define homogeneous vector field first.

Definition 5. *A vector field $h(x) = (h_1(x), ..., h_n(x))^T$ is homogeneous of degree $k \in \mathbb{R}$ with dilation $(r_2, ..., r_n) \in \mathbb{R}^n$ if*

$$h_i(\varepsilon^{r_1}x_1, ..., \varepsilon^{r_n}x_n) = \varepsilon^{k+r_i}h(x) \tag{5.4}$$

holds for all $\varepsilon > 0$ and all $i \in \{1, ..., n\}$, where $x = (x_1, ..., x_n)^T$.

The following lemma from [32] provides a sufficient condition for finite-time stabilization, and will be exploited to prove the main result of this section.

Lemma 7. *Consider the following system:*

$$\dot{x} = f(x) + \tilde{f}(x),\ x \in \mathbb{R}^n \tag{5.5}$$

where $f(x)$ is an n-dimensional continuous homogeneous vector field of degree $k < 0$ with dilation $(r_1, ..., r_n)$ satisfying $f(0) = 0$, and $\tilde{f}$ is also a continuous vector field satisfying $\tilde{f}(0) = 0$. Assume that the zero solution of $\dot{x} = f(x)$ is asymptotically stable. Then, the zero solution of (5.5) is locally finite-time stable if

$$\lim_{\varepsilon \to 0^+} \tilde{f}_i(\varepsilon^{r_1}x_1, ..., \varepsilon^{r_n}x_n)/\varepsilon^{k+r_i} = 0,\ i = 1, ..., n, \tag{5.6}$$

uniformly for any $x \in \{x \in \mathbb{R}^n : \|x\| = 1\}$

Assuming that all objective functions are convex, we provide the finite-time convergence result for the algorithm (5.1).

Theorem 11. *Under Assumptions 9 and 10, the virtual state x of the system (5.1) converges to $\mathbf{1} \otimes x^*$ within a finite time $T_1 < \infty$, where x^* is the optimal solution of problem (1.6), and $\alpha, \beta > 0$.*

Proof. **Step 1** Given that (x_e, v_e) is an equilibrium point of system (5.1), i.e.,

$$\begin{aligned} &\beta\bar{L}x_e = 0, \\ &\beta\bar{L}x_e + \beta\bar{L}v_e + \nabla(x_e) = 0, \end{aligned} \tag{5.7}$$

it will be shown that $x_e = x^*$.

From the first equality of (5.7) and Lemma 3 (3), it can be recognized that there exist $\gamma \in \mathbb{R}^n$ such that

$$x_e = \mathbf{1} \otimes \gamma. \tag{5.8}$$

Substituting equation (5.8) into the second line of (5.7), and having it left-multiplied by $\mathbf{1} \otimes I_m$, we have that

$$\textstyle\sum_i \nabla f_i(\gamma) = 0, \tag{5.9}$$

which implies that $\gamma = x^*$ and

$$x_e = \mathbf{1} \otimes x^*. \tag{5.10}$$

Therefore, it is only left to show the finite time stability of system (5.2) in the following steps.

Step 2. Define

$$g(q,l) = \begin{pmatrix} -\nabla^2(\mathbf{1} \otimes x^*) - \alpha \bar{L} & \beta \bar{L} \\ -\beta \bar{L} & 0 \end{pmatrix} \begin{pmatrix} \sigma(q) \\ \sigma(l) \end{pmatrix} \tag{5.11}$$

as well as

$$\tilde{g}(q,l) = \begin{pmatrix} -\nabla^2(x) + \nabla^2(\mathbf{1} \otimes x^*) & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \sigma(q) \\ \sigma(l) \end{pmatrix}, \tag{5.12}$$

which satisfy $g(0,0) = \tilde{g}(0,0) = 0$, and $g(\varepsilon q, \varepsilon l) = \varepsilon^{1+(\gamma-1)} g(q,l)$, i.e. $g(q,l)$ is homogeneous of degree $\gamma - 1 < 0$ with dilation $(1,...,1)$. It has been proved in proposition 1 that $(\dot{q}^T, \dot{l}^T)^T = g(q,l)$ is globally asymptotically stable.

Step 3. This step is devoted to show that $\tilde{f}$ satisfies (5.6).

Part 1. Let the point $(q^T, l^T)^T$ lie in $\{z \in \mathbb{R}^{2nm} : \|z\| = 1\}$ and $\varepsilon > 0$ be small enough. Consider the system

$$\begin{cases} \beta \bar{L} v_\varepsilon + \alpha \bar{L} x_\varepsilon + \nabla(x_\varepsilon) = \varepsilon q, \\ \beta \bar{L} x_\varepsilon = \varepsilon l. \end{cases} \tag{5.13}$$

Since $\bar{L}$ is not full rank, there exist some x_ε such that the second equality hold. For the same reason, there exists some v_ε solves $\varepsilon q - \beta \bar{L} x_\varepsilon - \nabla(x_\varepsilon) = \beta \bar{L} v_\varepsilon$.

Part 2. Now, we will prove that there is $\delta > 0$ such that $\|x_\varepsilon - \mathbf{1} \otimes x^*\| \leq c\varepsilon$ for $\varepsilon < \delta$ and some $c > 0$, i.e., there exist some neighborhood of $\mathbf{1} \otimes x^*$ such that x_ε converges to $\mathbf{1} \otimes x^*$ uniformly.

By step 1 and the continuity of ∇f_i, there exists $\delta > 0$ such that $x_\varepsilon \in B(\mathbf{1} \otimes x^*, 1)$ for all $\varepsilon < \delta$. The average of x_ε is denoted by $\bar{x}_\varepsilon = \frac{1}{n} \sum_i (x_\varepsilon)_i$. By the definition of $B(\cdot,\cdot)$, it can be verified that $x_\varepsilon \in B(\mathbf{1} \otimes x^*, 1)$ implies $\bar{x}_\varepsilon \in B(x^*, 1)$. Since $B(x^*, 1)$ and $B(\mathbf{1} \otimes x^*, 1)$ are compact, there exists

positive constants $M, N > 0$ such that

$$NI_{nm} \leq \nabla^2 f(\bar{x}_\varepsilon),\ \nabla^2 f_i((x_\varepsilon)_i) \leq MI_{nm}. \tag{5.14}$$

Denote $Null^\perp(\bar{L})$ by the orthogonal complement of $Null(\bar{L})$. For any $z \in \mathbb{R}^{nm}$, there exist unique vectors $z_p \in Null^\perp(\bar{L})$ and $z_o \in Null(\bar{L})$ such that $z = z_p + z_o$ as $Null^\perp(\bar{L}) \oplus Null(\bar{L}) = \mathbb{R}^{nm}$.

Denote β_2 by $\lambda_2(L)$, which equals to $\lambda_{m+1}(\bar{L})$ by Lemma 3. It is also the smallest positive eigenvalue of L and $\bar{L}$. Thus, $(\beta_2)^2 = \lambda_2(L^2) = \lambda_{m+1}(\bar{L}^2)$ as L and $\bar{L}$ are symmetric and positive semi-definite. By Courant–Fischer Theorem from [33], we know that

$$\begin{aligned}
\beta_2^2 &= \max_{\{S\subset R^{nm}:\dim S = mn-m\}} \min_{\{z:0\neq z\in S\}} \frac{z^T\bar{L}^2 z}{z^T z} \\
&= \max_{\{S\subset R^{nm}:\dim S = mn-m\}} \min_{\{z:0\neq z\in S\}} \frac{(z_p+z_o)^T\bar{L}^2(z_p+z_o)}{(z_p+z_o)^T(z_p+z_o)} \\
&\leq \max_{\{S\subset R^{nm}:\dim S = mn-m\}} \min_{\{z:0\neq z\in S\}} \frac{(z_p)^T\bar{L}^2 z_p}{(z_p)^T z_p} \\
&= \min_{\{z:0\neq z\in Null^\perp(\bar{L})\}} \frac{z^T\bar{L}^2 z}{z^T z} = \min_{\{z:0\neq z\in Null^\perp(\bar{L})\}} \frac{\|\bar{L}z\|^2}{\|z\|^2}.
\end{aligned} \tag{5.15}$$

When $\varepsilon < \delta$, we can conclude that

$$\begin{aligned}
& \|x_\varepsilon - \mathbf{1}\otimes x^*\| \\
\leq\ & \|\delta_x\| + \|\mathbf{1}\otimes(\bar{x}_\varepsilon - x^*)\| \\
\overset{(5.15)}{\leq}\ & \tfrac{\varepsilon}{\beta\beta_2}\|l\| + \sqrt{n\textstyle\sum_i(\bar{x}_\varepsilon - x^*)_i^2} \\
\leq\ & \tfrac{\varepsilon}{\beta\beta_2}\|l\| + \sqrt{n}\|\bar{x}_\varepsilon - x^*\| \\
\leq\ & \tfrac{\varepsilon}{\beta\beta_2}\|l\| + \tfrac{\sqrt{n}}{N}\|\nabla f(\bar{x}_\varepsilon)\| \\
\leq\ & \tfrac{\varepsilon}{\beta\beta_2}\|l\| + \tfrac{\sqrt{n}}{N}(\big\|\nabla f(\bar{x}_\varepsilon) - \tfrac{1}{n}\textstyle\sum_i \nabla f_i(x_\varepsilon^i)\big\| \\
& + \big\|\tfrac{1}{n}\textstyle\sum_i \nabla f_i(x_\varepsilon^i)\big\|) \\
\overset{(a)}{\leq}\ & \tfrac{\varepsilon}{\beta\beta_2}\|l\| + \tfrac{\sqrt{n}M}{N}\|\delta_x\| + \tfrac{\varepsilon}{n}\big\|(\mathbf{1}\otimes I_m)^T q\big\| \\
\leq\ & \varepsilon(\tfrac{1}{\beta\beta_2}\|l\| + \tfrac{\sqrt{n}M}{N\beta\beta_2}\|l\| + \tfrac{1}{\sqrt{n}}\|q\|) \leq c\varepsilon,
\end{aligned} \tag{5.16}$$

where $c = \frac{1}{\beta\beta_2} + \frac{\sqrt{n}M}{N\beta\beta_2} + \frac{1}{\sqrt{n}}$, and (a) is implied by

$$\left\|\sum_i \nabla f_i(x_\varepsilon^i)\right\| = \left\|(\mathbf{1}\otimes I_m)^T \nabla(x_\varepsilon)\right\| = \left\|(\mathbf{1}\otimes I_m)^T \varepsilon q\right\|. \tag{5.17}$$

Then, it is recognized by

$$\begin{aligned} |\tilde{g}(\varepsilon q, \varepsilon l)/\varepsilon^\gamma| &\leq M\|x_\varepsilon - \mathbf{1}\otimes x^*\|\|\varepsilon q\|^\gamma/\varepsilon^\gamma \\ &\leq M\|x_\varepsilon - \mathbf{1}\otimes x^*\| \\ &\leq cM\varepsilon, \end{aligned} \tag{5.18}$$

and we complete checking the last condition (5.6).

Step 4. According to Lemma 7, the origin of protocol (5.2) is locally finite-time stable, i.e., there is $R > 0$ such that any trajectory starting from $(q_1^T, l_1^T)^T \in B(0,R)$ would reach the origin in a finite time $T_1 = T_1(q_1, l_1)$.

According to proposition 1, the system (5.2) is globally asymptotically stable, i.e., every trajectory of (5.2) will encounter $B(0,R)$ in finite-time $T_2 = T_2(q_2, l_2)$, where $((q_2)^T, (l_2)^T)^T$ is the initial state.

Therefore, every trajectory of (5.2) will approach the optimal solution within finite-time $T = T_1(q_1, l_1) + T_2(q_2, l_2)$. By step 1, it can be derived that state x in system (5.1) converges to the optimal solution in finite time T as well. □

5.2 Fixed-time PI Algorithm

Motivated by some fixed-time algorithms for distributed optimization [12, 24, 41], which contain terms of the form $\sigma^\xi(\cdot)$ $(\xi > 1)$ and $\sigma^\gamma(\cdot)$ $(0 < \gamma < 1)$, we construct the second control scheme

$$\begin{aligned} \dot{v} &= -\sigma^\gamma(\beta\bar{L}x) - \sigma^\xi(\beta\bar{L}x), \\ \dot{x} &= -\sigma^\gamma(-\beta\bar{L}v + \alpha\bar{L}x + \nabla(x)) - \sigma^\xi(-\beta\bar{L}v + \alpha\bar{L}x + \nabla(x)), \end{aligned} \tag{5.19}$$

where $0 < \gamma < 1$ and $\xi > 1$. Intuitively, the $\sigma^\xi(\cdot)$ term would drive the virtual state to some neighborhood of $\mathbf{1}\otimes x^*$ in some fixed time, and afterwards, the $\sigma^\gamma(\cdot)$ term would drive it to the optimal solution in a fixed time.

5.2.1 Convergence Analysis

Define $q = \nabla(x) - \beta\bar{L}v + \alpha\bar{L}x$ and $l = \beta\bar{L}x$. Again, we take derivatives on q and l, and obtain a variation of (5.19)

$$\begin{pmatrix} \dot{q} \\ \dot{l} \end{pmatrix} = \begin{pmatrix} -\nabla^2(x) - \alpha\bar{L} & -\beta\bar{L} \\ \beta\bar{L} & 0 \end{pmatrix} \left(\begin{pmatrix} \sigma^{\gamma}(q) \\ \sigma^{\gamma}(l) \end{pmatrix} + \begin{pmatrix} \sigma^{\xi}(q) \\ \sigma^{\xi}(l) \end{pmatrix} \right), \tag{5.20}$$

where $\bar{m} = \frac{\mu+M}{2}$. Define $\Delta(x) = \nabla^2(x) - \bar{m}I_{nm}$. We have that $-\frac{M-\mu}{2}I_{nm} \leq \Delta(x) \leq \frac{M-\mu}{2}I_{nm}$ by Assumption 11.

We provide some lemmas that will be helpful for our analysis.

Lemma 8. *For any positive semi-definite matrix $Q \in \mathbb{R}^N$ ($N \in \mathbb{R}_+$), $Qz = 0$ ($z \in \mathbb{R}^N$) occurs if and only if $z^T Qz = 0$.*

Proof. ($\Longrightarrow$) Trivial.

($\Longleftarrow$) Noting that $z^T Qz = z^T Q^{1/2} Q^{1/2} z = \|Q^{1/2}z\|^2 = 0$, we have that $Q^{1/2}z = 0$. Left multiplying both side by $Q^{1/2}$ gives the desired inequality. □

The proofs of the following lemmas can be found in the Appendix.

Lemma 9. *For all $\nu, \alpha_1, \alpha_1 > 0$ and $z_1, z_2 \in \mathbb{R}^N$ ($N \in \mathbb{N}$), we have that*

$$\begin{array}{ll} & (\alpha_1)^{\nu+1}(\|z_1\|_{\nu+1})^{\nu+1} + (\alpha_2)^{\nu+1}(\|z_2\|_{\nu+1})^{\nu+1} \\ \geq & \alpha_1(\alpha_2)^{\nu}\|z_1\|_{\nu+1}(\|z_2\|_{\nu+1})^{\nu} + (\alpha_1)^{\nu}\alpha_2(\|z_1\|_{\nu+1})^{\nu}\|z_2\|_{\nu+1} \end{array} \tag{5.21}$$

Proof. If z_1 or z_2 is a zero vector, the inequality is trivially true.

If z_1 and z_2 are trivial, it is checked that

$$\begin{array}{ll} & ((\alpha_1 z_1) - (\alpha_2 z_2))^T(\sigma^{\nu}(\alpha_1 z_1) - \sigma^{\nu}(\alpha_2 z_2)) \\ = & \sum_i ((\alpha_1 z_1)_i - (\alpha_2 z_2)_i)(\sigma^{\nu}(\alpha_1 z_1)_i - \sigma^{\nu}(\alpha_2 z_2)_i) \\ \geq 0, & \end{array} \tag{5.22}$$

therefore, we obtain that

$$\begin{aligned}
& (\alpha_1)^{\nu+1}(\|z_1\|_{\nu+1})^{\nu+1} + (\alpha_2)^{\nu+1}(\|z_2\|_{\nu+1})^{\nu+1} \\
\geq\ & \alpha_1(\alpha_2)^{\nu}(z_1)^T\sigma^{\nu}(z_2) + (\alpha_1)^{\nu}\alpha_2(z_2)^T\sigma^{\nu}(z_1) \\
=\ & \alpha_1(\alpha_2)^{\nu}\frac{(z_1)^T\sigma^{\nu}(z_2)}{\|z_1\|\|z_2\|^{\nu}}\|z_1\|\|z_2\|^{\nu} \\
& +(\alpha_1)^{\nu}\alpha_2\frac{(z_2)^T\sigma^{\nu}(z_1)}{\|z_1\|^{\nu}\|z_2\|}\|z_1\|^{\nu}\|z_2\|.
\end{aligned} \tag{5.23}$$

We can further derive that

$$\begin{aligned}
& (\alpha_1)^{\nu+1}(\|z_1\|_{\nu+1})^{\nu+1} + (\alpha_2)^{\nu+1}(\|z_2\|_{\nu+1})^{\nu+1} \\
=\ & \alpha_1(\alpha_2)^{\nu}\|z_1\|\|z_2\|^{\nu}\max_{z_1,z_2\neq 0}\frac{(z_1)^T\sigma^{\nu}(z_2)}{\|z_1\|\|z_2\|^{\nu}} \\
& +(\alpha_1)^{\nu}\alpha_2\|z_1\|^{\nu}\|z_2\|\max_{z_1,z_2\neq 0}\frac{(z_2)^T\sigma^{\nu}(z_1)}{\|z_1\|^{\nu}\|z_2\|} \\
=\ & \alpha_1(\alpha_2)^{\nu}\|z_1\|\|z_2\|^{\nu}\cdot 1 + (\alpha_1)^{\nu}\alpha_2\|z_1\|^{\nu}\|z_2\|\cdot 1 \\
=\ & \alpha_1(\alpha_2)^{\nu}\|z_1\|\|z_2\|^{\nu} + (\alpha_1)^{\nu}\alpha_2\|z_1\|^{\nu}\|z_2\|,
\end{aligned} \tag{5.24}$$

where the maximums are reached when z_1 and z_2 are parallel and have the same direction. □

Lemma 10. *1. For all* $1 < \mu < 2$, $z \in \mathbb{R}^N$ $(\forall N \in \mathbb{R}_+)$

$$\|z\|_{\mu} \geq \|z\|_2. \tag{5.25}$$

2. For all $\mu > 2$, $z \in \mathbb{R}^N$ $(\forall N \in \mathbb{R}_+)$

$$\|z\|_2 \leq \|z\|_{\mu}(N)^{\frac{\mu-2}{2\mu}}. \tag{5.26}$$

Proof. 1. We will prove that $|a_1+a_2|^{\frac{\mu}{2}} \leq |a_1|^{\frac{\mu}{2}} + |a_2|^{\frac{\mu}{2}}$ holds for all $a_1, a_2 \in \mathbb{R}$. Define the function $g(x) = (|a_1|+x)^{\frac{\mu}{2}} - |a_1|^{\frac{\mu}{2}} - x^{\frac{\mu}{2}}$ $(x \in \mathbb{R}_+)$. Its derivative is $g'(x) = \frac{\mu}{2}(a_1+x)^{\frac{\mu}{2}-1} - \frac{\mu}{2}x^{\frac{\mu}{2}-1} \leq 0$, and thus, $g(x) \leq g(0) = 0$ on $\mathbb{R}_+$. We conclude that

$$|a_1+a_2|^{\frac{\mu}{2}} \leq ||a_1|+|a_2||^{\frac{\mu}{2}} \leq |a_1|^{\frac{\mu}{2}} + |a_2|^{\frac{\mu}{2}}. \tag{5.27}$$

Using (5.27), we can check that

$$\left|\sum_{i=1}^{N}(z_i)^2\right|^{\frac{\mu}{2}} \leq \left|\sum_{i=1}^{N-1}(z_i)^2\right|^{\frac{\mu}{2}} + \left|(z_N)^2\right|^{\frac{\mu}{2}} \leq \cdots \leq \sum_{i=1}^{N}\left|(z_i)^2\right|^{\frac{\mu}{2}}, \tag{5.28}$$

which gives that

$$\|z\|_2 = \left(\sum_{i=1}^{N}|z_i|^2\right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^{N}|z_i|^{\mu}\right)^{\frac{1}{\mu}} = \|z\|_{\mu}. \tag{5.29}$$

2. By Hölder's inequality, we know that

$$\left(\sum_{i=1}^{N}|z_i|^2|1|^{\mu-2}\right)^{\mu} \leq \left(\sum_{i=1}^{N}|z_i|^{\mu}\right)^{2}\left(\sum_{i=1}^{N}|1|^{\mu}\right)^{\mu-2}, \tag{5.30}$$

which further implies that

$$\left(\sum_{i=1}^{N}|z_i|^2\right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^{N}|z_i|^{\mu}\right)^{\frac{1}{\mu}}(N)^{\frac{\mu-2}{2\mu}}. \tag{5.31}$$

□

Assuming that all objective functions are strongly convex, we provide a fixed-time convergence result for the algorithm (5.19).

Theorem 12. *Given Assumptions 10 and 11, when there exists* $\alpha,\beta,c_1>0$ *such that*

$$\begin{cases} \beta\left(\frac{2\mu}{M+\mu}\right)^{\gamma} > (c_1)^{\gamma}\left(\frac{M-\mu}{2}\frac{1}{\beta_2}\right)^{\gamma+1}, \\ (c_1)^{\gamma}\frac{2\mu}{M+\mu} > \beta\left(\frac{\beta_n(1+c_1\beta)}{\bar{m}+\alpha\beta_n}\right)^{\gamma+1}, \\ \beta\left(\frac{2\mu}{M+\mu}N^{-\frac{\xi-1}{2}}\right)^{\xi} > (c_1)^{\xi}\left(\frac{M-\mu}{2}\frac{1}{\beta_2}\right)^{\xi+1}, \\ \left(c_1N^{-\frac{\xi-1}{2}}\right)^{\xi}\frac{2\mu}{M+\mu} > \beta\left(\frac{\beta_n(1+c_1\beta)}{\bar{m}+\alpha\beta_n}\right)^{\xi+1}, \end{cases} \tag{5.32}$$

the virtual state x of the system (5.19) converges to $\mathbf{1}\otimes x^*$ *within fixed time* $T_2<\infty$*, where* x^* *is the optimal solution of problem (1.6).*

Proof. Similar to the discussion from equations (5.7) to (5.10), it can be verified that $q = l = 0$ implies $x = \mathbf{1} \otimes x^*$. In the following analysis, it is sufficient to prove the fixed-time convergence of (5.20).

Step 1. Before analyzing the system (5.19), we would like to define the pseudoinverse of L.

Let β_i $(1 \leq i \leq n)$ be the i-th largest eigenvalues of L, and the eigen-decomposition $L = S_L D_L (S_L)^T$ holds for some diagonal matrix $D_L = \text{diag}\begin{pmatrix}\beta_1 & \cdots & \beta_n\end{pmatrix}$ and orthogonal matrix S_L. According Lemma 3, it holds that $0 = \beta_1 < \beta_2 \leq ... \leq \beta_n$. Denote the pseudoinverse of L by

$$L^+ = S_L \text{diag}\begin{pmatrix}0 & \frac{1}{\beta_2} & \cdots & \frac{1}{\beta_n}\end{pmatrix} (S_L)^T. \tag{5.33}$$

Matrices I_n, L and L^+ commute as they can both be diagonalized by S_L. Also, matrices I_{nm}, $\bar{L}$ and $\bar{L}^+$ commute as they can both be diagonalized by $S_L \otimes I_m$.

Step 2 Let $E = (I + c_1\beta\bar{L}^+\bar{L})(\bar{m}I + \alpha\bar{L})^{-1}$, $D = c_1\bar{L}^+$ and $C = \frac{c_1}{\beta}\bar{m}(\bar{L}^+)^2 + \frac{c_1}{\beta}\alpha\bar{L}^+ = \frac{c_1}{\beta}(\bar{L}^+)^2(\bar{m}I + \alpha\bar{L})$. We will prove that the Lyapunov candidate function

$$V(q,l) = \frac{1}{2}\begin{pmatrix}q^T & l^T\end{pmatrix}\begin{pmatrix}E & D^T\\ D & C\end{pmatrix}\begin{pmatrix}q\\ l\end{pmatrix} \tag{5.34}$$

is positive definite.

We will first show that V is positive semi-definite. Noting that $\bar{L}^+ + c_1\beta\bar{L}^+ \geq c_1\beta\bar{L}^+$ is always true, we obtain

$$(I + c_1\beta\bar{L}^+\bar{L})(\bar{m}I + \alpha\bar{L})^{-1}\left(\frac{1}{\beta}\bar{m}(\bar{L}^+)^2 + \frac{1}{\beta}\alpha(\bar{L}^+)^2\bar{L}\right) \geq (c_1)(\bar{L}^+)^2 \tag{5.35}$$

by multiplying both sides by $\frac{c_1}{\beta}L^+$. As $(\bar{L}^+)^2\bar{L} = \bar{L}^+$, it is verified that $(I + c_1\beta\bar{L}^+\bar{L})(\bar{m}I + \alpha\bar{L})^{-1}\left(\frac{c_1}{\beta}\bar{m}(\bar{L}^+)^2 + \frac{c_1}{\beta}\alpha\bar{L}^+\right) \geq c_1(\bar{L}^+)^2$, i.e., $EC - D^2 \geq 0$.

To investigate the pre-image of $V = 0$, it is sufficient to find the null space of

$$Q := \begin{pmatrix}E & D^T\\ D & C\end{pmatrix} \tag{5.36}$$

by Lemma 8. Let

$$\begin{cases}Eq + D^T l = 0,\\ Dq + Cl = 0.\end{cases} \tag{5.37}$$

Left multiplying the first equation by $\bar{m}I+\alpha\bar{L}$ and the second equation by $\beta\bar{L}$, we have the following argument:

$$\begin{cases}(I_{nm}+c_1\beta\bar{L}^+\bar{L})q+c_1\bar{L}^+(\bar{m}I+\alpha\bar{L})l=0,\\ c_1\beta\bar{L}\bar{L}^+q+c_1\bar{L}^+(\bar{m}I+\alpha\bar{L})l=0,\end{cases} \tag{5.38}$$

which gives that $I_{nm}q=q=0$ and l is consensus. By Lemma 6, we have that $l=0$ as well.

Remark 13. *Similar to the discussion on* (5.15)*, it can be verified that*

$$0<\lambda_{m+1}(Q)\leq \min_{\{z:0\neq z\in Null^{\perp}(Q)\}}\frac{z^TQz}{\|z\|^2}, \tag{5.39}$$

and thus

$$V(q,l)\geq\frac{\lambda_{m+1}(Q)}{2}\left((\|q\|_2)^2+(\|l\|_2)^2\right). \tag{5.40}$$

Step 3. The Lie derivative of V along (5.19) is

$$\begin{aligned}\dot{V} &= -\begin{pmatrix}q^T & l^T\end{pmatrix}\begin{pmatrix}E & D^T\\ D & C\end{pmatrix}\begin{pmatrix}\bar{m}I_{nm}+\Delta+\alpha\bar{L} & \beta\bar{L}\\ -\beta\bar{L} & 0\end{pmatrix}\begin{pmatrix}\sigma^{\gamma}(q)\\ \sigma^{\gamma}(l)\end{pmatrix}\\ &\quad -\begin{pmatrix}q^T & l^T\end{pmatrix}\begin{pmatrix}E & D^T\\ D & C\end{pmatrix}\begin{pmatrix}\bar{m}I_{nm}+\Delta+\alpha\bar{L} & \beta\bar{L}\\ -\beta\bar{L} & 0\end{pmatrix}\begin{pmatrix}\sigma^{\xi}(q)\\ \sigma^{\xi}(l)\end{pmatrix}\\ &= -V_{d1}-V_{d2},\end{aligned} \tag{5.41}$$

where V_{d1} represents the first term, and V_{d2} represents the second term.

Since $L(S_L)=S_LD_L$ and $S_L\in\mathbb{O}^{n\times n}$, the first column of P is $\frac{1}{\sqrt{n}}\mathbf{1}$. By showing

$$\begin{aligned}& l^T(I-\bar{L}^+\bar{L})\\ =\;& l^T(P\text{diag}(1,0,...,0)P^T)\otimes I\\ =\;& l^T\tfrac{1}{n}\mathbf{1}\mathbf{1}^T\otimes I=0,\end{aligned} \tag{5.42}$$

we could know that

$$l^T=l^T\bar{L}^+\bar{L}. \tag{5.43}$$

Therefore, we have that

$$
\begin{aligned}
V_{d1} &= -\begin{pmatrix} q^T & L^T \end{pmatrix} \begin{pmatrix} E & D^T \\ D & C \end{pmatrix} \begin{pmatrix} \bar{m} I_{nm} + \Delta + \alpha \bar{L} & \beta \bar{L} \\ -\beta \bar{L} & 0 \end{pmatrix} \begin{pmatrix} \sigma^\gamma(q) \\ \sigma^\gamma(l) \end{pmatrix} \\
&\leq -\begin{pmatrix} q^T & l^T \end{pmatrix} \begin{pmatrix} I & \frac{1+c_1\beta}{\bar{m}I+\alpha\bar{L}}\beta\bar{L} \\ 0 & c_1\beta\bar{L}^+\bar{L} \end{pmatrix} \begin{pmatrix} \sigma^\gamma(q) \\ \sigma^\gamma(l) \end{pmatrix} \\
&\quad +(\|q\|_2)^{\alpha+1}\|E\Delta\|_2 + \|l\|_2(\|q\|_2)^\alpha\|D\Delta\|_2 \\
&\overset{(5.43)}{\leq} -(\|q\|_{\gamma+1})^{\gamma+1} + \tfrac{M-\mu}{M+\mu}(\|q\|_{\gamma+1})^{\gamma+1} - c_1\beta(\|l\|_{\gamma+1})^{\gamma+1} \\
&\quad +\beta\tfrac{\beta_n(1+c_1\beta)}{\bar{m}+\alpha\beta_n}\|q\|_2(\|l\|_2)^\gamma + \tfrac{M-\mu}{2}\tfrac{c_1}{\beta_2}(\|q\|_2)^\gamma\|l\|_2 \\
&= -\tfrac{2\mu}{M+\mu}(\|q\|_{\gamma+1})^{\gamma+1} - c_1\beta(\|l\|_{\gamma+1})^{\gamma+1} \\
&\quad +\beta\tfrac{\beta_n(1+c_1\beta)}{\bar{m}+\alpha\beta_n}\|q\|_2(\|l\|_2)^\gamma + \tfrac{M-\mu}{2}\tfrac{c_1}{\beta_2}(\|q\|_2)^\gamma\|l\|_2 \\
&\overset{(a)}{\leq} -\tfrac{2\mu}{M+\mu}(\|q\|_2)^{\gamma+1} - c_1\beta(\|l\|_2)^{\gamma+1} \\
&\quad +\beta\tfrac{\beta_n(1+c_1\beta)}{\bar{m}+\alpha\beta_n}\|q\|_2(\|l\|_2)^\gamma + \tfrac{M-\mu}{2}\tfrac{c_1}{\beta_2}(\|q\|_2)^\gamma\|l\|_2 \\
&\leq -c_2((\|q\|_2)^{\gamma+1} + (\|l\|_2)^{\gamma+1}) \\
&= -c_2(\|(\|q\|_2, \|l\|_2)\|_{\gamma+1})^{\gamma+1} \\
&\overset{(b)}{\leq} -c_2(\|(\|q\|_2, \|l\|_2)\|_2)^{\gamma+1} \\
&\leq -c_2\left(\tfrac{2}{\lambda_{\max}(Q)}V\right)^{\frac{\gamma+1}{2}},
\end{aligned}
\tag{5.44}
$$

where (a) and (b) follow the first property of Lemma 10, and c_2 is some positive number. Let α_1 and α_2 be the solution to

$$
\begin{cases} \alpha_1(\alpha_2)^\gamma = \beta\frac{\beta_n(1+c_1\beta)}{\bar{m}+\alpha\beta_n}, \\ (\alpha_1)^\gamma\alpha_2 = \frac{M-\mu}{2}\frac{c_1}{\beta_2}. \end{cases}
\tag{5.45}
$$

According to Lemma 9, we can pick $c_2 = \min\{\frac{2\mu}{M+\mu} - (\alpha_1)^{\gamma+1}, c_1\beta - (\alpha_2)^{\gamma+1}\} > 0$, and the positivity of c_2 is guaranteed by the first two assumptions in (5.32).

Similarly, we obtain that

$$
V_{d2} = -\begin{pmatrix} q^T & L^T \end{pmatrix} \begin{pmatrix} E & D^T \\ D & C \end{pmatrix} \begin{pmatrix} \nabla(x) + \alpha\bar{L} & \beta\bar{L} \\ -\beta\bar{L} & 0 \end{pmatrix} \begin{pmatrix} \sigma^\xi(q) \\ \sigma^\xi(l) \end{pmatrix}
$$

$$
\begin{aligned}
&\leq -\tfrac{2\mu}{M+\mu}(\|q\|_{\xi+1})^{\xi+1} - c_1\beta(\|l\|_{\xi+1})^{\xi+1} \\
&\quad +\beta\tfrac{\beta_n(1+c_1\beta)}{\bar{m}+\alpha\beta_n}\|q\|_2(\|l\|_2)^{\xi} + \tfrac{M-\mu}{2}\tfrac{c_1}{\beta_2}(\|q\|_2)^{\xi}\|l\|_2 \\
&\overset{(c)}{\leq} -\tfrac{2\mu}{M+\mu}N^{-\frac{\xi-1}{2}}(\|q\|_2)^{\xi+1} - c_1\beta N^{-\frac{\xi-1}{2}}(\|l\|_2)^{\xi+1} \\
&\quad +\beta\tfrac{\beta_n(1+c_1\beta)}{\bar{m}+\alpha\beta_n}\|q\|_2(\|l\|_2)^{\xi} + \tfrac{M-\mu}{2}\tfrac{c_1}{\beta_2}(\|q\|_2)^{\xi}\|l\|_2 \\
&\leq -c_3((\|q\|_2)^{\xi+1} + (\|l\|_2)^{\xi+1}) \\
&= -c_3(\|(\|q\|_2, \|l\|_2)\|_{\xi+1})^{\frac{1}{\xi+1}} \\
&\overset{(d)}{\leq} -c_3((nm)^{-\frac{2\xi+2}{\xi-1}}\|(\|q\|_2, \|l\|_2)\|_2)^{\xi+1} \\
&\leq -c_3(nm)^{-\frac{2(\xi+1)^2}{\xi-1}}\left(\tfrac{2}{\lambda_{\max}(Q)}V\right)^{\frac{\xi+1}{2}},
\end{aligned}
\tag{5.46}
$$

where (c) and (d) follow the second property of Lemma 10, and $c_3 > 0$. The process of deriving c_3 is similar to that of deriving c_2.

Combining (5.44) and (5.46), we arrive at the following inequality:

$$
\dot{V} \leq -c_2\left(\tfrac{2}{\lambda_{\max}(Q)}V\right)^{\frac{\gamma+1}{2}} - c_3(nm)^{-\frac{2(\xi+1)^2}{\xi-1}}\left(\tfrac{2}{\lambda_{\max}(Q)}V\right)^{\frac{\xi+1}{2}}. \tag{5.47}
$$

Considering that $\dot{V} \leq -c_3(nm)^{-\frac{2(\xi+1)^2}{\xi-1}}\left(\tfrac{2}{\lambda_{\max}(Q)}V\right)^{\frac{\xi+1}{2}}$, we have

$$
\frac{\dot{V}}{V^{\frac{\xi+1}{2}}} \leq -c_3(nm)^{-\frac{2(\xi+1)^2}{\xi-1}}\left(\frac{2}{\lambda_{\max}(Q)}\right)^{\frac{\xi+1}{2}}, \tag{5.48}
$$

and integrating both sides yields

$$
c_3(nm)^{-\frac{2(\xi+1)^2}{\xi-1}}\left(\frac{2}{\lambda_{\max}(Q)}\right)^{\frac{\xi+1}{2}} t \leq \frac{1}{\frac{\xi+1}{2}-1}\left(V(q(t),l(t))^{\frac{1-\xi}{2}} - V(q(0),l(0))^{\frac{1-\xi}{2}}\right). \tag{5.49}
$$

When V reaches the set $\{V(q,l) \leq 1\}$, the right hand side of (5.49) is bounded by $\frac{1}{\frac{\xi+1}{2}-1}$, which is independent of the initial state. This implies that V reaches this set in a fixed time, denoted t_3, for any $q(0), l(0) \in \mathbf{R}^{nm}$.

On the other hand, $\dot{V}$ is also bounded by $-c_2\left(\frac{2}{\lambda_{\max}(Q)}V\right)^{\frac{2}{\gamma+1}}$. We have that

$$\frac{\dot{V}}{V^{\frac{\gamma+1}{2}}} \leq -c_2\left(\frac{2}{\lambda_{\max}(Q)}\right)^{\frac{\gamma+1}{2}} \tag{5.50}$$

and taking integral from $t = t_3$ gives

$$c_2\left(\frac{2}{\lambda_{\max}(Q)}\right)^{\frac{\gamma+1}{2}}(t-t_3) \leq \frac{1}{1-\frac{2}{\gamma+1}}\left(V(t_3)^{\frac{1-\gamma}{\gamma+1}} - V(t)^{\frac{1-\gamma}{\gamma+1}}\right) \leq \frac{1}{1-\frac{2}{\gamma+1}} \tag{5.51}$$

After $t = t_3$, V would reach the origin in a fixed time. The fixed-time convergence can be implied by (5.48) and (5.51). □

Remark 14. *Some observations on the role of the design parameters α, β and c_1 are given here. We can always find $\alpha, \beta, c_1 > 0$ satisfying the condition (5.32). Let $\beta = 1$ and let c_1 be sufficiently small to satisfy the first and the third inequalities. Also, we can always let α be sufficiently large such that the second and the fourth inequalities hold. Moreover, (5.32) is only a sufficient condition in simulation. When running our numerical examples, we observe that the algorithm converges in a fixed time for any positive α and β.*

5.3 Simulations

In this section, we discretize the algorithms (5.1) and (5.19) using Euler's method and apply them to two objective functions to demonstrate the convergence behavior established in Theorems 1 and 2. The step sizes are selected as 0.06 and 0.001 through trial and error.

5.3.1 Strictly Convex Case

Define some three-dimensional convex functions

$$f_i(x) = \frac{1}{4}\sum_{j=1}^{3} b_{ij}|x_j - a_{ij}|^4, i = 1,2,...,8, \tag{5.52}$$

where all a_{ij} and all entries of the initial values $x(0) \in \mathbb{R}^{60}$ are drawn from independent standard normal distributions, b_{ij} are uniformly distributed random numbers from $[0,1]$, and Assumption

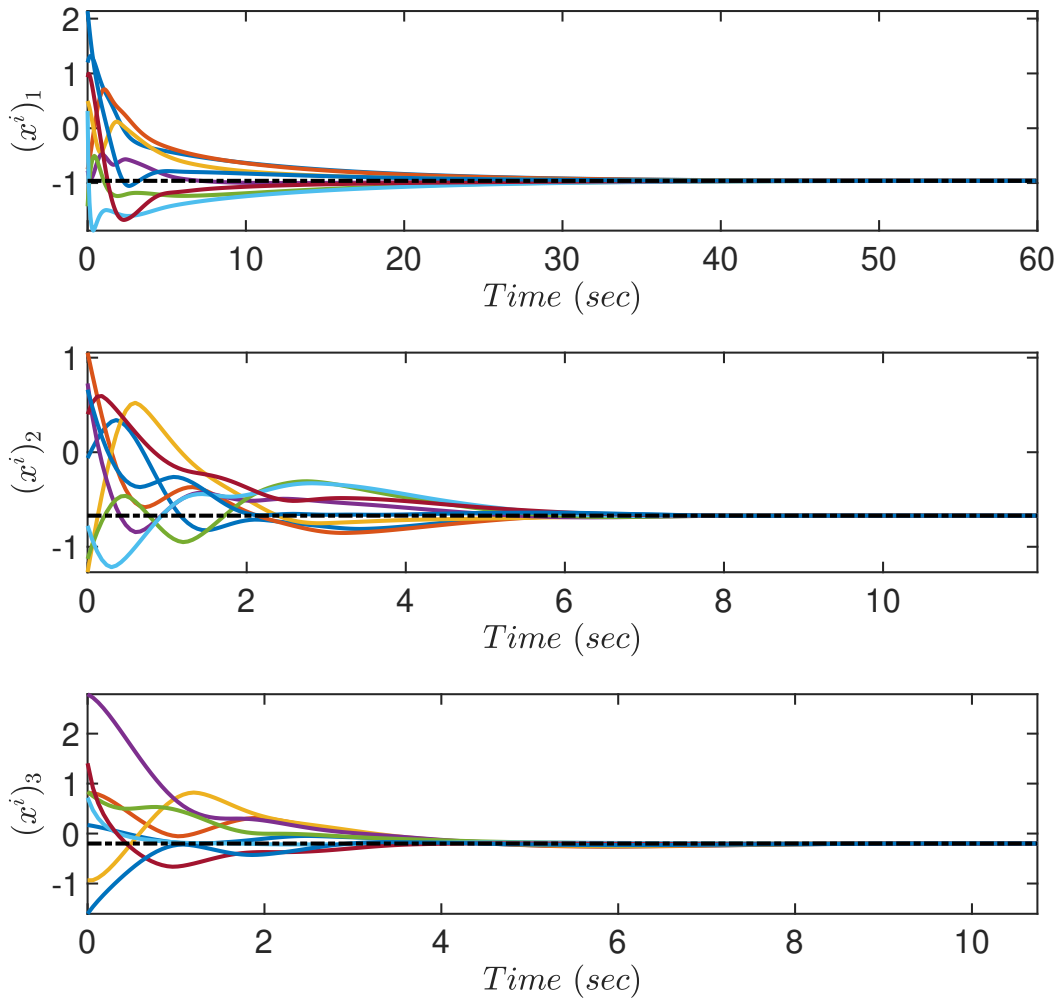

Figure 5.1: State trajectories of all agents using (5.1).

9 is satisfied. The graph G is a 8-cycle graph, that is to say, 8 agents are arranged into a cycle, and each agent is connected to its left and right agents.

Figures 5.3.1 and 5.3.1 show the evolutions of agents' states and the error of the total objective function in system (5.1), where $\alpha = \beta = 1$. We represent the optimal solution using dash-dot

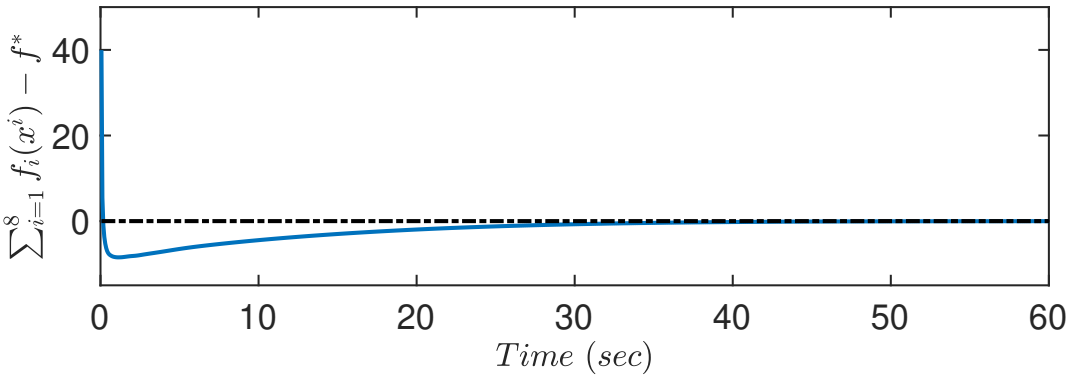

Figure 5.2: The error of the total objective function

lines. The three components of the optimal states are reached at about 35s, 7s and 8s, and the overall objective function is minimized at about 40s, which illustrate the finite-time behaviour guaranteed by Theorem 11.

5.3.2 Strongly Convex Case

Define some three-dimensional μ-strongly convex and M-smooth functions

$$f_i(x) = \frac{1}{2}\sum_{j=1}^{3} |x_j - a_{ij}|^2, i = 1,2,...,8, \tag{5.53}$$

where all entries of a_{ij} are drawn from independent standard normal distributions, and all entries of $x(0)$ and $v(0)$ are drawn from independent normal distributions with standard deviations of 10^{ν} ($\nu = 0,2,4,6$) and means of 0. The graph, G, is the same as before

Figure 5.3.2 shows the time evolutions of agents' states in system (5.19), where $\alpha = \beta = 1$, and $\nu = 0$. They converge in a finite-time, and it agrees with the analysis in Section V.

Figures 5.3.2 and 5.3.2 illustrate the time history of the distances between x and x^*, and the errors of the Lyapunov functions under different ν. Not surprisingly, the longer the distance between $x(0)$ and $\mathbf{1}\otimes x^*$, the more time it takes to reach the minimizer. In the plots, the distance between $x(0)$ and x^* would roughly increase by 100 times as ν increases by 2; however, the settling time grows slower. It is because the upper bound of the settling time is independent of the initial value, as shown in Theorem 12.

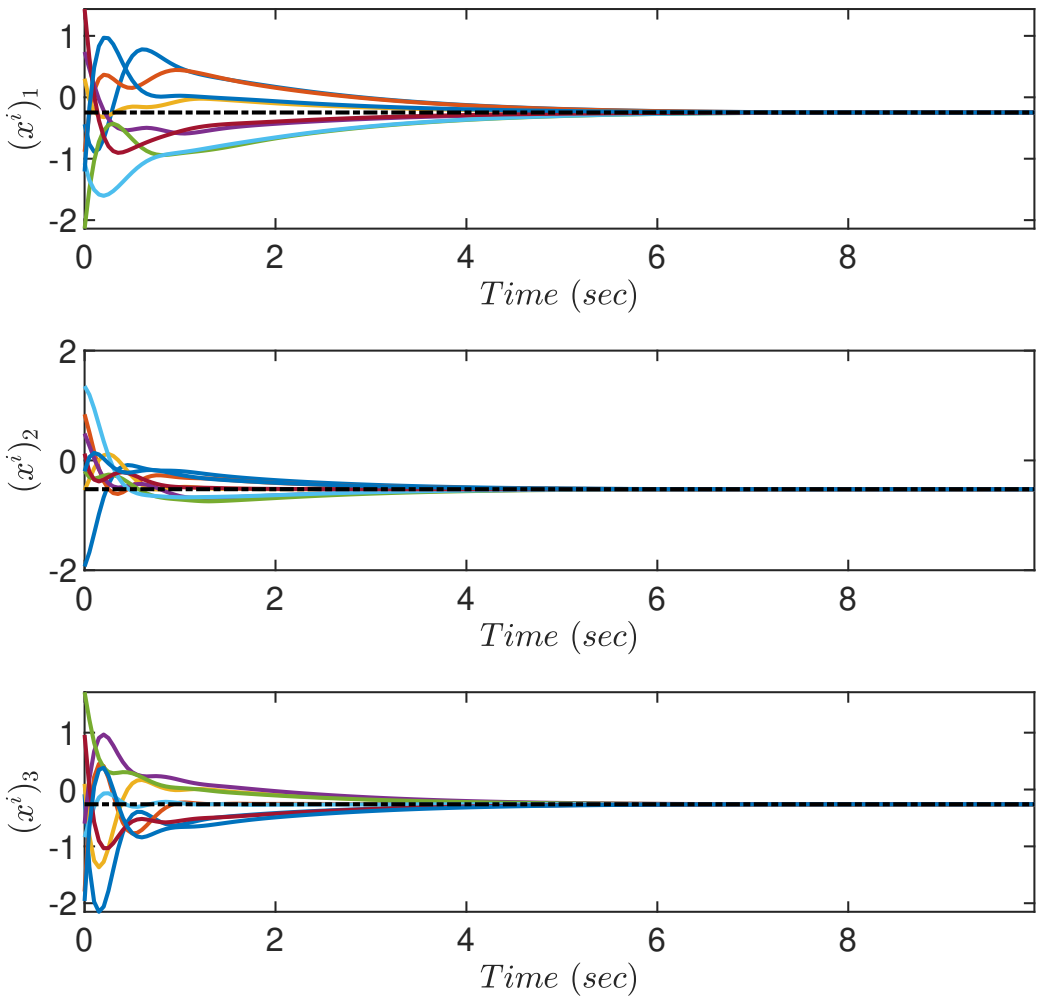

Figure 5.3: State trajectories of all agents using (5.19).

5.4 Summary

Combining the PI algorithm (4.2) from the last section with the powered function in [91, 92], we propose a finite-time distributed optimization algorithm (5.1). We are also motivated by some fixed-time optimization algorithm [12, 24, 41], and construct a fixed-time distributed optimiza-

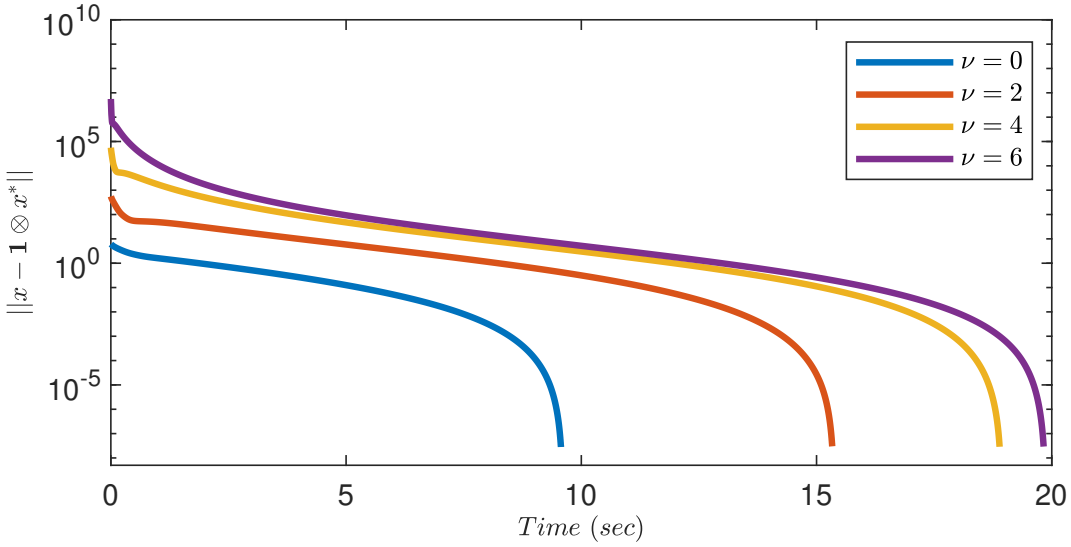

Figure 5.4: The distance between the state and the optimal solution.

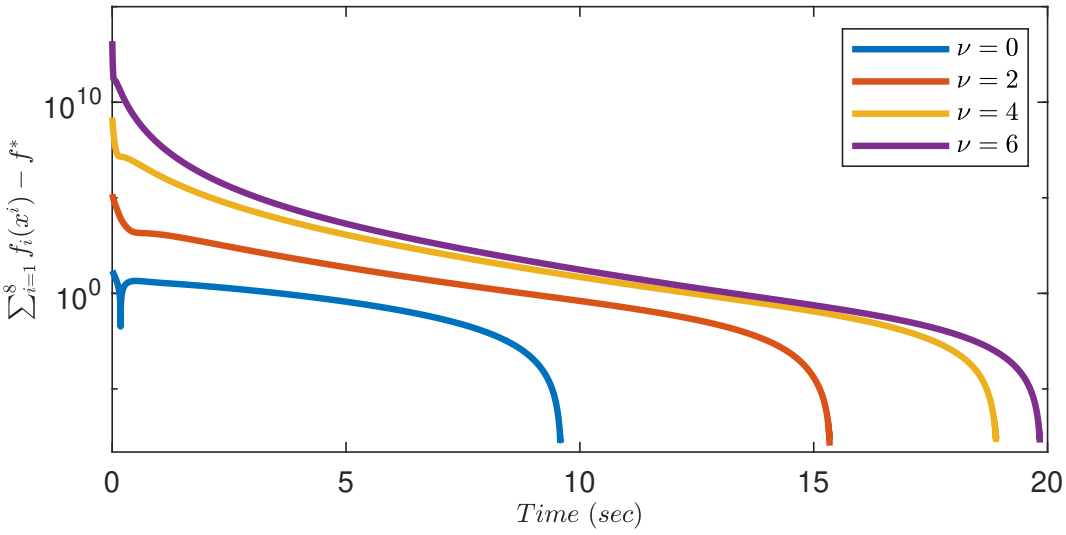

Figure 5.5: The errors of the total objective functions.

tion algorithm (5.19). Similar to the last chapter, instead of directly discussing the behavior of the algorithms, we build auxiliary systems by differentiating the original system first, and then establishing their connections.

The chapter demonstrates that the powered PI algorithm can converge within a finite time when applied to strictly convex functions. Another crucial aspect is the confirmation of fixed-time convergence when the algorithm is applied to strongly convex functions. This implies a consistent and predictable convergence timeline regardless of the initial conditions, enhancing the algorithm's practicality and reliability.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis has presented a comprehensive analysis of almost sure convergence rates for ASG methods, specifically RMSprop, Adadelta, Adam, Nadam, and AMSgrad. By offering a clear and rigorous proof of almost sure convergence to a critical point, this work contributes significantly to the field of machine learning optimization. For non-convex objective functions, our findings reveal that a weighted average of squared gradient norms achieves a unified convergence rate of $O(\frac{1}{t^{\frac{1}{2}-\theta}})$ for all $\theta \in (0, \frac{1}{2})$ Additionally, for strongly convex objectives, we have improved the convergence rates for RMSprop and Adadelta to $O(\frac{1}{t^{1-\theta}})$ within the same range of θ.

The Hogwild! algorithm, a locking-free parallel stochastic gradient descent method, was also analyzed under various loss function assumptions. We demonstrated its almost sure convergence rate for strongly convex functions, matching the optimal rate of classical SGD methods. For non-convex functions, we proved convergence to zero for both a weighted average of squared gradients and the algorithm's last iterations. We further extended our analysis to general convex smooth functions.

A new algorithm is propsed in chapter 4 for solving the distributed problem. First, with the assumptions of μ-strongly convex cost optimization function and connected topology, we propose a gradient-based algorithm whose convergence rate matches that of the centralized gradient method. Second, we prove the local exponential convergence behavior of this algorithm for strictly convex cost functions.

In chapter 5, we discuss algorithms for solving finite-time and fixed-time distributed optimization problems. We propose variants of the PI algorithm specifically designed for strictly

convex and strongly convex cost functions. These algorithms not only solve the distributed optimization problem but also guarantee finite and fixed settling times.

6.2 Future Work

Several avenues for future research emerge from this study:

- Improving the almost sure convergence rates for ASG methods such as Adam, AMSgrad, and Nadam presents a compelling challenge. These ASG methods frequently outperform SGD in practical applications, especially in deep learnin. Despite their practical success, the almost sure convergence rates for ASG methods, as discussed in Chapter 2, align with those of SGD noted in [45]. This observation leads to the anticipation of a more efficient convergence rate for ASG methods.

- Exploring the underlying principles and the physical intuition of ASG methods and understanding why they often outperform SGD is a fascinating topic.

  To gain a better insight of centralized Nesterov's Accelerated Gradient Method and Heavy-ball method, some papers utilize the second order ODE [69, 75, 96] derived by taking limit of these methods. Taking advantage of numerous tools from the field of differential equations and stochastic calculus, some papers use stochastic differential equations (SDEs) of SGD to gain new insights about non-trivial phenomena in non-convex optimization.

  In chapter 2, we have derived the limiting ODEs for ASG methods, providing a framework to examine the flow dynamics inherent to these approaches. A logical extension of this work involves delving into the corresponding Stochastic Differential Equations (SDEs) of ASG. Such an analysis could unveil deeper insights into the mechanisms, potentially revealing how ASG methods navigate the non-convex optimization with notable efficiency.

- If the cost function exhibits sufficient smoothness, [96] demonstrate that acceleration of Nesterov can be attained through a stable discretization of the ODE by employing conventional Runge-Kutta integrators. This insight suggests that applying a similar discretization strategy to the SDGs of ASG methods could potentially yield a more rapid approach.

- An interesting avenue for future research would be to explore whether the proof approaches in chapters 4 and 5 can be extended to continuous-time distributed stochastic gradient-based optimization algorithms.

References

[1] Hisham Abou-Kandil, Gerhard Freiling, Vlad Ionescu, and Gerhard Jank. *Matrix Riccati Equations in Control and Systems Theory*. Birkhäuser, 2012.

[2] Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

[3] Radu Balan. An extension of barbashin-krasovski-lasalle theorem to a class of nonautonomous systems. *arXiv preprint math/0506459*, 2005.

[4] Anas Barakat and Pascal Bianchi. Convergence and dynamical behavior of the adam algorithm for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 31(1):244–274, 2021.

[5] Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer, 2006.

[6] Dimitri P Bertsekas, Angelia Nedić, and Asuman E Ozdaglar. *Convex analysis and optimization*. Athena Scientific, 2003.

[7] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[8] Julius R Blum. Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, pages 382–386, 1954.

[9] Sebastian Bock and Martin Weiß. A proof of local convergence for the adam optimizer. In *2019 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2019.

[10] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. *Advances in neural information processing systems*, 20, 2007.

[11] Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Now Publishers Inc, 2011.

[12] Gang Chen and Zhiyong Li. A fixed-time convergent algorithm for distributed convex optimization in multi-agent systems. *Automatica*, 95:539–543, 2018.

[13] Weisheng Chen and Wei Ren. Event-triggered zero-gradient-sum distributed consensus optimization over directed networks. *Automatica*, 65:90–97, 2016.

[14] Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for rmsprop and adam in non-convex optimization and an empirical comparison to nesterov acceleration. *arXiv preprint arXiv:1807.06766*, 2018.

[15] Christopher M De Sa, Ce Zhang, Kunle Olukotun, and Christopher Ré. Taming the wild: A unified analysis of hogwild-style algorithms. *Advances in neural information processing systems*, 28, 2015.

[16] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale distributed deep networks. *Advances in neural information processing systems*, 25, 2012.

[17] Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

[18] Timothy Dozat. Incorporating nesterov momentum into adam. *ICLR 2016 workshop*, 2016.

[19] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[20] Paul Erdős and Alfréd Rényi. On random graph i. *Publ. Math. Debrecen*, 6:290–297, 1959.

[21] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

[22] Zhi Feng and Guoqiang Hu. Finite-time distributed optimization with quadratic objective functions under uncertain information. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 208–213. IEEE, 2017.

[23] Zhi Feng, Guoqiang Hu, and Christos G Cassandras. Finite-time distributed convex optimization for continuous-time multiagent systems with disturbance rejection. *IEEE Transactions on Control of Network Systems*, 7(2):686–698, 2019.

[24] Kunal Garg, Mayank Baranwal, Alfred O Hero, and Dimitra Panagou. Fixed-time distributed optimization: Consistent discretization, time-varying topology and non-convex functions. *arXiv preprint arXiv:1905.10472*, 2019.

[25] Bahman Gharesifard and Jorge Cortés. Distributed continuous-time convex optimization on weight-balanced digraphs. *IEEE Transactions on Automatic Control*, 59(3):781–786, 2013.

[26] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning, volume 1, 2016.

[27] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[28] Ge Guo and Jian Kang. Distributed optimization of multiagent systems against unmatched disturbances: A hierarchical integral control framework. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(6):3556–3567, 2021.

[29] Zhijun Guo and Gang Chen. Distributed zero-gradient-sum algorithm for convex optimization with time-varying communication delays and switching networks. *International Journal of Robust and Nonlinear Control*, 28(16):4900–4915, 2018.

[30] Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the adam family. *arXiv preprint arXiv:2112.03459*, 2021.

[31] Christopher RH Hanusa and Thomas Zaslavsky. Determinants in the kronecker product of matrices: The incidence matrix of a complete graph. *Linear and Multilinear Algebra*, 59(4):399–411, 2011.

[32] Yigwruang Hong, Jie Huang, and Yangsheng Xu. On an output feedback finite-time stabilization problem. *IEEE Transactions on Automatic Control*, 46(2):305–309, 2001.

[33] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge University Press, 2012.

[34] Zilun Hu and Jianying Yang. Distributed finite-time optimization for second order continuous-time multiple agents systems with time-varying cost function. *Neurocomputing*, 287:173–184, 2018.

[35] Kun Huang and Shi Pu. Improving the transient times for distributed stochastic gradient methods. *IEEE Transactions on Automatic Control*, 2022.

[36] Hassan K. Khalil. *Nonlinear systems*. Prentice Hall, Upper Saddle River, NJ, 3rd ed. edition, 2002.

[37] Solmaz S Kia, Jorge Cortés, and Sonia Martínez. Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication. *Automatica*, 55:254–264, 2015.

[38] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[39] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[40] Istvan Kovacs, Daniel S Silver, and Susan G Williams. Determinants of commuting-block matrices. *The American Mathematical Monthly*, 106(10):950–952, 1999.

[41] Chaojie Li, Xinghuo Yu, Xiaojun Zhou, and Wei Ren. A fixed time distributed optimization: A sliding mode perspective. In *IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*, pages 8201–8207. IEEE, 2017.

[42] Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd international conference on artificial intelligence and statistics*, pages 983–992. PMLR, 2019.

[43] Peng Lin, Wei Ren, and Jay A Farrell. Distributed continuous-time optimization: nonuniform gradient gains, finite-time convergence, and convex constraint set. *IEEE Transactions on Automatic Control*, 62(5):2239–2253, 2016.

[44] Jiayun Liu, Weisheng Chen, and Hao Dai. Distributed zero-gradient-sum (zgs) consensus optimisation over networks with time-varying topologies. *International Journal of Systems Science*, 48(9):1836–1843, 2017.

[45] Jun Liu and Ye Yuan. On almost sure convergence rates of stochastic gradient methods. In *Conference on Learning Theory*, pages 2963–2983. PMLR, 2022.

[46] Yang Liu, Zicong Xia, and Weihua Gui. Multi-objective distributed optimization via a predefined-time multi-agent approach. *IEEE Transactions on Automatic Control*, 2023.

[47] Jie Lu and Choon Yik Tang. Zero-gradient-sum algorithms for distributed convex optimization: The continuous-time case. *IEEE Transactions on Automatic Control*, 57(9):2348–2354, 2012.

[48] Qingguo Lü, Huaqing Li, and Dawen Xia. Distributed optimization of first-order discrete-time multi-agent systems with event-triggered communication. *Neurocomputing*, 235:255–263, 2017.

[49] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[50] Richard K Miller and Anthony N Michel. *Ordinary differential equations*. Academic press, 2014.

[51] Aryan Mokhtari, Qing Ling, and Alejandro Ribeiro. An approximate newton method for distributed optimization. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2959–2963. IEEE, 2015.

[52] Aryan Mokhtari, Qing Ling, and Alejandro Ribeiro. Network newton distributed optimization methods. *IEEE Transactions on Signal Processing*, 65(1):146–161, 2016.

[53] Michael Muehlebach and Michael Jordan. A dynamical systems perspective on nesterov acceleration. In *International Conference on Machine Learning*, pages 4656–4662. PMLR, 2019.

[54] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[55] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[56] Lam M Nguyen, Phuong Ha Nguyen, Peter Richtárik, Katya Scheinberg, Martin Takác, and Marten van Dijk. New convergence aspects of stochastic gradient algorithms. *Journal of Machine Learning Research 20*, 2019.

[57] Reza Olfati-Saber and Richard M Murray. Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on Automatic Control*, 49(9):1520–1533, 2004.

[58] Mihaela Oprea. Applications of multi-agent systems. In *Information Technology: Selected Tutorials*, pages 239–270. Springer, 2004.

[59] Alessandro Pilloni, Alessandro Pisano, Mauro Franceschelli, and Elio Usai. A discontinuous algorithm for distributed convex optimization. In *2016 14th International Workshop on Variable Structure Systems (VSS)*, pages 22–27. IEEE, 2016.

[60] Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

[61] Guannan Qu and Na Li. Accelerated distributed nesterov gradient descent. *IEEE Transactions on Automatic Control*, 65(6):2566–2581, 2019.

[62] Salar Rahili and Wei Ren. Distributed continuous-time convex optimization with time-varying cost functions. *IEEE Transactions on Automatic Control*, 62(4):1590–1605, 2016.

[63] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *Advances in neural information processing systems*, 24, 2011.

[64] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

[65] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

[66] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[67] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

[68] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[69] Bin Shi, Simon S Du, Weijie Su, and Michael I Jordan. Acceleration via symplectic discretization of high-resolution differential equations. *Advances in Neural Information Processing Systems*, 32, 2019.

[70] Xinli Shi, Guanghui Wen, and Xinghuo Yu. Finite-time convergent algorithms for time-varying distributed optimization. *IEEE Control Systems Letters*, 2023.

[71] John R Silvester. Determinants of block matrices. *The Mathematical Gazette*, 84(501):460–467, 2000.

[72] Navjot Singh, Deepesh Data, Jemin George, and Suhas Diggavi. Sparq-sgd: Event-triggered and compressed communication in decentralized stochastic optimization. *arXiv preprint arXiv:1910.14280*, 2019.

[73] Yanfei Song and Weisheng Chen. Finite-time convergent distributed consensus optimisation over networks. *IET Control Theory & Applications*, 10(11):1314–1318, 2016.

[74] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *Advances in neural information processing systems*, 27, 2014.

[75] Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[76] Rasul Tutunov, Haitham Bou-Ammar, and Ali Jadbabaie. Distributed newton method for large-scale consensus optimization. *IEEE Transactions on Automatic Control*, 64(10):3983–3994, 2019.

[77] Eugene E Tyrtyshnikov. *A Brief Introduction to Numerical Analysis*. Springer Science & Business Media, 2012.

[78] Jing Wang and Nicola Elia. Control approach to distributed optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 557–561. IEEE, 2010.

[79] Long Wang and Feng Xiao. Finite-time consensus problems for networks of dynamic agents. *IEEE Transactions on Automatic Control*, 55(4):950–955, 2010.

[80] Long Wang and Feng Xiao. Finite-time consensus problems for networks of dynamic agents. *IEEE Transactions on Automatic Control*, 55(4):950–955, 2010.

[81] Xiangyu Wang and Guodong Wang. Distributed finite-time optimisation algorithm for second-order multi-agent systems subject to mismatched disturbances. *IET Control Theory & Applications*, 14(18):2977–2988, 2020.

[82] Xiangyu Wang, Guodong Wang, and Shihua Li. Distributed finite-time optimization for disturbed second-order multiagent systems. *IEEE Transactions on Cybernetics*, 2020.

[83] Xiangyu Wang, Guodong Wang, and Shihua Li. Distributed finite-time optimization for integrator chain multiagent systems with disturbances. *IEEE Transactions on Automatic Control*, 65(12):5296–5311, 2020.

[84] Xiangyu Wang, Wei Xing Zheng, and Guodong Wang. Distributed finite-time optimization of second-order multiagent systems with unknown velocities and disturbances. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[85] Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *The Journal of Machine Learning Research*, 21(1):9047–9076, 2020.

[86] David Williams. *Probability with martingales*. Cambridge university press, 1991.

[87] Zizhen Wu, Zhongkui Li, and Junzhi Yu. Designing zero-gradient-sum protocols for finite-time distributed optimization problem. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021.

[88] Yijing Xie and Zongli Lin. Global optimal consensus for multi-agent systems with bounded controls. *Systems & Control Letters*, 102:104–111, 2017.

[89] Tao Yang, Xinlei Yi, Junfeng Wu, Ye Yuan, Di Wu, Ziyang Meng, Yiguang Hong, Hong Wang, Zongli Lin, and Karl H Johansson. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019.

[90] Hao Yu and Tongwen Chen. A new zeno-free event-triggered scheme for robust distributed optimal coordination. *Automatica*, 129:109639, 2021.

[91] Ye Yuan, Mu Li, Jun Liu, and Claire Tomlin. On the powerball method: Variants of descent methods for accelerated optimization. *IEEE Control Systems Letters*, 3(3):601–606, 2019.

[92] Ye Yuan, Mu Li, and Claire Tomlin. On the powerball method. In *2017 29th Chinese Control And Decision Conference (CCDC)*, pages 86–91. IEEE, 2017.

[93] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[94] Huan Zhang, Cho-Jui Hsieh, and Venkatesh Akella. Hogwild++: A new mechanism for decentralized asynchronous stochastic gradient descent. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 629–638. IEEE, 2016.

[95] Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct runge-kutta discretization achieves acceleration. *Advances in neural information processing systems*, 31, 2018.

[96] Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct runge-kutta discretization achieves acceleration. *Advances in neural information processing systems*, 31, 2018.

[97] Jingzhao Zhang, César A Uribe, Aryan Mokhtari, and Ali Jadbabaie. Achieving acceleration in distributed optimization via direct discretization of the heavy-ball ode. In *2019 American Control Conference (ACC)*, pages 3408–3413. IEEE, 2019.

[98] Mengyao Zhang, Xinzhi Liu, and Jun Liu. Convergence analysis of a continuous-time distributed gradient descent algorithm. *IEEE Control Systems Letters*, 5(4):1339–1344, 2020.

[99] Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11127–11135, 2019.

APPENDICES

Appendix A

Picard's Theorem

Consider an nonautonomous system of ODE

$$\dot{x} = f(t,x), \tag{A.1}$$

where $f : D \to \mathbb{R}^n$ and D is a subset of $\mathbb{R} \times \mathbb{R}^n$. For any $(t_0, x_0) \in D$, system (A.1) and the initial condition together define the initial value problem (IVP):

$$\dot{x} = f(t,x),\ x(t_0) = x_0. \tag{A.2}$$

A function $x : J \to \mathbb{R}^n$ is said to be a solution of the IVP on an interval J containing t_0 if $x(t_0) = x_0$ and $\dot{x} = f(x(t))$ for all $t \in J$.

Definition 6. *Let D be a subset of $\mathbb{R}^n$. A function is said to be*

- *Lipschitz continuous (or Lipschitz) in x on D, if there exists a constant $L > 0$ such that*

$$|f(t,x) - f(t,y)| \leq L|x - y| \tag{A.3}$$

  for all (t,x) and (t,y) in D;

- *locally Lipschitz in x at $(t_0, x_0) \in D$, if there exists a neighborhood U of x_0 such that f is Lipschitz continuous in x on U;*

- *locally Lipschitz in x on D, if f is locally Lipschitz in x at every $(t_0, x_0) \in D$;*

- *globally Lipschitz in x, if f is Lipschitz in x on $D = \mathbb{R}^n$.*

Local Lipschitz continuity of a function can be checked by the followng proposition.

Proposition 5. *Let $D \subset \mathbb{R} \times \mathbb{R}^n$ be an open set and $f(t,x)$ be a function defined on D and valued in $\mathbb{R}^n$. Suppose that $\frac{\partial f}{\partial x}$ exists and is continuous on D. Then f is locally Lipschitz in x on D.*

The following theorem establishes that local Lipschitz continuity ensures a unique solution to the IVP in a neighborhood of the initial point.

Theorem 13. *(Picard's Theorem) Let $D \subset \mathbb{R} \times \mathbb{R}^n$ be an open set. Suppose that f is continuous on D and Lipschitz continuous in x on D. For any $(t_0, x_0) \in D$, there exists $a, b > 0$ such that the set*

$$S = \{(t,x) \in \mathbb{R} \times \mathbb{R}^n : |t - t_0| \leq a, \|x - x_0\| \leq b\} \tag{A.4}$$

is contained in D. Define

$$M = \max_{(t,x) \in S} \|f(t,x)\|, \; c = \min\left(a, \frac{b}{M}\right). \tag{A.5}$$

Then the initial value problem (A.2) has a unique solution defined on the interval $[t_0 - c, t_0 + c]$.

Without the Lipschitz condition, Picard's theorem cannot guarantee the uniqueness of the solution. Peano's theorem states that if $f(t,x)$ in (A.2) is continuous in a neighborhood of the initial point, then at least one solution exists locally around the initial point.

Theorem 14. *(Peano's Existence Theorem) Let $D \subset \mathbb{R} \times \mathbb{R}^n$ be an open set. Suppose that $f : D \to \mathbb{R}^n$ is continuous on D. For any $(t_0, x_0) \in D$, let $a > 0$ and $b > 0$ be chosen such that the set*

$$S = \{(t,x) \in \mathbb{R} \times \mathbb{R}^n : |t - t_0| \leq a, \|x - x_0\| \leq b\} \tag{A.6}$$

is contained in D. Define

$$M = \max_{t,x} \in S \|f(t,x)\|, \; c = \min\left(a, \frac{b}{M}\right). \tag{A.7}$$

Then the initial value problem (A.2) has a solution defined on the interval $[t_0 - c, t_0 + c]$.

Example. (Leaky Bucket) Consider the leaky bucket described by

$$\dot{h}(t) = -a \cdot \sqrt{h(t)}, \; h(t_0) = 0, \tag{A.8}$$

where $h(t)$ represents the height of the water in the bucket at any time t, and a is a positive constant that depends on factors such as the size of the hole in the bucket. The initial value means that the bucket is empty at time t_0.

It can be verified that

$$h(t) = \begin{cases} \frac{a^2}{4}(t-\alpha)^2, t < \alpha \leq t_0, \\ 0, t > \alpha, \end{cases} \tag{A.9}$$

is a solution to the IVP for any $\alpha \in (-\infty, t_0]$. Because $-a\sqrt{h}$ is not locally Lipschitz at $h = 0$, Picard's theorem does not hold in this context. As a result, the uniqueness of the solution to the differential equation fails in this scenario.

Appendix B

Stability and LaSalle's Invariance Principle

Consider an autonomous system of ODE

$$\dot{x} = f(x), \tag{B.1}$$

where $f : D \to \mathbb{R}^n$ and D is a subset of $\mathbb{R}^n$. For any $x_0 \in D$, system (B.1) and the initial condition together define the initial value problem (IVP):

$$\dot{x} = f(x),\ x(0) = x_0. \tag{B.2}$$

A function $x(\cdot, x_0) : J \to \mathbb{R}^n$ is said to be a solution of the IVP on an interval J containing t_0 if it meets (B.1) for all $t \in J$. In this chapter, we assume the solution is unique.

Suppose that there exists some $\bar{x} \in D$ such that $f(\bar{x}) = 0$, then $\bar{x}$ is called the equilibrium point (or equilibrium) of (B.1). We can always use a change of variable to transform a non-zero equilibrium point to a zero one. Therefore, we can assume that the origin is an equilibrium point without loss of generality.

Definition 7. *The equilibrium point $x = 0$ of (B.2) is said to be*

- *stable, if for every $\varepsilon > 0$, there exists some $\delta > 0$ such that*

$$\|x_0\| \leq \delta \implies \|x(t, x_0)\| \leq \varepsilon \tag{B.3}$$

  for all $t \geq 0$;

- *asymptotically stable, if it is stable and there exists some* $\rho > 0$ *such that*

$$\|x_0\| \leq \rho \implies \lim_{t\to\infty} x(t,x_0) = 0; \tag{B.4}$$

- *globally asymptotically stable, if it is stable and*

$$\lim_{t\to\infty} x(t,x_0) = 0 \tag{B.5}$$

  for all $x_0 \in \mathbb{R}^n$*;*

- *unstable, if it is not stable.*

A reliable technique for checking stability is through the Lyapunov stability theorem.

Theorem 15. *(Lyapunov Stability Theorem for Autonomous System) Let* $D \subset \mathbb{R}^n$ *be open and contain the origin. Suppose that* $x = 0$ *be an equilibrium point of* (B.2)*. Let* $V : D \to \mathbb{R}$ *be continuously differentiable and positive definite on D.*

- *If* $\dot{V}$ *is negative semidefinite, then* $x = 0$ *is stable.*
- *If* $\dot{V}$ *is negative definite, then* $x = 0$ *is asymptotically stable.*
- *If* $\dot{V}$ *is negative definite and V is radially unbounded (with* $D = \mathbb{R}^n$*), then* $x = 0$ *is globally asymptotically stable.*

A set $\Omega \subset D$ is said to be positively invariant if all solutions $x(t,x_0)$ starting from Ω stay in Ω for all $t \geq 0$. Similarly, a set $\Omega \subset D$ is said to be invariant if all solutions $x(t,x_0)$ starting from Ω stay in Ω for all t.

Another approach for checking stability is the Barbashin–Krasovskii-LaSalle Theorem, where LaSalle's Invariance Principle us used to check the asymptotic tendency.

Theorem 16. *(LaSalle's Invariance Principle) Suppose that* $V : D \to \mathbb{R}$ *is continuously differentiable. Let* $\Omega \subset D$ *be a positively invariant compact set of* (B.2)*. If* $\dot{V}(x) \leq 0$ *for all* $x \in \Omega$*. Then, for any* $x_0 \in \Omega$*, we have that everey solution starting in* Ω *approaches the largest invariant set in*

$$E = \{x \in \Omega : \dot{V} = 0\}. \tag{B.6}$$

Theorem 17. *(Barbashin–Krasovskii-LaSalle Theorem) Let* $D \subset \mathbb{R}^n$ *be open and contain the origin. Suppose that* $x = 0$ *be an equilibrium point of* (B.2)*. Let* $V : D \to \mathbb{R}$ *be continuously*

differentiable and positive definite on D. Suppose that $\dot{V}(x) \leq 0$ *for all* $x \in D$. *Let* $S = \{x \in D : \dot{V} = 0\}$. *Suppose that no solution can stay identically in S except the trivial solution* $x = 0$. *Then* $x = 0$ *is asymptotically stable. If the above conditions hold with* $D = \mathbb{R}^n$ *and V is radially unbounded, then* $x = 0$ *is globally asymptotically stable.*

Remark 15. *In 2005, the Invariance Principle presented in [3, Theorem 2] extends the Barbashin-Krasovski-LaSalle Theorem to a class of non-autonomous Systems.*