Structure-Driven Algorithm Design in Optimization and Machine Learning

Tianyi Lin

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-71
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-71.html

May 8, 2023

Acknowledgement

I would like to express my heartfelt thanks to Mike, for his guidance through this journey. I only hope that I will be to my students as great of an advisor as Mike has been to me.

I benefited tremendously from many great teachers at Berkeley, especially the members of BAIR for the unique, friendly, and stimulating atmosphere.

The culture in Mike’s group is unique and innovative. I was encouraged by him to explore the different research directions, and enjoy the complete freedom to choose topics that interest me. I am also fortunate to have many amazing friends who are at Berkeley and who used to be at Berkeley.

I extend my heartfelt appreciation to my family for their unwavering support

during this journey. Their encouragement and belief in me kept me motivated and focused.

Structure-Driven Algorithm Design in Optimization and Machine Learning

by

Tianyi Lin

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael I. Jordan, Chair
Professor Peter L. Bartlett
Associate Professor Aditya Guntuboyina

Spring 2023

The dissertation of Tianyi Lin, titled **Structure-Driven Algorithm Design in Optimization and Machine Learning**, is approved:

Chair _____ Date _____

_____ Date _____

_____ Date _____

University of California, Berkeley

Structure-Driven Algorithm Design in Optimization and Machine Learning

Abstract

Structure-Driven Algorithm Design in Optimization and Machine Learning

by

Tianyi Lin

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Michael I. Jordan, Chair

A textbook property of optimization algorithms is their ability to solve the problems under generic regularity conditions. Two examples are simplex method and gradient descent (GD) method. However, the performance of these fundamental and general-purpose optimization algorithms is often unsatisfactory; they often run slowly and perhaps return the suboptimal solutions in generic settings. In my view, this is the *price of their generality*; indeed, the generic algorithms are an achievement, but for many problems, the gains from leveraging special structure can be huge. A basic question then arises: *how can we harness problem-specific structure within our algorithms to obtain fast, practical algorithms with strong performance guarantees?* As more structured data-driven decision-making models emerge, this question has become increasingly pressing and relevant to practitioners.

For example, the GD is known to get stuck at a suboptimal saddle points in nonconvex optimization. Nonetheless, a line of recent works have shown that random initialization or perturbation changes the dynamics of GD and makes it provably converge to a global optimal solution. In addition, both Markov decision process (MDP) and discrete optimal transport (OT) problems can be solved using large-scale linear programs. Rather than using generic LP algorithms, the policy iteration and the Sinkhorn iteration exploit special structures in MDP and OT and thus perform better in practice. Adapting algorithms to problem-specific structure is generally referred to as *structure-driven algorithm design*.

Although this line of research – which has been studied extensively for over 70 years – has enjoyed widespread success, the machine-learning success stories have introduced new formulations ripe for deep theoretical analysis and remarkable practical impact. My research pushes this frontier by identifying special structure of reliable machine learning (*minimax optimization*) and multi-agent machine learning (*high-order optimization and beyond*) and design optimal algorithms for computing the appropriately defined optimal solutions; and other structured problems, such as *efficient entropic regularized optimal transport*, *gradient-free nonsmooth nonconvex optimization*, and *adaptive and doubly optimal learning in games*.

To my family

Contents

List of Figures

List of Tables

Acknowledgments

About 10 years ago, while I was a research assistant at the Chinese University of Hong Kong (CUHK), I did my first research project on extending the classical topic model to analyze user-generated Web content and social media. I downloaded the monograph from Internet: *Graphical Models, Exponential Families, and Variational Inference*, and was impressed by Mike for his instrumental expertise and insights. Back then, I did not know that he is the highly respected and renowned pioneer in the field of machine learning, nor would I have imagined that 5 years later I will have the great fortune of having Mike as my doctoral advisor. I would like to express my heartfelt thanks to Mike, for his guidance and constant encouragement and support through this journey, as well as the many enjoyable discussions we had over the last 5 years. I only hope that I will be to my students as great of an advisor as Mike has been to me.

I benefited tremendously from many great teachers at Berkeley, especially the members of Berkeley artificial intelligence research (BAIR) lab for the unique, friendly, and stimulating atmosphere they have created. I am grateful to Laurent El Ghaoui and Martin Wainwright, who got me excited about various topics in convex optimization models and algorithms, and gave their time to discuss research problems with me. I thank Alper Atamturk and Shmuel Oren for linear and nonlinear optimization classes, and their very helpful advice. I thank Yi Ma for the class on low-dimensional models and Nika Haghtalab for introducing me to the intersection of machine learning and economics. Thanks to Ilan Adler for mentoring me in the early years and L. Craig Evans for two-semester wonderful course on the PDE theory. Zeyu Zheng showed me how to go from a graduate student to a professor and his pointers helped me navigate the academic job market without panic. Special thanks are due to Jim Pitman, who has been a great influence in turning my research interests towards the applications of probability theory in operations research and computer science. I sincerely thank him for his wonderful classes, his guidance and his inexhaustible enthusiasm. Peter Bartlett and Adytia Guntuboyinai, who helped by being on my thesis committee, have provided great feedback on my research and given good perspective. I would also like to thank Venkat Anantharam for being on my qualifying exam committee and providing many helpful discussions.

The culture in Mike's group (SAIL) is unique and innovative. I was encouraged by him to explore the different research directions, and enjoy the complete freedom to choose topics that interest me. It is this vibrant culture that creates an ideal environment for conducting high-quality research! Within SAIL, I have found several peer collaborators (and friends) who have made substantial impact on my research and life. In particular, Chi Jin served the dual roles of both a friend and mentor, and has been a great influence in shaping my research tastes for pursuing "simple" algorithms and elegant arguments. I was also really fortunate to collaborate with Nhat Ho, who deepened my knowledge of optimal transport and statistics, and Eric Mazumdar who taught me a lot about learning in games. Manolis Zampetakis and Emmanouil-Vasileios Vlatakis-Gkaragkounis provided me various guidance during my last two years and shared with me many funny stories about Greece. Nilesh Tripuraneni and I shared the precious memories of stimulating discussions about the academia and culture, and

Yaodong Yu showed me the importance of practical insights in deep learning. Outside SAIL, I would like to express deep gratitude for the discussion with Xi Chen, Marco Cuturi, George Lan, Panayotis Mertikopoulos, Wenpin Tang, Kaiqing Zhang, Jiawei Zhang and Zhengyuan Zhou, who shared with me their invaluable understandings on academia and until now they are my role models in being successful researchers.

In addition to all people mentioned above, I am fortunate to have many amazing friends who are at Berkeley and who used to be at Berkeley. I would like to express gratitude for the interaction with Tom Hu, who is not only a great collaborator but a great friend. I would like to thank my peer fellows in the EECS and Statistics departments: Anastasios Angelopoulos, Stephen Bates, Tatjana Chavdarova, Xiang Cheng, Melih Elibol, Paula Gradu, Wenshuo Guo, Meena Jagadeesan, Koulik Khamaru, Lihua Lei, Junchi Li, Lydia Liu, Romain Lopez, Aldo Pacchiano, Max Simchowitz, Neha Wadia, Serena Wang, Alex Wei and Tijana Zrnic. Special thanks are due to Wenlong Mou, Xiao Li, Yixuan Li, Feng Ruan and Zhiyi You for our endless debates that have made the life so much fun. I would also like to thank my peer fellows in the IEOR department: Eric Bertelli, Haoyang Cao, Junyu Cao, Yuhao Ding, Salar Fattahi, Han Feng, Hansheng Jiang, Yusuke Kikuchi, Anran Hu, Yundan Lin, Heyuan Liu, Alfonso Lobos, Pelagie Elimbi Moudio, Meng Qi, Xu Rao, Quico Spaen, Mark Velednitsky, Mengxin Wang, Jingxu Xu, Renyuan Xu, Nan Yang, Haoting Zhang, Jiacheng Zhang and Ruijie Zhou. Special thanks are due to Yuhao Ding and Haoting Zhang for providing great company as well as answering various questions. I thank many other peer fellows for making my time at Berkeley some of the best in my life. It was my great pleasure to meet you all!

I am greatly grateful to Shuzhong Zhang and Shiqian Ma, who introduced me to the field of optimization and guided me to take the first step. I would like to thank Hong Cheng for always trusting me and Yinyu Ye, whose big-picture thinking and papers play the pivot role in shaping my research style. I would like to thank my friends outside Berkeley for their encouragement. Special thanks are due to Wenjia Ba, Chenyou Fan, Chao Huang, Yan Qin, Jingchen Sun, Letian Wang, Yinqing Xu and Yi Yan for their support. Finally, I would like to thank numerous colleagues for their generous support during my job search. I am really looking forward to future times when our paths cross again!

Last but not least, I extend my heartfelt appreciation to my family for their unwavering support, love, and patience during this journey. Their constant encouragement and belief in me kept me motivated and focused.

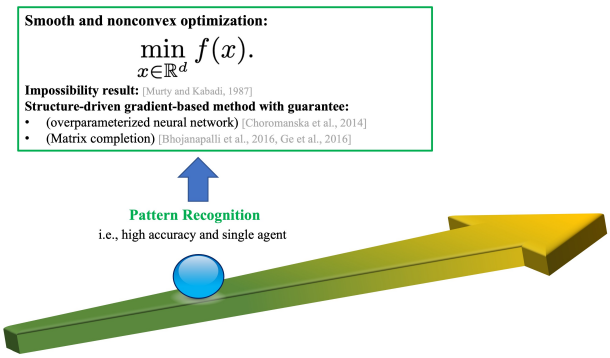
Chapter 1

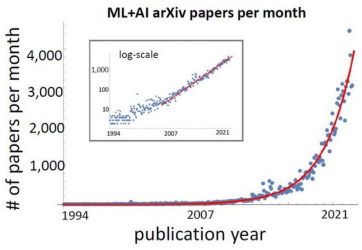
Introduction

A textbook property of optimization algorithms is their ability to optimize problems under generic regularity conditions. However, the performance of these fundamental and general-purpose optimization algorithms is often unsatisfactory; indeed, for many real problems, the gains from leveraging special structure can be huge. A basic question then arises: *how can we harness problem-specific structure within our algorithms to obtain fast, practical algorithms with strong performance guarantees?* Although this line of research – which has been studied extensively for over 70 years – has enjoyed widespread success, the recent reliable and/or multi-agent machine-learning success stories have introduced new formulations ripe for deep theoretical analysis and remarkable practical impact. My research that pushes this frontier can be summarized in two aspects: (i) *game-theoretic learning* where we identify some special structures and design simple and practical algorithms for computing reasonable solutions; and (ii) *distribution-based learning* where we seek the true data distribution from samples by introducing new formulations and designing structure-driven and sample-efficient algorithms.

1.1 Motivation

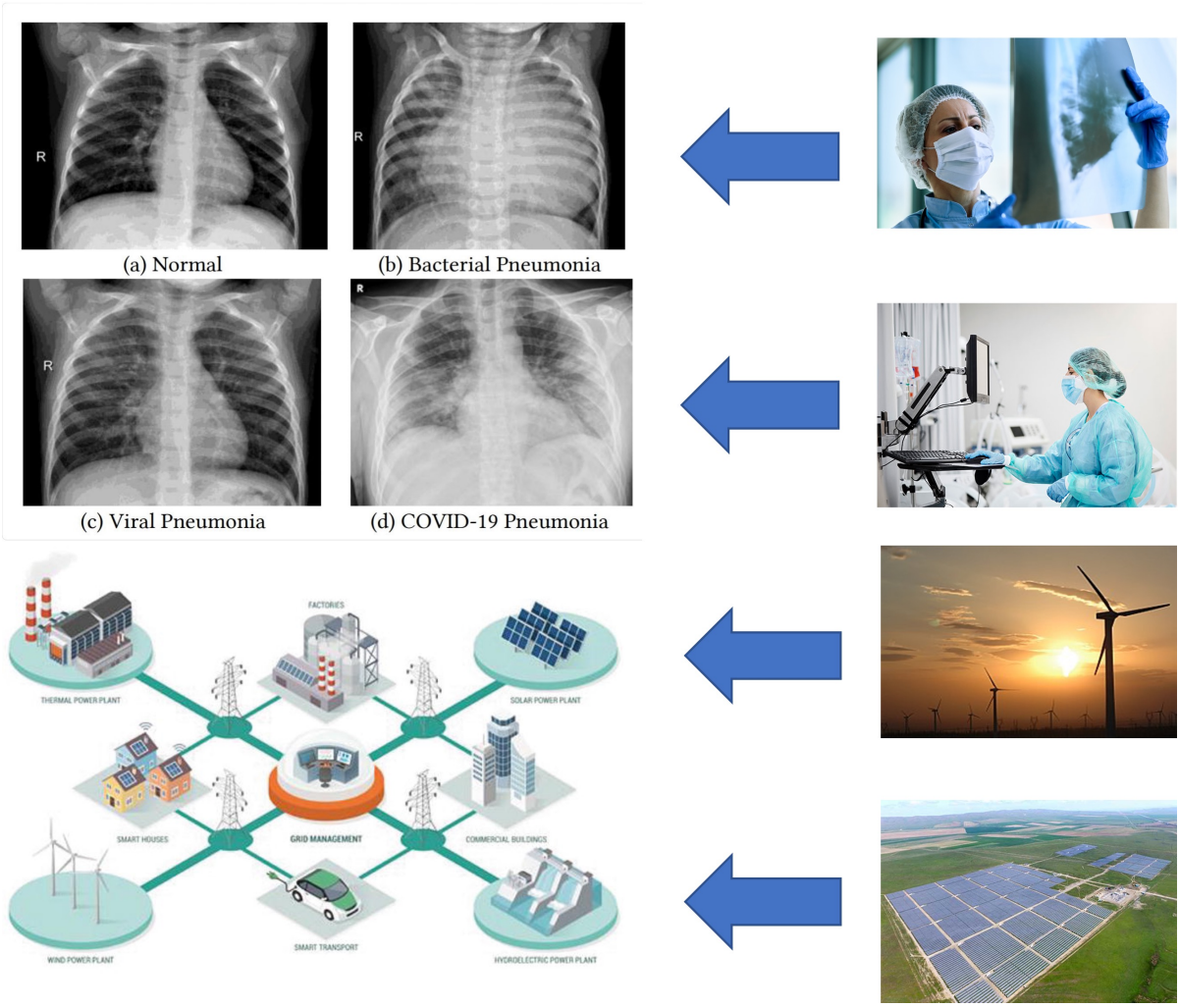
In optimization and machine learning, we often have structure about specific application problems – structure that we can potentially leverage to design and analyze specialized algorithms. For example, this is true for network flow problems [Ahuja et al., 1993], where the network simplex method is the method of choice in practice. Although the simplex method is not polynomial in general, the network simplex method is even strongly polynomial [Goldfarb and Hao, 1992, Orlin et al., 1993, Orlin, 1997, Armstrong and Jin, 1997]. It is true for Markov decision process problems [Bellman, 2013, Puterman, 2014], where the policy iteration method is the state-of-the-art approach. This is also the variant of the simplex method but enjoys a theoretical guarantee for solving discounted MDP with a fixed factor [Ye, 2011]. It is also true for low-rank optimization problems, where we use nuclear-norm constraint set. The Frank-Wolfe method perfectly fits such structure [Jaggi, 2013] and produces low-rank solutions with a finite-time guarantee [Freund and Grigas, 2016, Freund et al., 2017].



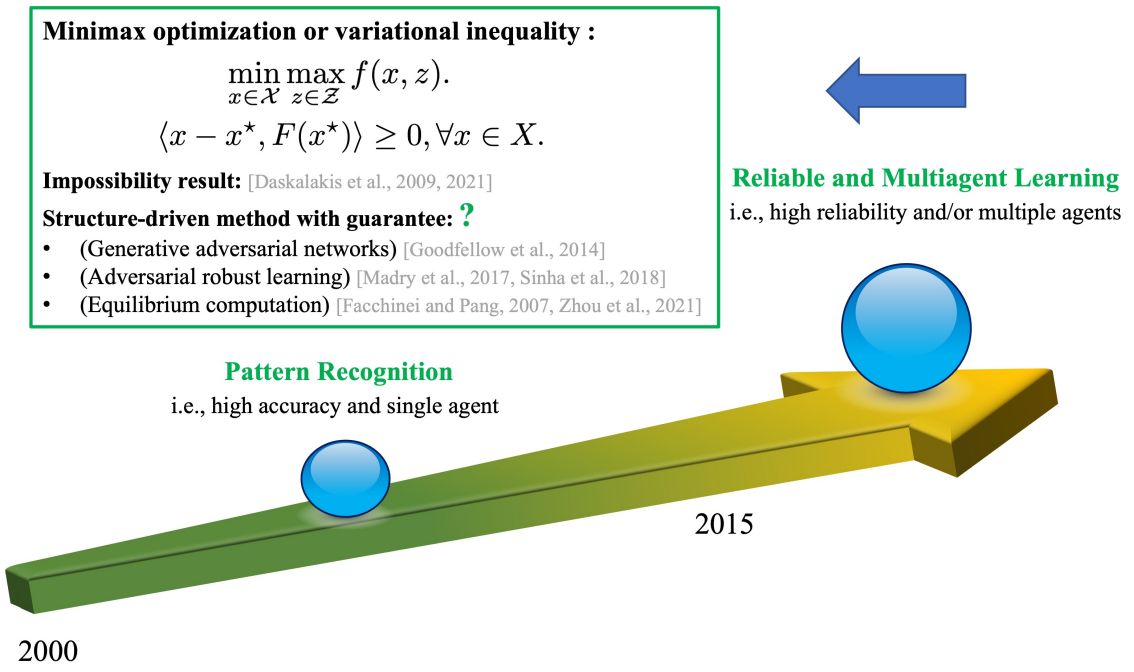


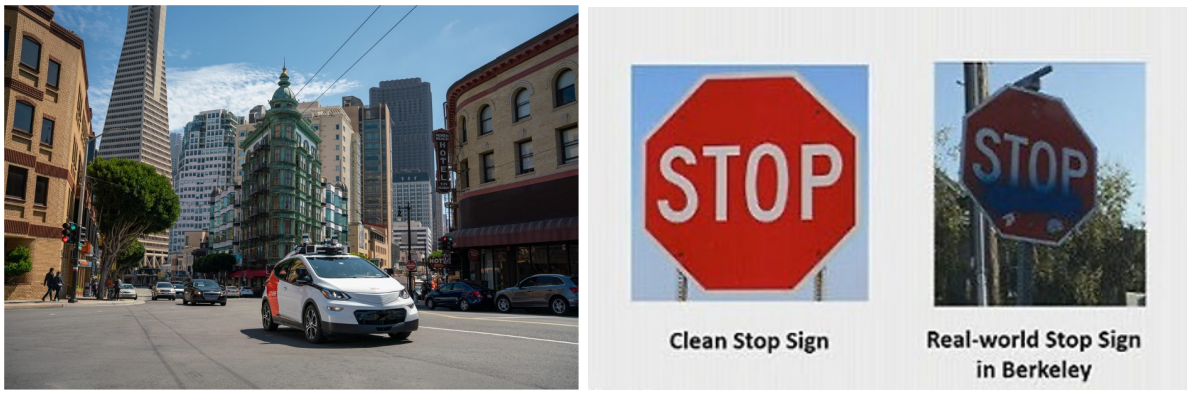
Beyond these classical problems, we have many other popular machine learning problems with special structures. Let us briefly review the trend of machine learning (ML) from an optimization viewpoint. The classical ML models are designed for pattern recognition [Bishop, 2006, Hastie et al., 2009], where the main goal is to do clustering and classification with high accuracy. These problems can be solved by smooth and nonconvex optimization toolbox. In this context, the impossibility results have stated that computing an approximate global solution is NP-hard [Murty and Kabadi, 1987]. However, many modern application problems, such as optimizing overparameterized neural networks [Choromanska et al., 2015] and matrix completion [Bhojanapalli et al., 2016, Ge et al., 2016], have special structure such that some specific forms of random initialization or random perturbation (but implementable) will change the gradient descent dynamics such that the generated iterates provably converge to a global optimal solution [Lee et al., 2019, Jin et al., 2021].

Such ML models and gradient-based methods have seen tremendous success in many application problems. In particular, the ML was used to decreasing student dropout, and evaluating applicants in college and graduate school admissions. The national institute of justice applied ML to address criminal justice needs, such as identifying individuals and their actions in image and video relating to criminal activity, DNA analysis, gunshot detection and crime forecasting. The U.S. department of transportation is also looking to increase public safety through developing and testing automatic traffic accident detection based on advanced ML systems. The ML was also being used in healthcare to interpret medical images or optimize queue systems, which could have important practical implications.

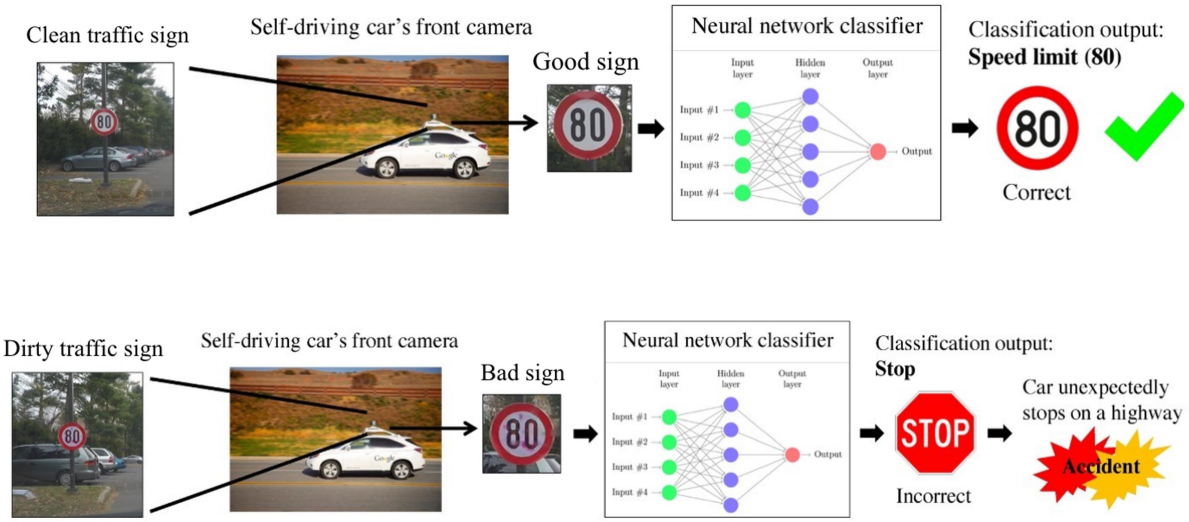


Reliable and/or multi-agent machine learning. One of the things I have learned over the years is that people become really, really excited about what ML can do and what ML can accomplish. From their viewpoint, there are so much benefits that come along with ML and they of it as, like look at that wonderful bridges that we can build. However, for someone who has conducted many experiments, we become suspicious about the ambitious goals that people are trying to achieve and feel that the current ML pipelines are really, really rickety. Maybe you can still get to the end but have to be really, really careful and skeptical. It is really none of the turkey solution yet. Historically, the ML tradition is defined by top conference publications and this is called paper-centric viewpoint: we are given the fixed dataset and we want to beat the benchmarks. As soon as we beat the benchmark, we declare the success. However, when we move to real application and use ML in production, the problem becomes more and more involved because of uncertainties and incentives from the environment. So moving from this so-called paper-centric viewpoint to a production system viewpoint, we need to be careful about the outcome of our ML models in practice.





To better illustrate the limitations of current ML models, we look at the concrete example from healthcare application [Zech et al., 2018]. What they do is to take the X-ray images and hope to detect the diagnose whether the patient has the pneumonia or not. The common observation is that the performance beyond training data degrades dramatically and the possible reason is that every image has its own mark, which varies from hospital to hospital and actually depends on other factors, such as the doctor who takes the X-ray and even the scanner which is used by the hospital. The embarrassing fact is that ML models effectively learn such hospital-specific pattern, which is the leakage that does not generalize well. Another example comes from power management, which allows the users of a wireless network to achieve their performance requirements while minimizing the power consumed by their equipment. This problem has long been the core aspect of network design and can be tackled by finding an unique equilibrium of a suitably designed continuous game [Zhou et al., 2021]. Moreover, in real scenario, the U.S. power grid has around 170,000 miles of



high-voltage transmission lines and almost 5000 generating units with capacity of at least 50 MW. In a system of such a scale and complexity, the key driving force behind uncertainty is the large-scale integration of wind and solar power generation into grids, both of which are highly stochastic in nature [Bertsimas et al., 2012, Sun and Conejo, 2021].

Minimax optimization, variational inequalities (VIs) and beyond. This leads to a natural transition from classical ML to reliable and multi-agent ML, where the main goal is to do multi-agent learning with high reliability. The basic setup changes from single-agent optimization model to minimax optimization model or VI model. In this context, the recent impossibility results have stated that determining whether or not an approximate global solution exists in general minimax optimization model is NP-hard [Daskalakis et al., 2021] and computing an approximate Nash equilibrium in general multi-player games (i.e., VI models)is PPAD-complete [Chen et al., 2009, Daskalakis et al., 2009]. In contrast, many modern application problems, such as generative adversarial networks (GANs) [Goodfellow et al., 2014], adversarial robust learning models [Madry et al., 2018, Sinha et al., 2018] and equilibrium computation [Facchinei and Pang, 2007], have introduced new formulations ripe for deep theoretical analysis and remarkable practical impact.

A typical example is autonomous driving. In particular, each self-driving car has multiple cameras at every angle to provide a perfect view of its surroundings and the detection accuracy would be important for self-driving cars to make the decisions. In the real scenario, the input for machine learning systems in self-driving car is often dirty. Such dirty sign could confuse self-driving cars and cause their ML systems to incorrectly classify signs, potentially putting the lives of passengers in danger. For example, the speed limit sign with strains fooled a self-driving car's ML system into classifying "speed limit 80" as "stop" by mistake, leading to an unexpected stop on a highway. The idea to make machine learning models reliable comes from robust and adaptive robust optimization [Ben-Tal et al., 2009, Bertsimas

et al., 2011] and is that we let our models to do well for all perturbed data using minimax optimization model [Madry et al., 2018, Sinha et al., 2018, Bertsimas et al., 2021].

The VI models have captured a wide range of problems in economics, operations research and computer science, including Nash equilibrium problems, Nash-Cournot production problems, oligopolistic electricity models, Markov perfect equilibrium models, economic equilibrium problems, Walrasian equilibrium models, invariant capital stock model, traffic equilibrium models, frictional contact problems and option pricing problems. This research has provided the foundation for work in machine learning in recent years, where general equilibrium problems have emerged in many real-world settings [Cesa-Bianchi and Lugosi, 2006, Jordan, 2018].

1.2 Overview of Our Results

During my Ph.D study, I have developed several structure-driven algorithms that can be used across various domains. For *game-theoretic learning*, I have studied minimax optimization for zero-sum, two-player games [Lin et al., 2020c,d, Jordan et al., 2022a, Lin et al., 2022c,d], highly smooth optimization, inclusions and variational inequalities (VI) for general-sum, multi-player games [Lin and Jordan, 2022a,b,c, 2023] and online and bandit learning in games [Lin et al., 2020e, Jordan et al., 2022c]. For *distribution-based learning*, most of my works have focused on optimal transport (OT) [Lin et al., 2019a, 2020a,b, 2021a, 2022a,b]. Some of the proposed structure-driven algorithms in these domains were the first to achieve optimal convergence guarantees for solving their corresponding problems [Lin et al., 2020d, Jordan et al., 2022a,c, Lin and Jordan, 2022b, Lin et al., 2022d] while others have been recognized as the state-of-the-art approaches in practice [Lin et al., 2019a, 2020a,c, 2022b,c]. In addition, I have studied some other problems, such as nonconvex nonsmooth optimization with Lipschitz objective functions [Lin et al., 2022f, Jordan et al., 2022b], and online and bandit nonsubmodular learning with delayed costs [Lin et al., 2022e].

In these papers, I leveraged classical techniques from optimization and variational analysis, and adapted them to the special structures arising in modern machine-learning problems. Concrete examples of structures include the *asymmetry* of players in min-max optimization, the *high-order Lipschitz continuity* in optimization, inclusions and VIs, the *montonicity* in multi-agent learning and the *low-dimensional* substructure in OT. The key challenge for structure-driven algorithm design, is identifying the algorithmic component that pairs to the problem-specific structure to obtain strong theoretical and practical performance.

Minimax optimization. In recent years, minimax optimization theory has begun to see applications in operations research and machine learning, with examples including generative adversarial networks (GANs), distributionally robust optimization (DRO) and online learning. Furthermore, learning in economic systems has increased the demand for algorithms to compute minimax optima and related equilibrium concepts. However, these new problem formulations are in essence nonconvex problems, where existing algorithms (e.g., gradient descent ascent (GDA) and extragradient method) run slowly and may not even converge in theory and practice.

We proposed to solve nonconvex-concave minimax problems using two-timescale GDA and proved the first nonasymptotic convergence rate guarantee to a Stackelberg equilibrium. The key observation is that nonconvex-concave minimax problems can be viewed as zero-sum games with asymmetrical players, motivating two-timescale rules: the nonconvex player is conservative with small stepsize while the concave player is somehow aggressive with large stepsize [Lin et al., 2020c]. In the extended version, we studied the effect of nonsmoothness on nonconvex-concave minimax problems and provided a more refined treatment of two-timescale GDA [Lin et al., 2022c].

We resolved a longstanding open question pertaining to the design of near-optimal first-order algorithms for minimax problems with asymmetrical players. We were the first to highlight the fundamental role that accelerated proximal point method played in optimal minimax optimization algorithm design [Lin et al., 2020d].

We answered an open conjecture about the performance gap of manifold extragradient (EG) methods. Our results showed that manifold GDA and EG achieved optimal convergence rates for smooth, nonsmooth and stochastic minimax optimization in geodesic metric spaces up to curvature factors [Jordan et al., 2022a].

We proposed exact and inexact regularized Newton-type methods for solving the convex-concave unconstrained min-max optimization problems. This is the first optimal convergence rate estimate for second-order methods in this setting [Lin et al., 2022d].

Highly smooth optimization and beyond. Optimization and variational inequalities capture a wide range of problems in optimization theory and beyond, including optimization problems, saddle-point problems and models of equilibria in games. While optimal first-order methods have been extensively studied in monotone setting, the investigations of optimal second-order and high-order methods are relatively rare, as exploiting the high-order derivative information is much more involved for algorithm design. In a series of recent works, Michael. I. Jordan and I have designed novel high-order methods for finding one solution at an optimal global rate in convex optimization [Lin and Jordan, 2022b], monotone equation [Lin and Jordan, 2022a], monotone VI [Lin and Jordan, 2022c] and a more general monotone inclusion [Lin and Jordan, 2023].

By appealing to a novel Lyapunov approach, we demonstrate the fundamental role that the closed-loop control and rescaled dynamical systems play in optimal acceleration and the clear advantage that the continuous-time perspective brings to algorithmic design.

Online and bandit game-theoretic learning. In the domains where the environment is rapidly changing or even adversarial, single-agent online learning offers useful tools for making sequential decisions agnostically. However, the environment consists of other agents that are engaged in online decision-making, with each agent's action impacting outcomes for others. Such interactions make predicting any agent's action difficult. While no-regret online learning algorithms help each agent maximize its transient performance (characterized by regret), the long-run behavior is ultimately determined by the equilibrium outcome: the two can be at odds with each other. For example, no-regret learning can converge to strictly dominated strategies in finite games.

We studied multi-agent learning via OGD in cocoercive games, which admitted multiple Nash equilibria and properly included unconstrained strongly monotone games. Indeed, we filled in several gaps, where three aspects – finite-time convergence guarantee, non-decreasing step-sizes, and fully adaptive algorithms – have been unexplored before [Lin et al., 2020e].

We also studied online no-regret learning in strongly monotone games with noisy gradient feedback. An important application is a learning version of newsvendor problem, where due to lost sales, only noisy gradient feedback can be observed and all problem parameters are unknown. Combining online gradient descent (OGD) with a structure-driven randomization, we designed the first fully adaptive and doubly optimal algorithm for both single-retailer and multi-retailer settings [Jordan et al., 2022c].

Efficient optimal transport (OT). OT – the problem of finding minimal cost couplings between pairs of probability distributions – has recently been used to learn the true data distribution from samples in numerous machine learning applications. The key challenge is computational and a new literature has begun to emerge to provide new algorithms for OT.

We proved a tight complexity bound for the greedy Sinkhorn algorithm, helping explain why such algorithm often outperformed the Sinkhorn algorithm in practice. By appealing to a novel primal-dual formulation of OT, we designed a new class of algorithms with theoretical guarantees [Lin et al., 2019a]. In the extended version, we investigated the structure of OT and proposed to accelerate Sinkhorn using estimate sequences [Lin et al., 2022b]. We continued to push structure-driven algorithm design for variants of OT, including multimarginal OT problem [Lin et al., 2022a], fixed-support OT barycenter problem [Lin et al., 2020b], and projection robust OT problem [Lin et al., 2021a]. Notably, these variants have diverse structures which make algorithm design challenging; indeed, the former two are large-scale LPs with additional structure, while the latter one is a nonconvex-concave min-max problem with a manifold constraint. We also proved the dimension-independent sample complexity and concentration results for projection OT (POT) under reasonable structural conditions, and derived consistency and central limit theorems for the estimators [Lin et al., 2021a].

Nonsmooth nonconvex optimization with Lipschitz functions. In these works, we designed randomized gradient-free methods with finite-time convergence guarantee regardless of noisy function value oracles [Lin et al., 2022f]. We also studied the class of determinis-

tic subgradient-based methods and showed (i) all of these methods suffer from dimension-dependent convergence rate, and (ii) the function value oracle could offer more information than the subgradient oracle [Jordan et al., 2022b].

1.3 Organization

This thesis is centered around four concrete questions in answering the general basic question – *in optimization and machine learning, how can we harness problem-specific structure within our algorithms to obtain fast, practical algorithms with strong performance guarantees?*

We start with minimax optimization in Part I, and ask whether the two-timescale (stochastic) gradient descent ascent is provably efficient or not. Chapter 2 provided the first nonasymptotic analysis for two-timescale GDA in this setting, shedding light on its superior practical performance in training generative adversarial networks (GANs) and other real applications. This chapter is based on two joint works with Chi Jin and Michael I. Jordan [Lin et al., 2020c, 2022c]. Then, we proceed to the next question: "Can we design gradient-based algorithms that achieve the lower bounds in both convex-concave and nonconvex-concave settings?" Chapter 3 highlighted the importance of accelerated proximal point method in this context. This chapter is based on a joint work with Chi Jin and Michael I. Jordan [Lin et al., 2020d]. Finally, we investigate if there is a necessary performance gap between the Riemannian and Euclidean optimal gradient-based algorithms in terms of accuracy and the condition number. Chapter 4 provided an analysis of extragradient method and gradient descent ascent adapted to the manifold-constrained setting. This chapter is based on a joint work with Michael I. Jordan and Emmanouil-Vasileios Vlatakis-Gkaragkounis [Jordan et al., 2022a].

The central topic of Part II is high-order optimization and beyond. Unlike gradient-based algorithms, the very basic questions remain open for high-order optimization, inclusions and variational inequalities (VIs), including what the dynamics of optimal algorithms look like and whether or not the binary search scheme is necessary. Chapter 5 provided a control-theoretical perspective on optimal tensor algorithms for minimizing a convex and highly smooth function in a finite-dimensional Euclidean space. This chapter is based on a joint work with Michael I. Jordan [Lin and Jordan, 2022b]. Chapter 6 proposed and analyzed a new dynamical system with a closed-loop control law in a Hilbert space, aiming to shed light on the acceleration phenomenon for monotone inclusion problems. This chapter is based on a joint work with Michael I. Jordan [Lin and Jordan, 2023]. Chapter 7 settled an open and challenging question pertaining to the design of simple and optimal high-order methods for solving smooth and monotone VIs. This chapter is based on a joint work with Michael I. Jordan [Lin and Jordan, 2022c].

Finally, Part III studies several other structured problems, covering optimal transport, nonsmooth nonconvex optimization, and no-regret learning in games. Chapter 8 presented several new complexity results for the entropic regularized algorithms that approximately solve the optimal transport (OT) problem between two discrete probability measures and showed the efficiency of these algorithms in practice. This chapter is based on a joint work

with Nhat Ho and Michael I. Jordan [Lin et al., 2022b]. Chapter 9 addressed two challenges that impede the development of efficient nonconvex nonsmooth optimization methods with finite-time convergence guarantee: the lack of computationally tractable optimality criterion and the lack of computationally powerful oracles. This chapter is based on a joint work with Michael I. Jordan and Zeyu Zheng [Lin et al., 2022f]. Chapter 10 designed the fully adaptive gradient-based algorithm that does not require a priori knowledge of parameters. It achieved near-optimal regret guarantee in single-agent setting and near-optimal last-iterate convergence rate guarantee in multi-agent setting. Our results also immediately yield the first feasible and near-optimal algorithm for solving a learning version of the newsvendor problem in both single-retailer and multi-retailer settings. This chapter is based on a joint work with Michael I. Jordan and Zhengyuan Zhou [Jordan et al., 2022c].

Part I

Minimax Optimization

Chapter 2

Two-Timescale Gradient Descent Ascent

We consider nonconvex-concave minimax problems, $\min_{\mathbf{x}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y})$ where f is nonconvex in $\mathbf{x}$ but concave in $\mathbf{y}$ and $\mathcal{Y}$ is a convex and bounded set. One of the most popular algorithms for solving this problem is the celebrated gradient descent ascent (GDA) algorithm, which has been widely used in machine learning, control theory and economics. Despite the extensive convergence results for the convex-concave setting, GDA with equal stepsize can converge to limit cycles or even diverge in a general setting. In this paper, we present the complexity results on two-timescale GDA for solving nonconvex-concave minimax problems, showing that the algorithm can find a stationary point of the function $\Phi(\cdot) := \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot,\mathbf{y})$ efficiently. To our knowledge, this is the first nonasymptotic analysis for two-timescale GDA in this setting, shedding light on its superior practical performance in training generative adversarial networks (GANs) and other real applications.

2.1 Introduction

We consider the following smooth minimax optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}), \tag{2.1}$$

where $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is nonconvex in $\mathbf{x}$ but concave in $\mathbf{y}$ and where $\mathcal{Y}$ is a convex set. Since von Neumann's seminal work [Neumann, 1928], the problem of finding the solution to problem (2.1) has been a major focus of research in mathematics, economics and computer science [Basar and Olsder, 1999, Nisan et al., 2007, Von Neumann and Morgenstern, 2007]. In recent years, minimax optimization theory has begun to see applications in machine learning, with examples including generative adversarial networks (GANs) [Goodfellow et al., 2014], statistics [Xu et al., 2009, Abadeh et al., 2015], online learning [Cesa-Bianchi and Lugosi, 2006], deep learning [Sinha et al., 2018] and distributed computing [Shamma, 2008, Mateos et al., 2010]. Moreover, there is increasing awareness that machine-learning systems are

embedded in real-world settings involving scarcity or competition that impose game-theoretic constraints [Jordan, 2018].

One of the simplest candidates for solving problem (2.1) is the natural generalization of gradient descent (GD) known as *gradient descent ascent* (GDA). At each iteration, this algorithm performs gradient descent over the variable $\mathbf{x}$ with the stepsize $\eta_{\mathbf{x}}$ and gradient ascent over the variable $\mathbf{y}$ with the stepsize $\eta_{\mathbf{y}}$. On the positive side, when the objective function f is convex in $\mathbf{x}$ and concave in $\mathbf{y}$, there is a vast literature establishing asymptotic and nonasymptotic convergence for the average iterates generated by GDA with the equal stepsizes ($\eta_{\mathbf{x}} = \eta_{\mathbf{y}}$); [see, e.g., Korpelevich, 1976, Chen and Rockafellar, 1997, Nedić and Ozdaglar, 2009, Nemirovski, 2004, Du and Hu, 2019]. Local linear convergence can also be shown under the additional assumption that f is locally strongly convex in $\mathbf{x}$ and strongly concave in $\mathbf{y}$ [Cherukuri et al., 2017, Liang and Stokes, 2019, Adolphs et al., 2019]. However, there has been no shortage of research highlighting the fact that in a general setting GDA with equal stepsizes can converge to limit cycles or even diverge [Benaım and Hirsch, 1999, Hommes and Ochea, 2012, Mertikopoulos et al., 2018].

Recent research has focused on alternative gradient-based algorithms that have guarantees beyond the convex-concave setting [Daskalakis et al., 2018, Heusel et al., 2017, Mertikopoulos et al., 2019, Mazumdar et al., 2019]. Two-timescale GDA [Heusel et al., 2017] has been particularly popular. This algorithm, which involves unequal stepsizes ($\eta_{\mathbf{x}} \neq \eta_{\mathbf{y}}$), has been shown to empirically to alleviate the issues of limit circles and it has theoretical support in terms of local asymptotic convergence to Nash equilibria [Heusel et al., 2017, Theorem 2].

This asymptotic result stops short of providing an understanding of algorithmic efficiency, and it would be desirable to provide a stronger, nonasymptotic, theoretical convergence rate for two-timescale GDA in a general setting. In particular, the following general structure arises in many applications: $f(\mathbf{x}, \cdot)$ is concave for any $\mathbf{x}$ and $\mathcal{Y}$ is a bounded set. Two typical examples include training of a neural network which is robust to adversarial examples [Madry et al., 2018] and learning of a robust classifier from multiple distributions [Sinha et al., 2018]. Both of these schemes can be posed as nonconvex-concave minimax problems. Based on this observation, it is natural to ask the question: *Are two-timescale GDA and stochastic GDA (SGDA) provably efficient for nonconvex-concave minimax problems?*

This paper presents an affirmative answer to this question, providing nonasymptotic complexity results for two-time scale GDA and SGDA in two settings. In the nonconvex-strongly-concave setting, two-time scale GDA and SGDA require $O(\kappa^2\epsilon^{-2})$ gradient evaluations and $O(\kappa^3\epsilon^{-4})$ stochastic gradient evaluations, respectively, to return an ϵ-stationary point of the function $\Phi(\cdot) = \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ where $\kappa > 0$ is a condition number. In the nonconvex-concave setting, two-time scale GDA and SGDA require $O(\epsilon^{-6})$ gradient evaluations and $O(\epsilon^{-8})$ stochastic gradient evaluations.

To motivate the proof ideas for analyzing two-time scale GDA and SGDA, it is useful to contrast our work with some of the strongest existing convergence analyses for nonconvex-concave problems. In particular, Jin et al. [2020] and Nouiehed et al. [2019] have provided complexity results for algorithms that have a nested-loop structure. Specifically, GDmax and multistep GDA are algorithms in which the outer loop can be interpreted as an inexact

Table 2.1: The gradient complexity of all algorithms for nonconvex-(strongly)-concave minimax problems. ϵ is a tolerance and $\kappa > 0$ is a condition number. The result denoted by * refers to the complexity bound after translating from ϵ-stationary point of f to our optimality measure; see Propositions 2.4.11 and 2.4.12. The result denoted by $^\circ$ is not presented explicitly but easily derived by standard arguments.

	Nonconvex-Strongly-Concave		Nonconvex-Concave		Simplicity
	Deterministic	Stochastic	Deterministic	Stochastic	
Jin et al. [2020]	$\tilde{O}\left(\kappa^2\epsilon^{-2}\right)^\circ$	$\tilde{O}\left(\kappa^3\epsilon^{-4}\right)$	$O(\epsilon^{-6})$	$O(\epsilon^{-8})^\circ$	Double-loop
Rafique et al. [2022]	$\tilde{O}(\kappa^2\epsilon^{-2})$	$\tilde{O}(\kappa^3\epsilon^{-4})$	$\tilde{O}(\epsilon^{-6})$	$\tilde{O}(\epsilon^{-6})$	Double-loop
Nouiehed et al. [2019]	$\tilde{O}(\kappa^4\epsilon^{-2})^{\star,\circ}$	–	$O(\epsilon^{-7})^\star$	–	Double-loop
Thekumparampil et al. [2019]	–	–	$\tilde{O}(\epsilon^{-3})$	–	Triple-loop
Kong and Monteiro [2021]	–	–	$\tilde{O}(\epsilon^{-3})$	–	Triple-loop
Lu et al. [2020]	$O(\kappa^4\epsilon^{-2})^\star$	–	$O(\epsilon^{-8})^\star$	–	Single-loop
This paper	$O(\kappa^2\epsilon^{-2})$	$O(\kappa^3\epsilon^{-4})$	$O(\epsilon^{-6})$	$O(\epsilon^{-8})$	Single-loop

gradient descent on a nonconvex function $\Phi(\cdot) = \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ while the inner loop provides an approximate solution to the maximization problem $\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})$ for a given $\mathbf{x} \in \mathbb{R}^m$. Strong convergence results are obtained when accelerated gradient ascent is used in the maximization problem.

Compared to GDmax and multistep GDA, two-time scale GDA and SGDA are harder to analyze. Indeed, $\mathbf{y}_t$ is not necessarily guaranteed to be close to $\mathbf{y}^\star(\mathbf{x}_t)$ at each iteration and thus it is unclear that $\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)$ might a reasonable descent direction. To overcome this difficulty, we develop a new technique which analyzes the concave optimization with a slowly changing objective function. This is the main technical contribution of this paper.

Notation. We use bold lower-case letters to denote vectors and caligraphic upper-case letter to denote sets. We use $\|\cdot\|$ to denote the ℓ_2-norm of vectors and spectral norm of matrices. For a function $f : \mathbb{R}^n \to \mathbb{R}$, $\partial f(\mathbf{z})$ denotes the subdifferential of f at $\mathbf{z}$. If f is differentiable, $\partial f(\mathbf{z}) = \{\nabla f(\mathbf{z})\}$ where $\nabla f(\mathbf{z})$ denotes the gradient of f at $\mathbf{z}$ and $\nabla_{\mathbf{x}} f(\mathbf{z})$ denotes the partial gradient of f with respect to $\mathbf{x}$ at $\mathbf{z}$. For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, the largest and smallest eigenvalue of A denoted by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$.

2.2 Related Works

Historically, an early concrete instantiation of problem (2.1) involved computing a pair of probability vectors $(\mathbf{x}, \mathbf{y})$, or equivalently solving $\min_{\mathbf{x}\in\Delta^m} \max_{\mathbf{y}\in\Delta^n} \mathbf{x}^\top A\mathbf{y}$ for a matrix $A \in \mathbb{R}^{m\times n}$ and probability simplices Δ^m and Δ^n. This bilinear minimax problem together with von Neumann's minimax theorem [Neumann, 1928] was a cornerstone in the development

of game theory. A simple and generic algorithm scheme was developed for solving this problem in which the min and max players each implemented a simple learning procedure in tandem [Robinson, 1951]. After then, Sion [1958] generalized von Neumann's result from bilinear games to general convex-concave games, $\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{y}} \min_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$, and triggered a line of algorithmic research on convex-concave minimax optimization in both continuous time [Kose, 1956, Cherukuri et al., 2017] and discrete time [Uzawa, 1958, Golshtein, 1974, Korpelevich, 1976, Nemirovski, 2004, Nedić and Ozdaglar, 2009, Mokhtari et al., 2020b,a, Azizian et al., 2020a]. It is well known that GDA can find an ϵ-approximate saddle point within $O(\kappa^2 \log(1/\epsilon))$ iterations for strongly-convex-strongly-concave games, and $O(\epsilon^{-2})$ iterations for convex-concave games if we impose the diminishing stepsizes [Nedić and Ozdaglar, 2009, Nemirovski, 2004].

Nonconvex-concave minimax problems appear to be a class of tractable problems in the form of problem (2.1) and have emerged as a focus in optimization and machine learning [Namkoong and Duchi, 2016, Sinha et al., 2018, Sanjabi et al., 2018, Grnarova et al., 2018, Nouiehed et al., 2019, Thekumparampil et al., 2019, Lu et al., 2020, Kong and Monteiro, 2021, Rafique et al., 2022]; see Table 2.1 for a comprehensive overview. We also wish to highlight the work of Grnarova et al. [2018], who proposed a variant of GDA for nonconvex-concave problem and the work of Sinha et al. [2018] and Sanjabi et al. [2018], who studied a class of inexact nonconvex SGD algorithms that can be categorized as variants of SGDmax for nonconvex-strongly-concave problem. Jin et al. [2020] analyzed the GDmax algorithm for nonconvex-concave problem and provided nonasymptotic convergence results.

Rafique et al. [2022] proposed "proximally guided stochastic mirror descent" and "variance reduced gradient" algorithms (PGSMD/PGSVRG) and proved that these algorithms find an approximate stationary point of $\Phi(\cdot) := \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$. However, PGSMD/PGSVRG are nested-loop algorithms and convergence results were established only in the special case where $f(\mathbf{x}, \cdot)$ is a linear function [Rafique et al., 2022, Assumption 2 D.2]. Nouiehed et al. [2019] developed a multistep GDA (MGDA) algorithm by incorporating accelerated gradient ascent as the subroutine at each iteration. This algorithm provably finds an approximate stationary point of $f(\cdot, \cdot)$ for nonconvex-concave problems with the fast rate of $O(\epsilon^{-3.5})$. Very recently, Thekumparampil et al. [2019] have proposed a proximal dual implicit accelerated gradient (ProxDIAG) algorithm for nonconvex-concave problems and proved that the algorithm find an approximate stationary point of $\Phi(\cdot)$ with the rate of $O(\epsilon^{-3})$. This complexity result is also achieved by an inexact proximal point algorithm [Kong and Monteiro, 2021]. All of these algorithms are, however, nested-loop algorithms and thus relatively complicated to implement. One would like to know whether the nested-loop structure is necessary or whether GDA, a single-loop algorithm, can be guaranteed to converge in the nonconvex-(strongly)-concave setting.

The most closest work is Lu et al. [2020] in which a single-loop HiBSA algorithm for nonconvex-(strongly)-concave problems is proposed with theoretical guarantees under a different notion of optimality. However, their analysis requires some restrictive assumptions; e.g., that $f(\cdot, \cdot)$ is lower bounded. We only require that $\max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$ is lower bounded. An example which meets our conditions and not those of Lu et al. [2020] is

$\min_{\mathbf{x}\in\mathbb{R}} \max_{\mathbf{y}\in[-1,1]} \mathbf{x}^\top\mathbf{y}$. Our less-restrictive assumptions make the problem more challenging and our technique is accordingly fundamentally difference from theirs.

During the past decade, the nonconvex-nonconcave minimax optimization has become a central topic in machine learning, inspired in part by the advent of generative adversarial networks [Goodfellow et al., 2014] and adversarial learning [Madry et al., 2018, Namkoong and Duchi, 2016, Sinha et al., 2018]. Most recent work aims at defining a notion of goodness or the development of new procedures for reducing oscillations [Daskalakis and Panageas, 2018b, Adolphs et al., 2019, Mazumdar et al., 2019] and speeding up the convergence of gradient dynamics [Heusel et al., 2017, Balduzzi et al., 2018, Mertikopoulos et al., 2019, Liu et al., 2021]. More specifically, Daskalakis and Panageas [2018b] studied minimax optimization (or zero-sum games) and show that the stable limit points of GDA are not necessarily Nash equilibria. Adolphs et al. [2019] and Mazumdar et al. [2019] proposed Hessian-based algorithms whose stable fixed points are exactly Nash equilibria. On the other hand, Balduzzi et al. [2018] developed a new symplectic gradient adjustment (SGA) algorithm for finding stable fixed points in potential games and Hamiltonian games. Heusel et al. [2017] proposed two-timescale GDA and show that Nash equilibria are stable fixed points of the continuous limit of two-timescale GDA under certain strong conditions. All of the existing convergence results are either local or asymptotic and can not be extended to cover our results in a nonconvex-concave setting. Very recently, Mertikopoulos et al. [2019] and Liu et al. [2021] provide nonasymptotic guarantees for a special class of nonconvex-nonconcave minimax problems under variational stability and the Minty condition. However, while both of these two conditions must hold in convex-concave setting, they do not necessarily hold in nonconvex-(strongly)-concave problem.

From the online learning perspective, it is crucial to understand if the proposed algorithm achieves no-regret property. For example, the optimistic algorithm [Daskalakis and Panageas, 2018a] is a no-regret algorithm, while the extragradient algorithm [Mertikopoulos et al., 2019] is not. In comparing limit behavior of zero-sum game dynamics, Bailey and Piliouras [2018] showed that the multiplicative weights update has similar property as GDA and specified the necessity of introducing the optimistic algorithms to study the last-iterate convergence.

2.3 Preliminaries

We recall basic definitions for smooth functions.

Definition 2.3.1 *A function f is L-**Lipschitz** if for $\forall \mathbf{x}, \mathbf{x}'$, $\|f(\mathbf{x}) - f(\mathbf{x}')\| \le L\|\mathbf{x} - \mathbf{x}'\|$.*

Definition 2.3.2 *A function f is ℓ-**smooth** if for $\forall \mathbf{x}, \mathbf{x}'$, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\| \le \ell\|\mathbf{x} - \mathbf{x}'\|$.*

Recall that the minimax problem (2.1) is equivalent to minimizing $\Phi(\cdot) = \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$. For nonconvex-concave minimax problems in which $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathbb{R}^m$, the maximization problem $\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})$ can be solved efficiently. However, it is still NP hard to find the global minimum of Φ in general since Φ is nonconvex.

We start by defining local surrogate for the global minimum of Φ. A surrogate in nonconvex optimization is the notion of stationarity, which is appropriate if Φ is differentiable.

Definition 2.3.3 *A point* $\mathbf{x}$ *is an* ϵ-stationary point *($\epsilon \geq 0$) of a differentiable function* Φ *if* $\|\nabla\Phi(\mathbf{x})\| \leq \epsilon$. *If* $\epsilon = 0$, *then* $\mathbf{x}$ *is a stationary point.*

Definition 2.3.3 is sufficient for nonconvex-strongly-concave minimax problem since $\Phi(\cdot) = \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ is differentiable in that setting. In contrast, a function Φ is not necessarily differentiable for general nonconvex-concave minimax problem even if f is Lipschitz and smooth. A weaker condition that we make use of is the following.

Definition 2.3.4 *A function* Φ *is* ℓ-weakly convex *if a function* $\Phi(\cdot) + (\ell/2)\|\cdot\|^2$ *is convex.*

For a ℓ-weakly convex function Φ, the subdifferential $\partial\Phi$ is uniquely determined by the subdifferential of $\Phi + (\ell/2)\|\cdot\|^2$. Thus, a naive measure of approximate stationarity can be defined as a point $\mathbf{x} \in \mathbb{R}^m$ such that at least one subgradient is small: $\min_{\xi\in\partial\Phi(\mathbf{x})} \|\xi\| \leq \epsilon$. However, this notion of stationarity can be very restrictive when optimizing nonsmooth functions. For example, when $\Phi(\cdot) = |\cdot|$ is a one-dimensional function, an ϵ-stationary point is zero for all $\epsilon \in [0, 1)$. This means that finding an approximate stationary point under this notion is as difficult as solving the problem exactly. Davis and Drusvyatskiy [2019] propose an alternative notion of stationarity based on the Moreau envelope. This has become recognized as standard for optimizing a weakly convex function.

Definition 2.3.5 *A function* $\Phi_\lambda : \mathbb{R}^m \to \mathbb{R}$ *is the Moreau envelope of* Φ *with a positive parameter* $\lambda > 0$ *if* $\Phi_\lambda(\mathbf{x}) = \min_{\mathbf{w}} \Phi(\mathbf{w}) + (1/2\lambda)\|\mathbf{w} - \mathbf{x}\|^2$ *for each* $\mathbf{x} \in \mathbb{R}^m$.

Lemma 2.3.6 *If* f *is* ℓ-*smooth and* $\mathcal{Y}$ *is bounded, the Moreau envelope* $\Phi_{1/2\ell}$ *of* $\Phi(\cdot) = \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ *is differentiable,* ℓ-*smooth and* ℓ-*strongly convex.*

Thus, an alternative measure of approximate stationarity of a function $\Phi(\cdot) = \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ can be defined as a point $\mathbf{x} \in \mathbb{R}^m$ such that the norm of the gradient of Moreau envelope is small: $\|\nabla\Phi_{1/2\ell}\| \leq \epsilon$. More generally, we have

Definition 2.3.7 *A point* $\mathbf{x}$ *is an* ϵ-*stationary point ($\epsilon \geq 0$) of a* ℓ-*weakly convex function* Φ *if* $\|\nabla\Phi_{1/2\ell}(\mathbf{x})\| \leq \epsilon$. *If* $\epsilon = 0$, *then* $\mathbf{x}$ *is a stationary point.*

Even though Definition 2.3.7 uses the language of Moreau envelopes, it also connects to the function Φ as follows.

Lemma 2.3.8 *If* $\mathbf{x}$ *is an* ϵ-*stationary point of a* ℓ-*weakly convex function* Φ *(Definition 2.3.7), there exists* $\hat{\mathbf{x}} \in \mathbb{R}^m$ *such that* $\min_{\xi\in\partial\Phi(\hat{\mathbf{x}})} \|\xi\| \leq \epsilon$ *and* $\|\mathbf{x} - \hat{\mathbf{x}}\| \leq \epsilon/2\ell$.

Lemma 2.3.8 shows that an ϵ-stationary point defined by Definition 2.3.7 can be interpreted as the relaxation or surrogate for $\min_{\xi\in\partial\Phi(\mathbf{x})} \|\xi\| \leq \epsilon$. In particular, if a point $\mathbf{x}$ is an ϵ-stationary point of an ℓ-weakly convex function Φ, then $\mathbf{x}$ is close to a point $\hat{\mathbf{x}}$ which has at least one small subgradient.

Algorithm 1 Two-Timescale GDA

Input: $(\mathbf{x}_0, \mathbf{y}_0)$, stepsizes $(\eta_{\mathbf{x}}, \eta_{\mathbf{y}})$.
for $t = 1, 2, \ldots, T$ **do**
  $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta_{\mathbf{x}} \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$,
  $\mathbf{y}_t \leftarrow \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_{t-1} + \eta_{\mathbf{y}} \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}))$.
Randomly draw $\hat{\mathbf{x}}$ from $\{\mathbf{x}_t\}_{t=1}^T$ at uniform.
Return: $\hat{\mathbf{x}}$.

Algorithm 2 Two-Timescale SGDA

Input: $(\mathbf{x}_0, \mathbf{y}_0)$, stepsizes $(\eta_{\mathbf{x}}, \eta_{\mathbf{y}})$, batch size M.
for $t = 1, 2, \ldots, T$ **do**
  Draw a collection of i.i.d. data samples $\{\xi_i\}_{i=1}^M$.
  $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta_{\mathbf{x}} \left(\frac{1}{M} \sum_{i=1}^M G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i) \right)$.
  $\mathbf{y}_t \leftarrow \mathcal{P}_{\mathcal{Y}} \left(\mathbf{y}_{t-1} + \eta_{\mathbf{y}} (\frac{1}{M} \sum_{i=1}^M G_{\mathbf{y}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i)) \right)$.
Randomly draw $\hat{\mathbf{x}}$ from $\{\mathbf{x}_t\}_{t=1}^T$ at uniform.
Return: $\hat{\mathbf{x}}$.

Remark 2.3.9 *We remark that our notion of stationarity is natural in real scenarios. Indeed, many applications arising from adversarial learning can be formulated as the minimax problem (2.1), and, in this setting,* $\mathbf{x}$ *is the classifier while* $\mathbf{y}$ *is the adversarial noise for the data. Practitioners are often interested in finding a robust classifier* $\mathbf{x}$ *instead of recovering the adversarial noise* $\mathbf{y}$*. Any stationary point of the function* $\Phi(\cdot) = \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$ *corresponds precisely to a robust classifier that achieves better classification error.*

Remark 2.3.10 *There are also other notions of stationarity based on* ∇f *are proposed for nonconvex-concave minimax problems in the literature [Lu et al., 2020, Nouiehed et al., 2019]. However, as pointed by Thekumparampil et al. [2019], these notions are weaker than that defined in Definition 2.3.3 and 2.3.7. For the sake of completeness, we specify the relationship between our notion of stationarity and other notions in Proposition 2.4.11 and 2.4.12.*

2.4 Main Results

We present complexity results for two-timescale GDA and SGDA in the setting of nonconvex-strongly-concave and nonconvex-concave minimax problems.

The algorithmic schemes that we study are extremely simple and are presented in Algorithm 1 and 2. In particular, each iteration comprises one (stochastic) gradient descent step over $\mathbf{x}$ with the stepsize $\eta_{\mathbf{x}} > 0$ and one (stochastic) gradient ascent step over $\mathbf{y}$ with the stepsize $\eta_{\mathbf{y}} > 0$. The choice of stepsizes $\eta_{\mathbf{x}}$ and $\eta_{\mathbf{y}}$ is crucial for the algorithms in both theoretical and practical senses. In particular, classical GDA and SGDA assume that $\eta_{\mathbf{x}} = \eta_{\mathbf{y}}$,

and the last iterate is only known convergent in strongly convex-concave problems [Liang and Stokes, 2019]. Even in convex-concave settings (or bilinear settings as special cases), GDA requires the assistance of averaging or other strategy [Daskalakis and Panageas, 2018a] to converge, otherwise, with fixed stepsize, the last iterate will always diverge and hit the constraint boundary eventually [Daskalakis et al., 2018, Mertikopoulos et al., 2018, Daskalakis and Panageas, 2018a]. In contrast, two-timescale GDA and SGDA ($\eta_{\mathbf{x}} \neq \eta_{\mathbf{y}}$) were shown to be locally convergent and practical in training GANs [Heusel et al., 2017].

One possible reason for this phenomenon is that the choice of $\eta_{\mathbf{x}} \neq \eta_{\mathbf{y}}$ reflects the non-symmetric nature of nonconvex-(strongly)-concave problems. For sequential problems such as robust learning, where the natural order of min-max is important (i.e., min-max is not equal to max-min), practitioners often prefer faster convergence for the inner max problem. Therefore, it is reasonable for us to choose $\eta_{\mathbf{x}} \ll \eta_{\mathbf{y}}$ rather than $\eta_{\mathbf{x}} = \eta_{\mathbf{y}}$.

Finally, we make the standard assumption that the oracle $G = (G_{\mathbf{x}}, G_{\mathbf{y}})$ is unbiased and has bounded variance.

Assumption 2.4.1 *The oracle G satisfies $\mathbb{E}[G(\mathbf{x}, \mathbf{y}, \xi) - \nabla f(\mathbf{x}, \mathbf{y}] = 0$ and $\mathbb{E}[\|G(\mathbf{x}, \mathbf{y}, \xi) - \nabla f(\mathbf{x}, \mathbf{y})\|^2] \leq \sigma^2$.*

Nonconvex-strongly-concave minimax problems. We present the complexity results for two-time-scale GDA and SGDA in the setting of nonconvex-strongly-concave minimax problems. The following assumption is made throughout.

Assumption 2.4.2 *The objective function and set ($f : \mathbb{R}^{m+n} \to \mathbb{R}, \mathcal{Y} \subseteq \mathbb{R}^n$) satisfy that (i) f is ℓ-smooth and $f(\mathbf{x}, \cdot)$ is μ-strongly concave; and (ii) $\mathcal{Y}$ is a convex and bounded set with a diameter $D \geq 0$.*

Let $\kappa = \ell/\mu$ denote the condition number and define

$$\Phi(\cdot) \;=\; \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y}), \quad \mathbf{y}^\star(\cdot) = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y}).$$

We present a technical lemma on the structure of the function Φ in the nonconvex-strongly-concave setting.

Lemma 2.4.3 *Under Assumption 2.4.2, $\Phi(\cdot)$ is $(\ell + \kappa\ell)$-smooth with $\nabla\Phi(\cdot) = \nabla_{\mathbf{x}} f(\cdot, \mathbf{y}^\star(\cdot))$. Also, $\mathbf{y}^\star(\cdot)$ is κ-Lipschitz.*

Since Φ is differentiable, the notion of stationarity in Definition 2.3.3 is our target given only access to the (stochastic) gradient of f. Denote $\Delta_\Phi = \Phi(\mathbf{x}_0) - \min_{\mathbf{x}} \Phi(\mathbf{x})$, we proceed to provide theoretical guarantees for Algorithm 1 and 2.

Theorem 2.4.4 (GDA) *Under Assumption 2.4.2 and letting the stepsizes be chosen as $\eta_{\mathbf{x}} = \Theta(1/\kappa^2\ell)$ and $\eta_{\mathbf{y}} = \Theta(1/\ell)$, the iteration complexity (also the gradient complexity) of Algorithm 1 to return an ϵ-stationary point is bounded by*

$$O\left(\frac{\kappa^2 \ell \Delta_\Phi + \kappa \ell^2 D^2}{\epsilon^2}\right).$$

Theorem 2.4.5 (SGDA) *Under Assumption 2.4.1 and 2.4.2 and letting the stepsizes $\eta_{\mathbf{x}}, \eta_{\mathbf{y}}$ be chosen as the same in Theorem 2.4.4 with the batch size $M = \Theta(\max\{1, \kappa\sigma^2\epsilon^{-2}\})$, the iteration complexity of Algorithm 2 to return an ϵ-stationary point is bounded by*

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2}{\epsilon^2}\right),$$

which gives the total stochastic gradient complexity:

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2}{\epsilon^2}\max\left\{1, \frac{\kappa\sigma^2}{\epsilon^2}\right\}\right).$$

We make several remarks. First, the two-timescale GDA and SGDA are guaranteed to find an ϵ-stationary point of $\Phi(\cdot)$ within $O(\kappa^2\epsilon^{-2})$ gradient evaluations and $O(\kappa^3\epsilon^{-4})$ stochastic gradient evaluations, respectively. The ratio of stepsizes $\eta_{\mathbf{y}}/\eta_{\mathbf{x}}$ is required to be $\Theta(\kappa^2)$ due to the nonsymmetric nature of our problem (min-max is not equal to max-min). The quantity $O(\kappa^2)$ reflects an efficiency trade-off in the algorithm. Furthermore, both of the algorithms are only guaranteed to visit an ϵ-stationary point within a certain number of iterations and return $\hat{\mathbf{x}}$ which is drawn from $\{\mathbf{x}_t\}_{t=1}^T$ at uniform. This does not mean that the last iterate $\mathbf{x}_T$ is the ϵ-stationary point. Such a scheme and convergence result are standard in nonconvex optimization for GD or SGD to find stationary points. In practice, one usually returns the iterate when the learning curve stops changing significantly. Finally, the minibatch size $M = \Theta(\epsilon^{-2})$ is necessary for the convergence property of two-timescale SGDA. Even though our proof technique can be extended to the purely stochastic setting ($M = 1$), the complexity result becomes worse, i.e., $O(\kappa^3\epsilon^{-5})$. It remains open whether this gap can be closed or not and we leave it as future work.

Nonconvex-concave minimax problems. We present the complexity results for two-timescale GDA and SGDA in the nonconvex-concave minimax setting. The following assumption is made throughout.

Assumption 2.4.6 *The objective function and constraint set, $(f : \mathbb{R}^{m+n} \to \mathbb{R}, \mathcal{Y} \subset \mathbb{R}^n)$ satisfy (i) f is ℓ-smooth and $f(\cdot, \mathbf{y})$ is L-Lipschitz for each $\mathbf{y} \in \mathcal{Y}$ and $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathbb{R}^m$; and (ii) $\mathcal{Y}$ is a convex and bounded set with a diameter $D \geq 0$.*

Since $f(\mathbf{x}, \cdot)$ is merely concave for each $\mathbf{x} \in \mathbb{R}^m$, the function $\Phi(\cdot) = \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ is possibly not differentiable. Fortunately, the following structural lemma shows that Φ is ℓ-weakly convex and L-Lipschitz.

Lemma 2.4.7 *Under Assumption 2.4.6, $\Phi(\cdot)$ is ℓ-weakly convex and L-Lipschitz with the gradient $\nabla_{\mathbf{x}} f(\cdot, \mathbf{y}^\star(\cdot)) \in \partial\Phi(\cdot)$ where $\mathbf{y}^\star(\cdot) \in \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$.*

Since Φ is ℓ-weakly convex, the notion of stationarity in Definition 2.3.7 is our target given only access to the (stochastic) gradient of f. Denote $\widehat{\Delta}_\Phi = \Phi_{1/2\ell}(\mathbf{x}_0) - \min_{\mathbf{x}} \Phi_{1/2\ell}(\mathbf{x})$ and $\widehat{\Delta}_0 = \Phi(\mathbf{x}_0) - f(\mathbf{x}_0, \mathbf{y}_0)$, we present complexity results for Algorithm 1 and 2.

Theorem 2.4.8 (GDA) *Under Assumption 2.4.6 and letting the step sizes be chosen as* $\eta_{\mathbf{x}} = \Theta(\epsilon^4/(\ell^3 L^2 D^2))$ *and* $\eta_{\mathbf{y}} = \Theta(1/\ell)$*, the iteration complexity (also the gradient complexity) of Algorithm 1 to return an* ϵ*-stationary point is bounded by*

$$O\left(\frac{\ell^3 L^2 D^2 \widehat{\Delta}_\Phi}{\epsilon^6} + \frac{\ell^3 D^2 \widehat{\Delta}_0}{\epsilon^4}\right).$$

Theorem 2.4.9 (SGDA) *Under Assumption 2.4.1 and 2.4.6 and letting the step sizes be chosen as* $\eta_{\mathbf{x}} = \Theta(\epsilon^4/(\ell^3 D^2 (L^2 + \sigma^2)))$ *and* $\eta_{\mathbf{y}} = \Theta(\epsilon^2/\ell\sigma^2)$ *with the batchsize* $M = 1$*, the iteration complexity (also the stochastic gradient complexity) of Algorithm 2 to return an* ϵ*-stationary point is bounded by*

$$\mathcal{O}\left(\left(\frac{\ell^3 (L^2 + \sigma^2) D^2 \widehat{\Delta}_\Phi}{\epsilon^6} + \frac{\ell^3 D^2 \widehat{\Delta}_0}{\epsilon^4}\right) \max\left\{1, \frac{\sigma^2}{\epsilon^2}\right\}\right).$$

We make several additional remarks. First, two-timescale GDA and SGDA are guaranteed to find an ϵ-stationary point in terms of Moreau envelopes within $O(\epsilon^{-6})$ gradient evaluations and $O(\epsilon^{-8})$ stochastic gradient evaluations, respectively. The ratio of stepsizes $\eta_{\mathbf{y}}/\eta_{\mathbf{x}}$ is required to be $\Theta(1/\epsilon^4)$ and this quantity reflects an efficiency trade-off in the algorithm. Furthermore, similar arguments hold for the output of the algorithms. Finally, the minibatch size $M = 1$ is allowed in Theorem 2.4.9, which is different from the result in Theorem 2.4.5.

Relationship between the stationarity notions. We provide additional technical results on the relationship between our notions of stationarity and other notions based on ∇f in the literature [Lu et al., 2020, Nouiehed et al., 2019]. In particular, we show that two notions can be translated in both directions with extra computational cost.

Definition 2.4.10 *A pair of points* $(\mathbf{x}, \mathbf{y})$ *is an* ϵ*-stationary point (*$\epsilon \geq 0$*) of a differentiable function* Φ *if, for* $\mathbf{y}^+ = \mathcal{P}_{\mathcal{Y}}(\mathbf{y} + (1/\ell)\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))$*, we have*

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^+)\| \leq \epsilon, \quad \|\mathbf{y}^+ - \mathbf{y}\| \leq \epsilon/\ell.$$

We present our results in the following two propositions.

Proposition 2.4.11 *Under Assumption 2.4.2, if a point* $\hat{\mathbf{x}}$ *is an* ϵ*-stationary point in terms of Definition 2.3.3, an* $O(\epsilon)$*-stationary point* $(\mathbf{x}', \mathbf{y}')$ *in terms of Definition 2.4.10 can be obtained using additional* $O(\kappa \log(1/\epsilon))$ *gradients or* $O(\epsilon^{-2})$ *stochastic gradients. Conversely, if a point* $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ *is an* ϵ/κ*-stationary point in terms of Definition 2.4.10, a point* $\hat{\mathbf{x}}$ *is an* $O(\epsilon)$*-stationary point in terms of Definition 2.3.3.*

Proposition 2.4.12 *Under Assumption 2.4.6, if a point* $\hat{\mathbf{x}}$ *is an* ϵ*-stationary point in terms of Definition 2.3.7, an* $O(\epsilon)$*-stationary point* $(\mathbf{x}', \mathbf{y}')$ *in terms of Definition 2.4.10 can be obtained using additional* $O(\epsilon^{-2})$ *gradients or* $O(\epsilon^{-4})$ *stochastic gradients. Conversely, if a point* $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ *is an* $\epsilon^2/\ell D$*-stationary point in terms of Definition 2.4.10, a point* $\hat{\mathbf{x}}$ *is an* $O(\epsilon)$*-stationary point in terms of Definition 2.3.3.*

To translate the notion of stationarity based on ∇f to our notion of stationarity, we need to pay an additional factor of $O(\kappa \log(1/\epsilon))$ or $O(\epsilon^{-2})$ in the two settings. In this sense, our notion of stationarity is stronger than that of Lu et al. [2020] and Nouiehed et al. [2019].

Discussion. Note that the focus of this paper is to provide basic nonasymptotic guarantees for the simple, and widely-used, two-timescale GDA and SGDA algorithms in the nonconvex-(strongly)-concave settings. We do not wish to imply that these algorithms are optimal in any sense, nor that acceleration should necessarily be achieved by incorporating momentum into the update for the variable $\mathbf{y}$. In fact, the optimal rate for optimizing a nonconvex-(strongly)-concave function remains open. The best known complexity bound has been presented by Thekumparampil et al. [2019] and Kong and Monteiro [2021]. Both of the analyses only require $\tilde{O}(\epsilon^{-3})$ gradient computations for solving nonconvex-concave problems but suffer from rather complicated algorithmic schemes.

Moreover, our complexity results are also valid in the convex-concave setting and this does not contradict results showing the divergence of GDA with fixed stepsize. We note a few distinctions: (1) our results guarantee that GDA will visit ϵ-stationary points at some iterates, which are not necessarily the last iterates; (2) our results only guarantee stationarity in terms of $\mathbf{x}_t$, not $(\mathbf{x}_t, \mathbf{y}_t)$. In fact, our proof permits the possibility of significant changes in $\mathbf{y}_t$ even when $\mathbf{x}_t$ is already close to stationarity. This together with our choice $\eta_{\mathbf{x}} \ll \eta_{\mathbf{y}}$, makes our results valid. To this end, we highlight that our algorithms can be used to achieve an approximate Nash equilibrium for convex-concave functions (i.e., optimality for both $\mathbf{x}$ and $\mathbf{y}$). Instead of averaging, we run two passes of two-timescale GDA or SGDA for min-max problem and max-min problem separately. That is, in the first pass we use $\eta_{\mathbf{x}} \ll \eta_{\mathbf{y}}$ while in the second pass we use $\eta_{\mathbf{x}} \gg \eta_{\mathbf{y}}$. Either pass will return an approximate stationary point for each players, which jointly forms an approximate Nash equilibrium.

2.5 Overview of Proofs

In the nonconvex-strongly-concave setting, our proof involves setting a pair of stepsizes, $(\eta_{\mathbf{x}}, \eta_{\mathbf{y}})$, which force $\{\mathbf{x}_t\}_{t\geq 1}$ to move much more slowly than $\{\mathbf{y}_t\}_{t\geq 1}$. Recall Lemma 2.4.3, which guarantees that $\mathbf{y}^\star(\cdot)$ is κ-Lipschitz:

$$\|\mathbf{y}^\star(\mathbf{x}_1) - \mathbf{y}^\star(\mathbf{x}_2)\| \leq \kappa \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

If $\{\mathbf{x}_t\}_{t\geq 1}$ moves slowly, then $\{\mathbf{y}^\star(\mathbf{x}_t)\}_{t\geq 1}$ also moves slowly. This allows us to perform gradient ascent on a slowly changing strongly-concave function $f(\mathbf{x}_t, \cdot)$, guaranteeing that $\|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_t)\|$ is small in an amortized sense. More precisely, letting the error be $\delta_t = \|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}_t\|^2$, the standard analysis of inexact nonconvex gradient descent implies a descent inequality in which the sum of δ_t provides control:

$$\Phi(\mathbf{x}_{T+1}) - \Phi(\mathbf{x}_0) \leq -\Omega(\eta_{\mathbf{x}}) \left(\sum_{t=0}^{T} \|\nabla \Phi(\mathbf{x}_t)\|^2 \right) + O(\eta_{\mathbf{x}} \ell^2) \left(\sum_{t=0}^{T} \delta_t \right).$$

The remaining step is to show that the second term is always small compared to the first term on the right-hand side. This can be done via a recursion for δ_t as follows:

$$\delta_t \le \gamma\delta_{t-1} + \beta\|\nabla\Phi(\mathbf{x}_{t-1})\|^2,$$

where $\gamma < 1$ and β is small. Thus, δ_t exhibits a linear contraction and $\sum_{t=0}^{T}\delta_t$ can be controlled by the term $\sum_{t=0}^{T}\|\nabla\Phi(\mathbf{x}_t)\|^2$.

In the nonconvex-concave setting, the idea is to set a pair of learning rates $(\eta_{\mathbf{x}}, \eta_{\mathbf{y}})$ which force $\{\mathbf{x}_t\}_{t\ge1}$ to move more slowly than $\{\mathbf{y}_t\}_{t\ge1}$. However, $f(\mathbf{x},\cdot)$ is merely concave and $\mathbf{y}^\star(\cdot)$ is not unique. This means that, even if $\mathbf{x}_1, \mathbf{x}_2$ are extremely close, $\mathbf{y}^\star(\mathbf{x}_1)$ can be dramatically different from $\mathbf{y}^\star(\mathbf{x}_2)$. Thus, $\|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_t)\|$ is no longer a viable error to control.

Fortunately, Lemma 2.4.7 implies that Φ is Lipschitz. That is to say, when the stepsize $\eta_{\mathbf{x}}$ is very small, $\{\Phi(\mathbf{x}_t)\}_{t\ge1}$ moves slowly:

$$|\Phi(\mathbf{x}_t) - \Phi(\mathbf{x}_{t-1})| \le L\|\mathbf{x}_t - \mathbf{x}_{t-1}\| \le \eta_{\mathbf{x}}L^2.$$

Again, this allows us to perform gradient ascent on a slowly changing concave function $f(\mathbf{x}_t,\cdot)$, and guarantees that $\Delta_t = f(\mathbf{x}_t,\mathbf{z}) - f(\mathbf{x}_t,\mathbf{y}_t)$ is small in an amortized sense where $\mathbf{z}\in\mathbf{y}^\star(\mathbf{x}_t)$. The analysis of Davis and Drusvyatskiy [2019] implies that Δ_t comes into the following descent inequality:

$$\Phi_{1/2\ell}(\mathbf{x}_{T+1}) - \Phi_{1/2\ell}(\mathbf{x}_0) \le O(\eta_{\mathbf{x}}\ell)\left(\sum_{t=0}^{T}\Delta_t\right) + O(\eta_{\mathbf{x}}^2\ell L^2(T+1)) - O(\eta_{\mathbf{x}})\left(\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\|^2\right),$$

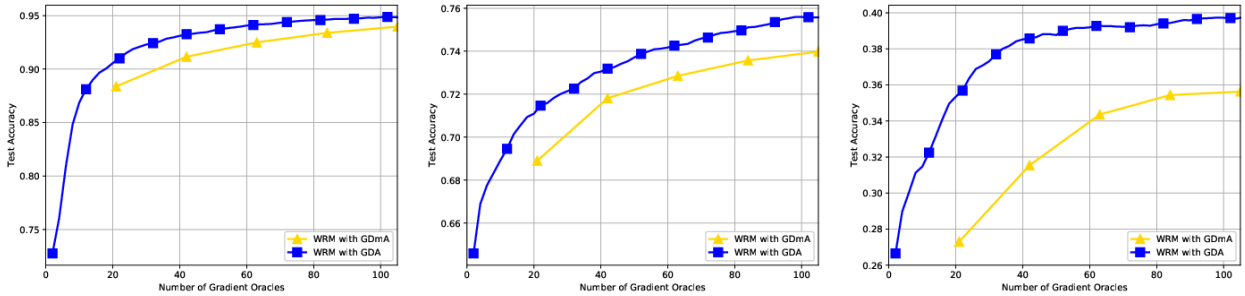
where the first term on the right-hand side is the error term. The remaining step is again to show the error term is small compared to the sum of the first two terms on the right-hand side. To bound the term $\sum_{t=0}^{T}\Delta_t$, we recall the following inequalities and use a telescoping argument (where the optimal point $\mathbf{y}^\star$ does not change):

$$\Delta_t \le \tfrac{\|\mathbf{y}_t-\mathbf{y}^\star\|^2 - \|\mathbf{y}_{t+1}-\mathbf{y}^\star\|^2}{\eta_{\mathbf{y}}}. \tag{2.2}$$

The major challenge here is that the optimal solution $\mathbf{y}^\star(\mathbf{x}_t)$ can change dramatically and the telescoping argument does not go through. An important observation is, however, that (2.2) can be proved if we replace the $\mathbf{y}^\star$ by any $\mathbf{y}\in\mathcal{Y}$, while paying an additional cost that depends on the difference in function value between $\mathbf{y}^\star$ and $\mathbf{y}$. More specifically, we pick a block of size $B = O(\epsilon^2/\eta_{\mathbf{x}})$ and show that the following statement holds for any $s \le \forall t < s+B$,

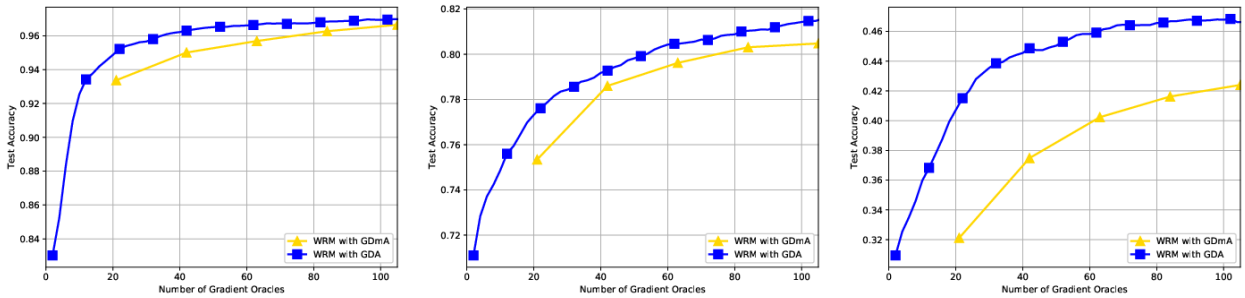
$$\Delta_{t-1} \le O(\ell)(\|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_s)\|^2 - \|\mathbf{y}_{t+1} - \mathbf{y}^\star(\mathbf{x}_s)\|^2) + O(\eta_{\mathbf{x}}L^2)(t-1-s).$$

We perform an analysis on the blocks where the concave problems are similar so the telescoping argument can now work. By carefully choosing $\eta_{\mathbf{x}}$, the term $\sum_{t=0}^{T}\Delta_t$ can also be well controlled.



(a) MNIST (b) Fashion-MNIST (c) CIFAR-10

Figure 2.1: Performance of WRM with GDmA and GDA on MNIST, Fashion-MNIST and CIFAR-10 datasets. We demonstrate test classification accuracy vs. time for different WRM models with GDmA and GDA. Note that $\gamma = 0.4$.



(a) MNIST (b) Fashion-MNIST (c) CIFAR-10

Figure 2.2: Performance of WRM with GDmA and GDA on MNIST, Fashion-MNIST and CIFAR-10 datasets. We demonstrate test classification accuracy vs. time for different WRM models with GDmA and GDA. Note that $\gamma = 1.3$.

2.6 Experiments

We present several empirical results to show that two-timescale GDA outperforms GDmax. The task is to train the empirical Wasserstein robustness model (WRM) [Sinha et al., 2018] over a collection of data samples $\{\xi_i\}_{i=1}^N$ with ℓ_2-norm attack and a penalty parameter $\gamma > 0$. Formally, we have

$$\min_{\mathbf{x}} \max_{\{\mathbf{y}_i\}_{i=1}^N \subseteq \mathcal{Y}} \frac{1}{N} \left[\sum_{i=1}^{N} (\ell(\mathbf{x}, \mathbf{y}_i) - \gamma \|\mathbf{y}_i - \xi_i\|^2) \right]. \tag{2.3}$$

As demonstrated in Sinha et al. [2018], we choose $\gamma > 0$ sufficiently large such that $\ell(\mathbf{x}, \mathbf{y}_i) - \gamma\|\mathbf{y}_i - \xi_i\|^2$ is strongly concave. Thus, this problem is nonconvex-strongly-concave.

We follow the setting of Sinha et al. [2018] and consider training a neural network classifier on three datasets[1]: MNIST, Fashion-MNIST, and CIFAR-10, with the default cross

[1] https://keras.io/datasets/

validation. The architecture consists of 8×8, 6×6 and 5×5 convolutional filter layers with ELU activations followed by a fully connected layer and softmax output. Small and large adversarial perturbation is set with $\gamma \in \{0.4, 1.3\}$ as the same as Sinha et al. [2018]. The baseline approach is denoted as *GDmA* in which $\eta_{\mathbf{x}} = \eta_{\mathbf{y}} = 10^{-3}$ and each inner loop contains 20 gradient ascent. Two-timescale GDA is denoted as *GDA* in which $\eta_{\mathbf{x}} = 5 \times 10^{-5}$ and $\eta_{\mathbf{y}} = 10^{-3}$. Figure 2.1 and 2.2 show that GDA consistently outperforms GDmA on all datasets. Compared to MNIST and Fashion-MNIST, the improvement on CIFAR-10 is more significant which is worthy further exploration in the future.

2.7 Conclusion

We show that two-time-scale GDA and SGDA algorithms return an ϵ-stationary point in $O(\kappa^2\epsilon^{-2})$ gradient evaluations and $O(\kappa^3\epsilon^{-4})$ stochastic gradient evaluations in the nonconvex-strongly-concave case, and $O(\epsilon^{-6})$ gradient evaluations and $O(\epsilon^{-8})$ stochastic gradient evaluations in the nonconvex-concave case. Therefore, these two algorithms are provably efficient in these settings. Future work aim to derive a lower bound for the complexity first-order algorithms in nonconvex-concave minimax problems.

2.8 Proof of Technical Lemmas

We provide complete proofs for the lemmas in Section 2.3 and Section 9.3.

Proof of Lemma 2.3.6. We provide a proof for an expanded version of Lemma 2.3.6.

Lemma 2.8.1 *If f is ℓ-smooth and $\mathcal{Y}$ is bounded, we have*

1. $\Phi_{1/2\ell}(\mathbf{x})$ *and* $\textsc{Prox}_{\Phi/2\ell}(\mathbf{x})$ *are well-defined for* $\forall \mathbf{x} \in \mathbb{R}^m$.
2. $\Phi(\textsc{Prox}_{\Phi/2\ell}(\mathbf{x})) \leq \Phi(\mathbf{x})$ *for any* $\mathbf{x} \in \mathbb{R}^m$.
3. $\Phi_{1/2\ell}$ *is ℓ-smooth with* $\nabla\Phi_{1/2\ell}(\mathbf{x}) = 2\ell(\mathbf{x} - \textsc{Prox}_{\Phi/2\ell}(\mathbf{x}))$.
4. $\Phi_{1/2\ell}(\mathbf{x}') - \Phi_{1/2\ell}(\mathbf{x}) - (\mathbf{x}' - \mathbf{x})^\top \nabla\Phi_{1/2\ell}(\mathbf{x}) \leq (\ell/2)\|\mathbf{x}' - \mathbf{x}\|^2$ *for any* $\mathbf{x}', \mathbf{x} \in \mathbb{R}^m$.

Proof. By the definition of Φ, we have

$$\Psi(\mathbf{x}) \doteq \Phi(\mathbf{x}) + \tfrac{\ell\|\mathbf{x}\|^2}{2} = \max_{\mathbf{y}\in\mathcal{Y}} \{f(\mathbf{x}, \mathbf{y}) + \tfrac{\ell\|\mathbf{x}\|^2}{2}\}.$$

Since f is ℓ-smooth, $f(\mathbf{x}, \mathbf{y}) + (\ell/2)\|\mathbf{x}\|^2$ is convex in $\mathbf{x}$ for any $\mathbf{y} \in \mathcal{Y}$. Since $\mathcal{Y}$ is bounded, Danskin's theorem [Rockafellar, 2015] implies that $\Psi(\mathbf{x})$ is convex. Putting these pieces yields that $\Phi(\mathbf{w}) + \ell\|\mathbf{w} - \mathbf{x}\|^2$ is $(\ell/2)$-strongly convex. This implies that $\Phi_{1/2\ell}(\mathbf{x})$ and $\textsc{Prox}_{\Phi/2\ell}(\mathbf{x})$ are well-defined. Furthermore, by the definition of $\textsc{Prox}_{\Phi/2\ell}(\mathbf{x})$, we have

$$\Phi(\textsc{Prox}_{\Phi/2\ell}(\mathbf{x})) \leq \Phi_{1/2\ell}(\textsc{Prox}_{\Phi/2\ell}(\mathbf{x})) \leq \Phi(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^m.$$

Moreover, Davis and Drusvyatskiy [2019, Lemma 2.2] implies that $\Phi_{1/2\ell}$ is ℓ-smooth with

$$\nabla\Phi_{1/2\ell}(\mathbf{x}) = 2\ell(\mathbf{x} - \textsc{Prox}_{\Phi/2\ell}(\mathbf{x})).$$

It follows from Nesterov [2013b, Theorem 2.1.5] that $\Phi_{1/2\ell}$ satisfies the last inequality. □

Proof of Lemma 2.3.8. Denote $\hat{\mathbf{x}} := \textsc{Prox}_{\Phi/2\ell}(\mathbf{x})$, we have $\nabla\Phi_{1/2\ell}(\mathbf{x}) = 2\ell(\mathbf{x} - \hat{\mathbf{x}})$ (cf. Lemma 2.3.6) and hence $\|\hat{\mathbf{x}} - \mathbf{x}\| = \|\nabla\Phi_{1/2\ell}(\mathbf{x})\|/2\ell$. Furthermore, the optimality condition for $\textsc{Prox}_{\Phi/2\ell}(\mathbf{x})$ implies that $2\ell(\mathbf{x} - \hat{\mathbf{x}}) \in \partial\Phi(\hat{\mathbf{x}})$. Putting these pieces together yields that $\min_{\xi\in\partial\Phi(\hat{\mathbf{x}})} \|\xi\| \le \|\nabla\Phi_{1/2\ell}(\mathbf{x})\|$.

Proof of Lemma 2.4.3. Since $f(\mathbf{x}, \mathbf{y})$ is strongly concave in $\mathbf{y}$ for each $\mathbf{x} \in \mathbb{R}^m$, a function $\mathbf{y}^\star(\cdot)$ is unique and well-defined. Then we claim that $\mathbf{y}^\star(\cdot)$ is κ-Lipschitz. Indeed, let $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^m$, the optimality of $\mathbf{y}^\star(\mathbf{x}_1)$ and $\mathbf{y}^\star(\mathbf{x}_2)$ implies that

$$(\mathbf{y} - \mathbf{y}^\star(\mathbf{x}_1))^\top \nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}^\star(\mathbf{x}_1)) \le 0, \qquad \forall \mathbf{y} \in \mathcal{Y}, \tag{2.4}$$

$$(\mathbf{y} - \mathbf{y}^\star(\mathbf{x}_2))^\top \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}^\star(\mathbf{x}_2)) \le 0, \qquad \forall \mathbf{y} \in \mathcal{Y}. \tag{2.5}$$

Letting $\mathbf{y} = \mathbf{y}^\star(\mathbf{x}_2)$ in (2.4) and $\mathbf{y} = \mathbf{y}^\star(\mathbf{x}_1)$ in (2.5) and summing the resulting two inequalities yields

$$(\mathbf{y}^\star(\mathbf{x}_2) - \mathbf{y}^\star(\mathbf{x}_1))^\top (\nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}^\star(\mathbf{x}_1)) - \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}^\star(\mathbf{x}_2))) \le 0. \tag{2.6}$$

Recall that $f(\mathbf{x}_1, \cdot)$ is μ-strongly concave, we have

$$(\mathbf{y}^\star(\mathbf{x}_2) - \mathbf{y}^\star(\mathbf{x}_1))^\top (\nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}^\star(\mathbf{x}_2)) - \nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}^\star(\mathbf{x}_1))) + \mu\|\mathbf{y}^\star(\mathbf{x}_2) - \mathbf{y}^\star(\mathbf{x}_1)\|^2 \le 0. \tag{2.7}$$

Then we conclude the desired result by combining (2.6) and (2.7) with ℓ-smoothness of f, i.e.,

$$\begin{aligned} \mu\|\mathbf{y}^\star(\mathbf{x}_2) - \mathbf{y}^\star(\mathbf{x}_1)\|^2 &\le (\mathbf{y}^\star(\mathbf{x}_2) - \mathbf{y}^\star(\mathbf{x}_1))^\top (\nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}^\star(\mathbf{x}_2)) - \nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}^\star(\mathbf{x}_2))) \\ &\le \ell\|\mathbf{y}^\star(\mathbf{x}_2) - \mathbf{y}^\star(\mathbf{x}_1)\|\|\mathbf{x}_2 - \mathbf{x}_1\|. \end{aligned}$$

Since $\mathbf{y}^\star(\mathbf{x})$ is unique and $\mathcal{Y}$ is convex and bounded, we conclude from Danskin's theorem [Rockafellar, 2015] that Φ is differentiable with $\nabla\Phi(\mathbf{x}) = \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^\star(\mathbf{x}))$. Since $\nabla\Phi(\mathbf{x}) = \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^\star(\mathbf{x}))$, we have

$$\|\nabla\Phi(\mathbf{x}) - \nabla\Phi(\mathbf{x}')\| = \|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^\star(\mathbf{x})) - \nabla_{\mathbf{x}} f(\mathbf{x}', \mathbf{y}^\star(\mathbf{x}'))\| \le \ell(\|\mathbf{x} - \mathbf{x}'\| + \|\mathbf{y}^\star(\mathbf{x}) - \mathbf{y}^\star(\mathbf{x}')\|).$$

Since $\mathbf{y}^\star(\cdot)$ is κ-Lipschitz, we conclude the desired result by plugging $\|\mathbf{y}^\star(\mathbf{x}) - \mathbf{y}^\star(\mathbf{x}')\| \le \kappa$. Since $\kappa \ge 1$, Φ is $2\kappa\ell$-smooth. Nesterov [2013b, Theorem 2.1.5] implies the last inequality.

Proof of Lemma 2.4.7. By the proof in Lemma 2.8.1, Φ is ℓ-weakly convex and $\partial\Phi(\mathbf{x}) = \partial\Psi(\mathbf{x}) - \ell\mathbf{x}$ where $\Psi(\mathbf{x}) = \max_{\mathbf{y}\in\mathcal{Y}}\{f(\mathbf{x}, \mathbf{y}) + (\ell/2)\|\mathbf{x}\|^2\}$. Since $f(\mathbf{x}, \mathbf{y}) + (\ell/2)\|\mathbf{x}\|^2$ is convex in $\mathbf{x}$ for each $\mathbf{y} \in \mathcal{Y}$ and $\mathcal{Y}$ is bounded, Danskin's theorem implies that $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^\star(\mathbf{x})) + \ell\mathbf{x} \in \partial\Psi(\mathbf{x})$. Putting these pieces together yields that $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^\star(\mathbf{x})) \in \partial\Phi(\mathbf{x})$.

Proof of Lemma on Stochastic Gradient. The following lemma establishes some properties of the stochastic gradients sampled at each iteration.

Lemma 2.8.2 *$\frac{1}{M}\sum_{i=1}^M G_{\mathbf{x}}(\mathbf{x}_t,\mathbf{y}_t,\xi_i)$ and $\frac{1}{M}\sum_{i=1}^M G_{\mathbf{y}}(\mathbf{x}_t,\mathbf{y}_t,\xi_i)$ are unbiased and have bounded variance,*

$$\mathbb{E}\left[\tfrac{1}{M}\textstyle\sum_{i=1}^M G_{\mathbf{x}}(\mathbf{x}_t,\mathbf{y}_t,\xi_i)\right] = \nabla_{\mathbf{x}} f(\mathbf{x}_t,\mathbf{y}_t), \quad \mathbb{E}\left[\left\|\tfrac{1}{M}\textstyle\sum_{i=1}^M G_{\mathbf{x}}(\mathbf{x}_t,\mathbf{y}_t,\xi_i)\right\|^2\right] \le \|\nabla_{\mathbf{x}} f(\mathbf{x}_t,\mathbf{y}_t)\|^2 + \tfrac{\sigma^2}{M},$$
$$\mathbb{E}\left[\tfrac{1}{M}\textstyle\sum_{i=1}^M G_{\mathbf{y}}(\mathbf{x}_t,\mathbf{y}_t,\xi_i)\right] = \nabla_{\mathbf{y}} f(\mathbf{x}_t,\mathbf{y}_t), \quad \mathbb{E}\left[\left\|\tfrac{1}{M}\textstyle\sum_{i=1}^M G_{\mathbf{y}}(\mathbf{x}_t,\mathbf{y}_t,\xi_i)\right\|^2\right] \le \|\nabla_{\mathbf{y}} f(\mathbf{x}_t,\mathbf{y}_t)\|^2 + \tfrac{\sigma^2}{M}.$$

Proof. Since $G = (G_{\mathbf{x}}, G_{\mathbf{y}})$ is unbiased, we have

$$\mathbb{E}\left[\frac{1}{M}\sum_{i=1}^M G_{\mathbf{x}}(\mathbf{x}_t,\mathbf{y}_t,\xi_i)\right] = \nabla_{\mathbf{x}} f(\mathbf{x}_t,\mathbf{y}_t), \quad \mathbb{E}\left[\frac{1}{M}\sum_{i=1}^M G_{\mathbf{y}}(\mathbf{x}_t,\mathbf{y}_t,\xi_i)\right] = \nabla_{\mathbf{y}} f(\mathbf{x}_t,\mathbf{y}_t).$$

Furthermore, we have

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^M G_{\mathbf{x}}(\mathbf{x}_t,\mathbf{y}_t,\xi_i) - \nabla_{\mathbf{x}} f(\mathbf{x}_t,\mathbf{y}_t)\right\|^2\right] = \frac{\sum_{i=1}^M \mathbb{E}\left[\|G_{\mathbf{x}}(\mathbf{x}_t,\mathbf{y}_t,\xi_i) - \nabla_{\mathbf{x}} f(\mathbf{x}_t,\mathbf{y}_t)\|^2\right]}{M^2} \le \frac{\sigma^2}{M},$$
$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^M G_{\mathbf{y}}(\mathbf{x}_t,\mathbf{y}_t,\xi_i) - \nabla_{\mathbf{y}} f(\mathbf{x}_t,\mathbf{y}_t)\right\|^2\right] = \frac{\sum_{i=1}^M \mathbb{E}\left[\|G_{\mathbf{y}}(\mathbf{x}_t,\mathbf{y}_t,\xi_i) - \nabla_{\mathbf{y}} f(\mathbf{x}_t,\mathbf{y}_t)\|^2\right]}{M^2} \le \frac{\sigma^2}{M}.$$

Putting these pieces together yields the desired result. □

2.9 Proof for Propositions 2.4.11 and 2.4.12

We provide the detailed proof of Propositions 2.4.11 and 2.4.12.

Proof of Proposition 2.4.11. Assume that a point $\hat{\mathbf{x}}$ satisfies that $\|\nabla\Phi(\hat{\mathbf{x}})\| \le \epsilon$, the optimization problem $\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y})$ is strongly concave (cf. Assumption 2.4.2) and $\mathbf{y}^\star(\hat{\mathbf{x}})$ is uniquely defined. We apply gradient descent for solving such problem and obtain a point $\mathbf{y}' \in \mathcal{Y}$ satisfying that

$$\mathbf{y}^+ = \mathcal{P}_{\mathcal{Y}}(\mathbf{y}' + (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}},\mathbf{y}')), \quad \|\mathbf{y}^+ - \mathbf{y}'\| \le \epsilon/\ell, \quad \|\mathbf{y}^+ - \mathbf{y}^\star(\hat{\mathbf{x}})\| \le \epsilon.$$

If $\|\nabla\Phi(\hat{\mathbf{x}})\| \le \epsilon$, we have

$$\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\mathbf{y}^+)\| \le \|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\mathbf{y}^+) - \nabla\Phi(\hat{\mathbf{x}})\| + \|\nabla\Phi(\hat{\mathbf{x}})\| = \|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\mathbf{y}^+) - \nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\mathbf{y}^\star(\hat{\mathbf{x}}))\| + \epsilon.$$

Since $f(\cdot,\cdot)$ is ℓ-smooth, we have

$$\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\mathbf{y}^+)\| \le \ell\|\mathbf{y}^+ - \mathbf{y}^\star(\hat{\mathbf{x}})\| + \epsilon = O(\epsilon).$$

The required number of gradient evaluations is $\mathcal{O}(\kappa \log(1/\epsilon))$. This argument holds for applying stochastic gradient with proper stepsize and the required number of stochastic gradient evaluations is $\mathcal{O}(\epsilon^{-2})$.

Conversely, if a point $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ satisfies that

$$\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+)\| \le \epsilon/\kappa, \quad \|\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}\| \le \epsilon/\kappa\ell,$$

where $\hat{\mathbf{y}}^+ = \mathcal{P}_{\mathcal{Y}}(\hat{\mathbf{y}} + (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}}))$. Then, we have

$$\|\nabla \Phi(\hat{\mathbf{x}})\| \le \|\nabla \Phi(\hat{\mathbf{x}}) - \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+)\| + \|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+)\| \le \ell \|\hat{\mathbf{y}}^+ - \mathbf{y}^\star(\hat{\mathbf{x}})\| + \epsilon/\kappa.$$

Since $f(\hat{\mathbf{x}}, \cdot)$ is μ-strongly-concave over $\mathcal{Y}$, the global error bound condition [Drusvyatskiy and Lewis, 2018] holds true here and we have

$$\|\hat{\mathbf{y}}^+ - \mathbf{y}^\star(\hat{\mathbf{x}})\| \le \|\hat{\mathbf{y}} - \mathbf{y}^\star(\hat{\mathbf{x}})\| \le \kappa \|\mathcal{P}_{\mathcal{Y}}(\hat{\mathbf{y}} + (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})) - \hat{\mathbf{y}}\| \le \epsilon/\ell.$$

Therefore, we conclude that

$$\|\nabla \Phi(\hat{\mathbf{x}})\| \le \epsilon + \epsilon/\kappa = O(\epsilon).$$

This completes the proof.

Proof of Proposition 2.4.12. Assume that a point $\hat{\mathbf{x}}$ satisfies that $\|\nabla \Phi_{1/2\ell}(\hat{\mathbf{x}})\| \le \epsilon$, the objective function $f(\mathbf{x}, \mathbf{y}) + \ell\|\mathbf{x} - \hat{\mathbf{x}}\|^2$ is strongly convex in $\mathbf{x}$ and concave in $\mathbf{y}$ (cf. Assumption 2.4.6) and $\mathbf{x}^\star(\hat{\mathbf{x}}) = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^m} \Phi(\mathbf{x}) + \ell\|\mathbf{x} - \hat{\mathbf{x}}\|^2$ is uniquely defined. We can apply extragradient algorithm for solving such problem and obtain a point $(\mathbf{x}', \mathbf{y}')$ satisfying that

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}', \mathbf{y}^+) + 2\ell(\mathbf{x}' - \hat{\mathbf{x}})\| \le \epsilon, \quad \|\mathbf{y}^+ - \mathbf{y}'\| \le \epsilon/\ell, \quad \|\mathbf{x}' - \mathbf{x}^\star(\hat{\mathbf{x}})\| \le \epsilon/\ell.$$

where $\mathbf{y}^+ = \mathcal{P}_{\mathcal{Y}}(\mathbf{y}' + (1/\ell)\nabla_{\mathbf{y}} f(\mathbf{x}', \mathbf{y}'))$. Since $2\ell\|\mathbf{x}^\star(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\| = \|\nabla \Phi_{1/2\ell}(\hat{\mathbf{x}})\| \le \epsilon$, we have

$$\begin{aligned} \|\nabla_{\mathbf{x}} f(\mathbf{x}', \mathbf{y}^+)\| &\le \|\nabla_{\mathbf{x}} f(\mathbf{x}', \mathbf{y}^+) + 2\ell(\mathbf{x}' - \hat{\mathbf{x}})\| + 2\ell\|\mathbf{x}' - \hat{\mathbf{x}}\| \le \epsilon + 2\ell\|\mathbf{x}' - \mathbf{x}^\star(\hat{\mathbf{x}})\| + 2\ell\|\mathbf{x}^\star(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\| \\ &\le 3\epsilon + \epsilon = O(\epsilon). \end{aligned}$$

The required number of gradient evaluations is indeed $O(\epsilon^{-2})$ [Mokhtari et al., 2020a]. This argument holds for applying stochastic mirror-prox algorithm and the required number of stochastic gradient evaluations is $O(\epsilon^{-4})$ [Juditsky et al., 2011].

Conversely, we let $\hat{\mathbf{y}}^+ = \mathcal{P}_{\mathcal{Y}}(\hat{\mathbf{y}} + (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}}))$ for simplicity. By definition, we have

$$\|\nabla \Phi_{1/2\ell}(\hat{\mathbf{x}})\|^2 = 4\ell^2 \|\hat{\mathbf{x}} - \mathbf{x}^\star(\hat{\mathbf{x}})\|^2.$$

Since $\Phi(\cdot) + \ell\|\cdot - \hat{\mathbf{x}}\|^2$ is $\ell/2$-strongly-convex, we have

$$\begin{aligned} &\max_{\mathbf{y} \in \mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}^\star(\hat{\mathbf{x}}), \mathbf{y}) - \ell\|\hat{\mathbf{x}} - \mathbf{x}^\star(\hat{\mathbf{x}})\|^2 \qquad (2.8) \\ &= \Phi(\hat{\mathbf{x}}) - \Phi(\mathbf{x}^\star(\hat{\mathbf{x}})) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\|^2 \ge \tfrac{\ell\|\hat{\mathbf{x}} - \mathbf{x}^\star(\hat{\mathbf{x}})\|^2}{4} = \tfrac{\|\nabla \Phi_{1/2\ell}(\hat{\mathbf{x}})\|^2}{16\ell}. \end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star(\hat{\mathbf{x}}),\mathbf{y}) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}})-\hat{\mathbf{x}}\|^2 \\
&\le \max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - f(\hat{\mathbf{x}},\hat{\mathbf{y}}^+) + f(\hat{\mathbf{x}},\hat{\mathbf{y}}^+) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star(\hat{\mathbf{x}}),\mathbf{y}) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}})-\hat{\mathbf{x}}\|^2 \\
&\le \max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - f(\hat{\mathbf{x}},\hat{\mathbf{y}}^+) + (f(\hat{\mathbf{x}},\hat{\mathbf{y}}^+) - f(\mathbf{x}^\star(\hat{\mathbf{x}}),\hat{\mathbf{y}}^+) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}})-\hat{\mathbf{x}}\|^2) \\
&\le \max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - f(\hat{\mathbf{x}},\hat{\mathbf{y}}^+) + (\|\hat{\mathbf{x}}-\mathbf{x}^\star(\hat{\mathbf{x}})\|\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\hat{\mathbf{y}}^+)\| - \ell\|\hat{\mathbf{x}}-\mathbf{x}^\star(\hat{\mathbf{x}})\|^2) \\
&\le \max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - f(\hat{\mathbf{x}},\hat{\mathbf{y}}^+) + \tfrac{\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\hat{\mathbf{y}}^+)\|^2}{4\ell}.
\end{aligned}$$

By the definition of $\hat{\mathbf{y}}^+$, we have

$$(\mathbf{y}-\hat{\mathbf{y}}^+)^\top(\hat{\mathbf{y}}^+ - \hat{\mathbf{y}} - (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}},\hat{\mathbf{y}})) \ge 0 \text{ for all } \mathbf{y}\in\mathcal{Y}.$$

Together with the ℓ-smoothness of the function $f(\hat{\mathbf{x}},\cdot)$ and the boundedness of $\mathcal{Y}$, we have

$$f(\hat{\mathbf{x}},\mathbf{y}) - f(\hat{\mathbf{x}},\hat{\mathbf{y}}^+) \le \tfrac{\ell}{2}(\|\mathbf{y}-\hat{\mathbf{y}}\|^2 - \|\mathbf{y}-\hat{\mathbf{y}}^+\|^2) \le \ell D\|\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}\| \text{ for all } \mathbf{y}\in\mathcal{Y}.$$

Putting these pieces together yields that

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star(\hat{\mathbf{x}}),\mathbf{y}) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}})-\hat{\mathbf{x}}\|^2 \le \ell D\|\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}\| + \tfrac{\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\hat{\mathbf{y}}^+)\|^2}{4\ell}.$$

Since a point $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ satisfies $\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\hat{\mathbf{y}}^+)\| \le \epsilon^2/(\ell D)$ and $\|\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}\| \le \epsilon^2/(\ell^2 D)$, we have

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star(\hat{\mathbf{x}}),\mathbf{y}) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}})-\hat{\mathbf{x}}\|^2 \le \tfrac{\epsilon^2}{\ell} + \tfrac{\epsilon^4}{4\ell^3 D^2}. \tag{2.9}$$

Putting these pieces together yields that $\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\| = O(\epsilon)$. This completes the proof.

2.10 Proofs for Nonconvex-Strongly-Concave Setting

We first specify the choice of parameters in Theorems 2.4.4 and 2.4.5. Then, we present the proof for nonconvex-strongly-concave setting with several technical lemmas. Note first that the case of $\ell D \lesssim \epsilon$ is trivial. Indeed, this means that the set $\mathcal{Y}$ is sufficiently small such that a single gradient ascent step is enough for approaching the ϵ-neighborhood of the optimal solution. In this case, the nonconvex-strongly-concave minimax problem reduces to a nonconvex smooth minimization problem, which has been studied in the existing literature.

We present the full version of Theorems 2.4.4 and 2.4.5 with the detailed choice of $\eta_{\mathbf{x}}$, $\eta_{\mathbf{y}}$ and M which are important to subsequent analysis.

Theorem 2.10.1 *Under Assumption 2.4.2 and letting the step sizes $\eta_{\mathbf{x}}$ and $\eta_{\mathbf{y}}$ be chosen as $\eta_{\mathbf{x}} = 1/[16(\kappa+1)^2\ell]$ and $\eta_{\mathbf{y}} = 1/\ell$, the iteration complexity of Algorithm 1 to return an ϵ-stationary point is bounded by*

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2}{\epsilon^2}\right),$$

which is also the total gradient complexity of the algorithm.

Theorem 2.10.2 *Under Assumptions 2.4.1 and 2.4.2 and letting the step sizes $\eta_{\mathbf{x}}$ and $\eta_{\mathbf{y}}$ be the same in Theorem 2.4.4 with the batch size $M = \max\{1, 48\kappa\sigma^2\epsilon^{-2}\}$, the number of iterations required by Algorithm 2 to return an ϵ-stationary point is bounded by $O((\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2)\epsilon^{-2})$ which gives the total gradient complexity of the algorithm:*

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2}{\epsilon^2}\max\left\{1, \frac{\kappa\sigma^2}{\epsilon^2}\right\}\right).$$

We present three key lemmas which are important for the subsequent analysis.

Lemma 2.10.3 *For Algorithm 1, the iterates $\{\mathbf{x}_t\}_{t\geq 1}$ satisfies the following inequality,*

$$\Phi(\mathbf{x}_t) \leq \Phi(\mathbf{x}_{t-1}) - \left(\tfrac{\eta_{\mathbf{x}}}{2} - 2\eta_{\mathbf{x}}^2\kappa\ell\right)\|\nabla\Phi(\mathbf{x}_{t-1})\|^2 + \left(\tfrac{\eta_{\mathbf{x}}}{2} + 2\eta_{\mathbf{x}}^2\kappa\ell\right)\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2.$$

For Algorithm 2, the iterates $\{\mathbf{x}_t\}_{t\geq 1}$ satisfy the following inequality:

$$\begin{aligned}\mathbb{E}\left[\Phi(\mathbf{x}_t)\right] \;\leq\; & \mathbb{E}\left[\Phi(\mathbf{x}_{t-1})\right] - \left(\tfrac{\eta_{\mathbf{x}}}{2} - 2\eta_{\mathbf{x}}^2\kappa\ell\right)\mathbb{E}[\|\nabla\Phi(\mathbf{x}_{t-1})\|^2] \\ & + \left(\tfrac{\eta_{\mathbf{x}}}{2} + 2\eta_{\mathbf{x}}^2\kappa\ell\right)\mathbb{E}[\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2] + \tfrac{\eta_{\mathbf{x}}^2\kappa\ell\sigma^2}{M}.\end{aligned}$$

Proof. We first consider the deterministic setting. Since Φ is $(\ell + \kappa\ell)$-smooth, we have

$$\Phi(\mathbf{x}_t) - \Phi(\mathbf{x}_{t-1}) - (\mathbf{x}_t - \mathbf{x}_{t-1})^\top\nabla\Phi(\mathbf{x}_{t-1}) \leq \kappa\ell\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2. \tag{2.10}$$

Plugging $\mathbf{x}_t - \mathbf{x}_{t-1} = -\eta_{\mathbf{x}}\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$ into (2.10) yields that

$$\begin{aligned}\Phi(\mathbf{x}_t) \;\leq\; & \Phi(\mathbf{x}_{t-1}) - \eta_{\mathbf{x}}\|\nabla\Phi(\mathbf{x}_{t-1})\|^2 + \eta_{\mathbf{x}}^2\kappa\ell\|\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 \\ & + \eta_{\mathbf{x}}(\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}))^\top\nabla\Phi(\mathbf{x}_{t-1}).\end{aligned} \tag{2.11}$$

By Young's inequality, we have

$$(\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}))^\top\nabla\Phi(\mathbf{x}_{t-1}) \leq \tfrac{\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 + \|\nabla\Phi(\mathbf{x}_{t-1})\|^2}{2}. \tag{2.12}$$

By the Cauchy-Schwartz inequality, we have

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 \leq 2(\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 + \|\nabla\Phi(\mathbf{x}_{t-1})\|^2). \tag{2.13}$$

Plugging (2.12) and (2.13) into (2.11) yields the first desired inequality.

We proceed to the stochastic setting. Plugging $\mathbf{x}_t - \mathbf{x}_{t-1} = -\eta_{\mathbf{x}}(\frac{1}{M}\sum_{i=1}^M G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i))$ into (2.10) yields that

$$\begin{aligned}\Phi(\mathbf{x}_t) \;\leq\; & \Phi(\mathbf{x}_{t-1}) - \eta_{\mathbf{x}}\|\nabla\Phi(\mathbf{x}_{t-1})\|^2 + \eta_{\mathbf{x}}^2\kappa\ell\left\|\frac{1}{M}\sum_{i=1}^M G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i)\right\|^2 \\ & + \eta_{\mathbf{x}}\left(\nabla\Phi(\mathbf{x}_{t-1}) - \left(\frac{1}{M}\sum_{i=1}^M G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i)\right)\right)^\top\nabla\Phi(\mathbf{x}_t).\end{aligned}$$

Taking the expectation on both sides, conditioned on $(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$, yields that

$$\begin{aligned}\mathbb{E}[\Phi(\mathbf{x}_t) \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1}] \leq \Phi(\mathbf{x}_{t-1}) - \eta_{\mathbf{x}}\|\nabla\Phi(\mathbf{x}_{t-1})\|^2 + \eta_{\mathbf{x}}^2\kappa\ell\|\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 \\ +\eta_{\mathbf{x}}(\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}))^\top \nabla\Phi(\mathbf{x}_{t-1}) + \eta_{\mathbf{x}}^2\kappa\ell\|\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 \\ +\eta_{\mathbf{x}}^2\kappa\ell\mathbb{E}\left[\left\|\tfrac{1}{M}\sum_{i=1}^{M} G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right\|^2 \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1}\right].\end{aligned} \tag{2.14}$$

Plugging (2.12) and (2.13) into (2.14) and taking the expectation of both sides yields the second desired inequality. □

Lemma 2.10.4 *For Algorithm 1, let $\delta_t = \|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}_t\|^2$, the following statement holds true,*

$$\delta_t \leq \left(1 - \tfrac{1}{2\kappa} + 4\kappa^3\ell^2\eta_{\mathbf{x}}^2\right)\delta_{t-1} + 4\kappa^3\eta_{\mathbf{x}}^2\|\nabla\Phi(\mathbf{x}_{t-1})\|^2.$$

For Algorithm 2, let $\delta_t = \mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}_t\|^2]$, the following statement holds true,

$$\delta_t \leq \left(1 - \tfrac{1}{2\kappa} + 8\kappa^3\ell^2\eta_{\mathbf{x}}^2\right)\delta_{t-1} + 8\kappa^3\eta_{\mathbf{x}}^2\mathbb{E}\left[\|\nabla\Phi(\mathbf{x}_{t-1})\|^2\right] + \tfrac{4\sigma^2\kappa^3\eta_{\mathbf{x}}^2}{M} + \tfrac{2\sigma^2}{\ell^2 M}.$$

Proof. We first prove the results for the deterministic setting. Since $f(\mathbf{x}_t, \cdot)$ is μ-strongly concave and $\eta_{\mathbf{y}} = 1/\ell$, we have

$$\|\mathbf{y}^\star(\mathbf{x}_{t-1}) - \mathbf{y}_t\|^2 \leq \left(1 - \tfrac{1}{\kappa}\right)\delta_{t-1}. \tag{2.15}$$

By Young's inequality, we have

$$\begin{aligned}\delta_t &\leq \left(1 + \tfrac{1}{2(\kappa-1)}\right)\|\mathbf{y}^\star(\mathbf{x}_{t-1}) - \mathbf{y}_t\|^2 + (1 + 2(\kappa-1))\|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}^\star(\mathbf{x}_{t-1})\|^2 \\ &\leq \left(\tfrac{2\kappa-1}{2\kappa-2}\right)\|\mathbf{y}^\star(\mathbf{x}_{t-1}) - \mathbf{y}_t\|^2 + 2\kappa\|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}^\star(\mathbf{x}_{t-1})\|^2 \\ &\overset{(2.15)}{\leq} \left(1 - \tfrac{1}{2\kappa}\right)\delta_{t-1} + 2\kappa\|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}^\star(\mathbf{x}_{t-1})\|^2.\end{aligned}$$

Since $\mathbf{y}^\star(\cdot)$ is κ-Lipschitz, $\|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}^\star(\mathbf{x}_{t-1})\| \leq \kappa\|\mathbf{x}_t - \mathbf{x}_{t-1}\|$. Furthermore, we have

$$\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 = \eta_{\mathbf{x}}^2\|\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 \leq 2\eta_{\mathbf{x}}^2\ell^2\delta_{t-1} + 2\eta_{\mathbf{x}}^2\|\nabla\Phi(\mathbf{x}_{t-1})\|^2.$$

Putting these pieces together yields the first desired inequality.

We proceed to the stochastic setting. Since $f(\mathbf{x}_t, \cdot)$ is μ-strongly concave and $\eta_{\mathbf{y}} = 1/\ell$, we have

$$\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_{t-1}) - \mathbf{y}_t\|^2] \leq \left(1 - \tfrac{1}{\kappa}\right)\delta_{t-1} + \tfrac{\sigma^2}{\ell^2 M}. \tag{2.16}$$

By Young's inequality, we have

$$\begin{aligned}\delta_t &\leq \left(1 + \tfrac{1}{2(\max\{\kappa,2\}-1)}\right)\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_{t-1}) - \mathbf{y}_t\|^2] + (1 + 2(\max\{\kappa,2\} - 1))\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}^\star(\mathbf{x}_{t-1})\|^2] \\ &\leq \left(\tfrac{2\max\{\kappa,2\}-1}{2\max\{\kappa,2\}-2}\right)\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_{t-1}) - \mathbf{y}_t\|^2] + 4\kappa\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}^\star(\mathbf{x}_{t-1})\|^2] \\ &\overset{(2.16)}{\leq} \left(1 - \tfrac{1}{2\kappa}\right)\delta_{t-1} + 4\kappa\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}^\star(\mathbf{x}_{t-1})\|^2] + \tfrac{2\sigma^2}{\ell^2 M}.\end{aligned}$$

Since $\mathbf{y}^\star(\cdot)$ is κ-Lipschitz, $\|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}^\star(\mathbf{x}_{t-1})\| \leq \kappa\|\mathbf{x}_t - \mathbf{x}_{t-1}\|$. Furthermore, we have

$$\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2] = \eta_{\mathbf{x}}^2 \mathbb{E}\left[\left\|\tfrac{1}{M}\sum_{i=1}^{M} G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i)\right\|^2\right] \leq 2\eta_{\mathbf{x}}^2\ell^2\delta_{t-1} + 2\eta_{\mathbf{x}}^2\mathbb{E}[\|\nabla\Phi(\mathbf{x}_{t-1})\|^2] + \tfrac{\eta_{\mathbf{x}}^2\sigma^2}{M}.$$

Putting these pieces together yields the second desired inequality. □

Lemma 2.10.5 *For Algorithm 1, let $\delta_t = \|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}_t\|^2$, the following statement holds true,*

$$\Phi(\mathbf{x}_t) \leq \Phi(\mathbf{x}_{t-1}) - \tfrac{7\eta_{\mathbf{x}}}{16}\|\nabla\Phi(\mathbf{x}_{t-1})\|^2 + \tfrac{9\eta_{\mathbf{x}}\ell^2\delta_{t-1}}{16}.$$

For Algorithm 2, let $\delta_t = \mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}_t\|^2]$, the following statement holds true,

$$\mathbb{E}[\Phi(\mathbf{x}_t)] \leq \mathbb{E}[\Phi(\mathbf{x}_{t-1})] - \tfrac{7\eta_{\mathbf{x}}}{16}\mathbb{E}[\|\nabla\Phi(\mathbf{x}_{t-1})\|^2] + \tfrac{9\eta_{\mathbf{x}}\ell^2\delta_{t-1}}{16} + \tfrac{\eta_{\mathbf{x}}^2\kappa\ell\sigma^2}{M}.$$

Proof. For two-timescale GDA and SGDA, $\eta_{\mathbf{x}} = 1/16(\kappa+1)\ell$ and hence

$$\tfrac{7\eta_{\mathbf{x}}}{16} \leq \tfrac{\eta_{\mathbf{x}}}{2} - 2\eta_{\mathbf{x}}^2\kappa\ell \leq \tfrac{\eta_{\mathbf{x}}}{2} + 2\eta_{\mathbf{x}}^2\kappa\ell \leq \tfrac{9\eta_{\mathbf{x}}}{16}. \tag{2.17}$$

Combining (2.17) with the first inequality in Lemma 2.10.3 yields that

$$\Phi(\mathbf{x}_t) \leq \Phi(\mathbf{x}_{t-1}) - \tfrac{7\eta_{\mathbf{x}}}{16}\|\nabla\Phi(\mathbf{x}_{t-1})\|^2 + \tfrac{9\eta_{\mathbf{x}}}{16}\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2.$$

Since $\nabla\Phi(\mathbf{x}_{t-1}) = \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1}))$, we have

$$\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 \leq \ell^2\|\mathbf{y}^\star(\mathbf{x}_{t-1}) - \mathbf{y}_{t-1}\|^2 = \ell^2\delta_{t-1}.$$

Putting these pieces together yields the first desired inequality.

We proceed to the results for the stochastic setting, combining (2.17) with the second inequality in Lemma 2.10.3 yields that

$$\mathbb{E}[\Phi(\mathbf{x}_t)] \leq \mathbb{E}[\Phi(\mathbf{x}_{t-1})] - \tfrac{7\eta_{\mathbf{x}}}{16}\mathbb{E}[\|\nabla\Phi(\mathbf{x}_{t-1})\|^2] + \tfrac{9\eta_{\mathbf{x}}}{16}\mathbb{E}[\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2] + \tfrac{\eta_{\mathbf{x}}^2\kappa\ell\sigma^2}{M}.$$

Since $\nabla\Phi(\mathbf{x}_{t-1}) = \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1}))$, we have

$$\mathbb{E}[\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2] \leq \ell^2\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_{t-1}) - \mathbf{y}_{t-1}\|^2] = \ell^2\delta_{t-1}.$$

Putting these pieces together yields the second desired inequality. □

Proof of Theorem 2.10.1. We define $\gamma = 1 - 1/2\kappa + 4\kappa^3\ell^2\eta_{\mathbf{x}}^2$ throughout. Performing the first inequality in Lemma 2.10.4 recursively yields that

$$\delta_t \leq \gamma^t\delta_0 + 4\kappa^3\eta_{\mathbf{x}}^2\left(\sum_{j=0}^{t-1}\gamma^{t-1-j}\|\nabla\Phi(\mathbf{x}_j)\|^2\right) \leq \gamma^t D^2 + 4\kappa^3\eta_{\mathbf{x}}^2\left(\sum_{j=0}^{t-1}\gamma^{t-1-j}\|\nabla\Phi(\mathbf{x}_j)\|^2\right). \quad (2.18)$$

Combining (2.18) with the first inequality in Lemma 2.10.5 yields that,

$$\Phi(\mathbf{x}_t) \leq \Phi(\mathbf{x}_{t-1}) - \tfrac{7\eta_{\mathbf{x}}}{16}\|\nabla\Phi(\mathbf{x}_{t-1})\|^2 + \tfrac{9\eta_{\mathbf{x}}\ell^2\gamma^{t-1}D^2}{16} + \tfrac{9\eta_{\mathbf{x}}^3\ell^2\kappa^3}{4}\left(\sum_{j=0}^{t-2}\gamma^{t-2-j}\|\nabla\Phi(\mathbf{x}_j)\|^2\right). \quad (2.19)$$

Summing up (2.19) over $t = 1, 2, \ldots, T+1$ and rearranging the terms yields that

$$\Phi(\mathbf{x}_{T+1}) \leq \Phi(\mathbf{x}_0) - \tfrac{7\eta_{\mathbf{x}}}{16}\sum_{t=0}^{T}\|\nabla\Phi(\mathbf{x}_t)\|^2 + \tfrac{9\eta_{\mathbf{x}}\ell^2D^2}{16}\left(\sum_{t=0}^{T}\gamma^t\right) + \tfrac{9\eta_{\mathbf{x}}^3\ell^2\kappa^3}{4}\left(\sum_{t=1}^{T+1}\sum_{j=0}^{t-2}\gamma^{t-2-j}\|\nabla\Phi(\mathbf{x}_j)\|^2\right).$$

Since $\eta_{\mathbf{x}} = 1/16(\kappa+1)^2\ell$, we have $\gamma \leq 1 - \frac{1}{4\kappa}$ and $\frac{9\eta_{\mathbf{x}}^3\ell^2\kappa^3}{4} \leq \frac{9\eta_{\mathbf{x}}}{1024\kappa}$. This implies that $\sum_{t=0}^{T}\gamma^t \leq 4\kappa$ and

$$\sum_{t=1}^{T+1}\sum_{j=0}^{t-2}\gamma^{t-2-j}\|\nabla\Phi(\mathbf{x}_j)\|^2 \leq 4\kappa\left(\sum_{t=0}^{T}\|\nabla\Phi(\mathbf{x}_t)\|^2\right)$$

Putting these pieces together yields that

$$\Phi(\mathbf{x}_{T+1}) \leq \Phi(\mathbf{x}_0) - \tfrac{103\eta_{\mathbf{x}}}{256}\left(\sum_{t=0}^{T}\|\nabla\Phi(\mathbf{x}_t)\|^2\right) + \tfrac{9\eta_{\mathbf{x}}\kappa\ell^2D^2}{4}.$$

By the definition of Δ_Φ, we have

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\|\nabla\Phi(\mathbf{x}_t)\|^2\right) \leq \tfrac{256(\Phi(\mathbf{x}_0)-\Phi(\mathbf{x}_{T+1}))}{103\eta_{\mathbf{x}}(T+1)} + \tfrac{576\kappa\ell^2D^2}{103(T+1)} \leq \tfrac{128\kappa^2\ell\Delta_\Phi + 5\kappa\ell^2D^2}{T+1}.$$

This implies that the number of iterations required by Algorithm 1 to return an ϵ-stationary point is bounded by

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2D^2}{\epsilon^2}\right),$$

which gives the same total gradient complexity.

Proof of Theorem 2.10.2. We define $\gamma = 1 - 1/2\kappa + 8\kappa^3\ell^2\eta_{\mathbf{x}}^2$ throughout. Performing the second inequality in Lemma 2.10.4 recursively together with $\delta_0 \leq D^2$ yields that

$$\delta_t \leq \gamma^t D^2 + 8\kappa^3\eta_{\mathbf{x}}^2 \left(\sum_{j=0}^{t-1} \gamma^{t-1-j} \mathbb{E}[\|\nabla\Phi(\mathbf{x}_j)\|^2] \right) + \left(\tfrac{4\sigma^2\kappa^3\eta_{\mathbf{x}}^2}{M} + \tfrac{2\sigma^2}{\ell^2 M} \right) \left(\sum_{j=0}^{t-1} \gamma^{t-1-j} \right). \tag{2.20}$$

Combining (2.20) with the second inequality in Lemma 2.10.5 yields that,

$$\begin{aligned} \mathbb{E}\left[\Phi(\mathbf{x}_t)\right]] \leq \mathbb{E}[\Phi(\mathbf{x}_{t-1})] - \tfrac{7\eta_{\mathbf{x}}}{16}\mathbb{E}[\|\nabla\Phi(\mathbf{x}_{t-1})\|^2] + \tfrac{9\eta_{\mathbf{x}}\ell^2\gamma^{t-1}D^2}{16} + \tfrac{\eta_{\mathbf{x}}^2\kappa\ell\sigma^2}{M} \\ + \tfrac{9\eta_{\mathbf{x}}^3\ell^2\kappa^3}{2} \left(\sum_{j=0}^{t-2} \gamma^{t-2-j} \mathbb{E}[\|\nabla\Phi(\mathbf{x}_j)\|^2] \right) + \tfrac{9\eta_{\mathbf{x}}\ell^2}{8} \left(\tfrac{2\sigma^2\kappa^3\eta_{\mathbf{x}}^2}{M} + \tfrac{\sigma^2}{\ell^2 M} \right) \left(\sum_{j=0}^{t-2} \gamma^{t-2-j} \right). \end{aligned} \tag{2.21}$$

Summing up (2.21) over $t = 1, 2, \ldots, T+1$ and rearranging the terms yields that

$$\begin{aligned} \mathbb{E}[\Phi(\mathbf{x}_{T+1})] \leq \Phi(\mathbf{x}_0) - \tfrac{7\eta_{\mathbf{x}}}{16} \sum_{t=0}^{T} \mathbb{E}[\|\nabla\Phi(\mathbf{x}_t)\|^2] + \tfrac{9\eta_{\mathbf{x}}\ell^2 D^2}{16} \left(\sum_{t=0}^{T} \gamma^t \right) \\ + \tfrac{\eta_{\mathbf{x}}^2\kappa\ell\sigma^2(T+1)}{M} + \tfrac{9\eta_{\mathbf{x}}^3\ell^2\kappa^3}{2} \left(\sum_{t=1}^{T+1} \sum_{j=0}^{t-2} \gamma^{t-2-j} \mathbb{E}[\|\nabla\Phi(\mathbf{x}_j)\|^2] \right) \\ + \tfrac{9\eta_{\mathbf{x}}\ell^2}{8} \left(\tfrac{2\sigma^2\kappa^3\eta_{\mathbf{x}}^2}{M} + \tfrac{\sigma^2}{\ell^2 M} \right) \left(\sum_{t=1}^{T+1} \sum_{j=0}^{t-2} \gamma^{t-2-j} \right). \end{aligned}$$

Since $\eta_{\mathbf{x}} = 1/16(\kappa+1)^2\ell$, we have $\gamma \leq 1 - \frac{1}{4\kappa}$ and $\frac{9\eta_{\mathbf{x}}^3\ell^2\kappa^3}{2} \leq \frac{9\eta_{\mathbf{x}}}{1024\kappa}$ and $\frac{2\sigma^2\kappa^3\eta_{\mathbf{x}}^2}{M} \leq \frac{\sigma^2}{\ell^2 M}$. This implies that $\sum_{t=0}^{T} \gamma^t \leq 4\kappa$ and

$$\begin{aligned} \sum_{t=1}^{T+1} \sum_{j=0}^{t-2} \gamma^{t-2-j} \mathbb{E}[\|\nabla\Phi(\mathbf{x}_j)\|^2] &\leq 4\kappa \left(\sum_{t=0}^{T} \mathbb{E}[\|\nabla\Phi(\mathbf{x}_t)\|^2] \right), \\ \left(\sum_{t=1}^{T+1} \sum_{j=0}^{t-2} \gamma^{t-1-j} \right) &\leq 4\kappa(T+1). \end{aligned}$$

Putting these pieces together yields that

$$\mathbb{E}[\Phi(\mathbf{x}_{T+1})] \leq \Phi(\mathbf{x}_0) - \tfrac{103\eta_{\mathbf{x}}}{256} \left(\sum_{t=0}^{T} \mathbb{E}[\|\nabla\Phi(\mathbf{x}_t)\|^2] \right) + \tfrac{9\eta_{\mathbf{x}}\kappa\ell^2 D^2}{4} + \tfrac{\eta_{\mathbf{x}}\sigma^2(T+1)}{16\kappa M} + \tfrac{9\eta_{\mathbf{x}}\kappa\sigma^2(T+1)}{M}.$$

By the definition of Δ_Φ, we have

$$\begin{aligned} \tfrac{1}{T+1} \left(\sum_{t=0}^{T} \mathbb{E}\left[\|\nabla\Phi(\mathbf{x}_t)\|^2\right] \right) &\leq \tfrac{256(\Phi(\mathbf{x}_0) - \mathbb{E}[\Phi(\mathbf{x}_{T+1})])}{103\eta_{\mathbf{x}}(T+1)} + \tfrac{576\kappa\ell^2 D^2}{103(T+1)} + \tfrac{16\sigma^2}{103\kappa M} + \tfrac{2304\kappa\sigma^2}{103M} \\ &\leq \tfrac{2\Delta_\Phi}{\eta_{\mathbf{x}}(T+1)} + \tfrac{5\kappa\ell^2 D^2}{T+1} + \tfrac{24\kappa\sigma^2}{M} \\ &\leq \tfrac{128\kappa^2\ell\Delta_\Phi + 5\kappa\ell^2 D^2}{T+1} + \tfrac{24\sigma^2\kappa}{M}. \end{aligned}$$

This implies that the number of iterations required by Algorithm 2 to return an ϵ-stationary point is bounded by

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2}{\epsilon^2}\right).$$

iterations, which gives the total gradient complexity of the algorithm:

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2}{\epsilon^2}\max\left\{1,\ \frac{\kappa\sigma^2}{\epsilon^2}\right\}\right).$$

This completes the proof.

2.11 Proofs for Nonconvex-Concave Setting

We first specify the choice of parameters in Theorems 2.4.8 and 2.4.9. Then we present the proofs for nonconvex-concave setting with several technical lemmas. Differently from the previous section, we include the case of $\ell D \lesssim \varepsilon$ in the analysis for nonconvex-concave minimax problems.

We present the full version of Theorems 2.4.8 and 2.4.9 with the detailed choice of $\eta_{\mathbf{x}}$, $\eta_{\mathbf{y}}$ and M which are important to subsequent analysis.

Theorem 2.11.1 *Under Assumption 2.4.6 and letting the step sizes $\eta_{\mathbf{x}}$ and $\eta_{\mathbf{y}}$ be chosen as $\eta_{\mathbf{x}} = \min\{\epsilon^2/[16\ell L^2], \epsilon^4/[4096\ell^3 L^2 D^2]\}$ and $\eta_{\mathbf{y}} = 1/\ell$, the iterations complexity of Algorithm 1 to return an ϵ-stationary point is bounded by*

$$O\left(\frac{\ell^3 L^2 D^2 \widehat{\Delta}_\Phi}{\epsilon^6} + \frac{\ell^3 D^2 \widehat{\Delta}_0}{\epsilon^4}\right).$$

which is also the total gradient complexity of the algorithm.

Theorem 2.11.2 *Under Assumptions 2.4.1 and 2.4.6 and letting the step sizes $\eta_{\mathbf{x}}$ and $\eta_{\mathbf{y}}$ be chosen as $\eta_{\mathbf{x}} = \min\{\epsilon^2/[16\ell(L^2+\sigma^2)], \epsilon^4/[8192\ell^3 D^2 L\sqrt{L^2+\sigma^2}],\ \epsilon^6/[65536\ell^3 D^2\sigma^2 L\sqrt{L^2+\sigma^2}]\}$ and $\eta_{\mathbf{y}} = \min\{1/2\ell, \epsilon^2/[16\ell\sigma^2]\}$ with a batch size $M = 1$, the iteration complexity of Algorithm 2 to return an ϵ-stationary point is bounded by*

$$O\left(\left(\frac{\ell^3\left(L^2+\sigma^2\right)D^2\widehat{\Delta}_\Phi}{\epsilon^6} + \frac{\ell^3 D^2\widehat{\Delta}_0}{\epsilon^4}\right)\max\left\{1,\ \frac{\sigma^2}{\epsilon^2}\right\}\right),$$

which is also the total gradient complexity of the algorithm.

We present three key lemmas which are important for the subsequent analysis.

Lemma 2.11.3 *For Algorithm 1, let $\Delta_t = \Phi(\mathbf{x}_t) - f(\mathbf{x}_t, \mathbf{y}_t)$, the following statement holds true,*

$$\Phi_{1/2\ell}(\mathbf{x}_t) \leq \Phi_{1/2\ell}(\mathbf{x}_{t-1}) + 2\eta_{\mathbf{x}}\ell\Delta_{t-1} - \tfrac{\eta_{\mathbf{x}}}{4}\|\nabla\Phi_{1/2\ell}(\mathbf{x}_{t-1})\|^2 + \eta_{\mathbf{x}}^2\ell L^2.$$

For Algorithm 2, let $\Delta_t = \mathbb{E}[\Phi(\mathbf{x}_t) - f(\mathbf{x}_t, \mathbf{y}_t)]$, the following statement holds true,

$$\mathbb{E}[\Phi_{1/2\ell}(\mathbf{x}_t)] \leq \mathbb{E}[\Phi_{1/2\ell}(\mathbf{x}_{t-1})] + 2\eta_{\mathbf{x}}\ell\Delta_{t-1} - \tfrac{\eta_{\mathbf{x}}}{4}\mathbb{E}[\|\nabla\Phi_{1/2\ell}(\mathbf{x}_{t-1})\|^2] + \eta_{\mathbf{x}}^2\ell(L^2 + \sigma^2).$$

Proof. We first consider the deterministic setting. Let $\hat{\mathbf{x}}_{t-1} = \textsc{Prox}_{\Phi/2\ell}(\mathbf{x}_{t-1})$, we have

$$\Phi_{1/2\ell}(\mathbf{x}_t) \leq \Phi(\hat{\mathbf{x}}_{t-1}) + \ell\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_t\|^2. \tag{2.22}$$

Since $f(\cdot, \mathbf{y})$ is L-Lipschitz for any $\mathbf{y} \in \mathcal{Y}$, we have

$$\begin{aligned}\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_t\|^2 &= \|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1} + \eta_{\mathbf{x}}\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 \\ &\leq \|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2 + 2\eta_{\mathbf{x}}\langle\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}, \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\rangle + \eta_{\mathbf{x}}^2 L^2.\end{aligned} \tag{2.23}$$

Plugging (2.23) into (2.22) yields that

$$\Phi_{1/2\ell}(\mathbf{x}_t) \leq \Phi_{1/2\ell}(\mathbf{x}_{t-1}) + 2\eta_{\mathbf{x}}\ell\langle\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}, \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\rangle + \eta_{\mathbf{x}}^2\ell L^2. \tag{2.24}$$

Since f is ℓ-smooth, we have

$$\langle\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}, \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\rangle \leq f(\hat{\mathbf{x}}_{t-1}, \mathbf{y}_{t-1}) - f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) + \tfrac{\ell}{2}\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2. \tag{2.25}$$

Furthermore, $\Phi(\hat{\mathbf{x}}_{t-1}) \geq f(\hat{\mathbf{x}}_{t-1}, \mathbf{y}_{t-1})$. By the definition of Δ_t, we have

$$f(\hat{\mathbf{x}}_{t-1}, \mathbf{y}_{t-1}) - f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \leq \Phi(\hat{\mathbf{x}}_{t-1}) - f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \leq \Delta_{t-1} - \tfrac{\ell}{2}\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2. \tag{2.26}$$

Plugging (2.25) and (2.26) into (2.24) together with $\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\| = \|\nabla\Phi_{1/2\ell}(\mathbf{x}_{t-1})\|/2\ell$ yields the first desired inequality.

We proceed to consider the stochastic setting. Indeed, we have

$$\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_t\|^2 \leq \|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2 + \eta_{\mathbf{x}}^2\left\|\tfrac{1}{M}\sum_{i=1}^{M} G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i)\right\| + 2\eta_{\mathbf{x}}\langle\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}, \tfrac{1}{M}\sum_{i=1}^{M} G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i)\rangle.$$

Taking the expectation of both sides of the above inequality, conditioned on $(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$, together with Lemma 2.8.2 and the Lipschitz property of $f(\cdot, \mathbf{y}_{t-1})$ yields that

$$\begin{aligned}\mathbb{E}[\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_t\|^2 \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1}] &\leq \|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2 + 2\eta_{\mathbf{x}}\langle\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}, \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\rangle + \eta_{\mathbf{x}}^2 L^2 \\ &+\eta_{\mathbf{x}}^2\mathbb{E}\left[\left\|\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) - \tfrac{1}{M}\sum_{i=1}^{M} G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i)\right\|^2 \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1}\right].\end{aligned}$$

Taking the expectation of both sides together with Lemma 2.8.2 yields that

$$\mathbb{E}[\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_t\|^2] \leq \mathbb{E}[\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2] + 2\eta_{\mathbf{x}}\mathbb{E}[\langle\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}, \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\rangle] + \eta_{\mathbf{x}}^2(L^2 + \sigma^2).$$

Combining with (2.25) and (2.26) yields that

$$\begin{aligned}\mathbb{E}[\Phi_{1/2\ell}(\mathbf{x}_t)] &\leq \mathbb{E}[\Phi_{1/2\ell}(\mathbf{x}_{t-1})] + 2\eta_{\mathbf{x}}\mathbb{E}[\langle \hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}, \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\rangle] + \eta_{\mathbf{x}}^2 \ell(L^2 + \sigma^2) \\ &\leq \mathbb{E}[\Phi_{1/2\ell}(\mathbf{x}_{t-1})] + 2\eta_{\mathbf{x}}\ell\Delta_{t-1} - \eta_{\mathbf{x}}\ell^2\mathbb{E}[\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2] + \eta_{\mathbf{x}}^2\ell(L^2+\sigma^2).\end{aligned}$$

Combined with $\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\| = \|\nabla\Phi_{1/2\ell}(\mathbf{x}_{t-1})\|/2\ell$ yields the second desired inequality. □

Lemma 2.11.4 *For Algorithm 1, let $\Delta_t = \Phi(\mathbf{x}_t) - f(\mathbf{x}_t, \mathbf{y}_t)$ and $s \leq t-1$, the following statement holds true,*

$$\Delta_{t-1} \leq \eta_{\mathbf{x}} L^2(2t - 2s - 1) + \tfrac{\ell}{2}(\|\mathbf{y}_{t-1} - \mathbf{y}^\star(\mathbf{x}_s)\|^2 - \|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_s)\|^2) + (f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})).$$

For Algorithm 2, let $\Delta_t = \mathbb{E}\left[\Phi(\mathbf{x}_t) - f(\mathbf{x}_t, \mathbf{y}_t)\right]$ and $s \leq t-1$, the following statement holds true,

$$\begin{aligned}\Delta_{t-1} &\leq \eta_{\mathbf{x}} L\sqrt{L^2+\sigma^2}(2t - 2s - 1) \\ &\quad + \tfrac{1}{2\eta_{\mathbf{y}}}(\mathbb{E}[\|\mathbf{y}_{t-1} - \mathbf{y}^\star(\mathbf{x}_s)\|^2] - \mathbb{E}[\|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_s)\|^2]) + \mathbb{E}[f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})] + \tfrac{\eta_{\mathbf{y}}\sigma^2}{2}.\end{aligned}$$

Proof. We first consider the deterministic setting. For any $\mathbf{y} \in \mathcal{Y}$, the convexity of $\mathcal{Y}$ and the update formula of $\mathbf{y}_t$ imply that

$$(\mathbf{y} - \mathbf{y}_t)^\top(\mathbf{y}_t - \mathbf{y}_{t-1} - \eta_{\mathbf{y}}\nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})) \geq 0.$$

Rearranging the inequality yields that

$$\|\mathbf{y}-\mathbf{y}_t\|^2 \leq 2\eta_{\mathbf{y}}(\mathbf{y}_{t-1}-\mathbf{y})^\top\nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) + 2\eta_{\mathbf{y}}(\mathbf{y}_t-\mathbf{y}_{t-1})^\top\nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) + \|\mathbf{y}-\mathbf{y}_{t-1}\|^2 - \|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2.$$

Since $f(\mathbf{x}_{t-1}, \cdot)$ is concave and ℓ-smooth and $\eta_{\mathbf{y}} = 1/\ell$, we have

$$f(\mathbf{x}_{t-1}, \mathbf{y}) - f(\mathbf{x}_{t-1}, \mathbf{y}_t) \leq \tfrac{\ell}{2}(\|\mathbf{y} - \mathbf{y}_{t-1}\|^2 - \|\mathbf{y} - \mathbf{y}_t\|^2).$$

Plugging $\mathbf{y} = \mathbf{y}^\star(\mathbf{x}_s)$ $(s \leq t-1)$ in the above inequality yields that

$$f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s)) - f(\mathbf{x}_{t-1}, \mathbf{y}_t) \leq \tfrac{\ell}{2}(\|\mathbf{y}_{t-1} - \mathbf{y}^\star(\mathbf{x}_s)\|^2 - \|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_s)\|^2).$$

By the definition of Δ_{t-1}, we have

$$\begin{aligned}\Delta_{t-1} &\leq (f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s))) \\ &\quad + (f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})) + (f(\mathbf{x}_{t-1}, \mathbf{y}_t) - f(\mathbf{x}_t, \mathbf{y}_t)) + \tfrac{\ell}{2}(\|\mathbf{y}_{t-1} - \mathbf{y}^\star(\mathbf{x}_s)\|^2 - \|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_s)\|^2).\end{aligned}$$

Since $f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_s)) \geq f(\mathbf{x}_s, \mathbf{y})$ for $\forall \mathbf{y} \in \mathcal{Y}$, we have

$$\begin{aligned}&f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s)) \\ &\leq f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_{t-1})) + f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s)) \\ &\leq f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_{t-1})) + f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_s)) - f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s)).\end{aligned} \tag{2.27}$$

Since $f(\cdot, \mathbf{y})$ is L-Lipschitz for any $\mathbf{y} \in \mathcal{Y}$, we have

$$\begin{aligned} f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_{t-1})) &\leq L\|\mathbf{x}_{t-1} - \mathbf{x}_s\| \leq \eta_{\mathbf{x}} L^2(t-1-s), \\ f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_s)) - f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s)) &\leq L\|\mathbf{x}_{t-1} - \mathbf{x}_s\| \leq \eta_{\mathbf{x}} L^2(t-1-s) \\ f(\mathbf{x}_{t-1}, \mathbf{y}_t) - f(\mathbf{x}_t, \mathbf{y}_t) &\leq L\|\mathbf{x}_{t-1} - \mathbf{x}_t\| \leq \eta_{\mathbf{x}} L^2. \end{aligned}$$

Putting these pieces together yields the first desired inequality.

We proceed to consider the stochastic setting. For $\forall \mathbf{y} \in \mathcal{Y}$, we use the similar argument and obtain that

$$\begin{aligned} \|\mathbf{y} - \mathbf{y}_t\|^2 \leq{}& 2\eta_{\mathbf{y}}(\mathbf{y}_{t-1} - \mathbf{y})^\top G_{\mathbf{y}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi) + 2\eta_{\mathbf{y}}(\mathbf{y}_t - \mathbf{y}_{t-1})^\top \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \\ &+ 2\eta_{\mathbf{y}}(\mathbf{y}_t - \mathbf{y}_{t-1})^\top (G_{\mathbf{y}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi) - \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})) + \|\mathbf{y} - \mathbf{y}_{t-1}\|^2 - \|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2. \end{aligned}$$

Using the Young's inequality, we have

$$\eta_{\mathbf{y}}(\mathbf{y}_t - \mathbf{y}_{t-1})^\top (G_{\mathbf{y}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi) - \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})) \leq \tfrac{\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2}{4} + \eta_{\mathbf{y}}^2 \|G_{\mathbf{y}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi) - \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2.$$

Taking the expectation of both sides of the above equality, conditioned on $(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$, together with Lemma 2.8.2 yields that

$$\begin{aligned} &\mathbb{E}[\|\mathbf{y} - \mathbf{y}_t\|^2 \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1}] \\ \leq{}& 2\eta_{\mathbf{y}}(\mathbf{y}_{t-1} - \mathbf{y})^\top \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) + 2\eta_{\mathbf{y}} \mathbb{E}[(\mathbf{y}_t - \mathbf{y}_{t-1})^\top \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1}] \\ &+ 2\eta_{\mathbf{y}}^2 \mathbb{E}[\|\nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) - G_{\mathbf{y}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi)\|^2 \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1}] + \|\mathbf{y} - \mathbf{y}_{t-1}\|^2 - \tfrac{\mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2 | \mathbf{x}_{t-1}, \mathbf{y}_{t-1}]}{2}. \end{aligned}$$

Taking the expectation of both sides together with Lemma 2.8.2 yields that

$$\begin{aligned} \mathbb{E}[\|\mathbf{y} - \mathbf{y}_t\|^2] \leq{}& 2\eta_{\mathbf{y}} \mathbb{E}[(\mathbf{y}_{t-1} - \mathbf{y})^\top \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) + (\mathbf{y}_t - \mathbf{y}_{t-1})^\top \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})] \\ &+ \mathbb{E}[\|\mathbf{y} - \mathbf{y}_{t-1}\|^2] - \tfrac{\mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2]}{2} + \eta_{\mathbf{y}}^2 \sigma^2. \end{aligned}$$

Since $f(\mathbf{x}_{t-1}, \cdot)$ is concave and ℓ-smooth, $\mathcal{Y}$ is convex and $\eta_{\mathbf{y}} \leq 1/2\ell$, we have

$$\mathbb{E}[\|\mathbf{y} - \mathbf{y}_t\|^2] \leq \mathbb{E}[\|\mathbf{y} - \mathbf{y}_{t-1}\|^2] + 2\eta_{\mathbf{y}}(f(\mathbf{x}_{t-1}, \mathbf{y}_t) - f(\mathbf{x}_{t-1}, \mathbf{y})) + \eta_{\mathbf{y}}^2 \sigma^2.$$

Plugging $\mathbf{y} = \mathbf{y}^\star(\mathbf{x}_s)$ $(s \leq t-1)$ in the above inequality yields that

$$\mathbb{E}[f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s)) - f(\mathbf{x}_{t-1}, \mathbf{y}_t)] \leq \tfrac{1}{2\eta_{\mathbf{y}}}(\mathbb{E}[\|\mathbf{y}_{t-1} - \mathbf{y}^\star(\mathbf{x}_s)\|^2] - \mathbb{E}[\|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_s)\|^2]) + \tfrac{\eta_{\mathbf{y}} \sigma^2}{2}.$$

By the definition of Δ_{t-1}, we have

$$\begin{aligned} \Delta_{t-1} \leq{}& \mathbb{E}[f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s)) + (f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})) + (f(\mathbf{x}_{t-1}, \mathbf{y}_t) - f(\mathbf{x}_t, \mathbf{y}_t))] \\ &+ \tfrac{\eta_{\mathbf{y}} \sigma^2}{2} + \tfrac{1}{2\eta_{\mathbf{y}}}(\mathbb{E}[\|\mathbf{y}_{t-1} - \mathbf{y}^\star(\mathbf{x}_s)\|^2] - \mathbb{E}[\|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_s)\|^2]). \end{aligned}$$

By the fact that $f(\cdot, \mathbf{y})$ is L-Lipschitz for $\forall \mathbf{y} \in \mathcal{Y}$ and Lemma 2.8.2, we have

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_{t-1}))] &\leq \eta_{\mathbf{x}} L\sqrt{L^2+\sigma^2}(t-1-s), \\
\mathbb{E}[f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_s)) - f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s))] &\leq \eta_{\mathbf{x}} L\sqrt{L^2+\sigma^2}(t-1-s), \\
\mathbb{E}[f(\mathbf{x}_{t-1}, \mathbf{y}_t) - f(\mathbf{x}_t, \mathbf{y}_t)] &\leq \eta_{\mathbf{x}} L\sqrt{L^2+\sigma^2}.
\end{aligned}$$

Putting these pieces together with (2.27) yields the second desired inequality. □

Without loss of generality, we assume that $B \leq T+1$ such that $(T+1)/B$ is an integer. The following lemma provides an upper bound for $\frac{1}{T+1}(\sum_{t=0}^{T} \Delta_t)$ for two-timescale GDA and SGDA using a localization technique.

Lemma 2.11.5 *For Algorithm 1, let $\Delta_t = \Phi(\mathbf{x}_t) - f(\mathbf{x}_t, \mathbf{y}_t)$, the following statement holds true,*

$$\tfrac{1}{T+1}\left(\sum_{t=0}^{T} \Delta_t\right) \leq \eta_{\mathbf{x}} L^2(B+1) + \tfrac{\ell D^2}{2B} + \tfrac{\widehat{\Delta}_0}{T+1}.$$

For Algorithm 2, let $\Delta_t = \mathbb{E}[\Phi(\mathbf{x}_t) - f(\mathbf{x}_t, \mathbf{y}_t)]$, the following statement holds true,

$$\tfrac{1}{T+1}\left(\sum_{t=0}^{T} \Delta_t\right) \leq \eta_{\mathbf{x}} L\sqrt{L^2+\sigma^2}(B+1) + \tfrac{D^2}{2B\eta_{\mathbf{y}}} + \tfrac{\eta_{\mathbf{y}}\sigma^2}{2} + \tfrac{\widehat{\Delta}_0}{T+1}.$$

Proof. We first consider the deterministic setting. In particular, we divide $\{\Delta_t\}_{t=0}^{T}$ into several blocks in which each block contains at most B terms, given by

$$\{\Delta_t\}_{t=0}^{B-1},\ \{\Delta_t\}_{t=B}^{2B-1}, \ldots, \{\Delta_t\}_{T-B+1}^{T}.$$

Then we have

$$\tfrac{1}{T+1}\left(\sum_{t=0}^{T} \Delta_t\right) \leq \tfrac{B}{T+1}\left[\sum_{j=0}^{(T+1)/B-1}\left(\tfrac{1}{B}\sum_{t=jB}^{(j+1)B-1} \Delta_t\right)\right]. \tag{2.28}$$

Furthermore, letting $s=0$ in the first inequality in Lemma (2.11.4) yields that

$$\begin{aligned}
\sum_{t=0}^{B-1} \Delta_t &\leq \eta_{\mathbf{x}} L^2 B^2 + \tfrac{\ell}{2}\|\mathbf{y}_0 - \mathbf{y}^\star(\mathbf{x}_0)\|^2 + (f(\mathbf{x}_B, \mathbf{y}_B) - f(\mathbf{x}_0, \mathbf{y}_0)) \qquad (2.29)\\
&\leq \eta_{\mathbf{x}} L^2 B^2 + \tfrac{\ell D^2}{2} + (f(\mathbf{x}_B, \mathbf{y}_B) - f(\mathbf{x}_0, \mathbf{y}_0)).
\end{aligned}$$

Similarly, letting $s = jB$ yields that, for $1 \leq j \leq \frac{T+1}{B} - 1$,

$$\sum_{t=jB}^{(j+1)B-1} \Delta_t \leq \eta_{\mathbf{x}} L^2 B^2 + \tfrac{\ell D^2}{2} + (f(\mathbf{x}_{jB+B}, \mathbf{y}_{jB+B}) - f(\mathbf{x}_{jB}, \mathbf{y}_{jB})). \tag{2.30}$$

Plugging (2.29) and (2.30) into (2.28) yields

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\Delta_t\right) \le \eta_{\mathbf{x}} L^2 B + \frac{\ell D^2}{2B} + \frac{f(\mathbf{x}_{T+1},\mathbf{y}_{T+1}) - f(\mathbf{x}_0,\mathbf{y}_0)}{T+1}. \tag{2.31}$$

Since $f(\cdot,\mathbf{y})$ is L-Lipschitz for any $\mathbf{y}\in\mathcal{Y}$, we have

$$\begin{aligned} f(\mathbf{x}_{T+1},\mathbf{y}_{T+1}) - f(\mathbf{x}_0,\mathbf{y}_0) &= f(\mathbf{x}_{T+1},\mathbf{y}_{T+1}) - f(\mathbf{x}_0,\mathbf{y}_{T+1}) + f(\mathbf{x}_0,\mathbf{y}_{T+1}) - f(\mathbf{x}_0,\mathbf{y}_0) \\ &\le \eta_{\mathbf{x}} L^2 (T+1) + \widehat{\Delta}_0. \end{aligned} \tag{2.32}$$

Plugging (2.32) into (2.31) yields the desired inequality. As for the stochastic case, letting $s = jB$ in the second inequality in Lemma 2.11.4 yields that

$$\sum_{t=jB}^{(j+1)B-1} \Delta_t \le \eta_{\mathbf{x}} L\sqrt{L^2+\sigma^2}B^2 + \frac{D^2}{2\eta_{\mathbf{y}}} + \frac{\eta_{\mathbf{y}}\sigma^2}{2}, \quad 0 \le j \le \frac{T+1}{B} - 1. \tag{2.33}$$

Using the similar argument with (2.33) and (2.28) yields the second desired inequality. □

Proof of Theorem 2.11.1. Summing up the first inequality in Lemma 2.11.3 over $t = 1, 2, \ldots, T+1$ yields that

$$\Phi_{1/2\ell}(\mathbf{x}_{T+1}) \le \Phi_{1/2\ell}(\mathbf{x}_0) + 2\eta_{\mathbf{x}}\ell\left(\sum_{t=0}^{T}\Delta_t\right) - \frac{\eta_{\mathbf{x}}}{4}\left(\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\|^2\right) + \eta_{\mathbf{x}}^2\ell L^2(T+1).$$

Combining the above inequality with the first inequality in Lemma 2.11.5 yields that

$$\begin{aligned} \Phi_{1/2\ell}(\mathbf{x}_{T+1}) &\le \Phi_{1/2\ell}(\mathbf{x}_0) + 2\eta_{\mathbf{x}}\ell(T+1)\left(\eta_{\mathbf{x}} L^2(B+1) + \frac{\ell D^2}{2B}\right) \\ &\quad + 2\eta_{\mathbf{x}}\ell\widehat{\Delta}_0 - \frac{\eta_{\mathbf{x}}}{4}\left(\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\|^2\right) + \eta_{\mathbf{x}}^2\ell L^2(T+1). \end{aligned}$$

By the definition of $\widehat{\Delta}_\Phi$, we have

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\|^2\right) \le \frac{4\widehat{\Delta}_\Phi}{\eta_{\mathbf{x}}(T+1)} + 8\ell\left(\eta_{\mathbf{x}}(B+1)L^2 + \frac{\ell D^2}{2B}\right) + \frac{8\ell\widehat{\Delta}_0}{T+1} + 4\eta_{\mathbf{x}}\ell L^2.$$

Letting $B = 1$ for $D = 0$ and $B = \frac{D}{2L}\sqrt{\frac{\ell}{\eta_{\mathbf{x}}}}$ for $D > 0$, we have

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\|^2\right) \le \frac{4\widehat{\Delta}_\Phi}{\eta_{\mathbf{x}}(T+1)} + \frac{8\ell\widehat{\Delta}_0}{T+1} + 16\ell L D\sqrt{\ell\eta_{\mathbf{x}}} + 4\eta_{\mathbf{x}}\ell L^2.$$

Since $\eta_{\mathbf{x}} = \min\{\frac{\epsilon^2}{16\ell L^2}, \frac{\epsilon^4}{4096\ell^3 L^2 D^2}\}$, we have

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\|^2\right) \leq \frac{4\widehat{\Delta}_\Phi}{\eta_{\mathbf{x}}(T+1)} + \frac{8\ell\widehat{\Delta}_0}{T+1} + \frac{\epsilon^2}{2}.$$

This implies that the number of iterations required by Algorithm 1 to return an ϵ-stationary point is bounded by

$$O\left(\left(\frac{\ell L^2\widehat{\Delta}_\Phi}{\epsilon^4} + \frac{\ell\widehat{\Delta}_0}{\epsilon^2}\right)\max\left\{1, \frac{\ell^2 D^2}{\epsilon^2}\right\}\right),$$

which gives the same total gradient complexity.

Proof of Theorem 2.11.2. Summing up the second inequality in Lemma 2.11.3 over $t = 1, 2, \ldots, T+1$ yields that

$$\mathbb{E}[\Phi_{1/2\ell}(\mathbf{x}_{T+1})] \leq \Phi_{1/2\ell}(\mathbf{x}_0) + 2\eta_{\mathbf{x}}\ell\sum_{t=0}^{T}\Delta_t - \frac{\eta_{\mathbf{x}}}{4}\sum_{t=0}^{T}\mathbb{E}[\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\|^2] + \eta_{\mathbf{x}}^2\ell(L^2+\sigma^2)(T+1).$$

Combining the above inequality with the second inequality in Lemma 2.11.5 yields that

$$\begin{aligned}\mathbb{E}[\Phi_{1/2\ell}(\mathbf{x}_{T+1})] \;\leq\; & \Phi_{1/2\ell}(\mathbf{x}_0) + 2\eta_{\mathbf{x}}\ell(T+1)\left(\eta_{\mathbf{x}}L\sqrt{L^2+\sigma^2}(B+1) + \frac{D^2}{2B\eta_{\mathbf{y}}} + \frac{\eta_{\mathbf{y}}\sigma^2}{2}\right)\\ & + 2\eta_{\mathbf{x}}\ell\widehat{\Delta}_0 - \frac{\eta_{\mathbf{x}}}{4}\sum_{t=0}^{T}\mathbb{E}[\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\|^2] + \eta_{\mathbf{x}}^2\ell(L^2+\sigma^2)(T+1).\end{aligned}$$

By the definition of $\widehat{\Delta}_\Phi$, we have

$$\begin{aligned}\frac{1}{T+1}\left(\sum_{t=0}^{T}\mathbb{E}[\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\|^2]\right) \;\leq\; & \frac{4\widehat{\Delta}_\Phi}{\eta_{\mathbf{x}}(T+1)} + 8\ell\left(\eta_{\mathbf{x}}L\sqrt{L^2+\sigma^2}(B+1) + \frac{D^2}{2B\eta_{\mathbf{y}}} + \frac{\eta_{\mathbf{y}}\sigma^2}{2}\right)\\ & + \frac{8\ell\widehat{\Delta}_0}{T+1} + 4\eta_{\mathbf{x}}\ell(L^2+\sigma^2).\end{aligned}$$

Letting $B = 1$ for $D = 0$ and $B = \frac{D}{2}\sqrt{\frac{1}{\eta_{\mathbf{x}}\eta_{\mathbf{y}}L\sqrt{L^2+\sigma^2}}}$ for $D > 0$, we have

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\|^2\right) \leq \frac{4\widehat{\Delta}_\Phi}{\eta_{\mathbf{x}}(T+1)} + \frac{8\ell\widehat{\Delta}_0}{T+1} + 16\ell D\sqrt{\frac{\eta_{\mathbf{x}}L\sqrt{L^2+\sigma^2}}{\eta_{\mathbf{y}}}} + 4\eta_{\mathbf{y}}\ell\sigma^2 + 4\eta_{\mathbf{x}}\ell(L^2+\sigma^2).$$

Since $\eta_{\mathbf{x}} = \min\{\frac{\epsilon^2}{16\ell(L^2+\sigma^2)}, \frac{\epsilon^4}{8192\ell^3 D^2 L\sqrt{L^2+\sigma^2}}, \frac{\epsilon^6}{65536\ell^3 D^2\sigma^2 L\sqrt{L^2+\sigma^2}}\}$ and $\eta_{\mathbf{y}} = \min\{\frac{1}{2\ell}, \frac{\epsilon^2}{16\ell\sigma^2}\}$, we have

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\|^2\right) \leq \frac{4\widehat{\Delta}_\Phi}{\eta_{\mathbf{x}}(T+1)} + \frac{8\ell\widehat{\Delta}_0}{T+1} + \frac{3\epsilon^2}{4}.$$

This implies that the number of iterations required by Algorithm 2 to return an ϵ-stationary point is bounded by

$$O\left(\left(\frac{\ell(L^2+\sigma^2)\widehat{\Delta}_\Phi}{\epsilon^4}+\frac{\ell\widehat{\Delta}_0}{\epsilon^2}\right)\max\left\{1,\ \frac{\ell^2D^2}{\epsilon^2},\ \frac{\ell^2D^2\sigma^2}{\epsilon^4}\right\}\right),$$

which gives the same total gradient complexity.

2.12 Results for GDmax and SGDmax

We present GDmax and SGDmax in Algorithm 3 and 4. For any $\mathbf{x}_t \in \mathbb{R}^m$, the max-oracle approximately solves $\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}_t, \mathbf{y})$ at each iteration. Although GDmax and SGDmax are easier to understand, they have two disadvantages over two-timescale GDA and SGDA: 1) Both GDmax and SGDmax are nested-loop algorithms. Since it is difficult to pre-determine the number iterations for the inner loop, these algorithms are not favorable in practice; 2) In the general setting where $f(\mathbf{x}, \cdot)$ is nonconcave, GDmax and SGDmax are inapplicable as we can not efficiently solve the maximization problem to a global optimum. Nevertheless, we present the complexity bound for GDmax and SGDmax for the sake of completeness. It is worth noting that a portion of results were derived before Jin et al. [2020] and Nouiehed et al. [2019] and our proof depends on the same techniques.

For nonconvex-strongly-convex problems, the target is to find an ϵ-stationary point (cf. Definition 2.3.3) given gradient (or stochastic gradient) access to f. Denote $\Delta_\Phi = \Phi(\mathbf{x}_0) - \min_{\mathbf{x}\in\mathbb{R}^m}\Phi(\mathbf{x})$, we present the gradient complexity for GDmax in the following theorem.

Theorem 2.12.1 *Under Assumption 2.4.2 and letting the step size $\eta_\mathbf{x} > 0$ and the tolerance for the max-oracle $\zeta > 0$ be $\eta_\mathbf{x} = 1/[8\kappa\ell]$ and $\zeta = \epsilon^2/[6\ell]$, the number of iterations required by Algorithm 3 to return an ϵ-stationary point is bounded by $O(\kappa\ell\Delta_\Phi\epsilon^{-2})$. Furthermore, the ζ-accurate max-oracle can be realized by gradient ascent (GA) with the stepsize $\eta_\mathbf{y} = 1/\ell$ for $O(\kappa\log(\ell D^2/\zeta))$ iterations, which gives the total gradient complexity of the algorithm:*

$$O\left(\frac{\kappa^2\ell\Delta_\Phi}{\epsilon^2}\log\left(\frac{\ell D}{\epsilon}\right)\right).$$

Theorem 2.12.1 demonstrates that, if we alternate between one-step gradient descent over $\mathbf{x}$ and $O(\kappa\log(\ell D/\epsilon))$ gradient ascent steps over $\mathbf{y}$ with a pair of proper learning rates $(\eta_\mathbf{x}, \eta_\mathbf{y})$, we find at least one stationary point of Φ within $O(\kappa^2\epsilon^{-2}\log(\ell/\epsilon))$ gradient evaluations. Then we present similar guarantees for stochastic setting in the following theorem.

Theorem 2.12.2 *Under Assumption 2.4.1 and 2.4.2 and letting the step size $\eta_\mathbf{x} > 0$ and the tolerance for the max-oracle $\zeta > 0$ be the same in Theorem 2.12.1 with the batch size $M = \max\{1, 12\kappa\sigma^2\epsilon^{-2}\}$, the number of iterations required by Algorithm 4 to return an ϵ-stationary point is bounded by $O(\kappa\ell\Delta_\Phi\epsilon^{-2})$. Furthermore, the ζ-accurate max-oracle can be realized by mini-batch stochastic gradient ascent (SGA) with the step size $\eta_\mathbf{y} = 1/\ell$ and the*

mini-batch size $M = \max\{1, 2\sigma^2\kappa\ell^{-1}\zeta^{-1}\}$ *for* $O(\kappa \log(\ell D^2/\zeta) \max\{1, 2\sigma^2\kappa\ell^{-1}\zeta^{-1}\})$ *gradient evaluations, which gives the total gradient complexity of the algorithm:*

$$O\left(\frac{\kappa^2 \ell \Delta_\Phi}{\epsilon^2} \log\left(\frac{\ell D}{\epsilon}\right) \max\left\{1, \frac{\kappa\sigma^2}{\epsilon^2}\right\}\right).$$

The sample size $M = O(\kappa\sigma^2\epsilon^{-2})$ shows that the variance is less than ϵ^2/κ so that the average stochastic gradients over the batch are sufficiently close to the true gradients $\nabla_{\mathbf{x}} f$ and $\nabla_{\mathbf{y}} f$.

We now proceed to the theoretical guarantee for GDmax and SGDmax algorithms for nonconvex-concave problems. The target is to find an ϵ-stationary point of a weakly convex function (Definition 2.3.7) given only gradient (or stochastic gradient) access to f. Denote $\widehat{\Delta}_\Phi = \Phi_{1/2\ell}(\mathbf{x}_0) - \min_{\mathbf{x}\in\mathbb{R}^m} \Phi_{1/2\ell}(\mathbf{x})$, we present the gradient complexity for GDmax and SGDmax in the following two theorems.

Theorem 2.12.3 *Under Assumption 2.4.6 and letting the step size* $\eta_{\mathbf{x}} > 0$ *and the tolerance for the max-oracle* $\zeta > 0$ *be* $\eta_{\mathbf{x}} = \epsilon^2/[\ell L^2]$ *and* $\zeta = \epsilon^2/[24\ell]$*, the number of iterations required by Algorithm 3 to return an* ϵ*-stationary point is bounded by* $O(\ell L^2 \widehat{\Delta}_\Phi \epsilon^{-4})$*. Furthermore, the* ζ*-accurate max-oracle is realized by GA with the step size* $\eta_{\mathbf{y}} = 1/2\ell$ *for* $O(\ell D^2/\zeta)$ *iterations, which gives the total gradient complexity of the algorithm:*

$$O\left(\frac{\ell^3 L^2 D^2 \widehat{\Delta}_\Phi}{\epsilon^6}\right).$$

Theorem 2.12.4 *Under Assumptions 2.4.1 and 2.4.6 and letting the tolerance for the max-oracle* $\zeta > 0$ *be chosen as the same as in Theorem 2.12.3 with a step size* $\eta_{\mathbf{x}} > 0$ *and a batch size* $M > 0$ *given by* $\eta_{\mathbf{x}} = \epsilon^2/[\ell(L^2 + \sigma^2)]$ *and* $M = 1$*, the number of iterations required by Algorithm 4 to return an* ϵ*-stationary point is bounded by* $O(\ell(L^2+\sigma^2)\widehat{\Delta}_\Phi \epsilon^{-4})$*. Furthermore, the* ζ*-accurate max-oracle is realized by SGA with the step size* $\eta_{\mathbf{y}} = \min\{1/2\ell, \epsilon^2/[\ell\sigma^2]\}$ *and a batch size* $M = 1$ *for* $O(\ell D^2 \zeta^{-1} \max\{1, \sigma^2\ell^{-1}\zeta^{-1}\})$ *iterations, which gives the following total gradient complexity of the algorithm:*

$$O\left(\frac{\ell^3(L^2 + \sigma^2) D^2 \widehat{\Delta}_\Phi}{\epsilon^6} \max\left\{1, \frac{\sigma^2}{\epsilon^2}\right\}\right).$$

When $\sigma^2 \lesssim \varepsilon^2$, the stochastic gradients are sufficiently close to the true gradients $\nabla_{\mathbf{x}} f$ and $\nabla_{\mathbf{y}} f$ and the gradient complexity of SGDmax matches that of GDmax.

Proof of Theorem 2.12.1. We present the gradient complexity bound of the gradient-ascent-based ζ-accurate max-oracle in the following lemma.

Lemma 2.12.5 *Let* $\zeta > 0$ *be given, the* ζ*-accurate max-oracle can be realized by running gradient ascent with a step size* $\eta_{\mathbf{y}} = 1/\ell$ *for* $O(\kappa \log(\ell D^2/\zeta))$ *gradient evaluations. In addition, the output* $\mathbf{y}$ *satisfies* $\|\mathbf{y}^\star - \mathbf{y}\|^2 \le \zeta/\ell$*, where* $\mathbf{y}^\star$ *is the exact maximizer.*

Algorithm 3 Gradient Descent with Max-oracle (GDmax)

Input: initial point $\mathbf{x}_0$, learning rate $\eta_{\mathbf{x}}$ and max-oracle accuracy ζ.
for $t = 1, 2, \ldots$ **do**
find $\mathbf{y}_{t-1} \in \mathcal{Y}$ so that $f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \geq \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}_{t-1}, \mathbf{y}) - \zeta$.
$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta_{\mathbf{x}} \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$.

Algorithm 4 Stochastic Gradient Descent with Max-oracle (SGDmax)

Input: initial point $\mathbf{x}_0$, learning rate $\eta_{\mathbf{x}}$ and max-oracle accuracy ζ.
for $t = 1, 2, \ldots$ **do**
Draw a collection of i.i.d. data samples $\{\xi_i\}_{i=1}^M$.
find $\mathbf{y}_{t-1} \in \mathcal{Y}$ so that $\mathbb{E}[f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \mid \mathbf{x}_{t-1}] \geq \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}_{t-1}, \mathbf{y}) - \zeta$.
$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta_{\mathbf{x}} \left(\frac{1}{M} \sum_{i=1}^M G_{\mathbf{x}} \left(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i \right) \right)$.

Proof. Since $f(\mathbf{x}_t, \cdot)$ is μ-strongly concave, we have

$$\begin{aligned} f(\mathbf{x}_t, \mathbf{y}^\star(\mathbf{x}_t)) - f(\mathbf{x}_t, \mathbf{y}_t) &\leq \left(1 - \tfrac{1}{\kappa}\right)^{N_t} \tfrac{\ell D^2}{2}, \\ \|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}_t\|^2 &\leq \left(1 - \tfrac{1}{\kappa}\right)^{N_t} D^2. \end{aligned}$$

The first inequality implies that the number of iterations required is $O(\kappa \log(\ell D^2/\zeta))$ which is also the number of gradient evaluations. This, together with the second inequality, yields the other results. □

It is easy to find that the first descent inequality in Lemma 2.10.3 is applicable to GDmax:

$$\Phi(\mathbf{x}_t) \leq \Phi(\mathbf{x}_{t-1}) - \left(\tfrac{\eta_{\mathbf{x}}}{2} - 2\eta_{\mathbf{x}}^2 \kappa \ell\right) \|\nabla \Phi(\mathbf{x}_{t-1})\|^2 + \left(\tfrac{\eta_{\mathbf{x}}}{2} + 2\eta_{\mathbf{x}}^2 \kappa \ell\right) \|\nabla \Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2.$$

Since $\nabla \Phi(\mathbf{x}_{t-1}) = \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1}))$, we have

$$\|\nabla \Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 \leq \ell^2 \|\mathbf{y}^\star(\mathbf{x}_{t-1}) - \mathbf{y}_{t-1}\|^2 \leq \ell \zeta.$$

Since $\eta_{\mathbf{x}} = 1/8\kappa\ell$, we have

$$\tfrac{\eta_{\mathbf{x}}}{4} \leq \tfrac{\eta_{\mathbf{x}}}{2} - 2\eta_{\mathbf{x}}^2 \kappa \ell \leq \tfrac{\eta_{\mathbf{x}}}{2} + 2\eta_{\mathbf{x}}^2 \kappa \ell \leq \tfrac{3\eta_{\mathbf{x}}}{4}.$$

Putting these pieces together yields that

$$\Phi(\mathbf{x}_t) \leq \Phi(\mathbf{x}_{t-1}) - \tfrac{\eta_{\mathbf{x}}}{4} \|\nabla \Phi(\mathbf{x}_{t-1})\|^2 + \tfrac{3\eta_{\mathbf{x}} \ell \zeta}{4}. \tag{2.34}$$

Summing up (2.34) over $t = 1, 2, \ldots, T+1$ and rearranging the terms yields that

$$\frac{1}{T+1} \sum_{t=0}^{T} \|\nabla \Phi(\mathbf{x}_t)\|^2 \leq \tfrac{4(\Phi(\mathbf{x}_0) - \Phi(\mathbf{x}_{T+1}))}{\eta_{\mathbf{x}}(T+1)} + 3\ell\zeta.$$

By the definition of $\eta_{\mathbf{x}}$ and Δ_Φ, we conclude that

$$\tfrac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi(\mathbf{x}_t)\|^2 \le \tfrac{32\kappa\ell\Delta_\Phi}{T+1} + 3\ell\zeta.$$

This implies that the number of iterations required by Algorithm 3 to return an ϵ-stationary point is bounded by

$$O\left(\frac{\kappa\ell\Delta_\Phi}{\epsilon^2}\right).$$

Combining Lemma 2.12.5 gives the total gradient complexity of Algorithm 3:

$$O\left(\frac{\kappa^2\ell\Delta_\Phi}{\epsilon^2}\log\left(\frac{\ell D}{\epsilon}\right)\right).$$

This completes the proof.

Proof of Theorem 2.12.2. We present the gradient complexity bound of the stochastic-gradient-ascent-based ζ-accurate max-oracle in the following lemma.

Lemma 2.12.6 *Let $\zeta > 0$ be given, the ζ-accurate max-oracle can be realized by running stochastic gradient ascent with a step size $\eta_{\mathbf{y}} = 1/\ell$ and a batch size $M = \max\{1, 2\sigma^2\kappa/\ell\zeta\}$ for*

$$O\left(\kappa\log\left(\frac{\ell D^2}{\zeta}\right)\max\left\{1,\ \frac{2\sigma^2\kappa}{\ell\zeta}\right\}\right)$$

stochastic gradient evaluations. In addition, the output $\mathbf{y}$ satisfies $\|\mathbf{y}^\star - \mathbf{y}\|^2 \le \zeta/\ell$ where $\mathbf{y}^\star$ is the exact maximizer.

Proof. Since $f(\mathbf{x}_t, \cdot)$ is μ-strongly concave, we have

$$\mathbb{E}[f(\mathbf{x}_t, \mathbf{y}^\star(\mathbf{x}_t)) - f(\mathbf{x}_t, \mathbf{y}_t)] \le \left(1-\tfrac{1}{\kappa}\right)^{N_t}\tfrac{\ell D^2}{2} + \tfrac{\eta_{\mathbf{y}}^2\ell\sigma^2}{M}\left(\sum_{j=0}^{N_{t-1}}\left(1-\mu\eta_{\mathbf{y}}\right)^{N_{t-1}-1-j}\right) \le \left(1-\tfrac{1}{\kappa}\right)^{N_t}\tfrac{\ell D^2}{2} + \tfrac{\sigma^2\kappa}{\ell M},$$

and

$$\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_t)) - \mathbf{y}_t\|^2] \le \left(1-\tfrac{1}{\kappa}\right)^{N_t} D^2 + \tfrac{\eta_{\mathbf{y}}^2\sigma^2}{M}\left(\sum_{j=0}^{N_{t-1}}\left(1-\mu\eta_{\mathbf{y}}\right)^{N_{t-1}-1-j}\right) \le \left(1-\tfrac{1}{\kappa}\right)^{N_t}\tfrac{\ell D^2}{2} + \tfrac{\sigma^2\kappa}{\ell^2 M}.$$

The first inequality implies that the number of iterations is $O(\kappa\log(\ell D^2/\zeta))$ and the number of stochastic gradient evaluation is $O(\kappa\log(\ell D^2/\zeta)\max\{1, 2\sigma^2\kappa/\ell\zeta\})$. This together with the second inequality yields the other results. □

It is easy to find that the second descent inequality in Lemma 2.10.3 is applicable to SGDmax:

$$\begin{aligned}\mathbb{E}[\Phi(\mathbf{x}_t)] \leq \mathbb{E}\left[\Phi(\mathbf{x}_{t-1})\right] - \left(\tfrac{\eta_{\mathbf{x}}}{2} - 2\eta_{\mathbf{x}}^2\kappa\ell\right)\mathbb{E}[\|\nabla\Phi(\mathbf{x}_{t-1})\|^2] \\ + \left(\tfrac{\eta_{\mathbf{x}}}{2} + 2\eta_{\mathbf{x}}^2\kappa\ell\right)\mathbb{E}[\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2] + \tfrac{\eta_{\mathbf{x}}^2\kappa\ell\sigma^2}{M}.\end{aligned}$$

Since $\nabla\Phi(\mathbf{x}_{t-1}) = \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1}))$, we have

$$\mathbb{E}[\|\nabla\Phi(\mathbf{x}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\|^2] \leq \ell^2\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}_t\|^2] \leq \ell\zeta.$$

Putting these pieces together with $\eta_{\mathbf{x}} = 1/8\kappa\ell$ yields that

$$\mathbb{E}[\Phi(\mathbf{x}_t)] \leq \mathbb{E}[\Phi(\mathbf{x}_{t-1})] - \tfrac{\eta_{\mathbf{x}}}{4}\mathbb{E}[\|\nabla\Phi(\mathbf{x}_{t-1})\|^2] + \tfrac{3\eta_{\mathbf{x}}\ell\zeta}{4} + \tfrac{\eta_{\mathbf{x}}^2\kappa\ell\sigma^2}{M}. \tag{2.35}$$

Summing up (2.35) over $t = 1, 2, \ldots, T+1$ and rearranging the terms yields that

$$\tfrac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla\Phi(\mathbf{x}_t)\|^2] \leq \tfrac{4(\Phi(\mathbf{x}_0) - \mathbb{E}[\Phi(\mathbf{x}_{T+1})])}{\eta_{\mathbf{x}}(T+1)} + 3\ell\zeta + \tfrac{4\eta_{\mathbf{x}}\kappa\ell\sigma^2}{M}.$$

By the definition of $\eta_{\mathbf{x}}$ and Δ_Φ, we conclude that

$$\tfrac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla\Phi(\mathbf{x}_t)\|^2] \leq \tfrac{32\kappa\ell\Delta_\Phi}{T+1} + 3\ell\zeta + \tfrac{\sigma^2}{2M}.$$

This implies that the number of iterations required by Algorithm 4 to return an ϵ-stationary point is bounded by

$$O\left(\frac{\kappa\ell\Delta_\Phi}{\epsilon^2}\right).$$

Note that the same batch set can be reused to construct the unbiased stochastic gradients for both $\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$ and $\nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$ at each iteration. Combining Lemma 2.12.6 gives the total gradient complexity of Algorithm 4:

$$O\left(\frac{\kappa^2\ell\Delta_\Phi}{\epsilon^2}\log\left(\frac{\sqrt{\kappa}\ell D}{\epsilon}\right)\max\left\{1, \frac{\sigma^2\kappa^2}{\epsilon^2}\right\}\right).$$

This completes the proof.

Proof of Theorem 2.12.3. We present the gradient complexity bound of the gradient-ascent-based ζ-accurate max-oracle in the following lemma.

Lemma 2.12.7 *Let $\zeta > 0$ be given, the ζ-accurate max-oracle can be realized by running gradient ascent with a step size $\eta_{\mathbf{y}} = 1/2\ell$ for*

$$O\left(\max\left\{1, \frac{2\ell D^2}{\zeta}\right\}\right)$$

gradient evaluations.

Proof. Since $f(\mathbf{x}_t, \cdot)$ is concave, we have

$$f(\mathbf{x}_t, \mathbf{y}^\star(\mathbf{x}_t)) - f(\mathbf{x}_t, \mathbf{y}_t) \leq \tfrac{2\ell D^2}{N_t},$$

which implies that the number of iterations required is $\mathcal{O}(\max\{1,\ 2\ell D^2/\zeta\})$ which is the number of gradient evaluation. □

It is easy to find that the first descent inequality in Lemma 2.11.3 is applicable to GDmax:

$$\Phi_{1/2\ell}(\mathbf{x}_t) \leq \Phi_{1/2\ell}(\mathbf{x}_{t-1}) + 2\eta_{\mathbf{x}}\ell\Delta_{t-1} - \tfrac{\eta_{\mathbf{x}}}{4}\|\nabla\Phi_{1/2\ell}(\mathbf{x}_{t-1})\|^2 + \eta_{\mathbf{x}}^2\ell L^2. \tag{2.36}$$

Summing up (2.36) over $T = 1, 2, \ldots, T+1$ together with $\Delta_{t-1} \leq \zeta$ and rearranging the terms yields that

$$\tfrac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\|^2 \leq \tfrac{4(\Phi_{1/2\ell}(\mathbf{x}_0) - \Phi_{1/2\ell}(\mathbf{x}_{T+1}))}{\eta_{\mathbf{x}}(T+1)} + 8\ell\zeta + 4\eta_{\mathbf{x}}\ell L^2.$$

By the definition of $\eta_{\mathbf{x}}$ and $\widehat{\Delta}_\Phi$, we have

$$\tfrac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\|^2 \leq \tfrac{48\ell L^2\widehat{\Delta}_\Phi}{\epsilon^2(T+1)} + 8\ell\zeta + \tfrac{\epsilon^2}{3}.$$

This implies that the number of iterations required by Algorithm 3 to return an ϵ-stationary point is bounded by

$$O\left(\frac{\ell L^2\widehat{\Delta}_\Phi}{\epsilon^4}\right).$$

Combining Lemma 2.12.7 gives the total gradient complexity of Algorithm 3:

$$O\left(\frac{\ell L^2\widehat{\Delta}_\Phi}{\epsilon^4}\max\left\{1,\ \frac{\ell^2 D^2}{\epsilon^2}\right\}\right).$$

This completes the proof.

Proof of Theorem 2.12.4. We present the gradient complexity bound of the stochastic-ascent-based ζ-accurate max-oracle in the following lemma.

Lemma 2.12.8 *Let $\zeta > 0$ be given, the ζ-accurate max-oracle can be realized by running stochastic gradient ascent with a step size $\eta_{\mathbf{y}} = \min\{1/2\ell, \zeta/2\sigma^2\}$ and a batch size $M = 1$ for*

$$O\left(\max\left\{1,\ \frac{4\ell D^2}{\zeta},\ \frac{4\sigma^2 D^2}{\zeta^2}\right\}\right) \tag{2.37}$$

stochastic gradient evaluations.

Proof. Since $f(\mathbf{x}_t, \cdot)$ is concave and $\eta_{\mathbf{y}} = \min\{\frac{1}{2\ell}, \frac{\zeta}{2\sigma^2}\}$, we have

$$\mathbb{E}[f(\mathbf{x}_t, \mathbf{y}^\star(\mathbf{x}_t))] - \mathbb{E}[f(\mathbf{x}_t, \mathbf{y}_t)] \leq \tfrac{D^2}{\eta_{\mathbf{y}} N_t} + \eta_{\mathbf{y}} \sigma^2.$$

which implies that the number of iterations required is $O(\max\{1, 4\ell D^2 \zeta^{-1}, 4\sigma^2 D^2 \zeta^{-2}\})$ which is also the number of stochastic gradient evaluations since $M = 1$. □

It is easy to find that the second descent inequality in Lemma 2.11.3 is applicable to SGDmax:

$$\mathbb{E}[\Phi_{1/2\ell}(\mathbf{x}_t)] \leq \mathbb{E}[\Phi_{1/2\ell}(\mathbf{x}_{t-1})] + 2\eta_{\mathbf{x}} \ell \Delta_{t-1} - \tfrac{\eta_{\mathbf{x}}}{4} \mathbb{E}[\|\nabla \Phi_{1/2\ell}(\mathbf{x}_{t-1})\|^2] + \eta_{\mathbf{x}}^2 \ell (L^2 + \sigma^2). \quad (2.38)$$

Summing up (2.38) over $T = 1, 2, \ldots, T+1$ together with $\Delta_{t-1} \leq \zeta$ and rearranging the terms yields that

$$\tfrac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}[\|\nabla \Phi_{1/2\ell}(\mathbf{x}_t)\|^2] \leq \tfrac{4(\Phi_{1/2\ell}(\mathbf{x}_0) - \mathbb{E}[\Phi_{1/2\ell}(\mathbf{x}_{T+1})])}{\eta_{\mathbf{x}}(T+1)} + 8\ell\zeta + 4\eta_{\mathbf{x}} \ell (L^2 + \sigma^2).$$

By the definition of $\eta_{\mathbf{x}}$ and $\widehat{\Delta}_\Phi$, we have

$$\tfrac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}[\|\nabla \Phi_{1/2\ell}(\mathbf{x}_t)\|^2] \leq \tfrac{48\ell(L^2+\sigma^2)\widehat{\Delta}_\Phi}{\epsilon^2(T+1)} + 8\ell\zeta + \tfrac{\epsilon^2}{3}.$$

This implies that the number of iterations required by Algorithm 4 to return an ϵ-stationary point is bounded by

$$O\left(\frac{\ell(L^2 + \sigma^2)\widehat{\Delta}_\Phi}{\epsilon^4}\right).$$

Combining Lemma 2.12.8 gives the total gradient complexity of Algorithm 3:

$$O\left(\frac{\ell(L^2 + \sigma^2)\widehat{\Delta}_\Phi}{\epsilon^4} \max\left\{1, \frac{\ell^2 D^2}{\epsilon^2}, \frac{\ell^2 D^2 \sigma^2}{\epsilon^4}\right\}\right).$$

This completes the proof.

Chapter 3

Near-Optimal Gradient-Based Algorithm

This chapter resolves a longstanding open question pertaining to the design of near-optimal first-order algorithms for smooth and strongly-convex-strongly-concave minimax problems. Current state-of-the-art first-order algorithms find an approximate Nash equilibrium using $\tilde{O}(\kappa_{\mathbf{x}} + \kappa_{\mathbf{y}})$ [Tseng, 1995] or $\tilde{O}(\min\{\kappa_{\mathbf{x}}\sqrt{\kappa_{\mathbf{y}}}, \sqrt{\kappa_{\mathbf{x}}}\kappa_{\mathbf{y}}\})$ [Alkousa et al., 2020] gradient evaluations, where $\kappa_{\mathbf{x}}$ and $\kappa_{\mathbf{y}}$ are the condition numbers for the strong-convexity and strong-concavity assumptions. A gap still remains between these results and the best existing lower bound $\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$ [Ibrahim et al., 2020, Zhang et al., 2022a]. This chapter presents the first algorithm with $\tilde{O}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$ gradient complexity, matching the lower bound up to logarithmic factors. Our algorithm is designed based on an accelerated proximal point method and an accelerated solver for minimax proximal steps. It can be easily extended to the settings of strongly-convex-concave, convex-concave, nonconvex-strongly-concave, and nonconvex-concave functions. This chapter also presents algorithms that match or outperform all existing methods in these settings in terms of gradient complexity, up to logarithmic factors.

3.1 Introduction

Let $\mathbb{R}^m$ and $\mathbb{R}^n$ be finite-dimensional Euclidean spaces and let the function $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ be smooth. Let $\mathcal{X}$ and $\mathcal{Y}$ are two nonempty closed convex sets in $\mathbb{R}^m$ and $\mathbb{R}^n$. Our problem of interest is the following minimax optimization problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}). \tag{3.1}$$

The theoretical study of solutions of problem (3.1) has been an focus of several decades of research in mathematics, statistics, economics and computer science [Basar and Olsder, 1999, Nisan et al., 2007, Von Neumann and Morgenstern, 2007, Facchinei and Pang, 2007, Berger, 2013]. Recently, this line of research has become increasingly relevant to algorithmic machine

learning, with applications including robustness in adversarial learning [Goodfellow et al., 2014, Sinha et al., 2018], prediction and regression problems [Cesa-Bianchi and Lugosi, 2006, Xu et al., 2009] and distributed computing [Shamma, 2008, Mateos et al., 2010]. Moreover, real-world machine-learning systems are increasingly embedded in multi-agent systems or matching markets and subject to game-theoretic constraints [Jordan, 2018].

Most existing work on minimax optimization focuses on the convex-concave setting, where the function $f(\cdot, \mathbf{y})$ is convex for each $\mathbf{y} \in \mathbb{R}^n$ and the function $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathbb{R}^m$. The best known convergence rate in a general convex-concave setting is $O(1/\epsilon)$ in terms of duality gap, which can be achieved by Nemirovski's mirror-prox algorithm [Nemirovski, 2004] (a special case of which is the extragradient algorithm [Korpelevich, 1976]), Nesterov's dual extrapolation algorithm [Nesterov, 2007] or Tseng's accelerated proximal gradient algorithm [Tseng, 2008]. This rate is known to be optimal for the class of smooth convex-concave problems [Ouyang and Xu, 2021]. Furthermore, optimal algorithms are known for special instances of convex-concave setting; e.g., for the affinely constrained smooth convex problem [Ouyang et al., 2015] and problems with a composite bilinear objective function, $f(\mathbf{x}, \mathbf{y}) = g(\mathbf{x}) + \mathbf{x}^\top A\mathbf{y} - h(\mathbf{y})$ [Chen et al., 2014].

Very recently, the lower complexity bound of first-order algorithms have been established for solving general strongly-convex-strongly-concave and strongly-convex-concave minimax optimization problems [Ouyang and Xu, 2021, Ibrahim et al., 2020, Zhang et al., 2022a]. For the strongly-convex-strongly-concave setting, in which $\kappa_{\mathbf{x}}, \kappa_{\mathbf{y}} > 0$ are the condition numbers for $f(\cdot, \mathbf{y})$ and $f(\mathbf{x}, \cdot)$, respectively, the complexity bound is $\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$ while the best known upper bounds are $\tilde{O}(\kappa_{\mathbf{x}} + \kappa_{\mathbf{y}})$ [Tseng, 1995, Gidel et al., 2019a, Mokhtari et al., 2020b] and $\tilde{O}(\min\{\kappa_{\mathbf{x}}\sqrt{\kappa_{\mathbf{y}}}, \kappa_{\mathbf{y}}\sqrt{\kappa_{\mathbf{x}}}\})$ [Alkousa et al., 2020]. For the strongly-convex-concave setting in which $\kappa_{\mathbf{x}} > 0$ and $\kappa_{\mathbf{y}} = 0$, the lower complexity bound is $\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$ while the best known upper bound is $O(\kappa_{\mathbf{x}}/\epsilon)$ [Thekumparampil et al., 2019]. The existing algorithms that obtain a rate of $O(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$ in this context are only for special case of strongly-convex-linear, where $\mathbf{x}$ and $\mathbf{y}$ are connected only through a bilinear term $\mathbf{x}^\top A\mathbf{y}$ or $f(\mathbf{x}, \cdot)$ is linear for each $\mathbf{x} \in \mathbb{R}^m$ [see, e.g., Nesterov, 2005, Chambolle and Pock, 2016, Juditsky and Nemirovski, 2011, Hamedani and Aybat, 2021]. Thus, a gap remains between the lower complexity bound and the upper complexity bound for existing algorithms in both the strongly-convex-strongly-concave setting and the strongly-convex-concave setting. Accordingly, we have the following open problem:

Can we design first-order algorithms that achieve the lower bounds in these settings?

We present an affirmative answer by resolving the above open problem up to logarithmic factors. More specifically, our contribution is as follows. We propose the first near-optimal algorithms for solving the strongly-convex-strongly-concave and strongly-convex-concave minimax optimization problems. In the former setting, our algorithm achieves a gradient complexity of $\tilde{O}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$ which matches the lower complexity bound [Ibrahim et al., 2020, Zhang et al., 2022a] up to logarithmic factors. In the latter setting, our algorithm attains a gra-

Table 3.1: Comparison of gradient complexities to find an ϵ-saddle point (Definition 3.3.4) in the convex-concave setting. This table highlights only the dependency on error tolerance ϵ and the strong-convexity and strong-concavity condition numbers, $\kappa_{\mathbf{x}}, \kappa_{\mathbf{y}}$.

Settings	References	Gradient Complexity
Strongly-Convex-Strongly-Concave	Tseng [1995]	$\tilde{O}(\kappa_{\mathbf{x}} + \kappa_{\mathbf{y}})$
	Nesterov and Scrimali [2006]	
	Gidel et al. [2019a]	
	Mokhtari et al. [2020b]	
	Alkousa et al. [2020]	$\tilde{O}(\min\{\kappa_{\mathbf{x}}\sqrt{\kappa_{\mathbf{y}}}, \kappa_{\mathbf{y}}\sqrt{\kappa_{\mathbf{x}}}\})$
	Theorem 3.5.1	$\tilde{O}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$
	Lower bound [Ibrahim et al., 2020]	$\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$
	Lower bound [Zhang et al., 2022a]	$\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$
Strongly-Convex-Linear (special case of strongly-convex-concave)	Juditsky and Nemirovski [2011]	$O(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$
	Hamedani and Aybat [2021]	
	Zhao [2022]	
Strongly-Convex-Concave	Thekumparampil et al. [2019]	$\tilde{O}(\kappa_{\mathbf{x}}/\sqrt{\epsilon})$
	Corollary 3.5.2	$\tilde{O}(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$
	Lower bound [Ouyang and Xu, 2021]	$\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$
Convex-Concave	Nemirovski [2004]	$O(\epsilon^{-1})$
	Nesterov [2007]	
	Tseng [2008]	
	Corollary 3.5.3	$\tilde{O}(\epsilon^{-1})$
	Lower bound [Ouyang and Xu, 2021]	$\Omega(\epsilon^{-1})$

dient complexity of $\tilde{O}(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$ which again matches the lower complexity bound [Ouyang and Xu, 2021] up to logarithmic factors. In addition, our algorithm extends to the general convex-concave setting, achieving a gradient complexity of $\tilde{O}(\epsilon^{-1})$, which matches the lower bound of Ouyang and Xu [2021] as well as the best existing upper bounds [Nemirovski, 2004, Nesterov, 2007, Tseng, 2008] up to logarithmic factors.

Our second contribution is a class of accelerated algorithms for the smooth nonconvex-strongly-concave and nonconvex-concave minimax optimization problems. In the former setting, our algorithm achieves a gradient complexity bound of $\tilde{O}(\sqrt{\kappa_{\mathbf{y}}}\epsilon^{-2})$ which improves

Table 3.2: Comparison of gradient complexities to find an ϵ-stationary point of f (Definition 3.3.5) or ϵ-stationary point of $\Phi(\cdot) := \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ (Definition 3.8.1 and Definition 3.8.5) in the nonconvex-concave settings. This table only highlights the dependence on tolerance ϵ and the condition number $\kappa_{\mathbf{y}}$.

Settings	References	Gradient Complexity
Nonconvex-Strongly-Concave (stationarity of f or stationarity of Φ)	Jin et al. [2020]	$\tilde{O}(\kappa_{\mathbf{y}}^2\epsilon^{-2})$
	Lu et al. [2020]	
	Lin et al. [2020c]	
	Rafique et al. [2022]	
	Theorem 3.6.1 & 3.8.7	$\tilde{O}(\sqrt{\kappa_{\mathbf{y}}}\epsilon^{-2})$
Nonconvex-Concave (stationarity of f)	Lu et al. [2020]	$\tilde{O}(\epsilon^{-4})$
	Nouiehed et al. [2019]	$\tilde{O}(\epsilon^{-3.5})$
	Ostrovskii et al. [2021]	$\tilde{O}(\epsilon^{-2.5})$
	Corollary 3.6.2	$\tilde{O}(\epsilon^{-2.5})$
Nonconvex-Concave (stationarity of Φ)	Jin et al. [2020]	$\tilde{O}(\epsilon^{-6})$
	Lin et al. [2020c]	
	Rafique et al. [2022]	
	Thekumparampil et al. [2019]	$\tilde{O}(\epsilon^{-3})$
	Kong and Monteiro [2021]	
	Zhao [2023]	
	Corollary 3.8.8	$\tilde{O}(\epsilon^{-3})$

the best known bound $\tilde{O}(\kappa_{\mathbf{y}}^2\epsilon^{-2})$ [Jin et al., 2020, Lin et al., 2020c, Lu et al., 2020, Rafique et al., 2022]. In the latter setting, our algorithms specialize to a range of different notions of optimality. In particular, expressing our results in terms of stationarity of f, our algorithm achieves a gradient complexity bound of $\tilde{O}(\epsilon^{-2.5})$, which improves the best known bound $\tilde{O}(\epsilon^{-3.5})$ [Nouiehed et al., 2019]. In terms of stationarity of the function $\Phi(\cdot) := \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$, our algorithm achieves a gradient complexity bound of $\tilde{O}(\epsilon^{-3})$ which matches the state-of-the-art results [Thekumparampil et al., 2019, Kong and Monteiro, 2021].

We provide a head-to-head comparison between our results and existing results in Table 3.1 for convex-concave settings, and Table 3.2 for nonconvex-concave settings.

3.2 Related Works

To the best of our knowledge, the earliest algorithmic schemes for solving the bilinear minimax problem, $\min_{\mathbf{x}\in\Delta^m}\max_{\mathbf{y}\in\Delta^n}\mathbf{x}^\top A\mathbf{y}$, date back to Brown's fictitious play [Brown, 1951] and Dantzig's simplex method [Dantzig, 1998]. This problem can also be solved by Korpelevich's extragradient (EG) algorithm [Korpelevich, 1976], which can be shown to be linearly convergent when A is square and full rank [Tseng, 1995]. There are several recent papers studying the convergence of EG and its variants, such as reflected gradient descent ascent [Chambolle and Pock, 2011, Malitsky, 2015, Yadav et al., 2018], optimistic gradient descent ascent (OGDA) [Daskalakis et al., 2018, Mokhtari et al., 2020b,a] and other variants [Rakhlin and Sridharan, 2013a,b, Mertikopoulos et al., 2019, Chavdarova et al., 2019, Hsieh et al., 2019, Mishchenko et al., 2020]. In the bilinear setting, Daskalakis et al. [2018] established the convergence of the optimistic gradient descent ascent (OGDA) method to a neighborhood of the solution; Liang and Stokes [2019] proved the linear convergence of the OGDA algorithm using a dynamical system approach. Very recently, Mokhtari et al. [2020b] have proposed a unified framework for achieving the sharpest convergence rates of both EG and OGDA algorithms.

For the convex-concave minimax problem, Nemirovski [2004] proved that his mirror-prox algorithm returns an ϵ-saddle point within the gradient complexity of $O(\epsilon^{-1})$ when $\mathcal{X}$ and $\mathcal{Y}$ are bounded. This algorithm was subsequently generalized by Auslender and Teboulle [2005] to a class of distance-generating functions, and the complexity result was extended to unbounded sets and composite objectives [Monteiro and Svaiter, 2010, 2011] using the hybrid proximal extragradient algorithm with different error criteria. Nesterov [2007] developed a dual extrapolation algorithm which possesses the same complexity bound as in Nemirovski [2004]. Later on, Tseng [2008] presented a unified treatment of these algorithms and a refined convergence analysis with same complexity result. Nedić and Ozdaglar [2009] analyzed the (sub)gradient descent ascent algorithm for convex-concave saddle point problems when the (sub)gradients are bounded over the constraint sets. Abernethy et al. [2021] presented a Hamiltonian gradient descent algorithm with last-iterate convergence under a "sufficiently bilinear" condition.

Several papers have studied special cases in the convex-concave setting. For the special case when the objective function is a composite bilinear form, $f(\mathbf{x},\mathbf{y}) = g(\mathbf{x}) + \mathbf{x}^\top A\mathbf{y} - h(\mathbf{y})$, Chambolle and Pock [2011] introduced a primal-dual algorithm that converges to a saddle point with the rate of $O(1/\epsilon)$ when the convex functions g and h are smooth. Nesterov [2005] proposed a smoothing technique and proved that the resulting algorithm achieves an improved rate with better dependence on Lipschitz constant of ∇g when h is the convex and smooth function and $\mathcal{X},\mathcal{Y}$ are both bounded. He and Monteiro [2016] and Kolossoski and Monteiro [2017] proved that such result also hold when $\mathcal{X},\mathcal{Y}$ are unbounded or the space is non-Euclidean. Chen et al. [2014, 2017] generalized Nesterov's technique to develop optimal algorithms for solving a class of stochastic saddle point problems and stochastic monotone variational inequalities. For a class of certain purely bilinear games where g and h are zero functions, Azizian et al. [2020b] demonstrated that linear convergence is possible for several

algorithms and their new algorithm achieved the tight bound. The second case is the so-called affinely constrained smooth convex problem, i.e., $\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x})$, s.t. $A\mathbf{x} = \mathbf{u}$. Esser et al. [2010] proposed a $O(\epsilon^{-1})$ primal-dual algorithm while Lan and Monteiro [2016] provided a first-order augmented Lagrangian method with the same $O(\epsilon^{-1})$ rate. By exploiting the structure, Ouyang et al. [2015] proposed a near-optimal algorithm in this setting.

For the strongly convex-concave minimax problem, Tseng [1995] and Nesterov and Scrimali [2006] proved that their algorithms find an ϵ-saddle point with a gradient complexity of $\tilde{O}(\kappa_{\mathbf{x}} + \kappa_{\mathbf{y}})$ using a variational inequality. Using a different approach, Gidel et al. [2019a] and Mokhtari et al. [2020b] derived the same complexity results for the OGDA algorithm. Very recently, Alkousa et al. [2020] proposed an accelerated gradient sliding algorithm with a gradient complexity of $\tilde{O}(\min\{\kappa_{\mathbf{x}}\sqrt{\kappa_{\mathbf{y}}}, \kappa_{\mathbf{y}}\sqrt{\kappa_{\mathbf{x}}}\})$ while Ibrahim et al. [2020] and Zhang et al. [2022a] established a lower complexity bound of $\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$ among all the first-order algorithms in this setting.

For strongly-convex-concave minimax problems, the best known general lower bound for first-order algorithm is $O(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$, as shown by Ouyang and Xu [2021]. Several papers have studied strongly-convex-concave minimax problem with additional structures. This includex optimizing a strongly convex function with linear constraints [Goldstein et al., 2014, Xu and Zhang, 2018, Xu, 2021], the case when $\mathbf{x}$ and $\mathbf{y}$ are connected only through a bilinear term $\mathbf{x}^\top A\mathbf{y}$ [Nesterov, 2005, Chambolle and Pock, 2016, Xie and Shi, 2019] and the case when $f(\mathbf{x}, \cdot)$ is linear for each $\mathbf{x} \in \mathbb{R}^m$ [Juditsky and Nemirovski, 2011, Hamedani and Aybat, 2021, Zhao, 2022]. The algorithms developed in these works were all guaranteed to return an ϵ-saddle point with a gradient complexity of $\tilde{O}(1/\sqrt{\epsilon})$ and some of them even achieve a near-optimal gradient complexity of $\tilde{O}(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$ [Nesterov, 2005, Chambolle and Pock, 2016]. However, the best known upper complexity bound for general strongly-convex-concave minimax problems is $O(\kappa_{\mathbf{x}}/\sqrt{\epsilon})$ which was shown using the *dual implicit accelerated gradient* algorithm [Thekumparampil et al., 2019].

For nonconvex-concave minimax problems, a line of recent work [Jin et al., 2020, Lin et al., 2020c, Rafique et al., 2022] has studied various algorithms and proved that they can find an approximate stationary point of $\Phi(\cdot) := \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$. In a deterministic setting, all of these algorithms guarantee a rate of $\tilde{O}(\kappa_{\mathbf{y}}^2\epsilon^{-2})$ and $\tilde{O}(\epsilon^{-6})$ when $f(\mathbf{x}, \cdot)$ is strongly concave and concave respectively. Thekumparampil et al. [2019] consider the same setting as ours and proposed a proximal dual implicit accelerated gradient algorithm and proved that it finds an approximate stationary point of $\Phi(\cdot)$ with the total gradient complexity of $\tilde{O}(\epsilon^{-3})$. Kong and Monteiro [2021] consider a general nonconvex minimax optimization model: $\min_{\mathbf{x}} h(\mathbf{x})+\rho(\mathbf{x})$, where h is a "simple" proper, lower semi-continuous and convex function and $\rho(\mathbf{x}) = \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})$ with f satisfying that $-f(\mathbf{x}, \cdot)$ is proper, convex, and lower semi-continuous. They propose to smooth ρ to $\rho_\xi(\mathbf{x}) = \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) - (1/2\xi)\|\mathbf{y} - \mathbf{y}_0\|^2$ and apply an accelerated inexact proximal point method to solve the smoothed problem $\min_{\mathbf{x}} h(\mathbf{x}) + \rho_\xi(\mathbf{x})$. The resulting AIPP-S algorithm attains the iteration complexity of $O(\epsilon^{-3})$ using a slightly different but equivalent notion of stationarity but requires the *exact* gradient of ρ_ξ at each iteration. This amounts to assuming that $\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) - (1/2\xi)\|\mathbf{y} - \mathbf{y}_0\|^2$ can be solved

exactly, which is restrictive due to the potentially complicated structure of $f(\mathbf{x}, \cdot)$ or $\mathcal{Y}$. If f is further assumed to be smooth, Zhao [2023] developed a variant of AIPP-S algorithm which only requires an inexact gradient of ρ_ξ at each iteration and attains the total gradient complexity of $\tilde{O}(\epsilon^{-3})$. On the other hand, the stationarity of $f(\cdot, \cdot)$ is proposed for quantifying the efficiency in nonconvex-concave minimax optimization [Nouiehed et al., 2019, Lu et al., 2020, Kong and Monteiro, 2021, Ostrovskii et al., 2021]. Using this notion of stationarity, Kong and Monteiro [2021] attains the rate of $O(\epsilon^{-2.5})$ but requires the *exact* gradient of ρ_ξ at each iteration. Without this assumption, the current state-of-the-art rate is $\tilde{O}(\epsilon^{-2.5})$ achieved by our Algorithm 9 and the algorithm proposed by a concurrent work [Ostrovskii et al., 2021]. Both algorithms are based on constructing an auxiliary function $f_{\epsilon,\mathbf{y}}$ and applying an accelerated solver for minimax proximal steps. Finally, several other algorithms have been developed either for specific nonconvex-concave minimax problems or in stochastic setting; see Namkoong and Duchi [2016], Sinha et al. [2018], Grnarova et al. [2018] for the details.

3.3 Preliminaries

We clarify the notation, review some background and provide formal definitions for the class of functions and optimality measure considered in this paper.

Notation. We use bold lower-case letters to denote vectors, as in $\mathbf{x}, \mathbf{y}, \mathbf{z}$ and calligraphic upper case letters to denote sets, as in $\mathcal{X}$ and $\mathcal{Y}$. For a differentiable function $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$, we le4t $\nabla f(\mathbf{z})$ denote the gradient of f at $\mathbf{z}$. For a function $f(\cdot, \cdot) : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ of two variables, $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$ (or $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$) to denote the partial gradient of f with respect to the first variable (or the second variable) at point $(\mathbf{x}, \mathbf{y})$. We also use $\nabla f(\mathbf{x}, \mathbf{y})$ to denote the full gradient at $(\mathbf{x}, \mathbf{y})$ where $\nabla f(\mathbf{x}, \mathbf{y}) = (\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))$. For a vector $\mathbf{x}$, we denote $\|\mathbf{x}\|$ as its ℓ_2-norm. For constraint sets $\mathcal{X}$ and $\mathcal{Y}$, we let $D_{\mathbf{x}}$ and $D_{\mathbf{y}}$ denote their diameters, where $D_{\mathbf{x}} = \max_{\mathbf{x},\mathbf{x}' \in \mathcal{X}} \|\mathbf{x} - \mathbf{x}'\|$ and $D_{\mathbf{y}} = \max_{\mathbf{y},\mathbf{y}' \in \mathcal{Y}} \|\mathbf{y} - \mathbf{y}'\|$. We use the notation $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{Y}}$ to denote projections onto the sets $\mathcal{X}$ and $\mathcal{Y}$. Finally, we use the notation $O(\cdot), \Omega(\cdot)$ to hide only absolute constants which do not depend on any problem parameter, and notation $\tilde{O}(), \tilde{\Omega}()$ to hide only absolute constants and log factors.

Minimax optimization. We are interested in the ℓ-smooth minimax optimization problems in the form (3.1). The regularity conditions for the function f are as follows.

Definition 3.3.1 *A function f is L-Lipschitz if for $\forall \mathbf{z}, \mathbf{z}' \in \mathbb{R}^n$, $|f(\mathbf{z}) - f(\mathbf{z}')| \le L\|\mathbf{z} - \mathbf{z}'\|$.*

Definition 3.3.2 *A function f is ℓ-smooth if for $\forall \mathbf{z}, \mathbf{z}' \in \mathbb{R}^n$, $\|\nabla f(\mathbf{z}) - \nabla f(\mathbf{z}')\| \le \ell\|\mathbf{z} - \mathbf{z}'\|$.*

Definition 3.3.3 *A differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly-convex if for $\forall \mathbf{x}', \mathbf{x} \in \mathbb{R}^d$, $\phi(\mathbf{x}') \ge \phi(\mathbf{x}) + (\mathbf{x}' - \mathbf{x})^\top \nabla\phi(\mathbf{x}) + (\mu/2)\|\mathbf{x}' - \mathbf{x}\|^2$. Furthermore, ϕ is μ-strongly-concave*

if $-\phi$ is μ-strongly-convex. If we set $\mu = 0$, then we recover the definitions of convexity and concavity for a continuous differentiable function.

Convex-concave setting: we assume that $f(\cdot, \mathbf{y})$ is convex for each $\mathbf{y} \in \mathcal{Y}$ and $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathcal{X}$. *Here $\mathcal{X}$ and $\mathcal{Y}$ are both convex and bounded.* Under these conditions, the Sion's minimax theorem [Sion, 1958] guarantees that

$$\max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}). \tag{3.2}$$

Furthermore, there exists at least one **saddle point (or Nash equilibrium)** $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{X} \times \mathcal{Y}$ such that the following equality holds true:

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \mathbf{y}^\star) = f(\mathbf{x}^\star, \mathbf{y}^\star) = \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star, \mathbf{y}). \tag{3.3}$$

Therefore, for any point $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$, the duality gap $\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}})$ forms the basis for a standard optimality criterion. Formally, we define

Definition 3.3.4 *A point $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$ is an ϵ-**saddle point** of a convex-concave function $f(\cdot, \cdot)$ if $\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}) \le \epsilon$. If $\epsilon = 0$, then $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is a saddle point.*

When $f(\cdot, \mathbf{y})$ is strongly convex for each $\mathbf{y} \in \mathcal{Y}$ and $f(\mathbf{x}, \cdot)$ is strongly concave for each $\mathbf{x} \in \mathcal{X}$, we let $\mu_\mathbf{x}$ and $\mu_\mathbf{y}$ be strongly-convex and strongly-concave modules. If f is ℓ-smooth, we denote $\kappa_\mathbf{x} = \ell/\mu_\mathbf{x}$ and $\kappa_\mathbf{y} = \ell/\mu_\mathbf{y}$ as the condition numbers of $f(\cdot, \mathbf{y})$ and $f(\mathbf{x}, \cdot)$.

Nonconvex-concave setting: we only assume that $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathbb{R}^m$. The function $f(\cdot, \mathbf{y})$ can be possibly nonconvex for some $\mathbf{y} \in \mathcal{Y}$. *Here $\mathcal{X}$ is convex but possibly unbounded while $\mathcal{Y}$ is convex and bounded.* In general, finding a global Nash equilibrium of f is intractable since in the special case where $\mathcal{Y}$ has only a single element, this problem reduces to a nonconvex optimization problem in which finding a global minimum is already NP-hard [Murty and Kabadi, 1987]. Similar to the literature in nonconvex constrained optimization, we opt to find local surrogates—stationary points—whose gradient mappings are zero. Formally, we define our optimality criterion as follows.

Definition 3.3.5 *A point $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$ is an ϵ-**stationary point** of an ℓ-smooth function $f(\cdot, \cdot)$ if*

$$\ell\|\mathcal{P}_\mathcal{X}[\hat{\mathbf{x}} - (1/\ell)\nabla_\mathbf{x} f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+)] - \hat{\mathbf{x}}\| \le \epsilon, \quad \ell\|\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}\| \le \epsilon.$$

where

$$\hat{\mathbf{y}}^+ = \mathcal{P}_\mathcal{Y}[\hat{\mathbf{y}} + (1/\ell)\nabla_\mathbf{y} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})].$$

If $\epsilon = 0$, then $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is a stationary point.

Algorithm 5 AGD$(g, \mathcal{X}, \mathbf{x}_0, \ell, \mu, \epsilon)$

1: **Input:** initial point $\mathbf{x}_0 \in \mathcal{X}$, smoothness ℓ, strongly-convex module μ and tolerance $\epsilon > 0$.
2: **Initialize:** set $t \leftarrow 0$, $\tilde{\mathbf{x}}_0 \leftarrow \mathbf{x}_0$, $\eta \leftarrow 1/\ell$, $\kappa \leftarrow \ell/\mu$ and $\theta \leftarrow \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$.
3: **repeat**
4: $\quad t \leftarrow t + 1$
5: $\quad \mathbf{x}_t \leftarrow \mathcal{P}_{\mathcal{X}}[\tilde{\mathbf{x}}_{t-1} - \eta \nabla g(\tilde{\mathbf{x}}_{t-1})]$.
6: $\quad \tilde{\mathbf{x}}_t \leftarrow \mathbf{x}_t + \theta(\mathbf{x}_t - \mathbf{x}_{t-1})$.
7: **until** $\|\mathbf{x}_t - \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - \eta \nabla g(\mathbf{x}_t))\|^2 \leq \frac{\epsilon}{2\kappa^2(\ell-\mu)}$ is satisfied.
8: **Output:** $\mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - \eta \nabla g(\mathbf{x}_t))$.

In the absence of the constraint set $\mathcal{X}$, Definition 3.3.5 reduces to the standard condition $\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+)\| \leq \epsilon$ and $\ell\|\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}\| \leq \epsilon$ which is consistent with Lin et al. [2020c, Definition 4.10]. Intuitively, the quantity $\|\mathcal{P}_{\mathcal{Y}}[\hat{\mathbf{y}} + (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})] - \hat{\mathbf{y}}\|$ represents the distance between a point $\hat{\mathbf{y}}$ and a point obtained by performing one-step projected partial gradient ascent at a point $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ starting from a point $\hat{\mathbf{y}}$. It also refers to the norm of gradient mapping at $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$; see Nesterov [2013a] for the details.

We note that this notion of stationarity of f (Definition 3.3.5) is closely related to an optimality notion in terms of stationary points of the function $\Phi(\cdot) := \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$ for nonconvex-concave functions.

Nesterov's accelerated gradient descent. Nesterov's Accelerated Gradient Descent (AGD) dates back to the seminal paper [Nesterov, 1983] where it is shown to be optimal among all the first-order algorithms for smooth and convex functions [Nesterov, 2018]. We present a version of AGD in Algorithm 5 which is frequently used to minimize an ℓ-smooth and μ-strongly convex function g over a convex set $\mathcal{X}$. The key steps of the AGD algorithm are Line 5-6, where Lines 5 performs a projected gradient descent step, while Line 6 performs a momentum step, which "overshoots" the iterate in the direction of momentum $(\mathbf{x}_t - \mathbf{x}_{t-1})$. Line 7 is the stopping condition to ensure that the output achieves the desired optimality.

The following theorem provides an upper bound on the gradient complexity of AGD; i.e., the total number of gradient evaluations to find an ϵ-optimal point.

Theorem 3.3.6 *Assume that g is ℓ-smooth and μ-strongly convex, $\hat{\mathbf{x}} = \text{AGD}(g, \mathbf{x}_0, \ell, \mu, \epsilon)$ satisfies $g(\hat{\mathbf{x}}) \leq \min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}) + \epsilon$ and the total number of gradient evaluations is bounded by*

$$O\left(\sqrt{\kappa} \log\left(\frac{\kappa^3 \ell \|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\epsilon}\right)\right),$$

where $\kappa = \ell/\mu$ is the condition number, and $\mathbf{x}^\star \in \mathcal{X}$ is the unique global minimum of g.

Compared with the classical result for Gradient Descent (GD), which requires $\tilde{O}(\kappa)$ gradient evaluations in the same setting, AGD improves over GD by a factor of $\sqrt{\kappa}$. AGD will be used as a basic component for acceleration.

Algorithm 6 INEXACT-APPA$(g, \mathbf{x}_0, \ell, \mu, \epsilon, T)$

1: **Input:** initial point $\mathbf{x}_0 \in \mathcal{X}$, proximal parameter ℓ, strongly-convex module μ, tolerance $\epsilon > 0$ and the maximum iteration number $T > 0$.
2: **Initialize:** set $\tilde{\mathbf{x}}_0 \leftarrow \mathbf{x}_0$, $\kappa \leftarrow \frac{\ell}{\mu}$, $\delta \leftarrow \frac{\epsilon}{(10\kappa)^2}$ and $\theta \leftarrow \frac{2\sqrt{\kappa}-1}{2\sqrt{\kappa}+1}$.
3: **for** $t = 1, \cdots, T$ **do**
4: find $\mathbf{x}_t$ so that $g(\mathbf{x}_t) + \ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 \le \min_{\mathbf{x}\in\mathcal{X}}\{g(\mathbf{x}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2\} + \delta$.
5: $\tilde{\mathbf{x}}_t \leftarrow \mathbf{x}_t + \theta(\mathbf{x}_t - \mathbf{x}_{t-1})$.
6: **Output:** $\mathbf{x}_T$.

3.4 Algorithm Components

We present two main algorithm components. Both of them are crucial for our final algorithms to achieve near-optimal convergence rates.

Inexact accelerated proximal point algorithm. Our first component is the Accelerated Proximal Point Algorithm (APPA, Algorithm 6) for minimizing a function $g(\cdot)$. Comparing APPA with classical AGD (Algorithm 5), we note that both of them have momentum steps which yield acceleration. The major difference is in Line 4 of Algorithm 6, where APPA solves a proximal subproblem

$$\mathbf{x}_t \leftarrow \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \; g(\mathbf{x}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2. \tag{3.4}$$

instead of performing a gradient-descent step as in AGD (Line 5 in Algorithm 5). We refer to the parameter ℓ in (3.4) as the *proximal parameter*.

We present an inexact version in Algorithm 6 where we tolerate a small error δ in terms of the function value in solving the proximal subproblem (3.4). That is, the solution $\mathbf{x}_t$ satisfies

$$g(\mathbf{x}_t) + \ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 \le \min_{\mathbf{x}\in\mathcal{X}}\{g(\mathbf{x}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2\} + \delta.$$

A theoretical guarantee for the inexact APPA algorithm is presented in the following theorem, which claims that as long as δ is sufficiently small, the algorithm finds an ϵ-optimal point of any μ-strongly-convex function g with proximal parameter ℓ in $\tilde{O}(\sqrt{\ell/\mu})$ iterations.

Theorem 3.4.1 *Assume that g is μ-strongly convex, $\epsilon \in (0, 1)$ and $\ell > \mu$. There exists $T > 0$ such that the output $\hat{\mathbf{x}} =$ INEXACT-APPA$(g, \mathbf{x}_0, \ell, \mu, \epsilon, T)$ satisfies $g(\hat{\mathbf{x}}) \le \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) + \epsilon$ and $T > 0$ satisfies the following inequality,*

$$T \ge c\sqrt{\kappa}\log\left(\frac{g(\mathbf{x}_0) - g(\mathbf{x}^\star) + (\mu/4)\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\epsilon}\right),$$

where $\kappa = \ell/\mu$ is an effective condition number, $\mathbf{x}^\star \in \mathcal{X}$ is the unique global minimum of g, and $c > 0$ is an absolute constant.

Algorithm 7 MAXIMIN-AG2$(g, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_{\mathbf{x}}, \mu_{\mathbf{y}}, \epsilon)$

Input: initial point $\mathbf{x}_0, \mathbf{y}_0$, smoothness ℓ, strongly convex module $\mu_{\mathbf{x}}, \mu_{\mathbf{y}}$ and tolerance $\epsilon > 0$.
Initialize: $t \leftarrow 0$, $\tilde{\mathbf{x}}_0 \leftarrow \mathbf{x}_0$, $\eta \leftarrow \frac{1}{2\kappa_{\mathbf{x}}\ell}$, $\kappa_{\mathbf{x}} \leftarrow \frac{\ell}{\mu_{\mathbf{x}}}$, $\kappa_{\mathbf{y}} \leftarrow \frac{\ell}{\mu_{\mathbf{y}}}$, $\theta \leftarrow \frac{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}-1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}+1}$, $\tilde{\epsilon} \leftarrow \frac{\epsilon}{(10\kappa_{\mathbf{x}}\kappa_{\mathbf{y}})^7}$.
repeat
  $t \leftarrow t + 1$.
  $\tilde{\mathbf{x}}_{t-1} \leftarrow \mathrm{AGD}(g(\cdot, \tilde{\mathbf{y}}_{t-1}), \mathbf{x}_0, \ell, \mu_{\mathbf{x}}, \tilde{\epsilon})$.
  $\mathbf{y}_t \leftarrow \mathcal{P}_{\mathcal{Y}}[\tilde{\mathbf{y}}_{t-1} + \eta\nabla_{\mathbf{y}} g(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1})]$.
  $\tilde{\mathbf{y}}_t \leftarrow \mathbf{y}_t + \theta(\mathbf{y}_t - \mathbf{y}_{t-1})$.
  $\mathbf{x}_t \leftarrow \mathrm{AGD}(g(\cdot, \mathbf{y}_t), \mathbf{x}_0, \ell, \mu_{\mathbf{x}}, \tilde{\epsilon})$.
until $\|\mathbf{y}_t - \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_t + \eta\nabla_{\mathbf{y}} g(\mathbf{x}_t, \mathbf{y}_t))\|^2 \le \frac{\epsilon}{(10\kappa_{\mathbf{x}}\kappa_{\mathbf{y}})^4\ell}$ is satisfied.
Output: $\mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - (1/2\kappa_{\mathbf{y}}\ell)\nabla_{\mathbf{x}} g(\mathbf{x}_t, \mathbf{y}_t))$.

Comparing with Theorem 3.3.6, the most important difference here is that Theorem 3.4.1 does not require the function g to have any smoothness property. In fact, ℓ is only a proximal parameter in proximal subproblem (3.4), which does not necessarily relate to the smoothness of g. On the flip side, the proximal subproblem (3.4) can not be easily solved in general. Theorem 3.4.1 guarantees the iteration complexity of Algorithm 5 while the complexity for solving these proximal steps is not discussed.

We conclude that APPA has a unique advantage over AGD in settings where g does not have a smoothness property but the proximal step (3.4) is easy to solve. These settings include LASSO [Beck and Teboulle, 2009], as well as minimax optimization problems.

Accelerated solver for minimax proximal steps. In minimax optimization problems of the form (3.1), we are interested in solving the following proximal subproblem as follows,

$$\mathbf{x}_{t+1} \leftarrow \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}\|^2, \quad \text{where} \quad \Phi(\mathbf{x}) := \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}), \tag{3.5}$$

which is equivalent to solving the following minimax problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \tilde{g}(\mathbf{x}, \mathbf{y}) := f(\mathbf{x}, \mathbf{y}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}\|^2. \tag{3.6}$$

For a generic strongly-convex-strongly-concave function $g(\cdot, \cdot)$, solving a minimax problem is equivalent to solving a maximin problem, due to Sion's minimax theorem:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} g(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \mathbf{y}).$$

A straightforward way of solving the maximin problem is to use a double-loop algorithm which solves the maximization and minimization problems on two different time scales. Specifically, the inner loop performs AGD on function $g(\cdot, \mathbf{y})$ to solve the inner minimization; i.e., to compute $\Psi(\mathbf{y}) := \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \mathbf{y})$ for each $\mathbf{y}$, and the outer loop performs

Accelerated Gradient Ascent (AGA) on the function $\Psi(\cdot)$ to solve the outer maximization. Since the algorithm aims to solve a maximin problem we use AGA-AGD, and we name the algorithm MAXIMIN-AG2. See Algorithm 7 for the formal version of this algorithm. We also incorporate Lines 8-9 to check termination conditions, which ensures that the output achieves the desired optimality. The theoretical guarantee for Algorithm 7 is given in the following theorem.

Theorem 3.4.2 *Assume that $g(\cdot,\cdot)$ is ℓ-smooth, $g(\cdot,\mathbf{y})$ is $\mu_{\mathbf{x}}$-strongly convex for each $\mathbf{y}\in\mathcal{Y}$ and $g(\mathbf{x},\cdot)$ is $\mu_{\mathbf{y}}$-strongly concave for each $\mathbf{x}\in\mathcal{X}$. Then $\hat{\mathbf{x}} = \text{MAXIMIN-AG2}(g,\mathbf{x}_0,\mathbf{y}_0,\ell,\mu_{\mathbf{x}},\mu_{\mathbf{y}},\epsilon)$ satisfies that $\max_{\mathbf{y}\in\mathcal{Y}} g(\hat{\mathbf{x}},\mathbf{y}) \le \min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} g(\mathbf{x},\mathbf{y}) + \epsilon$ and the total number of gradient evaluations is bounded by*

$$O\left(\kappa_{\mathbf{x}}\sqrt{\kappa_{\mathbf{y}}}\cdot\log^2\left(\frac{(\kappa_{\mathbf{x}}+\kappa_{\mathbf{y}})\ell(\tilde{D}_{\mathbf{x}}^2+D_{\mathbf{y}}^2)}{\epsilon}\right)\right),$$

where $\kappa_{\mathbf{x}} = \ell/\mu_{\mathbf{x}}$ and $\kappa_{\mathbf{y}} = \ell/\mu_{\mathbf{y}}$ are condition numbers, $\tilde{D}_{\mathbf{x}} = \|\mathbf{x}_0 - \mathbf{x}_g^\star(\mathbf{y}_0)\|$ is the initial distance where $\mathbf{x}_g^\star(\mathbf{y}_0) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\mathbf{y}_0)$ and $D_{\mathbf{y}} > 0$ is the diameter of the set $\mathcal{Y}$.

Theorem 3.4.2 claims that Algorithm 7 finds an ϵ-optimal point in $\tilde{O}(\kappa_{\mathbf{x}}\sqrt{\kappa_{\mathbf{y}}})$ iterations for strongly-convex-strongly-concave functions. This rate does not match the lower bound $\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$ [Ibrahim et al., 2020, Zhang et al., 2022a]. At a high level, it takes AGD $\tilde{O}(\sqrt{\kappa_{\mathbf{x}}})$ steps to solve the inner minimization problem and compute $\Psi(\mathbf{y}) := \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\mathbf{y})$. Despite the fact that the function g is ℓ-smooth, function Ψ is only guaranteed to be $(\kappa_{\mathbf{x}}\ell)$-smooth in the worst case, which makes the condition number of Ψ be $\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}$. Thus, AGA requires $\tilde{O}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$ iterations in the outer loop to solve the maximization of Ψ, which gives a total gradient complexity $\tilde{O}(\kappa_{\mathbf{x}}\sqrt{\kappa_{\mathbf{y}}})$.

The key observation here is that although Algorithm 7 is slow for general strongly-convex-strongly-concave functions, the functions $\tilde{g}$ of the form (3.6) in the proximal steps have a crucial property that $\kappa_{\mathbf{x}} = O(1)$ if the proximal parameter ℓ is chosen to be the smoothness parameter of function f. Therefore, when $f(\mathbf{x},\cdot)$ is strongly concave, by Theorem 3.4.2, it only takes Algorithm 7 $\tilde{O}(\sqrt{\kappa_{\mathbf{y}}})$ gradient evaluations to solve the proximal subproblem (3.6), which is very efficient. We will see the consequences of this fact in the following section.

3.5 Accelerating Convex-Concave Optimization

We present our main results for accelerating convex-concave optimization. We first present our new near-optimal algorithm and its theoretical guarantee for optimizing strongly-convex-strongly-concave functions. Then, we use simple reduction arguments to obtain results for strongly-convex-concave and convex-concave functions.

Algorithm 8 MINIMAX-APPA$(f, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_\mathbf{x}, \mu_\mathbf{y}, \epsilon, T)$

Input: initial point $\mathbf{x}_0, \mathbf{y}_0$, proximity ℓ, strongly-convex parameter μ, tolerance δ, iteration T.
Initialize: $\tilde{\mathbf{x}}_0 \leftarrow \mathbf{x}_0$, $\kappa_\mathbf{x} \leftarrow \frac{\ell}{\mu_\mathbf{x}}$, $\theta \leftarrow \frac{2\sqrt{\kappa_\mathbf{x}}-1}{2\sqrt{\kappa_\mathbf{x}}+1}$, $\delta \leftarrow \frac{\epsilon}{(10\kappa_\mathbf{x}\kappa_\mathbf{y})^4}$ and $\tilde{\epsilon} \leftarrow \frac{\epsilon}{10^2\kappa_\mathbf{x}\kappa_\mathbf{y}}$.
for $t = 1, \cdots, T$ **do**
  denote $g_t(\cdot,\cdot)$ where $g_t(\mathbf{x},\mathbf{y}) := f(\mathbf{x},\mathbf{y}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2$.
  $\mathbf{x}_t \leftarrow$ MAXIMIN-AG2$(g_t, \mathbf{x}_0, \mathbf{y}_0, 3\ell, 2\ell, \mu_\mathbf{y}, \delta)$
  $\tilde{\mathbf{x}}_t \leftarrow \mathbf{x}_t + \theta(\mathbf{x}_t - \mathbf{x}_{t-1})$.
$\tilde{\mathbf{y}} \leftarrow$ AGD$(-f(\mathbf{x}_T,\cdot), \mathbf{y}_0, \ell, \mu_\mathbf{y}, \tilde{\epsilon})$.
$\mathbf{y}_T \leftarrow \mathcal{P}_\mathcal{Y}\left(\tilde{\mathbf{y}} + (1/2\kappa_\mathbf{x}\ell)\nabla_\mathbf{y} f(\mathbf{x}_T, \tilde{\mathbf{y}})\right)$.
Output: $(\mathbf{x}_T, \mathbf{y}_T)$.

Strongly-convex-strongly-concave setting. With the algorithm components in hand, we are now ready to state our near-optimal algorithm. Algorithm 8 is a simple combination of Algorithm 6 and Algorithm 7. Its outer loop performs an inexact APPA to minimize the function $\Phi(\cdot) := \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot,\mathbf{y})$, while the inner loop uses Maximin-AG2 to solve the proximal subproblem (3.5), which is equivalent to solving (3.6). At the end, after finding a near-optimal $\mathbf{x}_T$, Algorithm 8 performs another AGD on the function $-f(\mathbf{x}_T,\cdot)$ to find a near-optimal $\mathbf{y}_T$. The theoretical guarantee for the algorithm is given in the following theorem.

Theorem 3.5.1 *Assume that f is ℓ-smooth and $\mu_\mathbf{x}$-strongly-convex-$\mu_\mathbf{y}$-strongly-concave. Then there exists $T > 0$ such that the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) =$ MINIMAX-APPA$(f, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_\mathbf{x}, \mu_\mathbf{y}, \epsilon, T)$ is an ϵ-saddle point, and the total number of gradient evaluations is bounded by*

$$O\left(\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}\log^3\left(\frac{(\kappa_\mathbf{x}+\kappa_\mathbf{y})\ell(D_\mathbf{x}^2 + D_\mathbf{y}^2)}{\epsilon}\right)\right),$$

where $\kappa_\mathbf{x} = \ell/\mu_\mathbf{x}$ and $\kappa_\mathbf{y} = \ell/\mu_\mathbf{y}$ are condition numbers.

Theorem 3.5.1 asserts that Algorithm 8 finds ϵ-saddle points in $\tilde{O}(\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}})$ gradient evaluations, matching the lower bound [Ibrahim et al., 2020, Zhang et al., 2022a], up to logarithmic factors. At a high level, despite the function Φ having undesirable smoothness properties, APPA minimizes Φ in the outer loop using $\tilde{O}(\sqrt{\kappa_\mathbf{x}})$ iterations according to Theorem 3.4.1, regardless of the smoothness of Φ. In addition, Maximin-AG2 solves the proximal step in the inner loop using $\tilde{O}(\sqrt{\kappa_\mathbf{y}})$ gradient evaluations, since the condition number of $g_t(\cdot,\mathbf{y})$ for any $\mathbf{y} \in \mathcal{Y}$ is $O(1)$. This gives the total gradient complexity $\tilde{O}(\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}})$.

Strongly-convex-concave setting. Our result in the strongly-convex-strongly-concave setting readily implies a near-optimal result in the strongly-convex-concave setting. Consider the following auxiliary function for an arbitrary $\mathbf{y}_0 \in \mathcal{Y}$ which is defined by

$$f_{\epsilon,\mathbf{y}}(\mathbf{x},\mathbf{y}) := f(\mathbf{x},\mathbf{y}) - (\epsilon/4D_\mathbf{y}^2)\|\mathbf{y} - \mathbf{y}_0\|^2. \tag{3.7}$$

It is clear that the difference between f and $f_{\epsilon,\mathbf{y}}$ is small in terms of function value:

$$\max_{(\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{Y}} |f(\mathbf{x},\mathbf{y}) - f_{\epsilon,\mathbf{y}}(\mathbf{x},\mathbf{y})| \le \epsilon/4.$$

This implies, according to Definition 3.3.4, that any $(\epsilon/2)$-saddle point of function $f_{\epsilon,\mathbf{y}}$ is also a ϵ-saddle point of function f, and thus it is sufficient to only solve the problem $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{x}\in\mathcal{Y}} f_{\epsilon,\mathbf{y}}(\mathbf{x},\mathbf{y})$. Finally, when f is a $\mu_{\mathbf{x}}$-strongly-convex-concave function, $f_{\epsilon,\mathbf{y}}$ becomes $\mu_{\mathbf{x}}$-strongly-convex-$\epsilon/(2D_{\mathbf{y}}^2)$-strongly-concave, which can be fed into Algorithm 8 to obtain the following result.

Corollary 3.5.2 *Assume that f is ℓ-smooth and $\mu_{\mathbf{x}}$-strongly-convex-concave. Then there exists $T > 0$ such that the output $(\hat{\mathbf{x}},\hat{\mathbf{y}}) = \text{MINIMAX-APPA}(f_{\epsilon,\mathbf{y}}, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_{\mathbf{x}}, \epsilon/(4D_{\mathbf{y}}^2), \epsilon/2, T)$ is an ϵ-saddle point, and the total number of gradient evaluations is bounded by*

$$O\left(\sqrt{\frac{\kappa_{\mathbf{x}}\ell}{\epsilon}} D_{\mathbf{y}} \log^3\left(\frac{\kappa_{\mathbf{x}}\ell(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right)$$

where $\kappa_{\mathbf{x}} = \ell/\mu_{\mathbf{x}}$ is the condition number, and $f_{\epsilon,\mathbf{y}}$ is defined as in (3.7).

Convex-concave setting. When f is only convex-concave, we can construct following strongly-convex-strongly-concave function f_ϵ:

$$f_\epsilon(\mathbf{x},\mathbf{y}) = f(\mathbf{x},\mathbf{y}) + (\epsilon/8D_{\mathbf{x}}^2)\|\mathbf{x}-\mathbf{x}_0\|^2 - (\epsilon/8D_{\mathbf{y}}^2)\|\mathbf{y}-\mathbf{y}_0\|^2, \tag{3.8}$$

which can be fed into Algorithm 8 to obtain the following result.

Corollary 3.5.3 *Assume function f is ℓ-smooth and convex-concave, then there exists $T > 0$, where the output $(\hat{\mathbf{x}},\hat{\mathbf{y}}) = \text{MINIMAX-APPA}(f_\epsilon, \mathbf{x}_0, \mathbf{y}_0, \ell, \epsilon/(4D_{\mathbf{x}}^2), \epsilon/(4D_{\mathbf{y}}^2), \epsilon/2, T)$ will be an ϵ-saddle point, and the total number of gradient evaluations is bounded by*

$$O\left(\frac{\ell D_{\mathbf{x}} D_{\mathbf{y}}}{\epsilon} \log^3\left(\frac{\ell(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right),$$

where f_ϵ is defined as in (3.8).

3.6 Accelerating Nonconvex-Concave Optimization

We present our methods for accelerating nonconvex-concave optimization. Similar to the previous section, we first present our algorithm and its theoretical guarantee for optimizing nonconvex-strongly-concave functions. We then use a simple reduction argument to obtain results for nonconvex-concave functions using the stationarity of the function f (Definition 3.3.5) as an optimality measure.

Algorithm 9 MINIMAX-PPA$(g, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_{\mathbf{y}}, \epsilon, T)$

Input: initial point $\mathbf{x}_0, \mathbf{y}_0$, proximity ℓ, strongly-convex parameter μ, tolerance δ, iteration T.
Initialize: set $\delta \leftarrow \frac{\epsilon^2}{(10\kappa_{\mathbf{y}})^4 \ell} \cdot (\frac{\epsilon}{\ell D_{\mathbf{y}}})^2$.
for $t = 1, \cdots, T$ **do**
  denote $g_t(\cdot, \cdot)$ where $g_t(\mathbf{x}, \mathbf{y}) := f(\mathbf{x}, \mathbf{y}) + \ell \|\mathbf{x} - \mathbf{x}_{t-1}\|^2$.
  $\mathbf{x}_t \leftarrow$ MAXIMIN-AG2$(g_t, \mathbf{x}_0, \mathbf{y}_0, 3\ell, \ell, \mu, \delta)$.
sample s uniformly from $\{1, 2, \cdots, T\}$.
$\mathbf{y}_s \leftarrow \mathrm{AGD}(-f(\mathbf{x}_s, \cdot), \mathbf{y}_0, \ell, \mu, \delta)$.
Output: $(\mathbf{x}_s, \mathbf{y}_s)$.

Nonconvex-strongly-concave setting. Our algorithm for nonconvex-strongly-concave optimization is described in Algorithm 9. Similar to Algorithm 8, we still use our accelerated solver Maximin-AG2 for the same proximal subproblem in the inner loop. The only minor difference is that, in the outer loop, Algorithm 9 only uses the Proximal Point Algorithm (PPA) on function $\Phi(\cdot) := \max_{y \in \mathcal{Y}} f(\cdot, \mathbf{y})$ without acceleration (or momentum steps). This is due to fact that gradient descent is already optimal among all first-order algorithm for finding stationary points of smooth nonconvex functions [Carmon et al., 2020]. The standard acceleration technique will not help for smooth nonconvex functions. We presents the theoretical guarantees for Algorithm 9 in the following theorem.

Theorem 3.6.1 *Assume that f is ℓ-smooth and $f(\mathbf{x}, \cdot)$ is $\mu_{\mathbf{y}}$-strongly-concave for all $\mathbf{x}$. Then there exists $T > 0$ such that the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) =$ MINIMAX-PPA$(f, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_{\mathbf{y}}, \epsilon, T)$ is an ϵ-stationary point of f with probability at least $2/3$, and the total number of gradient evaluations is bounded by*

$$O\left(\frac{\ell \Delta_\Phi}{\epsilon^2} \cdot \sqrt{\kappa_{\mathbf{y}}} \log^2\left(\frac{\kappa_{\mathbf{y}} \ell (\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right),$$

where $\kappa_{\mathbf{y}} = \ell / \mu_{\mathbf{y}}$ is the condition number, $\Delta_\Phi = \Phi(\mathbf{x}_0) - \min_{\mathbf{x} \in \mathbb{R}^m} \Phi(\mathbf{x})$ is the initial function value gap and $\tilde{D}_{\mathbf{x}} = \|\mathbf{x}_0 - \mathbf{x}_{g_1}^\star(\mathbf{y}_0)\|$ is the initial distance where $\mathbf{x}_g^\star(\mathbf{y}_0) = \mathrm{argmin}_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}, \mathbf{y}_0)$.

Theorem 3.6.1 claims that Algorithm 9 will find an ϵ-stationary point, with at least constant probability, in $\tilde{O}(\sqrt{\kappa_{\mathbf{y}}}/\epsilon^2)$ gradient evaluations. Similar to Theorem 3.5.1, the inner loop takes $\tilde{O}(\sqrt{\kappa_{\mathbf{y}}})$ gradient evaluations to solve the proximal step since the condition number of $g_t(\cdot, \mathbf{y})$ is $O(1)$ for any $\mathbf{y} \in \mathcal{Y}$. In the outer loop, regardless of the smoothness of $\Phi(\cdot)$, PPA with proximal parameter ℓ is capable of finding the stationary point in $\tilde{O}(1/\epsilon^2)$ iterations. In total, the gradient complexity is $\tilde{O}(\sqrt{\kappa_{\mathbf{y}}}/\epsilon^2)$.

Nonconvex-concave setting. Our result in the nonconvex-strongly-concave setting readily implies a fast result in the nonconvex-concave setting. Consider the following auxiliary

function for an arbitrary $\mathbf{y}_0 \in \mathcal{Y}$:

$$\tilde{f}_\epsilon(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y}) - (\epsilon/4D_{\mathbf{y}})\|\mathbf{y} - \mathbf{y}_0\|^2. \tag{3.9}$$

By construction, it is clear that the gradient of f and $\tilde{f}_\epsilon$ are close in the sense

$$\max_{(\mathbf{x},\mathbf{y})\in\mathbb{R}^m\times\mathcal{Y}} \|\nabla f(\mathbf{x}, \mathbf{y}) - \nabla\tilde{f}_\epsilon(\mathbf{x}, \mathbf{y})\| \le \epsilon/4.$$

This implies that any $(\epsilon/2)$-stationary point of $\tilde{f}_\epsilon$ is also a ϵ-stationary point of f, and thus it is sufficient to solve the problem $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{x}\in\mathcal{Y}} \tilde{f}_\epsilon(\mathbf{x}, \mathbf{y})$. Finally, the function $\tilde{f}_\epsilon(\mathbf{x}, \cdot)$ is always $\epsilon/(2D_{\mathbf{y}})$-strongly-concave, which can be fed into Algorithm 9 to obtain the following result.

Corollary 3.6.2 *Assume that f is ℓ-smooth and $f(\mathbf{x}, \cdot)$ is concave for all $\mathbf{x}$. Then there exists $T > 0$ such that the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \text{Minimax-PPA}(\tilde{f}_\epsilon, \mathbf{x}_0, \mathbf{y}_0, \ell, \epsilon/(2D_{\mathbf{y}}), \epsilon/2, T)$ is an ϵ-stationary point of f with probability at least $2/3$, and the total number of gradient evaluations is bounded by*

$$O\left(\frac{\ell\Delta_\Phi}{\epsilon^2} \cdot \sqrt{\frac{\ell D_{\mathbf{y}}}{\epsilon}} \log^2\left(\frac{\ell(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right),$$

where $D_{\mathbf{y}} > 0$, $\Delta_\Phi = \Phi(\mathbf{x}_0) - \min_{\mathbf{x}\in\mathbb{R}^m} \Phi(\mathbf{x})$ is the initial function value gap and $\tilde{D}_{\mathbf{x}} = \|\mathbf{x}_0 - \mathbf{x}_{g_1}^\star(\mathbf{y}_0)\|$ is the initial distance where $\mathbf{x}_g^\star(\mathbf{y}_0) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \mathbf{y}_0)$.

3.7 Conclusion

We have provided the first set of *near-optimal* algorithms for strongly-convex-(strongly)-concave minimax optimization problems and the state-of-the-art algorithms for nonconvex-(strongly)-concave minimax optimization problems. For the former class of problems, our algorithms match the lower complexity bound for first-order algorithms [Ouyang and Xu, 2021, Ibrahim et al., 2020, Zhang et al., 2022a] up to logarithmic factors. For the latter class of problems, our algorithms achieve the best known upper bound. In the future research, one important direction is to investigate the lower complexity bound of first-order algorithms for nonconvex-(strongly)-concave minimax problems. Despite several striking results on lower complexity bounds for nonconvex smooth problems [Carmon et al., 2020, 2021], this problem remains challenging as solving it requires a new construction of "chain-style" functions and resisting oracles.

3.8 Additional Results for Nonconvex-Concave Optimization

We present our results for nonconvex-concave optimization using the stationary of $\Phi(\cdot) := \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ (Definition 3.8.1 and Definition 3.8.5) as the optimality measure.

Optimality notion based on Moreau envelope. We present another optimality notion based on Moreau envelope for nonconvex-concave setting in which $f(\cdot, \mathbf{y})$ is not necessarily convex for each $\mathbf{y} \in \mathcal{Y}$ but $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathcal{X}$. For simplicity, we let $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{Y}$ be convex and bounded. In general, finding a global saddle point of f is intractable since solving the special case with a singleton $\mathcal{Y}$ globally is already NP-hard [Murty and Kabadi, 1987] as mentioned in the main text.

One approach, inspired by nonconvex optimization, is to equivalently reformulate problem (3.1) as the following nonconvex minimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^m} \left\{ \Phi(\mathbf{x}) := \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \right\}, \tag{3.10}$$

and define an optimality notion for the local surrogate of global optimum of Φ. In robust learning, $\mathbf{x}$ is the classifier while $\mathbf{y}$ is the adversarial noise. Practitioners are often only interested in finding a robust classifier $\mathbf{x}$ instead of an adversarial response $\mathbf{y}$ to each data point. Such a stationary point $\mathbf{x}$ precisely corresponds to a robust classifier that is stationary to the robust classification error.

If $f(\mathbf{x}, \cdot)$ is further assumed to be strongly concave for each $\mathbf{x} \in \mathbb{R}^m$, then Φ is smooth and a standard optimality notion is the stationary point.

Definition 3.8.1 *We call $\hat{\mathbf{x}}$ an ϵ-stationary point of a smooth function Φ if $\|\nabla\Phi(\hat{\mathbf{x}})\| \le \epsilon$. If $\epsilon = 0$, then $\hat{\mathbf{x}}$ is called a stationary point.*

In contrast, when $f(\mathbf{x}, \cdot)$ is merely concave for each $\mathbf{x} \in \mathcal{X}$, Φ is not necessarily smooth and even not differentiable. A weaker sufficient condition for the purpose of our paper is the weak convexity.

Definition 3.8.2 *A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is L-weakly convex if $\Phi(\cdot) + (L/2)\|\cdot\|^2$ is convex.*

First, a function Φ is ℓ-weakly convex if it is ℓ-smooth. Second, the subdifferential of a ℓ-weakly convex function Φ can be uniquely determined by the subdifferential of $\Phi(\cdot)+(\ell/2)\|\cdot\|^2$. This implies that the optimality notion can be defined by a point $\mathbf{x} \in \mathbb{R}^m$ with at least one small subgradient: $\min_{\xi\in\partial\Phi(\mathbf{x})} \|\xi\| \le \epsilon$. Unfortunately, this notion can be restrictive if Φ is nonsmooth. Considering a one-dimensional function $\Phi(\cdot) = |\cdot|$, a point $\mathbf{x}$ must be 0 if it satisfies the optimality notion with $\epsilon \in [0, 1)$. This means that finding a sufficiently accurate solution under such optimality notion is as difficult as solving the minimization exactly. Another popular optimality notion is based on the *Moreau envelope* of Φ when Φ is weakly convex [Davis and Drusvyatskiy, 2019].

Definition 3.8.3 *A function Φ_λ is the Moreau envelope of Φ with $\lambda > 0$ if for $\forall \mathbf{x} \in \mathbb{R}^m$, that $\Phi_\lambda(\mathbf{x}) = \min_{\mathbf{w}\in\mathbb{R}^m} \Phi(\mathbf{w}) + (1/2\lambda)\|\mathbf{w} - \mathbf{x}\|^2$.*

Lemma 3.8.4 (Properties of Moreau envelopes) *If $\Phi(\cdot)$ is ℓ-weakly convex, its Moreau envelope $\Phi_{1/2\ell}(\cdot)$ is 4ℓ-smooth with the gradient $\nabla\Phi_{1/2\ell}(\cdot) = 2\ell(\cdot - \textsc{Prox}_{\Phi/2\ell}(\cdot))$ in which a point $\textsc{Prox}_{\Phi/2\ell}(\cdot) = \operatorname{argmin}_{\mathbf{w}\in\mathbb{R}^m}\{\Phi(\mathbf{w}) + \ell\|\mathbf{w} - \cdot\|^2\}$ is defined.*

Thus, an ϵ-stationary point of an ℓ-weakly convex function Φ can be alternatively defined as a point $\hat{\mathbf{x}}$ satisfying that the gradient norm of Moreau envelope $\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\|$ is small.

Definition 3.8.5 *We call $\hat{\mathbf{x}}$ an ϵ-stationary point of a ℓ-weakly convex function Φ if $\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\| \le \epsilon$. If $\epsilon = 0$, then $\hat{\mathbf{x}}$ is called a stationary point.*

Lemma 3.8.6 (Properties of ϵ-stationary point) *If $\hat{\mathbf{x}}$ is an ϵ-stationary point of a ℓ-weakly convex function Φ, then there exists $\bar{\mathbf{x}} \in \mathbb{R}^m$ such that $\min_{\xi\in\partial\Phi(\bar{\mathbf{x}})}\|\xi\| \le \epsilon$ and $\|\hat{\mathbf{x}} - \bar{\mathbf{x}}\| \le \epsilon/2\ell$.*

Lemma 3.8.6 shows that an ϵ-stationary point defined by the Moreau envelope can be interpreted as the relaxation for a point with at least one small subgradient. In particular, if $\hat{\mathbf{x}}$ is an ϵ-stationary point of a ℓ-weakly convex function Φ, then it is close to a point which has small subgradient.

Nonconvex-strongly-concave setting. In the setting of nonconvex-strongly-concave function, we still use Algorithm 9. Similar to Theorem 3.6.1, we can obtain a guarantee, which finds a point $\hat{\mathbf{x}}$ satisfying $\|\nabla\Phi(\hat{\mathbf{x}})\| \le \epsilon$ in the same number of iterations as in Theorem 3.6.1.

Theorem 3.8.7 *Assume that f is ℓ-smooth and $f(\mathbf{x},\cdot)$ is $\mu_{\mathbf{y}}$-strongly-concave for all $\mathbf{x}$. Then there exists $T > 0$ such that the output $(\hat{\mathbf{x}},\hat{\mathbf{y}}) = \text{MINIMAX-PPA}(f,\mathbf{x}_0,\mathbf{y}_0,\ell,\mu_{\mathbf{y}},\epsilon,T)$ satisfies $\|\nabla\Phi(\hat{\mathbf{x}})\| \le \epsilon$ with probability at least $2/3$, and the total number of gradient evaluations is bounded by*

$$O\left(\frac{\ell\Delta_\Phi}{\epsilon^2}\cdot\sqrt{\kappa_{\mathbf{y}}}\log^2\left(\frac{\kappa_{\mathbf{y}}\ell(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right)$$

where $\kappa_{\mathbf{y}} = \ell/\mu_{\mathbf{y}}$ is the condition number, $\Delta_\Phi = \Phi(\mathbf{x}_0) - \min_{\mathbf{x}\in\mathbb{R}^m}\Phi(\mathbf{x})$ is the initial function value gap and $\tilde{D}_{\mathbf{x}} = \|\mathbf{x}_0 - \mathbf{x}^\star_{g_1}(\mathbf{y}_0)\|$ is the initial distance where $\mathbf{x}^\star_g(\mathbf{y}_0) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\mathbf{y}_0)$.

Nonconvex-concave setting. We can reduce the problem of optimizing a nonconvex-concave function to the problem of optimizing a nonconvex-strongly-concave function. The only caveat is that, in order to achieve the near-optimal point using Definition 3.8.5 as optimality measure, we can only add a $O(\epsilon^2)$ term as follows:

$$\bar{f}_\epsilon(\mathbf{x},\mathbf{y}) = f(\mathbf{x},\mathbf{y}) - (\epsilon^2/200\ell D_{\mathbf{y}}^2)\|\mathbf{y} - \mathbf{y}_0\|^2. \tag{3.11}$$

Now $\bar{f}_\epsilon(\mathbf{x},\cdot)$ is only $\epsilon^2/(100\ell D_{\mathbf{y}}^2)$-concave, by feeding it to Algorithm 9 and through a slightly more complicated reduction argument, we can only obtain gradient complexity bound of $\tilde{O}(\epsilon^{-3})$ instead of $\tilde{O}(\epsilon^{-2.5})$ as in Corollary 3.6.2. Formally, we have

Corollary 3.8.8 *Assume that f is ℓ-smooth, and $f(\mathbf{x},\cdot)$ is concave for all $\mathbf{x}$. Then there exists $T > 0$ such that the output $(\hat{\mathbf{x}},\hat{\mathbf{y}}) = \text{MINIMAX-PPA}(\bar{f}_\epsilon,\mathbf{x}_0,\mathbf{y}_0,\ell,\epsilon^2/(100\ell D_{\mathbf{y}}^2),\epsilon/10,T)$*

satisfies $\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\| \leq \epsilon$ *with probability at least* $2/3$*, and the total number of gradient evaluations is bounded by*

$$\mathcal{O}\left(\frac{\ell^2 D_{\mathbf{y}}\Delta_\Phi}{\epsilon^3}\log^2\left(\frac{\ell(\tilde{D}_{\mathbf{x}}^2+D_{\mathbf{y}}^2)}{\epsilon}\right)\right)$$

where $D_{\mathbf{y}} > 0$*,* $\Delta_\Phi = \Phi(\mathbf{x}_0) - \min_{\mathbf{x}\in\mathbb{R}^m}\Phi(\mathbf{x})$ *is the initial function value gap and* $\tilde{D}_{\mathbf{x}} = \|\mathbf{x}_0 - \mathbf{x}_{g_1}^\star(\mathbf{y}_0)\|$ *is the initial distance where* $\mathbf{x}_g^\star(\mathbf{y}_0) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\mathbf{y}_0)$*.*

3.9 Proofs for Algorithm Components

We present proofs for our algorithm components.

Proof of Theorem 3.3.6. We divide the proof into three parts. In the first part, we show that the output $\hat{\mathbf{x}}$ satisfies $g(\hat{\mathbf{x}}) \leq \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) + \epsilon$. In the second part, we derive the sufficient condition for guaranteeing the stopping criteria in Algorithm 5. In the third part, we derive the gradient complexity using the condition derived in the second part.

Part I. Let $\tilde{\mathbf{x}}_t = \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - (1/\ell)\nabla g(\mathbf{x}_t))$ be defined as the point achieved by one-step projected gradient descent from $\mathbf{x}_t$. Since g is ℓ-smooth and μ-strongly convex, it is straightforward to derive from Nesterov [2018, Corollary 2.3.2] that

$$g(\mathbf{x}) \geq g(\tilde{\mathbf{x}}_t) + \ell(\mathbf{x}_t - \tilde{\mathbf{x}}_t)^\top(\mathbf{x} - \mathbf{x}_t) + \tfrac{\ell}{2}\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2 + \tfrac{\mu}{2}\|\mathbf{x} - \mathbf{x}_t\|^2, \quad \text{for all } \mathbf{x}\in\mathcal{X}.$$

Using the Young's inequality, we have $(\mathbf{x}_t - \tilde{\mathbf{x}}_t)^\top(\mathbf{x} - \mathbf{x}_t) \geq -(1/2)(\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2 + \|\mathbf{x} - \mathbf{x}_t\|^2)$. Putting these pieces together with $\mathbf{x} = \mathbf{x}^\star$ yields that

$$g(\tilde{\mathbf{x}}_t) - \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) = g(\tilde{\mathbf{x}}_t) - g(\mathbf{x}^\star) \leq \left(\tfrac{\ell-\mu}{2}\right)\|\mathbf{x}_t - \mathbf{x}^\star\|^2.$$

Without loss of generality, we let $\ell > \mu$. Indeed, if $\ell = \mu$, then one-step projected gradient descent from any points in $\mathcal{X}$ guarantees that $g(\tilde{\mathbf{x}}_t) - \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) = 0$. Since $\hat{\mathbf{x}} = \tilde{\mathbf{x}}_t$ in Algorithm 5, it suffices to show that the following statement holds true,

$$\|\mathbf{x}_t - \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - (1/\ell)\nabla g(\mathbf{x}_t))\| \leq \sqrt{\tfrac{\epsilon}{2\kappa^2(\ell-\mu)}} \Longrightarrow \|\mathbf{x}_t - \mathbf{x}^\star\| \leq \sqrt{\tfrac{2\epsilon}{\ell-\mu}}. \tag{3.12}$$

Let $\tilde{\mathbf{x}}_t = \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - (1/\ell)\nabla g(\mathbf{x}_t))$ be defined as the point achieved by one-step projected gradient descent from $\mathbf{x}_t$, the ℓ-smoothness of g implies

$$\|\tilde{\mathbf{x}}_t - \mathbf{x}^\star\| \leq \|\mathbf{x}_k - \mathbf{x}^\star\|. \tag{3.13}$$

Using the definition of $\tilde{\mathbf{x}}_t$ and $\mathbf{x}^\star$, we have

$$(\mathbf{x}^\star - \tilde{\mathbf{x}}_t)^\top(\ell(\tilde{\mathbf{x}}_t - \mathbf{x}_t) + \nabla g(\mathbf{x}_t)) \geq 0, \qquad (\tilde{\mathbf{x}}_t - \mathbf{x}^\star)^\top\nabla g(\mathbf{x}^\star) \geq 0.$$

Summing up the above two inequalities and rearranging yields that

$$(\mathbf{x}^\star - \mathbf{x}_t)^\top(\nabla g(\mathbf{x}_t) - \nabla g(\mathbf{x}^\star)) \geq \ell(\mathbf{x}^\star - \tilde{\mathbf{x}}_t)^\top(\mathbf{x}_t - \tilde{\mathbf{x}}_t) + (\tilde{\mathbf{x}}_t - \mathbf{x}_t)^\top(\nabla g(\mathbf{x}_t) - \nabla g(\mathbf{x}^\star)).$$

Since g is ℓ-smooth and μ-strongly convex, we have

$$-\mu\|\mathbf{x}_t - \mathbf{x}^\star\|^2 \geq -\ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|(\|\mathbf{x}^\star - \tilde{\mathbf{x}}_t\| + \|\mathbf{x}^\star - \mathbf{x}_t\|) \overset{(3.13)}{\geq} -2\ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|\|\mathbf{x}_t - \mathbf{x}^\star\|.$$

Therefore, we conclude that

$$\|\mathbf{x}_t - \mathbf{x}^\star\| \leq 2\kappa\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\| = 2\kappa\|\mathbf{x}_t - \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - (1/\ell)\nabla g(\mathbf{x}_t))\| \overset{(3.12)}{\leq} \sqrt{\tfrac{2\epsilon}{\ell-\mu}}.$$

Part II. We first show that

$$\|\mathbf{x}_t - \mathbf{x}^\star\| \leq \tfrac{1}{3\kappa}\sqrt{\tfrac{\epsilon}{2(\ell-\mu)}} \Longrightarrow \|\mathbf{x}_t - \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - (1/\ell)\nabla g(\mathbf{x}_t))\| \leq \sqrt{\tfrac{\epsilon}{2\kappa^2(\ell-\mu)}}.$$

By the definition of $\mathbf{x}^\star$, we have $\mathbf{x}^\star = \mathcal{P}_{\mathcal{X}}(\mathbf{x}^\star - (1/\ell)\nabla g(\mathbf{x}^\star))$. This together with the triangle inequality and the nonexpansiveness of $\mathcal{P}_{\mathcal{X}}$ yields $\|\mathbf{x}_t - \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - (1/\ell)\nabla g(\mathbf{x}_t))\| \leq 3\|\mathbf{x}_t - \mathbf{x}^\star\|$ which implies the desired result. Then we derive a sufficient condition for guaranteeing that $\|\mathbf{x}_t - \mathbf{x}^\star\| \leq (1/(3\kappa))\sqrt{\epsilon/(2(\ell-\mu))}$. Since g is μ-strongly convex and $\mathbf{x}_t \in \mathcal{X}$, Nesterov [2018, Theorem 2.1.5] together with the fact that $(\mathbf{x}_t - \mathbf{x}^\star)^\top\nabla g(\mathbf{x}^\star) \geq 0$ implies that

$$\|\mathbf{x}_t - \mathbf{x}^\star\|^2 \leq \tfrac{2}{\mu}\left(g(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x})\right).$$

Putting these pieces together yields the desired sufficient condition as follows,

$$g(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) \leq \tfrac{\epsilon}{36\kappa^3}. \tag{3.14}$$

Part III. We proceed to derive the gradient complexity of the algorithm using the condition in Eq. (3.14). Since Algorithm 5 is exactly Nesterov's accelerated gradient descent, standard arguments based on estimate sequence [Nesterov, 2018] implies

$$g(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) \leq \left(1 - \tfrac{1}{\sqrt{\kappa}}\right)^t \left(g(\mathbf{x}_0) - \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) + \tfrac{\mu\|\mathbf{x}^\star - \mathbf{x}_0\|^2}{2}\right).$$

Therefore, the gradient complexity of Algorithm 5 to guarantee Eq. (3.14) is bounded by

$$O\left(1 + \sqrt{\kappa}\log\left(\frac{\kappa^3\ell\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\epsilon}\right)\right).$$

This completes the proof.

Proof of Theorem 3.4.1. Letting $\hat{\mathbf{x}} = \textsc{Inexact-APPA}(g, \mathbf{x}_0, \ell, \mu, \epsilon, T)$. Since $\hat{\mathbf{x}} = \mathbf{x}_T$, it suffices for us to estimate an lower bound for the maximum number of iterations T such that $g(\mathbf{x}_T) \leq \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) + \epsilon$. The following technical lemma is crucial to the analysis.

Lemma 3.9.1 *For any $\mathbf{x} \in \mathcal{X}$ and $\{(\mathbf{x}_t, \tilde{\mathbf{x}}_t)\}_{t\geq 0}$ generated by Algorithm 6, we have*

$$g(\mathbf{x}) \geq g(\mathbf{x}_t) - 2\ell(\mathbf{x} - \tilde{\mathbf{x}}_{t-1})^\top(\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) + 2\ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 + \tfrac{\mu\|\mathbf{x}-\mathbf{x}_t\|^2}{4} - 7\kappa\delta. \tag{3.15}$$

Proof. Using the definition of $\mathbf{x}_t$ in Algorithm 6, we have

$$g(\mathbf{x}_t) + \ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 \leq \min_{\mathbf{x}\in\mathcal{X}} \left\{g(\mathbf{x}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2\right\} + \delta.$$

Defining $\mathbf{x}_t^\star = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}}\{g(\mathbf{x}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2\}$ and using μ-strongly convexity of g, we have the following for any $\mathbf{x} \in \mathcal{X}$:

$$g(\mathbf{x}) \geq g(\mathbf{x}_t^\star) + \ell\|\mathbf{x}_t^\star - \tilde{\mathbf{x}}_{t-1}\|^2 - \ell\|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2 + \left(\ell + \tfrac{\mu}{2}\right)\|\mathbf{x} - \mathbf{x}_t^\star\|^2.$$

Equivalently, we have

$$\begin{aligned} g(\mathbf{x}) &\geq g(\mathbf{x}_t) + \ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 - \ell\|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2 + \left(\ell + \tfrac{\mu}{2}\right)\|\mathbf{x} - \mathbf{x}_t^\star\|^2 - \delta \\ &\geq g(\mathbf{x}_t) - 2\ell(\mathbf{x} - \mathbf{x}_t)^\top(\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) - \ell\|\mathbf{x} - \mathbf{x}_t\|^2 + \left(\ell + \tfrac{\mu}{2}\right)\|\mathbf{x} - \mathbf{x}_t^\star\|^2 - \delta. \end{aligned}$$

On the other hand, we have

$$\left(\ell + \tfrac{\mu}{2}\right)\|\mathbf{x} - \mathbf{x}_t^\star\|^2 - \ell\|\mathbf{x} - \mathbf{x}_t\|^2 = \tfrac{\mu\|\mathbf{x}-\mathbf{x}_t\|^2}{2} + (2\ell + \mu)(\mathbf{x} - \mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}_t^\star) + \left(\ell + \tfrac{\mu}{2}\right)\|\mathbf{x}_t - \mathbf{x}_t^\star\|^2$$

Using Young's inequality yields

$$(\mathbf{x} - \mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}_t^\star) \geq -\tfrac{\mu\|\mathbf{x}-\mathbf{x}_t\|^2}{4(2\ell+\mu)} - (1 + 2\kappa)\|\mathbf{x}_t - \mathbf{x}_t^\star\|^2.$$

Putting these pieces together yields that

$$g(\mathbf{x}) \geq g(\mathbf{x}_t) - 2\ell(\mathbf{x} - \mathbf{x}_t)^\top(\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) + \tfrac{\mu\|\mathbf{x}-\mathbf{x}_t\|^2}{4} - (2\ell + \mu)(1 + 2\kappa)\|\mathbf{x}_t - \mathbf{x}_t^\star\|^2 - \delta.$$

Furthermore, we have

$$(\mathbf{x} - \mathbf{x}_t)^\top(\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) = (\mathbf{x} - \tilde{\mathbf{x}}_{t-1})^\top(\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) - \|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2,$$

and

$$\|\mathbf{x}_t - \mathbf{x}_t^\star\|^2 \leq \tfrac{2}{\mu+2\ell}\left(g(\mathbf{x}_t) + \ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 - \min_{\mathbf{x}\in\mathcal{X}}\left\{g(\mathbf{x}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2\right\}\right) \leq \tfrac{2\delta}{\mu+2\ell}.$$

Putting these pieces together with $\kappa \geq 1$ yields the desired inequality. □

The remaining proof is based on Lemma 3.9.1. Indeed, we have

$$\begin{aligned}
& \left(1-\tfrac{1}{2\sqrt{\kappa}}\right) g(\mathbf{x}_{t-1}) + \tfrac{1}{2\sqrt{\kappa}} \left(g(\mathbf{x}^\star) + 14\kappa^{3/2}\delta\right) \\
\overset{\text{Eq. (3.15)}}{\geq} & \left(1-\tfrac{1}{2\sqrt{\kappa}}\right) \left(g(\mathbf{x}_t) - 2\ell(\mathbf{x}_{t-1}-\tilde{\mathbf{x}}_{t-1})^\top(\mathbf{x}_t-\tilde{\mathbf{x}}_{t-1}) + 2\ell\|\mathbf{x}_t-\tilde{\mathbf{x}}_{t-1}\|^2 + \tfrac{\mu\|\mathbf{x}_{t-1}-\mathbf{x}_t\|^2}{4} - 7\kappa\delta\right) \\
& + \tfrac{1}{2\sqrt{\kappa}} \left(g(\mathbf{x}_t) - 2\ell(\mathbf{x}^\star-\tilde{\mathbf{x}}_{t-1})^\top(\mathbf{x}_t-\tilde{\mathbf{x}}_{t-1}) + 2\ell\|\mathbf{x}_t-\tilde{\mathbf{x}}_{t-1}\|^2 + \tfrac{\mu\|\mathbf{x}^\star-\mathbf{x}_t\|^2}{4} - 7\kappa\delta\right) + 7\kappa\delta \\
= & \quad g(\mathbf{x}_t) - 2\ell\left(\left(1-\tfrac{1}{2\sqrt{\kappa}}\right)\mathbf{x}_{t-1} + \tfrac{\mathbf{x}^\star}{2\sqrt{\kappa}} - \tilde{\mathbf{x}}_{t-1}\right)^\top(\mathbf{x}_t-\tilde{\mathbf{x}}_{t-1}) + 2\ell\|\mathbf{x}_t-\tilde{\mathbf{x}}_{t-1}\|^2 + \tfrac{\mu\|\mathbf{x}^\star-\mathbf{x}_t\|^2}{8\sqrt{\kappa}}.
\end{aligned}$$

Equivalently, we have

$$\begin{aligned}
g(\mathbf{x}_t) - g(\mathbf{x}^\star) \leq & \left(1-\tfrac{1}{2\sqrt{\kappa}}\right)(g(\mathbf{x}_{t-1}) - g(\mathbf{x}^\star)) + 2\ell\left(\left(1-\tfrac{1}{2\sqrt{\kappa}}\right)\mathbf{x}_{t-1} + \tfrac{\mathbf{x}^\star}{2\sqrt{\kappa}} - \tilde{\mathbf{x}}_{t-1}\right)^\top(\mathbf{x}_t-\tilde{\mathbf{x}}_{t-1}) \\
& -2\ell\|\mathbf{x}_t-\tilde{\mathbf{x}}_{t-1}\|^2 - \tfrac{\mu\|\mathbf{x}^\star-\mathbf{x}_t\|^2}{8\sqrt{\kappa}} + 7\kappa\delta. \qquad (3.16)
\end{aligned}$$

Consider $\tilde{\mathbf{x}}_t = \mathbf{x}_t + \frac{2\sqrt{\kappa}-1}{2\sqrt{\kappa}+1}(\mathbf{x}_t - \mathbf{x}_{t-1})$, we let $\mathbf{w}_t = \tilde{\mathbf{x}}_t + 2\sqrt{\kappa}(\tilde{\mathbf{x}}_t - \mathbf{x}_t)$ and obtain that

$$\begin{aligned}
\mathbf{w}_t &= (1+2\sqrt{\kappa})\tilde{\mathbf{x}}_t - 2\sqrt{\kappa}\mathbf{x}_t = 2\sqrt{\kappa}\mathbf{x}_t - (2\sqrt{\kappa}-1)\mathbf{x}_{t-1} = \left(1-\tfrac{1}{2\sqrt{\kappa}}\right)\mathbf{w}_{t-1} + 2\sqrt{\kappa}\mathbf{x}_t - \tfrac{4\kappa-1}{2\sqrt{\kappa}}\tilde{\mathbf{x}}_{t-1} \\
&= \left(1-\tfrac{1}{2\sqrt{\kappa}}\right)\mathbf{w}_{t-1} + 2\sqrt{\kappa}\,(\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) + \tfrac{\tilde{\mathbf{x}}_{t-1}}{2\sqrt{\kappa}}.
\end{aligned}$$

This implies that

$$\begin{aligned}
\|\mathbf{w}_t - \mathbf{x}^\star\|^2 &= \left\|\left(1-\tfrac{1}{2\sqrt{\kappa}}\right)\mathbf{w}_{t-1} + \tfrac{\tilde{\mathbf{x}}_{t-1}}{2\sqrt{\kappa}} - \mathbf{x}^\star + 2\sqrt{\kappa}\,(\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1})\right\|^2 \qquad (3.17) \\
&= \left\|\left(1-\tfrac{1}{2\sqrt{\kappa}}\right)\mathbf{w}_{t-1} + \tfrac{\tilde{\mathbf{x}}_{t-1}}{2\sqrt{\kappa}} - \mathbf{x}^\star\right\|^2 + 4\sqrt{\kappa}\left(\left(1-\tfrac{1}{2\sqrt{\kappa}}\right)\mathbf{w}_{t-1} + \tfrac{\tilde{\mathbf{x}}_{t-1}}{2\sqrt{\kappa}} - \mathbf{x}^\star\right)^\top(\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) \\
&\quad + 4\kappa\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2.
\end{aligned}$$

Since $\mathbf{w}_{t-1} = \tilde{\mathbf{x}}_{t-1} + 2\sqrt{\kappa}(\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1})$, we have

$$\left(1-\tfrac{1}{2\sqrt{\kappa}}\right)\mathbf{w}_{t-1} + \tfrac{\tilde{\mathbf{x}}_{t-1}}{2\sqrt{\kappa}} = 2\sqrt{\kappa}\tilde{\mathbf{x}}_{t-1} - (2\sqrt{\kappa}-1)\mathbf{x}_{t-1}. \qquad (3.18)$$

Using the Young's inequality, we have

$$\begin{aligned}
& \left\|\left(1-\tfrac{1}{2\sqrt{\kappa}}\right)\mathbf{w}_{t-1} + \tfrac{\tilde{\mathbf{x}}_{t-1}}{2\sqrt{\kappa}} - \mathbf{x}^\star\right\|^2 \qquad (3.19) \\
\leq & \left(1-\tfrac{1}{2\sqrt{\kappa}}\right)^2\left(1+\tfrac{5}{8\sqrt{\kappa}-5}\right)\|\mathbf{w}_{t-1} - \mathbf{x}^\star\|^2 + \tfrac{1}{4\kappa}\left(1+\tfrac{8\sqrt{\kappa}-5}{5}\right)\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}^\star\|^2 \\
\leq & \left(1-\tfrac{1}{2\sqrt{\kappa}}\right)\left(1+\tfrac{1}{8\sqrt{\kappa}-5}\right)\|\mathbf{w}_{t-1} - \mathbf{x}^\star\|^2 + \tfrac{2\|\tilde{\mathbf{x}}_{t-1}-\mathbf{x}^\star\|^2}{5\sqrt{\kappa}} \\
\leq & \left(1-\tfrac{1}{6\sqrt{\kappa}}\right)\|\mathbf{w}_{t-1} - \mathbf{x}^\star\|^2 + \tfrac{2\|\tilde{\mathbf{x}}_{t-1}-\mathbf{x}^\star\|^2}{5\sqrt{\kappa}}.
\end{aligned}$$

Using the Young's inequality again, we have

$$\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}^\star\|^2 \leq \tfrac{5\|\mathbf{x}^\star - \mathbf{x}_t\|^2}{4} + 5\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_t\|^2. \tag{3.20}$$

Putting Eq. (3.17)-Eq. (3.20) together with $\kappa \geq 1$, we have

$$\begin{aligned}\|\mathbf{w}_t - \mathbf{x}^\star\|^2 \leq \left(1 - \tfrac{1}{6\sqrt{\kappa}}\right)\|\mathbf{w}_{t-1} - \mathbf{x}^\star\|^2 + \tfrac{\|\mathbf{x}^\star - \mathbf{x}_t\|^2}{2\sqrt{\kappa}} \\ +6\kappa\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 + 8\kappa\left(\tilde{\mathbf{x}}_{t-1} - \left(1 - \tfrac{1}{2\sqrt{\kappa}}\right)\mathbf{x}_{t-1} - \tfrac{\mathbf{x}^\star}{2\sqrt{\kappa}}\right)^\top (\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}).\end{aligned} \tag{3.21}$$

Combining Eq. (3.16) and Eq. (3.21) yields that

$$\begin{aligned} g(\mathbf{x}_t) - g(\mathbf{x}^\star) + \tfrac{\mu\|\mathbf{w}_t - \mathbf{x}^\star\|^2}{4} &\leq \left(1 - \tfrac{1}{2\sqrt{\kappa}}\right)(g(\mathbf{x}_{t-1}) - g(\mathbf{x}^\star)) + \left(1 - \tfrac{1}{6\sqrt{\kappa}}\right)\tfrac{\mu\|\mathbf{w}_{t-1} - \mathbf{x}^\star\|^2}{4} + 7\kappa\delta \\ &\leq \left(1 - \tfrac{1}{6\sqrt{\kappa}}\right)\left(g(\mathbf{x}_{t-1}) - g(\mathbf{x}^\star) + \tfrac{\mu\|\mathbf{w}_{t-1} - \mathbf{x}^\star\|^2}{4}\right) + 7\kappa\delta. \end{aligned}$$

Repeating the above inequality yields that

$$g(\mathbf{x}_T) - g(\mathbf{x}^\star) + \tfrac{\mu\|\mathbf{w}_T - \mathbf{x}^\star\|^2}{4} \leq \left(1 - \tfrac{1}{6\sqrt{\kappa}}\right)^\top \left(g(\mathbf{x}_0) - g(\mathbf{x}^\star) + \tfrac{\mu\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{4}\right) + 42\kappa^{3/2}\delta.$$

Therefore, we conclude that

$$g(\mathbf{x}_T) - g(\mathbf{x}^\star) \leq \left(1 - \tfrac{1}{6\sqrt{\kappa}}\right)^\top \left(g(\mathbf{x}_0) - g(\mathbf{x}^\star) + \tfrac{\mu\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{4}\right) + 42\kappa^{3/2}\delta.$$

Since the tolerance $\delta \leq \epsilon\kappa^{-3/2}/84$, we conclude that the iteration complexity of Algorithm 6 to guarantee that $g(\mathbf{x}_T) - \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) \leq \epsilon$ if there exists an absolute constant $c > 0$ such that

$$T \geq c\sqrt{\kappa}\log\left(\frac{g(\mathbf{x}_0) - g(\mathbf{x}^\star) + (\mu/4)\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\epsilon}\right).$$

This completes the proof.

Proof of Theorem 3.4.2. Before presenting the main proof, we define the following important functions:

$$\begin{aligned}\Phi_g(\cdot) &= \max_{\mathbf{y}\in\mathcal{Y}} g(\cdot, \mathbf{y}), & \mathbf{y}_g^\star(\cdot) &= \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} g(\cdot, \mathbf{y}), \\ \Psi_g(\cdot) &= \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \cdot), & \mathbf{x}_g^\star(\cdot) &= \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \cdot).\end{aligned}$$

All the above functions are well defined since $g(\cdot, \cdot)$ is strongly convex-concave. We provide their complete characterization in the following structural lemma.

Lemma 3.9.2 *Under the assumptions imposed in Theorem 3.4.2, we have*

(a) A function $\mathbf{y}_g^\star(\cdot)$ *is* $\kappa_{\mathbf{y}}$*-Lipschitz.*

(b) A function $\Phi_g(\cdot)$ *is* $2\kappa_{\mathbf{y}}\ell$*-smooth and* $\mu_{\mathbf{x}}$*-strongly convex with* $\nabla\Phi_g(\cdot) = \nabla_{\mathbf{x}} g(\cdot, \mathbf{y}_g^\star(\cdot))$.

(c) A function $\mathbf{x}_g^\star(\cdot)$ *is* $\kappa_{\mathbf{x}}$*-Lipschitz.*

(d) A function $\Psi_g(\cdot)$ *is* $2\kappa_{\mathbf{x}}\ell$*-smooth and* $\mu_{\mathbf{y}}$*-strongly concave with* $\nabla\Psi_g(\cdot) = \nabla_{\mathbf{y}} g(\mathbf{x}_g^\star(\cdot), \cdot)$.

where $\kappa_{\mathbf{x}} = \ell/\mu_{\mathbf{x}}$ *and* $\kappa_{\mathbf{y}} = \ell/\mu_{\mathbf{y}}$ *are condition numbers.*

Now we are ready to prove Theorem 3.4.2. We divide the proof into three parts. In the first part, we show that the output $\hat{\mathbf{x}} = \textsc{Maximin-AG2}(g, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_{\mathbf{x}}, \mu_{\mathbf{y}}, \epsilon)$ satisfies

$$\max_{\mathbf{y}\in\mathcal{Y}} \; g(\hat{\mathbf{x}}, \mathbf{y}) \;\leq\; \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \; g(\mathbf{x}, \mathbf{y}) + \epsilon \tag{3.22}$$

In the second part, we get the sufficient condition for guaranteeing the stopping criteria in Algorithm 7. In the third part, we estimate an upper bound for the gradient complexity of the algorithm using the condition derived in the second part. For the ease of presentation, we denote $(\mathbf{x}_g^\star, \mathbf{y}_g^\star)$ as the unique solution to the minimax optimization $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} g(\mathbf{x}, \mathbf{y})$.

Part I. By the definition of Φ_g, the inequality in Eq. (3.22) can be rewritten as follows,

$$\Phi_g(\hat{\mathbf{x}}) \leq \min_{\mathbf{x}\in\mathcal{X}} \; \Phi_g(\mathbf{x}) + \epsilon.$$

Since $\hat{\mathbf{x}} = \mathcal{P}_{\mathcal{X}}(\mathbf{x}_T - (1/2\kappa_{\mathbf{y}}\ell)\nabla_{\mathbf{x}} g(\mathbf{x}_T, \mathbf{y}_T))$, we have

$$\begin{aligned} 0 &\leq (\mathbf{x} - \hat{\mathbf{x}})^\top \left(2\kappa_{\mathbf{y}}\ell(\hat{\mathbf{x}} - \mathbf{x}_T) + \nabla_{\mathbf{x}} g(\mathbf{x}_T, \mathbf{y}_T)\right) \\ &= \; (\mathbf{x} - \hat{\mathbf{x}})^\top (2\kappa_{\mathbf{y}}\ell(\hat{\mathbf{x}} - \mathbf{x}_T) + \nabla\Phi_g(\mathbf{x}_T)) + (\mathbf{x} - \hat{\mathbf{x}})^\top (\nabla_{\mathbf{x}} g(\mathbf{x}_T, \mathbf{y}_T) - \nabla\Phi_g(\mathbf{x}_T)). \end{aligned}$$

Since $\nabla\Phi_g(\mathbf{x}_T) = \nabla_{\mathbf{x}} g(\mathbf{x}_T, \mathbf{y}_g^\star(\mathbf{x}_T))$, we have $\|\nabla_{\mathbf{x}} g(\mathbf{x}_T, \mathbf{y}_T) - \nabla\Phi_g(\mathbf{x}_T)\| \leq \ell\|\mathbf{y}_T - \mathbf{y}_g^\star(\mathbf{x}_T)\|$. Using the Young's inequality, we have

$$(\mathbf{x} - \hat{\mathbf{x}})^\top (\nabla_{\mathbf{x}} g(\mathbf{x}_T, \mathbf{y}_T) - \nabla\Phi_g(\mathbf{x}_T)) \leq \tfrac{\kappa_{\mathbf{y}}\ell\|\hat{\mathbf{x}} - \mathbf{x}_T\|^2}{2} + \tfrac{\kappa_{\mathbf{y}}\ell\|\mathbf{x} - \mathbf{x}_T\|^2}{2} + \mu_{\mathbf{y}}\|\mathbf{y}_T - \mathbf{y}_g^\star(\mathbf{x}_T)\|^2.$$

Since Φ_g is $2\kappa_{\mathbf{y}}\ell$-smooth and $\mu_{\mathbf{x}}$-strongly convex, we have

$$\begin{aligned} (\mathbf{x} - \hat{\mathbf{x}})^\top (2\kappa_{\mathbf{y}}\ell(\hat{\mathbf{x}} - \mathbf{x}_T) + \nabla\Phi_g(\mathbf{x}_T)) &\leq 2\kappa_{\mathbf{y}}\ell(\mathbf{x} - \mathbf{x}_T)^\top(\hat{\mathbf{x}} - \mathbf{x}_T) \\ +\Phi_g(\mathbf{x}) - \Phi_g(\hat{\mathbf{x}}) - \kappa_{\mathbf{y}}\ell\|\hat{\mathbf{x}} - \mathbf{x}_T\|^2 &- \tfrac{\mu_{\mathbf{x}}\|\mathbf{x} - \mathbf{x}_T\|^2}{2}. \end{aligned}$$

Using the Young's inequality, we have $(\mathbf{x} - \mathbf{x}_T)^\top(\hat{\mathbf{x}} - \mathbf{x}_T) \leq \|\mathbf{x} - \mathbf{x}_T\|^2 + (1/4)\|\hat{\mathbf{x}} - \mathbf{x}_T\|^2$. Putting these pieces together yields with $\mathbf{x} = \mathbf{x}_g^\star$ yields that

$$\Phi_g(\hat{\mathbf{x}}) - \min_{\mathbf{x}\in\mathcal{X}} \Phi_g(\mathbf{x}) \leq 3\kappa_{\mathbf{y}}\ell\|\mathbf{x}_T - \mathbf{x}_g^\star\|^2 + \mu_{\mathbf{y}}\|\mathbf{y}_T - \mathbf{y}_g^\star(\mathbf{x}_T)\|^2. \tag{3.23}$$

In what follows, we prove that $\Phi_g(\hat{\mathbf{x}}) \leq \min_{\mathbf{x}\in\mathcal{X}} \Phi_g(\mathbf{x}) + \epsilon$ if the following stopping conditions hold true,

$$g(\mathbf{x}_T, \mathbf{y}_T) - g(\mathbf{x}_g^\star(\mathbf{y}_T), \mathbf{y}_T) \leq \tfrac{\epsilon}{648\kappa_{\mathbf{x}}^3\kappa_{\mathbf{y}}^3}, \tag{3.24}$$

$$\|\mathbf{y}_T - \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla_{\mathbf{y}} g(\mathbf{x}_T, \mathbf{y}_T))\| \leq \tfrac{1}{24\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}}}\sqrt{\tfrac{\epsilon}{\kappa_{\mathbf{y}}\ell}}. \tag{3.25}$$

Indeed, we observe that $\|\mathbf{x}_T - \mathbf{x}_g^\star\| \leq \|\mathbf{x}_T - \mathbf{x}_g^\star(\mathbf{y}_T)\| + \|\mathbf{x}_g^\star(\mathbf{y}_T) - \mathbf{x}_g^\star(\mathbf{y}_g^\star)\| + \|\mathbf{x}_g^\star(\mathbf{y}_g^\star) - \mathbf{x}_g^\star\|$. By definition, we have $\mathbf{x}_g^\star(\mathbf{y}_g^\star) = \mathbf{x}_g^\star$. Also, $\mathbf{x}_g^\star(\cdot)$ is $\kappa_{\mathbf{x}}$-Lipschitz. Therefore, we have

$$\|\mathbf{x}_T - \mathbf{x}_g^\star\| \leq \|\mathbf{x}_T - \mathbf{x}_g^\star(\mathbf{y}_T)\| + \kappa_{\mathbf{x}}\|\mathbf{y}_T - \mathbf{y}_g^\star\|. \tag{3.26}$$

By the similar argument, we have

$$\|\mathbf{y}_T - \mathbf{y}_g^\star(\mathbf{x}_T)\| \leq \|\mathbf{y}_T - \mathbf{y}_g^\star\| + \kappa_{\mathbf{y}}\|\mathbf{x}_T - \mathbf{x}_g^\star\| \leq \kappa_{\mathbf{y}}\|\mathbf{x}_T - \mathbf{x}_g^\star(\mathbf{y}_T)\| + \kappa_{\mathbf{x}}\kappa_{\mathbf{y}}\|\mathbf{y}_T - \mathbf{y}_g^\star\|. \tag{3.27}$$

First, we bound the term $\|\mathbf{x}_T - \mathbf{x}_g^\star(\mathbf{y}_T)\|$. Since $g(\cdot, \mathbf{y}_T)$ is $\mu_{\mathbf{x}}$-strongly convex, we have

$$\|\mathbf{x}_T - \mathbf{x}_g^\star(\mathbf{y}_T)\| \leq \sqrt{\tfrac{2(g(\mathbf{x}_T,\mathbf{y}_T) - g(\mathbf{x}^\star(\mathbf{y}_T),\mathbf{y}_T))}{\mu_{\mathbf{x}}}} \leq \tfrac{1}{18\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}\sqrt{\tfrac{\epsilon}{\kappa_{\mathbf{y}}\ell}} \tag{3.28}$$

It remains to bound the term $\|\mathbf{y}_T - \mathbf{y}_g^\star\|$. Indeed, we have $\nabla\Psi_g(\mathbf{y}_T) = \nabla_{\mathbf{y}} g(\mathbf{x}_g^\star(\mathbf{y}_T), \mathbf{y}_T)$ and

$$\begin{aligned}\|\mathbf{y}_T - \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla\Psi_g(\mathbf{y}_T))\| &\leq \|\mathbf{y}_T - \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla_{\mathbf{y}} g(\mathbf{x}_T, \mathbf{y}_T))\| \\ &+\|\mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla_{\mathbf{y}} g(\mathbf{x}_T, \mathbf{y}_T)) - \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla\Psi_g(\mathbf{y}_T))\|.\end{aligned}$$

Since $\mathcal{P}_{\mathcal{Y}}$ is nonexpansive and $\nabla_{\mathbf{y}} g$ is ℓ-Lipschitz, we have

$$\|\mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla_{\mathbf{y}} g(\mathbf{x}_T, \mathbf{y}_T)) - \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla\Psi_g(\mathbf{y}_T))\| \leq \tfrac{\|\mathbf{x}_T - \mathbf{x}_g^\star(\mathbf{y}_T)\|}{2\kappa_{\mathbf{x}}}.$$

Putting these pieces together with Eq. (3.25) and Eq. (3.28) yields that

$$\|\mathbf{y}_T - \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla\Psi_g(\mathbf{y}_T))\| \leq \tfrac{1}{18\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}}}\sqrt{\tfrac{\epsilon}{\kappa_{\mathbf{y}}\ell}}. \tag{3.29}$$

Since $\mathbf{y}_g^\star = \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} \Psi_g(\mathbf{y})$ and $\tilde{\mathbf{y}}_T = \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla\Psi_g(\mathbf{y}_T))$ is achieved by one-step projected gradient ascent from $\mathbf{y}_T$, we derive from the $2\kappa_{\mathbf{x}}\ell$-smoothness of Ψ_g, we have

$$\|\tilde{\mathbf{y}}_T - \mathbf{y}_g^\star\| \leq \|\mathbf{y}_T - \mathbf{y}_g^\star\|. \tag{3.30}$$

Using the definition of $\tilde{\mathbf{y}}_T$ and $\mathbf{y}_g^\star$, we have

$$(\mathbf{y}_g^\star - \tilde{\mathbf{y}}_T)^\top(\tilde{\mathbf{y}}_T - \mathbf{y}_T - (1/2\kappa_{\mathbf{x}}\ell)\nabla\Psi_g(\mathbf{y}_T)) \geq 0, \qquad (\mathbf{y}_g^\star - \tilde{\mathbf{y}}_T)^\top\nabla\Psi_g(\mathbf{y}_g^\star) \geq 0.$$

Summing up the above two inequalities and rearranging yields that

$$(\mathbf{y}_g^\star - \mathbf{y}_T)^\top(\nabla\Psi_g(\mathbf{y}_g^\star) - \nabla\Psi_g(\mathbf{y}_T)) \geq 2\kappa_{\mathbf{x}}\ell(\mathbf{y}_g^\star - \tilde{\mathbf{y}}_T)^\top(\mathbf{y}_T - \tilde{\mathbf{y}}_T) + (\tilde{\mathbf{y}}_T - \mathbf{y}_T)^\top(\nabla\Psi_g(\mathbf{y}_g^\star) - \nabla\Psi_g(\mathbf{y}_T)).$$

Since Ψ_g is $2\kappa_{\mathbf{x}}\ell$-smooth and $\mu_{\mathbf{y}}$-strongly concave, we have

$$-\mu_{\mathbf{y}}\|\mathbf{y}_g^\star-\mathbf{y}_T\|^2 \geq -2\kappa_{\mathbf{x}}\ell\|\tilde{\mathbf{y}}_T-\mathbf{y}_T\|(\|\mathbf{y}_g^\star-\tilde{\mathbf{y}}_T\|+\|\mathbf{y}_g^\star-\mathbf{y}_T\|) \overset{(3.30)}{\geq} -4\kappa_{\mathbf{x}}\ell\|\tilde{\mathbf{y}}_T-\mathbf{y}_T\|\|\mathbf{y}_g^\star-\mathbf{y}_T\|.$$

This implies that

$$\|\mathbf{y}_g^\star-\mathbf{y}_T\| \leq 4\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}\|\mathbf{y}_T-\tilde{\mathbf{y}}_T\| \overset{(3.29)}{\leq} \tfrac{1}{4\kappa_{\mathbf{x}}}\sqrt{\tfrac{\epsilon}{\kappa_{\mathbf{y}}\ell}}. \tag{3.31}$$

Plugging Eq. (3.28) and Eq. (3.31) into Eq. (3.26) yields that

$$\|\mathbf{x}_T-\mathbf{x}_g^\star\| \leq \left(\tfrac{1}{18\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}+\tfrac{1}{4}\right)\sqrt{\tfrac{\epsilon}{\kappa_{\mathbf{y}}\ell}} \overset{\kappa_{\mathbf{x}},\kappa_{\mathbf{y}}\geq 1}{\leq} \tfrac{1}{2}\sqrt{\tfrac{\epsilon}{\kappa_{\mathbf{y}}\ell}}.$$

Plugging Eq. (3.28) and Eq. (3.31) into Eq. (3.27) yields that

$$\|\mathbf{y}_T-\mathbf{y}_g^\star(\mathbf{x}_T)\| \leq \left(\tfrac{1}{18\kappa_{\mathbf{x}}}+\tfrac{\kappa_{\mathbf{y}}}{4}\right)\sqrt{\tfrac{\epsilon}{\kappa_{\mathbf{y}}\ell}} \overset{\kappa_{\mathbf{x}},\kappa_{\mathbf{y}}\geq 1}{\leq} \tfrac{1}{2}\sqrt{\tfrac{\kappa_{\mathbf{y}}\epsilon}{\ell}}.$$

Putting these pieces together Eq. (3.23) yields the desired result.

Part II. We first show that $\|\mathbf{y}_T-\mathbf{y}_g^\star\| \leq (1/216\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}})\sqrt{\epsilon/\kappa_{\mathbf{y}}\ell}$ and Eq. (3.24) are sufficient to guarantee Eq. (3.25). Indeed, we have $\mathbf{y}_g^\star = \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_g^\star + (1/2\kappa_{\mathbf{x}}\ell)\nabla\Psi_g(\mathbf{y}_g^\star))$. This together with the triangle inequality and the nonexpansiveness of $\mathcal{P}_{\mathcal{Y}}$ yields

$$\|\mathbf{y}_T-\mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T+(1/2\kappa_{\mathbf{x}}\ell)\nabla_{\mathbf{y}}g(\mathbf{x}_T,\mathbf{y}_T))\| \leq 2\|\mathbf{y}_T-\mathbf{y}_g^\star\| + \tfrac{\|\nabla_{\mathbf{y}}g(\mathbf{x}_T,\mathbf{y}_T)-\nabla\Psi_g(\mathbf{y}_g^\star)\|}{2\kappa_{\mathbf{x}}\ell}.$$

Furthermore, $\nabla\Psi_g(\mathbf{y}_T) = \nabla_{\mathbf{y}}g(\mathbf{x}^\star(\mathbf{y}_T),\mathbf{y}_T)$ and

$$\|\nabla_{\mathbf{y}}g(\mathbf{x}_T,\mathbf{y}_T)-\nabla\Psi_g(\mathbf{y}_g^\star)\| \leq \|\nabla_{\mathbf{y}}g(\mathbf{x}_T,\mathbf{y}_T)-\nabla_{\mathbf{y}}g(\mathbf{x}_g^\star(\mathbf{y}_T),\mathbf{y}_T)\| + \|\nabla\Psi_g(\mathbf{y}_T)-\nabla\Psi_g(\mathbf{y}_g^\star)\|.$$

Since g is ℓ-smooth and Ψ_g is $2\kappa_{\mathbf{x}}\ell$-smooth, we have

$$\|\nabla_{\mathbf{y}}g(\mathbf{x}_T,\mathbf{y}_T)-\nabla\Psi_g(\mathbf{y}_g^\star)\| \leq \ell\|\mathbf{x}_T-\mathbf{x}_g^\star(\mathbf{y}_T)\| + 2\kappa_{\mathbf{x}}\ell\|\mathbf{y}_T-\mathbf{y}_g^\star\|.$$

Also, Eq. (3.24) guarantees that Eq. (3.28) holds true. Then we have

$$\|\mathbf{y}_T-\mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T+(1/2\kappa_{\mathbf{x}}\ell)\nabla_{\mathbf{y}}g(\mathbf{x}_T,\mathbf{y}_T))\| \leq 3\|\mathbf{y}_T-\mathbf{y}_g^\star\| + \tfrac{1}{36\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}}}\sqrt{\tfrac{\epsilon}{\kappa_{\mathbf{y}}\ell}}.$$

The above inequality together with $\|\mathbf{y}_T-\mathbf{y}_g^\star\| \leq (1/216\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}})\sqrt{\epsilon/\kappa_{\mathbf{y}}\ell}$ guarantees Eq. (3.25). Next we derive a sufficient condition for guaranteeing $\|\mathbf{y}_T-\mathbf{y}_g^\star\| \leq (1/216\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}})\sqrt{\epsilon/\kappa_{\mathbf{y}}\ell}$. Since Ψ_g is $\mu_{\mathbf{y}}$-strongly concave, Nesterov [2018, Theorem 2.1.5] implies that

$$\|\mathbf{y}_T-\mathbf{y}_g^\star\|^2 \leq \tfrac{2}{\mu_{\mathbf{y}}}\left(\max_{\mathbf{y}\in\mathcal{Y}}\Psi_g(\mathbf{y})-\Psi_g(\mathbf{y}_T)\right).$$

Putting these pieces together yields the desired condition as follows,

$$\max_{\mathbf{y}\in\mathcal{Y}}\Psi_g(\mathbf{y})-\Psi_g(\mathbf{y}_T) \leq \tfrac{\epsilon}{93312\kappa_{\mathbf{x}}^4\kappa_{\mathbf{y}}^4}. \tag{3.32}$$

Part III. We proceed to estimate an upper bound for the gradient complexity of Algorithm 7 using Eq. (3.32). Note that $\tilde{\epsilon} \leq \epsilon/(4477676(\kappa_{\mathbf{x}}\kappa_{\mathbf{y}})^{11/2})$ and we provide a key technical lemma which is crucial to the subsequent analysis.

Lemma 3.9.3 *For any $\mathbf{y} \in \mathcal{Y}$ and $\{(\mathbf{y}_t, \tilde{\mathbf{y}}_t)\}_{t\geq 0}$ generated by Algorithm 7, we have*

$$\Psi_g(\mathbf{y}) \leq 2\kappa_{\mathbf{x}}\ell(\mathbf{y} - \tilde{\mathbf{y}}_{t-1})^\top(\mathbf{y}_t - \tilde{\mathbf{y}}_{t-1}) + \Psi_g(\mathbf{y}_t) - \tfrac{\kappa_{\mathbf{x}}\ell\|\mathbf{y}_t - \tilde{\mathbf{y}}_{t-1}\|^2}{2} - \tfrac{\mu_{\mathbf{y}}\|\mathbf{y} - \tilde{\mathbf{y}}_{t-1}\|^2}{4} + 3\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}\tilde{\epsilon}.$$

Proof. For any $\mathbf{y} \in \mathcal{Y}$, the update formula $\mathbf{y}_t \leftarrow \mathcal{P}_{\mathcal{Y}}(\tilde{\mathbf{y}}_{t-1} + (1/2\kappa_{\mathbf{x}}\ell)\nabla_{\mathbf{y}} g(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1}))$ implies that

$$\begin{aligned} 0 &\leq (\mathbf{y} - \mathbf{y}_t)^\top(2\kappa_{\mathbf{x}}\ell(\mathbf{y}_t - \tilde{\mathbf{y}}_{t-1}) - \nabla_{\mathbf{y}} g(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1})) \\ &= (\mathbf{y} - \mathbf{y}_t)^\top(2\kappa_{\mathbf{x}}\ell(\mathbf{y}_t - \tilde{\mathbf{y}}_{t-1}) - \nabla\Psi_g(\tilde{\mathbf{y}}_{t-1})) + (\mathbf{y} - \mathbf{y}_t)^\top(\nabla\Psi_g(\tilde{\mathbf{y}}_{t-1}) - \nabla_{\mathbf{y}} g(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1})). \end{aligned}$$

Since $\nabla\Psi_g(\tilde{\mathbf{y}}_{t-1}) = \nabla_{\mathbf{y}} g(\mathbf{x}_g^\star(\tilde{\mathbf{y}}_{t-1}), \tilde{\mathbf{y}}_{t-1})$, we have

$$\|\nabla\Psi_g(\tilde{\mathbf{y}}_{t-1}) - \nabla_{\mathbf{y}} g(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1})\| \leq \ell\|\mathbf{x}_g^\star(\tilde{\mathbf{y}}_{t-1}) - \tilde{\mathbf{x}}_{t-1}\|.$$

Since $g(\cdot, \tilde{\mathbf{y}}_{t-1})$ is $\mu_{\mathbf{x}}$-strongly convex, we have

$$\|\mathbf{x}_g^\star(\tilde{\mathbf{y}}_{t-1}) - \tilde{\mathbf{x}}_{t-1}\| \leq \sqrt{\tfrac{2(g(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1}) - g(\mathbf{x}_g^\star(\tilde{\mathbf{y}}_{t-1}), \tilde{\mathbf{y}}_{t-1}))}{\mu_{\mathbf{x}}}} \leq \sqrt{\tfrac{2\tilde{\epsilon}}{\mu_{\mathbf{x}}}}.$$

Using Young's inequality, we have

$$(\mathbf{y} - \mathbf{y}_t)^\top(\nabla\Psi_g(\tilde{\mathbf{y}}_{t-1}) - \nabla_{\mathbf{y}} g(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1})) \leq \tfrac{\kappa_{\mathbf{x}}\ell\|\mathbf{y}_t - \tilde{\mathbf{y}}_{t-1}\|^2}{2} + \tfrac{\mu_{\mathbf{y}}\|\mathbf{y} - \tilde{\mathbf{y}}_{t-1}\|^2}{4} + 3\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}\tilde{\epsilon}.$$

Since Ψ_g is $2\kappa_{\mathbf{x}}\ell$-smooth and $\mu_{\mathbf{y}}$-strongly concave, we have

$$\begin{aligned} (\mathbf{y} - \mathbf{y}_t)^\top(2\kappa_{\mathbf{x}}\ell(\mathbf{y}_t - \tilde{\mathbf{y}}_{t-1}) - \nabla\Psi_g(\tilde{\mathbf{y}}_{t-1})) &\leq 2\kappa_{\mathbf{x}}\ell(\mathbf{y} - \tilde{\mathbf{y}}_{t-1})^\top(\mathbf{y}_t - \tilde{\mathbf{y}}_{t-1}) \\ +\Psi_g(\mathbf{y}_t) - \Psi_g(\mathbf{y}) - \kappa_{\mathbf{x}}\ell\|\mathbf{y}_t - \tilde{\mathbf{y}}_{t-1}\|^2 - \tfrac{\mu_{\mathbf{y}}\|\mathbf{y} - \tilde{\mathbf{y}}_{t-1}\|^2}{2}. \end{aligned}$$

Putting these pieces together yields the desired inequality. □

The remaining proof is based on the modification of Nesterov's techniques [Nesterov, 2018, Section 2.2.5]. Indeed, we define the estimate sequence as follows,

$$\begin{aligned} \Gamma_0(\mathbf{y}) &= \Psi_g(\mathbf{y}_0) - \tfrac{\mu_{\mathbf{y}}\|\mathbf{y} - \mathbf{y}_0\|^2}{2}, \\ \Gamma_{t+1}(\mathbf{y}) &= \tfrac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}}\Big(\Psi_g(\mathbf{y}_{t+1}) + 2\kappa_{\mathbf{x}}\ell(\mathbf{y} - \tilde{\mathbf{y}}_t)^\top(\mathbf{y}_{t+1} - \tilde{\mathbf{y}}_t) - \tfrac{\kappa_{\mathbf{x}}\ell\|\mathbf{y}_{t+1} - \tilde{\mathbf{y}}_t\|^2}{2} \\ &\quad - \tfrac{\mu_{\mathbf{y}}\|\mathbf{y} - \tilde{\mathbf{y}}_t\|^2}{4} - 12(\kappa_{\mathbf{x}}\kappa_{\mathbf{y}})^{3/2}\tilde{\epsilon}\Big) + \Big(1 - \tfrac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}}\Big)\Gamma_t(\mathbf{y}) \quad \text{for all } t \geq 0. \end{aligned}$$

We apply the inductive argument to prove,

$$\max_{\mathbf{y}\in\mathbb{R}^n} \Gamma_t(\mathbf{y}) \leq \Psi_g(\mathbf{y}_t), \quad \text{for all } t \geq 0. \tag{3.33}$$

Eq. (3.33) holds trivially when $t = 0$. Then, we show that Eq. (3.33) holds true when $t = T$ if Eq. (3.33) holds true for all $t \leq T-1$. Let $\mathbf{v}_t = \operatorname{argmax}_{\mathbf{y}\in\mathbb{R}^n} \Gamma_t(\mathbf{y})$ and $\Gamma_t^\star = \max_{\mathbf{y}\in\mathbb{R}^n} \Gamma_t(\mathbf{y})$, we have the canonical form $\Gamma_t(\mathbf{y}) = \Gamma_t^\star - (\mu_\mathbf{y}/4)\|\mathbf{y}-\mathbf{v}_t\|^2$. The following recursive rules hold for $\mathbf{v}_t$ and $\Gamma_t^\star$:

$$\begin{aligned}
\mathbf{v}_{t+1} &= \left(1-\tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)\mathbf{v}_t + \tfrac{\tilde{\mathbf{y}}_t}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}} + \sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}(\mathbf{y}_{t+1}-\tilde{\mathbf{y}}_t),\\
\Gamma_{t+1}^\star &= \left(1-\tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)\Gamma_t^\star + \tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\left(\Psi_g(\mathbf{y}_{t+1}) - 12(\kappa_\mathbf{x}\kappa_\mathbf{y})^{3/2}\tilde{\epsilon}\right) - \left(\tfrac{\ell}{8}\sqrt{\tfrac{\kappa_\mathbf{x}}{\kappa_\mathbf{y}}} - \tfrac{\kappa_\mathbf{x}\ell}{4}\right)\|\mathbf{y}_{t+1}-\tilde{\mathbf{y}}_t\|^2\\
&\quad -\tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\left(1-\tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)\left(\tfrac{\mu_\mathbf{y}\|\tilde{\mathbf{y}}_t-\mathbf{v}_t\|^2}{4} - 2\kappa_\mathbf{x}\ell(\mathbf{v}_t-\tilde{\mathbf{y}}_t)^\top(\mathbf{y}_{t+1}-\tilde{\mathbf{y}}_t)\right).
\end{aligned}$$

It follows from the recursive rule for Γ_t and its canonical form that

$$\nabla\Gamma_{t+1}(\mathbf{y}) = -\left(1-\tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)\tfrac{\mu_\mathbf{y}(\mathbf{y}-\mathbf{v}_t)}{2} + \tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\left(2\kappa_\mathbf{x}\ell(\mathbf{y}_{t+1}-\tilde{\mathbf{y}}_t) - \tfrac{\mu_\mathbf{y}(\mathbf{y}-\tilde{\mathbf{y}}_t)}{2}\right).$$

The recursive rule for $\mathbf{v}_t$ can be achieved by solving $\nabla\Gamma_{t+1}(\mathbf{v}_{t+1}) = 0$. Then we have

$$\begin{aligned}
&\Gamma_{t+1}^\star = \Gamma_{t+1}(\mathbf{v}_{t+1})\\
&= \left(1-\tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)\Gamma_t^\star - \left(1-\tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)\tfrac{\mu_\mathbf{y}\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2}{4} + \tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\left(\Psi_g(\mathbf{y}_{t+1}) - 12(\kappa_\mathbf{x}\kappa_\mathbf{y})^{3/2}\tilde{\epsilon}\right.\\
&\quad \left.-\tfrac{\kappa_\mathbf{x}\ell\|\mathbf{y}_{t+1}-\tilde{\mathbf{y}}_t\|^2}{2}\right) + \tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\left(2\kappa_\mathbf{x}\ell(\mathbf{v}_{t+1}-\tilde{\mathbf{y}}_t)^\top(\mathbf{y}_{t+1}-\tilde{\mathbf{y}}_t) - \tfrac{\mu_\mathbf{y}\|\mathbf{v}_{t+1}-\tilde{\mathbf{y}}_t\|^2}{4}\right).
\end{aligned}$$

Then, we conclude the recursive rule for $\Gamma_t^\star$ by plugging the recursive rule for $\mathbf{v}_k$ into the above equality. By the induction, Eq. (3.33) holds true when $t = T-1$ which implies

$$\begin{aligned}
\Gamma_T^\star &\leq \left(1-\tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)\Psi_g(\mathbf{y}_{T-1}) + \tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\left(\Psi_g(\mathbf{y}_T) - 12(\kappa_\mathbf{x}\kappa_\mathbf{y})^{3/2}\tilde{\epsilon}\right)\\
&\quad -\left(\tfrac{\ell}{8}\sqrt{\tfrac{\kappa_\mathbf{x}}{\kappa_\mathbf{y}}} - \tfrac{\kappa_\mathbf{x}\ell}{4}\right)\|\mathbf{y}_T-\tilde{\mathbf{y}}_{T-1}\|^2 - \tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\left(1-\tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)\left(\tfrac{\mu_\mathbf{y}\|\tilde{\mathbf{y}}_{T-1}-\mathbf{v}_{T-1}\|^2}{2}\right.\\
&\quad \left.-2\kappa_\mathbf{x}\ell(\mathbf{v}_{T-1}-\tilde{\mathbf{y}}_{T-1})^\top(\mathbf{y}_T-\tilde{\mathbf{y}}_{T-1})\right).
\end{aligned}$$

Applying Lemma 3.9.3 with $t = T$ and $\mathbf{y} = \mathbf{y}_{T-1}$ further implies that

$$\Psi_g(\mathbf{y}_{T-1}) \leq 2\kappa_\mathbf{x}\ell(\mathbf{y}_{T-1}-\tilde{\mathbf{y}}_{T-1})^\top(\mathbf{y}_T-\tilde{\mathbf{y}}_{T-1}) + \Psi(\mathbf{y}_T) - \tfrac{\kappa_\mathbf{x}\ell\|\mathbf{y}_T-\tilde{\mathbf{y}}_{T-1}\|^2}{2} - \tfrac{\mu_\mathbf{y}\|\mathbf{y}_{T-1}-\tilde{\mathbf{y}}_{T-1}\|^2}{2} + 3\kappa_\mathbf{x}\kappa_\mathbf{y}\tilde{\epsilon}.$$

Putting these pieces together yields that

$$\Gamma_T^\star \leq \Psi_g(\mathbf{y}_T) + \left(1-\tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)2\kappa_\mathbf{x}\ell(\mathbf{y}_T-\tilde{\mathbf{y}}_{T-1})^\top\left((\mathbf{y}_{T-1}-\tilde{\mathbf{y}}_{T-1}) + \tfrac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}(\mathbf{v}_{T-1}-\tilde{\mathbf{y}}_{T-1})\right).$$

Using the update formula $\tilde{\mathbf{y}}_t = \mathbf{y}_t + \frac{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}-1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}+1}(\mathbf{y}_t-\mathbf{y}_{t-1})$ and the recursive rule for $\mathbf{v}_t$ with the inductive argument, it is straightforward that $(\mathbf{y}_t-\tilde{\mathbf{y}}_t) + \frac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}(\mathbf{v}_t-\tilde{\mathbf{y}}_t) = 0$ for all $t \geq 0$. This implies that $\Gamma_T^\star \leq \Psi_g(\mathbf{y}_T)$. Therefore, we conclude that Eq. (3.33) holds for all $t \geq 0$.

On the other hand, Lemma 3.9.3 and the update formula for Γ_t implies that

$$\Gamma_{t+1}(\mathbf{y}) \geq \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}}\left(\Psi_g(\mathbf{y}) - 12(\kappa_{\mathbf{x}}\kappa_{\mathbf{y}})^{3/2}\tilde{\epsilon} - 3\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}\tilde{\epsilon}\right) + \left(1 - \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}}\right)\Gamma_t(\mathbf{y}).$$

Since $\kappa_{\mathbf{x}}, \kappa_{\mathbf{y}} \geq 1$, we have

$$\Psi_g(\mathbf{y}) - \Gamma_{t+1}(\mathbf{y}) \leq \left(1 - \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}}\right)(\Psi_g(\mathbf{y}) - \Gamma_t(\mathbf{y})) + 6\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}\tilde{\epsilon}.$$

Repeating the above inequality yields that

$$\Psi_g(\mathbf{y}) - \Gamma_T(\mathbf{y}) \leq \left(1 - \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}}\right)^{\top}(\Psi_g(\mathbf{y}) - \Gamma_0(\mathbf{y})) + 24(\kappa_{\mathbf{x}}\kappa_{\mathbf{y}})^{3/2}\tilde{\epsilon}.$$

Therefore, we conclude that

$$\max_{\mathbf{y}\in\mathcal{Y}} \Psi_g(\mathbf{y}) - \Psi_g(\mathbf{y}_T) \leq \left(1 - \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}}\right)^{\top}\frac{(2\kappa_{\mathbf{x}}\ell+\bar{\mu})D_{\mathbf{y}}^2}{2} + 24(\kappa_{\mathbf{x}}\kappa_{\mathbf{y}})^{3/2}\tilde{\epsilon}.$$

Since the tolerance $\tilde{\epsilon} \leq \frac{\epsilon}{4477676(\kappa_{\mathbf{x}}\kappa_{\mathbf{y}})^{11/2}}$, we conclude that the iteration complexity Algorithm 7 to guarantee Eq. (3.32) is bounded by $O(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}\log(\ell D_{\mathbf{y}}^2/\epsilon))$.

Now it suffices to establish the gradient complexity of the two AGD subroutines at each iteration. In particular, we use the gradient complexity of the AGD subroutine to guarantee that $g(\hat{\mathbf{x}}) \leq \min_{\mathcal{X}} g(\mathbf{x}) + \epsilon$ is bounded by

$$O\left(1 + \sqrt{\kappa}\log\left(\frac{\kappa^3\ell\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\epsilon}\right)\right),$$

where κ is the condition number of g and $\mathbf{x}^\star$ is the global optimum of g over $\mathcal{X}$. Since $\mathcal{Y}$ is a convex and bounded set, $\{\mathbf{y}_t\}_{t\geq 0}$ is a bounded sequence. Hence $\{\tilde{\mathbf{y}}_t\}_{t\geq 0}$ is also a bounded sequence. Since $\mathbf{x}_g^\star(\cdot)$ is $\kappa_{\mathbf{x}}$-Lipschitz (cf. Lemma 3.9.2), the sequences $\{\mathbf{x}_g^\star(\tilde{\mathbf{y}}_t)\}_{t\geq 0}$ and $\{\mathbf{x}_g^\star(\mathbf{y}_t)\}_{t\geq 0}$ are bounded. Thus, we have

$$\|\mathbf{x}_0 - \mathbf{x}_g^\star(\mathbf{y}_t)\|^2 = \|\mathbf{x}_0 - \mathbf{x}_g^\star(\tilde{\mathbf{y}}_t)\|^2 = O(\|\mathbf{x}_0 - \mathbf{x}_g^\star(\mathbf{y}_0)\|^2 + \kappa_{\mathbf{x}}^2 D_{\mathbf{y}}^2).$$

Putting these pieces together yields that the gradient complexity of every AGD subroutines at each iteration is bounded by $O(\sqrt{\kappa_{\mathbf{x}}}\log((\kappa_{\mathbf{x}}^3\ell(\|\mathbf{x}_0 - \mathbf{x}_g^\star(\mathbf{y}_0)\|^2 + \kappa_{\mathbf{x}}^2 D_{\mathbf{y}}^2)/\tilde{\epsilon}))$. Therefore, the gradient complexity of Algorithm 7 to guarantee Eq. (3.32) is bounded by

$$O\left(\kappa_{\mathbf{x}}\sqrt{\kappa_{\mathbf{y}}}\cdot\log^2\left(\frac{(\kappa_{\mathbf{x}}+\kappa_{\mathbf{y}})\ell(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right),$$

where $\kappa_{\mathbf{x}} = \ell/\mu_{\mathbf{x}}$ and $\kappa_{\mathbf{y}} = \ell/\mu_{\mathbf{y}}$ are condition numbers, $\tilde{D}_{\mathbf{x}} = \|\mathbf{x}_0 - \mathbf{x}_g^\star(\mathbf{y}_0)\|$ is the initial distance where $\mathbf{x}_g^\star(\mathbf{y}_0) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \mathbf{y}_0)$ and $D_{\mathbf{y}} > 0$ is the diameter of the set $\mathcal{Y}$.

3.10 Proofs for Convex-Concave Settings

Proof of Theorem 3.5.1. We first show that there exists $T > 0$ such that $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \textsc{Minimax-APPA}(f, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_{\mathbf{x}}, \mu_{\mathbf{y}}, \epsilon, T)$ is an ϵ-saddle point. Then we estimate the total number of gradient evaluations required to output an ϵ-approximate saddle point.

First, we note that Minimax-APPA in Algorithm 8 can be interpreted as an inexact accelerated proximal point algorithm Inexact-APPA with the inner loop solver Maximin-AG2 and AGD. Using Theorem 3.3.6 and Theorem 3.4.1, the point $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ satisfies

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \le \left(1 - \tfrac{1}{6\sqrt{\kappa_{\mathbf{x}}}}\right)^{\top} \left(\Phi(\mathbf{x}_0) - \Phi(\mathbf{x}^\star) + \tfrac{\mu_{\mathbf{x}}\|\mathbf{x}^\star - \mathbf{x}_0\|^2}{4}\right) + 42\kappa_{\mathbf{x}}^{3/2}\delta.$$

and $\hat{\mathbf{y}} \leftarrow \mathcal{P}_{\mathcal{Y}}\left(\tilde{\mathbf{y}} + (1/2\kappa_{\mathbf{x}}\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \tilde{\mathbf{y}})\right)$ where $\tilde{\mathbf{y}} \in \mathcal{Y}$ satisfies that

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - f(\hat{\mathbf{x}}, \tilde{\mathbf{y}}) \le \tilde{\epsilon}.$$

We let $\Phi(\cdot) = \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ and note that Φ is $\mu_{\mathbf{x}}$-strongly convex. Since f is $\mu_{\mathbf{x}}$-strongly-convex-$\mu_{\mathbf{y}}$-strongly-concave, the Nash equilibrium $(\mathbf{x}^\star, \mathbf{y}^\star)$ is unique and $\mathbf{x}^\star = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x})$. Therefore, we have

$$\|\hat{\mathbf{x}} - \mathbf{x}^\star\|^2 \le \tfrac{2}{\mu_{\mathbf{x}}}\left(\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})\right).$$

Since $f(\hat{\mathbf{x}}, \cdot)$ is $\mu_{\mathbf{y}}$-strongly concave, Nesterov [2018, Theorem 2.1.5] implies that

$$\|\tilde{\mathbf{y}} - \mathbf{y}^\star(\hat{\mathbf{x}})\|^2 \le \tfrac{2}{\mu_{\mathbf{y}}}\left(\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - f(\hat{\mathbf{x}}, \tilde{\mathbf{y}})\right) \le \tfrac{2\tilde{\epsilon}}{\mu_{\mathbf{y}}}.$$

Since $\mathbf{y}^\star(\cdot) = \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ is $\kappa_{\mathbf{y}}$-Lipschitz (cf. Lemma 3.9.2), we have $\|\mathbf{y}^\star - \mathbf{y}^\star(\hat{\mathbf{x}})\|^2 = \|\mathbf{y}^\star(\mathbf{x}^\star) - \mathbf{y}^\star(\hat{\mathbf{x}})\|^2 \le \kappa_{\mathbf{y}}^2\|\hat{\mathbf{x}} - \mathbf{x}^\star\|^2$. Thus, we have

$$\|\tilde{\mathbf{y}} - \mathbf{y}^\star\|^2 \le 2\kappa_{\mathbf{y}}^2\|\hat{\mathbf{x}} - \mathbf{x}^\star\|^2 + \tfrac{4\tilde{\epsilon}}{\mu_{\mathbf{y}}}.$$

Let $\Psi(\cdot) = \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \cdot)$. By the definition of $\hat{\mathbf{y}}$, the following inequality holds for $\forall \mathbf{y} \in \mathcal{Y}$,

$$\begin{aligned} 0 &\le (\mathbf{y} - \hat{\mathbf{y}})^\top (2\kappa_{\mathbf{x}}\ell(\hat{\mathbf{y}} - \tilde{\mathbf{y}}) - \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \tilde{\mathbf{y}})) \\ &= (\mathbf{y} - \hat{\mathbf{y}})^\top (2\kappa_{\mathbf{x}}\ell(\hat{\mathbf{y}} - \tilde{\mathbf{y}}) - \nabla\Psi(\tilde{\mathbf{y}})) + (\mathbf{y} - \hat{\mathbf{y}})^\top (\nabla\Psi(\tilde{\mathbf{y}}) - \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \tilde{\mathbf{y}})). \end{aligned}$$

Since $\nabla\Psi(\tilde{\mathbf{y}}) = \nabla_{\mathbf{y}} f(\mathbf{x}^\star(\tilde{\mathbf{y}}), \tilde{\mathbf{y}})$, we have $\|\nabla\Psi(\tilde{\mathbf{y}}) - \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \tilde{\mathbf{y}})\| \le \ell\|\mathbf{x}^\star(\tilde{\mathbf{y}}) - \hat{\mathbf{x}}\|$. Using the Young's inequality, we have

$$(\mathbf{y} - \hat{\mathbf{y}})^\top (\nabla\Psi(\tilde{\mathbf{y}}) - \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \tilde{\mathbf{y}})) \le \tfrac{\kappa_{\mathbf{x}}\ell\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2}{2} + \tfrac{\kappa_{\mathbf{x}}\ell\|\mathbf{y} - \tilde{\mathbf{y}}\|^2}{2} + \mu_{\mathbf{x}}\|\mathbf{x}^\star(\tilde{\mathbf{y}}) - \hat{\mathbf{x}}\|^2.$$

Since Ψ is $\mu_{\mathbf{y}}$-strongly concave and $2\kappa_{\mathbf{x}}\ell$-smooth, we have

$$(\mathbf{y} - \hat{\mathbf{y}})^\top (2\kappa_{\mathbf{x}}\ell(\hat{\mathbf{y}} - \tilde{\mathbf{y}}) - \nabla\Psi(\tilde{\mathbf{y}})) \le 2\kappa_{\mathbf{x}}\ell(\mathbf{y} - \tilde{\mathbf{y}})^\top (\hat{\mathbf{y}} - \tilde{\mathbf{y}}) + \Psi(\hat{\mathbf{y}}) - \Psi(\mathbf{y}) - \kappa_{\mathbf{x}}\ell\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2 - \tfrac{\mu_{\mathbf{x}}\|\mathbf{y} - \tilde{\mathbf{y}}\|^2}{2}.$$

Using the Young's inequality, we have $(\mathbf{y}-\tilde{\mathbf{y}})^\top(\hat{\mathbf{y}}-\tilde{\mathbf{y}}) \le \|\mathbf{y}-\tilde{\mathbf{y}}\|^2 + (1/4)\|\hat{\mathbf{y}}-\tilde{\mathbf{y}}\|^2$. Putting these pieces together with $\mathbf{y}=\mathbf{y}^\star$ yields that

$$\begin{aligned}
&\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\hat{\mathbf{y}}) = \Psi(\mathbf{y}^\star) - \Psi(\hat{\mathbf{y}}) \\
&\le\ 3\kappa_{\mathbf{x}}\ell\|\tilde{\mathbf{y}}-\mathbf{y}^\star\|^2 + \mu_{\mathbf{x}}\|\mathbf{x}^\star(\tilde{\mathbf{y}})-\hat{\mathbf{x}}\|^2 \\
&\le\ 3\kappa_{\mathbf{x}}\ell\|\tilde{\mathbf{y}}-\mathbf{y}^\star\|^2 + 2\mu_{\mathbf{x}}\|\mathbf{x}^\star(\tilde{\mathbf{y}})-\mathbf{x}^\star(\mathbf{y}^\star)\|^2 + 2\mu_{\mathbf{x}}\|\mathbf{x}^\star-\hat{\mathbf{x}}\|^2 \\
&\le\ 5\kappa_{\mathbf{x}}\ell\|\tilde{\mathbf{y}}-\mathbf{y}^\star\|^2 + 2\mu_{\mathbf{x}}\|\mathbf{x}^\star-\hat{\mathbf{x}}\|^2.
\end{aligned}$$

Therefore, we conclude that

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\hat{\mathbf{y}}) \le 20\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}\tilde{\epsilon} + (20\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}}^2+5)\left(\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y})\right).$$

Note that $\tilde{\epsilon} \le \epsilon/(80\kappa_{\mathbf{x}}\kappa_{\mathbf{y}})$ and $\delta \le \epsilon/(4200\kappa_{\mathbf{x}}^{7/2}\kappa_{\mathbf{y}}^2)$. This together with the above inequality implies that

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\hat{\mathbf{y}}) \le \tfrac{3\epsilon}{4} + (20\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}}^2+5)\left(1-\tfrac{1}{6\sqrt{\kappa_{\mathbf{x}}}}\right)^\top\left(\Phi(\mathbf{x}_0)-\Phi(\mathbf{x}^\star)+\tfrac{\mu_{\mathbf{x}}\|\mathbf{x}^\star-\mathbf{x}_0\|^2}{4}\right).$$

Thus, there exists an absolute constant $c>0$ such that $\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\hat{\mathbf{y}}) \le \epsilon$ if the maximum number of iterations $T \ge c\sqrt{\kappa_{\mathbf{x}}}\log(\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}}^2\ell\|\mathbf{x}^\star-\mathbf{x}_0\|^2/\epsilon)$. This implies that the total number of iterations is bounded by

$$O\left(\sqrt{\kappa_{\mathbf{x}}}\log\left(\frac{\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}}^2\ell\|\mathbf{x}^\star-\mathbf{x}_0\|^2}{\epsilon}\right)\right).$$

Furthermore, we call the solver MAXIMIN-AG2 at each iteration. Using Theorem 3.4.2 and $\delta = \epsilon/(10\kappa_{\mathbf{x}}\kappa_{\mathbf{y}})^4$, the number of gradient evaluations at each iteration is bounded by

$$O\left(\sqrt{\kappa_{\mathbf{y}}}\log\left(\frac{\kappa_{\mathbf{x}}^{7/2}\kappa_{\mathbf{y}}^3\ell(\tilde{D}_{\mathbf{x}}^2+D_{\mathbf{y}}^2)}{\epsilon}\right)\log\left(\frac{\kappa_{\mathbf{x}}^4\kappa_{\mathbf{y}}^4\ell D_{\mathbf{y}}^2}{\epsilon}\right)\right).$$

Recalling $D=\max\{D_{\mathbf{x}},D_{\mathbf{y}}\}<+\infty$, we conclude that the total number of gradient evaluations is bounded by

$$O\left(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}\log^3\left(\frac{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}\ell D^2}{\epsilon}\right)\right).$$

This completes the proof.

Proof of Corollary 3.5.2. $(\hat{\mathbf{x}},\hat{\mathbf{y}}) = \text{MINIMAX-APPA}(f_{\epsilon,\mathbf{y}},\mathbf{x}_0,\mathbf{y}_0,\ell,\mu_{\mathbf{x}},\epsilon/(4D_{\mathbf{y}}^2),\epsilon/2,T)$ can shown to be an ϵ-saddle point. Then we estimate the number of gradient evaluations to output an ϵ-saddle point using Theorem 3.5.1. By the definition of f_ϵ, the output $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ satisfies

$$\max_{\mathbf{y}\in\mathcal{Y}}\left\{f(\hat{\mathbf{x}},\mathbf{y}) - \tfrac{\epsilon\|\mathbf{y}-\mathbf{y}_0\|^2}{4D_{\mathbf{y}}^2}\right\} - \min_{\mathbf{x}\in\mathcal{X}}\left\{f(\mathbf{x},\hat{\mathbf{y}}) - \tfrac{\epsilon\|\hat{\mathbf{y}}-\mathbf{y}_0\|^2}{4D_{\mathbf{y}}^2}\right\} \le \tfrac{\epsilon}{2}.$$

Since the function $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathcal{X}$, we have

$$\max_{\mathbf{y}\in\mathcal{Y}} \left\{ f(\hat{\mathbf{x}}, \mathbf{y}) - \tfrac{\epsilon\|\mathbf{y}-\mathbf{y}_0\|^2}{4D_{\mathbf{y}}^2} \right\} \geq \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}_{T+1}, \mathbf{y}) - \tfrac{\epsilon}{4}.$$

On the other hand, we have

$$\min_{\mathbf{x}\in\mathcal{X}} \left\{ f(\mathbf{x}, \hat{\mathbf{y}}) - \tfrac{\epsilon\|\hat{\mathbf{y}}-\mathbf{y}_0\|^2}{4D_{\mathbf{y}}^2} \right\} \leq \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}) + \tfrac{\epsilon}{4}.$$

Putting these pieces together yields that $\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}) \leq \epsilon$.

Letting $\kappa_{\mathbf{y}} = 2\ell D_{\mathbf{y}}^2/\epsilon$ in the gradient complexity bound presented in Theorem 3.5.1, we conclude that the total number of gradient evaluations is bounded by

$$O\left(\sqrt{\frac{\kappa_{\mathbf{x}}\ell}{\epsilon}} D_{\mathbf{y}} \log^3\left(\frac{\kappa_{\mathbf{x}}\ell D^2}{\epsilon} \right) \right).$$

This completes the proof.

Proof of Corollary 3.5.3. $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \textsc{Minimax-APPA}(f_\epsilon, \mathbf{x}_0, \mathbf{y}_0, \ell, \epsilon/(4D_{\mathbf{x}}^2), \epsilon/(4D_{\mathbf{y}}^2), \epsilon/2, T)$ can shown to be an ϵ-saddle point. Then we estimate the number of gradient evaluations to output an ϵ-saddle point using Theorem 3.5.1. By the definition of f_ϵ, the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ satisfies

$$\max_{\mathbf{y}\in\mathcal{Y}} \left\{ f(\hat{\mathbf{x}}, \mathbf{y}) + \tfrac{\epsilon\|\hat{\mathbf{x}}-\mathbf{x}_0\|^2}{8D_{\mathbf{x}}^2} - \tfrac{\epsilon\|\mathbf{y}-\mathbf{y}_0\|^2}{8D_{\mathbf{y}}^2} \right\} - \min_{\mathbf{x}\in\mathcal{X}} \left\{ f(\mathbf{x}, \hat{\mathbf{y}}) + \tfrac{\epsilon\|\mathbf{x}-\mathbf{x}_0\|^2}{8D_{\mathbf{x}}^2} - \tfrac{\epsilon\|\hat{\mathbf{y}}-\mathbf{y}_0\|^2}{8D_{\mathbf{y}}^2} \right\} \leq \tfrac{\epsilon}{2}.$$

Since the function $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathcal{X}$, we have

$$\max_{\mathbf{y}\in\mathcal{Y}} \left\{ f(\hat{\mathbf{x}}, \mathbf{y}) + \tfrac{\epsilon\|\hat{\mathbf{x}}-\mathbf{x}_0\|^2}{8D_{\mathbf{x}}^2} - \tfrac{\epsilon\|\mathbf{y}-\mathbf{y}_0\|^2}{8D_{\mathbf{y}}^2} \right\} \geq \max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \tfrac{\epsilon}{4}.$$

On the other hand,

$$\min_{\mathbf{x}\in\mathcal{X}} \left\{ f(\mathbf{x}, \hat{\mathbf{y}}) + \tfrac{\epsilon\|\mathbf{x}-\mathbf{x}_0\|^2}{8D_{\mathbf{x}}^2} - \tfrac{\epsilon\|\hat{\mathbf{y}}-\mathbf{y}_0\|^2}{8D_{\mathbf{y}}^2} \right\} \leq \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}) + \tfrac{\epsilon}{4}.$$

Putting these pieces together yields that $\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}) \leq \epsilon$.

Furthermore, letting $\kappa_{\mathbf{x}} = 4\ell D_{\mathbf{x}}^2/\epsilon$ and $\kappa_{\mathbf{y}} = 2\ell D_{\mathbf{y}}^2/\epsilon$ in the gradient complexity bound presented in Theorem 3.5.1, we conclude that the total number of gradient evaluations is bounded by

$$O\left(\frac{\ell D_{\mathbf{x}} D_{\mathbf{y}}}{\epsilon} \log^3\left(\frac{\ell D^2}{\epsilon} \right) \right).$$

This completes the proof.

3.11 Proofs for Nonconvex-Concave Settings

Proof of Theorem 3.6.1. Using the definition of g_t, we have

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}_{t+1},\mathbf{y}) + \ell\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 \le \min_{\mathbf{x}\in\mathcal{X}}\left\{\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}) + \ell\|\mathbf{x}-\mathbf{x}_t\|^2\right\} + \delta.$$

This implies that

$$\Phi(\mathbf{x}_{t+1}) + \ell\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 \le \min_{\mathbf{x}\in\mathcal{X}}\left\{\Phi(\mathbf{x}) + \ell\|\mathbf{x}-\mathbf{x}_t\|^2\right\} + \delta \le \Phi(\mathbf{x}_t) + \delta.$$

Equivalently, we have

$$\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 \le \tfrac{\Phi(\mathbf{x}_t)-\Phi(\mathbf{x}_{t+1})+\delta}{\ell}. \tag{3.34}$$

Note that the function $\Phi(\cdot)+\ell\|\cdot-\mathbf{x}_t\|^2$ is ℓ-strongly convex and its minimizer $\mathbf{x}_t^\star$ is well defined and unique [Davis and Drusvyatskiy, 2019]. Since the function $\Phi(\cdot)+\ell\|\cdot-\mathbf{x}_t\|^2$ is ℓ-strongly convex, we derive from Nesterov [2018, Theorem 2.1.5] that

$$\|\mathbf{x}_{t+1}-\mathbf{x}_t^\star\|^2 \le \tfrac{2}{\ell}\left(\Phi(\mathbf{x}_{t+1}) + \ell\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 - \min_{\mathbf{x}\in\mathbb{R}^m}\left\{\Phi(\mathbf{x}) + \ell\|\mathbf{x}-\mathbf{x}_t\|^2\right\}\right) \le \tfrac{2\delta}{\ell}. \tag{3.35}$$

Since Φ is differentiable, we have

$$\mathbf{x}_t^\star = \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}_t^\star - \tfrac{\nabla\Phi(\mathbf{x}_t^\star)+2\ell(\mathbf{x}_t^\star-\mathbf{x}_t)}{\ell}\right).$$

Therefore, we have

$$\left\|\mathbf{x}_{t+1} - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}_{t+1} - \tfrac{\nabla\Phi(\mathbf{x}_{t+1})}{\ell}\right)\right\| \le 2\|\mathbf{x}_{t+1}-\mathbf{x}_t^\star\| + 2\|\mathbf{x}_t-\mathbf{x}_t^\star\| + \tfrac{\|\nabla\Phi(\mathbf{x}_t^\star)-\nabla\Phi(\mathbf{x}_{t+1})\|}{\ell}.$$

Since $\Phi(\cdot)$ is $2\kappa_{\mathbf{y}}\ell$-smooth, we have $\|\nabla\Phi(\mathbf{x}_{t+1})-\nabla\Phi(\mathbf{x}_t^\star)\| \le 2\kappa_{\mathbf{y}}\ell\|\mathbf{x}_{t+1}-\mathbf{x}_t^\star\|$. Putting these pieces together yields that

$$\begin{aligned}\left\|\mathbf{x}_{t+1} - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}_{t+1} - \tfrac{\nabla\Phi(\mathbf{x}_{t+1})}{\ell}\right)\right\| &\le (2\kappa_{\mathbf{y}}+2)\|\mathbf{x}_{t+1}-\mathbf{x}_t^\star\| + 2\|\mathbf{x}_t-\mathbf{x}_t^\star\| \\ &\overset{\kappa_{\mathbf{y}}\ge 1}{\le} 6\kappa_{\mathbf{y}}\|\mathbf{x}_{t+1}-\mathbf{x}_t^\star\| + 2\|\mathbf{x}_{t+1}-\mathbf{x}_t\|.\end{aligned} \tag{3.36}$$

Putting Eq. (3.34), Eq. (3.35) and Eq. (3.36) together with the Cauchy-Schwarz inequality yields

$$\begin{aligned}(\ell\|\mathbf{x}_{t+1} - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}_{t+1} - (1/\ell)\nabla\Phi(\mathbf{x}_{t+1})\right)\|)^2 &\le 72\kappa_{\mathbf{y}}^2\ell^2\|\mathbf{x}_{t+1}-\mathbf{x}_t^\star\|^2 + 8\ell^2\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 \\ &\le 8\ell(\Phi(\mathbf{x}_t)-\Phi(\mathbf{x}_{t+1})+\delta) + 144\kappa_{\mathbf{y}}^2\ell\delta.\end{aligned}$$

Summing up the above inequality over $t=0,1,\ldots,T-1$ and dividing it by T yields that

$$\begin{aligned}\tfrac{1}{T}\left(\sum_{t=0}^{T-1}(\ell\|\mathbf{x}_{t+1} - \mathcal{P}_{\mathcal{X}}(\mathbf{x}_{t+1} - (1/\ell)\nabla\Phi(\mathbf{x}_{t+1}))\|)^2\right) &\le \tfrac{8\ell(\Phi(\mathbf{x}_0)-\Phi(\mathbf{x}_T))}{T} + 8\ell\delta + 144\kappa_{\mathbf{y}}^2\ell\delta \\ &\overset{\kappa_{\mathbf{y}}\ge 1}{\le} \tfrac{8\ell(\Phi(\mathbf{x}_0)-\Phi(\mathbf{x}_T))}{T} + 152\kappa_{\mathbf{y}}^2\ell\delta.\end{aligned}$$

Since $\hat{\mathbf{x}} = \mathbf{x}_s$ is uniformly chosen from $\{\mathbf{x}_s\}_{1\le s\le T}$ and $\delta \le \epsilon^2/(10\kappa_{\mathbf{y}})^4\ell$, we have

$$\begin{aligned}\mathbb{E}[(\ell\|\hat{\mathbf{x}} - \mathcal{P}_{\mathcal{X}}(\hat{\mathbf{x}} - (1/\ell)\nabla\Phi(\hat{\mathbf{x}}))\|)^2] &= \frac{1}{T}\left(\sum_{t=0}^{T-1}(\ell\|\mathbf{x}_{t+1} - \mathcal{P}_{\mathcal{X}}(\mathbf{x}_{t+1} - (1/\ell)\nabla\Phi(\mathbf{x}_{t+1}))\|)^2\right)\\ &\le \frac{8\ell(\Phi(\mathbf{x}_0)-\Phi(\mathbf{x}_T))}{T} + 152\kappa_{\mathbf{y}}^2\ell\delta \le \frac{8\ell\Delta_\Phi}{T} + \frac{\epsilon^2}{8}.\end{aligned}$$

Using the Markov inequality, we conclude that there exists $T > c\ell\Delta_\Phi\epsilon^{-2}$, where the output $\hat{\mathbf{x}}$ will satisfy $\ell\|\hat{\mathbf{x}} - \mathcal{P}_{\mathcal{X}}(\hat{\mathbf{x}} - (1/\ell)\nabla\Phi(\hat{\mathbf{x}}))\| \le \epsilon/2$ with probability at least $2/3$.

For simplicity, we denote $\hat{\mathbf{y}}^+ = \mathcal{P}_{\mathcal{Y}}[\hat{\mathbf{y}} + (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]$. Since $\hat{\mathbf{y}}$ is obtained by running AGD on $-f(\hat{\mathbf{x}}, \cdot)$ to optimal with tolerance $\delta \le \epsilon^2/(10\kappa_{\mathbf{y}})^4\ell$, and $f(\hat{\mathbf{x}}, \cdot)$ is $\mu_{\mathbf{y}}$-concave function, we know that δ-optimality guarantees:

$$\ell\|\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}\| \le \epsilon, \quad \|\hat{\mathbf{y}}^+ - \mathbf{y}^\star(\hat{\mathbf{x}})\| \le \tfrac{\epsilon}{2\ell}.$$

Putting these pieces together yields that

$$\begin{aligned}\ell\|\hat{\mathbf{x}} - \mathcal{P}_{\mathcal{X}}(\hat{\mathbf{x}} - (1/\ell)\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+))\| &\le \ell\|\hat{\mathbf{x}} - \mathcal{P}_{\mathcal{X}}(\hat{\mathbf{x}} - (1/\ell)\nabla\Phi(\hat{\mathbf{x}}))\| + \|\nabla\Phi(\hat{\mathbf{x}}) - \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+)\|\\ &\le \ell\|\hat{\mathbf{x}} - \mathcal{P}_{\mathcal{X}}(\hat{\mathbf{x}} - (1/\ell)\nabla\Phi(\hat{\mathbf{x}}))\| + \ell\|\hat{\mathbf{y}}^+ - \mathbf{y}^\star(\hat{\mathbf{x}})\|\\ &\le \epsilon.\end{aligned}$$

This implies that $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is an ϵ-stationary point. Furthermore, we call the solver Maximin-AG2 at each iteration. Using Theorem 3.4.2 and $\delta \le \epsilon^2/(10\kappa_{\mathbf{y}})^4\ell$, the number of gradient evaluations at each iteration is bounded by

$$O\left(\sqrt{\kappa_{\mathbf{y}}}\log\left(\frac{\kappa_{\mathbf{y}}^5\ell^2(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon^2}\right)\log\left(\frac{\kappa_{\mathbf{y}}^4\ell^2 D_{\mathbf{y}}^2}{\epsilon^2}\right)\right).$$

Therefore, we conclude that the total number of gradient evaluations is bounded by

$$O\left(\frac{\ell\Delta_\Phi}{\epsilon^2}\cdot\sqrt{\kappa_{\mathbf{y}}}\log^2\left(\frac{\kappa_{\mathbf{y}}\ell(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right).$$

This completes the proof.

Proof of Corollary 3.6.2. Recall that the function $\tilde{f}_\epsilon$ is defined by

$$\tilde{f}_\epsilon(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y}) - \tfrac{\epsilon\|\mathbf{y}-\mathbf{y}_0\|^2}{4D_{\mathbf{y}}}.$$

This implies that the following statement holds for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$ that

$$\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}}\tilde{f}_\epsilon(\mathbf{x}, \mathbf{y}) = 0, \qquad \|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}}\tilde{f}_\epsilon(\mathbf{x}, \mathbf{y})\| \le \tfrac{\epsilon}{2}.$$

Since $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \textsc{Minimax-PPA}(\tilde{f}_\epsilon, \mathbf{x}_0, \mathbf{y}_0, \ell, \epsilon/(2D_{\mathbf{y}}), \epsilon/2, T)$, we have

$$\ell\|\mathcal{P}_{\mathcal{X}}[\hat{\mathbf{x}} - (1/\ell)\nabla_{\mathbf{x}}\tilde{f}_\epsilon(\hat{\mathbf{x}}, \hat{\mathbf{y}}_\epsilon^+)] - \hat{\mathbf{x}}\| \le \tfrac{\epsilon}{2}, \quad \ell\|\hat{\mathbf{y}}_\epsilon^+ - \hat{\mathbf{y}}\| \le \tfrac{\epsilon}{2}, \quad \hat{\mathbf{y}}_\epsilon^+ = \mathcal{P}_{\mathcal{Y}}[\hat{\mathbf{y}} + (1/\ell)\nabla_{\mathbf{y}}\tilde{f}_\epsilon(\hat{\mathbf{x}}, \hat{\mathbf{y}})].$$

Putting these pieces together with $\hat{\mathbf{y}}^+ = \mathcal{P}_{\mathcal{Y}}[\hat{\mathbf{y}} + (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]$ yields that

$$\ell\|\mathcal{P}_{\mathcal{X}}[\hat{\mathbf{x}}-(1/\ell)\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+)]-\hat{\mathbf{x}}\| \le \tfrac{\epsilon}{2}+\|\nabla_{\mathbf{x}}\tilde{f}_\epsilon(\hat{\mathbf{x}}, \hat{\mathbf{y}}_\epsilon^+)-\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+)\| \le \tfrac{\epsilon}{2}+\|\nabla_{\mathbf{y}}\tilde{f}_\epsilon(\hat{\mathbf{x}}, \hat{\mathbf{y}})-\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})\| \le \epsilon,$$

and

$$\ell\|\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}\| \le \ell\|\mathcal{P}_{\mathcal{Y}}[\hat{\mathbf{y}} + (1/\ell)\nabla_{\mathbf{y}}\tilde{f}_\epsilon(\hat{\mathbf{x}}, \hat{\mathbf{y}})] - \hat{\mathbf{y}}\| + \|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}}\tilde{f}_\epsilon(\mathbf{x}, \mathbf{y})\| \le \epsilon.$$

Therefore, we conclude that $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is an ϵ-stationary point of f. Furthermore, letting $\kappa_{\mathbf{y}} = 2\ell D_{\mathbf{y}}/\epsilon$ in the gradient complexity bound presented in Theorem 3.6.1, we conclude that the total number of gradient evaluations is bounded by

$$O\left(\frac{\ell\Delta_\Phi}{\epsilon^2}\cdot\sqrt{\frac{\ell D_{\mathbf{y}}}{\epsilon}}\log^2\left(\frac{\ell(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right).$$

This completes the proof.

Proof of Theorem 3.8.7. Using the same argument as in Theorem 3.6.1, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \le \tfrac{\Phi(\mathbf{x}_t)-\Phi(\mathbf{x}_{t+1})+\delta}{\ell}. \tag{3.37}$$

and

$$\|\mathbf{x}_{t+1} - \mathbf{x}_t^\star\|^2 \le \tfrac{2}{\ell}\left(\Phi(\mathbf{x}_{t+1}) + \ell\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 - \min_{\mathbf{x}\in\mathbb{R}^m}\left\{\Phi(\mathbf{x}) + \ell\|\mathbf{x} - \mathbf{x}_t\|^2\right\}\right) \le \tfrac{2\delta}{\ell}. \tag{3.38}$$

Since Φ is differentiable, we have $\nabla\Phi(\mathbf{x}_t^\star) + 2\ell(\mathbf{x}_t^\star - \mathbf{x}_t) = 0$ which implies that $\|\nabla\Phi(\mathbf{x}_t^\star)\| = 2\ell\|\mathbf{x}_t^\star - \mathbf{x}_t\|$. Since $\Phi(\cdot)$ is $2\kappa_{\mathbf{y}}\ell$-smooth, we have $\|\nabla\Phi(\mathbf{x}_{t+1}) - \nabla\Phi(\mathbf{x}_t^\star)\| \le 2\kappa_{\mathbf{y}}\ell\|\mathbf{x}_{t+1} - \mathbf{x}_t^\star\|$. Putting these pieces together yields that

$$\begin{aligned}\|\nabla\Phi(\mathbf{x}_{t+1})\| &\le 2\kappa_{\mathbf{y}}\ell\|\mathbf{x}_{t+1} - \mathbf{x}_t^\star\| + 2\ell\|\mathbf{x}_t^\star - \mathbf{x}_t\| \le (2\kappa_{\mathbf{y}}\ell + 2\ell)\|\mathbf{x}_{t+1} - \mathbf{x}_t^\star\| + 2\ell\|\mathbf{x}_{t+1} - \mathbf{x}_t\| \\ &\overset{\kappa_{\mathbf{y}}\ge 1}{\le} 4\kappa_{\mathbf{y}}\ell\|\mathbf{x}_{t+1} - \mathbf{x}_t^\star\| + 2\ell\|\mathbf{x}_{t+1} - \mathbf{x}_t\|.\end{aligned} \tag{3.39}$$

Putting Eq. (3.37), Eq. (3.38) and Eq. (3.39) together with the Cauchy-Schwarz inequality yields

$$\|\nabla\Phi(\mathbf{x}_{t+1})\|^2 \le 32\kappa_{\mathbf{y}}^2\ell^2\|\mathbf{x}_{t+1} - \mathbf{x}_t^\star\|^2 + 8\ell^2\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \le 8\ell(\Phi(\mathbf{x}_t) - \Phi(\mathbf{x}_{t+1}) + \delta) + 64\kappa_{\mathbf{y}}^2\ell\delta.$$

Summing up the above inequality over $t = 0, 1, \ldots, T-1$ and dividing it by T yields that

$$\tfrac{1}{T}\left(\sum_{t=0}^{T-1}\|\nabla\Phi(\mathbf{x}_{t+1})\|^2\right) \le \tfrac{8\ell(\Phi(\mathbf{x}_0)-\Phi(\mathbf{x}_T))}{T} + 8\ell\delta + 64\kappa_{\mathbf{y}}^2\ell\delta \overset{\kappa_{\mathbf{y}}\ge 1}{\le} \tfrac{8\ell(\Phi(\mathbf{x}_0)-\Phi(\mathbf{x}_T))}{T} + 72\kappa_{\mathbf{y}}^2\ell\delta.$$

Since $\hat{\mathbf{x}} = \mathbf{x}_s$ is uniformly chosen from $\{\mathbf{x}_s\}_{1\le s\le T}$ and $\delta \le \epsilon^2/144\kappa_{\mathbf{y}}^2\ell$, we have

$$\mathbb{E}[\|\nabla\Phi(\hat{\mathbf{x}})\|^2] = \tfrac{1}{T}\left(\sum_{t=0}^{T-1}\|\nabla\Phi(\mathbf{x}_{t+1})\|^2\right) \le \tfrac{8\ell(\Phi(\mathbf{x}_0)-\Phi(\mathbf{x}_T))}{T} + 72\kappa_{\mathbf{y}}^2\ell\delta \le \tfrac{8\ell\Delta_\Phi}{T} + \tfrac{\epsilon^2}{2}.$$

Using the Markov inequality, we conclude that there exists $T > c\ell\Delta_\Phi\epsilon^{-2}$, where the output $\hat{\mathbf{x}}$ will satisfy $\|\nabla\Phi(\hat{\mathbf{x}})\| \le \epsilon$ with probability at least 2/3. Furthermore, we call the solver MAXIMIN-AG2 at each iteration. Using Theorem 3.4.2 and $\delta \le \epsilon^2/144\kappa_{\mathbf{y}}^2\ell$, the number of gradient evaluations at each iteration is bounded by

$$O\left(\sqrt{\kappa_{\mathbf{y}}}\log\left(\frac{\kappa_{\mathbf{y}}^3\ell^2(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon^2}\right)\log\left(\frac{\kappa_{\mathbf{y}}^2\ell^2 D_{\mathbf{y}}^2}{\epsilon^2}\right)\right).$$

Therefore, we conclude that the total number of gradient evaluations is bounded by

$$O\left(\frac{\ell\Delta_\Phi}{\epsilon^2}\cdot\sqrt{\kappa_{\mathbf{y}}}\log^2\left(\frac{\kappa_{\mathbf{y}}\ell(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right).$$

This completes the proof.

Proof of Corollary 3.8.8. Recall that the function $\bar{f}_\epsilon$ is defined by

$$\bar{f}_\epsilon(\mathbf{x},\mathbf{y}) = f(\mathbf{x},\mathbf{y}) - \tfrac{\epsilon^2\|\mathbf{y}-\mathbf{y}_0\|^2}{200\ell D_{\mathbf{y}}^2}.$$

This implies that the following statement holds for all $(\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$ that

$$\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}}\bar{f}_\epsilon(\mathbf{x},\mathbf{y}) = 0, \qquad \|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}}\bar{f}_\epsilon(\mathbf{x},\mathbf{y})\| \le \tfrac{\epsilon^2}{100\ell D_{\mathbf{y}}}.$$

Using Theorem 3.8.7 and letting $\mathbf{y}_\epsilon^\star(\cdot) = \operatorname{argmin}_{\mathbf{y}\in\mathcal{Y}}\bar{f}_\epsilon(\cdot,\mathbf{y})$, we have

$$\|\nabla_{\mathbf{x}}\bar{f}_\epsilon(\hat{\mathbf{x}},\mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}))]\| \le \tfrac{\epsilon}{10}, \qquad \ell\|\mathcal{P}_{\mathcal{Y}}[\mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}) + (1/\ell)\nabla_{\mathbf{y}}\bar{f}_\epsilon(\hat{\mathbf{x}},\mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}))] - \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}})\| = 0.$$

For simplicity, we define $\mathbf{y}_\epsilon^+ = \mathcal{P}_{\mathcal{Y}}[\mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}) + (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}))]$. Then, we have

$$\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^+)\| \le \tfrac{\epsilon}{10} + \tfrac{\epsilon^2}{50\ell D_{\mathbf{y}}}, \qquad \ell\|\mathbf{y}_\epsilon^+ - \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}})\| \le \tfrac{\epsilon^2}{50\ell D_{\mathbf{y}}}.$$

Now let $\mathbf{x}^\star(\hat{\mathbf{x}}) = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^m}\Phi_{1/2\ell}(\mathbf{x}) := \Phi(\mathbf{x}) + \ell\|\mathbf{x}-\hat{\mathbf{x}}\|^2$, we have

$$\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\|^2 = 4\ell^2\|\hat{\mathbf{x}} - \mathbf{x}^\star(\hat{\mathbf{x}})\|^2$$

Since $\Phi(\cdot) + \ell\|\cdot - \hat{\mathbf{x}}\|^2$ is $\ell/2$-strongly-convex, we have

$$\begin{aligned}&\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star(\hat{\mathbf{x}}),\mathbf{y}) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\|^2\\ &= \Phi(\hat{\mathbf{x}}) - \Phi(\mathbf{x}^\star(\hat{\mathbf{x}})) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\|^2 \ge \tfrac{\ell\|\hat{\mathbf{x}}-\mathbf{x}^\star(\hat{\mathbf{x}})\|^2}{4} = \tfrac{\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\|^2}{16\ell}.\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star(\hat{\mathbf{x}}),\mathbf{y}) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}})-\hat{\mathbf{x}}\|^2 \\
&= \max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^+) + f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^+) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star(\hat{\mathbf{x}}),\mathbf{y}) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}})-\hat{\mathbf{x}}\|^2 \\
&\le \max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^+) + (f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^+) - f(\mathbf{x}^\star(\hat{\mathbf{x}}),\mathbf{y}_\epsilon^+) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}})-\hat{\mathbf{x}}\|^2) \\
&\le \max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^+) + (\|\hat{\mathbf{x}}-\mathbf{x}^\star(\hat{\mathbf{x}})\|\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^+)\| - \ell\|\hat{\mathbf{x}}-\mathbf{x}^\star(\hat{\mathbf{x}})\|^2) \\
&\le \max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^+) + \tfrac{\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^+)\|^2}{4\ell}.
\end{aligned}$$

Recall that $\mathbf{y}_\epsilon^+ = \mathcal{P}_{\mathcal{Y}}[\mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}) + (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}))]$, we have

$$(\mathbf{y}-\mathbf{y}_\epsilon^+)^\top(\mathbf{y}_\epsilon^+ - \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}) - (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}))) \ge 0, \text{ for all } \mathbf{y}\in\mathcal{Y}.$$

Together with the ℓ-smoothness of the function $f(\hat{\mathbf{x}},\cdot)$ and the boundedness of $\mathcal{Y}$, we have

$$f(\hat{\mathbf{x}},\mathbf{y}) - f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^+) \le \tfrac{\ell}{2}(\|\mathbf{y}-\mathbf{y}_\epsilon^\star(\hat{\mathbf{x}})\|^2 - \|\mathbf{y}-\mathbf{y}_\epsilon^+\|^2) \le \ell D_{\mathbf{y}}\|\mathbf{y}_\epsilon^+ - \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}})\|, \text{ for all } \mathbf{y}\in\mathcal{Y}.$$

Putting these pieces together yields that

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star(\hat{\mathbf{x}}),\mathbf{y}) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}})-\hat{\mathbf{x}}\|^2 \le \ell D_{\mathbf{y}}\|\mathbf{y}_\epsilon^+ - \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}})\| + \tfrac{\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^+)\|^2}{4\ell}.$$

Since a point $(\hat{\mathbf{x}},\mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}))$ satisfies that

$$\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^+)\| \le \tfrac{\epsilon}{10} + \tfrac{\epsilon^2}{50\ell D_{\mathbf{y}}}, \quad \ell\|\mathbf{y}_\epsilon^+ - \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}})\| \le \tfrac{\epsilon^2}{50\ell D_{\mathbf{y}}},$$

we have (assume that $\epsilon \precsim \ell D_{\mathbf{y}}$ without loss of generality)

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}},\mathbf{y}) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star(\hat{\mathbf{x}}),\mathbf{y}) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}})-\hat{\mathbf{x}}\|^2 \le \tfrac{\epsilon^2}{50\ell} + \tfrac{\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\mathbf{y}_\epsilon^+)\|^2}{4\ell}.$$

Putting these pieces together yields that $\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\| \le \epsilon$. Furthermore, letting $\kappa_{\mathbf{y}} = 100\ell^2 D_{\mathbf{y}}^2/\epsilon^2$ in the gradient complexity bound presented in Theorem 3.8.7, we conclude that the total number of gradient evaluations is bounded by

$$O\left(\frac{\ell^2 D_{\mathbf{y}}\Delta_\Phi}{\epsilon^3}\log^2\left(\frac{\ell(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right).$$

This completes the proof.

3.12 Proof of Technical Lemmas

We provide complete proofs for the lemmas in this chapter.

Proof of Lemma 3.8.4. We provide a proof for an expanded version of Lemma 3.8.4.

Lemma 3.12.1 *If Φ is ℓ-weakly convex, we have*

(a) $\Phi_{1/2\ell}(\mathbf{x})$ and $\textsc{Prox}_{\Phi/2\ell}(\mathbf{x}) = \operatorname{argmin} \Phi(\mathbf{w}) + \ell\|\mathbf{w}-\mathbf{x}\|^2$ are well defined for any $\mathbf{x} \in \mathbb{R}^m$.

(b) $\Phi(\textsc{Prox}_{\Phi/2\ell}(\mathbf{x})) \le \Phi(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^m$.

(c) $\Phi_{1/2\ell}$ is 4ℓ-smooth with $\nabla\Phi_{1/2\ell}(\mathbf{x}) = 2\ell(\mathbf{x} - \textsc{Prox}_{\Phi/2\ell}(\mathbf{x}))$.

Proof. Since Φ is ℓ-weakly convex, $\Phi(\cdot) + (\ell/2)\,\|\cdot - \mathbf{x}\|^2$ is convex for any $\mathbf{x} \in \mathbb{R}^m$. This implies that $\Phi(\cdot) + \ell\|\cdot - \mathbf{x}\|^2$ is $(\ell/2)$-strongly convex and $\Phi_{1/2\ell}(\mathbf{x})$ and $\textsc{Prox}_{\Phi/2\ell}(\mathbf{x})$ are well defined. For any $\mathbf{x} \in \mathbb{R}^m$, the definition of $\textsc{Prox}_{\Phi/2\ell}(\mathbf{x})$ implies that

$$\Phi(\textsc{Prox}_{\Phi/2\ell}(\mathbf{x})) \le \Phi_{1/2\ell}(\textsc{Prox}_{\Phi/2\ell}(\mathbf{x})) \le \Phi(\mathbf{x}).$$

By Davis and Drusvyatskiy [2019, Lemma 2.2], $\Phi_{1/2\ell}$ is differentiable with the gradient $\nabla\Phi_{1/2\ell}(\mathbf{x}) = 2\ell(\mathbf{x} - \textsc{Prox}_{\Phi/2\ell}(\mathbf{x}))$. Since $\textsc{Prox}_{\Phi/2\ell}$ is 1-Lipschitz, we $\|\nabla\Phi_{1/2\ell}(\mathbf{x}) - \nabla\Phi_{1/2\ell}(\mathbf{x}')\| \le 4\ell\|\mathbf{x} - \mathbf{x}'\|$. Therefore, the function $\Phi_{1/2\ell}$ is 4ℓ-smooth. □

Proof of Lemma 3.8.6. Denote $\hat{\mathbf{x}} := \textsc{Prox}_{\Phi/2\ell}(\mathbf{x})$, part (c) in Lemma 3.8.4 implies

$$\|\hat{\mathbf{x}} - \mathbf{x}\| = \tfrac{\|\nabla\Phi_{1/2\ell}(\mathbf{x})\|}{2\ell}.$$

Also, we have $2\ell(\mathbf{x} - \hat{\mathbf{x}}) \in \partial\Phi(\hat{\mathbf{x}})$. Putting these pieces together yields the desired result.

Proof of Lemma 3.9.2. For part (a), let $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^m$, the points $\mathbf{y}_g^\star(\mathbf{x})$ and $\mathbf{y}_g^\star(\mathbf{x}')$ satisfy

$$(\mathbf{y} - \mathbf{y}_g^\star(\mathbf{x}))^\top \nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y}_g^\star(\mathbf{x})) \le 0, \quad \forall \mathbf{y} \in \mathcal{Y}, \tag{3.40}$$

$$(\mathbf{y} - \mathbf{y}_g^\star(\mathbf{x}'))^\top \nabla_{\mathbf{y}} g(\mathbf{x}', \mathbf{y}_g^\star(\mathbf{x}')) \le 0, \quad \forall \mathbf{y} \in \mathcal{Y}. \tag{3.41}$$

Summing up Eq. (3.40) with $\mathbf{y} = \mathbf{y}_g^\star(\mathbf{x}')$ and Eq. (3.41) with $\mathbf{y} = \mathbf{y}_g^\star(\mathbf{x})$ yields

$$(\mathbf{y}_g^\star(\mathbf{x}') - \mathbf{y}_g^\star(\mathbf{x}))^\top (\nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y}_g^\star(\mathbf{x})) - \nabla_{\mathbf{y}} g(\mathbf{x}', \mathbf{y}_g^\star(\mathbf{x}'))) \le 0.$$

Since $g(\mathbf{x}, \cdot)$ is $\mu_{\mathbf{y}}$-strongly concave, we have

$$(\mathbf{y}_g^\star(\mathbf{x}') - \mathbf{y}_g^\star(\mathbf{x}))^\top (\nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y}_g^\star(\mathbf{x}')) - \nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y}_g^\star(\mathbf{x}))) + \mu_{\mathbf{y}}\|\mathbf{y}_g^\star(\mathbf{x}') - \mathbf{y}_g^\star(\mathbf{x})\|^2 \le 0.$$

Summing up the above two inequalities yields that

$$(\mathbf{y}_g^\star(\mathbf{x}') - \mathbf{y}_g^\star(\mathbf{x}))^\top (\nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y}_g^\star(\mathbf{x}')) - \nabla_{\mathbf{y}} g(\mathbf{x}', \mathbf{y}_g^\star(\mathbf{x}'))) + \mu_{\mathbf{y}}\|\mathbf{y}_g^\star(\mathbf{x}') - \mathbf{y}_g^\star(\mathbf{x})\|^2 \le 0.$$

Since $\nabla_{\mathbf{y}} g$ is ℓ-Lipschitz, we have $\mu_{\mathbf{y}}\|\mathbf{y}_g^\star(\mathbf{x}') - \mathbf{y}_g^\star(\mathbf{x})\|^2 \le \ell\|\mathbf{y}_g^\star(\mathbf{x}') - \mathbf{y}_g^\star(\mathbf{x})\|\|\mathbf{x}' - \mathbf{x}\|$. Therefore, we conclude that the function $\mathbf{y}_g^\star(\cdot)$ is $\kappa_{\mathbf{y}}$-Lipschitz.

For part (b), since the function $\mathbf{y}_g^\star(\cdot)$ is unique, Danskin's theorem [Rockafellar, 1970] implies that Φ_g is differentiable and $\nabla\Phi_g(\cdot) = \nabla_{\mathbf{x}} g(\cdot, \mathbf{y}_g^\star(\cdot))$. Let $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^m$, we have

$$\begin{aligned}\|\nabla\Phi_g(\mathbf{x}) - \nabla\Phi_g(\mathbf{x}')\| &= \|\nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}_g^\star(\mathbf{x})) - \nabla_{\mathbf{x}} g(\mathbf{x}', \mathbf{y}_g^\star(\mathbf{x}'))\| \le \ell\|\mathbf{x} - \mathbf{x}'\| + \ell\|\mathbf{y}_g^\star(\mathbf{x}) - \mathbf{y}_g^\star(\mathbf{x}')\| \\ &\overset{\bar\kappa \ge 1}{\le} \kappa_{\mathbf{y}}\ell\|\mathbf{x} - \mathbf{x}'\| + \ell\|\mathbf{y}_g^\star(\mathbf{x}) - \mathbf{y}_g^\star(\mathbf{x}')\|.\end{aligned}$$

Since $\mathbf{y}_g^\star(\cdot)$ is $\kappa_{\mathbf{y}}$-Lipschitz, the function Φ_g is $2\kappa_{\mathbf{y}}\ell$-smooth. Furthermore, let $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^m$, we have

$$\begin{aligned}\Phi_g(\mathbf{x}') - \Phi_g(\mathbf{x}) - (\mathbf{x}' - \mathbf{x})^\top\nabla\Phi_g(\mathbf{x}) &= g(\mathbf{x}', \mathbf{y}_g^\star(\mathbf{x}')) - g(\mathbf{x}, \mathbf{y}_g^\star(\mathbf{x})) - (\mathbf{x}' - \mathbf{x})^\top\nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}_g^\star(\mathbf{x})) \\ &\ge g(\mathbf{x}', \mathbf{y}_g^\star(\mathbf{x})) - g(\mathbf{x}, \mathbf{y}_g^\star(\mathbf{x})) - (\mathbf{x}' - \mathbf{x})^\top\nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}_g^\star(\mathbf{x})).\end{aligned}$$

Since $g(\cdot, \mathbf{y})$ is $\mu_{\mathbf{x}}$-strongly convex for each $\mathbf{y} \in \mathcal{Y}$, we have

$$\Phi_g(\mathbf{x}') - \Phi_g(\mathbf{x}) - (\mathbf{x}' - \mathbf{x})^\top\nabla\Phi_g(\mathbf{x}) \ge \tfrac{\mu_{\mathbf{x}}\|\mathbf{x}' - \mathbf{x}\|^2}{2}.$$

Therefore, the function Φ_g is $\mu_{\mathbf{x}}$-strongly convex.

For part (c), let $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^n$, the points $\mathbf{x}_g^\star(\mathbf{y})$ and $\mathbf{x}_g^\star(\mathbf{y}')$ satisfy

$$(\mathbf{x} - \mathbf{x}_g^\star(\mathbf{y}))^\top\nabla_{\mathbf{x}} g(\mathbf{x}_g^\star(\mathbf{y}), \mathbf{y}) \ge 0, \quad \forall \mathbf{x} \in \mathcal{X}, \tag{3.42}$$

$$(\mathbf{x} - \mathbf{x}_g^\star(\mathbf{y}'))^\top\nabla_{\mathbf{x}} g(\mathbf{x}_g^\star(\mathbf{y}'), \mathbf{y}') \ge 0, \quad \forall \mathbf{x} \in \mathcal{X}. \tag{3.43}$$

Summing up Eq. (3.42) with $\mathbf{x} = \mathbf{x}_g^\star(\mathbf{y}')$ and Eq. (3.43) with $\mathbf{x} = \mathbf{x}_g^\star(\mathbf{y})$ yields

$$(\mathbf{x}_g^\star(\mathbf{y}') - \mathbf{x}_g^\star(\mathbf{y}))^\top(\nabla_{\mathbf{x}} g(\mathbf{x}_g^\star(\mathbf{y}), \mathbf{y}) - \nabla_{\mathbf{x}} g(\mathbf{x}_g^\star(\mathbf{y}'), \mathbf{y}')) \ge 0.$$

Since $g(\cdot, \mathbf{y})$ is $\mu_{\mathbf{x}}$-strongly convex, we have

$$(\mathbf{x}_g^\star(\mathbf{y}') - \mathbf{x}_g^\star(\mathbf{y}))^\top(\nabla_{\mathbf{x}} g(\mathbf{x}_g^\star(\mathbf{y}'), \mathbf{y}') - \nabla_{\mathbf{x}} g(\mathbf{x}_g^\star(\mathbf{y}), \mathbf{y}')) - \mu_{\mathbf{x}}\|\mathbf{x}_g^\star(\mathbf{y}') - \mathbf{x}_g^\star(\mathbf{y})\|^2 \ge 0.$$

Summing up the above two inequalities yields that

$$(\mathbf{x}_g^\star(\mathbf{y}') - \mathbf{x}_g^\star(\mathbf{y}))^\top(\nabla_{\mathbf{x}} g(\mathbf{x}_g^\star(\mathbf{y}), \mathbf{y}) - \nabla_{\mathbf{x}} g(\mathbf{x}_g^\star(\mathbf{y}), \mathbf{y}')) - \mu_{\mathbf{x}}\|\mathbf{x}_g^\star(\mathbf{y}') - \mathbf{x}_g^\star(\mathbf{y})\|^2 \ge 0.$$

Since $\nabla_{\mathbf{x}} g$ is ℓ-smooth, we have $\mu_{\mathbf{x}}\|\mathbf{x}_g^\star(\mathbf{y}') - \mathbf{x}_g^\star(\mathbf{y})\|^2 \le \ell\|\mathbf{x}_g^\star(\mathbf{y}') - \mathbf{x}_g^\star(\mathbf{y})\|\|\mathbf{y}' - \mathbf{y}\|$. Therefore, we conclude that the function $\mathbf{x}_g^\star$ is $\kappa_{\mathbf{x}}$-Lipschitz.

For part (d), since the function $\mathbf{x}_g^\star(\cdot)$ is unique, Danskin's theorem [Rockafellar, 1970] implies that Ψ_g is differentiable and $\nabla\Psi_g(\cdot) = \nabla_{\mathbf{y}} g(\mathbf{x}_g^\star(\cdot), \cdot)$. Let $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^n$, we have

$$\|\nabla\Psi_g(\mathbf{y}) - \nabla\Psi_g(\mathbf{y}')\| = \|\nabla_{\mathbf{y}} g(\mathbf{x}_g^\star(\mathbf{y}), \mathbf{y}) - \nabla_{\mathbf{y}} g(\mathbf{x}_g^\star(\mathbf{y}'), \mathbf{y})\| \le \ell\|\mathbf{x}_g^\star(\mathbf{y}) - \mathbf{x}_g^\star(\mathbf{y}')\| + \ell\|\mathbf{y} - \mathbf{y}'\|.$$

Since $\mathbf{x}_g^\star(\cdot)$ is $\kappa_{\mathbf{x}}$-Lipschitz, the function Ψ_g is $2\kappa_{\mathbf{x}}\ell$-smooth. Let $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^n$, we have

$$\begin{aligned}\Psi_g(\mathbf{y}) - \Psi_g(\mathbf{y}') - (\mathbf{y} - \mathbf{y}')^\top\nabla\Psi_g(\mathbf{y}) &= g(\mathbf{x}_g^\star(\mathbf{y}), \mathbf{y}) - g(\mathbf{x}_g^\star(\mathbf{y}'), \mathbf{y}') - (\mathbf{y} - \mathbf{y}')^\top\nabla_{\mathbf{y}} g(\mathbf{x}_g^\star(\mathbf{y}), \mathbf{y}) \\ &\ge g(\mathbf{x}_g^\star(\mathbf{y}), \mathbf{y}) - g(\mathbf{x}_g^\star(\mathbf{y}), \mathbf{y}') - (\mathbf{y} - \mathbf{y}')^\top\nabla_{\mathbf{y}} g(\mathbf{x}_g^\star(\mathbf{y}), \mathbf{y}).\end{aligned}$$

Since $g(\mathbf{x}, \cdot)$ is $\mu_{\mathbf{y}}$-strongly concave for each $\mathbf{x} \in \mathcal{X}$, we have

$$\Psi_g(\mathbf{y}) - \Psi_g(\mathbf{y}') - (\mathbf{y} - \mathbf{y}')^\top\nabla\Psi_g(\mathbf{y}) \ge \tfrac{\mu_{\mathbf{y}}\|\mathbf{y}' - \mathbf{y}\|^2}{2}.$$

Therefore, the function Ψ_g is $\mu_{\mathbf{y}}$-strongly concave.

Chapter 4

Riemannian Gradient-Based Algorithm

From optimal transport to robust dimensionality reduction, a plethora of machine learning applications can be cast into the min-max optimization problems over Riemannian manifolds. Though many min-max algorithms have been analyzed in the Euclidean setting, it has proved elusive to translate these results to the Riemannian case. Zhang et al. [2022b] have recently shown that geodesic convex concave Riemannian problems always admit saddle-point solutions. Inspired by this result, we study whether a performance gap between Riemannian and optimal Euclidean space convex-concave algorithms is necessary. We answer this question in the negative—we prove that the Riemannian corrected extragradient (RCEG) method achieves last-iterate convergence at a linear rate in the geodesically strongly-convex-concave case, matching the Euclidean result. Our results also extend to the stochastic or non-smooth case where RCEG and Riemanian gradient ascent descent (RGDA) achieve near-optimal convergence rates up to factors depending on curvature of the manifold.

4.1 Introduction

Constrained optimization problems arise throughout machine learning, in classical settings such as dimension reduction [Boumal and Absil, 2011], dictionary learning [Sun et al., 2016a,b], and deep neural networks [Huang et al., 2018], but also in emerging problems involving decision-making and multi-agent interactions. While simple convex constraints (such as norm constraints) can be easily incorporated in standard optimization formulations, notably (proximal) gradient descent [Raskutti and Mukherjee, 2015, Giannou et al., 2021b,a, Antonakopoulos et al., 2020, Vlatakis-Gkaragkounis et al., 2020], in a range of other applications such as matrix recovery [Fornasier et al., 2011, Candes et al., 2008], low-rank matrix factorization [Han et al., 2021] and generative adversarial nets [Goodfellow et al., 2014], the constraints are fundamentally nonconvex and are often treated via special heuristics.

Thus, a general goal is to design algorithms that systematically take account of special

geometric structure of the feasible set [Mei et al., 2021, Lojasiewicz, 1963, Polyak, 1963]. A long line of work in the machine learning (ML) community has focused on understanding the geometric properties of commonly used constraints and how they affect optimization; [see, e.g., Ge et al., 2015, Anandkumar and Ge, 2016, Sra and Hosseini, 2016, Jin et al., 2017, Ge et al., 2017a, Du et al., 2017, Reddi et al., 2018a, Criscitiello and Boumal, 2019, Jin et al., 2021]. A prominent aspect of this agenda has been the re-expression of these constraints through the lens of Riemannian manifolds. This has given rise to new algorithms [Sra and Hosseini, 2015, Hosseini and Sra, 2015] with a wide range of ML applications, incldluing online principal component analysis (PCA), the computation of Mahalanobis distance from noisy measurements [Bonnabel, 2013], consensus distributed algorithms for aggregation in ad-hoc wireless networks [Tron et al., 2012] and maximum likelihood estimation for certain non-Gaussian (heavy- or light-tailed) distributions [Wiesel, 2012].

Going beyond simple minimization problems, the robustification of many ML tasks can be formulated as min-max optimization problems. Well-known examples in this domain include adversarial machine learning [Kumar et al., 2017, Chen et al., 2018], optimal transport [Lin et al., 2020a], and online learning [Mertikopoulos and Sandholm, 2018, Bomze et al., 2019, Antonakopoulos et al., 2020]. Similar to their minimization counterparts, non-convex constraints have been widely applicable to the min-max optimization as well [Heusel et al., 2017, Daskalakis and Panageas, 2018b, Balduzzi et al., 2018, Mertikopoulos et al., 2019, Jin et al., 2020]. Recently there has been significant effort in proving tighter results either under more structured assumptions [Thekumparampil et al., 2019, Nouiehed et al., 2019, Lu et al., 2020, Azizian et al., 2020a, Diakonikolas, 2020, Golowich et al., 2020b, Lin et al., 2020d,c, Liu et al., 2021, Ostrovskii et al., 2021, Kong and Monteiro, 2021], and/or obtaining last-iterate convergence guarantees [Daskalakis and Panageas, 2018b, 2019, Mertikopoulos et al., 2019, Adolphs et al., 2019, Liang and Stokes, 2019, Gidel et al., 2019b, Mazumdar et al., 2020, Liu et al., 2020, Mokhtari et al., 2020b, Lin et al., 2020d, Hamedani and Aybat, 2021, Abernethy et al., 2021, Cai et al., 2022] for computing min-max solutions in convex-concave settings. Nonetheless, the analysis of the iteration complexity in the general *non-convex non-concave* setting is still in its infancy [Vlatakis-Gkaragkounis et al., 2019, 2021]. In response, the optimization community has recently studied how to extend standard min-max optimization algorithms such as gradient descent ascent (GDA) and extragradient (EG) to the Riemannian setting. In mathematical terms, given two Riemannian manifolds $\mathcal{M}, \mathcal{N}$ and a function $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$, the Riemannian min-max optimization (RMMO) problem becomes

$$\min_{\mathbf{x}\in\mathcal{M}} \max_{\mathbf{y}\in\mathcal{N}} f(\mathbf{x}, \mathbf{y}).$$

The change of geometry from Euclidean to Riemannian poses several difficulties. Indeed, a fundamental stumbling block has been that this problem may not even have theoretically meaningful solutions. In contrast with minimization where an optimal solution in a bounded domain is always guaranteed [Fearnley et al., 2021], existence of such saddle points necessitates typically the application of topological fixed point theorems [Brouwer, 1911, Kakutani, 1941], KKM Theory [Knaster et al., 1929]). For the case of convex-concave f with com-

pact sets $\mathcal{X}$ and $\mathcal{Y}$, Sion [1958] generalized the celebrated theorem [Neumann, 1928] and guaranteed that a solution $(\mathbf{x}^\star, \mathbf{y}^\star)$ with the following property exists

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \mathbf{y}^\star) = f(\mathbf{x}^\star, \mathbf{y}^\star) = \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star, \mathbf{y}).$$

However, at the core of the proof of this result is an ingenuous application of Helly's lemma [Helly, 1923] for the sublevel sets of f, and, until the work of Ivanov [2014], it has been unclear how to formulate an analogous lemma for the Riemannian geometry. As a result, until recently have extensions of the min-max theorem been established, and only for restricted manifold families [Komiya, 1988, Kristály, 2014, Park, 2019].

Zhang et al. [2022b] was the first to establish a min-max theorem for a flurry of Riemannian manifolds equipped with unique geodesics. Notice that this family is not a mathematical artifact since it encompasses many practical applications of RMMO, including Hadamard and Stiefel ones used in PCA [Lee et al., 2022]. Intuitively, the unique geodesic between two points of a manifold is the analogue of the a linear segment between two points in convex set: For any two points $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, their connecting geodesic is the unique shortest path contained in $\mathcal{X}$ that connects them.

Even when the RMMO is well defined, transferring the guarantees of traditional min-max optimization algorithms like Gradient Ascent Descent (GDA) and Extra-Gradient (EG) to the Riemannian case is non-trivial. Intuitively speaking, in the Euclidean realm the main leitmotif of the last-iterate analyses the aforementioned algorithms is a proof that $\delta_t = \|\mathbf{x}_t - \mathbf{x}^\star\|^2$ is decreasing over time. To achieve this, typically the proof correlates δ_t and δ_{t-1} via a "square expansion," namely:

$$\underbrace{\|\mathbf{x}_{t-1} - \mathbf{x}^\star\|^2}_{\alpha^2} = \underbrace{\|\mathbf{x}_t - \mathbf{x}^\star\|^2}_{\beta^2} + \underbrace{\|\mathbf{x}_{t-1} - \mathbf{x}_t\|^2}_{\gamma^2} - \underbrace{2\langle \mathbf{x}_t - \mathbf{x}^\star, \mathbf{x}_{t-1} - \mathbf{x}_t\rangle}_{2\beta\gamma\cos(\hat{A})}. \tag{4.1}$$

Notice, however that the above expression relies strongly on properties of Euclidean geometry (and the flatness of the corresponding line), namely that the the lines connecting the three points x_t, x_{t-1} and x^* form a triangle; indeed, it is the generalization of the Pythagorean theorem, known also as the law of cosines, for the induced triangle $(ABC) := \{(\mathbf{x}_t, \mathbf{x}_{t-1}, x^\star)\}$. In a uniquely geodesic manifold such triangle may not belong to the manifold as discussed above. As a result, the difference of distances to the equilibrium using the geodesic paths $d^2_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star) - d^2_{\mathcal{M}}(\mathbf{x}_{t-1}, \mathbf{x}^\star)$ generally cannot be given in a closed form. The manifold's curvature controls how close these paths are to forming a Euclidean triangle. In fact, this so-called phenomenon of *distance distortion*, as it is typically called, was hypothesised by Zhang et al. [2022b, Section 4.2] to be the cause of exponential slowdowns when applying EG to RMMO problems when compared to their Euclidean counterparts.

Multiple attempts have been made to bypass this hurdle. Huang and Gao [2023] analyzed the Riemannian GDA (RGDA) for the non-convex non-concave setting. However, they do not present any last-iterate convergence results and, even in the average/best iterate setting, they only derive sub-optimal rates for the geodesic convex-concave setting due to the lack of

the machinery that convex analysis and optimization offers they derive sub-optimal rates for the geodesic convex-concave case, which is the problem of our interest. The analysis of Han et al. [2022] for Riemannian Hamiltonian Method (RHM), matches the rate of second-order methods in the Euclidean case. Although theoretically faster in terms of iterations, second-order methods are not preferred in practice since evaluating second order derivatives for optimization problems of thousands to millions of parameters quickly becomes prohibitive. Finally, Zhang et al. [2022b] leveraged the standard averaging output trick in EG to derive a sublinear convergence rate of $O(1/\epsilon)$ for the general geodesically convex-concave Riemannian framework. In addition, they conjectured that the use of a different method could close the exponential gap for the geodesically strongly-convex-strongly-convex scenario and its Euclidean counterpart.

Given this background, a crucial question underlying the potential for successful application of first-order algorithms to Riemannian settings is the following:

Is a performance gap necessary between Riemannian and Euclidean optimal convex-concave algorithms in terms of accuracy and the condition number?

Our aim in this paper is to provide an extensive analysis of the Riemannian counterparts of Euclidean optimal first-order methods adapted to the manifold-constrained setting. For the case of the smooth objectives, we consider the *Riemannian corrected extragradient* (RCEG) method while for non-smooth cases, we analyze the textbook *Riemannian gradient descent ascent* (RGDA) method. Our main results are summarized in the following table.

Alg: *RCEG*. Smooth setting with ℓ-*Lipschitz Gradient* (cf. Assumption 4.4.4, 4.5.1 and 4.5.2)

Perf. Measure	**Setting**	**Complexity**	**Theorem**
Last-Iterate	Det. GSCSC	$O\left(\kappa(\sqrt{\tau_0}+\frac{1}{\bar{\xi}_0})\log(\frac{1}{\epsilon})\right)$	**Thm.** 4.5.3
Last-Iterate	Stoc. GSCSC	$O\left(\kappa(\sqrt{\tau_0}+\frac{1}{\bar{\xi}_0})\log(\frac{1}{\epsilon})+\frac{\sigma^2\bar{\xi}_0}{\mu^2\epsilon}\log(\frac{1}{\epsilon})\right)$	**Thm.** 4.5.6
Avg-Iterate	Det. GCC	$O\left(\frac{\ell\sqrt{\tau_0}}{\epsilon}\right)$	[Zhang et al., 2022b, **Thm.1**]
Avg-Iterate	Stoc. GCC	$O\left(\frac{\ell\sqrt{\tau_0}}{\epsilon}+\frac{\sigma^2\bar{\xi}_0}{\epsilon^2}\right)$	**Thm.** 4.5.7
Alg: *RGDA*. Nonsmooth setting with L-*Lipschitz Function* (cf. Assumption 4.9.1 and 4.9.2)			
Last-Iterate	Det. GSCSC	$O\left(\frac{L^2\bar{\xi}_0}{\mu^2\epsilon}\right)$	**Thm.** 4.9.3
Last-Iterate	Stoc. GSCSC	$O\left(\frac{(L^2+\sigma^2)\bar{\xi}_0}{\mu^2\epsilon}\right)$	**Thm.** 4.9.6
Avg-Iterate	Det. GCC	$O\left(\frac{L^2\bar{\xi}_0}{\epsilon^2}\right)$	**Thm.** 4.9.4
Avg-Iterate	Stoc. GCC	$O\left(\frac{(L^2+\sigma^2)\bar{\xi}_0}{\epsilon^2}\right)$	**Thm.** 4.9.7

For the definition of acronyms, Det and Stoc stand for deterministic and stochastic, respectively. GSCSC and GCC stand for geodesically strongly-convex-strongly-concave (cf. Assumption 4.5.1 or Assumption 4.9.1) and geodesically convex-concave (cf. Assumption 4.5.2

or Assumption 4.9.2). Here $\epsilon \in (0,1)$ is the accuracy, L, ℓ the Lipschitzness of the objective and its gradient, $\kappa = \ell/\mu$ is the condition number of the function, where μ is the strong convexity parameter, $(\tau_0, \underline{\xi}_0, \overline{\xi}_0)$ are curvature parameters (cf. Assumption 4.4.4), and σ^2 is the variance of a Riemannian gradient estimator.

Our first main contribution is the derivation of a linear convergence rate for RCEG, answering the open conjecture of Zhang et al. [2022b] about the performance gap of single-loop extragradient methods. Indeed, while a direct comparison between $d^2_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star)$ and $d^2_{\mathcal{M}}(\mathbf{x}_{t-1}, \mathbf{x}^\star)$ is infeasible, we are able to establish a relationship between the iterates via appeal to the duality gap function and obtain a contraction in terms of $d^2_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star)$. In other words, the effect of Riemannian distance distortion is quantitative (the contraction ratio will depend on it) rather than qualitative (the geometric contraction still remains under a proper choice of constant stepsize). More specifically, we use $d^2_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star) + d^2_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star)$ and $d^2_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star) + d^2_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star)$ to bound a gap function defined by $f(\hat{\mathbf{x}}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \hat{\mathbf{y}}_t)$. Since the objective function is geodesically strongly-convex-strongly-concave, we have $f(\hat{\mathbf{x}}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \hat{\mathbf{y}}_t)$ is lower bounded by $\frac{\mu}{2}(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star)^2 + d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star)^2)$. Then, using the relationship between $(\mathbf{x}_t, \mathbf{y}_t)$ and $(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)$, we conclude the desired results in Theorem 4.5.3. Notably, our approach is not affected by the nonlinear geometry of the manifold.

Secondly, we endeavor to give a systematic analysis of aspects of the objective function, including its smoothness, its convexity and oracle access. As we shall see, similar to the Euclidean case, better finite-time convergence guarantees are connected with a geodesic smoothness condition. For the sake of completeness, we provide the analysis of Riemannian GDA for the full spectrum of stochasticity for the non-smooth case. More specifically, for the stochastic setting, the key ingredient to get the optimal convergence rate is to carefully select the step size such that the noise of the gradient estimator will not affect the final convergence rate significantly. As a highlight, such technique has been used for analyzing stochastic RCEG in the Euclidean setting [Kotsalis et al., 2022] and our analysis can be seen as the extension to the Riemannian setting. For the nonsmooth setting, the analysis is relatively simpler compared to smooth settings but we still need to deal with the issue caused by the nonlinear geometry of manifolds and the interplay between the distortion of Riemannian metrics, the gap function and the bounds of Lipschitzness of our bi-objective. Interestingly, the rates we derive are near optimal in terms of accuracy and condition number of the objective, and analogous to their Euclidean counterparts.

4.2 Related Works

The literature for the geometric properties of Riemannian Manifolds is immense and hence we cannot hope to survey them here; for an appetizer, we refer the reader to Burago et al. [2001] and Lee [2012] and references therein. On the other hand, as stated, it is not until recently that the long-run non-asymptotic behavior of optimization algorithms in Riemannian manifolds (even the smooth ones) has encountered a lot of interest.

Minimization on Riemannian manifolds. Many application problems can be formulated as the minimization or maximization of a smooth function over Riemannian manifold and has triggered a line of research on the extension of the classical first-order and second-order methods to Riemannian setting with asymptotic convergence to first-order stationary points in general [Absil et al., 2009]. Recent years have witnessed the renewed interests on nonasymptotic convergence analysis of solution methods. In particular, Boumal et al. [2019] proved the global sublinear convergence results for Riemannian gradient descent method and Riemannian trust region method, and further demonstrated that the Riemannian trust region method converges to a second-order stationary point in polynomial time; see also similar results in some other works [Kasai and Mishra, 2018, Hu et al., 2018, 2019]. We are also aware of recent works on problem-specific methods [Wen and Yin, 2013, Gao et al., 2018, Liu et al., 2019] and primal-dual methods [Zhang et al., 2020b].

Compared to the smooth counterpart, Riemannian nonsmooth optimization is harder and relatively less explored [Absil and Hosseini, 2019]. A few existing works focus on optimizing geodesically convex functions over Riemannian manifold with subgradient methods [Ferreira and Oliveira, 1998, Zhang and Sra, 2016, Bento et al., 2017]. In particular, Ferreira and Oliveira [1998] provided the first asymptotic convergence result while Zhang and Sra [2016] and [Bento et al., 2017] proved an nonasymptotic global convergence rate of $O(\epsilon^{-2})$ for Riemannian subgradient methods. Further, Ferreira and Oliveira [2002] assumed that the proximal mapping over Riemannian manifold is computationally tractable and proved the global sublinear convergence of Riemannian proximal point method. Focusing on optimization over Stiefel manifold, Chen et al. [2020] studied the composite objective function and proposed Riemannian proximal gradient method which only needs to compute the proximal mapping of nonsmooth component function over the tangent space of Stiefel manifold. Li et al. [2021] consider optimizing a weakly convex function over Stiefel manifold and proposed Riemannian subgradient methods that drive a near-optimal stationarity measure below ϵ within the number of iterations bounded by $O(\epsilon^{-4})$.

There are some results on stochastic optimization over Riemannian manifold. In particular, Bonnabel [2013] proved the first asymptotic convergence result for Riemannian stochastic gradient descent, which is extended by a line of subsequent works [Zhang et al., 2016, Tripuraneni et al., 2018, Becigneul and Ganea, 2019, Kasai et al., 2019]. If the Riemannian Hessian is not positive definite, some recent works have suggested frameworks to escape saddle points [Sun et al., 2019, Criscitiello and Boumal, 2019].

Min-Max optimization in Euclidean spaces. Focusing on solving specifically min-max problems, the algorithms under euclidean geometry have a very rich history in optimization that goes back at least to the original proximal point algorithms [Martinet, 1970, Rockafellar, 1976] for variational inequality (VI) problems; At a high level, if the objective function is Lipschitz and strictly convex-concave, the simple forward-backward schemes are known to converge – and if combined with a Polyak–Ruppert averaging scheme [Ruppert, 1988, Polyak

and Juditsky, 1992, Nemirovski et al., 2009], they achieve an $O(1/\epsilon^2)$ complexity[1] without the caveat of strictness [Bauschke and Combettes, 2011]. If, in addition, the objective admits Lipschitz continuous gradients, then the extragradient (EG) algorithm [Korpelevich, 1976] achieves trajectory convergence without strict monotonicity requirements, while the time-average iterate converges at $O(1/\epsilon)$ steps [Nemirovski, 2004]. Finally, if the problem is strongly convex-concave, forward-backward methods computes an ϵ-saddle point at $O(1/\epsilon)$ steps; and if the operator is also Lipschitz continuous, classical results in operator theory show that simple forward-backward methods suffice to achieve a linear convergence rate [Facchinei and Pang, 2007, Bauschke and Combettes, 2011].

Min-Max optimization on Riemannian manifolds. In the case of nonlinear geometry, the literature has been devoted on two different orthogonal axes: *a)* the existence of saddle point for min-max objective bi-functions and *b)* the design of algorithms for the computation of such points. For the existence of saddle point, a long line of recent work tried to generalize the seminal minima theorem for quasi-convex-quasi-concave problems of Sion [1958]. The crucial bottleneck of this generalization to Riemannian smooth manifolds had been the application of both Knaster–Kuratowski–Mazurkiewicz (KKM) theorem and Helly's theorem in non-flat spaces. Before Zhang et al. [2022b], the existence of saddle points had been identified for the special case of Hadamard manifolds [Komiya, 1988, Kristály, 2014, Bento et al., 2017, Park, 2019].

Similar with the existence results, initially the developed methods referred to the computation of singularities in monotone variational operators typically in hyperbolic Hadamard manifolds with negative curvature [Li et al., 2009]. More recently, Huang and Gao [2023] proposed a Riemannian gradient descent ascent method (RGDA), yet the analysis is restricted to $\mathcal{N}$ being a convex subset of the Euclidean space and $f(x, y)$ being strongly concave in y. It is worth mentioning that for the case Hadamard and generally hyperbolic manifolds, extra-gradient style algorithms have been proposed [Wang et al., 2010, Ferreira et al., 2005] in the literature, establishing mainly their asymptotic convergence. However it was not until recent Zhang et al. [2022b] that the riemannian correction trick has been analyzed for the case of the extra-gradient algorithm. Bearing in our mind the higher-order methods, Han et al. [2022] has recently proposed the Riemannian Hamiltonian Descent and versions of Newton's method for for geodesic convex geodesic concave functions. Since in this work, we focus only on first-order methods, we don't compare with the aforementioned Hamiltonian alternative since it incorporates always the extra computational burden of second-derivatives and hessian over a manifold.

[1]For the rest of presentation, we adopt the convention of presenting the *fine-grained complexity* performance measure for computing an $O(\epsilon)$-close solution instead of the *convergence rate* of a method. Thus a rate of the form $\|\mathbf{x}_t - \mathbf{x}^*\| \leq O(1/t^{1/p})$ typically corresponds to $O(1/\epsilon^p)$ gradient computations and the geometric rate $\|\mathbf{x}_t - \mathbf{x}^*\| \leq O(\exp(-\mu t))$ matches usually up with the $O(\ln(1/\epsilon))$ computational complexity.

4.3 Motivating Examples

We provide some examples of Riemannian min-max optimization to give a sense of their expressivity. Two of the examples are the generic models from the optimization literature [Ben-Tal et al., 2009, Absil et al., 2009, Hu et al., 2020] and the two others are the formulations of application problems arising from machine learning and data analytics [Pennec et al., 2006, Fletcher and Joshi, 2007, Lin et al., 2020a].

Example 4.3.1 (Riemannian optimization with nonlinear constraints) *We can consider a rather straightforward generalization of constrained optimization problem from Euclidean spaces to Riemannian manifolds [Bergmann and Herzog, 2019]. This formulation finds a wide range of real-world applications, e.g., non-negative principle component analysis, weighted max-cut and so on. Letting $\mathcal{M}$ be a finite-dimensional Riemannian manifold with unique geodesic, we focus on the following problem:*

$$\min_{\mathbf{x}\in\mathcal{M}} \; f(\mathbf{x}), \quad \text{s.t. } g(\mathbf{x}) \leq 0, \; h(\mathbf{x}) = 0,$$

where $g := (g_1, g_2, \ldots, g_m) : \mathcal{M} \mapsto \mathbb{R}^m$ and $h := (h_1, h_2, \ldots, h_n) : \mathcal{M} \mapsto \mathbb{R}^n$ are two mappings. Then, we can introduce the dual variables λ and μ and reformulate the aforementioned constrained optimization problem as follows,

$$\min_{\mathbf{x}\in\mathcal{M}} \max_{(\lambda,\mu)\in\mathbb{R}^m_+\times\mathbb{R}^n} \; f(\mathbf{x}) + \langle \lambda, g(\mathbf{x})\rangle + \langle \mu, h(\mathbf{x})\rangle.$$

Suppose that f and all of g_i are geodesically convex and smooth and h_i are geodesically linear, the above problem is a geodesic-convex-Euclidean-concave min-max optimization problem.

Example 4.3.2 (Distributionally robust Riemannian optimization) *Distributionally robust optimization (DRO) is an effective method to deal with the noisy data, adversarial data, and imbalanced data. We consider the problem of DRO over Riemannian manifold; indeed, given a set of data samples $\{\xi_i\}_{i=1}^N$, the problem of DRO over Riemannian manifold $\mathcal{M}$ can be written in the form of*

$$\min_{\mathbf{x}\in\mathcal{M}} \max_{\boldsymbol{p}\in\mathcal{S}} \; \sum_{i=1}^N p_i \ell(\mathbf{x};\xi_i) - \|\boldsymbol{p} - \tfrac{1}{N}\mathbf{1}\|^2,$$

where $\boldsymbol{p} = (p_1, p_2, \ldots, p_N)$ and $\mathcal{S} = \{\boldsymbol{p} \in \mathbb{R}^N : \sum_{i=1}^N p_i = 1, p_i \geq 0\}$. In general, $\ell(\mathbf{x};\xi_i)$ is denoted as the loss function over Riemannian manifold $\mathcal{M}$. If ℓ is geodesically convex and smooth, this is a geodesic-convex-Euclidean-concave min-max optimization problem.

Example 4.3.3 (Robust matrix Karcher mean problem) *We consider a robust version of classical matrix Karcher mean problem. More specifically, the Karcher mean of N symmetric positive definite matrices $\{A_i\}_{i=1}^N$ is defined as the matrix $X \in \mathcal{M} = \{X \in \mathbb{R}^{n\times n} :$*

$X \succ 0,\ X = X^\top\}$ that minimizes the sum of squared distance induced by the Riemannian metric:

$$d(X, Y) = \|\log(X^{-1/2} Y X^{-1/2})\|_F.$$

The loss function is thus defined by

$$f(X; \{A_i\}_{i=1}^N) = \sum_{i=1}^N (d(X, A_i))^2.$$

which is known to be nonconvex in Euclidean spaces but geodesically strongly convex. Then, the robust version of classical matrix Karcher mean problem is aiming at solving the following problem:

$$\min_{X \in \mathcal{M}} \max_{Y_i \in \mathcal{M}} \ f(X; \{Y_i\}_{i=1}^N) - \gamma \left(\sum_{i=1}^N (d(Y_i, A_i))^2 \right),$$

where $\gamma > 0$ stands for the trade-off between the computation of Karcher mean over a set of $\{Y_i\}_{i=1}^N$ and the difference between the observed samples $\{A_i\}_{i=1}^N$ and $\{Y_i\}_{i=1}^N$. It is clear that this is a geodesically strongly-convex-strongly-concave min-max optimization problem.

Example 4.3.4 (Projection robust optimal transport problem) *We consider the projection robust optimal transport (OT) problem – a robust variant of the OT problem – that achieves superior sample complexity bound [Lin et al., 2021a]. Let $\{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^d$ and $\{y_1, y_2, \ldots, y_n\} \subseteq \mathbb{R}^d$ denote sets of n atoms, and let $(r_1, r_2, \ldots, r_n)$ and $(c_1, c_2, \ldots, c_n)$ denote weight vectors. We define discrete probability measures $\mu = \sum_{i=1}^n r_i \delta_{x_i}$ and $\nu = \sum_{j=1}^n c_j \delta_{y_j}$. In this setting, the computation of the k-dimensional projection robust OT distance between μ and ν resorts to solving the following problem:*

$$\max_{U \in \mathrm{St}(d,k)} \min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^n \sum_{j=1}^n \pi_{i,j} \|U^\top x_i - U^\top y_j\|^2,$$

where $\mathrm{St}(d, k) = \{U \in \mathbb{R}^{d \times k} \mid U^\top U = I_k\}$ is a Stiefel manifold and $\Pi(r, c) = \{\pi \in \mathbb{R}_+^{n \times n} \mid \sum_{j=1}^n \pi_{ij} = r_i, \sum_{i=1}^n \pi_{ij} = c_j\}$ is a transportation polytope. It is worth mentioning that the above problem is a geodesically-nonconvex-Euclidean-concave min-max optimization problem with special structures, making the computation of stationary points tractable. While the global convergence guarantee for our algorithm does not apply, the above problem might be locally geodesically-convex-Euclidean-concave such that our algorithm with sufficiently good initialization works here.

In addition to these examples, it is worth mentioning that Riemannian min-max optimization problems contain all general min-max optimization problems in Euclidean spaces and all Riemannian minimization or maximization optimization problems. It is also an abstraction of many machine learning problems, e.g,. principle component analysis [Boumal and Absil, 2011], dictionary learning [Sun et al., 2016a,b], deep neural networks (DNNs) [Huang et al.,

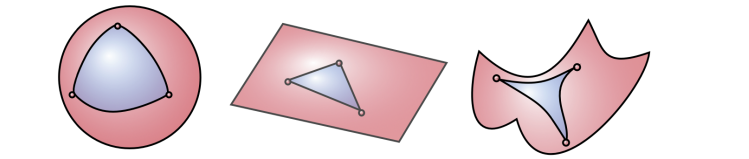
2018] and low-rank matrix learning [Vandereycken, 2013, Jawanpuria and Mishra, 2018]; indeed, the problem of principle component analysis resorts to optimization problems on Grassmann manifolds for example.

4.4 Preliminaries

We present the basic setup and optimality conditions for Riemannian min-max optimization. Indeed, we focus on some of key concepts that we need from Riemannian geometry, deferring a fuller presentation, including motivating examples and further discussion of related work, to the subsequent parts.

Riemannian geometry. An n-dimensional manifold $\mathcal{M}$ is a topological space where any point has a neighborhood that is homeomorphic to the n-dimensional Euclidean space. For each $\mathbf{x} \in \mathcal{M}$, each tangent vector is tangent to all parametrized curves passing through x and the tangent space $T_{\mathbf{x}}\mathcal{M}$ of a manifold $\mathcal{M}$ at this point is defined as the set of all tangent vectors. A Riemannian manifold $\mathcal{M}$ is a smooth manifold that is endowed with a smooth ("Riemannian") metric $\langle \cdot, \cdot \rangle_{\mathbf{x}}$ on the tangent space $T_{\mathbf{x}}\mathcal{M}$ for each point $\mathbf{x} \in \mathcal{M}$. The inner metric induces a norm $\| \cdot \|_{\mathbf{x}}$ on the tangent spaces.

A geodesic can be seen as the generalization of an Euclidean linear segment and is modeled as a smooth curve (map), $\gamma : [0, 1] \mapsto \mathcal{M}$, which is locally a distance minimizer. Additionally, because of the non-flatness of a manifold a different relation between the angles and the lengths of an arbitrary geodesic triangle is induced. This distortion can be quantified via the *sectional curvature* parameter $\kappa_{\mathcal{M}}$ thanks to Toponogov's theorem [Cheeger and Ebin, 1975, Burago et al., 1992]. A constructive consequence of this definition are the trigonometric



comparison inequalities (TCIs) that will be essential in our proofs; see Alimisis et al. [2020, Corollary 2.1] and Zhang and Sra [2016, Lemma 5] for detailed derivations. Assuming bounded sectional curvature, TCIs provide a tool for bounding Riemannian "inner products" that are more troublesome than classical Euclidean inner products.

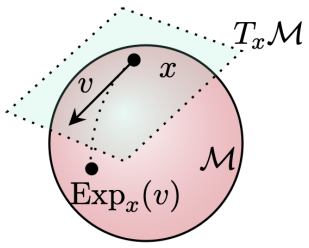
The following proposition summarizes the TCIs that we will need; note that if $\kappa_{\min} = \kappa_{\max} = 0$ (i.e., Euclidean spaces), then the proposition reduces to the law of cosines.

Proposition 4.4.1 *Suppose that $\mathcal{M}$ is a Riemannian manifold and let Δ be a geodesic triangle in $\mathcal{M}$ with the side length a, b, c and let A be the angle between b and c. Then, we have*

1. *If $\kappa_{\mathcal{M}}$ that is upper bounded by $\kappa_{\max} > 0$ and the diameter of $\mathcal{M}$ is bounded by $\frac{\pi}{\sqrt{\kappa_{\max}}}$, then*
$$a^2 \geq \underline{\xi}(\kappa_{\max}, c) \cdot b^2 + c^2 - 2bc\cos(A),$$
where $\underline{\xi}(\kappa, c) := 1$ for $\kappa \leq 0$ and $\underline{\xi}(\kappa, c) := c\sqrt{\kappa}\cot(c\sqrt{\kappa}) < 1$ for $\kappa > 0$.
2. *If $\kappa_{\mathcal{M}}$ is lower bounded by $\kappa_{\min}$, then*
$$a^2 \leq \overline{\xi}(\kappa_{\min}, c) \cdot b^2 + c^2 - 2bc\cos(A),$$
where $\overline{\xi}(\kappa, c) := c\sqrt{-\kappa}\coth(c\sqrt{-\kappa}) > 1$ if $\kappa < 0$ and $\overline{\xi}(\kappa, c) := 1$ if $\kappa \geq 0$.

Also, in contrast to the Euclidean case, $\mathbf{x}$ and $\nabla_{\mathbf{x}} f(\mathbf{x})$ do not lie in the same space, since $\mathcal{M}$ and $T_{\mathbf{x}}\mathcal{M}$ respectively are distinct entities. The interplay between these dual spaces typically is carried out via the *exponential maps*. An exponential map at a point $\mathbf{x} \in \mathcal{M}$ is a mapping from the tangent space $T_{\mathbf{x}}\mathcal{M}$ to $\mathcal{M}$. In particular, $\mathbf{y} := \mathrm{Exp}_{\mathbf{x}}(\mathbf{v}) \in \mathcal{M}$ is defined such that there exists a geodesic $\gamma : [0, 1] \mapsto \mathcal{M}$ satisfying $\gamma(0) = \mathbf{x}$, $\gamma(1) = \mathbf{y}$ and $\gamma'(0) = \mathbf{v}$. The inverse map exists since the manifold has a unique geodesic between any two points, which we denote as $\mathrm{Exp}_{\mathbf{x}}^{-1} : \mathcal{M} \mapsto T_{\mathbf{x}}\mathcal{M}$. Accordingly, we have $d_{\mathcal{M}}(\mathbf{x}, \mathbf{y}) = \|\mathrm{Exp}_x^{-1}(\mathbf{y})\|_{\mathbf{x}}$ is the Riemannian distance induced by the exponential map.

Finally, in contrast again to Euclidean spaces, we cannot compare the tangent vectors at different points $\mathbf{x}, \mathbf{y} \in \mathcal{M}$ since these vectors lie in different tangent spaces. To resolve the issue, it suffices to define a transport mapping that moves a tangent vector along the geodesics and also preserves the length and Riemannian metric $\langle \cdot, \cdot \rangle_{\mathbf{x}}$; indeed, we can define a parallel transport $\Gamma_{\mathbf{x}}^{\mathbf{y}} : T_{\mathbf{x}}\mathcal{M} \mapsto T_{\mathbf{y}}\mathcal{M}$ such that the inner product between any $\mathbf{u}, \mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$ is preserved; i.e., $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{x}} = \langle \Gamma_{\mathbf{x}}^{\mathbf{y}}(\mathbf{u}), \Gamma_{\mathbf{x}}^{\mathbf{y}}(\mathbf{v}) \rangle_{\mathbf{y}}$.



Riemannian min-max optimization and function classes. We let $\mathcal{M}$ and $\mathcal{N}$ be Riemannian manifolds with unique geodesic and bounded sectional curvature and assume that the function $f : \mathcal{M} \times \mathcal{N} \mapsto \mathbb{R}$ is defined on the product of these manifolds. The regularity conditions that we impose on the function f are as follows.

Definition 4.4.2 *A function $f : \mathcal{M} \times \mathcal{N} \mapsto \mathbb{R}$ is* geodesically L-Lipschitz *if for $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{M}$ and $\forall \mathbf{y}, \mathbf{y}' \in \mathcal{N}$, the following statement holds true: $|f(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}', \mathbf{y}')| \leq L(d_{\mathcal{M}}(\mathbf{x}, \mathbf{x}') + d_{\mathcal{N}}(\mathbf{y}, \mathbf{y}'))$. Additionally, if f is differentiable, it is called* geodesically ℓ-smooth *if for $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{M}$ and $\forall \mathbf{y}, \mathbf{y}' \in \mathcal{N}$, the following statement holds true,*

$$\begin{aligned} \|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) - \Gamma_{\mathbf{x}'}^{\mathbf{x}} \nabla_{\mathbf{x}} f(\mathbf{x}', \mathbf{y}')\| &\leq \ell(d_{\mathcal{M}}(\mathbf{x}, \mathbf{x}') + d_{\mathcal{N}}(\mathbf{y}, \mathbf{y}')), \\ \|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) - \Gamma_{\mathbf{y}'}^{\mathbf{y}} \nabla_{\mathbf{y}} f(\mathbf{x}', \mathbf{y}')\| &\leq \ell(d_{\mathcal{M}}(\mathbf{x}, \mathbf{x}') + d_{\mathcal{N}}(\mathbf{y}, \mathbf{y}')), \end{aligned}$$

where $(\nabla_{\mathbf{x}} f(\mathbf{x}', \mathbf{y}'), \nabla_{\mathbf{y}} f(\mathbf{x}', \mathbf{y}')) \in T_{\mathbf{x}'}\mathcal{M} \times T_{\mathbf{y}'}\mathcal{N}$ is the Riemannian gradient of f at $(\mathbf{x}', \mathbf{y}')$, $\Gamma_{\mathbf{x}'}^{\mathbf{x}}$ and $\Gamma_{\mathbf{y}'}^{\mathbf{y}}$ are the parallel transports of $\mathcal{M}$ from $\mathbf{x}'$ to $\mathbf{x}$ and of $\mathcal{N}$ from $\mathbf{y}'$ to $\mathbf{y}$, respectively.

Definition 4.4.3 *A function* $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ *is* geodesically strongly-convex-strongly-concave *with the modulus* $\mu > 0$ *if the following statement holds true,*

$$\begin{aligned} f(\mathbf{x}', \mathbf{y}) &\geq f(\mathbf{x}, \mathbf{y}) + \langle \partial_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), \mathrm{Exp}_{\mathbf{x}}^{-1}(\mathbf{x}') \rangle_{\mathbf{x}} + \tfrac{\mu}{2}(d_{\mathcal{M}}(\mathbf{x}, \mathbf{x}'))^2, \quad \text{for each } \mathbf{y} \in \mathcal{N}, \\ f(\mathbf{x}, \mathbf{y}') &\leq f(\mathbf{x}, \mathbf{y}) + \langle \partial_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}), \mathrm{Exp}_{\mathbf{y}}^{-1}(\mathbf{y}') \rangle_{\mathbf{y}} - \tfrac{\mu}{2}(d_{\mathcal{N}}(\mathbf{y}, \mathbf{y}'))^2, \quad \text{for each } \mathbf{x} \in \mathcal{M}. \end{aligned}$$

where $(\partial_{\mathbf{x}} f(\mathbf{x}', \mathbf{y}'), \partial_{\mathbf{y}} f(\mathbf{x}', \mathbf{y})) \in T_{\mathbf{x}'}\mathcal{M} \times T_{\mathbf{y}'}\mathcal{N}$ *is a Riemannian subgradient of* f *at a point* $(\mathbf{x}', \mathbf{y}')$*. A function* f *is* geodesically convex-concave *if the above holds true with* $\mu = 0$*.*

Following standard conventions in Riemannian optimization [Zhang and Sra, 2016, Alimisis et al., 2020, Zhang et al., 2022b], we make the following assumptions on the manifolds and objective functions:[2]

Assumption 4.4.4 *The objective function* $f : \mathcal{M} \times \mathcal{N} \mapsto \mathbb{R}$ *and manifolds* $\mathcal{M}$ *and* $\mathcal{N}$ *satisfy*

1. *The domain* $\{(\mathbf{x}, \mathbf{y}) \in \mathcal{M} \times \mathcal{N} : -\infty < f(\mathbf{x}, \mathbf{y}) < +\infty\}$ *is bounded by* $D > 0$*.*
2. $\mathcal{M}, \mathcal{N}$ *admit unique geodesic paths for any* $(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}') \in \mathcal{M} \times \mathcal{N}$*.*
3. *The sectional curvatures of* $\mathcal{M}$ *and* $\mathcal{N}$ *are both bounded in the range* $[\kappa_{\min}, \kappa_{\max}]$ *with* $\kappa_{\min} \leq 0$*. If* $\kappa_{\max} > 0$*, we assume that the diameter of manifolds is bounded by* $\frac{\pi}{\sqrt{\kappa_{\max}}}$*.*

Under the above conditions, Zhang et al. [2022b] proved an analog of the celebrated Sion's minimax theorem [Sion, 1958] in geodesic metric spaces. Formally, we have

$$\max_{\mathbf{y} \in \mathcal{N}} \min_{\mathbf{x} \in \mathcal{M}} f(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{x} \in \mathcal{M}} \max_{\mathbf{y} \in \mathcal{N}} f(\mathbf{x}, \mathbf{y}),$$

which guarantees that there exists at least one global saddle point $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{M} \times \mathcal{N}$ such that $\min_{\mathbf{x} \in \mathcal{M}} f(\mathbf{x}, \mathbf{y}^\star) = f(\mathbf{x}^\star, \mathbf{y}^\star) = \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}^\star, \mathbf{y})$. Note that the unicity of geodesics assumption is algorithm-independent and is imposed for guaranteeing that a saddle-point solution always exist. Even though this rules out many manifolds of interest, there are still many manifolds that satisfy such conditions. More specifically, the Hadamard manifold (manifolds with non-positive curvature, $\kappa_{\max} = 0$) has a unique geodesic between any two points. This also becomes a common regularity condition in Riemannian optimization [Zhang and Sra, 2016, Alimisis et al., 2020]. For any point $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{M} \times \mathcal{N}$, the duality gap $f(\hat{\mathbf{x}}, \mathbf{y}^\star) - f(\mathbf{x}^\star, \hat{\mathbf{y}})$ thus gives an optimality criterion.

Definition 4.4.5 *A point* $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{M} \times \mathcal{N}$ *is an* ϵ-saddle point *of a geodesically convex-concave function* $f(\cdot, \cdot)$ *if* $f(\hat{\mathbf{x}}, \mathbf{y}^\star) - f(\mathbf{x}^\star, \hat{\mathbf{y}}) \leq \epsilon$ *where* $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{M} \times \mathcal{N}$ *is a saddle point.*

[2] In particular, our assumed upper and lower bounds $\kappa_{\min}, \kappa_{\max}$ guarantee that TCIs in Proposition 4.4.1 can be used in our analysis for proving finite-time convergence.

In the setting where f is geodesically strongly-convex-strongly-concave with $\mu > 0$, it is not difficult to verify the uniqueness of a global saddle point $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{M} \times \mathcal{N}$. Then, we can consider the distance gap $(d(\hat{\mathbf{x}}, \mathbf{x}^\star))^2 + (d(\hat{\mathbf{y}}, \mathbf{y}^\star))^2$ as an optimality criterion for any point $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{M} \times \mathcal{N}$.

Definition 4.4.6 *A point $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{M} \times \mathcal{N}$ is an ϵ-saddle point of a geodesically strongly-convex-strongly-concave function $f(\cdot, \cdot)$ if $(d(\hat{\mathbf{x}}, \mathbf{x}^\star))^2 + (d(\hat{\mathbf{y}}, \mathbf{y}^\star))^2 \leq \epsilon$, where $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{M} \times \mathcal{N}$ is a saddle point. If f is geodesically ℓ-smooth, we denote $\kappa = \ell/\mu$ as the condition number.*

Given the above definitions, we can ask whether it is possible to find an ϵ-saddle point efficiently or not. In this context, Zhang et al. [2022b] have answered this question in the affirmative for the setting where f is geodesically ℓ-smooth and geodesically convex-concave; indeed, they derive the convergence rate of Riemannian corrected extragradient (RCEG) method in terms of time-average iterates and also conjecture that *RCEG does not guarantee convergence at a linear rate in terms of last iterates when f is geodesically ℓ-smooth and geodesically strongly-convex-strongly-concave, due to existence of distance distortion*; see Zhang et al. [2022b, Section 4.2]. Surprisingly, we show that RCEG with constant stepsize can achieve last-iterate convergence at a linear rate. Moreover, we establish the optimal convergence rates of stochastic RCEG for certain choices of stepsize for both geodesically convex-concave and geodesically strongly-convex-strongly-concave settings.

4.5 Riemannian Corrected Extragradient Method

We revisit the scheme of Riemannian corrected extragradient (RCEG) method proposed by Zhang et al. [2022b] and extend it to a stochastic algorithm that we refer to as *stochastic RCEG*. We present our main results on an optimal last-iterate convergence guarantee for the geodesically strongly-convex-strongly-concave setting (both deterministic and stochastic) and a time-average convergence guarantee for the geodesically convex-concave setting (stochastic). This complements the time-average convergence guarantee for geodesically convex-concave setting (deterministic) [Zhang et al., 2022b, Theorem 4.1] and resolves an open problem posted in Zhang et al. [2022b, Section 4.2].

Algorithmic scheme. The recently proposed *Riemannian corrected extragradient* (RCEG) method [Zhang et al., 2022b] is a natural extension of the celebrated extragradient (EG) method to the Riemannian setting. It resembles that of EG in Euclidean spaces but employs a simple modification in the extrapolation step to accommodate the nonlinear geometry of Riemannian manifolds. Let us provide some intuition how such modifications work.

Algorithm 10 RCEG

Input: initial points $(\mathbf{x}_0, \mathbf{y}_0)$ and stepsizes $\eta > 0$.
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Query $(\mathbf{g}_{\mathbf{x}}^t, \mathbf{g}_{\mathbf{y}}^t) \leftarrow (\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t))$, the Riemannian gradient of f at a point $(\mathbf{x}_t, \mathbf{y}_t)$
  $\hat{\mathbf{x}}_t \leftarrow \mathrm{Exp}_{\mathbf{x}_t}(-\eta \cdot \mathbf{g}_{\mathbf{x}}^t)$.
  $\hat{\mathbf{y}}_t \leftarrow \mathrm{Exp}_{\mathbf{y}_t}(\eta \cdot \mathbf{g}_{\mathbf{y}}^t)$.
  Query $(\hat{\mathbf{g}}_{\mathbf{x}}^t, \hat{\mathbf{g}}_{\mathbf{y}}^t) \leftarrow (\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t))$, the Riemannian gradient of f at a point $(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)$
  $\mathbf{x}_{t+1} \leftarrow \mathrm{Exp}_{\hat{\mathbf{x}}_t}(-\eta \cdot \hat{\mathbf{g}}_{\mathbf{x}}^t + \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t))$.
  $\mathbf{y}_{t+1} \leftarrow \mathrm{Exp}_{\hat{\mathbf{y}}_t}(\eta \cdot \hat{\mathbf{g}}_{\mathbf{y}}^t + \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t))$.

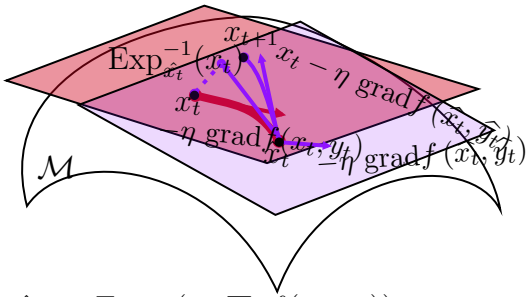
Algorithm 11 SRCEG

Input: initial points $(\mathbf{x}_0, \mathbf{y}_0)$ and stepsizes $\eta > 0$.
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Query $(\mathbf{g}_{\mathbf{x}}^t, \mathbf{g}_{\mathbf{y}}^t)$ as a **noisy** estimator of Riemannian gradient of f at a point $(\mathbf{x}_t, \mathbf{y}_t)$.
  $\hat{\mathbf{x}}_t \leftarrow \mathrm{Exp}_{\mathbf{x}_t}(-\eta \cdot \mathbf{g}_{\mathbf{x}}^t)$.
  $\hat{\mathbf{y}}_t \leftarrow \mathrm{Exp}_{\mathbf{y}_t}(\eta \cdot \mathbf{g}_{\mathbf{y}}^t)$.
  Query $(\hat{\mathbf{g}}_{\mathbf{x}}^t, \hat{\mathbf{g}}_{\mathbf{y}}^t)$ as a **noisy** estimator of Riemannian gradient of f at a point $(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)$.
  $\mathbf{x}_{t+1} \leftarrow \mathrm{Exp}_{\hat{\mathbf{x}}_t}(-\eta \cdot \hat{\mathbf{g}}_{\mathbf{x}}^t + \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t))$.
  $\mathbf{y}_{t+1} \leftarrow \mathrm{Exp}_{\hat{\mathbf{y}}_t}(\eta \cdot \hat{\mathbf{g}}_{\mathbf{y}}^t + \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t))$.

We start with a basic version of EG as follows, where $\mathcal{M}$ and $\mathcal{N}$ are classically restricted to be convex constraint sets in Euclidean spaces:

$$\begin{aligned} \hat{\mathbf{x}}_t &\leftarrow \mathcal{P}_{\mathcal{M}}(\mathbf{x}_t - \eta \cdot \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)), \quad & \hat{\mathbf{y}}_t &\leftarrow \mathcal{P}_{\mathcal{N}}(\mathbf{y}_t + \eta \cdot \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)), \\ \mathbf{x}_{t+1} &\leftarrow \mathcal{P}_{\mathcal{M}}(\mathbf{x}_t - \eta \cdot \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)), \quad & \mathbf{y}_{t+1} &\leftarrow \mathcal{P}_{\mathcal{N}}(\mathbf{y}_t + \eta \cdot \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)). \end{aligned} \tag{4.2}$$

Turning to the setting where $\mathcal{M}$ and $\mathcal{N}$ are Riemannian manifolds, the rather straightforward way to do the generalization is to replace the projection operator by the corresponding exponential map and the gradient by the corresponding Riemannian gradient. For the first line of Eq. (4.2), this approach works and leads to the following updates:



$$\hat{\mathbf{x}}_t \leftarrow \mathrm{Exp}_{\mathbf{x}_t}(-\eta \cdot \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)), \quad \hat{\mathbf{y}}_t \leftarrow \mathrm{Exp}_{\mathbf{y}_t}(\eta \cdot \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)).$$

However, we encounter some issues for the second line of Eq. (4.2): The above approach leads to problematic updates, $\mathbf{x}_{t+1} \leftarrow \mathrm{Exp}_{\mathbf{x}_t}(-\eta \cdot \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t))$ and $\mathbf{y}_{t+1} \leftarrow \mathrm{Exp}_{\mathbf{y}_t}(\eta \cdot \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t))$; indeed, the exponential maps $\mathrm{Exp}_{\mathbf{x}_t}(\cdot)$ and $\mathrm{Exp}_{\mathbf{y}_t}(\cdot)$ are defined from $T_{\mathbf{x}_t}\mathcal{M}$ to $\mathcal{M}$ and from $T_{\mathbf{y}_t}\mathcal{N}$ to $\mathcal{N}$ respectively. However, we have $-\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) \in T_{\hat{\mathbf{x}}_t}\mathcal{M}$ and $\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) \in T_{\hat{\mathbf{y}}_t}\mathcal{N}$. This motivates us to reformulate the second line of Eq. (4.2) as follows:

$$\mathbf{x}_{t+1} \leftarrow \mathcal{P}_{\mathcal{M}}(\hat{\mathbf{x}}_t - \eta \cdot \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + (\mathbf{x}_t - \hat{\mathbf{x}}_t)), \quad \mathbf{y}_{t+1} \leftarrow \mathcal{P}_{\mathcal{N}}(\hat{\mathbf{y}}_t + \eta \cdot \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + (\mathbf{y}_t - \hat{\mathbf{y}}_t)).$$

In the general setting with Riemannian manifolds, the terms $\mathbf{x}_t - \hat{\mathbf{x}}_t$ and $\mathbf{y}_t - \hat{\mathbf{y}}_t$ become $\mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t) \in T_{\hat{\mathbf{x}}_t}\mathcal{M}$ and $\mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t) \in T_{\hat{\mathbf{y}}_t}\mathcal{N}$. This observation yields the following updates:

$$\mathbf{x}_{t+1} \leftarrow \mathrm{Exp}_{\hat{\mathbf{x}}_t}(-\eta \cdot \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t)), \quad \mathbf{y}_{t+1} \leftarrow \mathrm{Exp}_{\hat{\mathbf{y}}_t}(\eta \cdot \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t)).$$

We then summarize the resulting RCEG method in Algorithm 10 and present the stochastic extension with noisy estimators of Riemannian gradients of f in Algorithm 11.

Main results. We present our main results on global convergence for Algorithms 10 and 11. To simplify the presentation, we treat separately the following two cases:

Assumption 4.5.1 *The function f is geodesically ℓ-smooth and geodesically strongly-convex-strongly-concave with $\mu > 0$.*

Assumption 4.5.2 *The function f is geodesically ℓ-smooth and geodesically convex-concave.*

Letting $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{M} \times \mathcal{N}$ be one global saddle point (it exists under either Assumption 4.5.1 or 4.5.2), we define $D_0 = (d_{\mathcal{M}}(\mathbf{x}_0, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_0, \mathbf{y}^\star))^2 > 0$ and $\kappa = \ell/\mu$ for the geodesically strongly-convex-strongly-concave setting. For ease of presentation, we also define a ratio $\tau(\cdot, \cdot)$ that measures how non-flatness changes in the spaces: $\tau([\kappa_{\min}, \kappa_{\max}], c) = \frac{\bar{\xi}(\kappa_{\min}, c)}{\underline{\xi}(\kappa_{\max}, c)} \geq 1$. We summarize our results for Algorithm 10 in the following theorem.

Theorem 4.5.3 *Given Assumptions 4.4.4 and 4.5.1, and letting $\eta = \min\{1/(2\ell\sqrt{\tau_0}), \underline{\xi}_0/(2\mu)\}$, there exists some $T > 0$ such that the output of Algorithm 10 satisfies that $(d(\mathbf{x}_T, \mathbf{x}^\star))^2 + (d(\mathbf{y}_T, \mathbf{y}^\star))^2 \leq \epsilon$ (i.e., an ϵ-saddle point of f in Definition 4.4.6) and the total number of Riemannian gradient evaluations is bounded by*

$$O\left(\left(\kappa\sqrt{\tau_0} + \frac{1}{\underline{\xi}_0}\right)\log\left(\frac{D_0}{\epsilon}\right)\right),$$

where $\tau_0 = \tau([\kappa_{\min}, \kappa_{\max}], D) \geq 1$ measures how non-flatness changes in $\mathcal{M}$ and $\mathcal{N}$ and $\underline{\xi}_0 = \underline{\xi}(\kappa_{\max}, D) \leq 1$ is properly defined in Proposition 4.4.1.

Remark 4.5.4 *Theorem 4.5.3 illustrates the last-iterate convergence of Algorithm 10 for solving geodesically strongly-convex-strongly-concave problems, thereby resolving an open problem delineated by Zhang et al. [2022b]. Further, the dependence on κ and $1/\epsilon$ cannot be improved since it matches the lower bound established for min-max optimization problems in Euclidean spaces [Zhang et al., 2022a]. However, we believe that the dependence on τ_0 and $\underline{\xi}_0$ is not tight, and it is of interest to either improve the rate or establish a lower bound for general Riemannian min-max optimization.*

Remark 4.5.5 *The current theoretical analysis covers local geodesic strong-convex-strong-concave settings. The key ingredient is how to define the local region; indeed, if we say the set of $\{(\mathbf{x}, \mathbf{y}) : d_{\mathcal{M}}(\mathbf{x}, \mathbf{x}^\star) \leq \delta, d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star) \leq \delta\}$ is a local region where the function is geodesically strong-convex-strong-concave. Then, the set of $\{(\mathbf{x}, \mathbf{y}) : (d_{\mathcal{M}}(\mathbf{x}, \mathbf{x}^\star)^2 + d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star)^2) \leq \delta^2\}$ must be contained in the above local region and the objective function is also geodesic strong-convex-strong-concave. If $(\mathbf{x}_0, \mathbf{y}_0) \in \{(\mathbf{x}, \mathbf{y}) : (d_{\mathcal{M}}(\mathbf{x}, \mathbf{x}^\star)^2 + d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star)^2) \leq \delta^2\}$, our theoretical analysis guarantees the last-iterate linear convergence rate. Such argument and definition of local region were standard for min-max optimization in the Euclidean setting; see Liang and Stokes [2019, Assumption 2.1].*

In the scheme of SRECG, we highlight that $(\mathbf{g}_{\mathbf{x}}^t, \mathbf{g}_{\mathbf{y}}^t)$ and $(\hat{\mathbf{g}}_{\mathbf{x}}^t, \hat{\mathbf{g}}_{\mathbf{y}}^t)$ are noisy estimators of Riemannian gradients of f at $(\mathbf{x}_t, \mathbf{y}_t)$ and $(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)$. It is necessary to impose the conditions such that these estimators are unbiased and has bounded variance. By abuse of notation, we assume that

$$\begin{aligned} \mathbf{g}_{\mathbf{x}}^t &= \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) + \xi_{\mathbf{x}}^t, \qquad \mathbf{g}_{\mathbf{y}}^t = \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) + \xi_{\mathbf{y}}^t, \\ \hat{\mathbf{g}}_{\mathbf{x}}^t &= \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + \hat{\xi}_{\mathbf{x}}^t, \qquad \hat{\mathbf{g}}_{\mathbf{y}}^t = \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + \hat{\xi}_{\mathbf{y}}^t. \end{aligned} \tag{4.3}$$

where the noises $(\xi_{\mathbf{x}}^t, \xi_{\mathbf{y}}^t)$ and $(\hat{\xi}_{\mathbf{x}}^t, \hat{\xi}_{\mathbf{y}}^t)$ are independent and satisfy that

$$\begin{aligned} &\mathbb{E}[\xi_{\mathbf{x}}^t] = 0, \quad \mathbb{E}[\xi_{\mathbf{y}}^t] = 0, \quad \mathbb{E}[\|\xi_{\mathbf{x}}^t\|^2 + \|\xi_{\mathbf{y}}^t\|^2] \leq \sigma^2, \\ &\mathbb{E}[\hat{\xi}_{\mathbf{x}}^t] = 0, \quad \mathbb{E}[\hat{\xi}_{\mathbf{y}}^t] = 0, \quad \mathbb{E}[\|\hat{\xi}_{\mathbf{x}}^t\|^2 + \|\hat{\xi}_{\mathbf{y}}^t\|^2] \leq \sigma^2. \end{aligned} \tag{4.4}$$

We are ready to summarize our results for Algorithm 11 in the following theorems.

Theorem 4.5.6 *Given Assumptions 4.4.4 and 4.5.1, letting Eq. (4.3) and Eq. (4.4) hold with $\sigma > 0$ and letting $\eta > 0$ satisfy $\eta = \min\{\frac{1}{24\ell\sqrt{\tau_0}}, \frac{\xi_0}{2\mu}, \frac{2(\log(T)+\log(\mu^2 D_0 \sigma^{-2}))}{\mu T}\}$, there exists some $T > 0$ so that the output of Algorithm 11 satisfies that $\mathbb{E}[(d(\mathbf{x}_T, \mathbf{x}^\star))^2 + (d(\mathbf{y}_T, \mathbf{y}^\star))^2] \leq \epsilon$ and the total number of noisy Riemannian gradient evaluations is bounded by*

$$O\left(\left(\kappa\sqrt{\tau_0} + \frac{1}{\underline{\xi}_0}\right)\log\left(\frac{D_0}{\epsilon}\right) + \frac{\sigma^2\overline{\xi}_0}{\mu^2\epsilon}\log\left(\frac{1}{\epsilon}\right)\right),$$

where $\tau_0 = \tau([\kappa_{\min}, \kappa_{\max}], D) \geq 1$ measures how non-flatness changes in $\mathcal{M}$ and $\mathcal{N}$ and $\underline{\xi}_0 = \underline{\xi}(\kappa_{\max}, D) \leq 1$ is properly defined in Proposition 4.4.1.

Theorem 4.5.7 *Given Assumptions 4.4.4 and 4.5.2 and assume that Eq. (4.3) and Eq. (4.4) hold with $\sigma > 0$ and let $\eta > 0$ satisfies that $\eta = \min\{\frac{1}{4\ell\sqrt{\tau_0}}, \frac{1}{\sigma}\sqrt{\frac{D_0}{\overline{\xi}_0 T}}\}$, there exists some $T > 0$ such that the output of Algorithm 11 satisfies that $\mathbb{E}[f(\bar{\mathbf{x}}_T, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}}_T)] \leq \epsilon$ and the total number of noisy Riemannian gradient evaluations is bounded by*

$$O\left(\frac{\ell D_0\sqrt{\tau_0}}{\epsilon} + \frac{\sigma^2\overline{\xi}_0}{\epsilon^2}\right),$$

where $\tau_0 = \tau([\kappa_{\min}, \kappa_{\max}], D)$ measures how non-flatness changes in $\mathcal{M}$ and $\mathcal{N}$ and $\overline{\xi}_0 = \overline{\xi}(\kappa_{\min}, D) \geq 1$ is properly defined in Proposition 4.4.1. The time-average iterates $(\bar{\mathbf{x}}_T, \bar{\mathbf{y}}_T) \in \mathcal{M} \times \mathcal{N}$ can be computed using the initial point $(\bar{\mathbf{x}}_0, \bar{\mathbf{y}}_0) = (0, 0)$ and the inductive formula: $\bar{\mathbf{x}}_{t+1} = \mathrm{Exp}_{\bar{\mathbf{x}}_t}(\frac{1}{t+1} \cdot \mathrm{Exp}_{\bar{\mathbf{x}}_t}^{-1}(\hat{\mathbf{x}}_t))$ and $\bar{\mathbf{y}}_{t+1} = \mathrm{Exp}_{\bar{\mathbf{y}}_t}(\frac{1}{t+1} \cdot \mathrm{Exp}_{\bar{\mathbf{y}}_t}^{-1}(\hat{\mathbf{y}}_t))$ for all $t = 0, 1, \ldots, T-1$.

Remark 4.5.8 *Theorem 4.5.6 presents the last-iterate convergence rate of Algorithm 11 for solving geodesically strongly-convex-strongly-concave problems while Theorem 4.5.7 gives the time-average convergence rate when the function f is only assumed to be geodesically convex-concave. Note that we carefully choose the stepsizes such that our upper bounds match the lower bounds established for stochastic min-max optimization problems in Euclidean spaces [Juditsky et al., 2011, Fallah et al., 2020, Kotsalis et al., 2022], in terms of the dependence on κ, $1/\epsilon$ and σ^2, up to log factors.*

Discussion: The last-iterate linear convergence rate in terms of Riemannian metrics is only limited to geodesically strongly convex-concave cases but other results, e.g., the average-iterate sublinear convergence rate, are derived under more mild conditions. This is consistent with classical results in the Euclidean setting where geodesic convexity reduces to convexity; indeed, the last-iterate linear convergence rate in terms of squared Euclidean norm is only obtained for strongly convex-concave cases. As such, our setting is not restrictive. Moreover, Zhang et al. [2022b] showed that the existence of a global saddle point is only guaranteed under the geodesically convex-concave assumption. For geodesically nonconvex-concave or geodesically nonconvex-nonconcave cases, a global saddle point might not exist and new optimality notions are required before algorithmic design. This question remains open in the Euclidean setting and is beyond the scope of this paper. However, we remark that an interesting class of robustification problems are nonconvex-nonconcave min-max problems in the Euclidean setting can be geodesically convex-concave in the Riemannian setting.

4.6 Experiments

We present numerical experiments on the task of robust principal component analysis (RPCA) for symmetric positive definite (SPD) matrices. In particular, we compare the performance of Algorithm 10 and 11 with different outputs, i.e., the last iterate $(\mathbf{x}_T, \mathbf{y}_T)$ versus the time-average iterate $(\bar{\mathbf{x}}_T, \bar{\mathbf{y}}_T)$ (see the precise definition in Theorem 4.5.7). Note that our implementations of both algorithms are based on the MANOPT package [Boumal et al., 2014]. All the experiments were implemented in MATLAB R2021b on a workstation with a 2.6 GHz Intel Core i7 and 16GB of memory.

Experimental setup. The problem of RPCA [Candès et al., 2011, Harandi et al., 2017] can be formulated as the Riemannian min-max optimization problem with an SPD manifold and a sphere manifold. Formally, we have

$$\max_{M \in \mathcal{M}_{\text{PSD}}^d} \min_{\mathbf{x} \in \mathcal{S}^d} \left\{ -\mathbf{x}^\top M \mathbf{x} - \tfrac{\alpha}{n} \sum_{i=1}^{n} d(M, M_i) \right\}. \tag{4.5}$$

In this formulation, $\alpha > 0$ denotes the penalty parameter, $\{M_i\}_{i \in [n]}$ is a sequence of given data SPD matrices, $\mathcal{M}_{\text{PSD}}^d = \{M \in \mathbb{R}^{d \times d} : M \succ 0, M = M^\top\}$ denotes the SPD manifold, $\mathcal{S}^d = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}$ denotes the sphere manifold and $d(\cdot, \cdot) : \mathcal{M}_{\text{PSD}}^d \times \mathcal{M}_{\text{PSD}}^d \mapsto \mathbb{R}$ is the Riemannian distance induced by the exponential map on the SPD manifold $\mathcal{M}_{\text{PSD}}^d$. As demonstrated by Zhang et al. [2022b], the problem of RPCA is nonconvex-nonconcave from a Euclidean perspective but is *locally geodesically strongly-convex-strongly-concave* and satisfies most of the assumptions that we make in this paper. In particular, the SPD manifold is complete with sectional curvature in $[-\frac{1}{2}, 1]$ [Criscitiello and Boumal, 2022] and the sphere manifold is complete with sectional curvature of 1. Other reasons why we use such example are: (i) it is a classical one in ML; (ii) Zhang et al. [2022b] also uses this example and

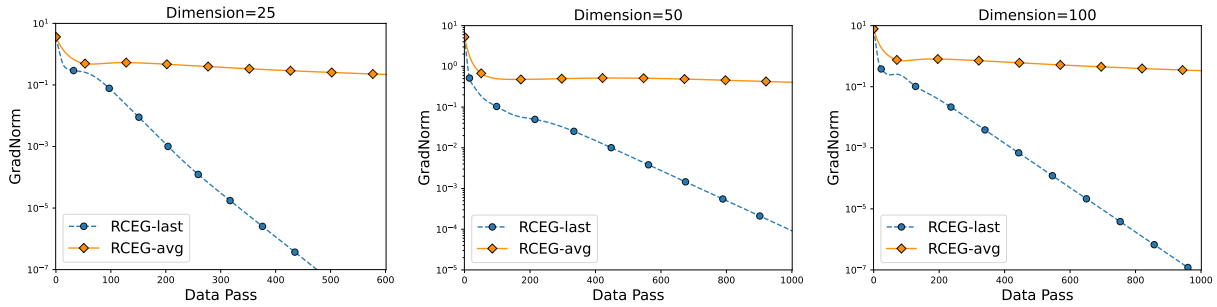


Figure 4.1: Comparison of last iterate (RCEG-last) and time-average iterate (RCEG-avg) for solving the RPCA problem in Eq. (4.5) with different problem dimensions $d \in \{25, 50, 100\}$. The horizontal axis represents the number of data passes and the vertical axis represents gradient norm.

observes the linear convergence behavior; (iii) the numerical results show that the unicity of geodesics assumption may not be necessary in practice; and (iv) this is an application where both min and max sides are done on Riemannian manifolds.

Following the previous works of Zhang et al. [2022b] and Han et al. [2022], we generate a sequence of data matrices M_i satisfying that their eigenvalues are in the range of $[0.2, 4.5]$. In our experiment, we fix $\alpha = 1.0$ and also vary the problem dimension $d \in \{25, 50, 100\}$. The evaluation metric is set as gradient norm. We set $n = 40$ and $n = 200$ in Figure 4.1 and 4.2. For RCEG, we set $\eta = \frac{1}{2\ell}$ where $\ell > 0$ is selected via grid search. For SRCEG, we set $\eta_t = \min\{\frac{1}{2\ell}, \frac{a}{t}\}$ where $\ell, a > 0$ are selected via grid search.

Experimental results. Figure 4.1 summarizes the effects of different outputs for RCEG; indeed, RCEG-last and RCEG-avg refer to Algorithm 10 with last iterate and time-average iterate respectively. It is clear that the last iterate of RCEG consistently exhibits linear convergence to an optimal solution in all the settings, verifying our theoretical results in Theorem 4.5.3. In contrast, the average iterate of RCEG converges much slower than the last iterate of RCEG. The possible reason is that the problem of RPCA is *only* locally geodesically strongly-convex-strongly-concave and averaging with the iterates generated during early stage will significantly slow down the convergence of RCEG.

Figure 4.2 presents the comparison between SRCEG (with either last iterate or time-average iterate) and RCEG with last-iterate; here, SRCEG-last and SRCEG-avg refer to Algorithm 11 with last iterate and time-average iterate respectively. We observe that SRCEG with either last iterate or average iterate converge faster than RCEG at the early stage and all of them finally converge to an optimal solution. This demonstrates the effectiveness and efficiency of SRCEG in practice. It is also worth mentioning that the difference between last-iterate convergence and time-average-iterate convergence is not as significant as in the deterministic setting. This is possibly because the technique of averaging help cancels the negative effect of imperfect information [Kingma and Ba, 2015, Yazıcı et al., 2019].

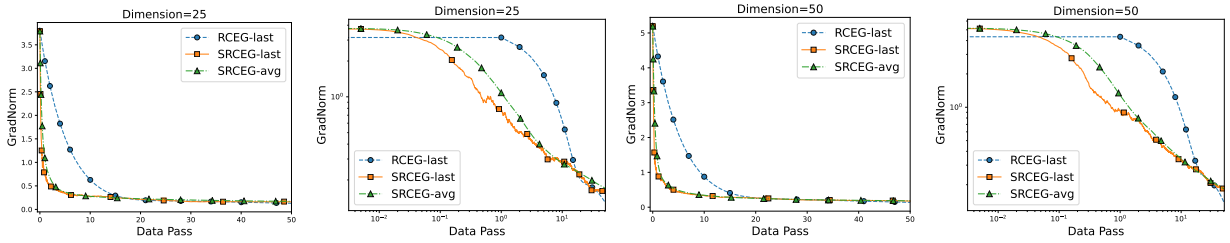


Figure 4.2: Comparison of RCEG and SRCEG for solving the RPCA problem in Eq. (4.5) with different problem dimensions $d \in \{25, 50\}$. The horizontal axis is the number of data passes and the vertical axis is gradient norm.

4.7 Conclusion

Inspired broadly by the structure of the complex competition that arises in many applications of robust optimization in ML, we focus on the problem of min-max optimization in the pure Riemannian setting (where both min and max player are constrained in a smooth manifold). Answering the open question of Zhang et al. [2022b] for the geodesically (strongly) convex-concave case, we showed that the Riemannian correction technique for EG matches the linear last-iterate complexity of their Euclidean counterparts in terms of accuracy and conditional number of objective for both deterministic and stochastic case. Additionally, we provide near-optimal guarantees for both smooth and non-smooth min-max optimization via Riemannian EG and GDA for the simple convex-concave case.

As a consequence of this work numerous open problems emerge; one immediate open question for future work is to explore whether the dependence on the curvature constant is also tight. Additionally, another generalization of interest would be to consider the performance of RCEG in the case of Riemannian Monotone Variational inequalities (RMVI) and examine the generalization of Zhang et al. [2022b] existence proof. Finally, there has been recent work in proving last-iterate convergence in the convex-concave setting via Sum-Of-Squares techniques [Cai et al., 2022]. It would be interesting to examine how one could leverage this machinery in a non-Euclidean but geodesic-metric-friendly framework.

4.8 Metric Geometry

To generalize the first-order methods in Euclidean setting, we introduce several basic concepts in metric geometry [Burago et al., 2001], which are known to include both Euclidean spaces and Riemannian manifolds as special cases. Formally, we have

Definition 4.8.1 (Metric Space) *A metric space (X, d) is a pair of a set X and a distance function $d(\cdot, \cdot)$ satisfying: (i) $d(\mathbf{x}, \mathbf{x}') \geq 0$ for any $\mathbf{x}, \mathbf{x}' \in X$; (ii) $d(\mathbf{x}, \mathbf{x}') = d(\mathbf{x}', \mathbf{x})$ for any $\mathbf{x}, \mathbf{x}' \in X$; and (iii) $d(\mathbf{x}, \mathbf{x}'') \leq d(\mathbf{x}, \mathbf{x}') + d(\mathbf{x}', \mathbf{x}'')$ for any $\mathbf{x}, \mathbf{x}', \mathbf{x}'' \in X$. In other words, the distance function $d(\cdot, \cdot)$ is non-negative, symmetrical and satisfies the triangle inequality.*

A *path* $\gamma : [0,1] \mapsto X$ is a continuous mapping from the interval $[0,1]$ to X and the *length* of γ is defined as $\text{length}(\gamma) := \lim_{n\to+\infty} \sup_{0=t_0<\ldots<t_n=1} \sum_{i=1}^{n} d(\gamma(t_{i-1}), \gamma(t_i))$. Note that the triangle inequality implies that $\sup_{0=t_0<\ldots<t_n=1} \sum_{i=1}^{n} d(\gamma(t_{i-1}), \gamma(t_i))$ is nondecreasing. Then, the length of a path γ is well defined since the limit is either $+\infty$ or a finite scalar. Moreover, for $\forall \epsilon > 0$, there exists $n \in \varsigma$ and the partition $0 = t_0 < \ldots < t_n = 1$ of the interval $[0,1]$ such that $\text{length}(\gamma) \leq \sum_{i=1}^{n} d(\gamma(t_{i-1}), \gamma(t_i)) + \epsilon$.

Definition 4.8.2 (Length Space) *A metric space (X,d) is a length space if, for any $\mathbf{x}, \mathbf{x}' \in X$ and $\epsilon > 0$, there exists $\gamma : [0,1] \mapsto X$ connecting $\mathbf{x}$ and $\mathbf{x}'$ such that* $\text{length}(\gamma) \leq d(x,x') + \epsilon$.

We can see from Definition 4.8.2 that a set of length spaces is strict subclass of metric spaces; indeed, for some $x, x' \in X$, there does not exist a path γ such that its length can be approximated by $d(x,x')$ for some tolerance $\epsilon > 0$. In metric geometry, a *geodesic* is a path which is locally a distance minimizer everywhere. More precisely, a path γ is a geodesic if there is a constant $\nu > 0$ such that for any $t \in [0,1]$ there is a neighborhood I of $[0,1]$ such that,

$$d(\gamma(t_1), \gamma(t_2)) = \nu |t_1 - t_2|, \quad \text{for any } t_1, t_2 \in I.$$

Note that the above generalizes the notion of geodesic for Riemannian manifolds. Then, we are ready to introduce the geodesic space and uniquely geodesic space [Bacak, 2014].

Definition 4.8.3 *A metric space (X,d) is a geodesic space if, for any $\mathbf{x}, \mathbf{x}' \in X$, there exists a geodesic $\gamma : [0,1] \mapsto X$ connecting $\mathbf{x}$ and $\mathbf{x}'$. Furthermore, it is called uniquely geodesic if the geodesic connecting $\mathbf{x}$ and $\mathbf{x}'$ is unique for any $\mathbf{x}, \mathbf{x}' \in X$.*

Trigonometric geometry in nonlinear spaces is intrinsically different from Euclidean space. In particular, we remark that the law of cosines in Euclidean space (with $\|\cdot\|$ as ℓ_2-norm) is crucial for analyzing the convergence property of optimization algorithms, e.g.,

$$\|a\|^2 = \|b\|^2 + \|c\|^2 - 2bc\cos(A),$$

where a, b, c are sides of *a geodesic triangle* in Euclidean space and A is the angle between b and c. However, such nice property does not hold for nonlinear spaces due to the lack of flat geometry, further motivating us to extend the law of cosines under nonlinear trigonometric geometry. That is to say, given a geodesic triangle in X with sides a, b, c where A is the angle between b and c, we hope to establish the relationship between a^2, b^2, c^2 and $2bc\cos(A)$ in nonlinear spaces; see the main context for the comparing inequalities.

Finally, we specify the definition of *section curvature* of Riemannian manifolds and clarify how such quantity affects the trigonometric comparison inequalities. More specifically, the sectional curvature is defined as the Gauss curvature of a 2-dimensional sub-manifold that are obtained from the image of a two-dimensional subspace of a tangent space after exponential mapping. It is worth mentioning that the above 2-dimensional sub-manifold is locally isometric to a 2-dimensional sphere, a Euclidean plane, and a hyperbolic plane with the same Gauss curvature if its sectional curvature is positive, zero and negative respectively. Then we

are ready to summarize the existing trigonometric comparison inequalities for Riemannian manifold with bounded sectional curvatures. Note that the following two propositions are the full version of Proposition 4.4.1 and will be used in our subsequent proofs.

Proposition 4.8.4 *Suppose that $\mathcal{M}$ is a Riemannian manifold with sectional curvature that is upper bounded by $\kappa_{\max}$ and let Δ be a geodesic triangle in $\mathcal{M}$ with the side length a, b, c and A which is the angle between b and c. If $\kappa_{\max} > 0$, we assume the diameter of $\mathcal{M}$ is bounded by $\frac{\pi}{\sqrt{\kappa_{\max}}}$. Then, we have*

$$a^2 \geq \underline{\xi}(\kappa_{\max}, c) \cdot b^2 + c^2 - 2bc\cos(A),$$

where $\underline{\xi}(\kappa, c) := 1$ for $\kappa \leq 0$ and $\underline{\xi}(\kappa, c) := c\sqrt{\kappa}\cot(c\sqrt{\kappa}) < 1$ for $\kappa > 0$.

Proposition 4.8.5 *Suppose that $\mathcal{M}$ is a Riemannian manifold with sectional curvature that is lower bounded by $\kappa_{\min}$ and let Δ be a geodesic triangle in $\mathcal{M}$ with the side length a, b, c and A which is the angle between b and c. Then, we have*

$$a^2 \leq \overline{\xi}(\kappa_{\min}, c) \cdot b^2 + c^2 - 2bc\cos(A),$$

where $\overline{\xi}(\kappa, c) := c\sqrt{-\kappa}\coth(c\sqrt{-\kappa}) > 1$ if $\kappa < 0$ and $\overline{\xi}(\kappa, c) := 1$ if $\kappa \geq 0$.

Remark 4.8.6 *Proposition 4.8.4 and 4.8.5 are simply the restatement of Alimisis et al. [2020, Corollary 2.1] and Zhang and Sra [2016, Lemma 5]. The former inequality is obtained when the sectional curvature is bounded from above while the latter inequality characterizes the relationship between the trigonometric lengths when the sectional curvature is bounded from below. If $\kappa_{\min} = \kappa_{\max} = 0$ (i.e., Euclidean spaces), we have $\overline{\xi}(\kappa_{\min}, c) = \underline{\xi}(\kappa_{\max}, c) = 1$. The proof is based on Toponogov's theorem and Riccati comparison estimate [Petersen, 2006, Proposition 25] and we refer the interested readers to Zhang and Sra [2016] and Alimisis et al. [2020] for the details.*

4.9 Riemannian Gradient Descent Ascent for Nonsmooth Setting

We propose and analyze Riemannian gradient descent ascent (RGDA) method for nonsmooth Riemannian min-max optimization and extend it to stochastic RGDA. We present our results on the optimal last-iterate convergence guarantee for geodesically strongly-convex-strongly-concave setting (both deterministic and stochastic) and time-average convergence guarantee for geodesically convex-concave setting (both deterministic and stochastic).

Algorithm 12 RGDA

Input: initial points $(\mathbf{x}_0, \mathbf{y}_0)$ and stepsizes $\eta_t > 0$.
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Query $(\mathbf{g}_{\mathbf{x}}^t, \mathbf{g}_{\mathbf{y}}^t) \leftarrow (\partial_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \partial_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t))$ as a Riemannian subgradient at a point $(\mathbf{x}_t, \mathbf{y}_t)$.
  $\mathbf{x}_{t+1} \leftarrow \mathrm{Exp}_{\mathbf{x}_t}(-\eta_t \cdot \mathbf{g}_{\mathbf{x}}^t)$.
  $\mathbf{y}_{t+1} \leftarrow \mathrm{Exp}_{\mathbf{y}_t}(\eta_t \cdot \mathbf{g}_{\mathbf{y}}^t)$.

Algorithm 13 SRGDA

Input: initial points $(\mathbf{x}_0, \mathbf{y}_0)$ and stepsizes $\eta_t > 0$.
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Query $(\mathbf{g}_{\mathbf{x}}^t, \mathbf{g}_{\mathbf{y}}^t)$ as a **noisy** estimator of Riemannian subgradient at a point $(\mathbf{x}_t, \mathbf{y}_t)$.
  $\mathbf{x}_{t+1} \leftarrow \mathrm{Exp}_{\mathbf{x}_t}(-\eta_t \cdot \mathbf{g}_{\mathbf{x}}^t)$.
  $\mathbf{y}_{t+1} \leftarrow \mathrm{Exp}_{\mathbf{y}_t}(\eta_t \cdot \mathbf{g}_{\mathbf{y}}^t)$.

Algorithmic scheme. Compared to Riemannian corrected extragradient (RCEG) method, our Riemannian gradient descent ascent (RGDA) method is a relatively straightforward generalization of GDA in Euclidean spaces. More specifically, we start with the scheme of GDA as follows (just consider $\mathcal{M}$ and $\mathcal{N}$ as convex constraint sets in Euclidean spaces),

$$\mathbf{x}_{t+1} \leftarrow \mathcal{P}_{\mathcal{M}}(\mathbf{x}_t - \eta_t \cdot \mathbf{g}_{\mathbf{x}}^t), \quad \mathbf{y}_{t+1} \leftarrow \mathcal{P}_{\mathcal{N}}(\mathbf{y}_t + \eta_t \cdot \mathbf{g}_{\mathbf{y}}^t). \tag{4.6}$$

where $(\mathbf{g}_{\mathbf{x}}^t, \mathbf{g}_{\mathbf{y}}^t) \in (\partial_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \partial_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t))$ is one subgradient. By replacing the projection operator with the corresponding exponential map and the gradient by the corresponding Riemannian gradient, we have

$$\mathbf{x}_{t+1} \leftarrow \mathrm{Exp}_{\mathbf{x}_t}(-\eta_t \cdot \mathbf{g}_{\mathbf{x}}^t), \quad \mathbf{y}_{t+1} \leftarrow \mathrm{Exp}_{\mathbf{y}_t}(\eta_t \cdot \mathbf{g}_{\mathbf{y}}^t).$$

where $(\mathbf{g}_{\mathbf{x}}^t, \mathbf{g}_{\mathbf{y}}^t) \leftarrow (\partial_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \partial_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t))$ is one Riemannian subgradient. We summarize the resulting scheme of RGDA method in Algorithm 12 and its stochastic extension with noisy estimators of Riemannian gradients of f in Algorithm 13.

Main results. We present our main results on the global convergence rate estimation for Algorithm 12 and 13 in terms of Riemannian gradient and noisy Riemannian gradient evaluations. The following assumptions are made throughout for geodesically strongly-convex-strongly-concave and geodesically convex-concave settings.

Assumption 4.9.1 *The objective function $f : \mathcal{M} \times \mathcal{N} \mapsto \mathbb{R}$ and manifolds $\mathcal{M}$ and $\mathcal{N}$ satisfy*

1. *f is geodesically L-Lipschitz and geodesically strongly-convex-strongly-concave with μ.*
2. *The domain $\{(\mathbf{x}, \mathbf{y}) \in \mathcal{M} \times \mathcal{N} : -\infty < f(\mathbf{x}, \mathbf{y}) < +\infty\}$ is bounded by $D > 0$.*
3. *The sectional curvatures of $\mathcal{M}$ and $\mathcal{N}$ are both bounded in the range $[\kappa_{\min}, +\infty)$ with $\kappa_{\min} \leq 0$.*

Assumption 4.9.2 *The objective function $f : \mathcal{M} \times \mathcal{N} \mapsto \mathbb{R}$ and manifolds $\mathcal{M}$ and $\mathcal{N}$ satisfy*

1. *f is geodesically L-Lipschitz and geodesically convex-concave.*
2. *The domain $\{(\mathbf{x}, \mathbf{y}) \in \mathcal{M} \times \mathcal{N} : -\infty < f(\mathbf{x}, \mathbf{y}) < +\infty\}$ is bounded by $D > 0$.*

3. *The sectional curvatures of $\mathcal{M}$ and $\mathcal{N}$ are both bounded in the range $[\kappa_{\min}, +\infty)$ with $\kappa_{\min} \leq 0$.*

Imposing the geodesically Lipschitzness condition is crucial to achieve finite-time convergence guarantee if we do not assume the geodesically smoothness condition. Note that we only require the lower bound for the sectional curvatures of manifolds and this is weaker than that presented in the main context.

Letting $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{M} \times \mathcal{N}$ be a global saddle point (it exists under either Assumption 4.9.1 or 4.9.2), we let $D_0 = (d_{\mathcal{M}}(\mathbf{x}_0, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_0, \mathbf{y}^\star))^2 > 0$ and summarize our results for Algorithm 12 in the following theorems.

Theorem 4.9.3 *Under Assumption 4.9.1 and let $\eta_t > 0$ satisfies that $\eta_t = \frac{1}{\mu}\min\{1, \frac{2}{t}\}$. There exists some $T > 0$ such that the output of Algorithm 12 satisfies that $(d(\mathbf{x}_T, \mathbf{x}^\star))^2 + (d(\mathbf{y}_T, \mathbf{y}^\star))^2 \leq \epsilon$ and the total number of Riemannian subgradient evaluations is bounded by*

$$O\left(\frac{\overline{\xi}_0 L^2}{\mu^2 \epsilon}\right),$$

where $\overline{\xi}_0 = \overline{\xi}(\kappa_{\min}, D)$ measures the lower bound for the non-flatness in $\mathcal{M}$ and $\mathcal{N}$.

Theorem 4.9.4 *Under Assumption 4.9.2 and let $\eta_t > 0$ satisfies that $\eta_t = \frac{1}{L}\sqrt{\frac{D_0}{2\overline{\xi}_0 T}}$. There exists some $T > 0$ such that the output of Algorithm 12 satisfies that $f(\bar{\mathbf{x}}_T, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}}_T) \leq \epsilon$ and the total number of Riemannian subgradient evaluations is bounded by*

$$O\left(\frac{\overline{\xi}_0 L^2 D_0}{\epsilon^2}\right),$$

where $\overline{\xi}_0 = \overline{\xi}(\kappa_{\min}, D)$ measures the lower bound for the non-flatness in $\mathcal{M}$ and $\mathcal{N}$, and the time-average iterates $(\bar{\mathbf{x}}_T, \bar{\mathbf{y}}_T) \in \mathcal{M} \times \mathcal{N}$ can be computed using the initial point $(\bar{\mathbf{x}}_0, \bar{\mathbf{y}}_0) = (0, 0)$ and the inductive formula: $\bar{\mathbf{x}}_{t+1} = \mathrm{Exp}_{\bar{\mathbf{x}}_t}(\frac{1}{t+1} \cdot \mathrm{Exp}^{-1}_{\bar{\mathbf{x}}_t}(\mathbf{x}_t))$ and $\bar{\mathbf{y}}_{t+1} = \mathrm{Exp}_{\bar{\mathbf{y}}_t}(\frac{1}{t+1} \cdot \mathrm{Exp}^{-1}_{\bar{\mathbf{y}}_t}(\mathbf{y}_t))$ for all $t = 0, 1, \ldots, T-1$.

Remark 4.9.5 *Theorem 4.9.3 and 4.9.4 establish the last-iterate and time-average rates of convergence of Algorithm 12 for solving Riemannian min-max optimization problems under Assumption 4.9.1 and 4.9.2 respectively. Further, the dependence on L and $1/\epsilon$ can not be improved since it has matched the lower bound established for the nonsmooth min-max optimization problems in Euclidean spaces.*

In the scheme of SRGDA, we highlight that $(\mathbf{g}^t_{\mathbf{x}}, \mathbf{g}^t_{\mathbf{y}})$ is a noisy estimators of Riemannian subgradient of f at $(\mathbf{x}_t, \mathbf{y}_t)$. It is necessary to impose the conditions such that these estimators are unbiased and has bounded variance. By abuse of notation, we assume that

$$\mathbf{g}^t_{\mathbf{x}} = \partial_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) + \xi^t_{\mathbf{x}}, \qquad \mathbf{g}^t_{\mathbf{y}} = \partial_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) + \xi^t_{\mathbf{y}}, \tag{4.7}$$

where the noises $(\xi_{\mathbf{x}}^t, \xi_{\mathbf{y}}^t)$ satisfy that

$$\mathbb{E}[\xi_{\mathbf{x}}^t] = 0, \qquad \mathbb{E}[\xi_{\mathbf{y}}^t] = 0, \qquad \mathbb{E}[\|\xi_{\mathbf{x}}^t\|^2 + \|\xi_{\mathbf{y}}^t\|^2] \leq \sigma^2. \tag{4.8}$$

We are ready to summarize our results for Algorithm 13 in the following theorems.

Theorem 4.9.6 *Under Assumption 4.9.1 and let Eq. (4.7) and Eq. (4.8) hold with $\sigma > 0$ and let $\eta_t > 0$ satisfies that $\eta_t = \frac{1}{\mu}\min\{1, \frac{2}{t}\}$. There exists some $T > 0$ such that the output of Algorithm 13 satisfies that $\mathbb{E}[(d(\mathbf{x}_T, \mathbf{x}^\star))^2 + (d(\mathbf{y}_T, \mathbf{y}^\star))^2] \leq \epsilon$ and the total number of noisy Riemannian gradient evaluations is bounded by*

$$O\left(\frac{\overline{\xi}_0(L^2 + \sigma^2)}{\mu^2\epsilon}\right),$$

where $\overline{\xi}_0 = \overline{\xi}(\kappa_{\min}, D)$ measures the lower bound for the non-flatness in $\mathcal{M}$ and $\mathcal{N}$.

Theorem 4.9.7 *Under Assumption 4.9.2 and let Eq. (4.7) and Eq. (4.8) hold with $\sigma > 0$ and let $\eta_t > 0$ satisfies that $\eta_t = \frac{1}{2}\sqrt{\frac{D_0}{\overline{\xi}_0(L^2+\sigma^2)T}}$. There exists some $T > 0$ such that the output of Algorithm 13 satisfies that $\mathbb{E}[f(\bar{\mathbf{x}}_T, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}}_T)] \leq \epsilon$ and the total number of noisy Riemannian gradient evaluations is bounded by*

$$O\left(\frac{\overline{\xi}_0(L^2 + \sigma^2)D_0}{\epsilon^2}\right),$$

where $\overline{\xi}_0 = \overline{\xi}(\kappa_{\min}, D)$ measures the lower bound for the non-flatness in $\mathcal{M}$ and $\mathcal{N}$, and the time-average iterates $(\bar{\mathbf{x}}_T, \bar{\mathbf{y}}_T) \in \mathcal{M} \times \mathcal{N}$ can be computed using the initial point $(\bar{\mathbf{x}}_0, \bar{\mathbf{y}}_0) = (0, 0)$ and the inductive formula: $\bar{\mathbf{x}}_{t+1} = \mathrm{Exp}_{\bar{\mathbf{x}}_t}(\frac{1}{t+1} \cdot \mathrm{Exp}_{\bar{\mathbf{x}}_t}^{-1}(\mathbf{x}_t))$ and $\bar{\mathbf{y}}_{t+1} = \mathrm{Exp}_{\bar{\mathbf{y}}_t}(\frac{1}{t+1} \cdot \mathrm{Exp}_{\bar{\mathbf{y}}_t}^{-1}(\mathbf{y}_t))$ for all $t = 0, 1, \ldots, T-1$.

Remark 4.9.8 *Theorem 4.9.6 and 4.9.7 establish the last-iterate and time-average rates of convergence of Algorithm 13 for solving Riemannian min-max optimization problems under Assumption 4.9.1 and 4.9.2. Moreover, the dependence on L and $1/\epsilon$ can not be improved since it has matched the lower bound established for nonsmooth stochastic min-max optimization problems in Euclidean spaces.*

4.10 Missing Proofs for Riemannian Corrected Extragradient Method

We present some technical lemmas for analyzing the convergence property of Algorithm 10 and 11. We also give the proofs of Theorem 4.5.3, 4.5.6 and 4.5.7.

Technical lemmas. We provide two technical lemmas for analyzing Algorithm 10 and 11 respectively. Parts of the first lemma were presented in Zhang et al. [2022b, Lemma C.1]. For the completeness, we provide the proof details.

Lemma 4.10.1 *Under Assumption 4.5.1 and let* $\{(\mathbf{x}_t, \mathbf{y}_t), (\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)\}_{t=0}^{T-1}$ *be generated by Algorithm 10 with the stepsize* $\eta > 0$*. Then, we have*

$$\begin{aligned} 0 \leq & \tfrac{1}{2}\left((d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right) \\ & + 2\overline{\xi}_0 \eta^2 \ell^2 ((d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2 - \tfrac{1}{2}\underline{\xi}_0 \left((d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2\right) \\ & - \tfrac{\mu\eta}{2}\left((d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2\right). \end{aligned}$$

where $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{M} \times \mathcal{N}$ *is a global saddle point of* f*.*

Proof. Since f is geodesically ℓ-smooth, we have the Riemannian gradients of f, i.e., $(\nabla_{\mathbf{x}} f, \nabla_{\mathbf{y}} f)$, are well defined. Since f is geodesically strongly-concave-strongly-concave with the modulus $\mu \geq 0$ (here $\mu = 0$ means that f is geodesically concave-concave), we have

$$\begin{aligned} & f(\hat{\mathbf{x}}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \hat{\mathbf{y}}_t) = f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) - f(\mathbf{x}^\star, \hat{\mathbf{y}}_t) - (f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) - f(\hat{\mathbf{x}}_t, \mathbf{y}^\star)) \\ & \overset{\text{Definition 4.4.3}}{\leq} -\langle \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle + \langle \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle - \tfrac{\mu}{2}(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 - \tfrac{\mu}{2}(d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2. \end{aligned}$$

Since $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{M} \times \mathcal{N}$ is a global saddle point of f, we have $f(\hat{\mathbf{x}}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \hat{\mathbf{y}}_t) \geq 0$. Recalling also from the scheme of Algorithm 10 that we have

$$\begin{aligned} \mathbf{x}_{t+1} &\leftarrow \mathrm{Exp}_{\hat{\mathbf{x}}_t}(-\eta \cdot \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t)), \\ \mathbf{y}_{t+1} &\leftarrow \mathrm{Exp}_{\hat{\mathbf{y}}_t}(\eta \cdot \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t)). \end{aligned}$$

By the definition of an exponential map, we have

$$\begin{aligned} \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_{t+1}) &= -\eta \cdot \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t), \\ \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_{t+1}) &= \eta \cdot \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t). \end{aligned} \tag{4.9}$$

This implies that

$$\begin{aligned} -\langle \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle &= \tfrac{1}{\eta}(\langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle - \langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle), \\ \langle \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle &= \tfrac{1}{\eta}(\langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle - \langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle). \end{aligned}$$

Putting these pieces together yields that

$$\begin{aligned} 0 \leq & \tfrac{1}{\eta}(\langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle - \langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle) - \tfrac{\mu}{2}(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 \\ & + \tfrac{1}{\eta}(\langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle - \langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle) - \tfrac{\mu}{2}(d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2. \end{aligned}$$

Equivalently, we have

$$\begin{aligned} 0 \leq \langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle - \langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle - \tfrac{\mu\eta}{2}(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 \\ + \langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle - \langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle - \tfrac{\mu\eta}{2}(d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2. \end{aligned} \tag{4.10}$$

It suffices to bound the terms in the right-hand side of Eq. (4.10) by leveraging the celebrated comparison inequalities on Riemannian manifold with bounded sectional curvature (see Proposition 4.8.4 and 4.8.5). More specifically, we define the constants using $\overline{\xi}(\cdot,\cdot)$ and $\underline{\xi}(\cdot,\cdot)$ from Proposition 4.8.4 and 4.8.5 as follows,

$$\overline{\xi}_0 = \overline{\xi}(\kappa_{\min}, D), \qquad \underline{\xi}_0 = \underline{\xi}(\kappa_{\max}, D).$$

By Proposition 4.8.4 and using that $\max\{d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star), d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star)\} \leq D$, we have

$$\begin{aligned} -\langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle &\leq -\tfrac{1}{2}\left(\underline{\xi}_0(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + (d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2\right), \\ -\langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle &\leq -\tfrac{1}{2}\left(\underline{\xi}_0(d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2\right). \end{aligned} \tag{4.11}$$

By Proposition 4.8.5 and using that $\max\{d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star), d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star)\} \leq D$, we have

$$\langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle \leq \tfrac{1}{2}\left(\overline{\xi}_0(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_{t+1}))^2 + (d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2\right).$$

and

$$\langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle \leq \tfrac{1}{2}\left(\overline{\xi}_0(d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_{t+1}))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right).$$

By the definition of an exponential map and Riemannian metric, we have

$$\begin{aligned} d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_{t+1}) &= \|\mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_{t+1})\| \overset{\text{Eq. (4.9)}}{=} \|\eta \cdot \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) - \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t)\|, \\ d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_{t+1}) &= \|\mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_{t+1})\| \overset{\text{Eq. (4.9)}}{=} \|\eta \cdot \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t)\|. \end{aligned} \tag{4.12}$$

Further, we see from the scheme of Algorithm 10 that we have

$$\begin{aligned} \hat{\mathbf{x}}_t &\leftarrow \mathrm{Exp}_{\mathbf{x}_t}(-\eta \cdot \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)), \\ \hat{\mathbf{y}}_t &\leftarrow \mathrm{Exp}_{\mathbf{y}_t}(\eta \cdot \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)). \end{aligned}$$

By the definition of an exponential map, we have

$$\mathrm{Exp}_{\mathbf{x}_t}^{-1}(\hat{\mathbf{x}}_t) = -\eta \cdot \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \qquad \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\hat{\mathbf{y}}_t) = \eta \cdot \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t).$$

Using the definition of a parallel transport map and the above equations, we have

$$\mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t) = \eta \cdot \Gamma_{\mathbf{x}_t}^{\hat{\mathbf{x}}_t} \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \qquad \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t) = -\eta \cdot \Gamma_{\mathbf{y}_t}^{\hat{\mathbf{y}}_t} \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)$$

Since f is geodesically ℓ-smooth, we have

$$\begin{aligned}\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) - \Gamma_{\mathbf{x}_t}^{\hat{\mathbf{x}}_t} \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\| &\leq \ell(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t) + d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t)),\\ \|\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) - \Gamma_{\mathbf{y}_t}^{\hat{\mathbf{y}}_t} \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)\| &\leq \ell(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t) + d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t)).\end{aligned}$$

Plugging the above inequalities into Eq. (4.12) yields that

$$\max\{d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_{t+1}), d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_{t+1})\} \leq \eta\ell(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t) + d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t)).$$

Therefore, we have

$$\begin{aligned}\langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle &\leq \tfrac{1}{2}\left(2\overline{\xi}_0\eta^2\ell^2((d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2) + (d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2\right),\\ \langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle &\leq \tfrac{1}{2}\left(2\overline{\xi}_0\eta^2\ell^2((d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2) + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right).\end{aligned}$$

Plugging these inequalities and Eq. (4.11) into Eq. (4.10) yields the desired inequality. □

The second lemma gives another key inequality that is satisfied by the iterates generated by Algorithm 11.

Lemma 4.10.2 *Under Assumption 4.5.1 (or Assumption 4.5.2) and the noisy model (cf. Eq. (4.3) and (4.4)) and let $\{(\mathbf{x}_t, \mathbf{y}_t), (\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)\}_{t=0}^{T-1}$ be generated by Algorithm 11 with the stepsize $\eta > 0$. Then, we have*

$$\begin{aligned}\mathbb{E}[f(\hat{\mathbf{x}}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \hat{\mathbf{y}}_t)] \leq\ & \tfrac{1}{2\eta}\mathbb{E}\left[(d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right]\\ &+6\overline{\xi}_0\eta\ell^2\mathbb{E}\left[(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2\right] - \tfrac{1}{2\eta}\underline{\xi}_0\mathbb{E}\left[(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2\right]\\ &-\tfrac{\mu}{2}\mathbb{E}\left[(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2\right] + 3\overline{\xi}_0\eta\sigma^2,\end{aligned}$$

where $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{M} \times \mathcal{N}$ is a global saddle point of f.

Proof. Using the same argument, we have ($\mu = 0$ refers to geodesically convex-concave case)

$$\begin{aligned}&f(\hat{\mathbf{x}}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \hat{\mathbf{y}}_t) = f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) - f(\mathbf{x}^\star, \hat{\mathbf{y}}_t) - (f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) - f(\hat{\mathbf{x}}_t, \mathbf{y}^\star))\\ &\leq\ -\langle \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle + \langle \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle - \tfrac{\mu}{2}(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 - \tfrac{\mu}{2}(d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2.\end{aligned}$$

Combining the arguments used in Lemma 4.10.1 and the scheme of Algorithm 11, we have

$$\begin{aligned}-\langle \hat{\mathbf{g}}_{\mathbf{x}}^t, \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle &= \tfrac{1}{\eta}(\langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle - \langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle),\\ \langle \hat{\mathbf{g}}_{\mathbf{y}}^t, \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle &= \tfrac{1}{\eta}(\langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle - \langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle).\end{aligned}$$

Putting these pieces together with Eq. (4.3) yields that

$$\begin{aligned}f(\hat{\mathbf{x}}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \hat{\mathbf{y}}_t) \leq\ & \tfrac{1}{\eta}(\langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle - \langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle) \qquad (4.13)\\ &+\tfrac{1}{\eta}(\langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle - \langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle) - \tfrac{\mu}{2}(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 - \tfrac{\mu}{2}(d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2\\ &+\langle \hat{\xi}_{\mathbf{x}}^t, \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle - \langle \hat{\xi}_{\mathbf{y}}^t, \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle.\end{aligned}$$

By the same argument as used in Lemma 4.10.1, we have

$$\begin{aligned} -\langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_t), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle &\leq -\tfrac{1}{2}\left(\underline{\xi}_0 (d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + (d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2\right), \\ -\langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_t), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle &\leq -\tfrac{1}{2}\left(\underline{\xi}_0 (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2\right), \end{aligned} \tag{4.14}$$

and

$$\begin{aligned} \langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle &\leq \tfrac{1}{2}\left(\overline{\xi}_0 \eta^2 \|\hat{\mathbf{g}}_{\mathbf{x}}^t - \Gamma_{\mathbf{x}_t}^{\hat{\mathbf{x}}_t} \mathbf{g}_{\mathbf{x}}^t\|^2 + (d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2\right), \\ \langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle &\leq \tfrac{1}{2}\left(\overline{\xi}_0 \eta^2 \|\hat{\mathbf{g}}_{\mathbf{y}}^t - \Gamma_{\mathbf{y}_t}^{\hat{\mathbf{y}}_t} \mathbf{g}_{\mathbf{y}}^t\|^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right). \end{aligned}$$

Since f is geodesically ℓ-smooth and Eq. (4.3) holds, we have

$$\begin{aligned} \|\hat{\mathbf{g}}_{\mathbf{x}}^t - \Gamma_{\mathbf{x}_t}^{\hat{\mathbf{x}}_t} \mathbf{g}_{\mathbf{x}}^t\|^2 &\leq 3\|\hat{\xi}_{\mathbf{x}}^t\|^2 + 3\|\xi_{\mathbf{x}}^t\|^2 + 6\ell^2 (d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + 6\ell^2 (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2, \\ \|\hat{\mathbf{g}}_{\mathbf{y}}^t - \Gamma_{\mathbf{y}_t}^{\hat{\mathbf{y}}_t} \mathbf{g}_{\mathbf{y}}^t\|^2 &\leq 3\|\hat{\xi}_{\mathbf{y}}^t\|^2 + 3\|\xi_{\mathbf{y}}^t\|^2 + 6\ell^2 (d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + 6\ell^2 (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2. \end{aligned}$$

Therefore, we have

$$\begin{aligned} &\langle \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle + \langle \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle \\ &\leq 6\overline{\xi}_0 \eta^2 \ell^2 ((d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2) + \tfrac{3}{2}\overline{\xi}_0 \eta^2 (\|\hat{\xi}_{\mathbf{x}}^t\|^2 + \|\xi_{\mathbf{x}}^t\|^2 + \|\hat{\xi}_{\mathbf{y}}^t\|^2 + \|\xi_{\mathbf{y}}^t\|^2) \\ &\quad + \tfrac{1}{2}\left((d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right). \end{aligned}$$

Plugging the above inequalities and Eq. (4.14) into Eq. (4.13) yields that

$$\begin{aligned} f(\hat{\mathbf{x}}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \hat{\mathbf{y}}_t) \leq\ & \tfrac{1}{2\eta}\left((d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right) \\ &+ 6\overline{\xi}_0 \eta \ell^2 ((d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2) + \tfrac{3}{2}\overline{\xi}_0 \eta (\|\hat{\xi}_{\mathbf{x}}^t\|^2 + \|\xi_{\mathbf{x}}^t\|^2 + \|\hat{\xi}_{\mathbf{y}}^t\|^2 + \|\xi_{\mathbf{y}}^t\|^2) \\ &- \tfrac{1}{2\eta}\underline{\xi}_0 \left((d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2\right) - \tfrac{\mu}{2}(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 - \tfrac{\mu}{2}(d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2 \\ &+ \langle \hat{\xi}_{\mathbf{x}}^t, \mathrm{Exp}_{\hat{\mathbf{x}}_t}^{-1}(\mathbf{x}^\star)\rangle - \langle \hat{\xi}_{\mathbf{y}}^t, \mathrm{Exp}_{\hat{\mathbf{y}}_t}^{-1}(\mathbf{y}^\star)\rangle. \end{aligned}$$

Taking the expectation of both sides and using Eq. (4.4) yields the desired inequality. □

Proof of Theorem 4.5.3. Since Riemannian metrics satisfy a triangle inequality, we have

$$(d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}^\star))^2 \geq \tfrac{1}{2}((d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2) - (d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2.$$

Plugging the above inequality into the inequality from Lemma 4.10.1 yields that

$$\begin{aligned} &(d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2 \\ &\leq \left(1 - \tfrac{\mu\eta}{2}\right)\left((d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2\right) + (4\overline{\xi}_0 \eta^2 \ell^2 + \mu\eta - \underline{\xi}_0)((d_{\mathcal{M}}(\hat{\mathbf{x}}_t, \mathbf{x}_t))^2 + (d_{\mathcal{N}}(\hat{\mathbf{y}}_t, \mathbf{y}_t))^2. \end{aligned}$$

Since $\eta = \min\{\frac{1}{4\ell\sqrt{\tau_0}}, \frac{\xi_0}{2\mu}\}$, we have $4\overline{\xi}_0 \eta^2 \ell^2 + \mu\eta - \underline{\xi}_0 \leq 0$. By the definition, we have $\tau_0 \geq 1$, $\kappa \geq 1$ and $\underline{\xi}_0 \leq 1$. This implies that

$$1 - \tfrac{\mu\eta}{2} = 1 - \min\left\{\tfrac{1}{8\kappa\sqrt{\tau_0}}, \tfrac{\xi_0}{4}\right\} > 0.$$

Putting these pieces together yields that

$$\begin{aligned}(d_{\mathcal{M}}(\mathbf{x}_t,\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\mathbf{y}_t,\mathbf{y}^\star))^2 &\leq \left(1-\min\left\{\tfrac{1}{8\kappa\sqrt{\tau_0}},\tfrac{\xi_0}{4}\right\}\right)^T (d_{\mathcal{M}}(\mathbf{x}_0,\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\mathbf{y}_0,\mathbf{y}^\star))^2\\ &\leq \left(1-\min\left\{\tfrac{1}{8\kappa\sqrt{\tau_0}},\tfrac{\xi_0}{4}\right\}\right)^T D_0.\end{aligned}$$

This completes the proof.

Proof of Theorem 4.5.6. Since Riemannian metrics satisfy a triangle inequality, we have

$(d_{\mathcal{M}}(\hat{\mathbf{x}}_t,\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\hat{\mathbf{y}}_t,\mathbf{y}^\star))^2 \geq \frac{1}{2}((d_{\mathcal{M}}(\mathbf{x}_t,\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\mathbf{y}_t,\mathbf{y}^\star))^2)-(d_{\mathcal{M}}(\hat{\mathbf{x}}_t,\mathbf{x}_t))^2+(d_{\mathcal{N}}(\hat{\mathbf{y}}_t,\mathbf{y}_t))^2.$

Plugging the above inequality into the inequality from Lemma 4.10.2 yields that

$$\begin{aligned}\mathbb{E}[f(\hat{\mathbf{x}}_t,\mathbf{y}^\star)-f(\mathbf{x}^\star,\hat{\mathbf{y}}_t)] &\leq \tfrac{1}{2\eta}\mathbb{E}\left[(d_{\mathcal{M}}(\mathbf{x}_t,\mathbf{x}^\star))^2-(d_{\mathcal{M}}(\mathbf{x}_{t+1},\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\mathbf{y}_t,\mathbf{y}^\star))^2-(d_{\mathcal{N}}(\mathbf{y}_{t+1},\mathbf{y}^\star))^2\right]\\ &\quad+(6\overline{\xi}_0\eta\ell^2+\tfrac{\mu}{2}-\tfrac{1}{2\eta}\underline{\xi}_0)\mathbb{E}\left[(d_{\mathcal{M}}(\hat{\mathbf{x}}_t,\mathbf{x}_t))^2+(d_{\mathcal{N}}(\hat{\mathbf{y}}_t,\mathbf{y}_t))^2\right]-\tfrac{\mu}{4}\mathbb{E}\left[(d_{\mathcal{M}}(\hat{\mathbf{x}}_t,\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\hat{\mathbf{y}}_t,\mathbf{y}^\star))^2\right]+3\overline{\xi}_0\eta\sigma^2.\end{aligned}$$

Since $(\mathbf{x}^\star,\mathbf{y}^\star)\in\mathcal{M}\times\mathcal{N}$ is a global saddle point of f, we have $\mathbb{E}[f(\hat{\mathbf{x}}_t,\mathbf{y}^\star)-f(\mathbf{x}^\star,\hat{\mathbf{y}}_t)]\geq 0$. Then, we have

$$\begin{aligned}\mathbb{E}\left[(d_{\mathcal{M}}(\mathbf{x}_{t+1},\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\mathbf{y}_{t+1},\mathbf{y}^\star))^2\right] &\leq \left(1-\tfrac{\mu\eta}{2}\right)\mathbb{E}\left[(d_{\mathcal{M}}(\mathbf{x}_t,\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\mathbf{y}_t,\mathbf{y}^\star))^2\right]\\ &\quad+(12\overline{\xi}_0\eta^2\ell^2+\mu\eta-\underline{\xi}_0)\mathbb{E}\left[(d_{\mathcal{M}}(\hat{\mathbf{x}}_t,\mathbf{x}_t))^2+(d_{\mathcal{N}}(\hat{\mathbf{y}}_t,\mathbf{y}_t))^2\right]+6\overline{\xi}_0\eta^2\sigma^2.\end{aligned}$$

Since $\eta\leq\min\{\frac{1}{24\ell\sqrt{\tau_0}},\frac{\xi_0}{2\mu}\}$, we have $12\overline{\xi}_0\eta^2\ell^2+\mu\eta-\underline{\xi}_0\leq 0$. This implies that

$$\mathbb{E}\left[(d_{\mathcal{M}}(\mathbf{x}_{t+1},\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\mathbf{y}_{t+1},\mathbf{y}^\star))^2\right]\leq\left(1-\tfrac{\mu\eta}{2}\right)\mathbb{E}\left[(d_{\mathcal{M}}(\mathbf{x}_t,\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\mathbf{y}_t,\mathbf{y}^\star))^2\right]+6\overline{\xi}_0\eta^2\sigma^2.$$

By the definition, we have $\tau_0\geq 1$, $\kappa\geq 1$ and $\underline{\xi}_0\leq 1$. This implies that

$$1-\tfrac{\mu\eta}{2}\geq 1-\min\left\{\tfrac{1}{48\kappa\sqrt{\tau_0}},\tfrac{\xi_0}{4}\right\}>0.$$

By the inductive arguments, we have

$$\begin{aligned}&\mathbb{E}\left[(d_{\mathcal{M}}(\mathbf{x}_t,\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\mathbf{y}_t,\mathbf{y}^\star))^2\right]\\ &\leq \left(1-\tfrac{\mu\eta}{2}\right)^T\left((d_{\mathcal{M}}(\mathbf{x}_0,\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\mathbf{y}_0,\mathbf{y}^\star))^2\right)+6\overline{\xi}_0\eta^2\sigma^2\left(\sum_{t=0}^{T-1}\left(1-\tfrac{\mu\eta}{2}\right)^t\right)\\ &\leq \left(1-\tfrac{\mu\eta}{2}\right)^T D_0+\tfrac{12\overline{\xi}_0\eta\sigma^2}{\mu}.\end{aligned}$$

Since $\eta=\min\{\frac{1}{24\ell\sqrt{\tau_0}},\frac{\xi_0}{2\mu},\frac{2(\log(T)+\log(\mu^2D_0\sigma^{-2}))}{\mu T}\}$, we have

$$\begin{aligned}\left(1-\tfrac{\mu\eta}{2}\right)^T D_0 &\leq \left(1-\min\left\{\tfrac{1}{48\kappa\sqrt{\tau_0}},\tfrac{\xi_0}{4}\right\}\right)^T D_0+\left(1-\tfrac{\log(\mu^2D_0\sigma^{-2}T)}{T}\right)^T D_0\\ &\overset{1+a\leq e^a}{\leq} \left(1-\min\left\{\tfrac{1}{48\kappa\sqrt{\tau_0}},\tfrac{\xi_0}{4}\right\}\right)^T D_0+\tfrac{\sigma^2}{\mu^2T},\end{aligned}$$

and

$$\frac{12\bar{\xi}_0\eta\sigma^2}{\mu} \leq \frac{24\bar{\xi}_0\sigma^2}{\mu^2 T}\log\left(\frac{\mu^2 D_0 T}{\sigma^2}\right).$$

Putting these pieces together yields that

$$\mathbb{E}\left[(d_{\mathcal{M}}(\mathbf{x}_t,\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\mathbf{y}_t,\mathbf{y}^\star))^2\right] \leq \left(1-\min\left\{\frac{1}{48\kappa\sqrt{\tau_0}},\frac{\xi_0}{4}\right\}\right)^T D_0+\frac{\sigma^2}{\mu^2 T}+\frac{24\bar{\xi}_0\sigma^2}{\mu^2 T}\log\left(\frac{\mu^2 D_0 T}{\sigma^2}\right).$$

This completes the proof.

Proof of Theorem 4.5.7. By the inductive formulas of $\bar{\mathbf{x}}_{t+1}=\operatorname{Exp}_{\bar{\mathbf{x}}_t}(\frac{1}{t+1}\cdot\operatorname{Exp}^{-1}_{\bar{\mathbf{x}}_t}(\hat{\mathbf{x}}_t))$ and $\bar{\mathbf{y}}_{t+1}=\operatorname{Exp}_{\bar{\mathbf{y}}_t}(\frac{1}{t+1}\cdot\operatorname{Exp}^{-1}_{\bar{\mathbf{y}}_t}(\hat{\mathbf{y}}_t))$ and using Zhang et al. [2022b, Lemma C.2], we have

$$f(\bar{\mathbf{x}}_T,\mathbf{y}^\star)-f(\mathbf{x}^\star,\bar{\mathbf{y}}_T)\leq\frac{1}{T}\left(\sum_{t=0}^{T-1}f(\hat{\mathbf{x}}_t,\mathbf{y}^\star)-f(\mathbf{x}^\star,\hat{\mathbf{y}}_t)\right).$$

Plugging the above inequality into the inequality from Lemma 4.10.2 yields that (recall that $\mu=0$ in geodesically convex-concave setting here)

$$\begin{aligned}\mathbb{E}[f(\bar{\mathbf{x}}_T,\mathbf{y}^\star)-f(\mathbf{x}^\star,\bar{\mathbf{y}}_T)] &\leq \frac{1}{2\eta T}\left((d_{\mathcal{M}}(\mathbf{x}_0,\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\mathbf{y}_0,\mathbf{y}^\star))^2\right)\\ &+\frac{1}{T}\left(6\bar{\xi}_0\eta\ell^2-\frac{1}{2\eta}\underline{\xi}_0\right)\left(\sum_{t=0}^{T-1}\mathbb{E}\left[(d_{\mathcal{M}}(\hat{\mathbf{x}}_t,\mathbf{x}_t))^2+(d_{\mathcal{N}}(\hat{\mathbf{y}}_t,\mathbf{y}_t))^2\right]\right)+3\bar{\xi}_0\eta\sigma^2.\end{aligned}$$

Since $\eta\leq\frac{1}{4\ell\sqrt{\tau_0}}$, we have $6\bar{\xi}_0\eta\ell^2-\frac{1}{2\eta}\underline{\xi}_0\leq 0$. This together with $(d_{\mathcal{M}}(\mathbf{x}_0,\mathbf{x}^\star))^2+(d_{\mathcal{N}}(\mathbf{y}_0,\mathbf{y}^\star))^2\leq D_0$ implies that

$$\mathbb{E}[f(\bar{\mathbf{x}}_T,\mathbf{y}^\star)-f(\mathbf{x}^\star,\bar{\mathbf{y}}_T)]\leq\frac{D_0}{2\eta T}+3\bar{\xi}_0\eta\sigma^2.$$

Since $\eta=\min\{\frac{1}{4\ell\sqrt{\tau_0}},\frac{1}{\sigma}\sqrt{\frac{D_0}{\bar{\xi}_0 T}}\}$, we have

$$\frac{D_0}{2\eta T}\leq\frac{2\ell D_0\sqrt{\tau_0}}{T}+\frac{\sigma}{2}\sqrt{\frac{\bar{\xi}_0 D_0}{T}},$$

and

$$3\bar{\xi}_0\eta\sigma^2\leq 3\sigma\sqrt{\frac{\bar{\xi}_0 D_0}{T}}.$$

Putting these pieces together yields that

$$\mathbb{E}[f(\bar{\mathbf{x}}_T,\mathbf{y}^\star)-f(\mathbf{x}^\star,\bar{\mathbf{y}}_T)]\leq\frac{2\ell D_0\sqrt{\tau_0}}{T}+\frac{7\sigma}{2}\sqrt{\frac{\bar{\xi}_0 D_0}{T}}.$$

This completes the proof.

4.11 Missing Proofs for Riemannian Gradient Descent Ascent

We present some technical lemmas for analyzing the convergence property of Algorithm 12 and 13. We also give the proofs of Theorem 4.9.3, 4.9.4, 4.9.6 and 4.9.7.

Technical lemmas. We provide two technical lemmas for analyzing Algorithm 12 and 13 respectively. The first lemma gives a key inequality that is satisfied by the iterates generated by Algorithm 12.

Lemma 4.11.1 *Under Assumption 4.9.1 (or Assumption 4.9.2) and let* $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=0}^{T-1}$ *be generated by Algorithm 12 with the stepsize* $\eta_t > 0$*. Then, we have*

$$\begin{aligned} f(\mathbf{x}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_t) \leq \tfrac{1}{2\eta_t}\left((d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2\right) \\ + \tfrac{1}{2\eta_t}\left((d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right) - \tfrac{\mu}{2}(d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - \tfrac{\mu}{2}(d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 + \overline{\xi}_0 \eta_t L^2, \end{aligned}$$

where $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{M} \times \mathcal{N}$ *is a global saddle point of* f*.*

Proof. Since f is geodesically strongly-concave-strongly-concave with the modulus $\mu \geq 0$ (here $\mu = 0$ means that f is geodesically concave-concave), we have

$$\begin{aligned} & f(\mathbf{x}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_t) = f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}^\star, \mathbf{y}_t) - (f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}_t, \mathbf{y}^\star)) \\ & \leq \; -\langle \partial_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}^\star)\rangle + \langle \partial_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t), \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}^\star)\rangle - \tfrac{\mu}{2}(d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - \tfrac{\mu}{2}(d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2. \end{aligned}$$

Recalling also from the scheme of Algorithm 12 that we have

$$\begin{aligned} \mathbf{x}_{t+1} &\leftarrow \mathrm{Exp}_{\mathbf{x}_t}(-\eta_t \cdot \partial_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)), \\ \mathbf{y}_{t+1} &\leftarrow \mathrm{Exp}_{\mathbf{y}_t}(\eta_t \cdot \partial_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)). \end{aligned}$$

By the definition of an exponential map, we have

$$\begin{aligned} \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}_{t+1}) &= -\eta_t \cdot \partial_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \\ \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}_{t+1}) &= \eta_t \cdot \partial_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t). \end{aligned} \tag{4.15}$$

This implies that

$$\begin{aligned} -\langle \partial_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}^\star)\rangle &= \tfrac{1}{\eta_t}\langle \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}^\star)\rangle, \\ \langle \partial_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t), \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}^\star)\rangle &= \tfrac{1}{\eta_t}\langle \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}^\star)\rangle. \end{aligned}$$

Putting these pieces together yields that

$$\begin{aligned} f(\mathbf{x}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_t) \leq \tfrac{1}{\eta_t}\langle \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}^\star)\rangle \\ + \tfrac{1}{\eta_t}\langle \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}^\star)\rangle - \tfrac{\mu}{2}(d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - \tfrac{\mu}{2}(d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2. \end{aligned} \tag{4.16}$$

It suffices to bound the terms in the right-hand side of Eq. (4.16) by leveraging the celebrated comparison inequalities on Riemannian manifold with lower bounded sectional curvature (see Proposition 4.8.5). More specifically, we define the constants using $\overline{\xi}(\cdot,\cdot)$ and $\underline{\xi}(\cdot,\cdot)$ from Proposition 4.8.5 as follows,

$$\overline{\xi}_0 = \overline{\xi}(\kappa_{\min}, D).$$

By Proposition 4.8.5 and using that $\max\{d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star), d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star)\} \le D$, we have

$$\begin{aligned}
\langle \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}^\star)\rangle &\le \tfrac{1}{2}\left(\overline{\xi}_0 (d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}_{t+1}))^2 + (d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2\right), \\
\langle \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}^\star)\rangle &\le \tfrac{1}{2}\left(\overline{\xi}_0 (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}_{t+1}))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right).
\end{aligned}$$

Since f is geodesically L-Lipschitz, we have

$$\|\partial_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\| \le L, \quad \|\partial_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)\| \le L.$$

By the definition of an exponential map and Riemannian metric, we have

$$\begin{aligned}
d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}_{t+1}) &= \|\mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}_{t+1})\| \overset{\text{Eq. (4.15)}}{=} \|\eta_t \cdot \partial_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\| \le \eta_t L, \\
d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}_{t+1}) &= \|\mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}_{t+1})\| \overset{\text{Eq. (4.15)}}{=} \|\eta_t \cdot \partial_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)\| \le \eta_t L.
\end{aligned}$$

Putting these pieces together yields that

$$\begin{aligned}
\langle \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}^\star)\rangle &\le \tfrac{1}{2}\left(\overline{\xi}_0 \eta_t^2 L^2 + (d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2\right), \\
\langle \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}^\star)\rangle &\le \tfrac{1}{2}\left(\overline{\xi}_0 \eta_t^2 L^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right).
\end{aligned}$$

Plugging the above inequalities into Eq. (4.16) yields the desired inequality. □

The second lemma gives another key inequality that is satisfied by the iterates generated by Algorithm 13.

Lemma 4.11.2 *Under Assumption 4.9.1 (or Assumption 4.9.2) and the noisy model (cf. Eq. (4.7) and (4.8)) and let $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=0}^{T-1}$ be generated by Algorithm 13 with the stepsize $\eta_t > 0$. Then, we have*

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{x}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_t)] \le \tfrac{1}{2\eta_t}\mathbb{E}\left[(d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2\right] \\
&\quad + \tfrac{1}{2\eta_t}\mathbb{E}\left[(d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right] - \tfrac{\mu}{2}\mathbb{E}\left[(d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2\right] + 2\overline{\xi}_0 \eta_t (L^2 + \sigma^2),
\end{aligned}$$

where $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{M} \times \mathcal{N}$ is a global saddle point of f.

Proof. Using the same argument, we have ($\mu = 0$ refers to geodesically convex-concave case)

$$\begin{aligned}
&f(\mathbf{x}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_t) = f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}^\star, \mathbf{y}_t) - (f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}_t, \mathbf{y}^\star)) \\
&\le -\langle \partial_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}^\star)\rangle + \langle \partial_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t), \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}^\star)\rangle - \tfrac{\mu}{2}(d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - \tfrac{\mu}{2}(d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2.
\end{aligned}$$

Combining the arguments used in Lemma 4.11.1 and the scheme of Algorithm 11, we have

$$\begin{aligned} -\langle \mathbf{g}_{\mathbf{x}}^t, \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}^\star)\rangle &= \tfrac{1}{\eta_t}\langle \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}^\star)\rangle, \\ \langle \mathbf{g}_{\mathbf{y}}^t, \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}^\star)\rangle &= \tfrac{1}{\eta_t}\langle \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}^\star)\rangle. \end{aligned}$$

Putting these pieces together with Eq. (4.7) yields that

$$\begin{aligned} f(\mathbf{x}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_t) \leq\ & \tfrac{1}{\eta_t}\langle \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}^\star) \\ & + \tfrac{1}{\eta_t}\langle \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}^\star)\rangle - \tfrac{\mu}{2}(d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - \tfrac{\mu}{2}(d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 \\ & + \langle \xi_{\mathbf{x}}^t, \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}^\star)\rangle - \langle \xi_{\mathbf{y}}^t, \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}^\star)\rangle. \end{aligned} \tag{4.17}$$

By the same argument as used in Lemma 4.11.1 and Eq. (4.7), we have

$$\begin{aligned} \langle \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}^\star)\rangle &\leq \tfrac{1}{2}\left(\bar{\xi}_0 (d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}_{t+1}))^2 + (d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2\right), \\ \langle \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}^\star)\rangle &\leq \tfrac{1}{2}\left(\bar{\xi}_0 (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}_{t+1}))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right), \end{aligned}$$

and

$$\begin{aligned} d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}_{t+1}) &= \|\mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}_{t+1})\| = \|\eta_t \cdot \mathbf{g}_{\mathbf{x}}^t\| \leq \eta_t(L + \|\xi_{\mathbf{x}}^t\|), \\ d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}_{t+1}) &= \|\mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}_{t+1})\| = \|\eta_t \cdot \mathbf{g}_{\mathbf{y}}^t\| \leq \eta_t(L + \|\xi_{\mathbf{y}}^t\|). \end{aligned}$$

Therefore, we have

$$\begin{aligned} \langle \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}_{t+1}), \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}^\star)\rangle + \langle \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}_{t+1}), \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}^\star)\rangle \leq\ & \tfrac{1}{2}\bar{\xi}_0\eta_t^2(4L^2 + 2\|\xi_{\mathbf{x}}^t\|^2 + 2\|\xi_{\mathbf{y}}^t\|^2) \\ & + \tfrac{1}{2}\left((d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right). \end{aligned}$$

Plugging the above inequalities into Eq. (4.17) yields that

$$\begin{aligned} f(\mathbf{x}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_t) \leq\ & \tfrac{1}{2\eta_t}\left((d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - (d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 - (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right) \\ & + \bar{\xi}_0\eta_t(2L^2 + \|\xi_{\mathbf{x}}^t\|^2 + \|\xi_{\mathbf{y}}^t\|^2) - \tfrac{\mu}{2}(d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 - \tfrac{\mu}{2}(d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 + \langle \xi_{\mathbf{x}}^t, \mathrm{Exp}_{\mathbf{x}_t}^{-1}(\mathbf{x}^\star)\rangle - \langle \xi_{\mathbf{y}}^t, \mathrm{Exp}_{\mathbf{y}_t}^{-1}(\mathbf{y}^\star)\rangle. \end{aligned}$$

Taking the expectation of both sides and using Eq. (4.8) yields the desired inequality. □

Proof of Theorem 4.9.3. Since $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{M} \times \mathcal{N}$ is a global saddle point of f, we have $f(\mathbf{x}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_t) \geq 0$. Plugging this inequality into the inequality from Lemma 4.11.1 yields that

$$(d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2 \leq (1 - \mu\eta_t)\left((d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2\right) + 2\bar{\xi}_0\eta_t^2 L^2.$$

Since $\eta_t = \frac{1}{\mu}\min\{1, \frac{2}{t}\}$, we have

$$(d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2 \leq (1 - \tfrac{2}{t})\left((d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2\right) + \tfrac{8\bar{\xi}_0 L^2}{\mu^2 t^2}, \quad \text{for all } t \geq 2.$$

Letting $\{b_t\}_{t\geq 1}$ be a nonnegative sequence such that $a_{t+1} \leq (1 - \frac{P}{t})a_t + \frac{Q}{t^2}$ where $P > 1$ and $Q > 0$. Then, Chung [1954] proved that $a_t \leq \frac{Q}{P-1}\frac{1}{t}$. Therefore, we have

$$(d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 \leq \tfrac{8\bar{\xi}_0 L^2}{\mu^2 t}, \quad \text{for all } t \geq 2.$$

This completes the proof.

Proof of Theorem 4.9.4. By the inductive formulas of $\bar{\mathbf{x}}_{t+1} = \mathrm{Exp}_{\bar{\mathbf{x}}_t}(\frac{1}{t+1} \cdot \mathrm{Exp}^{-1}_{\bar{\mathbf{x}}_t}(\mathbf{x}_t))$ and $\bar{\mathbf{y}}_{t+1} = \mathrm{Exp}_{\bar{\mathbf{y}}_t}(\frac{1}{t+1} \cdot \mathrm{Exp}^{-1}_{\bar{\mathbf{y}}_t}(\mathbf{y}_t))$ and using Zhang et al. [2022b, Lemma C.2], we have

$$f(\bar{\mathbf{x}}_T, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}}_T) \leq \tfrac{1}{T}\left(\sum_{t=0}^{T-1} f(\mathbf{x}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_t)\right).$$

Plugging the above inequality into the inequality from Lemma 4.11.1 yields that (recall that $\mu = 0$ in geodesically convex-concave setting and $\eta_t = \eta = \frac{1}{L}\sqrt{\frac{D_0}{2\bar{\xi}_0 T}}$)

$$f(\bar{\mathbf{x}}_T, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}}_T) \leq \tfrac{1}{2\eta T}\left((d_{\mathcal{M}}(\mathbf{x}_0, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_0, \mathbf{y}^\star))^2\right) + \bar{\xi}_0 \eta L^2.$$

This together with $(d_{\mathcal{M}}(\mathbf{x}_0, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_0, \mathbf{y}^\star))^2 \leq D_0$ implies that

$$f(\bar{\mathbf{x}}_T, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}}_T) \leq \tfrac{D_0}{2\eta T} + \bar{\xi}_0 \eta L^2.$$

Since $\eta = \frac{1}{L}\sqrt{\frac{D_0}{2\bar{\xi}_0 T}}$, we have

$$f(\bar{\mathbf{x}}_T, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}}_T) \leq L\sqrt{\tfrac{2\bar{\xi}_0 D_0}{T}}.$$

This completes the proof.

Proof of Theorem 4.9.6. Since $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{M} \times \mathcal{N}$ is a global saddle point of f, we have $\mathbb{E}[f(\mathbf{x}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_t)] \geq 0$. Plugging this inequality into the inequality from Lemma 4.11.2 yields that

$$\mathbb{E}\left[(d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right] \leq (1-\mu\eta_t)\mathbb{E}\left[(d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2\right] + 4\bar{\xi}_0\eta_t^2(L^2+\sigma^2).$$

Since $\eta_t = \frac{1}{\mu}\min\{1, \frac{2}{t}\}$, we have

$$\mathbb{E}\left[(d_{\mathcal{M}}(\mathbf{x}_{t+1}, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_{t+1}, \mathbf{y}^\star))^2\right] \leq (1-\tfrac{2}{t})\mathbb{E}\left[(d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2\right] + \tfrac{16\bar{\xi}_0(L^2+\sigma^2)}{\mu^2 t^2}, \text{ for all } t \geq 2.$$

Applying the same argument as used in Theorem 4.9.3, we have

$$(d_{\mathcal{M}}(\mathbf{x}_t, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_t, \mathbf{y}^\star))^2 \leq \tfrac{16\bar{\xi}_0(L^2+\sigma^2)}{\mu^2 t}, \text{ for all } t \geq 2.$$

This completes the proof.

Proof of Theorem 4.9.7. Using the same argument, we have

$$f(\bar{\mathbf{x}}_T, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}}_T) \leq \tfrac{1}{T}\left(\sum_{t=0}^{T-1} f(\mathbf{x}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_t)\right).$$

Plugging the above inequality into the inequality from Lemma 4.11.2 yields that (recall that $\mu = 0$ in geodesically convex-concave setting and $\eta_t = \eta = \frac{1}{2}\sqrt{\frac{D_0}{\bar{\xi}_0(L^2+\sigma^2)T}}$)

$$\mathbb{E}[f(\bar{\mathbf{x}}_T, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}}_T)] \leq \frac{1}{2\eta T}\left((d_{\mathcal{M}}(\mathbf{x}_0, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_0, \mathbf{y}^\star))^2\right) + 2\bar{\xi}_0\eta(L^2+\sigma^2).$$

This together with $(d_{\mathcal{M}}(\mathbf{x}_0, \mathbf{x}^\star))^2 + (d_{\mathcal{N}}(\mathbf{y}_0, \mathbf{y}^\star))^2 \leq D_0$ implies that

$$\mathbb{E}[f(\bar{\mathbf{x}}_T, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}}_T)] \leq \frac{D_0}{2\eta T} + 2\bar{\xi}_0\eta(L^2+\sigma^2).$$

Since $\eta = \frac{1}{2}\sqrt{\frac{D_0}{\bar{\xi}_0(L^2+\sigma^2)T}}$, we have

$$f(\bar{\mathbf{x}}_T, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}}_T) \leq 2\sqrt{\frac{\bar{\xi}_0(L^2+\sigma^2)D_0}{T}}.$$

This completes the proof.

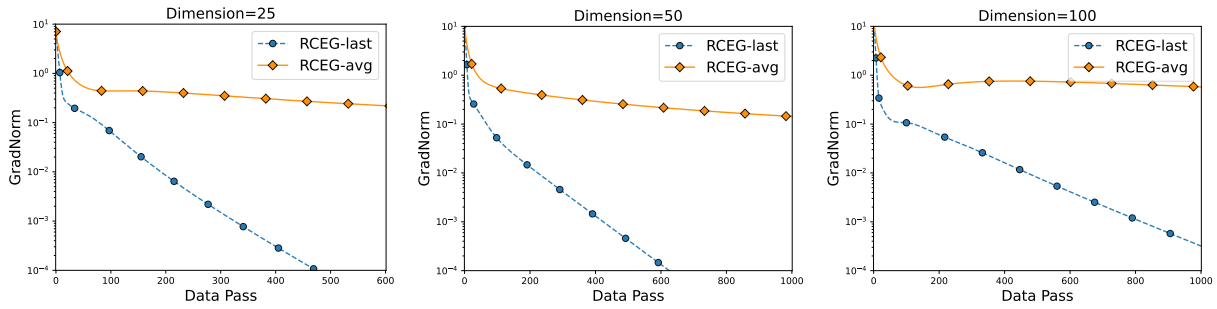


Figure 4.3: Comparison of last iterate (RCEG-last) and time-average iterate (RCEG-avg) for solving the RPCA problem when $\alpha = 2.0$. The horizontal axis represents the number of data passes and the vertical axis represents gradient norm.

4.12 Additional Experimental Results

We present some additional experimental results for the effect of different choices of α as well the effect of different choices of η for for RCEG. In our experiment, we set $n = 40$.

Figure 4.3 presents the performance of RCEG when $\alpha = 2.0$. We find that the results are similar to that summarized in Figure 4.1. In particular, the last iterate of RCEG consistently achieves the linearly convergence to an optimal solution in all the settings. In contrast, the average iterate of RCEG converges much slower than the last iterate of RCEG. Figure 4.4 summarizes the effect of different choices of η in RCEG. We observe that setting η as a relatively larger value will speed up the convergence to an optimal solution while all of the choices here lead to the linear convergence. This suggests that the choice of stepsize η in RCEG can be aggressive in practice.

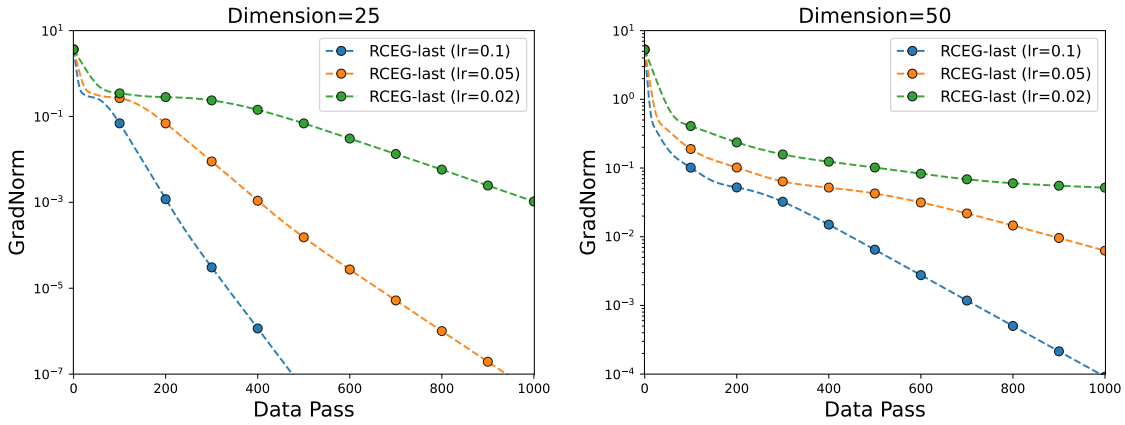


Figure 4.4: Comparison of different step sizes ($\eta \in \{0.1, 0.05, 0.02\}$) for solving the RPCA problem with different dimensions when $\alpha = 2.0$. The horizontal axis represents the number of data passes and the vertical axis represents gradient norm.

Part II

High-Order Optimization and Beyond

Chapter 5

A Closed-Loop Control Approach to High-Order Optimization

We provide a control-theoretic perspective on optimal tensor algorithms for minimizing a convex function in a finite-dimensional Euclidean space. Given a function $\Phi : \mathbb{R}^d \to \mathbb{R}$ that is convex and twice continuously differentiable, we study a closed-loop control system that is governed by the operators $\nabla\Phi$ and $\nabla^2\Phi$ together with a feedback control law $\lambda(\cdot)$ satisfying the algebraic equation $(\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta$ for some $\theta \in (0, 1)$. Our first contribution is to prove the existence and uniqueness of a local solution to this system via the Banach fixed-point theorem. We present a simple yet nontrivial Lyapunov function that allows us to establish the existence and uniqueness of a global solution under certain regularity conditions and analyze the convergence properties of trajectories. The rate of convergence is $O(1/t^{(3p+1)/2})$ in terms of objective function gap and $O(1/t^{3p})$ in terms of squared gradient norm. Our second contribution is to provide two algorithmic frameworks obtained from discretization of our continuous-time system, one of which generalizes the large-step A-HPE framework of Monteiro and Svaiter [2013] and the other of which leads to a new optimal p-th order tensor algorithm. While our discrete-time analysis can be seen as a simplification and generalization of Monteiro and Svaiter [2013], it is largely motivated by the aforementioned continuous-time analysis, demonstrating the fundamental role that the feedback control plays in optimal acceleration and the clear advantage that the continuous-time perspective brings to algorithmic design. A highlight of our analysis is that we show that all of the p-th order optimal tensor algorithms that we discuss minimize the squared gradient norm at a rate of $O(k^{-3p})$, which complements the recent analysis [Gasnikov et al., 2019a, Jiang et al., 2019, Bubeck et al., 2019].

5.1 Introduction

The interplay between continuous-time and discrete-time perspectives on dynamical systems has made a major impact on optimization theory. Classical examples include (1) the inter-

pretation of steepest descent, heavy ball and proximal algorithms as the explicit and implicit discretization of gradient-like dissipative systems [Polyak, 1987, Antipin, 1994, Attouch and Cominetti, 1996, Alvarez, 2000, Attouch et al., 2000, Alvarez and Attouch, 2001]; and (2) the explicit discretization of Newton-like and Levenberg-Marquardt regularized systems [Alvarez and Pérez C, 1998, Attouch and Redont, 2001, Alvarez et al., 2002, Attouch and Svaiter, 2011, Attouch et al., 2012, Maingé, 2013, Attouch et al., 2013b, Abbas et al., 2014, Attouch et al., 2016a, Attouch and László, 2020b,a], which give standard and regularized Newton algorithms. One particularly salient way that these connections have spurred research is via the use of Lyapunov functions to transfer asymptotic behavior and rates of convergence between continuous time and discrete time.

Recent years have witnessed a flurry of new research focusing on continuous-time perspectives on Nesterov's accelerated gradient algorithm (NAG) [Nesterov, 1983] and related methods [Güler, 1992, Beck and Teboulle, 2009, Tseng, 2010, Nesterov, 2013a]. These perspectives arise from derivations that obtain differential equations as limits of discrete dynamics [Su et al., 2016, Krichene et al., 2015b, Attouch and Peypouquet, 2016, Vassilis et al., 2018, Muehlebach and Jordan, 2019, Diakonikolas and Orecchia, 2019, Attouch and Peypouquet, 2019, Sebbouh et al., 2020, Shi et al., 2022], including quasi-gradient formulations and Kurdyka-Lojasiewicz theory [Bégout et al., 2015, Attouch et al., 2022a] (see the references [Huang, 2006, Chergui, 2008, Chill and Fašangová, 2010, Bárta et al., 2012, Bárta and Fašangová, 2016] for geometrical perspective on the topic), inertial gradient systems with constant or asymptotic vanishing damping [Su et al., 2016, Attouch and Cabot, 2017, Attouch et al., 2018, 2019a] and their extension to maximally monotone operators [Bot and Csetnek, 2016, Attouch and Cabot, 2018, 2020], Hessian-driven damping [Alvarez et al., 2002, Attouch et al., 2012, 2016b, Boţ et al., 2021, Attouch et al., 2022d,a, Shi et al., 2022], time scaling [Attouch et al., 2019a,c, 2022a,c], dry friction damping [Adly and Attouch, 2020, 2022], closed-loop damping [Attouch et al., 2022b,a], control-theoretic design [Lessard et al., 2016, Hu and Lessard, 2017, Fazlyab et al., 2018] and Lagrangian and Hamiltonian frameworks [Wibisono et al., 2016, Betancourt et al., 2018, Maddison et al., 2018, O'Donoghue and Maddison, 2019, França et al., 2020, Diakonikolas and Jordan, 2021, Muehlebach and Jordan, 2021, França et al., 2021]. Examples of hitherto unknown results that have arisen from this line of research include the fact that NAG achieves a fast rate of $o(k^{-2})$ in terms of objective function gap [May, 2017, Attouch and Peypouquet, 2016, Attouch et al., 2018] and $O(k^{-3})$ in terms of squared gradient norm [Shi et al., 2022].

The introduction of the Hessian-driven damping into continuous-time dynamics has been a particular milestone in optimization and mechanics. The precursor of this perspective can be found in the variational characterization of the Levenberg-Marquardt method and Newton's method [Alvarez and Pérez C, 1998], a development that inspired work on continuous-time Newton-like approaches for convex minimization [Alvarez and Pérez C, 1998, Attouch and Redont, 2001] and monotone inclusions [Attouch and Svaiter, 2011, Maingé, 2013, Attouch et al., 2013b, Abbas et al., 2014, Attouch et al., 2016a, Attouch and László, 2020b,a]. Building on these works, Alvarez et al. [2002] distinguished Hessian-driven damping from classical continuous Newton formulations and showed its importance in optimization and

mechanics. Subsequently, Attouch et al. [2016b] demonstrated the connection between Hessian-driven damping and the forward-backward algorithms in Nesterov acceleration (e.g., FISTA), and combined Hessian-driven damping with asymptotically vanishing damping [Su et al., 2016]. The resulting dynamics takes the following form:

$$\ddot{x}(t) + \tfrac{\alpha}{t}\dot{x}(t) + \beta\nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0, \tag{5.1}$$

where it is worth mentioning that the presence of the Hessian does not entail numerical difficulties since it arises in the form $\nabla^2\Phi(x(t))\dot{x}(t)$, which is the time derivative of the function $t \mapsto \nabla\Phi(x(t))$. Further work in this vein appeared in Shi et al. [2022], where Nesterov acceleration was interpreted via multiscale limits that yield high-resolution differential equations:

$$\ddot{x}(t) + \tfrac{3}{t}\dot{x}(t) + \sqrt{s}\nabla^2\Phi(x(t))\dot{x}(t) + \left(1 + \tfrac{3\sqrt{s}}{2t}\right)\nabla\Phi(x(t)) = 0. \tag{5.2}$$

These limits were used in particular to distinguish between Polyak's heavy-ball method and NAG, which are not distinguished by naive limiting arguments that yield the same differential equation for both.

Althought the coefficients are different in Eq. (5.1) and Eq. (5.2), both contain Hessian-driven damping, which corresponds to a correction term obtained via discretization, and which provides fast convergence to zero of the gradients and reduces the oscillatory aspects. Using this viewpoint, several subtle analyses have been recently provided in work independent of ours [Attouch et al., 2022b,a]. In particular, they develop a convergence theory for a general inertial system with asymptotic vanishing damping and Hessian-driven damping. Under certain conditions, the fast convergence is guaranteed in terms of both objective function gap and squared gradient norm. Beyond the aforementioned line of work, however, most of the focus in using continuous-time perspectives to shed light on acceleration has been restricted to the setting of first-order optimization algorithms. As noted in a line of recent work [Monteiro and Svaiter, 2013, Nesterov, 2018, Arjevani et al., 2019, Gasnikov et al., 2019a, Jiang et al., 2019, Bubeck et al., 2019, Song et al., 2021], there is a significant gap in our understanding of optimal p-th order tensor algorithms with $p \geq 2$, with existing algorithms and analysis being much more involved than NAG.

In this paper, we show that a continuous-time perspective helps to bridge this gap and yields a unified perspective on first-order and higher-order acceleration. We refer to our work as a *control-theoretic perspective*, as it involves the study of a closed-loop control system that can be viewed as a differential equation that is governed by a feedback control law, $\lambda(\cdot)$, satisfying the algebraic equation $(\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta$ for some $\theta \in (0, 1)$. Our approach is similar to that of Attouch et al. [2013b, 2016a], for the case without inertia, and it provides a first step into a theory of the autonomous inertial systems that link closed-loop control and optimal high-order tensor algorithms. Mathematically, our system can be written as follows:

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \beta(t)\nabla^2\Phi(x(t))\dot{x}(t) + b(t)\nabla\Phi(x(t)) = 0, \tag{5.3}$$

where (α, β, b) explicitly depends on the variables (x, λ, a), the parameters $c > 0$, $\theta \in (0, 1)$ and the order $p \in \{1, 2, \ldots\}$:

$$\begin{aligned} &\alpha(t) = \tfrac{2\dot{a}(t)}{a(t)} - \tfrac{\ddot{a}(t)}{\dot{a}(t)}, \quad \beta(t) = \tfrac{(\dot{a}(t))^2}{a(t)}, \quad b(t) = \tfrac{\dot{a}(t)(\dot{a}(t)+\ddot{a}(t))}{a(t)}, \\ &a(t) = \tfrac{1}{4}(\textstyle\int_0^t \sqrt{\lambda(s)}ds + c)^2, \quad (\lambda(t))^p \|\nabla\Phi(x(t))\|^{p-1} = \theta. \end{aligned} \tag{5.4}$$

The initial condition is $x(0) = x_0 \in \{x \in \mathbb{R}^d \mid \|\nabla\Phi(x)\| \neq 0\}$ and $\dot{x}(0) \in \mathbb{R}^d$. Note that this condition is not restrictive since $\|\nabla\Phi(x_0)\| = 0$ implies that the optimization problem has been already solved. A key ingredient in our system is the algebraic equation $(\lambda(t))^p \|\nabla\Phi(x(t))\|^{p-1} = \theta$, which links the feedback control law $\lambda(\cdot)$ and the gradient norm $\|\nabla\Phi(x(\cdot))\|$, and which generalizes an equation appearing in Attouch et al. [2016a] for modeling the proximal Newton algorithm. We recall that Eq. (5.3) has also been studied in Attouch et al. [2022b,a], who provide a general convergence result when (α, β, b) satisfies certain conditions. However, when $p \geq 2$, the specific choice of (α, β, b) in Eq. (5.4) does not have an analytic form and it thus seems difficult to verify whether (α, β, b) in our control system satisfies that condition (see Attouch et al. [2022a, Theorem 2.1])). This topic is beyond the scope of this paper and we leave its investigation to future work.

Our contribution. Throughout the paper, unless otherwise indicated, we assume that

$$\Phi : \mathbb{R}^d \to \mathbb{R} \text{ *is convex and twice continuously differentiable and the set of global minimizers of* } \Phi \text{ is nonempty.}$$

As we shall see, our main results on the existence and uniqueness of solutions and convergence properties of trajectories are valid under this general assumption. We also believe that this general setting paves the way for extensions to nonsmooth convex functions or maximal monotone operators (replacing the gradient by the subdifferential or the operator) [Alvarez et al., 2002, Attouch et al., 2012, 2016b]. This is evidenced by the equivalent first-order reformulations of our closed-loop control system in time and space (without the occurrence of the Hessian). However, we do not pursue these extensions in the current paper.

The main contributions of our work are the following:

1. We study the closed-loop control system of Eq. (5.3) and Eq. (5.4) and prove the existence and uniqueness of a local solution. We show that when $p = 1$ and $c = 0$, our feedback law reduces to $\lambda(t) = \theta$ and our overall system reduces to the high-resolution differential equation studied in Shi et al. [2022], showing explicitly that our system extends the high-resolution framework from first-order optimization to high-order optimization.

2. We construct a simple yet nontrivial Lyapunov function that allows us to establish the existence and uniqueness of a global solution under regularity conditions (see Theorem 6.2.6). We also use the Lyapunov function to analyze the convergence rates of the solution trajectories; in particular, we show that the convergence rate is $O(t^{-(3p+1)/2})$ in terms of objective function gap and $O(t^{-3p})$ in terms of squared gradient norm.

3. We provide two algorithmic frameworks based on the implicit discretization of our closed-looped control system, one of which generalizes the large-step A-HPE in Monteiro and Svaiter [2013]. Our iteration complexity analysis is largely motivated by the aforementioned continuous-time analysis, simplifying the analysis in Monteiro and Svaiter [2013] for the case of $p = 2$ and generalizing it to $p > 2$ in a systematic manner (see Theorem 5.4.3 and 5.4.6 for the details).

4. We combine the algorithmic frameworks with an approximate tensor subroutine, yielding a suite of optimal p-th order tensor algorithms for minimizing a convex smooth function Φ which has Lipschitz p-th order derivatives. The resulting algorithms include not only the algorithms studied in the previous works [Gasnikov et al., 2019a, Jiang et al., 2019, Bubeck et al., 2019] but also yield a new optimal p-th order tensor algorithm. A highlight of our analysis is to show that all these p-th order optimal algorithms minimize the squared gradient norm at a rate of $O(k^{-3p})$, complementing the recent analysis in the aforementioned works.

Further related work. In addition to the aforementioned works, we provide a few additional remarks regarding related work on accelerated first-order and high-order algorithms for convex optimization.

A significant body of recent work in convex optimization focuses on understanding the underlying principle behind Nesterov's accelerated first-order algorithm (NAG) [Nesterov, 1983, 2018], with a particular focus on the interpretation of Nesterov acceleration as a temporal discretization of a continuous-time dynamical system [Krichene et al., 2015b, Su et al., 2016, Attouch and Peypouquet, 2016, May, 2017, Vassilis et al., 2018, Diakonikolas and Orecchia, 2019, Muehlebach and Jordan, 2019, Attouch et al., 2018, 2019a,b, Attouch and Peypouquet, 2019, Sebbouh et al., 2020, Attouch and Cabot, 2020, Adly and Attouch, 2022, Attouch et al., 2022a,b,d, Shi et al., 2022]. A line of new first-order algorithms have been obtained from the continuous-time dynamics by various advanced numerical integration strategies [Scieur et al., 2017, Betancourt et al., 2018, Zhang et al., 2018, Maddison et al., 2018, Shi et al., 2019, Wilson et al., 2019]. In particular, Scieur et al. [2017] showed that a basic gradient flow system and multi-step integration scheme yields a class of accelerated first-order optimization algorithms. Zhang et al. [2018] applied Runge-Kutta integration to an inertial gradient system without Hessian-driven damping [Wibisono et al., 2016] and showed that the resulting algorithm is faster than NAG when the objective function is sufficiently smooth and when the order of the integrator is sufficiently large. Maddison et al. [2018] and França et al. [2020] both considered conformal Hamiltonian systems and showed that the resulting discrete-time algorithm achieves fast convergence under certain smoothness conditions. Very recently, Shi et al. [2019] have rigorously justified the use of symplectic Euler integrators compared to explicit and implicit Euler integration, which was further studied by Muehlebach and Jordan [2021] and França et al. [2021]. Unfortunately, none of these approaches are suitable for interpreting optimal high-order tensor algorithms.

Research on acceleration in the second-order setting dates back to Nesterov's accelerated cubic regularized Newton algorithm (ACRN) [Nesterov, 2008] and Monteiro and Svaiter's accelerated Newton proximal extragradient (A-NPE) [Monteiro and Svaiter, 2013]. The ACRN algorithm was extended to a p-th order tensor algorithm with the improved convergence rate of $O(k^{-(p+1)})$ [Baes, 2009] and an adaptive p-th order tensor algorithm with essentially the same rate [Jiang et al., 2020]. This novel extension was also revisited by Nesterov [2021b] with a discussion on the efficient implementation of a third-order tensor algorithm. Meanwhile, within the alternative A-NPE framework, a p-th order tensor algorithm was studied in a line of works [Gasnikov et al., 2019a, Jiang et al., 2019, Bubeck et al., 2019] and was shown to achieve a convergence rate of $O(k^{-(3p+1)/2})$, matching the lower bound [Arjevani et al., 2019]. Subsequently, a high-order coordinate descent algorithm was studied in Amaral et al. [2022], and very recently, the high-order A-NPE framework has been specialized to the strongly convex setting [Marques Alves, 2022], generalizing the discrete-time algorithms in this paper with an improved convergence rate. Beyond the setting of Lipschitz continuous derivatives, high-order algorithms and their accelerated variants have been adapted for more general setting with Hölder continuous derivatives [Grapiglia and Nesterov, 2017, 2019, 2020, 2022b, Doikov and Nesterov, 2022] and an optimal algorithm has been proposed in Song et al. [2021]. Other settings include structured convex non-smooth minimization [Bullins, 2020], convex-concave minimax optimization and monotone variational inequalities [Bullins and Lai, 2022, Ostroukhov et al., 2020], and structured smooth convex minimization [Kamzolov, 2020, Nesterov, 2021d, 2023]. In the nonconvex setting, high-order algorithms have been proposed and analyzed [Birgin et al., 2016, 2017, Martínez, 2017, Cartis et al., 2018, 2019].

Unfortunately, the derivations of these algorithms do not flow from a single underlying principle but tend to involve case-specific algebra. As in the case of first-order algorithms, one would hope that a continuous-time perspective would offer unification, but the only work that we are aware of in this regard is Song et al. [2021], and the connection to dynamical systems in that work is unclear. In particular, some aspects of the UAF algorithm (see Song et al. [2021, Algorithm 5.1]), including the conditions in Eq. (5.31) and Eq. (5.32), do not have a continuous-time interpretation but rely on case-specific algebra. Moreover, their continuous-time framework reduces to an inertial system without Hessian-driven damping in the first-order setting, which has been proven to be an inaccurate surrogate.

We have been also aware of other type of discrete-time algorithms [Zhang et al., 2018, Maddison et al., 2018, Wilson et al., 2019] which were derived from continuous-time perspective with theoretical guarantee under certain condition. In particular, Wilson et al. [2019] derived a family of first-order algorithms by appeal to the explicit time discretization of the accelerated rescaled gradient dynamics. Their new algorithms are guaranteed to (surprisingly) achieve the same convergence rate as the existing optimal tensor algorithms [Gasnikov et al., 2019a, Jiang et al., 2019, Bubeck et al., 2019]. However, the strong smoothness assumption is necessary and might rule out many interesting application problems. In contrast, all the optimization algorithms developed in this paper are applicable for *general* convex and smooth problems with the optimal rate of convergence.

Notation. We use bold lower-case letters such as x to denote vectors, and upper-case letters such as X to denote tensors. For a vector $x \in \mathbb{R}^d$, we let $\|x\|$ denote its ℓ_2 Euclidean norm and let $\mathbb{B}_\delta(x) = \{x' \in \mathbb{R}^d \mid \|x' - x\| \le \delta\}$ denote its δ-neighborhood. For a tensor $X \in \mathbb{R}^{d_1 \times \cdots \times d_p}$, we define

$$X[z^1, \cdots, z^p] = \sum_{1 \le i_j \le d_j, 1 \le j \le p} \left[X_{i_1, \cdots, i_p}\right] z^1_{i_1} \cdots z^p_{i_p},$$

and denote by $\|X\|_{\mathrm{op}} = \max_{\|z^i\|=1, 1 \le j \le p} X[z^1, \cdots, z^p]$ its operator norm.

Fix $p \ge 1$, we define $\mathcal{F}^p_\ell(\mathbb{R}^d)$ as the class of convex functions on $\mathbb{R}^d$ with ℓ-Lipschitz p-th order derivatives; that is, $f \in \mathcal{F}^p_\ell(\mathbb{R}^d)$ if and only if f is convex and $\|\nabla^{(p)} f(x') - \nabla^{(p)} f(x)\|_{\mathrm{op}} \le \ell \|x' - x\|$ for all $x, x' \in \mathbb{R}^d$ in which $\nabla^{(p)} f(x)$ is the p-th order derivative tensor of f at $x \in \mathbb{R}^d$. More specifically, for $\{z^1, z^2, \ldots, z^p\} \subseteq \mathbb{R}^d$, we have

$$\nabla^{(p)} f(x)[z^1, \ldots, z^p] = \sum_{1 \le i_1, \cdots, i_p \le d} \left[\frac{\partial^p f}{\partial x_{i_1} \cdots \partial x_{i_p}}(x)\right] z^1_{i_1} \cdots z^p_{i_p}.$$

Given a tolerance $\epsilon \in (0, 1)$, the notation $a = O(b(\epsilon))$ stands for an upper bound, $a \le C b(\epsilon)$, in which $C > 0$ is independent of ϵ.

5.2 The Closed-Loop Control System

We study the closed-loop control system in Eq. (5.3) and Eq. (5.4). We start by rewriting our system as a first-order system in time and space (without the occurrence of the Hessian) which is important to our subsequent analysis and implicit time discretization. Then, we analyze the algebraic equation $(\lambda(t))^p \|\nabla \Phi(x(t))\|^{p-1} = \theta$ for $\theta \in (0, 1)$ and prove the existence and uniqueness of a local solution using the Banach fixed-point theorem. We conclude by discussing other systems in the literature that exemplify our general framework.

First-order system in time and space. We rewrite the closed-loop control system in Eq. (5.3) and Eq. (5.4) as follows:

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \beta(t)\nabla^2 \Phi(x(t))\dot{x}(t) + b(t)\nabla \Phi(x(t)) = 0,$$

where (α, β, b) explicitly depend on the variables (x, λ, a), the parameters $c > 0$, $\theta \in (0, 1)$ and the order $p \in \{1, 2, \ldots\}$:

$$\begin{aligned} &\alpha(t) = \tfrac{2\dot{a}(t)}{a(t)} - \tfrac{\ddot{a}(t)}{\dot{a}(t)}, \quad \beta(t) = \tfrac{(\dot{a}(t))^2}{a(t)}, \quad b(t) = \tfrac{\dot{a}(t)(\dot{a}(t) + \ddot{a}(t))}{a(t)}, \\ &a(t) = \tfrac{1}{4}(\textstyle\int_0^t \sqrt{\lambda(s)} ds + c)^2, \quad (\lambda(t))^p \|\nabla \Phi(x(t))\|^{p-1} = \theta. \end{aligned}$$

By multiplying both sides of the first equation by $\frac{a(t)}{\dot{a}(t)}$ and using the definition of $\alpha(t)$, $\beta(t)$ and $b(t)$, we have

$$\frac{a(t)}{\dot{a}(t)}\ddot{x}(t) + \left(2 - \tfrac{a(t)\ddot{a}(t)}{(\dot{a}(t))^2}\right)\dot{x}(t) + \dot{a}(t)\nabla^2 \Phi(x(t))\dot{x}(t) + (\dot{a}(t) + \ddot{a}(t))\nabla \Phi(x(t)) = 0.$$

Defining $z_1(t) = \frac{a(t)}{\dot{a}(t)}\dot{x}(t)$ and $z_2(t) = \dot{a}(t)\nabla\Phi(x(t))$, we have

$$\dot{z}_1(t) = \frac{a(t)}{\dot{a}(t)}\ddot{x}(t) + \left(1 - \frac{a(t)\ddot{a}(t)}{(\dot{a}(t))^2}\right)\dot{x}(t), \quad \dot{z}_2(t) = \dot{a}(t)\nabla^2\Phi(x(t))\dot{x}(t) + \ddot{a}(t)\nabla\Phi(x(t)).$$

Putting these pieces together yields

$$\dot{z}_1(t) + \dot{x}(t) + \dot{z}_2(t) = -\dot{a}(t)\nabla\Phi(x(t)).$$

Integrating this equation over the interval $[0, t]$, we have

$$z_1(t) + x(t) + z_2(t) = z_1(0) + x(0) + z_2(0) - \int_0^t \dot{a}(s)\nabla\Phi(x(s))ds. \tag{5.5}$$

Since $x(0) = x_0 \in \{x \in \mathbb{R}^d \mid \|\nabla\Phi(x)\| \neq 0\}$, it is easy to verify that $\lambda(0)$ is well defined and determined by the algebraic equation $\lambda(0) = \theta^{\frac{1}{p}}\|\nabla\Phi(x_0)\|^{-\frac{p-1}{p}}$. Using the definition of $a(t)$, we have $a(0) = \frac{c^2}{4}$ and $\dot{a}(0) = \frac{c\theta^{\frac{1}{2p}}\|\nabla\Phi(x_0)\|^{-\frac{p-1}{2p}}}{2}$. Putting these pieces together with the definition of $z_1(t)$ and $z_2(t)$, we have

$$\begin{aligned} z_1(0) + x(0) + z_2(0) &= \frac{a(0)}{\dot{a}(0)}\dot{x}(0) + x(0) + \dot{a}(0)\nabla\Phi(x(0)) \\ &= x(0) + \frac{c\theta^{-\frac{1}{2p}}\dot{x}(0)\|\nabla\Phi(x(0))\|^{\frac{p-1}{2p}} + c\theta^{\frac{1}{2p}}\|\nabla\Phi(x(0))\|^{-\frac{p-1}{2p}}\nabla\Phi(x(0))}{2}. \end{aligned}$$

This implies that $z_1(0) + x(0) + z_2(0)$ is completely determined by the initial condition and parameters $c > 0$ and $\theta \in (0, 1)$. For simplicity, we define $v_0 := z_1(0) + x(0) + z_2(0)$ and rewrite Eq. (5.5) in the following form:

$$\frac{a(t)}{\dot{a}(t)}\dot{x}(t) + x(t) + \dot{a}(t)\nabla\Phi(x(t)) = v_0 - \int_0^t \dot{a}(s)\nabla\Phi(x(s))ds. \tag{5.6}$$

By introducing a new variable $v(t) = v_0 - \int_0^t \dot{a}(s)\nabla\Phi(x(s))ds$, we rewrite Eq. (5.6) in the following equivalent form:

$$\dot{v}(t) + \dot{a}(t)\nabla\Phi(x(t)) = 0, \quad \dot{x}(t) + \frac{\dot{a}(t)}{a(t)}(x(t) - v(t)) + \frac{(\dot{a}(t))^2}{a(t)}\nabla\Phi(x(t)) = 0.$$

Summarizing, the closed-loop control system in Eq. (5.3) and Eq. (5.4) can be written as a first-order system in time and space as follows:

$$\begin{cases} \dot{v}(t) + \dot{a}(t)\nabla\Phi(x(t)) = 0 \\ \dot{x}(t) + \frac{\dot{a}(t)}{a(t)}(x(t) - v(t)) + \frac{(\dot{a}(t))^2}{a(t)}\nabla\Phi(x(t)) = 0 \\ a(t) = \frac{1}{4}(\int_0^t \sqrt{\lambda(s)}ds + c)^2 \\ (\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta \\ (x(0), v(0)) = (x_0, v_0). \end{cases} \tag{5.7}$$

We also provide another first-order system in time and space with different variable (x, v, λ, γ). We study this system because its implicit time discretization leads to a new algorithmic framework which does not appear in the literature. This first-order system is summarized as follows:

$$\begin{cases} \dot{v}(t) - \frac{\dot{\gamma}(t)}{\gamma^2(t)}\nabla\Phi(x(t)) = 0 \\ \dot{x}(t) - \frac{\dot{\gamma}(t)}{\gamma(t)}(x(t) - v(t)) + \frac{(\dot{\gamma}(t))^2}{(\gamma(t))^3}\nabla\Phi(x(t)) = 0 \\ \gamma(t) = 4(\int_0^t \sqrt{\lambda(s)}ds + c)^{-2} \\ (\lambda(t))^p \|\nabla\Phi(x(t))\|^{p-1} = \theta \\ (x(0), v(0)) = (x_0, v_0). \end{cases} \tag{5.8}$$

Remark 5.2.1 *The first-order systems in Eq. (5.7) and Eq. (5.8) are equivalent. It suffices to show that*

$$\dot{a}(t) = -\frac{\dot{\gamma}(t)}{\gamma^2(t)}, \quad \frac{\dot{a}(t)}{a(t)} = -\frac{\dot{\gamma}(t)}{\gamma(t)}, \quad \frac{(\dot{a}(t))^2}{a(t)} = \frac{(\dot{\gamma}(t))^2}{(\gamma(t))^3}.$$

By the definition of $a(t)$ and $\gamma(t)$, we have $a(t) = \frac{1}{\gamma(t)}$ which implies that $\dot{a}(t) = -\frac{\dot{\gamma}(t)}{\gamma^2(t)}$.

Remark 5.2.2 *The first-order systems in Eq. (5.7) and Eq. (5.8) pave the way for extensions to nonsmooth convex functions or maximal monotone operators (replacing the gradient by the subdifferential or the operator), as done in Alvarez et al. [2002] and Attouch et al. [2012, 2016b]. In this setting, either the open-loop case or the closed-loop case without inertia has been studied in the literature [Attouch and Svaiter, 2011, Maingé, 2013, Attouch et al., 2013b, Abbas et al., 2014, Attouch et al., 2016a, Bot and Csetnek, 2016, Attouch and Cabot, 2018, 2020, Attouch and László, 2020b], but there is significantly less work on the case of a closed-loop control system with inertia. For recent progress in this direction, see Attouch et al. [2022b] and references therein.*

Algebraic equation. We study the algebraic equation,

$$(\lambda(t))^p \|\nabla\Phi(x(t))\|^{p-1} = \theta \in (0, 1), \tag{5.9}$$

which links the feedback control $\lambda(\cdot)$ and the solution trajectory $x(\cdot)$ in the closed-loop control system. To streamline the presentation, we define a function $\varphi : [0, +\infty) \times \mathbb{R}^d \mapsto [0, +\infty)$ such that

$$\varphi(\lambda, x) = \lambda \|\nabla\Phi(x)\|^{\frac{p-1}{p}}, \quad \varphi(0, x) = 0.$$

By definition, Eq. (6.5) is equivalent to $\varphi(\lambda(t), x(t)) = \theta^{1/p}$. Our first proposition presents a property of the mapping $\varphi(\cdot, x)$, for a fixed $x \in \mathbb{R}^d$ satisfying $\nabla\Phi(x) \neq 0$. We have:

Proposition 5.2.3 *Fixing $x \in \mathbb{R}^d$ with $\nabla\Phi(x) \neq 0$, the mapping $\varphi(\cdot, x)$ satisfies*

1. *$\varphi(\cdot, x)$ is linear, strictly increasing and $\varphi(0, x) = 0$.*

2. $\varphi(\lambda, x) \to +\infty$ *as* $\lambda \to +\infty$.

Proof. By the definition of φ, the mapping $\varphi(\cdot, x)$ is linear and $\varphi(0, x) = 0$. Since $\nabla\Phi(x) \neq 0$, we have $\|\nabla\Phi(x)\| > 0$ and $\varphi(\cdot, x)$ is thus strictly increasing. Since $\varphi(\cdot, x)$ is linear and strictly increasing, $\varphi(\lambda, x) \to +\infty$ as $\lambda \to +\infty$. □

In view of Proposition 5.2.3, for any fixed point x with $\nabla\Phi(x) \neq 0$, there exists a unique $\lambda > 0$ such that $\varphi(\lambda, x) = \theta^{1/p}$ for some $\theta \in (0, 1)$. We accordingly define $\Omega \subseteq \mathbb{R}^d$ and the mapping $\Lambda_\theta : \Omega \mapsto (0, \infty)$ as follows:

$$\Omega = \{x \in \mathbb{R}^d \mid \|\nabla\Phi(x)\| \neq 0\}, \quad \Lambda_\theta(x) = \theta^{\frac{1}{p}} \|\nabla\Phi(x)\|^{-\frac{p-1}{p}}. \tag{5.10}$$

We now provide several basic results concerning Ω and $\Lambda_\theta(\cdot)$ which are crucial to the proof of existence and uniqueness presented in this chapter.

Proposition 5.2.4 *The set* Ω *is open.*

Proof. Given $x \in \Omega$, it suffices to show that $\mathbb{B}_\delta(x) \subseteq \Omega$ for some $\delta > 0$. Since Φ is twice continuously differentiable, $\nabla\Phi$ is locally Lipschitz; that is, there exists $\tilde{\delta} > 0$ and $L > 0$ such that

$$\|\nabla\Phi(z) - \nabla\Phi(x)\| \leq L\|z - x\|, \quad \forall z \in \mathbb{B}_{\delta_1}(x).$$

Combining this inequality with the triangle inequality, we have

$$\|\nabla\Phi(z)\| = \|\nabla\Phi(x)\| - \|\nabla\Phi(z) - \nabla\Phi(x)\| \geq \|\nabla\Phi(x)\| - L\|z - x\|.$$

Let $\delta = \min\{\tilde{\delta}, \frac{\|\nabla\Phi(x)\|}{2L}\}$. Then, for any $z \in \mathbb{B}_\delta(x)$, we have

$$\|\nabla\Phi(z)\| \geq \tfrac{\|\nabla\Phi(x)\|}{2} > 0 \Longrightarrow z \in \Omega.$$

This completes the proof. □

Proposition 5.2.5 *Fixing* $\theta \in (0, 1)$*, the mappings* $\Lambda_\theta(\cdot)$ *and* $\sqrt{\Lambda_\theta(\cdot)}$ *are continuous and locally Lipschitz over* Ω.

Proof. By the definition of $\Lambda_\theta(\cdot)$, it suffices to show that $\Lambda_\theta(\cdot)$ is continuous and locally Lipschitz over Ω since the same argument works for $\sqrt{\Lambda_\theta(\cdot)}$.

First, we prove the continuity of $\Lambda_\theta(\cdot)$ over Ω. Since $\|\nabla\Phi(x)\| > 0$ for any $x \in \Omega$, the function $\|\nabla\Phi(\cdot)\|^{-\frac{p-1}{p}}$ is continuous over Ω. By the definition of $\Lambda_\theta(\cdot)$, we achieve the desired result. Second, we prove that $\Lambda_\theta(\cdot)$ is locally Lipschitz over Ω. Since Φ is twice continuously differentiable, $\nabla\Phi$ is locally Lipschitz. For $p = 1$, $\Lambda_\theta(\cdot)$ is a constant everywhere and thus locally Lipschitz over Ω. For $p \geq 2$, the function $x^{-\frac{p-1}{p}}$ is locally Lipschitz at any point $x > 0$. Also, by Proposition 5.2.4, Ω is an open set. Putting these pieces together yields that $\|\nabla\Phi(\cdot)\|^{-\frac{p-1}{p}}$ is locally Lipschitz over Ω; that is, there exist $\delta > 0$ and $L > 0$ such that

$$|\|\nabla\Phi(x')\|^{-\frac{p-1}{p}} - \|\nabla\Phi(x'')\|^{-\frac{p-1}{p}}| \leq L\|x' - x''\|, \quad \forall x', x'' \in \mathbb{B}_\delta(x),$$

which implies that

$$|\Lambda_\theta(x') - \Lambda_\theta(x'')| \leq \theta^{\frac{1}{p}} L \|x' - x''\|, \quad \forall x', x'' \in \mathbb{B}_\delta(x).$$

This completes the proof. □

Existence and uniqueness of a local solution. We prove the existence and uniqueness of a local solution of the closed-loop control system in Eq. (5.3) and Eq. (5.4) by appeal to the Banach fixed-point theorem. Using the previous arguments (see Eq. (5.6)), our system can be equivalently written as follows:

$$\begin{cases} \dot{x}(t) + \frac{\dot{a}(t)}{a(t)}(x(t) + \int_0^t \dot{a}(s)\nabla\Phi(x(s))ds - v_0) + \frac{(\dot{a}(t))^2}{a(t)}\nabla\Phi(x(t)) = 0 \\ a(t) = \frac{1}{4}(\int_0^t \sqrt{\lambda(s)}ds + c)^2 \\ (\lambda(t))^p \|\nabla\Phi(x(t))\|^{p-1} = \theta \\ x(0) = x_0. \end{cases}$$

Using the mapping $\Lambda_\theta : \Omega \mapsto (0, \infty)$ (see Eq. (5.10)), this system can be further formulated as an autonomous system. Indeed, we have

$$\lambda(t) = \Lambda_\theta(x(t)) \Longleftrightarrow \lambda(t)]^p \|\nabla\Phi(x(t))\|^{p-1} = \theta,$$

which implies that

$$a(t) = \frac{1}{4}\left(\int_0^t \sqrt{\Lambda_\theta(x(s))}ds + c\right)^2, \quad \dot{a}(t) = \frac{1}{2}\sqrt{\Lambda_\theta(x(t))}\left(\int_0^t \sqrt{\Lambda_\theta(x(s))}\, ds + c\right).$$

Putting these pieces together, we arrive at an autonomous system in the following compact form:

$$\dot{x}(t) = F(t, x(t)), \quad x(0) = x_0 \in \Omega, \tag{5.11}$$

where the vector field $F : [0, +\infty) \times \Omega \mapsto \mathbb{R}^d$ is given by

$$F(t, x(t)) = -\frac{\sqrt{\Lambda_\theta(x(t))}(2x(t) + \int_0^t \sqrt{\Lambda_\theta(x(s))}(\int_0^s \sqrt{\Lambda_\theta(x(w))}dw + c)\nabla\Phi(x(s))ds - v_0)}{\int_0^t \sqrt{\Lambda_\theta(x(s))}\, ds + c} - \Lambda_\theta(x(t))\nabla\Phi(x(t)). \tag{5.12}$$

A common method for proving the existence and uniqueness of a local solution is via appeal to the Cauchy-Lipschitz theorem [Coddington and Levinson, 1955, Theorem I.3.1]. This theorem, however, requires that $F(t, x)$ be continuous in t and Lipschitz in x, and this is not immediate in our case due to the appearance of $\int_0^t \sqrt{\Lambda_\theta(x(s))}ds$. We instead recall that the proof of the Cauchy-Lipschitz theorem is generally based on the Banach fixed-point theorem [Granas and Dugundji, 2013], and we avail ourselves directly of the latter theorem. In particular, we construct Picard iterates ψ_k whose limit is a fixed point of a contraction T. We have the following theorem.

Theorem 5.2.6 *There exists $t_0 > 0$ such that the autonomous system in Eq. (5.11) and Eq. (5.12) has a unique solution $x : [0, t_0] \mapsto \mathbb{R}^d$.*

Proof. By Proposition 5.2.4 and the initial condition $x_0 \in \Omega$, there exists $\delta > 0$ such that $\mathbb{B}_\delta(x_0) \subseteq \Omega$. Note that Φ is twice continuously differentiable. By the definition of Λ_θ, we obtain that $\Lambda_\theta(z)$ and $\nabla\Phi(z)$ are both bounded for any $z \in \mathbb{B}_\delta(x_0)$. Putting these pieces together shows that there exists $M > 0$ such that, for any continuous function $x : [0, 1] \mapsto \mathbb{B}_\delta(x_0)$, we have

$$\|F(t, x(t))\| \leq M, \quad \forall t \in [0, 1]. \tag{5.13}$$

The set of such functions is not empty since a constant function $x = x_0$ is one element. Letting $t_1 = \min\{1, \frac{\delta}{M}\}$, we define $\mathcal{X}$ as the space of all continuous functions x on $[0, t_0]$ for some $t_0 < t_1$ whose graph is contained entirely inside the rectangle $[0, t_0] \times \mathbb{B}_\delta(x_0)$. For any $x \in \mathcal{X}$, we define

$$z(t) = Tx = x_0 + \int_0^t F(s, x(s))ds.$$

Note that $z(\cdot)$ is well defined and continuous on $[0, t_0]$. Indeed, $x \in \mathcal{X}$ implies that $x(t) \in \mathbb{B}_\delta(x_0) \subseteq \Omega$ for $\forall t \in [0, t_0]$. Thus, the integral of $F(s, x(s))$ is well defined and continuous. Second, the graph of $z(t)$ lies entirely inside the rectangle $[0, t_0] \times \mathbb{B}_\delta(x_0)$. Indeed, since $t \leq t_0 < t_1 = \min\{1, \frac{\delta}{M}\}$, we have

$$\|z(t) - x_0\| = \left\| \int_0^t F(s, x(s))ds \right\| \overset{\text{Eq. (5.13)}}{\leq} Mt \leq Mt_0 \leq Mt_1 \leq \delta.$$

Putting these pieces together yields that T maps $\mathcal{X}$ to itself. By the fundamental theorem of calculus, we have $\dot{z}(t) = F(t, x(t))$. By a standard argument from ordinary differential equation theory, $\dot{x}(t) = F(t, x(t))$ and $x(0) = x_0$ if and only if x is a fixed point of T. Thus, it suffices to show the existence and uniqueness of a fixed point of T.

We consider the Picard iterates $\{\psi_k\}_{k\geq 0}$ with $\psi_0(t) = x_0$ for $\forall t \in [0, t_0]$ and $\psi_{k+1} = T\psi_k$ for all $k \geq 0$. By the Banach fixed-point theorem [Granas and Dugundji, 2013], the Picard iterates converge to a unique fixed point of T if $\mathcal{X}$ is an nonempty and complete metric space and T is a contraction from $\mathcal{X}$ to $\mathcal{X}$.

First, we show that $\mathcal{X}$ is an nonempty and complete metric space. Indeed, we define $d(x, x') = \max_{t\in[0,t_0]} \|x(t) - x'(t)\|$. It is easy to verify that d is a metric and $(\mathcal{X}, d)$ is a complete metric space (see Sutherland [2009] for the details). In addition, $\mathcal{X}$ is nonempty since the constant function $x = x_0$ is one element.

It remains to prove that T is a contraction for some $t_0 < t_1$. Indeed, $\Lambda_\theta(z)$ and $\nabla\Phi(z)$ are bounded for $\forall z \in \mathbb{B}_\delta(x_0)$; that is, there exists $M_1 > 0$ such that $\max\{\Lambda_\theta(z), \|\nabla\Phi(z)\|\} \leq M_1$ for $\forall z \in \mathbb{B}_\delta(x_0)$. By Proposition 5.2.5, Λ_θ and $\sqrt{\Lambda_\theta}$ are continuous and locally Lipschitz over Ω. Since $\mathbb{B}_\delta(x_0) \subseteq \Omega$ is bounded, there exists $L_1 > 0$ such that, for any $x', x'' \in \mathbb{B}_\delta(x_0)$, we have

$$\max\{|\Lambda_\theta(x') - \Lambda_\theta(x'')|, |\sqrt{\Lambda_\theta}(x') - \sqrt{\Lambda_\theta}(x'')|\} \leq L_1\|x' - x''\|. \tag{5.14}$$

Note that Φ is twice continuously differentiable. Thus, there exists $L_2 > 0$ such that $\|\nabla\Phi(x') - \nabla\Phi(x'')\| \leq L_2\|x' - x''\|$ for $\forall x', x'' \in \mathbb{B}_\delta(x_0)$. In addition, for any $t \in [0, t_0]$, we have $\|x(t)\| \leq \|x_0\| + \delta = M_2$.

We now proceed to the main proof. By the triangle inequality, we have

$$\begin{aligned}\|Tx'(t) - Tx''(t)\| \leq & \underbrace{\int_0^t \|\Lambda_\theta(x'(s))\nabla\Phi(x'(s)) - \Lambda_\theta(x''(s))\nabla\Phi(x''(s))\|ds}_{\mathbf{I}} \\ & + \int_0^t \Bigg\| \frac{\sqrt{\Lambda_\theta(x'(s))}}{\int_0^s \sqrt{\Lambda_\theta(x'(w))}dw + c} \left(\int_0^s \left(\sqrt{\Lambda_\theta(x'(w))} \left(\int_0^w \sqrt{\Lambda_\theta(x'(v))}\, dv + c \right) \right) \nabla\Phi(x'(w))dw \right) \\ & \underbrace{- \frac{\sqrt{\Lambda_\theta(x''(s))}}{\int_0^s \sqrt{\Lambda_\theta(x''(w))}dw + c} \left(\int_0^s \left(\sqrt{\Lambda_\theta(x''(w))} \left(\int_0^w \sqrt{\Lambda_\theta(x''(v))}\, dv + c \right) \right) \nabla\Phi(x''(w))dw \right) \Bigg\| ds}_{\mathbf{II}} \\ & + \underbrace{\int_0^t \left\| \frac{2\sqrt{\Lambda_\theta(x'(s))}}{\int_0^s \sqrt{\Lambda_\theta(x'(w))}dw + c}(x'(s) - v_0) - \frac{2\sqrt{\Lambda_\theta(x''(s))}}{\int_0^s \sqrt{\Lambda_\theta(x''(w))}dw + c}(x''(s) - v_0) \right\| ds}_{\mathbf{III}}.\end{aligned}$$

The key inequality for the subsequent analysis is as follows:

$$\|a_1b_1 - a_2b_2\| \leq \|a_1\|\|b_1 - b_2\| + \|b_2\|\|a_1 - a_2\|. \tag{5.15}$$

First, by combining Eq. (5.15) with $\max\{\Lambda_\theta(x(t)), \|\nabla\Phi(x(t))\|\} \leq M_1$, $\|\nabla\Phi(x') - \nabla\Phi(x'')\| \leq L_2\|x' - x''\|$ and Eq. (5.14), we obtain:

$$\mathbf{I} \leq M_1(L_1 + L_2)t_0 d(x', x'').$$

Second, we combine Eq. (5.15) with $\sqrt{\Lambda_\theta(x(t))} \leq \sqrt{M_1}$, Eq. (5.14) and $0 < s \leq t_0 < t_1 < 1$ to obtain:

$$\left\| \frac{\sqrt{\Lambda_\theta(x'(s))}}{\int_0^s \sqrt{\Lambda_\theta(x'(w))}dw + c} - \frac{\sqrt{\Lambda_\theta(x''(s))}}{\int_0^s \sqrt{\Lambda_\theta(x''(w))}dw + c} \right\| \leq \left(\tfrac{1}{c} + \tfrac{2\sqrt{M_1}}{c^2} \right) L_1 d(x', x'').$$

We also obtain by combining Eq. (5.15) with $\max\{\Lambda_\theta(x(t)), \|\nabla\Phi(x(t))\|\} \leq M_1$, $\|\nabla\Phi(x') - \nabla\Phi(x'')\| \leq L_2\|x' - x''\|$, Eq. (5.14) and $0 < w \leq s \leq t_0 < t_1 < 1$ that

$$\begin{aligned}& \left\| \int_0^s \left(\sqrt{\Lambda_\theta(x'(w))} \left(\int_0^w \sqrt{\Lambda_\theta(x'(v))}\, dv + c \right) \right) \nabla\Phi(x'(w))dw \right. \\ & \quad \left. - \int_0^s \left(\sqrt{\Lambda_\theta(x''(w))} \left(\int_0^w \sqrt{\Lambda_\theta(x''(v))}\, dv + c \right) \right) \nabla\Phi(x''(w))dw \right\| \\ & \leq (M_1L_2 + c\sqrt{M_1}L_2 + 2(M_1)^{3/2}L_1 + cM_1L_1)d(x', x'').\end{aligned}$$

In addition, by using $\max\{\Lambda_\theta(x(t)), \|\nabla\Phi(x(t))\|\} \leq M_1$ and $0 < w \leq s \leq t_0 < t_1 < 1$, we have

$$\left\| \frac{\sqrt{\Lambda_\theta(x'(s))}}{\int_0^s \sqrt{\Lambda_\theta(x'(w))}\, dw + c} \right\| \leq \frac{\sqrt{M_1}}{c},$$
$$\left\| \int_0^s \left(\sqrt{\Lambda_\theta(x''(w))} \left(\int_0^w \sqrt{\Lambda_\theta(x''(v))}\, dv + c \right) \right) \nabla\Phi(x''(w)) dw \right\| \leq (M_1)^2 + c(M_1)^{3/2}.$$

Putting these pieces together yields that

$$\mathbf{II} \leq \left(\frac{2(M_1)^{5/2}L_1}{c^2} + \frac{(M_1)^{3/2}L_2 + 5(M_1)^2 L_1}{c} + M_1 L_2 + 2(M_1)^{3/2} L_1 \right) t_0 d(x', x'').$$

Finally, by a similar argument, we have

$$\mathbf{III} \leq \left(\frac{2\sqrt{M_1} + 2(M_2 + \|v_0\|)L_1}{c} + \frac{4\sqrt{M_1}(M_2 + \|v_0\|)L_1}{c^2} \right) t_0 d(x', x'').$$

Combining the upper bounds for **I**, **II** and **III**, we have

$$d(Tx', Tx'') = \max_{t \in [0, t_0]} \|Tx'(t) - Tx''(t)\| \leq \bar{M} t_0 d(x', x''),$$

where $\bar{M}$ is a constant that does not depend on t_0 (in fact it depends on c, x_0, δ, $\Phi(\cdot)$ and $\Lambda_\theta(\cdot)$) and is defined as follows:

$$\bar{M} = \frac{2((M_1)^2 + 2M_2 + 2\|v_0\|)\sqrt{M_1}L_1}{c^2} + \frac{2\sqrt{M_1} + (2M_2 + 2\|v_0\| + 5(M_1)^2)L_1 + (M_1)^{3/2}L_2}{c} + 2M_1 L_2 + (M_1 + 2(M_1)^{3/2})L_1.$$

Therefore, the mapping T is a contraction if $t_0 \in (0, t_1]$ satisfies $t_0 \leq \frac{1}{2\bar{M}}$. □

Discussion. We compare the closed-loop control system in Eq. (5.3) and Eq. (5.4) with four main classes of systems in the literature.

Hessian-driven damping. The formal introduction of Hessian-driven damping in optimization dates to Alvarez et al. [2002], with many subsequent developments; see, e.g., Attouch et al. [2016b]. The system studied in this literature takes the following form:

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0.$$

In a Hilbert space setting and when $\alpha > 3$, the literature has established the weak convergence of any solution trajectory to a global minimizer of Φ and the convergence rate of $o(1/t^2)$ in terms of objective function gap.

Recall also that Shi et al. [2022] interpreted Nesterov acceleration as the discretization of a high-resolution differential equation:

$$\ddot{x}(t) + \frac{3}{t}\dot{x}(t) + \sqrt{s}\nabla^2\Phi(x(t))\dot{x}(t) + \left(1 + \frac{3\sqrt{s}}{2t}\right)\nabla\Phi(x(t)) = 0,$$

and showed that this equation distinguishes between Polyak's heavy-ball method and Nesterov's accelerated gradient method. In the special case in which $c = 0$ and $p = 1$, our system in Eq. (5.3) and Eq. (5.4) becomes

$$\ddot{x}(t) + \tfrac{3}{t}\dot{x}(t) + \theta\nabla^2\Phi(x(t))\dot{x}(t) + \left(\theta + \tfrac{\theta}{t}\right)\nabla\Phi(x(t)) = 0. \tag{5.16}$$

which also belongs to the class of high-resolution differential equations. Moreover, for $c = 0$ and $p = 1$, our system can be studied within the recently-proposed framework of Attouch et al. [2022b,a]; indeed, in this case (α, β, b) in Attouch et al. [2022a, Theorem 2.1] has an analytic form. However, the choice of (α, β, b) in our general setting in Eq. (5.4), for $p \geq 2$, does not have an analytic form and it is difficult to verify whether (α, β, b) in this case satisfies their condition.

Newton and Levenberg-Marquardt regularized systems. The precursor of this perspective was developed by Alvarez and Pérez C [1998] in a variational characterization of general regularization algorithms. By constructing the regularization of the potential function $\Phi(\cdot, \epsilon)$ satisfying $\Phi(\cdot, \epsilon) \to \Phi$ as $\epsilon \to 0$, they studied the following system:

$$\nabla^2\Phi(x(t), \epsilon(t))\dot{x}(t) + \dot{\epsilon}(t)\tfrac{\partial^2\Phi}{\partial\epsilon\partial x}(x(t), \epsilon(t)) + \nabla\Phi(x(t), \epsilon(t)) = 0.$$

Subsequently, Attouch and Redont [2001] and Attouch and Svaiter [2011] studied Newton dissipative and Levenberg-Marquardt regularized systems:

$$\begin{aligned} \textbf{(Newton)} \qquad & \ddot{x}(t) + \nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0. \\ \textbf{(Levenberg-Marquardt)} \qquad & \lambda(t)\dot{x}(t) + \nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0. \end{aligned}$$

These systems have been shown to be well defined and stable with robust asymptotic behavior [Attouch and Svaiter, 2011, Attouch et al., 2013b, Abbas et al., 2014], further motivating the study of the following inertial gradient system with constant damping and Hessian-driven damping [Alvarez et al., 2002]:

$$\ddot{x}(t) + \alpha\dot{x}(t) + \beta\nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0.$$

This system attains strong asymptotic stabilization and fast convergence properties [Alvarez et al., 2002, Attouch et al., 2012] and can be extended to solve the monotone inclusion problems with theoretical guarantee [Attouch and Svaiter, 2011, Maingé, 2013, Attouch et al., 2013b, Abbas et al., 2014, Attouch et al., 2016a, Attouch and László, 2020b,a]. However, these systems are aimed at interpreting standard and regularized Newton algorithms and fail to model optimal acceleration for the second-order algorithms [Monteiro and Svaiter, 2013].

Recently, Attouch et al. [2016a] proposed a proximal Newton algorithm for solving monotone inclusions, which is motivated by a closed-loop control system without inertia. It attains a suboptimal convergence rate of $O(t^{-2})$ in terms of objective function gap.

Closed-loop control systems. The closed-loop damping approach in Attouch et al. [2013b, 2016a] closely resembles ours. In particular, they interpret various Newton-type methods as the discretization of the closed-loop control system without inertia and prove the existence and uniqueness of a solution as well as the convergence rate of the solution trajectory. There are, however, some significant differences between our work and theirs. In particular, the appearance of inertia is well known to make analysis much more challenging. Standard existence and uniqueness proofs based on the Cauchy-Schwarz theorem suffice to analyze the system of Attouch et al. [2013b, 2016a] thanks to the lack of inertia, while Picard iterates and the Banach fixed-point theorem are necessary for our analysis. The construction of the Lyapunov function is also more difficult for the system with inertia.

This is an active research area and we refer the interested reader to a recent article of Attouch et al. [2022b] for a comprehensive treatment of this topic.

Continuous-time interpretation of high-order tensor algorithms. There is comparatively little work on continuous-time perspectives on high-order tensor algorithms; indeed, we are aware of only Wibisono et al. [2016] and Song et al. [2021].

By appealing to a variational formulation, Wibisono et al. [2016] derived the following inertial gradient system with asymptotic vanishing damping:

$$\ddot{x}(t) + \tfrac{p+2}{t}\dot{x}(t) + C(p+1)^2 t^{p-1}\nabla\Phi(x(t)) = 0. \tag{5.17}$$

Compared to our closed-loop control system, in Eq. (5.3) and Eq. (5.4), the system in Eq. (5.17) is an open-loop system without the algebra equation and does not contain Hessian-driven damping. These differences yield solution trajectories that only attain a suboptimal convergence rate of $O(t^{-(p+1)})$ in terms of objective function gap.

Very recently, Song et al. [2021] have proposed and analyzed the following dynamics (we consider the Euclidean setting for simplicity):

$$\begin{cases} a(t)\dot{x}(t) = \dot{a}(t)(z(t) - x(t)) \\ z(t) = \operatorname{argmin}_{x\in\mathbb{R}^d} \int_0^t \dot{a}(s)(\Phi(x(s)) + \langle\nabla\Phi(x(s)), x - x(s)\rangle)ds + \frac{1}{2}\|x - x_0\|^2. \end{cases}$$

Solving the minimization problem yields $z(t) = x_0 - \int_0^t \dot{a}(s)\nabla\Phi(x(s))ds$. Substituting and rearranging yields:

$$\ddot{x}(t) + \left(\tfrac{2\dot{a}(t)}{a(t)} - \tfrac{\ddot{a}(t)}{\dot{a}(t)}\right)\dot{x}(t) + \left(\tfrac{(\dot{a}(t))^2}{a(t)}\right)\nabla\Phi(x(t)) = 0. \tag{5.18}$$

Compared to our closed-loop control system, the system in (5.18) is open-loop and lacks Hessian-driven damping. Moreover, $a(t)$ needs to be determined by hand and Song et al. [2021] do not establish existence or uniqueness of solutions.

5.3 Lyapunov Function

We construct a Lyapunov function that allows us to prove existence and uniqueness of a global solution of our closed-loop control system and to analyze convergence rates. As we will see, an analysis of the rate of decrease of the Lyapunov function together with the algebraic equation permit the derivation of new convergence rates for both the objective function gap and the squared gradient norm.

Existence and uniqueness of a global solution. Our main theorem on the existence and uniqueness of a global solution is summarized as follows.

Theorem 5.3.1 *Suppose that λ is absolutely continuous on any finite bounded interval. Then the closed-loop control system in Eq. (5.3) and Eq. (5.4) has a unique global solution, $(x, \lambda, a) : [0, +\infty) \mapsto \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$.*

Remark 5.3.2 *Intuitively, the feedback law $\lambda(\cdot)$, which we will show satisfies $\lambda(t) \to +\infty$ as $t \to +\infty$, links to the gradient norm $\|\nabla\Phi(x(\cdot))\|$ via the algebraic equation. Since we are interested in the worst-case convergence rate of solution trajectories, which corresponds to the worst-case iteration complexity of discrete-time algorithms, it is necessary that λ does not dramatically change. In open-loop Levenberg-Marquardt systems, Attouch and Svaiter [2011] impose the same condition on the regularization parameters. In closed-loop control systems, however, λ is not a given datum but an emergent component of the dynamics. Thus, it is preferable to prove that λ satisfies this condition rather than assuming it, as done in Attouch et al. [2013b, Theorem 5.2] and Attouch et al. [2016a, Theorem 2.4] for a closed-loop control system without inertia. The key step in their proof is to show that $\lambda(t) \leq \lambda(0)e^t$ locally by exploiting the specific structure of their system. This technical approach is, however, not applicable to our system due to the incorporation of the inertia term.*

Recall that the system in Eq. (5.3) and Eq. (5.4) can be equivalently written as the first-order system in time and space, as in Eq. (5.7). Accordingly, we define the following simple Lyapunov function:

$$\mathcal{E}(t) = a(t)(\Phi(x(t)) - \Phi(x^\star)) + \tfrac{1}{2}\|v(t) - x^\star\|^2, \tag{5.19}$$

where $x^\star$ is a global optimal solution of Φ.

Remark 5.3.3 *Note that the Lyapunov function (5.19) is composed of a sum of the mixed energy $\frac{1}{2}\|v(t) - x^*\|$ and the potential energy $a(t)(\Phi(x(t)) - \Phi(x^*))$. This function is similar to Lyapunov functions developed for analyzing the convergence of Newton-like dynamics [Attouch and Svaiter, 2011, Attouch et al., 2013b, Abbas et al., 2014, Attouch et al., 2016a] and the inertial gradient system with asymptotic vanishing damping [Su et al., 2016, Attouch et al., 2016b, Wilson et al., 2021, Shi et al., 2022]. Indeed, Wilson et al. [2021] construct a unified time-dependent Lyapunov function using the Bregman divergence and showed that*

their approach is equivalent to Nesterov's estimate sequence technique in a number of cases, including quasi-monotone subgradient, accelerated gradient descent and conditional gradient. Our Lyapunov function differs from existing choices in that v is not a standard momentum term depending on $\dot{x}$, but depends on x, λ and $\nabla\Phi$; see Eq. (5.7).

We provide two technical lemmas that characterize the descent property of $\mathcal{E}$ and the boundedness of the local solution $(x, v) : [0, t_0] \mapsto \mathbb{R}^d \times \mathbb{R}^d$.

Lemma 5.3.4 *Suppose that $(x, v, \lambda, a) : [0, t_0] \mapsto \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$ is a local solution of the first-order system in Eq. (5.7). Then, we have*

$$\tfrac{d\mathcal{E}(t)}{dt} \leq -a(t)\theta^{\frac{1}{p}}\|\nabla\Phi(x(t))\|^{\frac{p+1}{p}}, \quad \forall t \in [0, t_0].$$

Proof. By the definition, we have

$$\tfrac{d\mathcal{E}(t)}{dt} = \dot{a}(t)\Phi(x(t)) - \dot{a}(t)\Phi(x^\star) + \langle a(t)\dot{x}(t), \nabla\Phi(x(t))\rangle + \langle \dot{v}(t), v(t) - x^\star\rangle.$$

In addition, we have $\langle \dot{v}(t), v(t) - x^\star\rangle = \langle \dot{v}(t), v(t) - x(t)\rangle + \langle \dot{v}(t), x(t) - x^\star\rangle$ and $\dot{v}(t) = -\dot{a}(t)\nabla\Phi(x(t))$. Putting these pieces together yields:

$$\begin{aligned}\tfrac{d\mathcal{E}(t)}{dt} = &\underbrace{\dot{a}(t)(\Phi(x(t)) - \Phi(x^\star) - \langle\nabla\Phi(x(t)), x(t) - x^\star\rangle)}_{\mathbf{I}} \\ &+ \underbrace{\langle a(t)\dot{x}(t), \nabla\Phi(x(t))\rangle + \dot{a}(t)\langle x(t) - v(t), \nabla\Phi(x(t))\rangle}_{\mathbf{II}}.\end{aligned}$$

By the convexity of Φ, we have $\Phi(x(t)) - \Phi(x^\star) - \langle\nabla\Phi(x(t)), x(t) - x^\star\rangle \leq 0$. Since $\dot{a}(t) \geq 0$, we have $\mathbf{I} \leq 0$. Furthermore, Eq. (5.7) implies that

$$\dot{x}(t) + \tfrac{\dot{a}(t)}{a(t)}(x(t) - v(t)) = -\lambda(t)\nabla\Phi(x(t)),$$

which implies that

$$\mathbf{II} = \langle a(t)\dot{x}(t) + \dot{a}(t)x(t) - \dot{a}(t)v(t), \nabla\Phi(x(t))\rangle = -\lambda(t)a(t)\|\nabla\Phi(x(t))\|^2.$$

This together with the algebraic equation implies $\mathbf{II} \leq -a(t)\theta^{\frac{1}{p}}\|\nabla\Phi(x(t))\|^{\frac{p+1}{p}}$. Putting all these pieces together yields the desired inequality. □

Lemma 5.3.5 *Suppose that $(x, v, \lambda, a) : [0, t_0] \mapsto \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$ is a local solution of the first-order system in Eq. (5.7). Then, $(x(\cdot), v(\cdot))$ is bounded over the interval $[0, t_0]$ and the upper bound only depends on the initial condition.*

Proof. By Lemma 5.3.4, the function $\mathcal{E}$ is nonnegative and nonincreasing on the interval $[0, t_0]$. This implies that, for any $t \in [0, t_0]$, we have

$$\tfrac{1}{2}\|v(t) - x^\star\|^2 \leq a(t)(\Phi(x(t)) - \Phi(x^\star)) + \tfrac{1}{2}\|v(t) - x^\star\|^2 \leq \mathcal{E}(0).$$

Therefore, $v(\cdot)$ is bounded on the interval $[0, t_0]$ and the upper bound only depends on the initial condition. Furthermore, we have

$$a(t)(x(t) - x^\star) - a(0)(x_0 - x^\star) = \int_0^t (\dot{a}(s)(x(s) - x^\star) + a(s)\dot{x}(s))ds.$$

Using the triangle inequality and $a(0) = c^2$, we have

$$\begin{aligned}\|a(t)(x(t) - x^\star)\| &\leq c^2\|x_0 - x^\star\| + \int_0^t \|a(s)\dot{x}(s) + \dot{a}(t)x(s) - \dot{a}(s)x^\star\|ds \\ &\overset{\text{Eq. (5.7)}}{\leq} c^2\|x_0 - x^\star\| + \int_0^t \|\dot{a}(s)v(s) - \dot{a}(s)x^\star\|ds + \int_0^t \|\lambda(s)a(s)\nabla\Phi(x(s))\|ds.\end{aligned}$$

Note that $\|v(t) - x^\star\| \leq \sqrt{2\mathcal{E}(0)}$ is proved for all $t \in [0, t_0]$ and $a(t)$ is monotonically increasing with $a(0) = c^2$. Thus, the following inequality holds:

$$\begin{aligned}\|x(t) - x^\star\| &\leq \frac{c^2\|x_0 - x^\star\| + (a(t) - c^2)\sqrt{2\mathcal{E}(0)} + \int_0^t \lambda(s)a(s)\|\nabla\Phi(x(s))\|ds}{a(t)} \\ &\leq \|x_0 - x^\star\| + \sqrt{2\mathcal{E}(0)} + \tfrac{1}{a(t)}\int_0^t \lambda(s)a(s)\|\nabla\Phi(x(s))\|ds.\end{aligned}$$

By the Hölder inequality and using the fact that $a(t)$ is monotonically increasing, we have

$$\begin{aligned}&\int_0^t \lambda(s)a(s)\|\nabla\Phi(x(s))\|ds = \int_0^t \sqrt{\lambda(s)a(s)}(\sqrt{\lambda(s)a(s)}\|\nabla\Phi(x(s))\|)ds \\ &\leq \left(\int_0^t \lambda(s)a(s)ds\right)^{1/2}\left(\int_0^t \lambda(s)a(s)\|\nabla\Phi(x(s))\|^2 ds\right)^{1/2} \\ &\leq \sqrt{a(t)}\left(\int_0^t \sqrt{\lambda(s)}ds\right)\left(\int_0^t \lambda(s)a(s)\|\nabla\Phi(x(s))\|^2 ds\right)^{1/2} \\ &\leq a(t)\left(\int_0^t \lambda(s)a(s)\|\nabla\Phi(x(s))\|^2 ds\right)^{1/2}.\end{aligned}$$

The algebra equation implies that $\lambda(t)\|\nabla\Phi(x(t))\|^2 = \theta^{\frac{1}{p}}\|\nabla\Phi(x(t))\|^{\frac{p+1}{p}}$. Thus, by Lemma 5.3.4 again, we have

$$\int_0^t \lambda(s)a(s)\|\nabla\Phi(x(s))\|^2 ds = \int_0^t a(s)\theta^{\frac{1}{p}}\|\nabla\Phi(x(s))\|^{\frac{p+1}{p}} ds \leq \mathcal{E}(0).$$

Putting these pieces together yields that $\|x(t) - x^\star\| \leq \|x_0 - x^\star\| + 3\sqrt{\mathcal{E}(0)}$. Therefore, $x(t)$ is bounded on the interval $[0, t_0]$ and the upper bound only depends on the initial condition. This completes the proof. □

Proof of Theorem 6.2.6: We are ready to prove our main result on the existence and uniqueness of a global solution. In particular, let us consider a maximal solution of the closed-loop control system in Eq. (5.3) and Eq. (5.4):

$$(x, \lambda, a) : [0, T_{\max}) \mapsto \Omega \times (0, +\infty) \times (0, +\infty).$$

The existence of a maximal solution follows from a classical argument relying on the existence and uniqueness of a local solution (see Theorem 6.2.4).

It remains to show that the maximal solution is a global solution; that is, $T_{\max} = +\infty$, if λ is absolutely continuous on any finite bounded interval. Indeed, the property of λ guarantees that $\lambda(\cdot)$ is bounded on the interval $[0, T_{\max})$. By Lemma 5.3.5 and the equivalence between the closed-loop control system in Eq. (5.3) and Eq. (5.4) and the first-order system in Eq. (5.7), the solution trajectory $x(\cdot)$ is bounded on the interval $[0, T_{\max})$ and the upper bound only depends on the initial condition. This implies that $\dot{x}(\cdot)$ is also bounded on the interval $[0, T_{\max})$ by considering the system in the autonomous form of Eq. (5.11) and (5.12). Putting these pieces together yields that $x(\cdot)$ is Lipschitz continuous on $[0, T_{\max})$ and there exists $\bar{x} = \lim_{t \to T_{\max}} x(t)$.

If $T_{\max} < +\infty$, the absolute continuity of λ on any finite bounded interval implies that $\lambda(\cdot)$ is bounded on $[0, T_{\max}]$. This together with the algebraic equation implies that $\bar{x} \in \Omega$. However, by Theorem 6.2.4 with initial data $\bar{x}$, we can extend the solution to a strictly larger interval which contradicts the maximality of the aforementioned solution.

Rate of convergence. We establish a convergence rate for a global solution of the closed-loop control system in Eq. (5.3) and Eq. (5.4).

Theorem 5.3.6 *Suppose that $(x, \lambda, a) : [0, +\infty) \mapsto \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$ is a global solution of the closed-loop control system in Eq. (5.3) and Eq. (5.4). Then, the objective function gap satisfies*

$$\Phi(x(t)) - \Phi(x^\star) = O(t^{-\frac{3p+1}{2}}).$$

and the squared gradient norm satisfies

$$\inf_{0 \le s \le t} \|\nabla \Phi(x(s))\|^2 = O(t^{-3p}).$$

Remark 5.3.7 *This theorem shows that the convergence rate is $O(t^{-(3p+1)/2})$ in terms of objective function gap and $O(t^{-3p})$ in terms of squared gradient norm. Note that the former result does not imply the latter result but only gives a rate of $O(t^{-(3p+1)/2})$ for the squared gradient norm minimization even when $\Phi \in \mathcal{F}_\ell^1(\mathbb{R}^d)$ is assumed with $\|\nabla \Phi(x(t))\|^2 \le 2\ell(\Phi(x(t)) - \Phi(x^\star))$. In fact, the squared gradient norm minimization is generally of independent interest [Nesterov, 2012, Grapiglia and Nesterov, 2022b, Shi et al., 2022] and its analysis involves different techniques.*

The following lemma is a global version of Lemma 5.3.4 and the proof is exactly the same. Thus, we only state the result.

Lemma 5.3.8 *Suppose that* $(x, v, \lambda, a) : [0, +\infty) \mapsto \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$ *is a global solution of the first-order system in Eq.* (5.7)*. Then, we have*

$$\frac{d\mathcal{E}(t)}{dt} \leq -a(t)\theta^{\frac{1}{p}} \|\nabla\Phi(x(t))\|^{\frac{p+1}{p}}.$$

In view of Lemma 5.3.8, the key ingredient for analyzing the convergence rate in terms of both the objective function gap and the squared gradient norm is a lower bound on $a(t)$. We summarize this result in the following lemma.

Lemma 5.3.9 *Suppose that* $(x, v, \lambda, a) : [0, +\infty) \mapsto \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$ *is a global solution of the first-order system in Eq.* (5.7)*. Then, we have*

$$a(t) \geq \left(\frac{c}{2} + \left(\frac{\theta^{\frac{2}{3p+1}}}{(p+1)(\mathcal{E}(0))^{\frac{p-1}{3p+1}}} \right)^{\frac{3p+1}{4}} t^{\frac{3p+1}{4}} \right)^2.$$

Proof. For $p = 1$, the feedback control law is given by $\lambda(t) = \theta$, for $\forall t \in [0, +\infty)$, and

$$a(t) = \left(\frac{c}{2} + \frac{\sqrt{\theta}t}{2} \right)^2 = \left(\frac{c}{2} + \left(\frac{\theta^{\frac{2}{3p+1}}}{(p+1)(\mathcal{E}(0))^{\frac{p-1}{3p+1}}} \right)^{\frac{3p+1}{4}} t^{\frac{3p+1}{4}} \right)^2.$$

For $p \geq 2$, the algebraic equation implies that $\|\nabla\Phi(x(t))\| = (\frac{\theta^{1/p}}{\lambda(t)})^{\frac{p}{p-1}}$ since $\lambda(t) > 0$ for $\forall t \in [0, +\infty)$. This together with Lemma 5.3.8 implies that

$$\frac{d\mathcal{E}(t)}{dt} \leq -a(t)\theta^{\frac{1}{p}} \|\nabla\Phi(x(t))\|^{\frac{p+1}{p}} = -a(t)\theta^{\frac{2}{p-1}} [\lambda(t)]^{-\frac{p+1}{p-1}}.$$

Since $\mathcal{E}(t) \geq 0$, we have

$$\int_0^t a(s)\theta^{\frac{2}{p-1}} (\lambda(s))^{-\frac{p+1}{p-1}} ds \leq \mathcal{E}(0).$$

By the Hölder inequality, we have

$$\begin{aligned} \int_0^t (a(s))^{\frac{p-1}{3p+1}} ds &= \int_0^t (a(s)(\lambda(s))^{-\frac{p+1}{p-1}})^{\frac{p-1}{3p+1}} (\lambda(s))^{\frac{p+1}{3p+1}} ds \\ &\leq \left(\int_0^t a(s)(\lambda(s))^{-\frac{p+1}{p-1}} ds \right)^{\frac{p-1}{3p+1}} \left(\int_0^t \sqrt{\lambda(s)} ds \right)^{\frac{2p+2}{3p+1}}. \end{aligned}$$

Combining these results with the definition of a yields:

$$\begin{aligned} &\int_0^t (a(s))^{\frac{p-1}{3p+1}} ds \leq \theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}} \left(\int_0^t \sqrt{\lambda(s)} ds \right)^{\frac{2p+2}{3p+1}} \\ &\leq \theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}} (2\sqrt{a(t)} - c)^{\frac{2p+2}{3p+1}} \leq 2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}} \left(\sqrt{a(t)} - \frac{c}{2} \right)^{\frac{2p+2}{3p+1}}. \end{aligned}$$

Since $a(t)$ is nonnegative and nondecreasing with $\sqrt{a(0)} = \frac{c}{2}$, we have

$$\int_0^t \left(\sqrt{a(s)} - \tfrac{c}{2}\right)^{\frac{2p-2}{3p+1}} ds \leq 2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}} \left(\sqrt{a(t)} - \tfrac{c}{2}\right)^{\frac{2p+2}{3p+1}}. \tag{5.20}$$

The remaining steps in the proof are based on the Bihari-LaSalle inequality [LaSalle, 1949, Bihari, 1956]. In particular, we denote $y(\cdot)$ by $y(t) = \int_0^t (\sqrt{a(s)} - \frac{c}{2})^{\frac{2p-2}{3p+1}} ds$. Then, $y(0) = 0$ and Eq. (5.20) implies that

$$y(t) \leq 2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}} (\dot{y}(t))^{\frac{p+1}{p-1}}.$$

This implies that

$$\dot{y}(t) \geq \left(\frac{y(t)}{2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}}}\right)^{\frac{p-1}{p+1}} \Longrightarrow \frac{\dot{y}(t)}{(y(t))^{\frac{p-1}{p+1}}} \geq \left(\frac{1}{2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}}}\right)^{\frac{p-1}{p+1}}.$$

Integrating this inequality over $[0, t]$ yields:

$$(y(t))^{\frac{2}{p+1}} \geq \tfrac{2}{p+1} \left(\frac{1}{2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}}}\right)^{\frac{p-1}{p+1}} t.$$

Equivalently, by the definition of $y(t)$, we have

$$\int_0^t \left(\sqrt{a(s)} - \tfrac{c}{2}\right)^{\frac{2p-2}{3p+1}} ds \geq \left(\tfrac{2}{p+1}\right)^{\frac{p+1}{2}} \left(\frac{1}{2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}}}\right)^{\frac{p-1}{2}} t^{\frac{p+1}{2}}.$$

This together with Eq. (5.20) yields that

$$\begin{aligned}
\sqrt{a(t)} &\geq \tfrac{c}{2} + \left(\frac{1}{2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}}} \int_0^t \left(\sqrt{a(s)} - \tfrac{c}{2}\right)^{\frac{2p-2}{3p+1}} ds\right)^{\frac{3p+1}{2p+2}} \\
&\geq \tfrac{c}{2} + \left(\frac{\theta^{\frac{2}{3p+1}}}{(p+1)(\mathcal{E}(0))^{\frac{p-1}{3p+1}}}\right)^{\frac{3p+1}{4}} t^{\frac{3p+1}{4}}.
\end{aligned}$$

This completes the proof. □

Proof of Theorem 6.3.7: Since the first-order system in Eq. (5.7) is equivalent to the closed-loop control system in Eq. (5.3) and Eq. (5.4), $(x, \lambda, a) : [0, +\infty) \to \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$ is a global solution of the latter system with $x(0) = x_0 \in \Omega$. By Lemma 5.3.8, we have $\mathcal{E}(t) \leq \mathcal{E}(0)$ for $\forall t \geq 0$; that is,

$$a(t)(\Phi(x(t)) - \Phi(x^\star)) + \tfrac{1}{2}\|v(t) - x^\star\|^2 \leq \mathcal{E}(0).$$

Since $(x(0), v(0)) = (x_0, v_0)$ and $\|v(t) - x^\star\| \geq 0$, we have $a(t)(\Phi(x(t)) - \Phi(x^\star)) \leq \mathcal{E}(0)$. By Lemma 5.3.9, we have

$$\Phi(x(t)) - \Phi(x^\star) \leq \mathcal{E}(0) \left(\frac{c}{2} + \left(\frac{\theta^{\frac{2}{3p+1}}}{(p+1)(\mathcal{E}(0))^{\frac{p-1}{3p+1}}} \right)^{\frac{3p+1}{4}} t^{\frac{3p+1}{4}} \right)^{-2} = O(t^{-\frac{3p+1}{2}}).$$

By Lemma 5.3.8 and using the fact that $\mathcal{E}(t) \geq 0$ for $\forall t \in [0, +\infty)$, we have

$$\int_0^t a(s)\theta^{\frac{1}{p}} \|\nabla\Phi(x(s))\|^{\frac{p+1}{p}} ds \leq \mathcal{E}(0),$$

which implies that

$$\left(\inf_{0\leq s\leq t} \|\nabla\Phi(x(s))\|^{\frac{p+1}{p}} \right) \left(\int_0^t a(s)ds \right) \leq \theta^{-\frac{1}{p}} \mathcal{E}(0).$$

By Lemma 5.3.9, we obtain

$$\int_0^t a(s)ds \geq \int_0^t \left(\frac{c}{2} + \left(\frac{\theta^{\frac{2}{3p+1}}}{(p+1)(\mathcal{E}(0))^{\frac{p-1}{3p+1}}} \right)^{\frac{3p+1}{4}} s^{\frac{3p+1}{4}} \right)^2 ds.$$

In addition, $\inf_{0\leq s\leq t} \|\nabla\Phi(x(s))\|^{\frac{p+1}{p}} = (\inf_{0\leq s\leq t} \|\nabla\Phi(x(s))\|^2)^{\frac{p+1}{2p}}$. Putting these pieces together yields

$$\inf_{0\leq s\leq t} \|\nabla\Phi(x(s))\|^2 \leq \left(\frac{\theta^{-\frac{1}{p}}\mathcal{E}(0)}{\int_0^t (\frac{c}{2} + (\frac{\theta^{\frac{2}{3p+1}}}{(p+1)(\mathcal{E}(0))^{\frac{p-1}{3p+1}}})^{\frac{3p+1}{4}} s^{\frac{3p+1}{4}})^2 ds} \right)^{\frac{2p}{p+1}} = O(t^{-3p}).$$

This completes the proof.

Discussion. It is useful to compare our approach to time scaling methods [Attouch et al., 2019a,c, 2022a,c] and quasi-gradient methods [Bégout et al., 2015, Attouch et al., 2022b].

Regularity condition. Why is proving the existence and uniqueness of a global solution of the closed-loop control system in Eq. (5.3) and Eq. (5.4) hard without the regularity condition? Our system differs from the existing systems in three respects: (i) the appearance of both $\ddot{x}$ and $\dot{x}$; (ii) the algebraic equation that links λ and $\nabla\Phi(x)$; and (iii) the evolution dynamics depends on λ via a and $\dot{a}$. From a technical point of view, the combination of these features makes it challenging to control a lower bound on gradient norm $\|\nabla\Phi(x(\cdot))\|$ or an upper bound on the feedback control $\lambda(\cdot)$ on the local interval. In sharp contrast, $\|\nabla\Phi(x(t))\| \geq \|\nabla\Phi(x(0))\|e^{-t}$ or $\lambda(t) \leq \lambda(0)e^t$ can readily be derived for the Levenberg-Marquardt regularized system in Attouch and Svaiter [2011, Corollary 3.3] and even the

closed-loop control systems without inertia in Attouch et al. [2013b, Theorem 5.2] and Attouch et al. [2016a, Theorem 2.4]. Thus, we can not exclude the case of $\lambda(t) \to +\infty$ on the bounded interval without the regularity condition and we accordingly fail to establish global existence and uniqueness. We consider it an interesting open problem to derive the regularity condition rather than imposing it as an assumption.

Infinite-dimensional setting. It is promising to study our system using the techniques developed by Attouch et al. [2016b] for an infinite-dimensional setting. Our convergence analysis can in fact be extended directly, yielding the same rate of $O(1/t^{(3p+1)/2})$ in terms of objective function gap and $O(1/t^{3p})$ in terms of squared gradient norm in the Hilbert-space setting. However, the weak convergence of the solution trajectories is another matter. Note that Attouch et al. [2016b] studied the following open-loop system with the parameters (α, β):

$$\ddot{x}(t) + \tfrac{\alpha}{t}\dot{x}(t) + \beta\nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0.$$

The condition $\alpha > 3$ is crucial for proving weak convergence of solution trajectories and establishing strong convergence in various practical situations. Indeed, the convergence of the solution trajectory has not been established so far when $\alpha = 3$ (except in the one-dimensional case with $\beta = 0$; see Attouch et al. [2019b] for the reference). Unfortunately, when $c = 0$ and $p = 1$, the closed-loop control system in Eq. (5.3) and Eq. (5.4) becomes

$$\ddot{x}(t) + \tfrac{3}{t}\dot{x}(t) + \theta\nabla^2\Phi(x(t))\dot{x}(t) + \left(\theta + \tfrac{\theta}{t}\right)\nabla\Phi(x(t)) = 0.$$

The asymptotic damping coefficient $\frac{3}{t}$ does not satisfy the aforementioned condition in Attouch et al. [2016b], leaving doubt as to whether weak convergence holds true for the closed-loop control system in Eq. (5.3) and Eq. (5.4).

Time scaling. In the context of non-autonomous dissipative systems, time scaling is a simple yet universally powerful tool to accelerate the convergence of solution trajectories [Attouch et al., 2019a,c, 2022a,c]. Considering the general inertial gradient system in Eq. (5.3):

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \beta(t)\nabla^2\Phi(x(t))\dot{x}(t) + b(t)\nabla\Phi(x(t)) = 0,$$

the effect of time scaling is characterized by the coefficient parameter $b(t)$ which comes in as a factor of $\nabla\Phi(x(t))$. In Attouch et al. [2019a,c], the authors conducted an in-depth study of the convergence of this above system without Hessian-driven damping ($\beta = 0$). For the case $\alpha(t) = \frac{\alpha}{t}$, the convergence rate turns out to be $O(\frac{1}{t^2 b(t)})$ under certain conditions on the scalar α and $b(\cdot)$. Thus, a clear improvement can be achieved by taking $b(t) \to +\infty$. This demonstrates the power and potential of time scaling, as further evidenced by recent work on systems with Hessian damping [Attouch et al., 2022a] and other systems which are associated with the augmented Lagrangian formulation of the affine constrained convex minimization problem [Attouch et al., 2022c].

Comparing to our approach, the time scaling technique is based on an open-loop control regime, and indeed $b(t)$ is chosen by hand. In contrast, $\lambda(t)$ in our system is determined by the gradient of $\nabla\Phi(x(t))$ via the algebraic equation, and the evolution dynamics depend on λ via a and $\dot{a}$. The time scaling methodology accordingly does not capture the continuous-time interpretation of optimal acceleration in high-order optimization [Monteiro and Svaiter, 2013, Gasnikov et al., 2019a, Jiang et al., 2019, Bubeck et al., 2019]. In contrast, our algebraic equation provides a rigorous justification for the large-step condition in the existing algorithms [Monteiro and Svaiter, 2013, Gasnikov et al., 2019a, Jiang et al., 2019, Bubeck et al., 2019] when $p \geq 2$ and demonstrates the fundamental role that the feedback control plays in optimal acceleration, a role clarified by the continuous-time perspective.

Quasi-gradient approach and Kurdyka-Lojasiewicz (KL) theory. The quasi-gradient approach to inertial gradient systems were developed in Bégout et al. [2015] and recently applied by Attouch et al. [2022b] to analyze inertial dynamics with closed-loop control of the velocity. Recall that a vector field F is called a quasi-gradient for a function E if it has the same singular point as E and if the angle between the field F and the gradient ∇E remains acute and bounded away from $\frac{\pi}{2}$ (see the references [Huang, 2006, Chergui, 2008, Chill and Fašangová, 2010, Bárta et al., 2012, Bárta and Fašangová, 2016] for further geometrical interpretation).

Recent results in Bégout et al. [2015, Theorem 3.2] and Attouch et al. [2022b, Theorem 7.2] have suggested that the convergence properties for the bounded trajectories of quasi-gradient systems have been established if the function E is KL [Kurdyka, 1998, Bolte et al., 2010]. In Attouch et al. [2022b], the authors considered two closed-loop velocity control systems with a damping potential ϕ:

$$\ddot{x}(t) + \nabla\phi(\dot{x}(t)) + \nabla\Phi(x(t)) = 0. \tag{5.21}$$

$$\ddot{x}(t) + \nabla\phi(\dot{x}(t)) + \beta\nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0. \tag{5.22}$$

They proposed to use the Hamiltonian formulation of these systems and accordingly defined a function E_λ for $(x, v) = (x, \dot{x}(t))$ by

$$E_\eta(x, v) := \tfrac{1}{2}\|v\|^2 + \Phi(x) + \eta\langle\nabla\Phi(x), v\rangle.$$

If ϕ satisfies some certain growth conditions (see Attouch et al. [2022b, Theorem 7.3 and 9.2]), the systems in Eq. (5.21) and Eq. (5.22) both have a quasi-gradient structure for E_η for sufficiently small $\eta > 0$. This provides an elegant framework for analyzing the convergence properties of the systems in the form of Eq. (5.21) and Eq. (5.22).

Why is analyzing our system hard using the quasi-gradient approach? Our system differs from the systems in Eq. (5.21) and Eq. (5.22) in two aspects: (i) the closed-loop control law is designed for the gradient of Φ rather than the velocity $\dot{x}$; (ii) the damping coefficients are time dependent, depending on λ via a and $\dot{a}$, and do not have an analytic form for $p \geq 2$. Considering the systems in Eq. (5.7) and Eq. (5.8), we find that F is a time-dependent vector field which can not be tackled by the current quasi-gradient approach. We consider it an interesting open problem to develop a quasi-gradient approach for analyzing our system.

5.4 Implicit Time Discretization and Optimal Acceleration

We propose two conceptual algorithmic frameworks that arise via implicit time discretization of the closed-loop system in Eq. (5.7) and Eq. (5.8). Our approach demonstrates the importance of the large-step condition [Monteiro and Svaiter, 2013] for optimal acceleration, interpreting it as the discretization of the algebraic equation. This allows us to further clarify why this condition is unnecessary for first-order optimization algorithms in the case of $p = 1$ (the algebraic equation disappears). With an approximate tensor subroutine [Nesterov, 2021b], we derive two class of p-th order tensor algorithms, one of which recovers existing optimal p-th order tensor algorithms [Gasnikov et al., 2019a, Jiang et al., 2019, Bubeck et al., 2019] and the other of which leads to a new optimal p-th order tensor algorithm.

Conceptual algorithmic frameworks. We study two conceptual algorithmic frameworks which are derived by implicit time discretization of Eq. (5.7) with $c = 0$ and Eq. (5.8) with $c = 2$.

First algorithmic framework. By the definition of $a(t)$, we have $(\dot{a}(t))^2 = \lambda(t)a(t)$ and $a(0) = 0$. This implies an equivalent formulation of the first-order system in Eq. (5.7) with $c = 0$ as follows,

$$
\begin{cases}
\dot{v}(t) + \dot{a}(t)\nabla\Phi(x(t)) = 0 \\
\dot{x}(t) + \frac{\dot{a}(t)}{a(t)}(x(t) - v(t)) + \frac{(\dot{a}(t))^2}{a(t)}\nabla\Phi(x(t)) = 0 \\
a(t) = \frac{1}{4}(\int_0^t \sqrt{\lambda(s)}ds)^2 \\
(\lambda(t))^p \|\nabla\Phi(x(t))\|^{p-1} = \theta \\
(x(0), v(0)) = (x_0, v_0)
\end{cases}
$$

$$
\iff
\begin{cases}
\dot{v}(t) + \dot{a}(t)\nabla\Phi(x(t)) = 0 \\
a(t)\dot{x}(t) + \dot{a}(t)(x(t) - v(t)) + \lambda(t)a(t)\nabla\Phi(x(t)) = 0 \\
(\dot{a}(t))^2 = \lambda(t)a(t) \\
(\lambda(t))^p \|\nabla\Phi(x(t))\|^{p-1} = \theta \\
(x(0), v(0), a(0)) = (x_0, v_0, 0).
\end{cases}
$$

Algorithm 14 Conceptual Algorithmic Framework I

STEP 0: Let $x_0, v_0 \in \mathbb{R}^d$, $\sigma \in (0,1)$ and $\theta > 0$ be given, and set $A_0 = 0$ and $k = 0$.
STEP 1: If $0 = \nabla\Phi(x_k)$, then **stop**.
STEP 2: Otherwise, compute $\lambda_{k+1} > 0$ and a triple $(x_{k+1}, w_{k+1}, \epsilon_{k+1}) \in \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty)$ such that

$$\begin{aligned} &w_{k+1} \in \partial_{\epsilon_{k+1}}\Phi(x_{k+1}),\\ &\|\lambda_{k+1}w_{k+1} + x_{k+1} - \tilde{v}_k\|^2 + 2\lambda_{k+1}\epsilon_{k+1} \le \sigma^2\|x_{k+1} - \tilde{v}_k\|^2,\\ &\lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \ge \theta. \end{aligned}$$

where $\tilde{v}_k = \frac{A_k}{A_k+a_{k+1}}x_k + \frac{a_{k+1}}{A_k+a_{k+1}}v_k$ and $a_{k+1}^2 = \lambda_{k+1}(A_k + a_{k+1})$.
STEP 3: Compute $A_{k+1} = A_k + a_{k+1}$ and $v_{k+1} = v_k - a_{k+1}w_{k+1}$.
STEP 4: Set $k \leftarrow k+1$, and go to **STEP 1**.

We define discrete-time sequences, $\{(x_k, v_k, \lambda_k, a_k, A_k)\}_{k\ge0}$, that correspondx to the continuous-time sequences $\{(x(t), v(t), \lambda(t), \dot{a}(t), a(t))\}_{t\ge0}$. By implicit time discretization, we have

$$\begin{cases} v_{k+1} - v_k + a_{k+1}\nabla\Phi(x_{k+1}) = 0\\ A_{k+1}(x_{k+1} - x_k) + a_{k+1}(x_k - v_k) + \lambda_{k+1}A_{k+1}\nabla\Phi(x_{k+1}) = 0\\ (a_{k+1})^2 = \lambda_{k+1}(A_k + a_{k+1}),\ a_{k+1} = A_{k+1} - A_k,\ a_0 = 0\\ (\lambda_{k+1})^p\|\nabla\Phi(x_{k+1})\|^{p-1} = \theta. \end{cases} \tag{5.23}$$

By introducing a new variable $\tilde{v}_k = \frac{A_k}{A_k+a_{k+1}}x_k + \frac{a_{k+1}}{A_k+a_{k+1}}v_k$, the second and fourth lines of Eq. (5.23) can be equivalently reformulated as follows:

$$\lambda_{k+1}\nabla\Phi(x_{k+1}) + x_{k+1} - \tilde{v}_k = 0, \qquad \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} = \theta.$$

We propose to solve these two equations inexactly and replace $\nabla\Phi(x_{k+1})$ by a sufficiently accurate approximation in the first line of Eq. (5.23). In particular, the first equation can be equivalently written in the form of $\lambda_{k+1}w_{k+1}+x_{k+1}-\tilde{v}_k = 0$, where $w_{k+1} \in \{\nabla\Phi(x_{k+1})\}$. This motivates us to introduce a relative error tolerance [Solodov and Svaiter, 1999a, Monteiro and Svaiter, 2010]. In particular, we define the ε-subdifferential of a function f by

$$\partial_\epsilon f(x) := \{w \in \mathbb{R}^d \mid f(y) \ge f(x) + \langle y - x, w\rangle - \epsilon,\ \forall y \in \mathbb{R}^d\}, \tag{5.24}$$

and find $\lambda_{k+1} > 0$ and a triple $(x_{k+1}, w_{k+1}, \varepsilon_{k+1})$ such that $\|\lambda_{k+1}w_{k+1} + x_{k+1} - \tilde{v}_k\|^2 + 2\lambda_{k+1}\epsilon_{k+1} \le \sigma^2\|x_{k+1} - \tilde{v}_k\|^2$, where $w_{k+1} \in \partial_{\epsilon_{k+1}}\Phi(x_{k+1})$. To this end, w_{k+1} is a sufficiently accurate approximation of $\nabla\Phi(x_{k+1})$. Moreover, the second equation can be relaxed to $\lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \ge \theta$.

Remark 5.4.1 *We present our first conceptual algorithmic framework in Algorithm 14. This scheme includes the large-step A-HPE framework [Monteiro and Svaiter, 2013] as a special instance. Indeed, it reduces to the large-step A-HPE framework if we set $y = \tilde{y}$ and $p = 2$ and change the notation of $(x, v, \tilde{v}, w)$ to $(y, x, \tilde{x}, v)$ in Monteiro and Svaiter [2013].*

Second algorithmic framework. By the definition of $\gamma(t)$, we have $(\frac{\dot{\gamma}(t)}{\gamma(t)})^2 = \lambda(t)\gamma(t)$ and $\gamma(0) = 1$. This implies an equivalent formulation of the first-order system in Eq. (5.8) with $c = 2$:

$$\begin{cases} \dot{v}(t) - \frac{\dot{\gamma}(t)}{\gamma^2(t)}\nabla\Phi(x(t)) = 0 \\ \dot{x}(t) - \frac{\dot{\gamma}(t)}{\gamma(t)}(x(t) - v(t)) + \frac{(\dot{\gamma}(t))^2}{(\gamma(t))^3}\nabla\Phi(x(t)) = 0 \\ \gamma(t) = 4(\int_0^t \sqrt{\lambda(s)}ds + c)^{-2} \\ (\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta \\ (x(0), v(0)) = (x_0, v_0) \end{cases}$$

$$\iff \begin{cases} \dot{v}(t) + \frac{\alpha(t)}{\gamma(t)}\nabla\Phi(x(t)) = 0 \\ \dot{x}(t) + \alpha(t)(x(t) - v(t)) + \lambda(t)\nabla\Phi(x(t)) = 0 \\ (\alpha(t))^2 = \lambda(t)\gamma(t),\ \dot{\gamma}(t) + \alpha(t)\gamma(t) = 0 \\ (\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta \\ (x(0), v(0), \gamma(0)) = (x_0, v_0, 1). \end{cases}$$

We define discrete-time sequences, $\{(x_k, v_k, \lambda_k, \alpha_k, \gamma_k)\}_{k\geq 0}$, that correspondx to the continuous-time sequences $\{(x(t), v(t), \lambda(t), \alpha(t), \gamma(t))\}_{t\geq 0}$. From implicit time discretization, we have

$$\begin{cases} v_{k+1} - v_k + \frac{\alpha_{k+1}}{\gamma_{k+1}}\nabla\Phi(x_{k+1}) = 0 \\ x_{k+1} - x_k + \alpha_{k+1}(x_k - v_k) + \lambda_{k+1}\nabla\Phi(x_{k+1}) = 0 \\ (\alpha_{k+1})^2 = \lambda_{k+1}\gamma_{k+1},\ \gamma_{k+1} = (1 - \alpha_{k+1})\gamma_k,\ \gamma_0 = 1 \\ (\lambda_{k+1})^p\|\nabla\Phi(x_{k+1})\|^{p-1} = \theta. \end{cases} \tag{5.25}$$

By introducing a new variable $\tilde{v}_k = (1 - \alpha_{k+1})x_k + \alpha_{k+1}v_k$, the second and fourth lines of Eq. (5.23) can be equivalently reformulated as

$$\lambda_{k+1}\nabla\Phi(x_{k+1}) + x_{k+1} - \tilde{v}_k = 0, \qquad \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} = \theta.$$

By the same approximation strategy as before, we solve these two equations inexactly and replace $\nabla\Phi(x_{k+1})$ by a sufficiently accurate approximation in the first line of Eq. (5.25).

Remark 5.4.2 *We present our second conceptual algorithmic framework formally in Algorithm 15. To the best of our knowledge, this scheme does not appear in the literature and is based on an estimate sequence which differs from the one used in Algorithm 14. However, from a continuous-time perspective, these two algorithms are equivalent up to a constant $c > 0$, demonstrating that they achieve the same convergence rate in terms of both objective function gap and squared gradient norm.*

Algorithm 15 Conceptual Algorithmic Framework II

STEP 0: Let $x_0, v_0 \in \mathbb{R}^d$, $\sigma \in (0,1)$ and $\theta > 0$ be given, and set $\gamma_0 = 1$ and $k = 0$.
STEP 1: If $0 = \nabla\Phi(x_k)$, then **stop**.
STEP 2: Otherwise, compute $\lambda_{k+1} > 0$ and a triple $(x_{k+1}, w_{k+1}, \epsilon_{k+1}) \in \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty)$ such that

$$\begin{aligned} &w_{k+1} \in \partial_{\epsilon_{k+1}}\Phi(x_{k+1}), \\ &\|\lambda_{k+1}w_{k+1} + x_{k+1} - \tilde{v}_k\|^2 + 2\lambda_{k+1}\epsilon_{k+1} \leq \sigma^2\|x_{k+1} - \tilde{v}_k\|^2, \\ &\lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \geq \theta. \end{aligned}$$

where $\tilde{v}_k = (1-\alpha_{k+1})x_k + \alpha_{k+1}v_k$ and $(\alpha_{k+1})^2 = \lambda_{k+1}(1-\alpha_{k+1})\gamma_k$.
STEP 3: Compute $\gamma_{k+1} = (1-\alpha_{k+1})\gamma_k$ and $v_{k+1} = v_k - \frac{\alpha_{k+1}}{\gamma_{k+1}}w_{k+1}$.
STEP 4: Set $k \leftarrow k+1$, and go to **STEP 1**.

Comparison with Güler's accelerated proximal point algorithm. Algorithm 15 is related to Güler's accelerated proximal point algorithm (APPA) [Güler, 1992], which combines Nesterov acceleration [Nesterov, 1983] and Martinet's PPA [Martinet, 1970, 1972]. Indeed, the analogs of update formulas $\tilde{v}_k = (1-\alpha_{k+1})x_k + \alpha_{k+1}v_k$ and $(\alpha_{k+1})^2 = \lambda_{k+1}(1-\alpha_{k+1})\gamma_k$ appear in Güler's algorithm, suggesting similar dynamics. However, Güler's APPA does not specify how to choose $\{\lambda_k\}_{k\geq 0}$ but regard them as parameters, while our algorithm links its choice with the gradient norm of Φ via the large-step condition.

Such difference is emphasized by recent studies on the continuous-time perspective of Güler's APPA [Attouch et al., 2019c,a]. More specifically, Attouch et al. [2019a] proved that Güler's APPA can be interpreted as the implicit time discretization of an open-loop inertial gradient system (see Attouch et al. [2019a, Eq. (53)]):

$$\ddot{x}(t) + \left(g(t) - \tfrac{\dot{g}(t)}{g(t)}\right)\dot{x}(t) + \beta(t)\nabla\Phi(x(t)) = 0.$$

where g_k and β_k in their notation correspond to α_k and λ_k in Algorithm 15. By using $\gamma_{k+1} - \gamma_k = -\alpha_{k+1}\gamma_k$ and standard continuous-time arguments, we have $g(t) = -\frac{\dot{\gamma}(t)}{\gamma(t)}$ and $\beta(t) = \lambda(t) = \frac{(\dot{\gamma}(t))^2}{(\gamma(t))^3}$. By further defining $a(t) = \frac{1}{\gamma(t)}$, the above system is in the form of

$$\ddot{x}(t) + \left(\tfrac{2\dot{a}(t)}{a(t)} - \tfrac{\ddot{a}(t)}{\dot{a}(t)}\right)\dot{x}(t) + \left(\tfrac{(\dot{a}(t))^2}{a(t)}\right)\nabla\Phi(x(t)) = 0, \tag{5.26}$$

where a explicitly depends on the variable λ as follows,

$$a(t) = \tfrac{1}{4}(\textstyle\int_0^t \sqrt{\lambda(s)}ds + 2)^2.$$

Compared to our closed-loop control system, the one in Eq. (5.26) is open-loop without the algebra equation and does not contain Hessian-driven damping. The coefficient for the gradient term is also different, standing for different time rescaling in the evolution dynamics [Attouch et al., 2022a].

Complexity analysis. We study the iteration complexity of Algorithm 14 and 15. Our analysis is largely motivated by the aforementioned continuous-time analysis, simplifying the analysis in Monteiro and Svaiter [2013] for the case of $p = 2$ and generalizing it to the case of $p > 2$ in a systematic manner (see Theorem 5.4.3 and Theorem 5.4.6). We denote $x^\star$ as the projection of v_0 onto the solution set of Φ.

Algorithm 14. We start with the presentation of our main results for Algorithm 14, which in fact generalizes Monteiro and Svaiter [2013, Theorem 4.1] to the case of $p > 2$.

Theorem 5.4.3 *For every integer $k \geq 1$, the objective function gap satisfies*

$$\Phi(x_k) - \Phi(x^\star) = O(k^{-\frac{3p+1}{2}}),$$

and

$$\inf_{1\leq i\leq k} \|w_i\|^2 = O(k^{-3p}), \quad \inf_{1\leq i\leq k} \epsilon_i = O(k^{-\frac{3p+3}{2}}).$$

Note that the only difference between Algorithm 14 and large-step A-HPE framework in Monteiro and Svaiter [2013] is the order in the algebraic equation. As such, many of the technical results derived in Monteiro and Svaiter [2013] also hold for Algorithm 14; more specifically, Monteiro and Svaiter [2013, Theorem 3.6, Lemma 3.7 and Proposition 3.9].

We also present a technical lemma that provides a lower bound for A_k.

Lemma 5.4.4 *For $p \geq 1$ and every integer $k \geq 1$, we have*

$$A_k \geq \left(\frac{\theta(1-\sigma^2)^{\frac{p-1}{2}}}{(p+1)^{\frac{3p+1}{2}} \|v_0 - x^\star\|^{p-1}} \right) k^{\frac{3p+1}{2}}.$$

Proof. For $p = 1$, the large-step condition implies that $\lambda_k \geq \theta$ for all $k \geq 0$. By Monteiro and Svaiter [2013, Lemma 3.7], we have $A_k \geq \frac{\theta k^2}{4}$.

For $p \geq 2$, the large-step condition implies that

$$\begin{aligned}\sum_{i=1}^k A_i(\lambda_i)^{-\frac{p+1}{p-1}}\theta^{\frac{2}{p-1}} &\leq \sum_{i=1}^k A_i(\lambda_i)^{-\frac{p+1}{p-1}}(\lambda_i\|x_i - \tilde{v}_{i-1}\|^{p-1})^{\frac{2}{p-1}} \\ &= \sum_{i=1}^k \frac{A_i}{\lambda_i}\|x_i - \tilde{v}_{i-1}\|^2 \overset{\text{Monteiro and Svaiter [2013, Theorem 3.6]}}{\leq} \frac{\|v_0 - x^\star\|^2}{1-\sigma^2}.\end{aligned}$$

By the Hölder inequality, we have

$$\sum_{i=1}^k (A_i)^{\frac{p-1}{3p+1}} = \sum_{i=1}^k (A_i(\lambda_i)^{-\frac{p+1}{p-1}})^{\frac{p-1}{3p+1}}(\lambda_i)^{\frac{p+1}{3p+1}} \leq (\sum_{i=1}^k A_i(\lambda_i)^{-\frac{p+1}{p-1}})^{\frac{p-1}{3p+1}}(\sum_{i=1}^k \sqrt{\lambda_i})^{\frac{2p+2}{3p+1}}.$$

For the ease of presentation, we define $C = \theta^{-\frac{2}{3p+1}}(\frac{\|v_0-x^\star\|^2}{1-\sigma^2})^{\frac{p-1}{3p+1}}$. Putting these pieces together yields:

$$\sum_{i=1}^{k}(A_i)^{\frac{p-1}{3p+1}} \leq C\left(\sum_{i=1}^{k}\sqrt{\lambda_i}\right)^{\frac{2p+2}{3p+1}} \overset{\text{Monteiro and Svaiter [2013, Lemma 3.7]}}{\leq} 2C(A_k)^{\frac{p+1}{3p+1}}. \tag{5.27}$$

The remaining proof is based on the Bihari-LaSalle inequality in discrete time. In particular, we define $\{y_k\}_{k\geq 0}$ by $y_k = \sum_{i=1}^{k}(A_i)^{\frac{p-1}{3p+1}}$. Then, $y_0 = 0$ and Eq. (5.27) implies that

$$y_k \leq 2C(y_k - y_{k-1})^{\frac{p+1}{p-1}}.$$

This implies that

$$y_k - y_{k-1} \geq \left(\tfrac{y_k}{2C}\right)^{\frac{p-1}{p+1}} \Longrightarrow \tfrac{y_k-y_{k-1}}{(y_k)^{\frac{p-1}{p+1}}} \geq \left(\tfrac{1}{2C}\right)^{\frac{p-1}{p+1}}. \tag{5.28}$$

Inspired by the continuous-time inequality in Lemma 5.4.4, we claim that the following discrete-time inequality holds for every integer $k \geq 1$:

$$(y_k)^{\frac{2}{p+1}} - (y_{k-1})^{\frac{2}{p+1}} \geq \tfrac{2}{p+1}\left(\tfrac{y_k-y_{k-1}}{(y_k)^{\frac{p-1}{p+1}}}\right). \tag{5.29}$$

Indeed, we define $g(t) = 1 - t^{\frac{2}{p+1}}$ and find that this function is convex for $\forall t \in (0,1)$ since $p \geq 1$. Thus, we have

$$1 - t^{\frac{2}{p+1}} = g(t) - g(1) \geq (t-1)\nabla g(1) = \tfrac{2(1-t)}{p+1} \Longrightarrow \tfrac{1-t^{\frac{2}{p+1}}}{1-t} \geq \tfrac{2}{p+1}.$$

Since y_k is increasing, we have $\frac{y_{k-1}}{y_k} \in (0,1)$. Then, the desired Eq. (5.28) follows from setting $t = \frac{y_{k-1}}{y_k}$. Combining Eq. (5.28) and Eq. (5.29) yields that

$$(y_k)^{\frac{2}{p+1}} - (y_{k-1})^{\frac{2}{p+1}} \geq \tfrac{2}{p+1}\left(\tfrac{1}{2C}\right)^{\frac{p-1}{p+1}}.$$

Therefore, we conclude that

$$(y_k)^{\frac{2}{p+1}} = (y_0)^{\frac{2}{p+1}} + \left(\sum_{i=1}^{k}(y_i)^{\frac{2}{p+1}} - (y_{i-1})^{\frac{2}{p+1}}\right) \geq \tfrac{2}{p+1}\left(\tfrac{1}{2C}\right)^{\frac{p-1}{p+1}} k.$$

By the definition of y_k, we have

$$\sum_{i=1}^{k}(A_i)^{\frac{p-1}{3p+1}} \geq \left(\tfrac{2}{p+1}\right)^{\frac{p+1}{2}}\left(\tfrac{1}{2C}\right)^{\frac{p-1}{2}} k^{\frac{p+1}{2}}.$$

This together with Eq. (5.27) yields that

$$A_k \geq \left(\tfrac{1}{2C}\sum_{i=1}^{k}(A_i)^{\frac{p-1}{3p+1}}\right)^{\frac{3p+1}{p+1}} \geq \left(\tfrac{1}{(p+1)C}\right)^{\frac{3p+1}{2}} k^{\frac{3p+1}{2}}.$$

This completes the proof. □

Remark 5.4.5 *The proof of Lemma 5.4.4 is much simpler than the existing analysis; e.g., Monteiro and Svaiter [2013, Lemma 4.2] for the case of $p = 2$ and Jiang et al. [2019, Theorem 3.4] and Bubeck et al. [2019, Lemma 3.3] for the case of $p \geq 2$. Notably, it is not a generalization of the highly technical proof in Monteiro and Svaiter [2013, Lemma 4.2] but can be interpreted as the discrete-time counterpart of the proof of Lemma 5.3.9.*

Proof of Theorem 5.4.3: For every integer $k \geq 1$, by Monteiro and Svaiter [2013, Theorem 3.6] and Lemma 5.4.4, we have

$$\Phi(x_k) - \Phi(x^\star) \leq \tfrac{\|v_0 - x^\star\|^2}{2A_k} = O(k^{-\frac{3p+1}{2}}).$$

Combining Monteiro and Svaiter [2013, Proposition 3.9] and Lemma 5.4.4, we have

$$\inf_{1\leq i\leq k} \lambda_i \|w_i\|^2 \leq \tfrac{1+\sigma}{1-\sigma} \tfrac{\|v_0 - x^\star\|^2}{\sum_{i=1}^k A_i} = O(k^{-\frac{3p+3}{2}}),$$
$$\inf_{1\leq i\leq k} \varepsilon_i \leq \tfrac{\sigma^2}{2(1-\sigma^2)} \tfrac{\|v_0 - x^\star\|^2}{\sum_{i=1}^k A_i} = O(k^{-\frac{3p+3}{2}}).$$

In addition, we have $\|\lambda_i w_i + x_i - \tilde{v}_{i-1}\| \leq \sigma \|x_i - \tilde{v}_{i-1}\|$ and $\lambda_i \|x_i - \tilde{v}_{i-1}\|^{p-1} \geq \theta$. This implies that $\lambda_i \|w_i\|^{\frac{p-1}{p}} \geq \theta^{\frac{1}{p}}(1-\sigma)^{\frac{p-1}{p}}$. Putting these pieces together yields that $\inf_{1\leq i\leq k} \|w_i\|^{\frac{p+1}{p}} = O(k^{-\frac{3p+3}{2}})$ which implies that

$$\inf_{1\leq i\leq k} \|w_i\|^2 = \left(\inf_{1\leq i\leq k} \|w_i\|^{\frac{p+1}{p}}\right)^{\frac{2p}{p+1}} = O(k^{-3p}).$$

This completes the proof.

Algorithm 15. We now present our main results for Algorithm 15. The proof is analogous to that of Theorem 5.4.3 and based on another estimate sequence.

Theorem 5.4.6 *For every integer $k \geq 1$, the objective function gap satisfies*

$$\Phi(x_k) - \Phi(x^\star) = O(k^{-\frac{3p+1}{2}})$$

and

$$\inf_{1\leq i\leq k} \|w_i\|^2 = O(k^{-3p}), \quad \inf_{1\leq i\leq k} \epsilon_i = O(k^{-\frac{3p+3}{2}}).$$

Inspired by the continuous-time Lyapunov function in Eq. (5.19), we construct a discrete-time Lypanunov function for Algorithm 15 as follows:

$$\mathcal{E}_k = \tfrac{1}{\gamma_k}(\Phi(x_k) - \Phi(x^\star)) + \tfrac{1}{2}\|v_k - x^\star\|^2. \tag{5.30}$$

We use this function to prove technical results that pertain to Algorithm 15 and which are the analogs of Monteiro and Svaiter [2013, Theorem 3.6, Lemma 3.7 and Proposition 3.9].

Lemma 5.4.7 *For every integer $k \geq 1$,*

$$\frac{1-\sigma^2}{2}\left(\sum_{i=1}^{k}\frac{1}{\lambda_i\gamma_i}\|x_i-\tilde{v}_{i-1}\|^2\right)\leq \mathcal{E}_0-\mathcal{E}_k,$$

which implies that

$$\Phi(x_k)-\Phi(x^\star)\leq \gamma_k\mathcal{E}_0,\quad \|v_k-x^\star\|\leq\sqrt{2\mathcal{E}_0}.$$

Assuming that $\sigma<1$, we have $\sum_{i=1}^{k}\frac{1}{\lambda_i\gamma_i}\|x_i-\tilde{v}_{i-1}\|^2\leq\frac{2\mathcal{E}_0}{1-\sigma^2}$.

Proof. It suffices to prove the first inequality which implies the other results. Based on the discrete-time Lyapunov function, we define two functions $\phi_k:\mathbb{R}^d\mapsto\mathbb{R}$ and $\Gamma_k:\mathbb{R}^d\mapsto\mathbb{R}$ by (Γ_k is related to $\mathcal{E}_k$ and defined recursively):

$$\begin{aligned}&\phi_k(v)=\Phi(x_k)+\langle v-x_k,w_k\rangle-\epsilon_k-\Phi(x^\star),\ \forall k\geq 0,\\ &\Gamma_0(v)=\tfrac{1}{\gamma_0}(\Phi(x_0)-\Phi(x^\star))+\tfrac{1}{2}\|v-v_0\|^2,\ \Gamma_{k+1}=\Gamma_k+\tfrac{\alpha_{k+1}}{\gamma_{k+1}}\phi_{k+1},\ \forall k\geq 0.\end{aligned}$$

First, by definition, ϕ_k is affine. Since $w_{k+1}\in\partial_{\epsilon_{k+1}}\Phi(x_{k+1})$, Eq. (5.24) implies that $\phi_k(v)\leq\Phi(v)-\Phi(x^\star)$. Furthermore, Γ_k is quadratic and $\nabla^2\Gamma_k=\nabla^2\Gamma_0$ since ϕ_k is affine. Then, we prove that $\Gamma_k(v)\leq\Gamma_0(v)+\frac{1-\gamma_k}{\gamma_k}(\Phi(v)-\Phi(x^\star))$ using induction. Indeed, it holds when $k=0$ since $\gamma_0=1$. Assuming that this inequality holds for $\forall i\leq k$, we derive from $\phi_k(v)\leq\Phi(v)-\Phi(x^\star)$ and $\gamma_{k+1}=(1-\alpha_{k+1})\gamma_k$ that

$$\Gamma_{k+1}(v)\leq\Gamma_0(v)+\left(\tfrac{1-\gamma_k}{\gamma_k}+\tfrac{\alpha_{k+1}}{\gamma_{k+1}}\right)(\Phi(v)-\Phi(x^\star))=\Gamma_0(v)+\tfrac{1-\gamma_k}{\gamma_k}(\Phi(v)-\Phi(x^\star)).$$

Finally, we prove that $v_k=\operatorname{argmin}_{v\in\mathbb{R}^d}\Gamma_k(v)$ using the induction. Indeed, it holds when $k=0$. Suppose that this inequality holds for $\forall i\leq k$, we have

$$\nabla\Gamma_{k+1}(v)=\nabla\Gamma_k(v)+\tfrac{\alpha_{k+1}}{\gamma_{k+1}}\nabla\phi_{k+1}(v)=v-v_k+\tfrac{\alpha_{k+1}}{\gamma_{k+1}}w_{k+1}.$$

Using the definition of v_k and the fact that $\gamma_{k+1}=(1-\alpha_{k+1})\gamma_k$, we have $\nabla\Gamma_{k+1}(v)=0$ if and only if $v=v_{k+1}$.

The remaining proof is based on the gap sequence $\{\beta_k\}_{k\geq 0}$ which is defined by $\beta_k=\inf_{v\in\mathbb{R}^d}\Gamma_k(v)-\frac{1}{\gamma_k}(\Phi(x_k)-\Phi(x^\star))$. Using the previous facts that Γ_k is quadratic with $\nabla^2\Gamma_k=1$ and the upper bound for $\Gamma_k(v)$, we have

$$\beta_k=\Gamma_k(x^\star)-\tfrac{1}{\gamma_k}(\Phi(x_k)-\Phi(x^\star))-\tfrac{1}{2}\|x^\star-v_k\|^2\leq\Gamma_0(x^\star)-\mathcal{E}_k=\mathcal{E}_0-\mathcal{E}_k.$$

By definition, we have $\beta_0=0$. Thus, it suffices to prove that the following recursive inequality holds true for every integer $k\geq 0$,

$$\beta_{k+1}\geq\beta_k+\tfrac{1-\sigma^2}{2\lambda_{k+1}\gamma_{k+1}}\|x_{k+1}-\tilde{v}_k\|^2. \tag{5.31}$$

In particular, we define $\tilde{v} = (1-\alpha_{k+1})x_k + \alpha_{k+1}v$ for any given $v \in \mathbb{R}^d$. Using the definition of $\tilde{v}_k$ and the affinity of ϕ_{k+1}, we have

$$\phi_{k+1}(\tilde{v}) = (1-\alpha_{k+1})\phi_{k+1}(x_k) + \alpha_{k+1}\phi_{k+1}(v), \tag{5.32}$$
$$\tilde{v} - \tilde{v}_k = \alpha_{k+1}(v - v_k). \tag{5.33}$$

Since Γ_k is quadratic with $\nabla^2\Gamma_k = 1$, we have $\Gamma_k(v) = \Gamma_k(v_k) + \frac{1}{2}\|v - v_k\|^2$. Plugging this into the recursive equation for Γ_k yields that

$$\Gamma_{k+1}(v) = \Gamma_k(v_k) + \tfrac{1}{2}\|v - v_k\|^2 + \tfrac{\alpha_{k+1}}{\gamma_{k+1}}\phi_{k+1}(v).$$

By the definition of β_k, we have $\Gamma_k(v_k) = \beta_k + \frac{1}{\gamma_k}(\Phi(x_k) - \Phi(x^\star))$. Putting these pieces together with the definition of $\mathcal{E}_k$ yields that

$$\Gamma_{k+1}(v) = \beta_k + \tfrac{\alpha_{k+1}}{\gamma_{k+1}}\phi_{k+1}(v) + \tfrac{1}{\gamma_k}(\Phi(x_k) - \Phi(x^\star)) + \tfrac{1}{2}\|v - v_k\|^2.$$

Since $\phi_{k+1}(v) \le \Phi(v) - \Phi(x^\star)$, we have

$$\begin{aligned}
\Gamma_{k+1}(v) &\ge \beta_k + \tfrac{\alpha_{k+1}}{\gamma_{k+1}}\phi_{k+1}(v) + \tfrac{1}{\gamma_k}\phi_{k+1}(x_k) + \tfrac{1}{2}\|v - v_k\|^2 \\
&\overset{\text{Eq. (5.32)}}{=} \beta_k + \tfrac{1}{\gamma_{k+1}}\phi_{k+1}(\tilde{v}) + \tfrac{1}{2}\|v - v_k\|^2 \\
&= \beta_k + \tfrac{1}{\gamma_{k+1}}\left(\phi_{k+1}(\tilde{v}) + \tfrac{\gamma_{k+1}}{2}\|v - v_k\|^2\right) \\
&\overset{\text{Eq. (5.33)}}{=} \beta_k + \tfrac{1}{\gamma_{k+1}}\left(\phi_{k+1}(\tilde{v}) + \tfrac{\gamma_{k+1}}{2(\alpha_{k+1})^2}\|\tilde{v} - \tilde{v}_k\|^2\right) \\
&= \beta_k + \tfrac{1}{\gamma_{k+1}}\left(\phi_{k+1}(\tilde{v}) + \tfrac{1}{2\lambda_{k+1}}\|\tilde{v} - \tilde{v}_k\|^2\right).
\end{aligned}$$

Using Monteiro and Svaiter [2013, Lemma 3.3] with $\lambda = \lambda_{k+1}$, $\tilde{v} = \tilde{v}_k$, $\tilde{x} = x_{k+1}$, $\tilde{w} = w_{k+1}$ and $\epsilon = \epsilon_{k+1}$, we have

$$\inf_{v\in\mathbb{R}^d}\left\{\langle v - x_{k+1}, w_{k+1}\rangle - \epsilon_{k+1} + \tfrac{1}{2\lambda_{k+1}}\|v - \tilde{v}_k\|^2\right\} \ge \tfrac{1-\sigma^2}{2\lambda_{k+1}}\|x_{k+1} - \tilde{v}_k\|^2.$$

which implies that

$$\phi_{k+1}(\tilde{v}) + \tfrac{1}{2\lambda_{k+1}}\|\tilde{v} - \tilde{v}_k\|^2 - \tfrac{1}{\gamma_{k+1}}(\Phi(x_{k+1}) - \Phi(x^\star)) \ge \tfrac{1-\sigma^2}{2\lambda_{k+1}}\|x_{k+1} - \tilde{v}_k\|^2.$$

Putting these pieces together yields that

$$\inf_{v\in\mathbb{R}^d}\Gamma_{k+1}(v) - \tfrac{1}{\gamma_{k+1}}(\Phi(x_{k+1}) - \Phi(x^\star)) \ge \beta_k + \tfrac{1-\sigma^2}{2\lambda_{k+1}\gamma_{k+1}}\|x_{k+1} - \tilde{v}_k\|^2.$$

which together with the definition of β_k yields the desired inequality in Eq. (5.31). □

Lemma 5.4.8 *For every integer $k \ge 0$, it holds that*

$$\sqrt{\tfrac{1}{\gamma_{k+1}}} \ge \sqrt{\tfrac{1}{\gamma_k}} + \tfrac{1}{2}\sqrt{\lambda_{k+1}}.$$

As a consequence, the following statements hold: (i) For every integer $k \ge 0$, it holds that $\gamma_k \le (1 + \frac{1}{2}\sum_{j=1}^k \sqrt{\lambda_j})^{-2}$; (ii) If $\sigma < 1$ is further assumed, we have $\sum_{j=1}^k \|x_j - \tilde{v}_{j-1}\|^2 \le \frac{2\mathcal{E}_0}{1-\sigma^2}$.

Proof. It suffices to prove the first inequality which implies the other results. By the definition of $\{\gamma_k\}_{k\geq 0}$ and $\{\alpha_k\}_{k\geq 0}$, we have $\gamma_{k+1} = (1-\alpha_{k+1})\gamma_k$ and $(\alpha_{k+1})^2 = \lambda_{k+1}\gamma_{k+1}$. This implies that

$$\frac{1}{\gamma_k} = \frac{1}{\gamma_{k+1}} - \frac{\alpha_{k+1}}{\gamma_{k+1}} = \frac{1}{\gamma_{k+1}} - \sqrt{\frac{\lambda_{k+1}}{\gamma_{k+1}}}.$$

Since $\gamma_k > 0$ and $\lambda_k > 0$, we have $\sqrt{\frac{1}{\gamma_{k+1}}} \geq \frac{1}{2}\sqrt{\lambda_{k+1}}$ and

$$\frac{1}{\gamma_k} \leq \frac{1}{\gamma_{k+1}} - \sqrt{\frac{\lambda_{k+1}}{\gamma_{k+1}}} + \frac{\lambda_{k+1}}{4} = \left(\sqrt{\frac{1}{\gamma_{k+1}}} - \frac{1}{2}\sqrt{\lambda_{k+1}}\right)^2.$$

which implies the desired inequality. □

Lemma 5.4.9 *For every integer $k \geq 1$ and $\sigma < 1$, there exists $1 \leq i \leq k$ such that*

$$\inf_{1\leq i\leq k} \sqrt{\lambda_i}\|w_i\| \leq \sqrt{\frac{1+\sigma}{1-\sigma}}\sqrt{\frac{2\mathcal{E}_0}{\sum_{i=1}^k \frac{1}{\gamma_i}}}, \quad \inf_{1\leq i\leq k} \epsilon_i \leq \frac{\sigma^2}{2(1-\sigma^2)}\frac{2\mathcal{E}_0}{\sum_{i=1}^k \frac{1}{\gamma_i}}.$$

Proof. With the convention $0/0 = 0$, we define $\tau_k = \max\{\frac{2\epsilon_k}{\sigma^2}, \frac{\lambda_k\|w_k\|^2}{(1+\sigma)^2}\}$ for every integer $k \geq 1$. Then, we have

$$\begin{aligned}
&2\lambda_k\epsilon_k \leq \sigma^2\|x_k - \tilde{v}_{k-1}\|^2,\\
&\|\lambda_k w_k\| \leq \|\lambda_k w_k + x_k - \tilde{v}_{k-1}\| + \|x_k - \tilde{v}_{k-1}\| \leq (1+\sigma)\|x_k - \tilde{v}_{k-1}\|.
\end{aligned}$$

which implies that $\lambda_k\tau_k \leq \|x_k - \tilde{v}_{k-1}\|^2$ for every integer $k \geq 1$. This together with Lemma 5.4.7 yields that

$$\frac{2\mathcal{E}_0}{1-\sigma^2} \geq \sum_{i=1}^k \frac{1}{\lambda_i\gamma_i}\|x_i - \tilde{v}_{i-1}\|^2 \geq \left(\inf_{1\leq i\leq k} \tau_i\right)\left(\sum_{i=1}^k \frac{1}{\gamma_i}\right).$$

Combining this inequality with the definition of τ_k yields the desired results. □

As the analog of Lemma 5.4.4, we provide a technical lemma on the upper bound for γ_k. The analysis is based on the same idea for proving Lemma 5.4.4 and is motivated by continuous-time analysis for the first-order system in Eq. (5.8).

Lemma 5.4.10 *For $p \geq 1$ and every integer $k \geq 1$, we have*

$$\gamma_k \leq \frac{(p+1)^{\frac{3p+1}{2}}}{\theta}\left(\frac{2\mathcal{E}_0}{1-\sigma^2}\right)^{\frac{p-1}{2}} k^{-\frac{3p+1}{2}}.$$

Proof. For $p = 1$, the large-step condition implies that $\lambda_k \geq \theta$ for all $k \geq 0$. By Lemma 5.4.8, we have $\gamma_k \leq \frac{4}{\theta k^2}$. For $p \geq 2$, the large-step condition implies that

$$\begin{aligned}
\sum_{i=1}^k (\gamma_i)^{-1}(\lambda_i)^{-\frac{p+1}{p-1}}\theta^{\frac{2}{p-1}} &\leq \sum_{i=1}^k (\gamma_i)^{-1}(\lambda_i)^{-\frac{p+1}{p-1}}(\lambda_i\|x_i - \tilde{v}_{i-1}\|^{p-1})^{\frac{2}{p-1}}\\
&= \sum_{i=1}^k \frac{1}{\lambda_i\gamma_i}\|x_i - \tilde{v}_{i-1}\|^2 \overset{\text{Lemma 5.4.7}}{\leq} \frac{2\mathcal{E}_0}{1-\sigma^2}.
\end{aligned}$$

By the Hölder inequality, we have

$$\sum_{i=1}^{k}(\gamma_i)^{-\frac{p-1}{3p+1}} = \sum_{i=1}^{k}\left(\frac{1}{(\lambda_i)^{\frac{p+1}{p-1}}\gamma_i}\right)^{\frac{p-1}{3p+1}}(\lambda_i)^{\frac{p+1}{3p+1}} \leq \left(\sum_{i=1}^{k}\frac{1}{(\lambda_i)^{\frac{p+1}{p-1}}\gamma_i}\right)^{\frac{p-1}{3p+1}}\left(\sum_{i=1}^{k}\sqrt{\lambda_i}\right)^{\frac{2p+2}{3p+1}}.$$

For ease of presentation, we define $C = \theta^{-\frac{2}{3p+1}}(\frac{2\mathcal{E}_0}{1-\sigma^2})^{\frac{p-1}{3p+1}}$. Putting these pieces together yields that

$$\sum_{i=1}^{k}(\gamma_i)^{-\frac{p-1}{3p+1}} \leq C\left(\sum_{i=1}^{k}\sqrt{\lambda_i}\right)^{\frac{2p+2}{3p+1}} \overset{\text{Lemma 5.4.8}}{\leq} 2C(\gamma_k)^{-\frac{p+1}{3p+1}}. \tag{5.34}$$

Using the same argument for proving Lemma 5.4.4, we have

$$\sum_{i=1}^{k}(\gamma_i)^{-\frac{p-1}{3p+1}} \geq \left(\frac{2}{p+1}\right)^{\frac{p+1}{2}}\left(\frac{1}{2C}\right)^{\frac{p-1}{2}}k^{\frac{p+1}{2}}.$$

This together with Eq. (5.34) yields that

$$\frac{1}{\gamma_k} \geq \left(\frac{1}{2C}\sum_{i=1}^{k}(\gamma_i)^{-\frac{p-1}{3p+1}}\right)^{\frac{3p+1}{p+1}} \geq \left(\frac{1}{(p+1)C}\right)^{\frac{3p+1}{2}}k^{\frac{3p+1}{2}}.$$

This completes the proof. □

Proof of Theorem 5.4.6: For every integer $k \geq 1$, by Lemma 5.4.7 and Lemma 5.4.10, we have

$$\Phi(x_k) - \Phi(x^\star) \leq \gamma_k\mathcal{E}_0 = O(k^{-\frac{3p+1}{2}}).$$

By Lemma 5.4.9 and Lemma 5.4.10, we have

$$\inf_{1\leq i\leq k}\lambda_i\|w_i\|^2 \leq \frac{1+\sigma}{1-\sigma}\frac{2\mathcal{E}_0}{\sum_{i=1}^{k}\frac{1}{\gamma_i}} = O(k^{-\frac{3p+3}{2}}),$$
$$\inf_{1\leq i\leq k}\epsilon_i \leq \frac{\sigma^2}{2(1-\sigma^2)}\frac{2\mathcal{E}_0}{\sum_{i=1}^{k}\gamma_i} = O(k^{-\frac{3p+3}{2}}).$$

As in the proof of Theorem 5.4.3, we conclude that $\inf_{1\leq i\leq k}\|w_i\|^2 = O(k^{-3p})$.

Remark 5.4.11 *The discrete-time analysis here is based on a discrete-time Lyapunov function in Eq. (5.30), which is closely related to the continuous one in Eq. (5.19), and two simple yet nontrivial technical lemmas (see Lemma 5.4.4 and 5.4.10), which are both discrete-time versions of Lemma 5.3.9. Notably, the proofs of Lemma 5.4.4 and 5.4.10 follows the same path for proving Lemma 5.3.9 and have demanded the use of the Bihari-LaSalle inequality in discrete time.*

Optimal algorithms and gradient norm minimization. By instantiating Algorithm 14 and 15 with approximate tensor subroutines, we develop two families of optimal p-th order tensor algorithms for minimizing the function $\Phi \in \mathcal{F}_\ell^p(\mathbb{R}^d)$. The former one include all of existing optimal p-th order tensor algorithms [Gasnikov et al., 2019a, Jiang et al., 2019, Bubeck et al., 2019] while the latter one is new to our knowledge. Moreover, we provide one hitherto unknown result that the optimal p-th order tensor algorithms minimize the squared gradient norm at a rate of $O(k^{-3p})$. Our results extend those for first-order algorithms [Shi et al., 2022] and second-order algorithms [Monteiro and Svaiter, 2013].

Approximate tensor subroutine. The celebrated proximal point algorithms [Rockafellar, 1976, Güler, 1992] (corresponding to implicit time discretization of certain systems) require solving an exact proximal iteration with proximal coefficient $\lambda > 0$ at each iteration:

$$x = \operatorname*{argmin}_{u \in \mathbb{R}^d} \left\{ \Phi(u) + \tfrac{1}{2\lambda}\|u - v\|^2 \right\}. \tag{5.35}$$

In general, Eq. (6.38) can be as hard as minimizing the function Φ when the proximal coefficient $\lambda \to +\infty$. Fortunately, when $\Phi \in \mathcal{F}_\ell^p(\mathbb{R}^d)$, it suffices to solve the subproblem that minimizes the sum of the p-th order Taylor approximation of Φ and a regularization term, motivating a line of p-th order tensor algorithms [Baes, 2009, Birgin et al., 2016, 2017, Martínez, 2017, Jiang et al., 2020, Nesterov, 2021b, Gasnikov et al., 2019a, Jiang et al., 2019, Bubeck et al., 2019]. More specifically, we define

$$\Phi_v(u) = \Phi(v) + \langle \nabla\Phi(v), u - v\rangle + \sum_{j=2}^{p} \tfrac{1}{j!}\nabla^{(j)}\Phi(v)[u-v]^j + \tfrac{\ell\|u-v\|^{p+1}}{(p+1)!}.$$

Our proposed algorithms are based on either an inexact solution of Eq. (5.36a), used in Jiang et al. [2019], or an exact solution of Eq. (5.36b), used in Gasnikov et al. [2019a] and Bubeck et al. [2019]:

$$\min_{u \in \mathbb{R}^d} \ \Phi_v(u) + \tfrac{1}{2\lambda}\|u - v\|^2, \tag{5.36a}$$

$$\min_{u \in \mathbb{R}^d} \ \Phi_v(u). \tag{5.36b}$$

In particular, the solution x_v of Eq. (5.36a) is unique and satisfies $\lambda\nabla\Phi_v(x_v) + x_v - v = 0$. Thus, we denote a *$\hat{\sigma}$-inexact solution* of Eq. (5.36a) by a vector $x \in \mathbb{R}^d$ satisfying that $\|\lambda\nabla\Phi_v(x) + x - v\| \le \hat{\sigma}\|x - v\|$ use either it or an exact solution of Eq. (5.36b).

First algorithm. We present the first optimal p-th order tensor algorithm in Algorithm 16 and prove that it is Algorithm 14 with specific choice of θ.

Proposition 5.4.12 *Algorithm 16 is Algorithm 14 with $\theta = \frac{\sigma_I p!}{2\ell}$ or $\theta = \frac{(p-1)!}{2\ell}$.*

Algorithm 16 Optimal p-th order Tensor Algorithm I [Gasnikov et al., 2019a, Jiang et al., 2019, Bubeck et al., 2019]

STEP 0: Let $x_0, v_0 \in \mathbb{R}^d$, $\hat{\sigma} \in (0,1)$ and $0 < \sigma_l < \sigma_u < 1$ such that $\sigma_l(1+\hat{\sigma})^{p-1} < \sigma_u(1-\hat{\sigma})^{p-1}$ and $\sigma = \hat{\sigma} + \sigma_u < 1$ be given, and set $A_0 = 0$ and $k = 0$.
STEP 1: If $0 = \nabla\Phi(x_k)$, then **stop**.
STEP 2: Otherwise, compute a positive scalar λ_{k+1} with a $\hat{\sigma}$-inexact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (5.36a) satisfying that

$$\tfrac{\sigma_l p!}{2\ell} \leq \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \leq \tfrac{\sigma_u p!}{2\ell},$$

or an exact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (5.36b) satisfying that

$$\tfrac{(p-1)!}{2\ell} \leq \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \leq \tfrac{p!}{\ell(p+1)},$$

where $\tilde{v}_k = \frac{A_k}{A_k + a_{k+1}} x_k + \frac{a_{k+1}}{A_k + a_{k+1}} v_k$ and $a_{k+1}^2 = \lambda_{k+1}(A_k + a_{k+1})$.
STEP 3: Compute $A_{k+1} = A_k + a_{k+1}$ and $v_{k+1} = v_k - a_{k+1}\nabla\Phi(x_{k+1})$.
STEP 4: Set $k \leftarrow k+1$, and go to **STEP 1**.

Proof. Given that a pair $(x_k, v_k)_{k\geq 1}$ is generated by Algorithm 16, we define $w_k = \nabla\Phi(x_k)$ and $\varepsilon_k = 0$. Then $v_{k+1} = v_k - a_{k+1}\nabla\Phi(x_{k+1}) = v_k - a_{k+1}w_{k+1}$. Using Jiang et al. [2019, Proposition 3.2] with a $\hat{\sigma}$-inexact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (5.36a) at $(\lambda_{k+1}, \tilde{v}_k)$, a triple $(x_{k+1}, w_{k+1}, \varepsilon_{k+1}) \in \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty)$ satisfies that

$$w_{k+1} \in \partial_{\epsilon_{k+1}}\Phi(x_{k+1}), \quad \|\lambda_{k+1}w_{k+1} + x_{k+1} - \tilde{v}_k\|^2 + 2\lambda_{k+1}\epsilon_{k+1} \leq \sigma^2\|x_{k+1} - \tilde{v}_k\|^2.$$

Since $\theta = \frac{\sigma_l p!}{2\ell} \in (0,1)$ and $\sigma = \hat{\sigma} + \sigma_u < 1$, we have

$$\begin{aligned}
\lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \leq \tfrac{\sigma_u p!}{2\ell} \quad &\Longrightarrow \quad \hat{\sigma} + \tfrac{2\ell\lambda_{k+1}}{p!}\|x_{k+1} - \tilde{v}_k\|^{p-1} \leq \hat{\sigma} + \sigma_u = \sigma, \\
\lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \geq \tfrac{\sigma_l p!}{2\ell} \quad &\Longrightarrow \quad \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \geq \theta.
\end{aligned}$$

Using the same argument with Bubeck et al. [2019, Lemma 3.1] instead of Jiang et al. [2019, Proposition 3.2] and an exact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (5.36b), we obtain the same result with $\theta = \frac{(p-1)!}{2\ell}$. Putting these pieces together yields the desired conclusion. □

In view of Proposition 5.4.12, the iteration complexity derived for Algorithm 14 hold for Algorithm 16. We summarize the results in the following theorem.

Theorem 5.4.13 *For every integer $k \geq 1$, the objective function gap satisfies*

$$\Phi(x_k) - \Phi(x^\star) = O(k^{-\frac{3p+1}{2}}),$$

and the squared gradient norm satisfies

$$\inf_{1\leq i\leq k} \|\nabla\Phi(x_i)\|^2 = O(k^{-3p}).$$

Algorithm 17 Optimal p-th order Tensor Algorithm II

STEP 0: Let $x_0, v_0 \in \mathbb{R}^d$, $\hat{\sigma} \in (0,1)$ and $0 < \sigma_l < \sigma_u < 1$ such that $\sigma_l(1+\hat{\sigma})^{p-1} < \sigma_u(1-\hat{\sigma})^{p-1}$ and $\sigma = \hat{\sigma} + \sigma_u < 1$ be given, and set $\gamma_0 = 1$ and $k = 0$.
STEP 1: If $0 = \nabla\Phi(x_k)$, then **stop**.
STEP 2: Otherwise, compute a positive scalar λ_{k+1} with a $\hat{\sigma}$-inexact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (5.36a) satisfying that

$$\tfrac{\sigma_l p!}{2\ell} \leq \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \leq \tfrac{\sigma_u p!}{2\ell},$$

or an exact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (5.36b) satisfying that

$$\tfrac{(p-1)!}{2\ell} \leq \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \leq \tfrac{p!}{\ell(p+1)},$$

where $\tilde{v}_k = (1-\alpha_{k+1})x_k + \alpha_{k+1}v_k$ and $(\alpha_{k+1})^2 = \lambda_{k+1}(1-\alpha_{k+1})\gamma_k$.
STEP 3: Compute $\gamma_{k+1} = (1-\alpha_{k+1})\gamma_k$ and $v_{k+1} = v_k - \frac{\alpha_{k+1}\nabla\Phi(x_{k+1})}{\gamma_{k+1}}$.
STEP 4: Set $k \leftarrow k+1$, and go to **STEP 1**.

Remark 5.4.14 *Theorem 5.4.13 has been derived in Monteiro and Svaiter [2013, Theorem 6.4] for the special case of $p = 2$, and a similar result for Nesterov's accelerated gradient descent (the special case of $p = 1$) has also been derived in Shi et al. [2022]. For $p \geq 3$ in general, the first inequality on the objective function gap has been derived independently in Gasnikov et al. [2019a, Theorem 1], Jiang et al. [2019, Theorem 3.5] and Bubeck et al. [2019, Theorem 1.1], while the second inequality on the squared gradient norm is new.*

Second algorithm. We present the second optimal p-th order tensor algorithm in Algorithm 17 which is Algorithm 15 with specific choice of θ. The proof is omitted since it is the same as the aforementioned analysis for Algorithm 16.

Proposition 5.4.15 *Algorithm 17 is Algorithm 15 with $\theta = \frac{\sigma_l p!}{2\ell}$ or $\theta = \frac{(p-1)!}{2\ell}$.*

Theorem 5.4.16 *For every integer $k \geq 1$, the objective gap satisfies*

$$\Phi(x_k) - \Phi(x^\star) = O(k^{-\frac{3p+1}{2}}),$$

and the squared gradient norm satisfies

$$\inf_{1\leq i\leq k} \|\nabla\Phi(x_i)\|^2 = O(k^{-3p}).$$

Remark 5.4.17 *The approximate tensor subroutine in Algorithm 16 and 17 can be efficiently implemented usinga novel bisection search scheme. We refer the interested readers to Jiang et al. [2019] and Bubeck et al. [2019] for the details.*

5.5 Conclusion

We have presented a closed-loop control system for modeling optimal tensor algorithms for smooth convex optimization and provided continuous-time and discrete-time Lyapunov functions for analyzing the convergence properties of this system and its discretization. Our framework provides a systematic way to derive discrete-time p-th order optimal tensor algorithms, for $p \geq 2$, and simplify existing analyses via the use of a Lyapunov function. A key ingredient in our framework is the algebraic equation, which is not present in the setting of $p = 1$, but is essential for deriving optimal acceleration methods for $p \geq 2$. Our framework allows us to infer that a certain class of p-th order tensor algorithms minimize the squared norm of the gradient at a fast rate of $O(k^{-3p})$ for smooth convex functions.

It is worth noting that one could also consider closed-loop feedback control of the velocity. This is called nonlinear damping in the PDE literature; see Attouch et al. [2022b] for recent progress in this direction. There are also several other avenues for future research. In particular, it is of interest to bring our perspective into register with the Lagrangian and Hamiltonian frameworks that have proved productive in recent work [Wibisono et al., 2016, Diakonikolas and Jordan, 2021, Muehlebach and Jordan, 2021, França et al., 2021] and the control-theoretic viewpoint of Lessard et al. [2016] and Hu and Lessard [2017]. We would hope for this study to provide additional insight into the geometric or dynamical role played by the algebraic equation for modeling the continuous-time dynamics. Moreover, we wish to study possible extensions of our framework to nonsmooth optimization by using differential inclusions Vassilis et al. [2018] and monotone inclusions. The idea is to consider the setting in which $0 \in T(x)$ where T is a maximally monotone operator in a Hilbert space [Alvarez and Attouch, 2001, Attouch and Svaiter, 2011, Maingé, 2013, Attouch et al., 2013b, Abbas et al., 2014, Attouch et al., 2016a, Bot and Csetnek, 2016, Attouch and Cabot, 2018, 2020, Attouch and László, 2020b,a]. Finally, given that we know that direct discretization of our closed-loop control system cannot recover Nesterov's optimal high-order tensor algorithms [Nesterov, 2018, Section 4.3], it is of interest to investigate the continuous-time limit of Nesterov's algorithms and see whether the algebraic equation plays a role in their analysis.

Chapter 6

A Closed-Loop Control Approach to High-Order Inclusion

We propose and analyze a new dynamical system with *a closed-loop control law* in a Hilbert space $\mathcal{H}$, aiming to shed light on the acceleration phenomenon for *monotone inclusion* problems, which unifies a broad class of optimization, saddle point and variational inequality (VI) problems under a single framework. Given an operator $A : \mathcal{H} \rightrightarrows \mathcal{H}$ that is maximal monotone, we propose a closed-loop control system that is governed by the operator $I - (I + \lambda(t)A)^{-1}$, where a feedback law $\lambda(\cdot)$ is tuned by the resolution of the algebraic equation $\lambda(t)\|(I + \lambda(t)A)^{-1}x(t) - x(t)\|^{p-1} = \theta$ for some $\theta > 0$. Our first contribution is to prove the existence and uniqueness of a global solution via the Cauchy-Lipschitz theorem. We present a simple Lyapunov function for establishing the weak convergence of trajectories via the Opial lemma and strong convergence results under additional conditions. We then prove a global ergodic convergence rate of $O(t^{-(p+1)/2})$ in terms of a gap function and a global pointwise convergence rate of $O(t^{-p/2})$ in terms of a residue function. Local linear convergence is established in terms of a distance function under an error bound condition. Further, we provide an algorithmic framework based on the implicit discretization of our system in a Euclidean setting, generalizing the large-step HPE framework [Monteiro and Svaiter, 2012]. Even though the discrete-time analysis is a simplification and generalization of existing analyses for a bounded domain, it is largely motivated by the aforementioned continuous-time analysis, illustrating the fundamental role that the closed-loop control plays in acceleration in monotone inclusion. A highlight of our analysis is a new result concerning p^{th}-order tensor algorithms for monotone inclusion problems, complementing the recent analysis for saddle point and VI problems [Bullins and Lai, 2022].

6.1 Introduction

Monotone inclusion refers to the problem of finding a root of a point-to-set maximal monotone operator $A : \mathcal{H} \rightrightarrows \mathcal{H}$ (see the definition in Rockafellar [1970]), where $\mathcal{H}$ is a real Hilbert

space. Formally, we have

$$\text{Find } x \in \mathcal{H} \text{ such that } 0 \in Ax. \tag{6.1}$$

Monotone inclusion is a fundamental problem in applied mathematics, unifying a broad class of optimization, saddle point and variational inequality problems in a single framework. In particular, the minimization of a convex function f consists in finding $x \in \mathcal{H}$ such that $0 \in \partial\Phi(x)$, where $\partial\Phi(\cdot)$—the subdifferential of Φ—is known to be maximal monotone if Φ is proper, lower semi-continuous and convex. As a further example, letting $A = F + \partial \mathbf{1}_{\mathcal{X}}$ where $F : \mathcal{H} \mapsto \mathcal{H}$ is continuous and monotone and $\mathbf{1}_{\mathcal{X}}$ is an indicator function of a closed and convex set $\mathcal{X} \subseteq \mathcal{H}$, the monotone inclusion problem becomes

$$\text{Finding } x \in \mathcal{X} \text{ such that } \langle F(x), y - x\rangle \geq 0 \text{ for all } y \in \mathcal{X}.$$

This is known as the variational inequality (VI) problem [Facchinei and Pang, 2007], and it also covers many classical problems as special cases [Karamardian, 1972, Kelley, 1995]. Over several decades, the monotone inclusion problem has found applications in a wide set of fields, including partial differential equations [Polyanin and Zaitsev, 2003], game theory [Osborne, 2004], signal/image processing [Bose and Meyer, 2003] and location theory [Farahani and Hekmatfar, 2009]; see also Facchinei and Pang [2007, Section 1.4] for additional applications. Recently, the model has begun to see applications in machine learning as an abstraction of saddle point problems, with examples including generative adversarial networks (GANs) [Goodfellow et al., 2014], online learning in games [Cesa-Bianchi and Lugosi, 2006], adversarial learning [Sinha et al., 2018] and distributed computing [Shamma, 2008]. These applications have made significant demands with respect to computational feasibility, and the design of efficient algorithms for solving monotone inclusions has moved to the fore in the past decade [Eckstein and Svaiter, 2009, Briceno-Arias and Combettes, 2011, Combettes, 2013, Combettes and Eckstein, 2018, Davis, 2015, Briceno-Arias and Davis, 2018].

A simple and basic tool for solving monotone inclusion problems is the celebrated proximal point algorithm (PPA) [Martinet, 1970, 1972, Rockafellar, 1976]. The idea is to reformulate Eq. (6.1) as a fixed-point problem given by

$$\text{Find } x \in \mathcal{H} \text{ such that } x - (I + \lambda A)^{-1}x = 0, \tag{6.2}$$

where $\lambda > 0$ is a parameter and $(I + \lambda A)^{-1}$ is the resolvent of index λ of A. Letting $x_0 \in \mathcal{H}$ be an initial point, the PPA scheme is implemented by

$$x_{k+1} = (I + \lambda A)^{-1}x_k, \quad \text{for all } k \geq 0.$$

In the special case of convex optimization, where $A = \partial f$, the convergence rate of PPA is $O(1/k)$ in terms of objective function gap [Güler, 1991]. It has been accelerated to $O(1/k^2)$ by Güler [1992]. However, this acceleration can not be extended to monotone inclusion problems in full generality, although extensions have been found under certain conditions (e.g., cocoercivity) [Alvarez and Attouch, 2001, Attouch and Peypouquet, 2019, Attouch and Cabot, 2020]. In the context of monotone VIs, the ergodic convergence rate is $O(1/k)$

in terms of a gap function and the pointwise convergence rate is $O(1/\sqrt{k})$ in terms of a residue function [Facchinei and Pang, 2007]. The former rate has matched the lower bound for first-order methods [Diakonikolas, 2020] while the latter rate can be improved using new acceleration techniques [Kim, 2021]. This line of work focuses, however, on first-order algorithms and does not regard acceleration as a general phenomenon to be realized via appeal to high-order smoothness structure of an operator. As noted in a seminal work [Monteiro and Svaiter, 2012], there remains a gap in our understanding of accelerated p^{th}-order tensor algorithms for monotone inclusion problems, for the case of $p \geq 2$, where the algorithmic design and convergence analysis is much more delicate.

In this paper, we avail ourselves of a continuous-time viewpoint for formulating acceleration in monotone inclusion, making use of a closed-loop control mechanism. We build on a two-decade trend that exploits the interplay between continuous-time and discrete-time perspectives on dynamical systems for monotone inclusion problems [Alvarez and Pérez C, 1998, Attouch and Redont, 2001, Alvarez et al., 2002, Attouch and Svaiter, 2011, Attouch et al., 2012, Maingé, 2013, Attouch et al., 2013b, Abbas et al., 2014, Attouch et al., 2016a, Attouch and László, 2020b, 2021]. As in these papers, our work makes use of Lyapunov functions to transfer asymptotic behavior and rates of convergence between continuous time and discrete time.

Our point of departure is the following continuous-time problem that incorporates a time-varying function $\lambda(\cdot)$ in place of λ in the fixed-point formulation of monotone inclusion in Eq. (6.2):

$$\dot{x}(t) + x(t) - (I + \lambda(t)A)^{-1}x(t) = 0. \tag{6.3}$$

The time evolution of $\lambda(\cdot)$ is specified by a closed-loop control law:

$$\lambda(t)\|\dot{x}(t)\|^{p-1} = \theta, \tag{6.4}$$

where $\theta > 0$ and the order $p \in \{1, 2, \ldots\}$ are parameters. We assume that $x(0) \in \{x \in \mathcal{H} \mid 0 \notin Ax\}$. This is not restrictive since $0 \in Ax(0)$ implies that the monotone inclusion problem has been solved. Throughout the paper, we assume that A *is maximal monotone and* $A^{-1}(0) = \{x \in \mathcal{H} \mid 0 \in Ax\}$ *is a nonempty set.* As we shall see, our main results on the existence and uniqueness of global solutions and the convergence properties of trajectories are valid under this general assumption. Finally, we remark that our control law in Eq. (6.4) is a natural generalization of a similar equation in Attouch et al. [2016a] that models the proximal Newton algorithm specialized to convex optimization.

Contributions. We first study the closed-loop control system in Eq. (6.3) and (6.4) and prove the existence and uniqueness of a global solution via the Cauchy-Lipschitz theorem (see Theorem 6.2.6). For $p = 1$, we have $\lambda(t) = \theta$ and our system becomes the continuous-time PPA dynamics, indicating that our system extends PPA from first-order monotone inclusion to high-order monotone inclusion. Further, we provide a Lyapunov function that allows us to establish weak convergence of trajectories via the Opial lemma (see Theorem 6.3.1) and yield strong convergence results under additional conditions (see Theorem 6.3.5). We

obtain an ergodic convergence rate of $O(t^{-(p+1)/2})$ in terms of a gap function and a pointwise convergence rate of $O(t^{-p/2})$ in terms of a residue function (see Theorem 6.3.7). Local linear convergence guarantee is established under an error-bound condition (see Theorem 6.3.9). Moreover, we provide an algorithmic framework based on the implicit discretization of our closed-looped control system and remark that it generalizes the large-step HPE framework of Monteiro and Svaiter [2012]. Our iteration complexity analysis, which is largely motivated by our continuous-time analysis, can be viewed as a simplification and generalization of the analysis in Monteiro and Svaiter [2012] for bounded domains (see Theorem 6.4.1). Finally, we combine our algorithmic framework with an approximate tensor subroutine, yielding a suite of accelerated p^{th}-order tensor algorithms for monotone inclusion problems with $A = F + H$, where F has Lipschitz $(p-1)^{\text{th}}$-order derivative and H is simple and maximal monotone. A highlight of our analysis is a set of new theoretical results concerning the convergence rate of p^{th}-order tensor algorithms for monotone inclusion problems, complementing the previous analysis in Bullins and Lai [2022].

Notation. We use bold lower-case letters such as x to denote vectors, and upper-case letters such as X to denote tensors. We let $\mathcal{H}$ be a real Hilbert space that is endowed with the scalar product $\langle \cdot, \cdot \rangle$. For a vector $x \in \mathcal{H}$, we let $\|x\|$ denote its norm induced by $\langle \cdot, \cdot \rangle$ and let $\mathbb{B}_\delta(x) = \{x' \in \mathcal{H} \mid \|x' - x\| \le \delta\}$ denote its δ-neighborhood. For the operator $A : \mathcal{H} \rightrightarrows \mathcal{H}$, we let $\mathrm{dom}(A) = \{x \in \mathcal{H} : Ax \neq \emptyset\}$. If $\mathcal{H} = \mathbb{R}^d$ is a real Euclidean space, $\|x\|$ refers to the ℓ_2-norm of x. For a tensor $X \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_p}$, we define

$$X[z^1, \cdots, z^p] = \sum_{1 \le i_j \le d_j, 1 \le j \le p} \left[X_{i_1, \cdots, i_p}\right] z^1_{i_1} \cdots z^p_{i_p},$$

and denote by $\|X\|_{\text{op}} = \max_{\|z^i\|=1, 1 \le j \le p} X[z^1, \cdots, z^p]$ its operator norm induced by $\|\cdot\|$. Fix $p \ge 1$, we let $\mathcal{G}^p_L(\mathbb{R}^d)$ be a class of maximal monotone single-valued operators $F : \mathbb{R}^d \to \mathbb{R}^d$ where the $(p-1)^{\text{th}}$-order Jacobian are L-Lipschitz. In other words, $F \in \mathcal{G}^p_L(\mathbb{R}^d)$ if F is maximal monotone and $\|D^{(p-1)}F(x') - D^{(p-1)}F(x)\| \le L\|x' - x\|$ for all $x, x' \in \mathbb{R}^d$ where $D^{(p-1)}F(x)$ is the $(p-1)^{\text{th}}$-order Jacobian of F at $x \in \mathbb{R}^d$ and $D^{(0)}F = F$ for all $F \in \mathcal{G}^1_L(\mathbb{R}^d)$. To be more specific, for $\{z_1, z_2, \ldots, z_p\} \subseteq \mathbb{R}^d$, we have

$$D^{(p-1)}F(x)[z^1, \cdots, z^p] = \sum_{1 \le i_1, \ldots, i_p \le d} \left[\frac{\partial F_{i_1}}{\partial x_{i_2} \cdots \partial x_{i_p}}(x)\right] z^1_{i_1} \cdots z^p_{i_p}.$$

Given an iteration count $k \ge 1$, the notation $a = O(b(k))$ stands for $a \le C \cdot b(k)$ where the constant $C > 0$ is independent of k.

6.2 The Closed-Loop Control System

We study the closed-loop control system in Eq. (6.3) and Eq. (6.4). Indeed, we start by analyzing the algebraic equation $\lambda(t)\|(I + \lambda(t)A)^{-1}x(t) - x(t)\|^{p-1} = \theta$ for $\theta \in (0, 1)$. Then,

we prove the existence and uniqueness of a local solution by appeal to the Cauchy-Lipschitz theorem and extend the local solution to a global solution using properties of the closed-loop control law $\lambda(\cdot)$. We conclude by discussing other systems in the literature that exemplify our general framework.

Algebraic equation. We study the algebraic equation,

$$\lambda(t)\|(I+\lambda(t)A)^{-1}x(t)-x(t)\|^{p-1}=\theta\in(0,1), \tag{6.5}$$

which links the feedback control law $\lambda(\cdot)$ and the solution trajectory $x(\cdot)$. To streamline the presentation, for the case of $p\geq 2$ we define a function $\varphi:[0,+\infty)\times\mathcal{H}\mapsto[0,+\infty)$, such that

$$\varphi(\lambda,x)=\lambda^{\frac{1}{p-1}}\|x-(I+\lambda A)^{-1}x\|,\quad \varphi(0,x)=0.$$

By the definition of φ, Eq. (6.5) is equivalent to $\varphi(\lambda(t),x(t))=\theta^{1/(p-1)}$. Our first lemma shows that the mapping $x\mapsto\varphi(\lambda,x)$ is Lipschitz continuous for fixed $\lambda>0$. We have:

Lemma 6.2.1 *For $p\geq 2$, we have $|\varphi(\lambda,x_1)-\varphi(\lambda,x_2)|\leq\lambda^{\frac{1}{p-1}}\|x_1-x_2\|$ for $\forall x_1,x_2\in\mathcal{H}$ and $\forall\lambda>0$.*

The next lemma presents a key property of the mapping $\lambda\mapsto\varphi(\lambda,x)$ for a fixed $x\in\mathcal{H}$. It can be interpreted as a generalization of Monteiro and Svaiter [2012, Lemma 4.3] and Attouch et al. [2016a, Lemma 1.3] from $p=2$ to $p\geq 2$.

Lemma 6.2.2 *For $p\geq 2$, we have*

$$\left(\tfrac{\lambda_2}{\lambda_1}\right)^{\frac{1}{p-1}}\varphi(\lambda_1,x)\leq\varphi(\lambda_2,x)\leq\left(\tfrac{\lambda_2}{\lambda_1}\right)^{\frac{p}{p-1}}\varphi(\lambda_1,x),$$

for all $x\in\mathcal{H}$ and $0<\lambda_1\leq\lambda_2$. In addition, $\varphi(\lambda,x)=0$ if and only if $0\in Ax$ for any fixed $\lambda>0$.

The following proposition provides a property of the mapping $\lambda\mapsto\varphi(\lambda,x)$, for any fixed $x\in\mathcal{H}$ satisfying $x\notin A^{-1}(0)=\{x'\in\mathcal{H}:0\in Ax'\}$. We have:

Proposition 6.2.3 *Suppose that $p\geq 2$ and $x\notin A^{-1}(0)$ is fixed, the mapping $\varphi(\cdot,x)$ is continuous and strictly increasing. Further, we have $\varphi(0,x)=0$ and $\varphi(\lambda,x)\to+\infty$ as $\lambda\to+\infty$.*

Proof. By definition of φ, we have $\varphi(0,x)=0$ for any fixed $x\notin A^{-1}(0)$. Since $x\notin A^{-1}(0)$, Lemma 6.2.2 guarantees that $\varphi(\lambda,x)>0$ for all $\lambda>0$ and $\varphi(\lambda_1,x)<\varphi(\lambda_2,x)$ for all $0<\lambda_1<\lambda_2$. That is to say, the mapping $\varphi(\cdot,x)$ is strictly increasing. In addition, we fix $\lambda_1>0$ and let $\lambda_2\to+\infty$ in Lemma 6.2.2, yielding that $\varphi(\lambda,x)\to+\infty$ as $\lambda\to+\infty$ for any fixed $x\notin A^{-1}(0)$. Finally, we prove the continuity of the mapping $\lambda\mapsto\varphi(\lambda,x)$. In

particular, Lemma 6.2.2 implies that $\varphi(\lambda, x) \leq \lambda^{p/(p-1)}\varphi(1, x)$ for any fixed $\lambda \in (0, 1]$. This together with the definition of φ implies that

$$0 \leq \limsup_{\lambda \to 0^+} \varphi(\lambda, x) \leq \lim_{\lambda \to 0^+} \lambda^{\frac{p}{p-1}}\varphi(1, x) = 0,$$

which implies the continuity of the mapping $\lambda \mapsto \varphi(\lambda, x)$ at $\lambda = 0$. Left continuity and right continuity of $\lambda \mapsto \varphi(\lambda, x)$ at $\lambda > 0$ follow from the first and the second inequality in Lemma 6.2.2. □

In view of Proposition 6.2.3, for any fixed $x \notin A^{-1}(0)$, there exists a unique $\lambda > 0$ so that $\varphi(\lambda, x) = \theta^{1/(p-1)}$ for some $\theta \in (0, 1)$. We accordingly define $\Omega \subseteq \mathcal{H}$ and the mapping $\Lambda_\theta : \Omega \mapsto (0, \infty)$ as follows:

$$\Omega = \mathcal{H} \setminus A^{-1}(0) \doteq \{x \in \mathcal{H} : 0 \notin Ax\}, \qquad \Lambda_\theta(x) = (\varphi(\cdot, x))^{-1}(\theta^{1/(p-1)}). \tag{6.6}$$

Since A is maximal monotone, we have that $A^{-1}(0)$ is closed and thus Ω is open. Note that this simple fact is crucial to the subsequent analysis of the existence and uniqueness of a local solution.

Existence and uniqueness of a local solution. We prove the existence and uniqueness of a local solution of the closed-loop control system in Eq. (6.3) and Eq. (6.4) by appeal to the Cauchy-Lipschitz theorem. The system considered in this paper can be written in the following form:

$$\begin{cases} \dot{x}(t) + x(t) - (I + \lambda(t)A)^{-1}x(t) = 0, \\ \lambda(t)\|(I + \lambda(t)A)^{-1}x(t) - x(t)\|^{p-1} = \theta, \\ x(0) = x_0 \in \Omega. \end{cases}$$

Using the mapping $\Lambda_\theta : \Omega \mapsto (0, \infty)$ (see Eq. (6.6)), this system can be expressed as an autonomous system. Indeed, we have

$$\lambda(t) = \Lambda_\theta(x(t)) \Longleftrightarrow \lambda(t)\|(I + \lambda(t)A)^{-1}x(t) - x(t)\|^{p-1} = \theta.$$

Putting these pieces together, we arrive at an autonomous system in the compact form of

$$\dot{x}(t) = F(x(t)), \quad x(0) = x_0 \in \Omega, \tag{6.7}$$

where the vector field $F : \Omega \mapsto \mathcal{H}$ is given by

$$F(x) = (I + \Lambda_\theta(x)A)^{-1}x - x. \tag{6.8}$$

Our theorem on the existence and uniqueness of a local solution is summarized as follows.

Theorem 6.2.4 *There exists $t_0 > 0$ such that the autonomous system in Eq. (6.7) and Eq. (6.8) has a unique solution $x : [0, t_0] \mapsto \mathcal{H}$. Equivalently, the closed-loop control system in Eq. (6.3) and Eq. (6.4) has a unique solution, $(x, \lambda) : [0, t_0] \mapsto \mathcal{H} \times (0, +\infty)$. In addition, $x(\cdot)$ is continuously differentiable and $\lambda(\cdot)$ is locally Lipschitz continuous.*

A standard approach for proving the existence and uniqueness of a local solution is via appeal to the Cauchy-Lipschitz theorem [Coddington and Levinson, 1955, Theorem I.3.1]. This theorem requires that the vector field $F(\cdot)$ is Lipschitz continuous, which is not immediate in our case due to the appearance of $(I + \Lambda_\theta(x)A)^{-1}$. In order to avail ourselves directly of the Cauchy-Lipschitz theorem, the first step is to study the properties of the function $\Lambda_\theta(x)$. We have the following lemma.

Lemma 6.2.5 *Suppose that $p \geq 2$ and the function $\Gamma_\theta : \mathcal{H} \mapsto (0, +\infty)$ is given by*

$$\Gamma_\theta(x) = \left(\inf\{\alpha > 0 : \left\|x - (I + \alpha^{-1}A)^{-1}x\right\| \leq \alpha^{\frac{1}{p-1}}\theta^{\frac{1}{p-1}}\}\right)^{\frac{1}{p-1}}.$$

Then, we have

$$\Gamma_\theta(x) = \begin{cases} \left(\frac{1}{\Lambda_\theta(x)}\right)^{\frac{1}{p-1}}, & \text{if } x \in \Omega, \\ 0, & \text{otherwise,} \end{cases}$$

and Γ_θ itself is Lipschitz continuous with a constant $\theta^{-1/(p-1)} > 0$.

Proof of Theorem 6.2.4: For simplicity, we define $A_\theta = I - (I + \theta A)^{-1}$. In what follows, we first prove that $F : \Omega \mapsto \mathcal{H}$ defined in Eq. (6.8) is locally Lipschitz continuous case by case.

Case of $p = 1$: The algebraic equation in Eq. (6.4) implies that $\lambda(\cdot)$ is a constant function such that $\lambda(t) \equiv \theta$ for all $t \geq 0$. Then, by the definition of F in Eq. (6.8), we have

$$F(x) = (I + \theta A)^{-1}x - x.$$

By the definition of A_λ, we have $\|F(x_1) - F(x_2)\| = \|A_\theta x_1 - A_\theta x_2\|$. It is straightforward to derive that A_θ is 1-Lipschitz continuous (see the proof of Lemma 6.2.1). Putting these pieces together yields the desired result.

Case of $p \geq 2$: Taking $x_0 \in \Omega$ and $0 < \delta < (\frac{\theta}{\Lambda_\theta(x_0)})^{1/(p-1)}$, we have $B_\delta(x_0, \delta) \subseteq \Omega$ since Ω is open. For any $x \in B_\delta(x_0, \delta)$, Lemma 6.2.5 implies

$$\left|\left(\frac{1}{\Lambda_\theta(x)}\right)^{\frac{1}{p-1}} - \left(\frac{1}{\Lambda_\theta(x_0)}\right)^{\frac{1}{p-1}}\right| = |\Gamma_\theta(x) - \Gamma_\theta(x_0)| \leq \theta^{-\frac{1}{p-1}}\|x - x_0\| \leq \delta\theta^{-\frac{1}{p-1}}.$$

In view of the choice of $\delta > 0$ and the definition of λ_0, we have

$$0 < \left(\frac{1}{\Lambda_\theta(x_0)}\right)^{\frac{1}{p-1}} - \left(\frac{\delta^{p-1}}{\theta}\right)^{\frac{1}{p-1}} \leq \left(\frac{1}{\Lambda_\theta(x)}\right)^{\frac{1}{p-1}} = \Gamma_\theta(x) \leq \left(\frac{1}{\Lambda_\theta(x_0)}\right)^{\frac{1}{p-1}} + \left(\frac{\delta^{p-1}}{\theta}\right)^{\frac{1}{p-1}}. \tag{6.9}$$

Taking $x_1, x_2 \in B_\delta(x_0, \delta) \subseteq \Omega$, we let $\lambda_1 = \Lambda_\theta(x_1)$ and $\lambda_2 = \Lambda_\theta(x_2)$. Then, we have

$$\|F(x_1) - F(x_2)\| = \|A_{\lambda_1}x_1 - A_{\lambda_2}x_2\| \leq \|A_{\lambda_1}x_1 - A_{\lambda_1}x_2\| + \|A_{\lambda_1}x_2 - A_{\lambda_2}x_2\|. \tag{6.10}$$

By the definition of λ_1, we obtain that $\lambda_1 > 0$ and A_{λ_1} is 1-Lipschitz continuous. This implies

$$\|A_{\lambda_1}x_1 - A_{\lambda_1}x_2\| \leq \|x_1 - x_2\|. \tag{6.11}$$

Further, we obtain from the proof of Attouch and Peypouquet [2019, Lemma A.4] that

$$\|A_{\lambda_1}x_2 - A_{\lambda_2}x_2\| \leq \left|1 - \tfrac{\lambda_1}{\lambda_2}\right| \|x_2 - (I + \lambda_2 A)^{-1}x_2)\|.$$

This together with the definition of λ_2 and $\Lambda_\theta(\cdot)$ yields

$$\|A_{\lambda_1}x_2 - A_{\lambda_2}x_2\| \leq \left|1 - \tfrac{\lambda_1}{\lambda_2}\right| \left(\tfrac{\theta}{\lambda_2}\right)^{\frac{1}{p-1}}. \tag{6.12}$$

Plugging Eq. (6.11) and Eq. (6.12) into Eq. (6.10) and using the definition of λ_1 and λ_2, we have

$$\|F(x_1) - F(x_2)\| \leq \|x_1 - x_2\| + \left|1 - \tfrac{\Lambda_\theta(x_1)}{\Lambda_\theta(x_2)}\right| \left(\tfrac{\theta}{\Lambda_\theta(x_2)}\right)^{\frac{1}{p-1}}. \tag{6.13}$$

Using $\Gamma_\theta(x) = (\frac{1}{\Lambda_\theta(x)})^{\frac{1}{p-1}}$ for all $x \in \Omega$ and the Lipschitz continuity of Γ_θ (cf. Lemma 6.2.5), we have

$$\begin{aligned}
\left|1 - \tfrac{\Lambda_\theta(x_1)}{\Lambda_\theta(x_2)}\right| \left(\tfrac{1}{\Lambda_\theta(x_2)}\right)^{\frac{1}{p-1}} &= \tfrac{\Gamma_\theta(x_2)}{(\Gamma_\theta(x_1))^{p-1}} \left|(\Gamma_\theta(x_1))^{p-1} - (\Gamma_\theta(x_2))^{p-1}\right| \\
&= \tfrac{\Gamma_\theta(x_2)}{(\Gamma_\theta(x_1))^{p-1}} |\Gamma_\theta(x_1) - \Gamma_\theta(x_2)| \left(\sum_{i=1}^{p-1} (\Gamma_\theta(x_1))^{p-1-i}(\Gamma_\theta(x_2))^{i-1}\right) \\
&\leq \tfrac{\Gamma_\theta(x_2)}{(\Gamma_\theta(x_1))^{p-1}} \left(\sum_{i=1}^{p-1} (\Gamma_\theta(x_1))^{p-1-i}(\Gamma_\theta(x_2))^{i-1}\right) \theta^{-\frac{1}{p-1}} \|x_1 - x_2\|.
\end{aligned}$$

Plugging this inequality into Eq. (6.13) yields

$$\|F(x_1) - F(x_2)\| \leq \left(1 + \tfrac{\Gamma_\theta(x_2)}{(\Gamma_\theta(x_1))^{p-1}} \left(\sum_{i=1}^{p-1} (\Gamma_\theta(x_1))^{p-1-i}(\Gamma_\theta(x_2))^{i-1}\right)\right) \|x_1 - x_2\|.$$

Since $x_1, x_2 \in B_\delta(x_0, \delta)$, Eq. (6.9) implies

$$0 < \left(\tfrac{1}{\Lambda_\theta(x_0)}\right)^{\frac{1}{p-1}} - \left(\tfrac{\delta^{p-1}}{\theta}\right)^{\frac{1}{p-1}} \leq \Gamma_\theta(x_i) \leq \left(\tfrac{1}{\Lambda_\theta(x_0)}\right)^{\frac{1}{p-1}} + \left(\tfrac{\delta^{p-1}}{\theta}\right)^{\frac{1}{p-1}}, \quad \text{for all } i = 1, 2.$$

Therefore, we conclude that

$$\|F(x_1) - F(x_2)\| \leq C\|x_1 - x_2\|,$$

where $C > 0$ is a constant that is independent of the choice of x_1 and x_2 but only depends on the value of δ, θ, p and $\Lambda_\theta(x_0)$. This proves the claim.

We are ready to prove main results. Indeed, by the Cauchy-Lipschitz theorem (local version), for any $x_0 \in \Omega$, there exists a unique local solution $x : [0, t_1] \mapsto \mathcal{H}$ of the autonomous system in Eq. (6.7) and Eq. (6.8) for some $t_1 > 0$. Thus, there exists a unique local solution, $(x, \lambda) : [0, t_1] \mapsto \mathcal{H} \times (0, +\infty)$, of the closed-loop control system in Eq. (6.3) and Eq. (6.4) with $\lambda(t) = \Lambda_\theta(x(t))$. By Cauchy-Lipschitz theorem, we have that $x(\cdot)$ is continuously differentiable and $x(t) \in \Omega$ for all $t \in [0, t_1]$. For the case of $p = 1$, Eq. (6.4) implies that $\lambda(\cdot)$ is a constant function and thus locally Lipschitz continuous. For the case of $p \geq 2$, Lemma 6.2.5 together with $x(t) \in \Omega$ for all $t \in [0, t_1]$ implies

$$\lambda(t) = \Lambda_\theta(x(t)) = \left(\frac{1}{\Gamma_\theta(x(t))}\right)^{\frac{1}{p-1}} \quad \text{for all } t \in [0, t_1].$$

Since $\Gamma_\theta(x)$ is Lipschitz continuous in x, we obtain that $\lambda(\cdot)$ is Lipschitz continuous on $[0, t_2]$ for some sufficiently small $t_2 > 0$. Then, by taking $t_0 = \min\{t_1, t_2\} > 0$, we achieve the desired results. This completes the proof.

Existence and uniqueness of a global solution. Our theorem on the existence and uniqueness of a global solution is summarized as follows.

Theorem 6.2.6 *The closed-loop control system in Eq. (6.3) and Eq. (6.4) has a unique global solution, $(x, \lambda) : [0, +\infty) \mapsto \mathcal{H} \times (0, +\infty)$. Moreover, $x(\cdot)$ is continuously differentiable and $\lambda(\cdot)$ is locally Lipschitz continuous. If $p \geq 2$, we have*

$$\|x(t) - (I + \lambda(t)A)^{-1}x(t)\| \geq \|x(0) - (I + \lambda(0)A)^{-1}x(0)\|e^{-t}, \quad \text{for all } t \geq 0.$$

Remark 6.2.7 *Theorem 6.2.6 demonstrates that $x(t) - (I + \lambda(t)A)^{-1}x(t) \neq 0$ for all $t \geq 0$. After some straightforward calculations, it is clear that the aforementioned argument is equivalent to the assertion that the orbit $x(\cdot)$ stays in Ω. In other words, if $x_0 \in \Omega$, our closed-loop control system in Eq. (6.3) and Eq. (6.4) is not stabilized in finite time, which helps clarify the asymptotic convergence behavior of many discrete-time algorithms to a solution of monotone inclusion problems (see Monteiro and Svaiter [2010, 2012] for examples).*

Remark 6.2.8 *The feedback law $\lambda(\cdot)$, which we will show satisfies $\lambda(t) \to +\infty$ as $t \to +\infty$, links to $\|\dot{x}(\cdot)\| = \|x(\cdot) - (I + \lambda(\cdot)A)^{-1}x(\cdot)\|$ via Eq. (6.4). Intuitively, if $\lambda(\cdot)$ changes dramatically, we can not globalize a local solution using classical arguments. In the Levenberg-Marquardt regularized systems, Attouch and Svaiter [2011] resolved this issue by assuming that $\lambda(\cdot)$ is absolutely continuous on any finite bounded interval and proving that $\lambda(t) \leq \lambda(0)e^{ct}$ holds true for some constant $c > 0$. However, $\lambda(\cdot)$ is not a given datum in our closed-loop control system but an emergent component of the evolution dynamics. As such, it is preferable to prove that $\lambda(t) \leq \lambda(0)e^{ct}$ hold true without imposing any condition, as done in the works [Attouch et al., 2013b, 2016a]. Recently, Lin and Jordan [2022b] have studied a closed-loop control system which characterized accelerated p^{th}-order tensor algorithms for*

convex optimization and established the global existence and uniqueness results under the condition used in Attouch and Svaiter [2011]. They also clarified why this condition is necessary and considered it an open problem to remove it. In the subsequent analysis, we prove that $|\dot{\lambda}(t)| \leq (p-1)\lambda(t)$ *holds for our system in Eq. (6.3) and Eq. (6.4) without imposing any condition, demonstrating that acceleration in monotone inclusion problems is intrinsically different from that in convex optimization.*

We provide two lemmas that characterize further properties of the feedback law $\lambda(\cdot)$.

Lemma 6.2.9 *Suppose that* $(x, \lambda) : [0, t_0] \mapsto \mathcal{H} \times (0, +\infty)$ *is a solution of the closed-loop control system in Eq. (6.3) and Eq. (6.4). Then, we have* $|\dot{\lambda}(t)| \leq (p-1)\lambda(t)$ *for almost all* $t \in [0, t_0]$.

Lemma 6.2.10 *Suppose that* $(x, \lambda) : [0, t_0] \mapsto \mathcal{H} \times (0, +\infty)$ *is a solution of the closed-loop control system in Eq. (6.3) and Eq. (6.4). Then, we have that* $\lambda(\cdot)$ *is nondecreasing.*

Proof of Theorem 6.2.6: We are ready to prove our main result on the existence and uniqueness of a global solution. In particular, for the case of $p = 1$, it is clear that $\lambda(t) = \theta$ is a constant function and the vector field $F : \Omega \mapsto \mathcal{H}$ is in fact global Lipschitz continuous (see the proof of Theorem 6.2.4). Thus, by the Cauchy-Lipschitz theorem (global version), we achieve the desired result.

For the case of $p \geq 2$, let us consider a maximal solution of the closed-loop control system in Eq. (6.3) and Eq. (6.4) as follows,

$$(x, \lambda) : [0, T_{\max}) \mapsto \Omega \times (0, +\infty).$$

Using the existence and uniqueness of a local solution (see Theorem 6.2.4) and a classical argument, we obtain that the aforementioned maximal solution must exist. Further, by using Lemma 6.2.9 and 6.2.10, we obtain that $\lambda(\cdot)$ is nondecreasing with $0 \leq \dot{\lambda}(t) \leq (p-1)\lambda(t)$ for almost all $t \in [0, T_{\max})$.

It remains to show that the maximal solution is a global solution; that is, $T_{\max} = +\infty$. Indeed, the property of λ guarantees

$$0 < \lambda(0) \leq \lambda(t) \leq \lambda(0)e^{(p-1)t}. \tag{6.14}$$

If $T_{\max} < +\infty$, this inequality implies that $\lambda(t) \leq \lambda(0)e^{(p-1)T_{\max}}$ for all $t \in [0, T_{\max}]$. This together with the fact that $\lambda(\cdot)$ is nondecreasing on $[0, T_{\max})$ implies that $\bar{\lambda} = \lim_{t \to T_{\max}} \lambda(t)$ exists and is finite and strictly positive. Using Eq. (6.3) and Eq. (6.4), we have

$$\|\dot{x}(t)\| = \|(I + \lambda(t)A)^{-1}x(t) - x(t)\| = \left(\frac{\theta}{\lambda(t)}\right)^{\frac{1}{p-1}}. \tag{6.15}$$

Combining Eq. (6.14) and Eq. (6.15) implies that $\|\dot{x}(\cdot)\|$ is bounded on $[0, T_{\max})$. Thus, $x(\cdot)$ is Lipschitz continuous on $[0, T_{\max})$ and this implies that $\bar{x} = \lim_{t \to T_{\max}} x(t)$ exists. We claim

that $\bar{x} \in \Omega$. Indeed, the function $g(\lambda, x) = \|(I+\lambda A)^{-1}x - x\|$ is continuous in (λ, x). Since $\lambda(\cdot)$ and $x(\cdot)$ are continuous on $[0, T_{\max}]$, we have $\|(I+\lambda(t)A)^{-1}x(t) - x(t)\| \to \|(I+\bar{\lambda}A)^{-1}\bar{x} - \bar{x}\|$ as $t \to T_{\max}$. Then, Eq. (6.15) implies

$$\|(I+\bar{\lambda}A)^{-1}\bar{x} - \bar{x}\| = \lim_{t\to T_{\max}} \|(I+\lambda(t)A)^{-1}x(t) - x(t)\| = \lim_{t\to T_{\max}} \left(\frac{\theta}{\lambda(t)}\right)^{\frac{1}{p-1}} = \left(\frac{\theta}{\bar{\lambda}}\right)^{\frac{1}{p-1}} > 0.$$

By appeal to Theorem 6.2.4 with an initial point $\bar{x}$, we can then extend the solution to a strictly larger interval which contradicts the maximality of the aforementioned solution.

Using Eq. (6.15) again, we have

$$\|x(t) - (I+\lambda(t)A)^{-1}x(t)\| = \left(\frac{\lambda(0)}{\lambda(t)}\right)^{\frac{1}{p-1}} \|x(0) - (I+\lambda(0)A)^{-1}x(0)\|.$$

Further, it is clear that Eq. (6.14) holds true for all $t \in [0, +\infty)$. That is to say, we have $\frac{\lambda(0)}{\lambda(t)} \geq e^{-(p-1)t}$. Putting these pieces together yields

$$\|x(t) - (I+\lambda(t)A)^{-1}x(t)\| \geq e^{-t}\|x(0) - (I+\lambda(0)A)^{-1}x(0)\|,$$

which completes the proof.

Discussion. We compare the system in Eq. (6.3) and Eq. (6.4) to other systems for convex optimization and monotone inclusion. We also give an overview of the closed-loop control approach and the continuous-time interpretation of high-order tensor algorithms.

Existing systems for optimization and inclusion problems. In the context of optimization with a convex potential function $\Phi : \mathcal{H} \mapsto \mathbb{R}$, Polyak [1964] was the first to use inertial dynamics to accelerate gradient methods. However, the convergence rate of $O(1/t)$ he obtained is not better than the steepest descent method. A decisive step to obtain a faster convergence rate was taken by Su et al. [2016] who considered using *asymptotically vanishing damping* for modeling Nesterov's acceleration [Nesterov, 1983, Güler, 1992], triggering a productive line of research on the dynamical systems foundations of accelerated first-order algorithms [Attouch and Peypouquet, 2016, Attouch and Cabot, 2017, Attouch et al., 2018, Diakonikolas and Orecchia, 2019, Apidopoulos et al., 2020, Muehlebach and Jordan, 2021]. Another important ingredient for obtaining acceleration is so-called Hessian-driven damping [Alvarez et al., 2002, Attouch et al., 2016b, Lin and Jordan, 2022b, Attouch et al., 2022a,d] which originated from a variational characterization of general regularization optimization algorithms [Alvarez and Pérez C, 1998]. This involved the study of Newton and Levenberg-Marquardt regularized systems as follows:

$$\textbf{(Newton)} \qquad \ddot{x}(t) + \nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0,$$
$$\textbf{(Levenberg-Marquardt)} \qquad \lambda(t)\dot{x}(t) + \nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0.$$

These systems are well defined and admit robust asymptotic behavior [Attouch and Redont, 2001, Attouch and Svaiter, 2011]. Based on this work, Alvarez et al. [2002] distinguished Hessian-driven damping from continuous-time Newton dynamics and Attouch et al. [2016b] interpreted Nesterov's acceleration in the forward-backward algorithms by combining Hessian-driven damping with asymptotically vanishing damping. The resulting dynamics takes the following general form:

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \beta(t)\nabla^2\Phi(x(t))\dot{x}(t) + b(t)\nabla\Phi(x(t)) = 0. \tag{6.16}$$

Further work in this vein appeared in Shi et al. [2022], where Nesterov's acceleration was interpreted via multiscale limits that distinguish it from heavy ball method, and Attouch et al. [2022a], where time scaling was introduced with Hessian-driven damping. Unfortunately, none of the above approaches are suitable for deriving optimal accelerated versions of high-order tensor algorithms in convex smooth optimization [Monteiro and Svaiter, 2013, Gasnikov et al., 2019b]. Recently, Lin and Jordan [2022b] provided an initial foray into analyzing the continuous-time dynamics of high-order tensor algorithms using the system in Eq. (6.16) in which the tuning of $(\alpha(\cdot), \beta(\cdot), b(\cdot))$ is done in a closed loop by resolution of the algebraic equation. Their approach gives a systematic way to derive discrete-time optimal high-order tensor algorithms, further simplifying and generalizing the existing analysis in Monteiro and Svaiter [2013] via appeal to the construction of a unified discrete-time Lyapunov function.

The extension of the continuous-time dynamics and Lyapunov analysis from convex optimization to monotone inclusion problems has been pursued during the last two decades [Alvarez and Attouch, 2001, Attouch and Maingé, 2011, Attouch and Svaiter, 2011, Maingé, 2013, Attouch et al., 2013b, 2016a, Abbas et al., 2014, Bot and Csetnek, 2016, Attouch and Peypouquet, 2019, Attouch and Cabot, 2018, 2020, Attouch and László, 2020b, 2021]. In particular, Attouch and Svaiter [2011] considered a generalization of Levenberg-Marquardt regularized systems for monotone inclusion problems as follows,

$$\begin{cases} v(t) \in Ax(t), \\ \lambda(t)\dot{x}(t) + \dot{v}(t) + v(t) = 0. \end{cases}$$

This system yields weak convergence to $A^{-1}(0)$ under a certain condition on $\lambda(\cdot)$. Subsequent work has obtained convergence rates for various first-order algorithms obtained by the implicit discretization of this system or its variants [Attouch et al., 2013b, Abbas et al., 2014, Attouch et al., 2016a].

Under the assumption that A is point-to-point and *cocoercive*, inertial systems taking the following form have been considered in the literature [Alvarez and Attouch, 2001, Attouch and Maingé, 2011, Maingé, 2013, Bot and Csetnek, 2016]:

$$\ddot{x}(t) + \alpha\dot{x}(t) + A(x(t)) = 0.$$

It is worth mentioning that cocoercivity is necessary for guaranteeing weak asymptotic stabilization, and a fast convergence rate. For $\lambda > 0$, the operator $A_\lambda = \frac{1}{\lambda}(I - (I + \lambda A)^{-1})$ is λ-cocoercive and $A^{-1}(0) = A_\lambda^{-1}(0)$. This motivates us to study the following inertial system:

$$\ddot{x}(t) + \alpha\dot{x}(t) + A_\lambda(x(t)) = 0.$$

In the quest for faster convergence, Attouch and Peypouquet [2019] combined this system with asymptotically vanishing damping and a time-dependent regularizing parameter $\lambda(\cdot)$:

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + A_{\lambda(t)}(x(t)) = 0.$$

The discretization of these dynamics gives the relaxed inertial proximal algorithm [Attouch and Cabot, 2018, Attouch et al., 2019c, Attouch and Peypouquet, 2019, Attouch and Cabot, 2020]. Recently, Attouch and László [2020b, 2021] have proposed to study Newton-like inertial dynamics which generalizes the system in Eq. (6.16) from convex optimization to monotone inclusion problems. The resulting dynamics takes the following general form:

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \beta(t)\tfrac{d}{dt}(A_{\lambda(t)}x(t)) + b(t)A_{\lambda(t)}x(t) = 0. \tag{6.17}$$

The introduction of Newton-like correction term $\frac{d}{dt}(A_{\lambda(t)}x(t))$—the generalization of the Hessian-driven damping – provides a well-posed system for which we can derive the weak convergence of trajectories to $A^{-1}(0)$. The convergence rates have also been obtained in both continuous-time and discrete-time cases using the metric $\|A_{\lambda(t)}x(t)\|$(see Attouch and Peypouquet [2019] and Attouch and László [2020b, 2021]). In contrast, we investigate different dynamical systems and derive different rates in terms of two more intuitive metrics—a gap function and a residue function.

All of the aforementioned dynamical systems study first-order algorithms for monotone inclusion problems and do not aim to capture the acceleration that may be obtainable from high-order smoothness structures. The only exception that we are aware of is Attouch et al. [2016a] who proposed a proximal Newton method for solving monotone inclusion problems but conducted the convergence rate estimation when an operator is the subdifferential of a convex function. Meanwhile, Monteiro and Svaiter [2012] and Bullins and Lai [2022] have demonstrated that high-order tensor algorithms can achieve faster convergence rate than first-order algorithms, but their derivations depend heavily on case-specific algebra. As such, there remains a gap in our understanding; in particular, we are missing a continuous-time perspective on acceleration in monotone inclusion.

Closed-loop control systems. Closed-loop control systems have been studied in the context of convex optimization [Lin and Jordan, 2022b, Attouch et al., 2022b] and monotone inclusion [Attouch et al., 2013b, 2016a]. Even though Attouch et al. [2013b, 2016a] closely resembles our work, some differences exist. In particular, their convergence analysis of Newton-type methods targets the solution of convex optimization rather than monotone inclusion problems. Our focus, on the other hand, is to link closed-loop control with acceleration in monotone inclusion, especially when $p > 2$. From a technical viewpoint, the

construction of the gap function and the convergence rate estimation that we provide do not appear in these earlier works.

Continuous-time perspective on high-order tensor algorithms. To the best of our knowledge, all the existing work on continuous-time interpretations of high-order tensor algorithms focus on convex optimization [Wibisono et al., 2016, Song et al., 2021, Lin and Jordan, 2022b]. In particular, Wibisono et al. [2016] studied the following inertial gradient system with asymptotically vanishing damping:

$$\ddot{x}(t) + \tfrac{p+2}{t}\dot{x}(t) + C(p+1)^2 t^{p-1}\nabla\Phi(x(t)) = 0,$$

which is an open-loop system without Hessian-driven damping. They derived a class of p^{th}-order tensor algorithms by implicit discretization and established a convergence rate of $O(k^{-(p+1)})$ in terms of the objective function gap. Song et al. [2021] proposed another form of open-loop dynamics (we consider the simplified form in a Euclidean setting):

$$\ddot{x}(t) + \left(\tfrac{2\dot{a}(t)}{a(t)} - \tfrac{\ddot{a}(t)}{\dot{a}(t)}\right)\dot{x}(t) + \left(\tfrac{(\dot{a}(t))^2}{a(t)}\right)\nabla\Phi(x(t)) = 0,$$

which is also open-loop and lacks Hessian-driven damping. Recently, Lin and Jordan [2022b] provided a control-theoretic perspective on optimal acceleration for high-order tensor algorithms. They considered the following closed-loop control system with Hessian-driven damping:

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \beta(t)\nabla^2\Phi(x(t))\dot{x}(t) + b(t)\nabla\Phi(x(t)) = 0,$$

where (α, β, b) are defined by

$$\alpha(t) = \tfrac{2\dot{a}(t)}{a(t)} - \tfrac{\ddot{a}(t)}{\dot{a}(t)}, \quad \beta(t) = \tfrac{(\dot{a}(t))^2}{a(t)}, \quad b(t) = \tfrac{\dot{a}(t)(\dot{a}(t)+\ddot{a}(t))}{a(t)},$$
$$a(t) = \tfrac{1}{4}(\textstyle\int_0^t \sqrt{\lambda(s)}ds)^2, \quad (\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta,$$

and recovered a class of optimal high-order tensor algorithms [Monteiro and Svaiter, 2013, Gasnikov et al., 2019b] from implicit discretization of the above system. Here the rate is $O(t^{-(3p+1)/2})$ in terms of the objective function gap.

There is comparatively little work on the development of high-order tensor algorithms for monotone inclusion problems; indeed, we are only aware of the high-order mirror-prox method [Bullins and Lai, 2022]. However, the derivation of this algorithm does not flow from a single underlying principle but again involves case-specific algebra. It has been an open challenge to extend earlier work on open-loop and closed-loop systems from convex optimization to monotone inclusion problems.

6.3 Convergence Properties of Trajectories

We give a Lyapunov function for analyzing the convergence properties of solution trajectories of our system in Eq. (6.3) and Eq. (6.4). In particular, we prove the weak convergence of

trajectories to equilibrium by appeal to the Opial lemma as well as strong convergence results under additional conditions. We also derive new global convergence rates by estimating the rate of decrease of Lyapunov function. Finally, we use another Lyapunov function to establish local linear convergence under an error bound condition.

Weak convergence. We present our results on the weak convergence of trajectories.

Theorem 6.3.1 *Suppose that $(x, \lambda) : [0, +\infty) \mapsto \mathcal{H} \times (0, +\infty)$ is a global solution of the closed-loop control system in Eq. (6.3) and Eq. (6.4). Then, there exists some $\bar{x} \in A^{-1}(0)$ such that the trajectory $x(t)$ weakly converges to $\bar{x}$ as $t \to +\infty$.*

Remark 6.3.2 *In the Hilbert-space setting (possibly infinite dimensional), the weak convergence of $x(\cdot)$ to some $\bar{x} \in A^{-1}(0)$ in Theorem 6.3.1 is the best we can expect without additional conditions. The same result was established for other systems for convex optimization, including an open-loop inertial system for first-order algorithms [Attouch et al., 2016b] and a closed-loop control system (which is the special instance of our system for $p = 2$), and for second-order algorithms [Attouch et al., 2016a].*

We define the following Lyapunov function for the system in Eq. (6.3) and Eq. (6.4):

$$\mathcal{E}(t) = \tfrac{1}{2}\|x(t) - z\|^2, \tag{6.18}$$

where $z \in \mathcal{H}$ is a point in the Hilbert space. Note that this function measures the distance between $x(t)$ and any fixed point $z \in \mathcal{H}$. It is simpler than that used for analyzing the convergence of Newton-like inertial dynamics for monotone inclusion problems and different from the ones developed for the systems with asymptotically vanishing damping. The closest Lyapunov function to ours is the one employed by Attouch et al. [2016a], which is defined as the distance between $x(t)$ and $z \in A^{-1}(0)$. We note that the seemingly minor modification to the form in Eq. (6.18) is key to deriving new results on the ergodic convergence of trajectories in terms of a gap function (see Theorem 6.3.7 and its proof).

The following proposition gives the Opial lemma in its continuous form [Opial, 1967]. It has become a basic analytical tool to study the weak convergence of trajectories of dynamical systems associated with discrete-time algorithms for convex optimization [Alvarez, 2000, Attouch et al., 2000] and monotone inclusion [Attouch and Redont, 2001, Attouch et al., 2016a].

Proposition 6.3.3 (Opial Lemma) *Suppose that $S \subseteq \mathcal{H}$ is a nonempty subset and $x : [0, +\infty) \mapsto \mathcal{H}$ is a mapping. Then, there exists some $\bar{x} \in S$ such that $x(t)$ weakly converges to $\bar{x}$ as $t \to +\infty$ if both of the following assumptions hold true:*

1. *For every $z \in S$, we have that $\lim_{t\to+\infty} \|x(t) - z\|$ exists.*
2. *Every weak sequential cluster point of $x(\cdot)$ belongs to S.*

To prove Theorem 6.3.1, we provide one technical lemma that characterizes a descent property of $\mathcal{E}(\cdot)$ which is crucial to our subsequent analysis in this paper.

Lemma 6.3.4 *Suppose that $(x,\lambda):[0,+\infty)\mapsto\mathcal{H}\times(0,+\infty)$ is a global solution of the closed-loop control system in Eq. (6.3) and Eq. (6.4) and let $z\in A^{-1}(0)$ in Eq. (6.18). Then, we have*

$$\tfrac{d\mathcal{E}(t)}{dt}\leq-\|x(t)-(I+\lambda(t)A)^{-1}x(t)\|^2.$$

Proof of Theorem 6.3.1: By Proposition 6.3.3, it suffices to prove that (i) $\lim_{t\to+\infty}\|x(t)-z\|$ exists for every $z\in A^{-1}(0)$ and (ii) every weak sequential cluster point of $x(\cdot)$ belongs to $A^{-1}(0)$.

By Lemma 6.3.4, we have $\mathcal{E}(t)=\frac{1}{2}\|x(t)-z\|^2$ is nonincreasing for any fixed $z\in A^{-1}(0)$. This implies that (i) holds true. Further, let $\bar{x}\in\mathcal{H}$ be a weak sequential cluster point of $x(\cdot)$, we claim that it is also a weak sequential cluster point of $y(\cdot)=(I+\lambda(\cdot)A)^{-1}x(\cdot)$. Indeed, Lemma 6.3.4 implies

$$\mathcal{E}(0)-\mathcal{E}(t)\geq\int_0^t\|x(s)-y(s)\|^2\,ds,\quad\text{for all } t\geq0.$$

Since $\mathcal{E}(t)\geq0$, we have $\mathcal{E}(0)-\mathcal{E}(t)\leq\mathcal{E}(0)$. As a direct consequence of Lemma 6.2.10 and Theorem 6.2.6, we have that $\lambda:[0,+\infty)$ is nondecreasing. This together with Eq. (6.4) implies that $t\mapsto\|x(t)-y(t)\|$ is nonincreasing. Putting these pieces together yields

$$\|x(t)-y(t)\|^2\leq\tfrac{\mathcal{E}(0)}{t}\to0,\quad\text{as } t\to+\infty.$$

This implies the desired result that $\bar{x}$ is also a weak sequential cluster point of $y(\cdot)$. In addition, we have $\frac{1}{\lambda(t)}(x(t)-y(t))\in Ay(t)$. Combining $\|x(t)-y(t)\|\to0$ and Eq. (6.4) implies that $\lambda(t)\to+\infty$. Therefore, we have that $\frac{1}{\lambda(t)}\|x(t)-y(t)\|\to0$ as $t\to+\infty$. Since $\bar{x}$ is a weak sequential cluster point of $y(\cdot)$ and the graph of A is demi-closed [Goebel and Kirk, 1990, Chapter 10], we have $0\in A\bar{x}$ and hence $\bar{x}\in A^{-1}(0)$.

Strong convergence. We further establish strong convergence of a global solution of the closed-loop control system in Eq. (6.3) and Eq. (6.4) under additional conditions.

Theorem 6.3.5 *Suppose that $(x,\lambda):[0,+\infty)\mapsto\mathcal{H}\times(0,+\infty)$ is a global solution of the closed-loop control system in Eq. (6.3) and Eq. (6.4). Then, there exists some $\bar{x}\in A^{-1}(0)$ such that the trajectory $x(t)$ converges strongly to $\bar{x}$ as $t\to+\infty$ if either of the following conditions holds true:*

1. *$A=\nabla\Phi$ where $\Phi:\mathcal{H}\mapsto\mathbb{R}\cup\{+\infty\}$ is convex, differentiable and inf-compact.*[1]

[1] A function Φ is *inf-compact* if for any $r>0$ and $\kappa\in\mathbb{R}$, the set $\{x\in\mathcal{H}:\|x\|\leq r,\Phi(x)\leq\kappa\}$ is a relatively compact set in $\mathcal{H}$, i.e., the set whose closure is compact.

2. $A^{-1}(0)$ has a nonempty interior.

Remark 6.3.6 *In the Hilbert-space setting, the strong convergence is desirable since it guarantees that $\|x(t)-\bar{x}\|$ eventually becomes arbitrarily small [Bauschke and Combettes, 2001]. It has been studied for various discrete-time algorithms in convex optimization [Solodov and Svaiter, 2000] and the realization of the importance of strong convergence dates to Güler [1991] who showed that the convergence rate of the sequence of objectives $\{\Phi(x_k)\}_{k\geq 0}$ is better when $\{x_k\}_{k\geq 0}$ with strong convergence than weak convergence. In addition, the conditions assumed in Theorem 6.3.5 can be verifiable by hand and the similar results can be obtained using the generalized chain rule for the subdifferential in the case $A=\partial\Phi$.*

Proof of Theorem 6.3.5: For the first case, we claim that $t\mapsto\Phi(x(t))$ is nonincreasing. Indeed, we let $y(t)=(I+\lambda(t)\nabla\Phi)^{-1}x(t)$ and deduce from the convexity of Φ that

$$\tfrac{d\Phi(x(t))}{dt}=\langle\dot{x}(t),\nabla\Phi(x(t))\rangle\overset{\text{Eq. (6.3)}}{=}\langle y(t)-x(t),\nabla\Phi(x(t))\rangle\leq\langle y(t)-x(t),\nabla\Phi(y(t))\rangle.$$

By the definition of $y(t)$, we have $\lambda(t)\nabla\Phi(y(t))+y(t)-x(t)=0$. This implies

$$\langle y(t)-x(t),\nabla\Phi(y(t))\rangle=-\lambda(t)\|\nabla\Phi(y(t))\|^2\leq 0.$$

Putting these pieces together yields the desired result. As such, it is immediate to see that $x(\cdot)$ is contained in $\{x\in\mathcal{H}:\Phi(x)\leq\Phi(x_0)\}$. By Lemma 6.3.4, we have $\mathcal{E}(t)=\frac{1}{2}\|x(t)-z\|^2$ is nonincreasing for any fixed $z\in A^{-1}(0)$. Thus, letting $x^\star\in A^{-1}(0)$ with $\|x^\star\|$ finite, we have that $\mathcal{E}(0)$ is finite and

$$x(t)\in S_0\doteq\left\{x\in\mathcal{H}:\Phi(x)\leq\Phi(x_0),\|x\|\leq\|x^\star\|+\sqrt{2\mathcal{E}(0)}\right\},\quad\text{for all } t\in[0,+\infty).$$

Since $\Phi:\mathcal{H}\mapsto\mathbb{R}\cup\{+\infty\}$ is inf-compact on any bounded set, we have that S_0 is relatively compact. This implies that the trajectory $x(\cdot)$ is relatively compact. By Theorem 6.3.1, there exists some $\bar{x}\in A^{-1}(0)$ such that $x(t)$ converges weakly to $\bar{x}$ as $t\to+\infty$. As such, we conclude the desired result.

For the second case, letting $x^\star\in\mathcal{H}$ be a point in the interior of $A^{-1}(0)$, there exists $\delta>0$ such that $\mathbb{B}_\delta(x^\star)\subseteq A^{-1}(0)$. Denoting $A_\lambda=I-(I+\lambda A)^{-1}$, we have

$$x\in A^{-1}(0)\Longleftrightarrow 0\in Ax\Longleftrightarrow x=(I+\lambda A)^{-1}x\Longleftrightarrow 0=A_\lambda x\Longleftrightarrow x\in A_\lambda^{-1}(0),$$

which implies that $A^{-1}(0)=A_\lambda^{-1}(0)$ and $\mathbb{B}_\delta(x^\star)\subseteq A_\lambda^{-1}(0)$ for any $\lambda>0$. It is also well known that A_λ is monotone [Rockafellar, 1970] Since $\mathbb{B}_\delta(x^\star)\subseteq A_\lambda^{-1}(0)$, we have $x^\star+\delta h\in A_\lambda^{-1}(0)$ for any $h\in\mathcal{H}$ with $\|h\|\leq 1$. This together with the monotonicity of A_λ yields

$$\langle A_\lambda x(t),x(t)-(x^\star+\delta h)\rangle\geq 0,$$

which implies

$$\delta\langle A_\lambda x(t),h\rangle\leq\langle A_\lambda x(t),x(t)-x^\star\rangle.\tag{6.19}$$

Combining the above inequality with Eq. (6.3) yields

$$\|\dot{x}(t)\| = \|A_{\lambda(t)}x(t)\| = \sup_{\|h\|\leq 1} \langle A_{\lambda(t)}x(t), h\rangle \overset{\text{Eq. (6.19)}}{\leq} \tfrac{1}{\delta}\langle A_{\lambda(t)}x(t), x(t)-x^\star\rangle = -\tfrac{1}{\delta}\langle \dot{x}(t), x(t)-x^\star\rangle.$$

Then, we let $0 \leq t_1 \leq t_2$ and deduce from the above inequality that

$$\|x(t_2)-x(t_1)\| \leq \int_{t_1}^{t_2} \|\dot{x}(s)\|\, ds \leq -\tfrac{1}{\delta}\left(\int_{t_1}^{t_2} \langle \dot{x}(s), x(s)-x^\star\rangle\right) \leq \tfrac{1}{2\delta}\left(\|x(t_1)-x^\star\|^2 - \|x(t_2)-x^\star\|^2\right).$$

Since $x^\star \in A^{-1}(0)$, we deduce from Lemma 6.3.4 that $\|x(t) - x^\star\|$ is nonincreasing and convergent. Thus, the trajectory $x(\cdot)$ has the Cauchy property.

Rate of convergence. We prove the ergodic convergence rate of $O(t^{-(p+1)/2})$ for a global solution of the closed-loop control system in Eq. (6.3) and Eq. (6.4) in terms of a gap function. We also prove a pointwise convergence rate of $O(t^{-p/2})$ in terms of a residue function, and then establish local linear convergence for a global solution in terms of a distance function.

Before stating our results, we provide the gap function and the residue function for monotone inclusion problems. Indeed, the following gap function originates from the Fitzpatrick function [Borwein and Lewis, 2010] and is also defined in the concurrent work of Cui et al. [2022]. Formally, we have

$$\textsc{gap}(x) = \sup_{z\in \mathrm{dom}(A)} \sup_{\xi\in Az} \langle \xi, x - z\rangle. \tag{6.20}$$

Clearly, $\textsc{gap}(\cdot)$ is closed[2] and convex. Moreover, if A is maximal monotone, we have that $\textsc{gap}(x) \geq 0$ for all $x \in \mathcal{H}$ with equality if and only if $x \in A^{-1}(0)$ holds. The residue function is derived from the monotone inclusion problem as follows,

$$\textsc{res}(x) = \inf_{\xi\in Ax} \|\xi\|. \tag{6.21}$$

We are now ready to present our main results on the global convergence rate estimation in terms of the gap function in Eq. (6.20) and the residue function in Eq. (6.21).

Theorem 6.3.7 *Suppose that $(x, \lambda) : [0, +\infty) \mapsto \mathcal{H} \times (0, +\infty)$ is a global solution of the closed-loop control system in Eq. (6.3) and Eq. (6.4) and let $\mathrm{dom}(A)$ be closed and bounded. Then, we have*

$$\textsc{gap}(\tilde{z}(t)) = O(t^{-\frac{p+1}{2}}),$$

and

$$\textsc{res}(z(t)) = O(t^{-\frac{p}{2}}),$$

[2] A function $\Phi : \mathcal{H} \mapsto \mathbb{R}$ is closed if the sublevel set $\{x \in \mathcal{H} : \Phi(x) \leq \alpha\}$ is closed for any $\alpha \in \mathbb{R}$; see Rockafellar [1970].

where $\tilde{z}(\cdot)$ and $z(\cdot)$ are uniquely determined by $\lambda(\cdot)$ and $x(\cdot)$ as follows,

$$\textbf{(Ergodic Iterate)} \qquad \tilde{z}(t) = \frac{1}{\int_0^t \lambda(s)\,ds}\left(\int_0^t \lambda(s)(I+\lambda(s)A)^{-1}x(s)\,ds\right),$$
$$\textbf{(Pointwise Iterate)} \qquad z(t) = (I+\lambda(t)A)^{-1}x(t).$$

Remark 6.3.8 *Theorem 6.3.7 is new to the best of our knowledge and extends several classical results concerning discrete-time algorithms for monotone inclusion problems. Indeed, the discrete-time version of our results have been obtained by the extragradient method for $p = 1$ [Nemirovski, 2004, Monteiro and Svaiter, 2010, 2011] and the Newton proximal extragradient method for $p = 2$ [Monteiro and Svaiter, 2012]. A similar ergodic convergence result was achieved by high-order mirror-prox method [Bullins and Lai, 2022] for saddle point and VI problems for $p \geq 3$. Notably, our theorem demonstrates the importance of averaging for monotone inclusion problems by showing that the convergence rate can be faster in the ergodic sense for all $p \geq 1$. The idea of averaging for convex optimization and monotone VIs goes back to at least the mid-seventies [Bruck Jr, 1977, Lions, 1978, Nemirovski and Yudin, 1978, Nemirovski, 1981]. Its advantage was also recently justified for saddle point problems and VIs by establishing lower bounds [Golowich et al., 2020a, Ouyang and Xu, 2021]. Our theorem provides another way to understand averaging from a continuous-time point of view.*

We define the so-called error bound condition as an inequality that bounds the distance between $x \in \mathcal{H}$ and $A^{-1}(0)$ by a residual function at x. This condition has been proven to be useful in proving the linear convergence of discrete-time algorithms for solving convex optimization and monotone VI problems [Lewis and Pang, 1998, Drusvyatskiy and Lewis, 2018, Drusvyatskiy et al., 2021]. We adapt this condition for monotone inclusion problems as follows. We assume that there exists $\delta > 0$ and $\kappa > 0$ such that

$$\textsc{dist}(0, Ax) \leq \delta \Longrightarrow \textsc{dist}(x, A^{-1}(0)) \leq \kappa \cdot \textsc{dist}(0, Ax), \tag{6.22}$$

where $\textsc{dist}(x, S) = \inf_{z\in S} \|x - z\|$ is a distance function. The corresponding Lyapunov function used for analyzing a global solution under the error bound condition is as follows:

$$\tilde{\mathcal{E}}(t) = \frac{1}{2}(\textsc{dist}(x(t), A^{-1}(0)))^2 \doteq \inf_{z\in A^{-1}(0)}\left\{\frac{1}{2}\|x(t) - z\|^2\right\}. \tag{6.23}$$

The Lyapunov function in Eq. (6.23) can be interpreted as a continuous version of a function used by various authors; see e.g., Tseng [1995]. The convergence rate estimation intuitively depends on the descent inequality. This requires the differentiation of $\tilde{\mathcal{E}}(\cdot)$ which is not immediate since $A^{-1}(0)$ is not a singleton set and the projection of $x(t)$ onto $A^{-1}(0)$ will change as t varies. We instead upper bound the difference $\tilde{\mathcal{E}}(t') - \mathcal{E}(t)$ for any $t' \geq t$ given a fixed t. We have the following theorem.

Theorem 6.3.9 *Suppose that $(x, \lambda) : [0, +\infty) \mapsto \mathcal{H} \times (0, +\infty)$ is a global solution of the closed-loop control system in Eq. (6.3) and Eq. (6.4) and let the error bound condition in Eq. (6.22) hold true. Then, there exists a sufficiently large $t_0 > 0$ such that*

$$\textsc{dist}(x(t), A^{-1}(0)) = O(e^{-ct/2}), \quad \text{for all } t > t_0.$$

where $c > 0$ is a constant and upper bounded by $c \leq 2\left(1 + \frac{\kappa}{\lambda(0)}\right)^{-2}$.

Remark 6.3.10 *Theorem 6.3.9 establishes the strong convergence of $x(\cdot)$ to some $x^\star \in A^{-1}(0)$ under the error bound condition and establishes local linear convergence in terms of a distance function. This improves the results in Theorem 6.3.7 and demonstrates the value of the error bound condition. The same linear convergence guarantee is established in Csetnek et al. [2021] under similar conditions. In fact, the convergence analysis of discrete-time algorithms under an error bound condition is of independent interest [Solodov, 2003] and its analysis involves different techniques.*

Proof of Theorem 6.3.7: Using the definition of $\mathcal{E}(\cdot)$ in Eq. (6.18) and the same argument as applied in Lemma 6.3.4, we have

$$\tfrac{d\mathcal{E}(t)}{dt} = -\|x(t) - (I + \lambda(t)A)^{-1}x(t)\|^2 - \langle x(t) - (I + \lambda(t)A)^{-1}x(t), (I + \lambda(t)A)^{-1}x(t) - z\rangle.$$

Using the definition of $z(\cdot)$ and the fact that $\|x(t) - (I + \lambda(t)A)^{-1}x(t)\|^2 \geq 0$, we have

$$\tfrac{d\mathcal{E}(t)}{dt} \leq -\langle x(t) - z(t), z(t) - z\rangle.$$

Since A is monotone and $\frac{1}{\lambda(t)}(x(t) - z(t)) \in Az(t)$, we have

$$\langle x(t) - z(t), z(t) - z\rangle \geq \lambda(t)\langle \xi, z(t) - z\rangle, \quad \text{for all } \xi \in Az.$$

Putting these pieces together yields that, for any $z \in \mathrm{dom}(A)$ and any $\xi \in Az$, we have

$$\tfrac{d\mathcal{E}(t)}{dt} \leq -\lambda(t)\langle \xi, z(t) - z\rangle.$$

Integrating this inequality over $[0, t]$ yields

$$\int_0^t \lambda(s)\langle \xi, z(s) - z\rangle \, ds \leq \mathcal{E}(0) - \mathcal{E}(t) \leq \tfrac{1}{2}\|x_0 - z\|^2, \quad \text{for all } t \geq 0.$$

Equivalently, we have

$$\langle \xi, \tilde{z}(t) - z\rangle \leq \tfrac{1}{\int_0^t \lambda(s)\, ds}\left(\tfrac{1}{2}\|x_0 - z\|^2\right) \leq \tfrac{1}{\int_0^t \lambda(s)\, ds}\left(\sup_{z \in \mathrm{dom}(A)} \|x_0 - z\|^2\right).$$

By the definition of $\textsc{gap}(\cdot)$ and using the boundedness of $\mathrm{dom}(A)$, we have

$$\textsc{gap}(\tilde{z}(t)) = O\left(\frac{1}{\int_0^t \lambda(s)\, ds}\right). \tag{6.24}$$

Since $z(t) = (I + \lambda(t)A)^{-1}x(t)$, we can obtain from the proof of Theorem 6.3.1 that

$$\|x(t) - z(t)\|^2 \leq \frac{\mathcal{E}(0)}{t}, \quad \text{for all } t \geq 0. \tag{6.25}$$

Since $\frac{1}{\lambda(t)}(x(t) - z(t)) \in Az(t)$, we have

$$\textsc{res}(z(t)) \leq \frac{1}{\lambda(t)}\|x(t) - z(t)\| \leq \frac{\mathcal{E}(0)}{\lambda(t)\sqrt{t}} = O\left(\frac{1}{\lambda(t)\sqrt{t}}\right). \tag{6.26}$$

It remain to estimate the lower bound for the feedback law $\lambda(\cdot)$. Indeed, by combining Eq. (6.25) and the algebraic equation in Eq. (6.4), we have

$$\lambda(t) = \frac{\theta}{\|x(t)-z(t)\|^{p-1}} \geq \theta\left(\frac{t}{\mathcal{E}(0)}\right)^{\frac{p-1}{2}}. \tag{6.27}$$

Plugging Eq. (6.27) into Eq. (6.24) and Eq. (6.26) yields the desired results.

Proof of Theorem 6.3.9: Fixing $t \geq 0$ and using the definition of $\tilde{\mathcal{E}}$ in Eq. (6.23), we have

$$\tilde{\mathcal{E}}(t') - \tilde{\mathcal{E}}(t) = \tfrac{1}{2}\left((\textsc{dist}(x(t'), A^{-1}(0)))^2 - (\textsc{dist}(x(t), A^{-1}(0)))^2\right), \quad \text{for all } t' \geq t.$$

We let $x^\star(t)$ denote the projection of $x(t)$ onto $A^{-1}(0)$ and deduce from the above inequality that

$$\frac{\tilde{\mathcal{E}}(t')-\tilde{\mathcal{E}}(t)}{t'-t} \leq \frac{\|x(t')-x^\star(t)\|^2-\|x(t)-x^\star(t)\|^2}{2(t'-t)} = \langle \frac{x(t')-x(t)}{t'-t}, \frac{x(t')+x(t)}{2} - x^\star(t)\rangle.$$

Letting $t' \to^+ t$, we have

$$\limsup_{t'\to^+ t} \frac{\tilde{\mathcal{E}}(t')-\tilde{\mathcal{E}}(t)}{t'-t} \leq \langle \dot{x}(t), x(t) - x^\star(t)\rangle. \tag{6.28}$$

For simplicity, we let $y(t) = (I + \lambda(t)A)^{-1}x(t)$ and deduce that $\frac{1}{\lambda(t)}(x(t) - y(t)) \in Ay(t)$. In addition, $0 \in Ax^\star(t)$. Putting these pieces together with the monotonicity of A yields

$$\begin{aligned} \langle \dot{x}(t), x(t) - x^\star(t)\rangle &\overset{\text{Eq. (6.3)}}{=} -\langle x(t) - y(t), x(t) - x^\star(t)\rangle \\ &= -\|x(t) - y(t)\|^2 - \langle x(t) - y(t), y(t) - x^\star(t)\rangle \leq -\|x(t) - y(t)\|^2. \end{aligned} \tag{6.29}$$

Using the same argument in the proof of Theorem 6.3.1, we have $t \mapsto \frac{1}{\lambda(t)}\|x(t) - y(t)\|$ is nonincreasing and converges to zero as $t \to +\infty$. So there exists a sufficiently large $t_0 > 0$ such that

$$\frac{1}{\lambda(t)}\|x(t) - y(t)\| \leq \delta, \quad \text{for all } t \geq t_0,$$

where $\delta > 0$ is defined in the error bound condition (cf. Eq. (6.22)). Recall that $\frac{1}{\lambda(t)}(x(t) - y(t)) \in Ay(t)$, we have $\textsc{dist}(0, Ay(t)) \leq \delta$. Since the error bound condition in Eq. (6.22) holds true, we have

$$\textsc{dist}(y(t), A^{-1}(0)) \leq \kappa \cdot \textsc{dist}(0, Ay(t)),$$

We let $y^\star(t)$ denote the projection of $y(t)$ onto $A^{-1}(0)$ and deduce from the triangle inequality that

$$\textsc{dist}(x(t), A^{-1}(0)) \leq \|x(t) - y^\star(t)\| \leq \|x(t) - y(t)\| + \textsc{dist}(y(t), A^{-1}(0)).$$

Putting these pieces together yields

$$\textsc{dist}(x(t), A^{-1}(0)) \leq \|x(t) - y(t)\| + \kappa \cdot \textsc{dist}(0, Ay(t)) \leq \left(1 + \tfrac{\kappa}{\lambda(t)}\right) \|x(t) - y(t)\|.$$

Since $\lambda(t)$ is nondecreasing (cf. Lemma 6.2.10), we have $\lambda(t) \geq \lambda(0)$. By the definition of $\tilde{\mathcal{E}}$, we have

$$\tilde{\mathcal{E}}(t) = \tfrac{1}{2}(\textsc{dist}(x(t), A^{-1}(0)))^2 \leq \tfrac{1}{2}\left(1 + \tfrac{\kappa}{\lambda(0)}\right)^2 \|x(t) - y(t)\|^2. \tag{6.30}$$

Plugging Eq. (6.29) and Eq. (6.30) into Eq. (6.28), we have

$$\limsup_{t' \to^+ t} \tfrac{\tilde{\mathcal{E}}(t') - \tilde{\mathcal{E}}(t)}{t' - t} \leq -2\left(1 + \tfrac{\kappa}{\lambda(0)}\right)^{-2} \tilde{\mathcal{E}}(t) \leq -c \cdot \tilde{\mathcal{E}}(t). \tag{6.31}$$

Fixing $t > 0$, we define a partition of an interval $[0, t)$,

$$0 = t_0 < t_1 < t_2 < \ldots < t_i < \ldots < t_n = t,$$

with $\sup_{0 \leq i \leq n-1} |t_{i+1} - t_i| \leq h$. Here, $h > 0$ is sufficiently small such that Eq. (6.31) guarantees

$$\tfrac{\tilde{\mathcal{E}}(t_{i+1}) - \tilde{\mathcal{E}}(t_i)}{t_{i+1} - t_i} \leq -c \cdot \tilde{\mathcal{E}}(t_i), \quad \text{for all } i \in \{0, 1, 2, \ldots, n-1\}.$$

This inequality implies

$$\tilde{\mathcal{E}}(t) - \tilde{\mathcal{E}}(0) = \sum_{i=0}^{n-1} (\tilde{\mathcal{E}}(t_{i+1}) - \tilde{\mathcal{E}}(t_i)) \leq -c \cdot \left(\sum_{i=0}^{n-1} \tilde{\mathcal{E}}(t_i)(t_{i+1} - t_i)\right).$$

Since $\tilde{\mathcal{E}}(\cdot) : [0, +\infty) \to [0, +\infty)$ is a continuous function, it is integrable (possibly not differentiable). Letting $h \to 0$, we have

$$\sum_{i=0}^{n-1} \tilde{\mathcal{E}}(t_i)(t_{i+1} - t_i) \to \int_0^t \tilde{\mathcal{E}}(s)\, ds.$$

Putting these pieces together yields

$$\tilde{\mathcal{E}}(t) - \tilde{\mathcal{E}}(0) \leq -c \left(\int_0^t \tilde{\mathcal{E}}(s)\, ds\right).$$

Recall the Grönwall–Bellman inequality in the integral form [Gronwall, 1919, Bellman, 1943]: if $u(\cdot)$ and $\beta(\cdot)$ are both continuous and satisfy the integral inequality: $u(t) \leq u_0 + \int_0^t \beta(s)u(s)ds$, we have

$$u(t) \leq u_0 \exp\left(\int_0^t \beta(s)\, ds\right).$$

This implies that $\tilde{\mathcal{E}}(t) \leq \tilde{\mathcal{E}}(0)e^{-ct}$. Therefore, we conclude that there exists a sufficiently large $t_0 > 0$ such that

$$\textsc{dist}(x(t), A^{-1}(0)) = O(e^{-ct/2}), \quad \text{for all } t > t_0.$$

This completes the proof.

Discussion. We comment on the main techniques for analyzing the system in Eq. (6.3) and Eq. (6.4), including Lyapunov analysis and weak versus strong convergence. We also compare our approach to other approaches based on time scaling and dry friction.

Lyapunov analysis. Key to the continuous-time approach is to derive inertial gradient systems as limits of discrete-time algorithms and interpret the acceleration as the effect of asymptotically vanishing damping and Hessian-driven damping. Analyzing such a dynamical system requires a more complicated Lyapunov function than Eq. (6.18). In this context, Wilson et al. [2021] have constructed a unified Lyapunov function and their analysis was shown to be equivalent to Nesterov's estimate sequence analysis for a variety of first-order algorithms, including quasi-monotone subgradient, accelerated gradient descent and conditional gradient. In contrast, the associated dynamical systems for general monotone inclusion problems need not contain any inertial term [Attouch and Svaiter, 2011, Attouch et al., 2013b, Abbas et al., 2014, Attouch et al., 2016a]. However, this does not mean that inertia is not relevant outside optimization. Indeed, the inertial dynamical systems and their discretization [Attouch and Maingé, 2011] give a family of accelerated first-order algorithms for monotone inclusion problems under the cocoercive condition. Nonetheless, the Lyapunov analysis in the current paper becomes quite simple since our system does not involve any inertial term. In addition, the analysis of the convergence rate estimation under an error bound condition involves a new Lyapunov function that can be of independent interest.

Weak versus strong convergence. In the Hilbert-space setting, the (generalized) steepest descent dynamical system associated to a convex potential function Φ has the following form:

$$\begin{cases} -\dot{x}(t) \in \partial\Phi(x(t)), \\ x(0) = x_0. \end{cases}$$

It is well known that the trajectory converges to a point $\bar{x} \in \{x : f(x) = \inf_{x\in\mathcal{H}} f(x)\} \neq \emptyset$ [Brézis, 1973, 1978]. However, the theoretical understanding is far from being complete.

In particular, it remains open how to characterize the relationship between $\bar{x}$ and the initial point x_0 [Lemaire, 1996]. There is also a famous counterexample [Baillon, 1978] which shows that the trajectories of the above system converge weakly but not strongly. Despite the progress on weak versus strong convergence of a regularized Newton dynamic for monotone inclusion problems [Attouch and Baillon, 2018], we are not aware of any discussion about these properties for closed-loop control systems and consider it an interesting open problem to find a counterexample (weak versus strong convergence) for the system in Eq. (6.3) and Eq. (6.4).

It is worth mentioning that the convergence results of trajectories are important aspects of the convergence analysis, especially in an infinite-dimensional setting; indeed, these results have been established for the trajectories of various dynamical systems for monotone inclusion problems in earlier research [Attouch and Svaiter, 2011, Attouch et al., 2013b, 2016a, Abbas et al., 2014, Bot and Csetnek, 2016, Attouch and Peypouquet, 2019, Attouch and Cabot, 2020, Attouch and László, 2020b, 2021]. A few results are valid only for weak convergence and become true for strong convergence only under additional conditions. Some results are only valid in the ergodic sense, e.g., the rate of $O(t^{-(p+1)/2})$ in Theorem 6.3.7.

Time scaling and dry friction. In the context of dissipative dynamical systems associated with convex optimization algorithms, there have been two simple yet universally powerful techniques to strengthen the convergence properties of trajectories: time scaling [Attouch et al., 2019c, 2022a] and dry friction [Adly and Attouch, 2020, 2022]. In particular, the effect of time scaling is revealed by the coefficient parameter $b(t)$ which comes in as a factor of $\nabla\Phi(x(t))$ in the following open-loop inertial gradient system:

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \beta(t)\nabla^2\Phi(x(t))\dot{x}(t) + b(t)\nabla\Phi(x(t)) = 0.$$

In Attouch et al. [2019c], the authors investigated the above system without Hessian-driven damping ($\beta(t) = 0$). They proved that the convergence rate of a solution trajectory is $O(1/(t^2 b(t)))$ if $\alpha(\cdot)$ and $b(\cdot)$ satisfy certain conditions. As such, a clear improvement is attained by taking $b(t) \to +\infty$. This demonstrates the power and potential of time scaling, as further evidenced by recent work on systems with Hessian damping [Attouch et al., 2022a]. Furthermore, some recent work studied another open-loop inertial gradient system in the form of

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \partial\phi(\dot{x}(t)) + \beta(t)\nabla^2\Phi(x(t))\dot{x}(t) + b(t)\nabla\Phi(x(t)) \ni 0,$$

where the dry friction function ϕ is convex with a sharp minimum at the origin, e.g., $\phi(x) = r\|x\|$ with $r > 0$. In Adly and Attouch [2022], the authors provided a study of the convergence of this system without Hessian-driven damping ($\beta(t) = 0$) and derived a class of appealing first-order algorithms that achieve a finite convergence guarantee. Subsequently, Adly and Attouch [2020] derived similar results for systems with Hessian-driven damping.

Unfortunately, the aforementioned works on time scaling and dry friction techniques are restricted to the study of open-loop systems associated with convex optimization algorithms. As such, it remains unknown if these methodologies can be extended to monotone inclusion

problems and further capture the continuous-time interpretation of acceleration in high-order monotone inclusion [Monteiro and Svaiter, 2012, Bullins and Lai, 2022]. In contrast, our closed-loop control system provides a rigorous justification for the large-step condition in the algorithm of Monteiro and Svaiter [2012] and Bullins and Lai [2022] when $p \geq 2$, explaining why the closed-loop control is key to acceleration in monotone inclusion.

6.4 Implicit Discretization and Acceleration

We propose an algorithmic framework that arises via implicit discretization of our system in Eq. (6.3) and Eq. (6.4) in a Euclidean setting. It demonstrates the importance of the large-step condition [Monteiro and Svaiter, 2012] for acceleration in monotone inclusion problems, interpreting it as discretization of the algebraic equation. Our framework clarifies why this condition is unnecessary for acceleration in monotone inclusion problems when $p = 1$ (the algebraic equation vanishes). With an approximate tensor subroutine for smooth operator A, we derive a specific class of p^{th}-order tensor algorithms which generalize p^{th}-order tensor algorithms for convex-concave saddle point and monotone variational inequality problems [Bullins and Lai, 2022].

Conceptual algorithmic frameworks. We study a conceptual algorithmic framework which is derived by implicit discretization of the closed-loop control system in Eq. (6.3) and Eq. (6.4) in a Euclidean setting. Indeed, our system takes the form of

$$\begin{cases} \dot{x}(t) + x(t) - (I + \lambda(t)A)^{-1}x(t) = 0, \\ \lambda(t)\|(I + \lambda(t)A)^{-1}x(t) - x(t)\|^{p-1} = \theta, \\ x(0) = x_0 \in \Omega. \end{cases}$$

We define the discrete-time sequence $\{(x_k, \lambda_k)\}_{k\geq 0}$ that corresponds to its continuous-time counterpart $\{(x(t), \lambda(t))\}_{t\geq 0}$. By an implicit discretization, we have

$$\begin{cases} x_{k+1} - (I + \lambda_{k+1}A)^{-1}x_k = 0, \\ \lambda_{k+1}\|x_{k+1} - x_k\|^{p-1} = \theta, \\ x_0 \in \Omega. \end{cases} \tag{6.32}$$

By introducing two new variables y_{k+1} and $v_{k+1} \in Ay_{k+1}$, the first and second lines of Eq. (6.32) can be equivalently reformulated as follows:

$$\begin{cases} \lambda_{k+1}v_{k+1} + y_{k+1} - x_k = 0, \\ \lambda_{k+1}\|y_{k+1} - x_k\|^{p-1} = \theta, \\ x_{k+1} = x_k - \lambda_{k+1}v_{k+1}, \\ x_0 \in \Omega. \end{cases}$$

Algorithm 18 Conceptual Algorithmic Framework

STEP 0: Let $x_0, v_0 \in \mathbb{R}^d$, $\sigma \in (0,1)$ and $\theta > 0$ be given, and set $k = 0$.
STEP 1: If $0 \in Ax_k$, then **stop**.
STEP 2: Otherwise, compute $\lambda_{k+1} > 0$ and a triple $(y_{k+1}, v_{k+1}, \epsilon_{k+1}) \in \mathcal{H} \times \mathcal{H} \times (0, +\infty)$ such that

$$\begin{aligned} &v_{k+1} \in A^{\epsilon_{k+1}}(y_{k+1}), \\ &\|\lambda_{k+1}v_{k+1} + y_{k+1} - x_k\|^2 + 2\lambda_{k+1}\epsilon_{k+1} \le \sigma^2\|y_{k+1} - x_k\|^2, \\ &\lambda_{k+1}\|y_{k+1} - x_k\|^{p-1} \ge \theta. \end{aligned}$$

STEP 3: Compute $x_{k+1} = x_k - \lambda_{k+1}v_{k+1}$.
STEP 4: Set $k \leftarrow k + 1$, and go to **STEP 1**.

We propose to solve the above equations inexactly with an accurate approximation of A. Following a suggestion of Monteiro and Svaiter [2010], we introduce the relative error tolerance condition with an ε-enlargement of maximal monotone operators given by

$$A^{\epsilon}(x) = \{v \in \mathbb{R}^d \mid \langle x - \tilde{x}, v - \tilde{v}\rangle \ge -\epsilon,\ \forall \tilde{x} \in \mathbb{R}^d, \forall \tilde{v} \in A\tilde{x}\}. \tag{6.33}$$

Our subroutine is to find $\lambda_{k+1} > 0$ and a triple $(y_{k+1}, v_{k+1}, \varepsilon_{k+1})$ such that

$$\|\lambda_{k+1}v_{k+1} + y_{k+1} - x_k\|^2 + 2\lambda_{k+1}\epsilon_{k+1} \le \sigma^2\|y_{k+1} - x_k\|^2, \quad v_{k+1} \in A^{\epsilon_{k+1}}(y_{k+1}).$$

From the above condition, we see that v_{k+1} is sufficiently close to an element in $A(y_{k+1})$. In addition, we relax the discrete-time algebraic equation by using $\lambda_{k+1}\|y_{k+1} - x_k\|^{p-1} \ge \theta$.

We present our conceptual algorithmic framework formally in Algorithm 18. It includes the large-step HPE framework of Monteiro and Svaiter [2012] as a special instance. In fact, we can recover the large-step HPE framework if we set $p = 2$ and change the notation of A to T in Algorithm 18.

Convergence rate estimation. We present both an ergodic and a pointwise estimate of convergence rate for Algorithm 18. Our analysis is motivated by the aforementioned continuous-time analysis, simplifying the analysis in Monteiro and Svaiter [2012] for the case of $p = 2$ and generalizing it to the case of $p > 2$ in a systematic manner.

We start with the presentation of our main results for Algorithm 18, which generalizes the results in Monteiro and Svaiter [2012, Theorem 2.5 and 2.7] in terms of a gap function for bounded domain from $p = 2$ to $p \ge 2$. To streamline the presentation, we rewrite a gap function in Eq. (6.20):

$$\textsc{gap}(x) = \sup_{z \in \mathrm{dom}(A)} \sup_{\xi \in Az} \langle \xi, x - z\rangle.$$

It is worth mentioning that the theoretical results in Monteiro and Svaiter [2012, Theorem 2.5 and 2.7] are presented for unbounded domains using the modified optimality criterion. Our analysis can be extended using the relationship between a gap function and a relative tolerance error criterion [Monteiro and Svaiter, 2010]. However, the proof becomes significantly longer and its link with continuous-time analysis becomes unclear (the continuous-time version of the modified optimality criterion is unclear). Accordingly, we focus on the bounded domain and present the results for simplicity.

Theorem 6.4.1 *Let $k \geq 1$ be an integer and let $\mathrm{dom}(A)$ be closed and bounded. Then, we have*

$$\textsc{gap}(\tilde{y}_k) = O(k^{-\frac{p+1}{2}}),$$

and

$$\inf_{1\leq i\leq k} \|v_i\| = O(k^{-\frac{p}{2}}), \qquad \inf_{1\leq i\leq k} \epsilon_i = O(k^{-\frac{p+1}{2}}),$$

where the ergodic iterates $\{\tilde{y}_k\}_{k\geq 1}$ are defined by

$$\tilde{y}_k = \frac{1}{\sum_{i=1}^k \lambda_k}\left(\sum_{i=1}^k \lambda_i y_i\right).$$

In addition, if we let $\epsilon_k = 0$ for all $k \geq 1$ and assume that the error bound condition in Eq. (6.22) holds true, the iterates $\{x_k\}_{k\geq 1}$ converge to $A^{-1}(0)$ with a local linear rate.

Since the only difference between Algorithm 18 and the large-step HPE framework in Monteiro and Svaiter [2012] is the order in the algebraic equation, many technical results still hold for Algorithm 18 but their proofs tend to involve case-specific algebra. Our key contribution is to provide a simple proof which flows from the unified underlying continuous-time principle, and also to derive local linear convergence under the error bound condition.

We present a discrete-time Lypanunov function for Algorithm 18 as follows:

$$\mathcal{E}_k = \tfrac{1}{2}\|x_k - z\|^2, \tag{6.34}$$

which will be used to prove technical results that pertain to Algorithm 18.

Lemma 6.4.2 *For every integer $k \geq 1$, we have*

$$\sum_{i=1}^k \lambda_i\langle v, y_i - z\rangle + \tfrac{1-\sigma^2}{2}\left(\sum_{i=1}^k \|x_{i-1} - y_i\|^2\right) \leq \mathcal{E}_0 - \mathcal{E}_k, \quad \text{for all } v \in Az, \tag{6.35}$$

Letting $\tilde{y}_k = \frac{1}{\sum_{i=1}^k \lambda_i}(\sum_{i=1}^k \lambda_i y_i)$ be the ergodic iterates, we have $\sup_{v\in Az}\langle v, \tilde{y}_k - z\rangle \leq \frac{\mathcal{E}_0}{\sum_{i=1}^k \lambda_i}$. If we further assume that $\sigma < 1$, we have $\sum_{i=1}^k \|x_{i-1} - y_i\|^2 \leq \frac{\|x_0 - z^\star\|^2}{1-\sigma^2}$ for any $z^\star \in A^{-1}(0)$.

Lemma 6.4.3 *For every integer $k \geq 1$ and $\sigma < 1$, there exists $1 \leq i \leq k$ such that*

$$\inf_{1\leq i\leq k} \sqrt{\lambda_i}\|v_i\| \leq \sqrt{\tfrac{1+\sigma}{1-\sigma}} \left(\sum_{i=1}^{k} \lambda_i\right)^{-\frac{1}{2}} \left(\inf_{z^\star \in A^{-1}(0)} \|x_0 - z^\star\|\right),$$
$$\inf_{1\leq i\leq k} \epsilon_i \leq \tfrac{\sigma^2}{2(1-\sigma^2)} \left(\sum_{i=1}^{k} \lambda_i\right)^{-1} \left(\inf_{z^\star \in A^{-1}(0)} \|x_0 - z^\star\|^2\right).$$

We provide a lemma giving a lower bound for $\sum_{i=1}^{k} \lambda_i$. The analysis is motivated by continuous-time analysis for the system in Eq. (6.3) and Eq. (6.4).

Lemma 6.4.4 *For $p \geq 1$ and every integer $k \geq 1$, we have*

$$\sum_{i=1}^{k} \lambda_i \geq \theta \left((1-\sigma^2)\left(\inf_{z^\star \in A^{-1}(0)} \|x_0 - z^\star\|^2\right)\right)^{\frac{p-1}{2}} k^{\frac{p+1}{2}}.$$

Proof of Theorem 6.4.1: For every integer $k \geq 1$, combining Lemma 6.4.2 and Lemma 6.4.4 implies

$$\text{GAP}(\tilde{y}_k) = \sup_{z\in \text{dom}(A)} \sup_{v\in Az} \langle v, \tilde{y}_k - z\rangle \leq \tfrac{1}{2(\sum_{i=1}^{k} \lambda_i)} \left(\sup_{z\in \text{dom}(A)} \|z - x_0\|^2\right) = O(k^{-\frac{p+1}{2}}).$$

Combining Lemma 6.4.3 and Lemma 6.4.4, we have

$$\inf_{1\leq i\leq k} \sqrt{\lambda_i}\|v_i\| \leq \sqrt{\tfrac{1+\sigma}{1-\sigma}} \left(\sum_{i=1}^{k} \lambda_i\right)^{-\frac{1}{2}} \left(\inf_{z^\star \in A^{-1}(0)} \|x_0 - z^\star\|\right) = O(k^{-\frac{p+1}{4}}),$$
$$\inf_{1\leq i\leq k} \epsilon_i \leq \tfrac{\sigma^2}{2(1-\sigma^2)} \left(\sum_{i=1}^{k} \lambda_i\right)^{-1} \left(\inf_{z^\star \in A^{-1}(0)} \|x_0 - z^\star\|^2\right) = O(k^{-\frac{p+1}{2}}).$$

From Step 2 of Algorithm 1, we have

$$\|\lambda_i v_i + y_i - x_{i-1}\|^2 + 2\lambda_i\epsilon_i \leq \sigma^2\|y_i - x_{i-1}\|^2, \quad \lambda_i\|y_i - x_{i-1}\|^{p-1} \geq \theta.$$

Since $\lambda_i \geq 0$ and $\epsilon_i \geq 0$, the first inequality implies

$$\|\lambda_i v_i + y_i - x_{i-1}\| \leq \sigma\|y_i - x_{i-1}\|.$$

By the triangle inequality, we have

$$\sigma\|y_i - x_{i-1}\| \geq \|y_i - x_{i-1}\| - \lambda_i\|v_i\| \Longrightarrow \lambda_i\|v_i\| \geq (1-\sigma)\|y_i - x_{i-1}\|.$$

This inequality together with $\lambda_i\|y_i - x_{i-1}\|^{p-1} \geq \theta$ implies

$$\lambda_i\|v_i\|^{\frac{p-1}{p}} = (\lambda_i)^{\frac{1}{p}}(\lambda_i\|v_i\|)^{\frac{p-1}{p}} \geq \left(\frac{\theta}{\|y_i - x_{i-1}\|^{p-1}}\right)^{\frac{1}{p}} ((1-\sigma)\|y_i - x_{i-1}\|)^{\frac{p-1}{p}} = \theta^{\frac{1}{p}}(1-\sigma)^{\frac{p-1}{p}}.$$

Equivalently, we have

$$\sqrt{\lambda_i} \geq \theta^{\frac{1}{2p}}(1-\sigma)^{\frac{p-1}{2p}}\|v_i\|^{-\frac{p-1}{2p}}.$$

This implies

$$\left(\theta^{\frac{1}{2p}}(1-\sigma)^{\frac{p-1}{2p}}\right)\inf_{1\leq i\leq k}\|v_i\|^{\frac{p+1}{2p}} \leq \inf_{1\leq i\leq k}\sqrt{\lambda_i}\|v_i\| = O(k^{-\frac{p+1}{4}}) \Longrightarrow \inf_{1\leq i\leq k}\|v_i\|^{\frac{p+1}{2p}} = O(k^{-\frac{p+1}{4}}).$$

Therefore, we conclude that

$$\inf_{1\leq i\leq k}\|v_i\| = \left(\inf_{1\leq i\leq k}\|v_i\|^{\frac{p+1}{2p}}\right)^{\frac{2p}{p+1}} = O(k^{-\frac{p}{2}}).$$

It remains to prove that the iterates $\{x_k\}_{k\geq 1}$ converge to $A^{-1}(0)$ with a local linear rate under the error bound condition in Eq. (6.22) and that $\epsilon_k = 0$ for all $k \geq 1$. Indeed, it follows from the proof of Lemma 6.4.2 that

$$\mathcal{E}_k - \mathcal{E}_{k+1} \geq \lambda_{k+1}\langle v, y_{k+1} - z\rangle + \tfrac{1-\sigma^2}{2}\|x_k - y_{k+1}\|^2, \quad \text{for all } v \in Az.$$

Recall that $\mathcal{E}_k = \frac{1}{2}\|x_k - z\|^2$. Thus, we have

$$\|x_k - z\|^2 - \|x_{k+1} - z\|^2 \geq 2\lambda_{k+1}\langle v, y_{k+1} - z\rangle + (1-\sigma^2)\|x_k - y_{k+1}\|^2, \quad \text{for all } v \in Az.$$

Here $z \in \mathbb{R}^d$ can be any point. Then, we set $z = x_k^\star = \operatorname{argmin}_{x\in A^{-1}(0)}\|x - x_k\|$ and choose $v = 0 \in Az$. Plugging into the above inequality implies

$$\|x_k - x_k^\star\|^2 - \|x_{k+1} - x_k^\star\|^2 \geq 2\lambda_{k+1}\langle 0, y_{k+1} - x_k^\star\rangle + (1-\sigma^2)\|x_k - y_{k+1}\|^2 = (1-\sigma^2)\|x_k - y_{k+1}\|^2.$$

By definition, we have $\|x_{k+1} - x_{k+1}^\star\| \leq \|x_{k+1} - x_k^\star\|$ and $\textsc{dist}(x_k, A^{-1}(0)) = \|x_k - x_k^\star\|$. Putting these pieces together yields that, for all $k \geq 1$, we have

$$(\textsc{dist}(x_k, A^{-1}(0)))^2 - (\textsc{dist}(x_{k+1}, A^{-1}(0)))^2 \geq (1-\sigma^2)\|x_k - y_{k+1}\|^2. \tag{6.36}$$

It is worth mentioning that Eq. (6.36) implies that $\|x_k - y_{k+1}\| \to 0$. Using the large step condition that $\lambda_k\|y_k - x_{k-1}\|^{p-1} \geq \theta$, we have $\{\lambda_k\}_{k\geq 1}$ is lower bounded by a constant $\underline{\lambda} > 0$. Further, we have

$$\|\lambda_k v_k\| \leq \|\lambda_k v_k + y_k - x_{k-1}\| + \|y_k - x_{k-1}\| \leq (1+\sigma)\|y_k - x_{k-1}\|,$$

which implies that $\|v_k\| \to 0$ as $k \to +\infty$. Since $\epsilon_k = 0$ for all $k \geq 1$, we have $v_k \in Ay_k$. So there exists a sufficiently large $k_0 > 0$ such that $\textsc{dist}(0, Ay_k) \leq \delta$ for all $k \geq k_0$ where $\delta > 0$

is defined in the error bound condition (cf. Eq. (6.22)). Since the error bound condition in Eq. (6.22) holds true, we have

$$\textsc{dist}(y_{k+1}, A^{-1}(0)) \leq \kappa \cdot \textsc{dist}(0, Ay_{k+1}) \leq \kappa \|v_{k+1}\|.$$

We let $y_{k+1}^\star = \operatorname{argmin}_{y \in A^{-1}(0)} \|y - y_{k+1}\|$ and deduce from the triangle inequality that

$$\textsc{dist}(x_k, A^{-1}(0)) \leq \|x_k - y_{k+1}^\star\| \leq \|x_k - y_{k+1}\| + \textsc{dist}(y_{k+1}, A^{-1}(0)) \leq \|x_k - y_{k+1}\| + \kappa \|v_{k+1}\|.$$

Putting these pieces together yields that

$$\textsc{dist}(x_k, A^{-1}(0)) \leq \left(1 + \tfrac{\kappa}{\lambda_{k+1}}\right) \|y_{k+1} - x_k\| \leq \left(1 + \tfrac{\kappa(1+\sigma)}{\underline{\lambda}}\right) \|y_{k+1} - x_k\|. \tag{6.37}$$

Plugging Eq. (6.37) into Eq. (6.36) yields that

$$(\textsc{dist}(x_k, A^{-1}(0)))^2 - (\textsc{dist}(x_{k+1}, A^{-1}(0)))^2 \geq (1 - \sigma^2) \left(\tfrac{\underline{\lambda}}{\kappa(1+\sigma)+\underline{\lambda}}\right)^2 (\textsc{dist}(x_k, A^{-1}(0)))^2.$$

This completes the proof.

Remark 6.4.5 *The discrete-time analysis in Theorem 6.4.1 is based on the Lyapunov function from Eq. (7.9), which is inspired by the one in Eq. (6.18) and Eq. (6.23). Notably, the proofs of these technical results follow the same path for the continuous-time analysis in Theorem 6.3.7 and 6.3.9.*

Global acceleration and local linear convergence. By instantiating Algorithm 18 with approximate tensor subroutines [Nesterov, 2021b], we develop a new family of p^{th}-order tensor algorithms for monotone inclusion problems with $A = F + H$ in which $F \in \mathcal{G}_L^p(\mathbb{R}^d)$ is a point-to-point operator and H is simple and maximal monotone. We provide new convergence results concerning these tensor algorithms, including an ergodic rate of $O(k^{-(p+1)/2})$ in terms of a gap function, a pointwise rate of $O(k^{-p/2})$ in terms of a residue function, and establish local linear convergence under an error bound condition. Our results extend those results in Monteiro and Svaiter [2012] for second-order algorithms for monotone inclusion problems and complement the analysis in Bullins and Lai [2022] concerning high-order tensor algorithms for saddle point and variational inequality problems.

The proximal point algorithm (PPA) (corresponding to implicit discretization of certain systems) requires solving an exact proximal iteration with proximal coefficient $\lambda > 0$ at each iteration:

$$y = (I + \lambda(F + H))^{-1}(x). \tag{6.38}$$

In many application problem, $H = \partial \mathbf{1}_{\mathcal{X}}$, where $\partial \mathbf{1}_{\mathcal{X}}$ is the subdifferential of an indicator function onto a closed and convex set $\mathcal{X}$. Nevertheless, Eq. (6.38) is still hard when the

Algorithm 19 Accelerated p^{th}-order Tensor Algorithm

STEP 0: Let $x_0 \in \mathbb{R}^d$, $\hat{\sigma} \in (0,1)$ and $0 < \sigma_l < \sigma_u < 1$ such that $\sigma_l(1+\hat{\sigma})^{p-1} < \sigma_u(1-\hat{\sigma})^{p-1}$ and $\sigma = \hat{\sigma} + \sigma_u < 1$ be given, and set $k = 0$.
STEP 1: Compute $x'_k = \mathcal{P}_{\mathrm{dom}(A)}(x_k)$. If $0 \in A(x_k)$, then **stop**.
STEP 2: Otherwise, compute a positive scalar $\lambda_{k+1} > 0$ with a $\hat{\sigma}$-inexact solution $y_{k+1} \in \mathbb{R}^d$ of Eq. (6.40) at (λ_{k+1}, x_k) satisfying that

$$u_{k+1} \in (F_{x'_k} + H)(y_{k+1}), \quad \|\lambda_{k+1}u_{k+1} + y_{k+1} - x_k\| \le \hat{\sigma}\|y_{k+1} - x_k\|,$$

and

$$\tfrac{\sigma_l p!}{L} \le \lambda_{k+1}\|y_{k+1} - x_k\|^{p-1} \le \tfrac{\sigma_u p!}{L}.$$

STEP 3: Compute $v_{k+1} = F(y_{k+1}) + u_{k+1} - F_{x'_k}(y_{k+1})$.
STEP 4: Compute $x_{k+1} = x_k - \lambda_{k+1}v_{k+1}$.
STEP 5: Set $k \leftarrow k+1$, and go to **STEP 1**.

proximal coefficient $\lambda \to +\infty$. Fortunately, when $F \in \mathcal{G}^p_L(\mathbb{R}^d)$, it suffices to solve the subproblem with the $(p-1)^{\text{th}}$-order approximation of F. More specifically, we define

$$F_x(u) = F(x) + \langle DF(x), u - x\rangle + \sum_{j=2}^{p-1} \tfrac{1}{j!} D^{(j)}F(x)[u-x]^j. \tag{6.39}$$

Our proposed algorithms are based on an inexact solution of the following subproblem:

$$y = (I + \lambda(F_x + H))^{-1}(x). \tag{6.40}$$

Clearly, the solution x_v of Eq. (6.40) is unique and satisfies $\lambda F_x(x_v) + H(x_v) + x_v - x = 0$. Thus, we denote a *$\hat{\sigma}$-inexact solution* of Eq. (6.40) at (λ, x) by a vector $y \in \mathbb{R}^d$ satisfying that $u \in (F_{x'} + H)(y)$ and $\|\lambda u + y - x\| \le \hat{\sigma}\|y - x\|$ for some $\hat{\sigma} \in (0,1)$ and $x' = \mathcal{P}_{\mathrm{dom}(A)}(x)$ (recall $A = F + H$).

We summarize our accelerated p^{th}-order tensor algorithm in Algorithm 19 and prove that it is an application of Algorithm 18 with a specific choice of θ.

Proposition 6.4.6 *Algorithm 19 is Algorithm 18 with $\theta = \frac{\sigma_l p!}{L}$ and $\epsilon_k = 0$ for all $k \ge 1$.*

Proof. Letting a tuple $(x_k, v_k, u_k)_{k\ge1}$ be generated by Algorithm 19, it is clear that

$$v_{k+1} = F(y_{k+1}) + u_{k+1} - F_{x'_k}(y_{k+1}) \in (F + H)(y_{k+1}).$$

This is equivalent to that $v_{k+1} \in A^{\epsilon_{k+1}}(y_{k+1})$ in Algorithm 18 with $\varepsilon_k = 0$ for all $k \ge 1$. Further, since $\theta = \frac{\sigma_l p!}{L} > 0$, we have

$$\lambda_{k+1}\|y_{k+1} - x_k\|^{p-1} \ge \tfrac{\sigma_l p!}{L} \Longrightarrow \lambda_{k+1}\|y_{k+1} - x_k\|^{p-1} \ge \theta.$$

It suffices to show that $\|\lambda_{k+1}v_{k+1} + y_{k+1} - x_k\|^2 + 2\lambda_{k+1}\epsilon_{k+1} \leq \sigma^2\|y_{k+1} - x_k\|^2$. Since $\epsilon_k = 0$ for all $k \geq 1$, the above inequality is equivalent to

$$\|\lambda_{k+1}v_{k+1} + y_{k+1} - x_k\| \leq \sigma\|y_{k+1} - x_k\|. \tag{6.41}$$

Indeed, we have

$$\begin{aligned}\|\lambda_{k+1}v_{k+1} + y_{k+1} - x_k\| &\leq \lambda_{k+1}\|v_{k+1} - u_{k+1}\| + \|\lambda_{k+1}u_{k+1} + y_{k+1} - x_k\| \\ &\leq \lambda_{k+1}\|F(y_{k+1}) - F_{x'_k}(y_{k+1})\| + \hat{\sigma}\|y_{k+1} - x_k\|.\end{aligned}$$

Using the definition of $F_{x'_k}$ (cf. Eq. (7.8)) and the fact that $F \in \mathcal{G}^p_L(\mathbb{R}^d)$, we have

$$\lambda_{k+1}\|F(y_{k+1}) - F_{x'_k}(y_{k+1})\| \leq \tfrac{\lambda_{k+1}L}{p!}\|y_{k+1} - x'_k\|^p.$$

Since $y_{k+1} \in \mathrm{dom}(A)$ and $x'_k = \mathcal{P}_{\mathrm{dom}(A)}(x_k)$, we have $\|y_{k+1} - x'_k\| \leq \|y_{k+1} - x_k\|$. This implies

$$\lambda_{k+1}\|F(y_{k+1}) - F_{x'_k}(y_{k+1})\| \leq \tfrac{\lambda_{k+1}L}{p!}\|y_{k+1} - x_k\|^p. \tag{6.42}$$

Since $\sigma = \hat{\sigma} + \sigma_u < 1$, we have

$$\lambda_{k+1}\|y_{k+1} - x_k\|^{p-1} \leq \tfrac{\sigma_u p!}{L} \Longrightarrow \hat{\sigma} + \tfrac{\lambda_{k+1}L}{p!}\|y_{k+1} - x_k\|^{p-1} \leq \hat{\sigma} + \sigma_u = \sigma. \tag{6.43}$$

Combing Eq. (6.42) and Eq. (6.43), we have

$$\lambda_{k+1}\|F(y_{k+1}) - F_{x'_k}(y_{k+1})\| + \hat{\sigma}\|y_{k+1} - x_k\| \leq \left(\tfrac{\lambda_{k+1}L}{p!}\|y_{k+1} - x_k\|^{p-1} + \hat{\sigma}\right)\|y_{k+1} - x_k\| \leq \sigma\|y_{k+1} - x_k\|.$$

Putting these pieces together yields the desired equation in Eq. (6.41). □

In view of Proposition 6.4.6, the iteration complexity derived for Algorithm 18 holds for Algorithm 19. Furthermore, we have $v_k \in (F + H)(y_k) = Ay_k$ for all $k \geq 1$ which implies (see the definition of a residue function in Eq. (6.21))

$$\textsc{res}(y_k) = \inf_{\xi \in Ay_k} \|\xi\| \leq \|v_k\|.$$

As a consequence of Theorem 6.4.1, we summarize the results in the following theorem.

Theorem 6.4.7 *For every integer $k \geq 1$ and let $\mathrm{dom}(A)$ be closed and bounded. Then, we have*

$$\textsc{gap}(\tilde{y}_k) = O(k^{-\frac{p+1}{2}}),$$

and

$$\textsc{res}(y_k) = O(k^{-\frac{p}{2}}),$$

where the ergodic iterates $\{\tilde{y}_k\}_{k\geq 1}$ are defined by

$$\tilde{y}_k = \tfrac{1}{\sum_{i=1}^k \lambda_k}\left(\sum_{i=1}^k \lambda_i y_i\right).$$

In addition, if we assume that the error bound condition in Eq. (6.22) holds true, the iterates $\{x_k\}_{k\geq 1}$ converge to $A^{-1}(0)$ with a local linear rate.

Remark 6.4.8 *The ergodic and pointwise convergence results in Theorem 6.4.7 have been obtained in Monteiro and Svaiter [2012, Theorem 3.5 and 3.6] for the case of $p = 2$ and derived by Nemirovski [2004] and Monteiro and Svaiter [2010] for the extragradient method (the case of $p = 1$). For $p \geq 3$ in general, these global convergence results generalize Bullins and Lai [2022, Theorem 4.5] from saddle point and variational inequality problems to monotone inclusion problems. The local linear convergence results under an error bound are well known for the extragradient method in the literature [Tseng, 1995, Monteiro and Svaiter, 2010] but are new for the case of $p \geq 2$ to our knowledge.*

Remark 6.4.9 *The approximate tensor subroutine in Algorithm 19 has been implemented using binary search procedures efficiently specialized to the case of $p = 2$; see Monteiro and Svaiter [2012, Section 4] and Bullins and Lai [2022, Section 5]. Could we generalize this scheme to handle the more general case of $p \geq 3$, similar to what has been accomplished in convex optimization [Gasnikov et al., 2019b]? We leave the answer to future work.*

6.5 Conclusion

We propose a new closed-loop control system for capturing the acceleration phenomenon in monotone inclusion problems. In terms of theoretical guarantee, we obtain ergodic and pointwise convergence rates via appeal to simple and intuitive Lyapunov functions. Our framework based on implicit discretization of the aforementioned system gives a systematic way to derive discrete-time p^{th}-order accelerated tensor algorithms for all $p \geq 1$ and simplify existing analyses via the use of a discrete-time Lyapunov function. Key to our framework is the algebraic equation, which disappears for the case of $p = 1$, but is essential for achieving the acceleration for the case of $p \geq 2$. We also infer that a certain class of p^{th}-order tensor algorithms can achieve local linear convergence under an error bound condition.

Notably, our closed-loop control system is related to the nonlinear damping in the PDE literature where the closed-loop feedback control in fact depends on the velocity [Attouch et al., 2022a]; indeed, it is demonstrated by the algebraic equation $\lambda(t)\|\dot{x}(t)\|^{p-1} = \theta$.

There are several other avenues for future research. For example, it is interesting to study the monotone inclusion problems via appeal to the Lagrangian and Hamiltonian frameworks that have proved productive in recent work [Wibisono et al., 2016, Diakonikolas and Jordan, 2021, Muehlebach and Jordan, 2021, França et al., 2021]. Moreover, we would hope for this study to provide additional insight into the geometric or dynamical role played by the algebraic equation for shaping the continuous-time dynamics. Indeed, it is of interest to investigate the continuous-time limit of Newton methods for Bouligand-differentiable equations [Robinson, 1987], which is another generalization of complementarity and VI problems, and see whether the closed-loop control approach leads to efficient algorithms or not.

6.6 Proof of Technical Lemmas

Proof of Lemma 6.2.1. For simplicity, we denote by $A_\lambda = I - (I + \lambda A)^{-1}$ and write $\varphi(\lambda, x) = \lambda^{1/(p-1)}\|A_\lambda x\|$. Then it suffices to show

$$|\|A_\lambda x_1\| - \|A_\lambda x_2\|| \le \|x_1 - x_2\|. \tag{6.44}$$

It is known in convex analysis (see Rockafellar [1970] for example) that $\|A_\lambda x_1 - A_\lambda x_2\| \le \|x_1 - x_2\|$. This together with the triangle inequality yields Eq. (6.44).

Proof of Lemma 6.2.2. For simplicity, we define $z_1 = (I + \lambda_1 A)^{-1}x$, $z_2 = (I + \lambda_2 A)^{-1}x$, $v_1 \in Az_1$ and $v_2 \in Az_2$. In view of the definitions, we have

$$\begin{aligned} \varphi(\lambda_1, x) &= (\lambda_1)^{\frac{1}{p-1}}\|x - z_1\|, \qquad \lambda_1 v_1 + z_1 - x = 0, \\ \varphi(\lambda_2, x) &= (\lambda_2)^{\frac{1}{p-1}}\|x - z_2\|, \qquad \lambda_2 v_2 + z_2 - x = 0. \end{aligned} \tag{6.45}$$

After straightforward calculation, we have

$$\lambda_1(v_1 - v_2) + z_1 - z_2 = (\lambda_1 v_1 + z_1 - x) - (\lambda_2 v_2 + z_2 - x) + (\lambda_2 - \lambda_1)v_2 = (\lambda_2 - \lambda_1)v_2.$$

Since A is a maximal monotone operator, $v_1 \in Az_1$ and $v_2 \in Az_2$, we have $\langle v_1 - v_2, z_1 - z_2\rangle \ge 0$. Putting these pieces together yields

$$(\lambda_2 - \lambda_1)\langle v_1 - v_2, v_2\rangle = \lambda_1\|v_1 - v_2\|^2 + \langle v_1 - v_2, z_1 - z_2\rangle \ge 0.$$

This together with $\lambda_1 \le \lambda_2$ implies that $\langle v_1 - v_2, v_2\rangle \ge 0$ and thus we have $\|v_1\| \ge \|v_2\|$. Combining the last inequality with Eq. (6.45), we have

$$\varphi(\lambda_2, x) = (\lambda_2)^{\frac{1}{p-1}}\|\lambda_2 v_2\| = (\lambda_2)^{\frac{p}{p-1}}\|v_2\| \le (\lambda_2)^{\frac{p}{p-1}}\|v_1\| = \left(\tfrac{\lambda_2}{\lambda_1}\right)^{\frac{p}{p-1}}\varphi(\lambda_1, x). \tag{6.46}$$

After a short calculation, we have

$$v_1 - v_2 + \tfrac{1}{\lambda_2}(z_1 - z_2) = \tfrac{1}{\lambda_1}(\lambda_1 v_1 + z_1 - x) - \tfrac{1}{\lambda_2}(\lambda_2 v_2 + z_2 - x) + \left(\tfrac{1}{\lambda_2} - \tfrac{1}{\lambda_1}\right)(z_1 - x) = \left(\tfrac{1}{\lambda_2} - \tfrac{1}{\lambda_1}\right)(z_1 - x).$$

Since $\langle v_1 - v_2, z_1 - z_2\rangle \ge 0$, we have

$$\left(\tfrac{1}{\lambda_2} - \tfrac{1}{\lambda_1}\right)\langle z_1 - x, z_1 - z_2\rangle = \langle v_1 - v_2, z_1 - z_2\rangle + \tfrac{1}{\lambda_2}\|z_1 - z_2\|^2 \ge 0.$$

This together with $\lambda_1 \le \lambda_2$ implies that $\langle z_1 - x, z_1 - z_2\rangle \le 0$ and thus we have $\|x - z_2\| \ge \|x - z_1\|$. Combining the last inequality with Eq. (6.45), we have

$$\varphi(\lambda_2, x) = (\lambda_2)^{\frac{1}{p-1}}\|x - z_2\| \ge (\lambda_2)^{\frac{1}{p-1}}\|x - z_1\| = \left(\tfrac{\lambda_2}{\lambda_1}\right)^{\frac{1}{p-1}}\varphi(\lambda_1, x). \tag{6.47}$$

Combining Eq. (6.46) and Eq. (6.47) yields the desired inequality. The last statement of the lemma follows trivially from the maximal monotonicity of A and the definition of φ.

Proof of Lemma 6.2.5. Rearranging the first inequality in Lemma 6.2.2 implies

$$\frac{\varphi(\lambda_1,x)}{(\lambda_1)^{\frac{1}{p-1}}} \leq \frac{\varphi(\lambda_2,x)}{(\lambda_2)^{\frac{1}{p-1}}}, \quad \text{for any } x \in \mathcal{H} \text{ and } 0 < \lambda_1 \leq \lambda_2.$$

This yields that the mapping $\lambda \mapsto \|x - (I + \lambda A)^{-1}x\|$ is nondecreasing and $\alpha \mapsto \|x - (I + \alpha^{-1}A)^{-1}x\|$ is a (continuous) nonincreasing function. As a consequence, we obtain that Γ_θ is a real-valued nonnegative function. Further, by definition, if $x \in \Omega$, we have

$$\Gamma_\theta(x) \;=\; \left(\inf\{\alpha > 0 \mid \alpha^{-\frac{1}{p-1}}\|x - (I + \alpha^{-1}A)^{-1}x\| \leq \theta^{\frac{1}{p-1}}\}\right)^{\frac{1}{p-1}} = \left(\frac{1}{\Lambda_\theta(x)}\right)^{\frac{1}{p-1}}.$$

Moreover, if $x \notin \Omega$, we have $x-(I+\alpha^{-1}A)^{-1}x = 0$ for all $\alpha > 0$ which implies that $\Gamma_\theta(x) = 0$. Putting these pieces together yields the desired relationship between $\Gamma_\theta(x)$ and $\Lambda_\theta(x)$.

It remains to show that $\Gamma_\theta : \mathcal{H} \mapsto (0, +\infty)$ is Lipschitz continuous with $\theta^{-1/(p-1)} > 0$. Take $x_1, x_2 \in \mathcal{H}$ and $\bar{\alpha} > 0$, we suppose that

$$\|x_1 - (I + \bar{\alpha}^{-1}A)^{-1}x_1\| \leq \bar{\alpha}^{\frac{1}{p-1}}\theta^{\frac{1}{p-1}}.$$

It is straightforward to deduce that $I-(I+\bar{\alpha}^{-1}A)^{-1}$ is 1-Lipschitz continuous (see Rockafellar [1970] or the proof of Lemma 6.2.1) which implies

$$\|x_2 - (I + \bar{\alpha}^{-1}A)^{-1}x_2\| - \|x_1 - (I + \bar{\alpha}^{-1}A)^{-1}x_1\| \leq \|x_2 - x_1\|.$$

Therefore, we have

$$\|x_2 - (I + \bar{\alpha}^{-1}A)^{-1}x_2\| \leq \bar{\alpha}^{\frac{1}{p-1}}\theta^{\frac{1}{p-1}} + \|x_2 - x_1\| = (\bar{\alpha}^{\frac{1}{p-1}} + \|x_2 - x_1\|\theta^{-\frac{1}{p-1}})\theta^{\frac{1}{p-1}}.$$

Let $\bar{\beta} = (\bar{\alpha}^{1/(p-1)} + \|x_2 - x_1\|\theta^{-1/(p-1)})^{p-1}$. Since $\bar{\beta} \geq \bar{\alpha}$ and the mapping $\alpha \mapsto \|x - (I + \alpha^{-1}A)^{-1}x\|$ is nonincreasing, we have

$$\|x_2 - (I + \bar{\beta}^{-1}A)^{-1}x_2\| \leq \|x_2 - (I + \bar{\alpha}^{-1}A)^{-1}x_2\| \leq \bar{\beta}^{\frac{1}{p-1}}\theta^{\frac{1}{p-1}}.$$

By the definition of Γ_θ, we have $\Gamma_\theta(x_2) \leq \bar{\beta}^{1/(p-1)} = \bar{\alpha}^{1/(p-1)} + \|x_2 - x_1\|\theta^{-1/(p-1)}$ and this inequality holds true for all $\bar{\alpha} > 0$ satisfying that $\|x_1 - (I + \bar{\alpha}^{-1}A)^{-1}x_1\| \leq \bar{\alpha}^{1/(p-1)}\theta^{1/(p-1)}$; that is, all $\bar{\alpha} > 0$ satisfying that $\bar{\alpha}^{1/(p-1)} \geq \Gamma_\theta(x_1)$. Putting these pieces together, we have

$$\Gamma_\theta(x_2) \leq \Gamma_\theta(x_1) + \|x_2 - x_1\|\theta^{-\frac{1}{p-1}}.$$

Using the symmetry of x_1 and x_2, we have $\Gamma_\theta(x_1) \leq \Gamma_\theta(x_2) + \|x_1 - x_2\|\theta^{-1/(p-1)}$. Therefore, we have

$$|\Gamma_\theta(x_1) - \Gamma_\theta(x_2)| \leq \theta^{-\frac{1}{p-1}}\|x_1 - x_2\|.$$

This completes the proof.

Proof of Lemma 6.2.9. For the case of $p = 1$, we have $\lambda(t) = \theta$ is a constant function and the desired result holds true. For the case of $p \geq 2$, let $t, t' \in [0, t_0]$ and $t \neq t'$. Then, we have

$$\left|(\lambda(t'))^{\frac{1}{p-1}} - (\lambda(t))^{\frac{1}{p-1}}\right| = (\lambda(t')\lambda(t))^{\frac{1}{p-1}} \left|\left(\frac{1}{\lambda(t)}\right)^{\frac{1}{p-1}} - \left(\frac{1}{\lambda(t')}\right)^{\frac{1}{p-1}}\right|.$$

Using the definition of $\Gamma_\theta(\cdot)$ and the fact that it is Lipschitz continuous with a constant $\theta^{-1/(p-1)} > 0$ (cf. Lemma 6.2.5), we have

$$\left|\left(\frac{1}{\lambda(t)}\right)^{\frac{1}{p-1}} - \left(\frac{1}{\lambda(t')}\right)^{\frac{1}{p-1}}\right| = |\Gamma_\theta(x(t)) - \Gamma_\theta(x(t'))| \leq \frac{\|x(t)-x(t')\|}{\theta^{1/(p-1)}}.$$

Putting these pieces together yields

$$\left|\frac{(\lambda(t'))^{\frac{1}{p-1}} - (\lambda(t))^{\frac{1}{p-1}}}{t'-t}\right| \leq \left(\frac{\lambda(t')\lambda(t)}{\theta}\right)^{\frac{1}{p-1}} \frac{\|x(t)-x(t')\|}{|t-t'|}.$$

Fix $t \in [0, t_0]$ and let $t' \to t$. Then, we have

$$\limsup_{t' \to t} \left|\frac{(\lambda(t'))^{\frac{1}{p-1}} - (\lambda(t))^{\frac{1}{p-1}}}{t'-t}\right| \leq \left(\frac{(\lambda(t))^2}{\theta}\right)^{\frac{1}{p-1}} \|\dot{x}(t)\|.$$

Using Eq. (6.3) and Eq. (6.4), we have

$$\|\dot{x}(t)\| = \|(I + \lambda(t)A)^{-1}x(t) - x(t)\| = \left(\frac{\theta}{\lambda(t)}\right)^{\frac{1}{p-1}}.$$

In addition, for almost all $t \in [0, t_0]$, we have

$$\limsup_{t' \to t} \left|\frac{(\lambda(t'))^{\frac{1}{p-1}} - (\lambda(t))^{\frac{1}{p-1}}}{t'-t}\right| \geq \frac{1}{p-1}|\dot{\lambda}(t)(\lambda(t))^{\frac{1}{p-1}-1}|.$$

Putting these pieces together yields the desired result.

Proof of Lemma 6.2.10. For the case of $p = 1$, we have $\lambda(t) = \theta$ is a constant function and the desired result holds true. For the case of $p \geq 2$, since $\lambda(\cdot)$ is locally Lipschitz continuous, it suffices to show that $\dot{\lambda}(t) \geq 0$ for almost all $t \in [0, t_0)$. Indeed, let $y(t) = (I+\lambda(t)A)^{-1}x(t)$, we deduce from Eq. (6.3) that $\dot{x}(t) = y(t) - x(t)$. Then, for any fixed $t \in (0, t_0)$, we have $0 < h < \min\{t_0 - t, 1\}$ exists and the following inequality holds:

$$x(t) + h\dot{x}(t) = (1-h)x(t) + hy(t) \Longrightarrow x(t) + h\dot{x}(t) - y(t) = (1-h)(x(t) - y(t)).$$

By the definition of $y(\cdot)$, we have $\frac{1}{\lambda(t)}(x(t) - y(t)) \in Ay(t)$. Combining this with the above equality yields that $y(t) = (I + (1-h)\lambda(t)A)^{-1}(x(t) + h\dot{x}(t))$. Then, by the definition of φ,

we have

$$\begin{aligned}
&\varphi((1-h)\lambda(t), x(t)+h\dot{x}(t)) \\
&= (1-h)^{\frac{1}{p-1}}(\lambda(t))^{\frac{1}{p-1}}\|x(t)+h\dot{x}(t)-(I+(1-h)\lambda(t)A)^{-1}(x(t)+h\dot{x}(t))\| \\
&= (1-h)^{\frac{1}{p-1}}(\lambda(t))^{\frac{1}{p-1}}\|x(t)+h\dot{x}(t)-y(t)\| \\
&= (1-h)^{\frac{p}{p-1}}(\lambda(t))^{\frac{1}{p-1}}\|x(t)-y(t)\|.
\end{aligned}$$

In addition, Eq. (6.4) implies that $(\lambda(t))^{1/(p-1)}\|x(t)-y(t)\| = \theta^{1/(p-1)}$. Putting these pieces together yields

$$\varphi((1-h)\lambda(t), x(t)+h\dot{x}(t)) = (1-h)^{\frac{p}{p-1}}\theta^{\frac{1}{p-1}}. \tag{6.48}$$

Using the triangle inequality and Lemma 6.2.1, we have

$$\begin{aligned}
\varphi(\lambda(t), x(t+h)) &\leq \varphi(\lambda(t), x(t)+h\dot{x}(t)) + |\varphi(\lambda(t), x(t+h)) - \varphi(\lambda(t), x(t)+h\dot{x}(t))| \\
&\leq \varphi(\lambda(t), x(t)+h\dot{x}(t)) + (\lambda(t))^{\frac{1}{p-1}}\|x(t+h)-x(t)-h\dot{x}(t)\|. \qquad (6.49)
\end{aligned}$$

Using the second inequality in Lemma 6.2.2 and $0 < h < 1$, we have

$$\varphi(\lambda(t), x(t)+h\dot{x}(t)) \leq \left(\tfrac{1}{1-h}\right)^{\frac{p}{p-1}} \varphi((1-h)\lambda(t), x(t)+h\dot{x}(t)) \overset{\text{Eq. (6.48)}}{=} \theta^{\frac{1}{p-1}}. \tag{6.50}$$

For the ease of presentation, we define the function $\omega : (0, \min\{t_0 - t, 1\}) \mapsto (0, +\infty)$ by

$$\omega(h) = \left(\frac{\lambda(t)\|x(t+h)-x(t)-h\dot{x}(t)\|^{p-1}}{\theta}\right)^{\frac{1}{p-1}}.$$

Plugging Eq. (6.50) into Eq. (6.49) and simplifying the resulting inequality using the definition of $\omega(\cdot)$ yields

$$\varphi(\lambda(t), x(t+h)) \leq \theta^{\frac{1}{p-1}}\left(1+\omega(h)\right).$$

Using the first inequality in Lemma 6.2.2 and $\omega(h) \geq 0$ for all $h \in (0, \min\{t_0 - t, 1\})$, we have

$$\varphi\left(\tfrac{\lambda(t)}{(1+\omega(h))^{p-1}}, x(t+h)\right) \leq \left(\tfrac{1}{(1+\omega(h))^{p-1}}\right)^{\frac{1}{p-1}} \varphi(\lambda(t), x(t+h)).$$

Putting these pieces together yields

$$\varphi\left(\tfrac{\lambda(t)}{(1+\omega(h))^{p-1}}, x(t+h)\right) \leq \theta^{\frac{1}{p-1}}.$$

Since $\varphi(\cdot, x(t+h))$ is increasing and $\varphi(\lambda(t+h), x(t+h)) = \theta^{1/(p-1)}$, we have $\lambda(t+h) \geq \frac{\lambda(t)}{(1+\omega(h))^{p-1}}$. Equivalently, we have

$$\liminf_{h\to 0^+} \tfrac{\lambda(t+h)-\lambda(t)}{h} \geq -\lim_{h\to 0^+} \tfrac{\lambda(t)}{(1+\omega(h))^{p-1}} \cdot \tfrac{(1+\omega(h))^{p-1}-1}{h}.$$

By the definition of $\omega(h)$ and using the continuity of $x(\cdot)$, we have $\omega(h) \to 0$ as $h \to 0^+$. Since $p \geq 2$ is an integer, we have

$$(1+\omega(h))^{p-1} - 1 = \sum_{i=1}^{p-1} \tfrac{(p-1)!}{i!(p-1-i)!}(\omega(h))^i = \omega(h)\left(p-1+\sum_{i=1}^{p-2}\tfrac{(p-1)!}{(i+1)!(p-2-i)!}(\omega(h))^i\right).$$

Further, we have

$$\tfrac{\omega(h)}{h} = \left(\tfrac{\lambda(t)}{\theta}\right)^{\frac{1}{p-1}} \tfrac{\|x(t+h)-x(t)-h\dot{x}(t)\|}{h} \to 0, \quad \text{as } h \to 0^+.$$

Putting these pieces together yields

$$\tfrac{\lambda(t)}{(1+\omega(h))^{p-1}} \to \lambda(t), \quad \tfrac{(1+\omega(h))^{p-1}-1}{h} \to 0, \quad \text{as } h \to 0^+.$$

Therefore, we conclude that $\dot{\lambda}(t) \geq 0$ for almost all $t \in [0, t_0)$ by achieving

$$\liminf_{h\to 0^+} \tfrac{\lambda(t+h)-\lambda(t)}{h} \geq 0.$$

This completes the proof.

Proof of Lemma 6.3.4. By the definition, we have

$$\tfrac{d\mathcal{E}(t)}{dt} = \langle \dot{x}(t), x(t) - z\rangle.$$

In addition, Eq. (6.3) implies that $\dot{x}(t) = -x(t) + (I+\lambda(t)A)^{-1}x(t)$. Then, we have

$$\tfrac{d\mathcal{E}(t)}{dt} = -\|x(t)-(I+\lambda(t)A)^{-1}x(t)\|^2 - \langle x(t)-(I+\lambda(t)A)^{-1}x(t), (I+\lambda(t)A)^{-1}x(t)-z\rangle. \quad (6.51)$$

Letting $y(t) = (I+\lambda(t)A)^{-1}x(t)$, we have $\frac{1}{\lambda(t)}(x(t)-y(t)) \in Ay(t)$. Since $z \in A^{-1}(0)$, we have $0 \in Az$. By the monotonicity of A, we have

$$\tfrac{1}{\lambda(t)}\langle x(t)-y(t), y(t)-z\rangle \geq 0.$$

Using $\lambda(t) > 0$ and the definition of $y(t)$, we have

$$\langle x(t)-(I+\lambda(t)A)^{-1}x(t), (I+\lambda(t)A)^{-1}x(t)-z\rangle \geq 0. \quad (6.52)$$

Plugging Eq. (6.52) into Eq. (6.51) yields the desired inequality.

Proof of Lemma 6.4.2. It suffices to prove the first inequality in Eq. (6.35) which implies the other results. Indeed, we have

$$\begin{aligned}\mathcal{E}_k - \mathcal{E}_{k+1} &= \langle x_k - x_{k+1}, x_{k+1} - z\rangle + \tfrac{1}{2}\|x_{k+1} - x_k\|^2 \\ &= \langle x_k - x_{k+1}, y_{k+1} - z\rangle + \langle x_k - x_{k+1}, x_{k+1} - y_{k+1}\rangle + \tfrac{1}{2}\|x_{k+1} - x_k\|^2 \\ &= \underbrace{\langle x_k - x_{k+1}, y_{k+1} - z\rangle}_{\mathbf{I}} + \tfrac{1}{2}\left(\underbrace{\|x_k - y_{k+1}\|^2 - \|x_{k+1} - y_{k+1}\|^2}_{\mathbf{II}}\right).\end{aligned} \tag{6.53}$$

Using the update $x_{k+1} = x_k - \lambda_{k+1}v_{k+1}$ and letting $v \in Az$, we have

$$\mathbf{I} = \lambda_{k+1}\langle v_{k+1}, y_{k+1} - z\rangle = \lambda_{k+1}\langle v_{k+1} - v, y_{k+1} - z\rangle + \lambda_{k+1}\langle v, y_{k+1} - z\rangle.$$

Using $v_{k+1} \in A^{\epsilon_{k+1}}y_{k+1}$ and Eq. (6.33), we have $\langle v_{k+1} - v, y_{k+1} - z\rangle \geq -\epsilon_{k+1}$. This implies

$$\mathbf{I} \geq \lambda_{k+1}\langle v, y_{k+1} - z\rangle - \lambda_{k+1}\epsilon_{k+1}. \tag{6.54}$$

Since $x_{k+1} = x_k - \lambda_{k+1}v_{k+1}$ and $\|\lambda_{k+1}v_{k+1} + y_{k+1} - x_k\|^2 + 2\lambda_{k+1}\epsilon_{k+1} \leq \sigma^2\|y_{k+1} - x_k\|^2$, we have

$$\mathbf{II} = \|x_k - y_{k+1}\|^2 - \|\lambda_{k+1}v_{k+1} + y_{k+1} - x_k\|^2 \geq (1-\sigma^2)\|x_k - y_{k+1}\|^2 + 2\lambda_{k+1}\epsilon_{k+1}. \tag{6.55}$$

Plugging Eq. (6.54) and Eq. (6.55) into Eq. (6.53), we have

$$\mathcal{E}_k - \mathcal{E}_{k+1} \geq \lambda_{k+1}\langle v, y_{k+1} - z\rangle + \tfrac{1-\sigma^2}{2}\|x_k - y_{k+1}\|^2,$$

which implies the desired inequality.

Proof of Lemma 6.4.3. By the convention $0/0 = 0$, we define $\tau_k = \max\{\frac{2\epsilon_k}{\sigma^2}, \frac{\lambda_k\|v_k\|^2}{(1+\sigma)^2}\}$ for every integer $k \geq 1$. Then, we have

$$\begin{aligned}2\lambda_k\epsilon_k &\leq \sigma^2\|y_k - x_{k-1}\|^2, \\ \|\lambda_k v_k\| &\leq \|\lambda_k v_k + y_k - x_{k-1}\| + \|y_k - x_{k-1}\| \leq (1+\sigma)\|y_k - x_{k-1}\|.\end{aligned}$$

which implies that $\lambda_k\tau_k \leq \|y_k - x_{k-1}\|^2$ for every integer $k \geq 1$. This together with Lemma 6.4.2 yields

$$\frac{\inf_{z^\star \in A^{-1}(0)}\|x_0 - z^\star\|^2}{1-\sigma^2} \geq \sum_{i=1}^{k}\|y_i - x_{i-1}\|^2 \geq \left(\inf_{1\leq i\leq k}\tau_i\right)\left(\sum_{i=1}^{k}\lambda_i\right).$$

Combining this inequality with the definition of τ_k yields the desired results.

Proof of Lemma 6.4.4. For $p = 1$, the large-step condition implies that $\lambda_k \geq \theta$ for all $k \geq 0$. For $p \geq 2$, the large-step condition implies

$$\begin{aligned}\sum_{i=1}^{k}(\lambda_i)^{-\frac{2}{p-1}}\theta^{\frac{2}{p-1}} &\leq \sum_{i=1}^{k}(\lambda_i)^{-\frac{2}{p-1}}(\lambda_i\|x_{i-1}-y_i\|^{p-1})^{\frac{2}{p-1}} \\ &= \sum_{i=1}^{k}\|x_{i-1}-y_i\|^2 \overset{\text{Lemma 6.4.2}}{\leq} \frac{1}{1-\sigma^2}\left(\inf_{z^\star\in A^{-1}(0)}\|x_0-z^\star\|^2\right).\end{aligned}$$

By the Hölder inequality, we have

$$\sum_{i=1}^{k}1 = \sum_{i=1}^{k}\left(\frac{1}{(\lambda_i)^{\frac{2}{p-1}}}\right)^{\frac{p-1}{p+1}}(\lambda_i)^{\frac{2}{p+1}} \leq \left(\sum_{i=1}^{k}\frac{1}{(\lambda_i)^{\frac{2}{p-1}}}\right)^{\frac{p-1}{p+1}}\left(\sum_{i=1}^{k}\lambda_i\right)^{\frac{2}{p+1}}.$$

For the ease of presentation, we define $C = \frac{1}{(1-\sigma^2)}\theta^{-\frac{2}{p-1}}(\inf_{z^\star\in A^{-1}(0)}\|x_0-z^\star\|^2)$. Putting these pieces together yields

$$k \leq C^{\frac{p-1}{p+1}}\left(\sum_{i=1}^{k}\lambda_i\right)^{\frac{2}{p+1}},$$

which implies

$$\sum_{i=1}^{k}\lambda_i \geq \left(\frac{1}{C}\right)^{\frac{p-1}{2}}k^{\frac{p+1}{2}}.$$

This completes the proof.

Chapter 7

An Optimal Algorithm for High-Order Variational Inequality

This paper settles an open and challenging question that pertaining to the design of simple and optimal high-order methods for solving smooth and monotone variational inequalities (VIs). A VI involves finding $x^\star \in \mathcal{X}$ such that $\langle F(x), x - x^\star \rangle \geq 0$ for all $x \in \mathcal{X}$ and we consider the setting in which $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ is smooth with up to $(p-1)^{\text{th}}$-order derivatives. For $p = 2$, the cubic regularization of Newton's method has been extended to VIs with a global rate of $O(\epsilon^{-1})$ [Nesterov, 2006]. An improved rate of $O(\epsilon^{-2/3} \log\log(1/\epsilon))$ can be obtained via an alternative second-order method, but this method requires a nontrivial line-search procedure as an inner loop. Similarly, the high-order methods based on similar line-search procedures have been shown to achieve a rate of $O(\epsilon^{-2/(p+1)} \log\log(1/\epsilon))$ [Bullins and Lai, 2022, Lin and Jordan, 2023, Jiang and Mokhtari, 2022], but the inner loop requires fine-tuning of parameters and can be computationally complex. As highlighted by Nesterov, it would be desirable to develop a simple high-order VI method that retains the optimality of the more complex methods [Nesterov, 2018]. We propose a p^{th}-order method that does *not* require any search procedure and provably converges to a weak solution at a rate of $O(\epsilon^{-2/(p+1)})$. We prove that our p^{th}-order method is optimal in the monotone setting by establishing a lower bound of $\Omega(\epsilon^{-2/(p+1)})$ under a linear span assumption. Our method with restarting attains a global linear and local superlinear convergence rate for smooth and strongly monotone VIs. Further, our method achieves a global rate of $O(\epsilon^{-2/p})$ for solving smooth and nonmonotone VIs satisfying the Minty condition and our method with restarting attains a global linear and local superlinear convergence rate for smooth and nonmonotone VIs satisfying the strong Minty condition.

7.1 Introduction

Let $\mathbb{R}^d$ be a finite-dimensional Euclidean space and let $\mathcal{X} \subseteq \mathbb{R}^d$ be a closed, convex and bounded set with a diameter $D > 0$. Given that $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a continuous opera-

tor, a fundamental assumption in optimization theory, generalizing convexity, is that F is *monotone*:

$$\langle F(x) - F(x'), x - x' \rangle \geq 0, \quad \text{for all } x, x' \in \mathbb{R}^d.$$

Another useful assumption in this context is that F is $(p-1)^{\text{th}}$-order L-smooth; in particular, that it has Lipschitz-continuous $(p-1)^{\text{th}}$-order derivative ($p \geq 1$) in the sense that there exists a constant $L > 0$ such that

$$\|\nabla^{(p-1)} F(x) - \nabla^{(p-1)} F(x')\|_{\text{op}} \leq L\|x - x'\|, \quad \text{for all } x, x' \in \mathbb{R}^d. \tag{7.1}$$

With these assumptions, we can formulate the main problem of interest in this paper—the *Minty variational inequality* problem [Minty, 1962]. This consists in finding a point $x^\star \in \mathcal{X}$ such that

$$\langle F(x), x - x^\star \rangle \geq 0, \quad \text{for all } x \in \mathcal{X}. \tag{7.2}$$

The solution to Eq. (7.2) is referred to as a *weak* solution to the variational inequality (VI) corresponding to F and $\mathcal{X}$ [Facchinei and Pang, 2007]. By way of comparison, the *Stampacchia variational inequality* problem [Hartman and Stampacchia, 1966] consists in finding a point $x^\star \in \mathcal{X}$ such that

$$\langle F(x^\star), x - x^\star \rangle \geq 0, \quad \text{for all } x \in \mathcal{X}, \tag{7.3}$$

and the solution to Eq. (7.3) is called a *strong* solution to the VI corresponding to F and $\mathcal{X}$. In the setting where F is continuous and monotone, the solution sets of Eq. (7.2) and Eq. (7.3) are equivalent. However, these two solution sets are different in general and a weak solution need not exist when a strong solution exists. In addition, computing an approximate strong solution involves a higher computational burden than finding an approximate weak solution [Monteiro and Svaiter, 2010, 2011, Chen et al., 2017]. Earlier work has focused on the asymptotic global convergence analysis of various VI methods under mild conditions [Lemke and Howson, 1964, Scarf, 1967, Todd, 2013, Hammond and Magnanti, 1987, Fukushima, 1992, Magnanti and Perakis, 1997b]. Two notable exceptions are the generalizations of the ellipsoid method [Magnanti and Perakis, 1995] and the interior-point method [Ralph and Wright, 1997], both of which have been the subject of nonasymptotic complexity analysis.

VIs capture a wide range of problems in optimization theory and beyond, including saddle-point problems and models of equilibria in game-theoretic settings [Cottle et al., 1980, Kinderlehrer and Stampacchia, 2000, Trémolières et al., 2011]. Moreover, the challenge of designing solution methods for VIs with provable worst-case bounds has driven significant research over several decades; see Harker and Pang [1990] and Facchinei and Pang [2007]. This research has provided a foundation for work in machine learning in recent years, where general saddle-point problems have emerged in many settings, including generative adversarial networks (GANs) [Goodfellow et al., 2014] and multi-agent learning in games [Cesa-Bianchi and Lugosi, 2006, Mertikopoulos and Zhou, 2019]. Some of these applications in ML induce a nonmonotone structure, with representative examples including the training of robust neural networks [Madry et al., 2018] or robust classifiers [Sinha et al., 2018].

Building on seminal work in the context of high-order optimization [Baes, 2009, Birgin et al., 2017], we tackle the challenge of developing p^{th}-order methods for VIs via an inexact solution of regularized subproblems obtained from a $(p-1)^{\text{th}}$-order Taylor expansion of F. Accordingly, we make the following assumptions throughout this paper.

A1. $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ is $(p-1)^{\text{th}}$-order L-smooth.

A2. The subproblem based on a $(p-1)^{\text{th}}$-order Taylor expansion of F and a convex and bounded set $\mathcal{X}$ can be computed approximately in an efficient manner.

For the first-order VI methods ($p = 1$), Nemirovski [2004] has proved that the extragradient (EG) method [Korpelevich, 1976, Antipin, 1978] converges to a weak solution with a global rate of $O(\epsilon^{-1})$ if F is monotone and Eq. (7.1) holds. There are other methods with the same global rate guarantee, including forward-backward splitting method [Tseng, 2000], optimistic gradient (OG) [Popov, 1980, Mokhtari et al., 2020a, Kotsalis et al., 2022] and dual extrapolation [Nesterov, 2007]. All these methods match the lower bound of Ouyang and Xu [2021] and are thus optimal. In addition, a general adaptive line search framework has been proposed to unify and extend several convergence results from the VI literature [Magnanti and Perakis, 2004].

The investigation of second-order and high-order ($p \geq 2$) counterparts of these first-order methods is less advanced, as exploiting high-order derivative information is much more involved for VIs [Nesterov, 2006, Monteiro and Svaiter, 2012]. Aiming to fill this gap, some work has been recently devoted to studying high-order extensions of first-order VI methods [Bullins and Lai, 2022, Lin and Jordan, 2023, Jiang and Mokhtari, 2022]. These extensions attain a rate of $O(\epsilon^{-2/(p+1)} \log\log(1/\epsilon))$ but require the nontrivial line-search procedures at each iteration. Although an additional $\log\log(1/\epsilon)$ factor is normally regarded as modest, the associated line-search procedures require fine tuning of parameters and can be prohibitive from a computational viewpoint. Thus, the problem of designing a **simple** and **optimal** high-order method remains open. In particular, Nesterov [2018, page 305] noted the difficulty of removing the line-search procedure without sacrificing the rate of convergence and highlighted this as an open and challenging question. We summarize the problem as follows:

Can we design a simple and optimal p^{th}-order VI method without line search?

In this paper, we present an affirmative answer to this problem by identifying a p^{th}-order method that achieves a global rate of $O(\epsilon^{-2/(p+1)})$ while dispensing entirely with the line-search inner loop. The core idea of the proposed method is to incorporate a simple adaptive strategy into a high-order generalization of the dual extrapolation method.

There are two reasons why we choose the dual extrapolation method as a base algorithm for our high-order methods. First, the dual extrapolation method has its own merits as summarized in Nesterov [2007], and the first second-order VI method to attain a global convergence rate of $O(\epsilon^{-1})$ [Nesterov, 2006] was developed based on a dual extrapolation step.

Our method can be interpreted as a simplification and generalization of this method. Second, the dual extrapolation step is an important ingredient for algorithm design in optimization, given the close relationship between extrapolation and acceleration in the context of first-order methods for smooth convex optimization [Lan and Zhou, 2018a,b]. This is in contrast to the EG method, which is an approximate proximal point method [Mokhtari et al., 2020a]. It would deepen our understanding of the scope of dual extrapolation if we could design a simple and optimal high-order VI method based on this scheme.

Contributions. The contribution of this paper consists in fully closing the gap between the upper and lower bounds in the monotone setting and improving the state-of-the-art upper bounds in the strongly monotone and/or structured non-monotone settings. In further detail:

1. We present a new p^{th}-order method for solving smooth and monotone VIs where F has a Lipschitz continuous $(p-1)^{\text{th}}$-order derivative and $\mathcal{X}$ is convex and bounded. We prove that the number of calls of subproblem solvers required by our method to find an ϵ-weak solution is bounded by

$$O\left(\left(\frac{LD^{p+1}}{\epsilon}\right)^{\frac{2}{p+1}}\right).$$

   We prove that our method is optimal by establishing a matching lower bound under a linear span assumption. Moreover, we present a restarted version of our method for solving smooth and strongly monotone VIs. That is, we show that there exists a constant $\mu > 0$ such that

$$\langle F(x) - F(x'), x - x' \rangle \geq \mu \|x - x'\|^2, \quad \text{for all } x, x' \in \mathbb{R}^d.$$

   We show that the number of calls of subproblem solvers required to find $\hat{x} \in \mathcal{X}$ satisfying $\|\hat{x} - x^\star\| \leq \epsilon$ is bounded by

$$O\left((\kappa D^{p-1})^{\frac{2}{p+1}} \log_2\left(\frac{D}{\epsilon}\right)\right),$$

   where $\kappa = L/\mu$ refers to the condition number of F. The restarted version also achieves local superlinear convergence for the case of $p \geq 2$.

2. We show how to modify our framework such that it can be used for solving smooth and nonmonotone VIs satisfying the so-called Minty condition (see Definition 7.2.5). Again, we note that a line-search procedure is not required. We prove that the number of calls of subproblem solvers to find an ϵ-strong solution is bounded by

$$O\left(\left(\frac{LD^{p+1}}{\epsilon}\right)^{\frac{2}{p}}\right).$$

   Our methods with restarting attain a global linear and local superlinear convergence rate (for the case of $p \geq 2$) under the strong Minty condition.

Concurrently appearing on arXiv, Adil et al. [2022] has established the same upper bounds as ours for a high-order generalization of the EG method for solving smooth and monotone VIs. However, it still remains open whether or not their method can be extended to solve strongly monotone VIs[1] or nonmonotone VIs satisfying the Minty condition.

A lower bound has been established in Adil et al. [2022] for *a class of p^{th}-order methods restricted to solving the primal problem*. This is a rather strong limitation that excludes both our method and their method. We derive the same lower bound for a broader class of p^{th}-order methods that include both our method and their method thanks to the construction of a new hard instance. Although the hard instance function is different (and the lower bound does improve), we do wish to acknowledge that the proof techniques from Adil et al. [2022] inspired our analysis.

Related works. In addition to the aforementioned work, we review relevant research on high-order convex optimization. We focus on p^{th}-order methods for $p \geq 2$.

To the best of our knowledge, the systematic investigation of the global convergence rate of second-order methods originates in work on the cubic regularization of Newton's method (CRN) [Nesterov and Polyak, 2006] and its accelerated counterpart (ACRN) [Nesterov, 2008]. The ACRN method was then extended with a p^{th}-order regularization model, yielding an improved global rate of $O(\epsilon^{-1/(p+1)})$ [Baes, 2009] while an adaptive p^{th}-order method was proposed in Jiang et al. [2020] with the same rate. This extension was recently revisited by Nesterov [2021b], Grapiglia and Nesterov [2022a] with a discussion on an efficient implementation of a third-order method. Meanwhile, within the accelerated Newton proximal extragradient (ANPE) framework [Monteiro and Svaiter, 2013], a p^{th}-order method was also proposed by Gasnikov et al. [2019b] with a global rate of $O(\epsilon^{-2/(3p+1)} \log(1/\epsilon))$ for minimizing a convex function whose p^{th}-order derivative is Lipschitz continuous. In this context, an additional log factor remains between the above upper bound and the lower bound of $O(\epsilon^{-2/(3p+1)})$ [Arjevani et al., 2019]. This gap was recently closed by two works [Kovalev and Gasnikov, 2022, Carmon et al., 2022] that offer a complementary viewpoint to that of Monteiro and Svaiter [2013], Gasnikov et al. [2019b] on how to remove the line-search procedure. Subsequently, the p^{th}-order ANPE framework was extended to a strongly convex setting [Marques Alves, 2022] and shown to achieve an optimal global linear rate. Beyond the setting with Lipschitz continuous p^{th}-order derivatives, these p^{th}-order methods have been adapted to a setting with Hölder continuous p^{th}-order derivatives [Grapiglia and Nesterov, 2017, 2019, 2020, Song et al., 2021, Doikov and Nesterov, 2022]. Further settings include smooth nonconvex minimization [Cartis et al., 2010, 2011a,b, 2019, Birgin et al., 2016, 2017, Martínez, 2017] and structured nonsmooth minimization [Bullins, 2020]. There is also a complementary line of research that studies the favorable properties of lower-order methods in the setting of higher-order smoothness [Nesterov, 2021d,a,c].

[1] We are also aware of very recent work [Huang and Zhang, 2022a] that analyzes a high-order extragradient method for solving smooth and strongly monotone VIs. They have established the same convergence rate guarantee as our method in this setting.

We are aware of various high-order methods obtained via discretization of continuous-time dynamical systems [Wibisono et al., 2016, Lin and Jordan, 2022b]. In particular, Wibisono et al. [2016] showed that the ACRN method and its p^{th}-order variants can be obtained from implicit discretization of an open-loop system without Hessian-driven damping. Lin and Jordan [2022b] have provided a control-theoretic perspective on p^{th}-order ANPE methods by recovering them from implicit discretization of a closed-loop system with Hessian-driven damping. Both of these two works proved the convergence rate of p^{th}-order ACRN and ANPE methods using Lyapunov functions.

Notation. We use lower-case letters such as x to denote vectors and upper-case letters such as X to denote tensors. Let $\mathbb{R}^d$ be a finite-dimensional Euclidean space (the dimension is $d \in \{1, 2, \ldots\}$), endowed with the scalar product $\langle \cdot, \cdot \rangle$. For $x \in \mathbb{R}^d$, we let $\|x\|$ denote its ℓ_2-norm. For $X \in \mathbb{R}^{d_1 \times \cdots \times d_p}$, we define

$$X[z^1, \cdots, z^p] = \sum_{1 \le i_j \le d_j, 1 \le j \le p} (X_{i_1, \cdots, i_p}) z^1_{i_1} \cdots z^p_{i_p},$$

and $\|X\|_{\text{op}} = \max_{\|z^i\|=1, 1 \le j \le p} X[z^1, \cdots, z^p]$ as well. Fixing $p \ge 0$ and letting $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ be a continuous and high-order differentiable operator, we define $\nabla^{(p)} F(x)$ as the p^{th}-order derivative at a point $x \in \mathbb{R}^d$ and write $\nabla^{(0)} F = F$. To be more precise, letting $z_1, \ldots, z_k \in \mathbb{R}^d$, we have

$$\nabla^{(k)} F(x)[z^1, \cdots, z^k] = \sum_{1 \le i_1, \ldots, i_k \le d} \left(\frac{\partial F_{i_1}}{\partial x_{i_2} \cdots \partial x_{i_k}}(x) \right) z^1_{i_1} \cdots z^k_{i_k}.$$

For a closed and convex set $\mathcal{X} \subseteq \mathbb{R}^d$, we let $\mathcal{P}_{\mathcal{X}}$ be the orthogonal projection onto $\mathcal{X}$ and let $\text{dist}(x, \mathcal{X}) = \inf_{x' \in \mathcal{X}} \|x' - x\|$ denote the distance between x and $\mathcal{X}$. Finally, $a = O(b(L, \mu, \epsilon))$ stands for an upper bound $a \le C \cdot b(L, \mu, \epsilon)$, where $C > 0$ is independent of parameters L, μ and the tolerance $\epsilon \in (0, 1)$, and $a = \tilde{O}(b(L, \mu, \epsilon))$ indicates the same inequality where $C > 0$ depends on logarithmic factors of $1/\epsilon$.

7.2 Preliminaries

We present the basic formulation of variational inequality (VI) problems and provide definitions for the class of operators and optimality criteria considered in this paper. We further give an overview of Nesterov's dual extrapolation concept from which our method originates.

Variational inequality problem. The regularity conditions that we consider for $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ are as follows.

Definition 7.2.1 *F is k^{th}-order L-smooth if*

$$\|\nabla^{(k)} F(x) - \nabla^{(k)} F(x')\|_{\text{op}} \le L \|x - x'\|,$$

for all x, x'.

Definition 7.2.2 *F is* μ-strongly monotone *if* $\langle F(x) - F(x'), x - x' \rangle \geq \mu \|x - x'\|^2$ *for all* x, x'. *If* $\mu = 0$, *we recover the definition of monotone operator.*

With the definitions in mind, we state the assumptions that impose in addition to **A1** and **A2** in order to define highly smooth VI problems.

Assumption 7.2.3 *We assume that (i)* $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ *is* $(p-1)^{\text{th}}$*-order L-smooth, and (ii)* $\mathcal{X}$ *is convex and bounded with a diameter* $D = \max_{x,x' \in \mathcal{X}} \|x - x'\| > 0$.

The convergence of derivative-based optimization methods to a weak solution $x^\star \in \mathcal{X}$ depends on properties of F near this point, and in particular some form of smoothness condition is needed. As for the boundedness condition for $\mathcal{X}$, it is standard in the VI literature [Facchinei and Pang, 2007]. This condition not only guarantees the validity of the most natural optimality criterion in the monotone setting—the gap function [Nemirovski, 2004, Nesterov, 2007]—but additionally it is satisfied in real application problems Facchinei and Pang [2007]. On the other hand, there is another line of work focusing on relaxing the boundedness condition via appeal to other notions of approximate solution [Monteiro and Svaiter, 2010, 2011, 2012, Chen et al., 2017]. For simplicity, we retain the boundedness condition and leave the analysis for cases with unbounded constraint sets to future work.

Monotone setting. For some of our results we focus on operators F that are monotone in addition to Assumption 7.2.3. Under monotonicity, it is well known that any ϵ-strong solution is an ϵ-weak solution but the reverse does not hold true in general. Accordingly, we formally define $\hat{x} \in \mathcal{X}$ as an ϵ-weak solution or an ϵ-strong solution as follows:

$$\begin{aligned} &\textbf{(ϵ-weak solution)} \quad \langle F(x), \hat{x} - x \rangle \leq \epsilon, \quad \text{for all } x \in \mathcal{X}, \\ &\textbf{(ϵ-strong solution)} \quad \langle F(\hat{x}), \hat{x} - x \rangle \leq \epsilon, \quad \text{for all } x \in \mathcal{X}. \end{aligned}$$

These definitions motivate the use of a gap function, $\textsc{gap}(\cdot) : \mathcal{X} \mapsto \mathbb{R}_+$, defined by

$$\textsc{gap}(\hat{x}) = \sup_{x \in \mathcal{X}} \langle F(x), \hat{x} - x \rangle, \tag{7.4}$$

to measure the optimality of a point $\hat{x} \in \mathcal{X}$ that is output by various iterative solution methods; see, e.g., Tseng [2000], Nemirovski [2004], Nesterov [2007], Mokhtari et al. [2020a]. Note that the boundedness of $\mathcal{X}$ and the existence of a strong solution guarantee that the gap function is well defined. Formally, we have

Definition 7.2.4 *A point* $\hat{\mathbf{x}} \in \mathcal{X}$ *is an* ϵ-weak solution *to the monotone VI that corresponds to* $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ *and* $\mathcal{X} \subseteq \mathbb{R}^d$ *if we have* $\textsc{gap}(\hat{x}) \leq \epsilon$. *If* $\epsilon = 0$, *then* $\hat{x} \in \mathcal{X}$ *is a* weak solution.

In the strongly monotone setting, we let $\mu > 0$ denote the modulus of strong monotonicity for F. Under Assumption 7.2.3, we define $\kappa := L/\mu$ as the *condition number* of F. It is worth mentioning that the condition number quantifies the difficulty of solving the optimization

problem [Nesterov, 2018] and appears in the iteration complexity bound of derivative-based methods for optimizing a smooth and strongly convex function. Accordingly, the VI that corresponds to F and $\mathcal{X}$ is more computationally challenging as $\kappa > 0$ increases.

Structured nonmonotone setting. We study the case where F is nonmonotone but satisfies the Minty condition. Imposing such a condition is crucial since the smoothness of F is not sufficient to guarantee that the problem is computationally tractable. This has been shown by Daskalakis et al. [2021] who established that even deciding whether an approximate min-max solution exists is NP hard in smooth and nonconvex-nonconcave min-max optimization (which is a special instance of nonmonotone VIs).

Recent work has shown that the nonmonotone VI problem satisfying the Minty condition is computationally tractable [Solodov and Svaiter, 1999b, Dang and Lan, 2015, Iusem et al., 2017, Kannan and Shanbhag, 2019, Song et al., 2020, Liu et al., 2021, Diakonikolas et al., 2021]. We thus make the following formal definition.

Definition 7.2.5 *The VI corresponding to $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ and $\mathcal{X} \subseteq \mathbb{R}^d$ satisfies the* Minty condition *if there exists a point $x^\star \in \mathcal{X}$ such that $\langle F(x), x - x^\star \rangle \geq 0$ for all $x \in \mathcal{X}$.*

We make some comments on the Minty condition. First, this condition simply assumes the existence of at least one weak solution. Second, Harker and Pang [1990, Theorem 3.1] guarantees that there is at least one strong solution since F is continuous and $\mathcal{X}$ is closed and bounded. However, the set of weak solutions is only a subset of the set of strong solutions if F is not necessarily monotone, and the weak solution might not exist. From this perspective, the Minty condition gives a favorable structure. Furthermore, the Minty condition is weaker than generalized monotone assumptions [Dang and Lan, 2015, Iusem et al., 2017, Kannan and Shanbhag, 2019] that imply that the computation of an ϵ-strong solution of nonmonotone VIs is tractable for first-order methods. Finally, we say the VI satisfies the *μ_M-strong Minty condition* [Song et al., 2020] if there exists a point $x^\star \in \mathcal{X}$ such that $\langle F(x), x - x^\star \rangle \geq \mu_M \|x - x^\star\|^2$ for all $x \in \mathcal{X}$.

Accordingly, we define the *residue function* $\textsc{res}(\cdot) : \mathcal{X} \mapsto \mathbb{R}_+$ given by

$$\textsc{res}(\hat{x}) = \sup_{x \in \mathcal{X}} \langle F(\hat{x}), \hat{x} - x \rangle, \tag{7.5}$$

which measures the optimality of a point $\hat{x} \in \mathcal{X}$ achieved by iterative solution methods; see, e.g., Dang and Lan [2015], Iusem et al. [2017], Kannan and Shanbhag [2019], Song et al. [2020]. It is worth noting that the boundedness of $\mathcal{X}$ and the continuity of F guarantee that the residue function is well defined. Formally, we have

Definition 7.2.6 *A point $\hat{\mathbf{x}} \in \mathcal{X}$ is an* ϵ-strong solution *to the nonmonotone VI corresponding to $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ and $\mathcal{X} \subseteq \mathbb{R}^d$ if we have $\textsc{res}(\hat{x}) \leq \epsilon$. If $\epsilon = 0$, then $\hat{x} \in \mathcal{X}$ is a* strong solution.

There are many application problems that can be formulated as nonmonotone VIs satisfying the Minty condition, such as competitive exchange economies [Brighi and John, 2002] and product pricing [Choi et al., 1990, Gallego and Hu, 2014, Ewerhart, 2014]. Also, the Minty condition restricted to nonconvex optimization was adopted for analyzing the convergence of stochastic gradient descent for deep learning [Li and Yuan, 2017] and it has found real-world applications [Kleinberg et al., 2018].

Comments on weak versus strong solutions. First, the montonicity assumption is assumed such that the *averaged iterates* make sense and we have proved that the averaged iterates converge to an ϵ-weak solution with a faster convergence rate of $O(\epsilon^{-2/(p+1)})$ in this setting (see Theorem 7.3.1). Such a bound is stronger than that for convergence rate of *best iterates* to an ϵ-strong solution under only the Minty condition (see Theorem 7.3.7). Further, if we impose the monotonicity assumption, we conjecture that the rate of convergence to an ϵ-strong solution can be improved from $O(\epsilon^{-2/p})$ to $O(\epsilon^{-2/(p+1)})$. Such a result has been achieved for the case of $p = 1$ [Diakonikolas, 2020]. However, it is worth mentioning that the first-order method in [Diakonikolas, 2020] is different from the first-order extragradient method and the first-order dual extrapolation method which are known to achieve an optimal convergence to an ϵ-weak solution. It remains unclear how to design a high-order generalization of such new Halpern iteration methods. Finally, the complexity bound of $O(\epsilon^{-2/(p+1)})$ can not be extended beyond the monotone setting if only the Minty condition holds. Indeed, the key ingredient for proving the complexity bound of $O(\epsilon^{-2/(p+1)})$ is the use of averaged iterates in our new method. Such an averaging technique is known to be crucial for the monotone setting [Magnanti and Perakis, 1997a] but is not known to be valid when only the Minty condition holds. In addition, the fast convergence of Halpern iteration in Diakonikolas [2020] for achieving an ϵ-strong solution heavily relies on the monotonicity assumption and does not extend to the setting when only the Minty condition holds. We would be very surprised if the optimal complexity bound for the monotone setting (note that we have established the matching lower bound) can be achieved for the setting when only the Minty condition holds. Even for the case of $p = 1$, we are not aware of any relevant supporting evidence. Further exploration of this topic is beyond the scope of our paper.

Comments on Euclidean versus non-Euclidean settings. The non-Euclidean generalization of the first-order dual extrapolation method has been shown to outperform the original method in various application problem (e.g., the case where $\mathcal{X}$ is a simplex) [Nesterov, 2007]. It remains a possibility that such a benefit also occurs for the case of $p \geq 2$ and thus it seems promising to study the high-order dual extrapolation method in non-Euclidean settings. In fact, we can follow the approach from Adil et al. [2022] and extend our methods to the non-Euclidean setting using Bregman divergence. However, we can not say much about the superiority of high-order dual extrapolation methods in the non-Euclidean setting since the solution of the subproblem will become much more involved. This is different from

the first-order case where each subproblem has a closed-form solution even in non-Euclidean settings. This is also a intriguing topic but again beyond the scope of our paper.

Nesterov's dual extrapolation method. Dual extrapolation method [Nesterov, 2007] has been shown to be an optimal first-order method for computing the weak solution of the VI when F is zeroth-order L-smooth and monotone [Ouyang and Xu, 2021]. We recall the basic formulation in our setting of a VI defined via an operator $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ and a closed, convex and bounded set $\mathcal{X} \subseteq \mathbb{R}^d$. Starting with the initial points $x_0 \in \mathcal{X}$ and $s_0 = 0 \in \mathbb{R}^d$, the k^{th} iteration of the scheme is given by ($k \geq 1$):

$$\begin{aligned}
&\text{Find } v_k \in \mathcal{X} \text{ s.t. } v_k = \text{argmax}_{v\in\mathcal{X}} \langle s_{k-1}, v - x_0\rangle - \tfrac{\beta}{2}\|v - x_0\|^2,\\
&\text{Find } x_k \in \mathcal{X} \text{ s.t. } \langle F(v_k) + \beta(x_k - v_k), x - x_k\rangle \geq 0 \text{ for all } x \in \mathcal{X},\\
&s_k = s_{k-1} - \lambda F(x_k).
\end{aligned}$$

This method can be viewed as an instance of the celebrated extragradient method in the dual space (we refer to $s \in \mathbb{R}^d$ as the dual variable). Indeed, the rule which transforms a point s_{k-1} into the next point s_k at the k^{th} iteration is called a dual extrapolation step. Nesterov [2007, Theorem 2] showed that the dual extrapolation method, with $\beta = L$ and $\lambda = 1$, generates a sequence $\{x_k\}_{k\geq 0} \subseteq \mathcal{X}$ satisfying the condition that the average iterate, $\tilde{x}_k = \frac{1}{k+1}\sum_{i=0}^{k} x_i$, is an ϵ-weak solution after $O(\epsilon^{-1})$ iterations. Here, $L > 0$ is the Lipschitz constant of F.

Nesterov also considered the setting where F is monotone and first-order L-smooth and proposed a second-order variant of dual extrapolation method for computing the weak solution of the VI [Nesterov, 2006]. Starting with the initial points $x_0 \in \mathcal{X}$ and $s_0 = 0 \in \mathbb{R}^d$, the k^{th} iteration of the scheme is given by ($k \geq 1$):

$$\begin{aligned}
&\text{Find } v_k \in \mathcal{X} \text{ s.t. } v_k = \text{argmax}_{v\in\mathcal{X}} \langle s_{k-1}, v - x_0\rangle - \tfrac{\beta}{3}\|v - x_0\|^3,\\
&\text{Find } x_k \in \mathcal{X} \text{ s.t. } \langle F^1_{v_k}(x_k) + \tfrac{M}{2}\|x_k - v_k\|(x_k - v_k), x - x_k\rangle \geq 0 \text{ for all } x \in \mathcal{X},\\
&s_k = s_{k-1} - \lambda F(x_k),
\end{aligned}$$

where $F^1_v(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^d$ is defined as a first-order Taylor expansion of F at a point $v \in \mathcal{X}$:

$$F^1_v(x) = F(v) + \nabla F(v)(x - v).$$

This scheme is based on the dual extrapolation step but with a different regularization and with a first-order Taylor expansion of F. This makes sense since we have zeroth-order and first-order derivative information available and hope to use both of them to accelerate convergence. Similar ideas have been studied for convex optimization [Nesterov and Polyak, 2006], leading to a simple second-order method with a faster global rate of convergence [Nesterov, 2008] than the optimal first-order method [Nesterov, 1983]. Unfortunately, this second-order dual extrapolation method with $\beta = 6L$, $M = 5L$ and $\lambda = 1$ is only guaranteed to achieve an iteration complexity of $O(\epsilon^{-1})$ [Nesterov, 2006, Theorem 4].

Algorithm 20 Perseus(p, x_0, L, T, opt)

Input: order p, initial point $x_0 \in \mathcal{X}$, parameter L, iteration number T and opt $\in \{0, 1, 2\}$.

Initialization: set $s_0 = 0_d \in \mathbb{R}^d$.
for $k = 0, 1, 2, \ldots, T$ **do**
STEP 1: If $x_k \in \mathcal{X}$ is a solution of the VI, then **stop**.
STEP 2: Compute $v_{k+1} = \text{argmax}_{v \in \mathcal{X}} \{\langle s_k, v - x_0 \rangle - \frac{1}{2}\|v - x_0\|^2\}$.
STEP 3: Compute $x_{k+1} \in \mathcal{X}$ such that Eq. (7.7) holds true.
STEP 4: Compute $\lambda_{k+1} > 0$ such that $\frac{1}{20p-8} \leq \frac{\lambda_{k+1} L \|x_{k+1} - v_{k+1}\|^{p-1}}{p!} \leq \frac{1}{10p+2}$.
STEP 5: Compute $s_{k+1} = s_k - \lambda_{k+1} F(x_{k+1})$.

Output: $\hat{x} = \begin{cases} \tilde{x}_T = \frac{1}{\sum_{k=1}^T \lambda_k} \sum_{k=1}^T \lambda_k x_k, & \text{if opt} = 0, \\ x_T, & \text{else if opt} = 1, \\ x_{k_T} \text{ for } k_T = \text{argmin}_{1 \leq k \leq T} \|x_k - v_k\|, & \text{else if opt} = 2. \end{cases}$

7.3 A Regularized High-Order Model and Algorithm

We present our algorithmic derivation of Perseus and provide a theoretical convergence guarantee for the method. We provide intuition into why Perseus and its restarted version yield fast rates of convergence for VI problems. We present a full treatment of the global and local convergence of Perseus and its restarted version for both the monotone setting and the nonmonotone setting under the Minty condition.

Algorithmic scheme. We present our p^{th}-order method—Perseus(p, x_0, L, T, opt)—in Algorithm 20. Here $p \in \{1, 2, \ldots\}$ is the order, $x_0 \in \mathcal{X}$ is an initial point, $L > 0$ is a Lipschitz constant for $(p-1)^{\text{th}}$-order smoothness, T is the maximum iteration number and opt $\in \{0, 1, 2\}$ is the type of output. Our method is a generalization of the dual extrapolation method [Nesterov, 2007] from first order to general p^{th} order.

The novelty of our method lies in an adaptive strategy used for updating λ_{k+1} (see **Step 4**). This modification is simple yet important. It is the key for obtaining a global rate of $O(\epsilon^{-2/(p+1)})$ (monotone) and that of $O(\epsilon^{-2/p})$ (nonmonotone with the Minty condition) under Assumption 7.2.3. Focusing on the case of $p = 2$ and the monotone setting, our results improve on the best existing global convergence rates of $O(\epsilon^{-1})$ [Nesterov, 2006] and that of $O(\epsilon^{-2/3} \log\log(1/\epsilon))$ [Monteiro and Svaiter, 2012] under Assumption 7.2.3, while not sacrificing algorithmic simplicity. In addition, our methods allow the subproblem to be solved inexactly, and we give options for choosing the type of outputs under different assumptions.

Comments on inexact solution of subproblems. We remark that **Step 3** involves computing an approximate strong solution to the VI where we define the operator $F_{v_{k+1}}(x)$

as the sum of a high-order polynominal and a regularization term. Indeed, we have[2]

$$\begin{aligned} F_{v_{k+1}}(x) = F(v_{k+1}) + \langle \nabla F(v_{k+1}), x - v_{k+1} \rangle \\ + \ldots + \tfrac{1}{(p-1)!} \nabla^{(p-1)} F(v_{k+1})[x - v_{k+1}]^{p-1} + \tfrac{5L}{(p-1)!} \|x - v_{k+1}\|^{p-1}(x - v_{k+1}), \end{aligned}$$

where we write the VI of interest in the subproblem as follows:

$$\text{Find } x_{k+1} \in \mathcal{X} \text{ such that } \langle F_{v_{k+1}}(x_{k+1}), x - x_{k+1} \rangle \geq 0 \text{ for all } x \in \mathcal{X}. \tag{7.6}$$

Since $F_{v_{k+1}}$ is continuous and $\mathcal{X}$ is convex and bounded, Harker and Pang [1990, Theorem 3.1] guarantees that a strong solution to the VI in Eq. (7.6) exists and the problem of finding an approximate strong solution is well defined.

In the monotone setting, we can prove that the p^{th}-order regularization subproblem in Eq. (7.6) is monotone (in fact, it is relatively strongly monotone) if the original VI is p^{th}-order L-smooth and monotone. Indeed, the VI with F is monotone if and only if the symmetric part of the Jacobian matrix $\nabla F(x)$ is *positive semidefinite* for all $x \in \mathbb{R}^d$ [Rockafellar and Wets, 2009, Proposition 12.3]. That is to say,

$$\tfrac{1}{2}(\nabla F(x) + \nabla F(x)^\top) \succeq 0_{d \times d}, \quad \text{for all } x \in \mathbb{R}^d.$$

For the case of $p = 1$, we have $\nabla F_{v_{k+1}}(x) = 5L \cdot I_{d \times d} \succeq 0_{d \times d}$ for all $x \in \mathbb{R}^d$ where $I_{d \times d} \in \mathbb{R}^{d \times d}$ is an identity matrix. Thus, the VI in Eq. (7.6) is $5L$-strongly monotone. For the case of $p \geq 2$, we have

$$\begin{aligned} \nabla F_{v_{k+1}}(x) = \nabla F(v_{k+1}) + \ldots + \tfrac{1}{(p-2)!} \nabla^{(p-1)} F(v_{k+1})[x - v_{k+1}]^{p-2} \\ + \tfrac{5L}{(p-1)!} \|x - v_{k+1}\|^{p-1} I_{d \times d} + \tfrac{5L}{(p-2)!} \|x - v_{k+1}\|^{p-2}(x - v_{k+1})(x - v_{k+1})^\top. \end{aligned}$$

Since the original VI is p^{th}-order L-smooth, we obtain from Jiang and Mokhtari [2022, Eq. (7)] that

$$\begin{aligned} & \|\nabla F(x) - (\nabla F(v_{k+1}) + \ldots + \tfrac{1}{(p-2)!} \nabla^{(p-1)} F(v_{k+1})[x - v_{k+1}]^{p-2})\|_{\text{op}} \\ \leq \; & \tfrac{L}{(p-1)!} \|x - v_{k+1}\|^{p-1}. \end{aligned}$$

This implies that

$$\begin{aligned} \tfrac{1}{2}(\nabla F_{v_{k+1}}(x) + \nabla F_{v_{k+1}}(x)^\top) &\succeq \tfrac{1}{2}(\nabla F(x) + \nabla F(x)^\top) \\ &\quad + \tfrac{4L}{(p-1)!} \|x - v_{k+1}\|^{p-1} I_{d \times d} + \tfrac{5L}{(p-2)!} \|x - v_{k+1}\|^{p-2}(x - v_{k+1})(x - v_{k+1})^\top \\ &\succeq \tfrac{4L}{(p-1)!}(\|x - v_{k+1}\|^{p-1} I_{d \times d} + \|x - v_{k+1}\|^{p-2}(x - v_{k+1})(x - v_{k+1})^\top), \end{aligned}$$

where the second inequality holds since the original VI is monotone. Thus, the VI in Eq. (7.6) is monotone and $4L$-relatively strongly monotone with respect to the reference function

[2]For ease of presentation, we choose the factor of 5 here. It is worth noting that other large coefficients also suffice to achieve the same global convergence rate guarantee.

Algorithm 21 Perseus-restart(p, x_0, L, σ, D, T, opt)

Input: order p, initial point $x_0 \in \mathcal{X}$, parameters L, σ, D, iteration number T and opt $\in \{0, 1\}$.

Initialization: set $T_{\text{inner}} = \begin{cases} \lceil (\frac{2^{p+1}(5p-2)}{p!} \frac{LD^{p-1}}{\sigma})^{\frac{2}{p+1}} \rceil, & \text{if opt} = 0, \\ 1, & \text{else if opt} = 1. \end{cases}$

for $k = 0, 1, 2, \ldots, T$ **do**

STEP 1: If $x_k \in \mathcal{X}$ is a solution of the VI, then **stop**.

STEP 2: Compute $x_{k+1} = \text{Perseus}(p, x_k, L, T_{\text{inner}}, \text{opt})$.

Output: x_{T+1}.

$h(x) = \frac{1}{p!}\|x - v_{k+1}\|^p$ (see Nesterov [2021b] for the precise definition). Putting these pieces together yields the desired result. From a computational viewpoint, we can use the generalized mirror-prox method in Titov et al. [2022] to compute $x_{k+1} \in \mathcal{X}$ satisfying the following approximation condition:

$$\sup_{x \in \mathcal{X}} \langle F_{v_{k+1}}(x_{k+1}), x_{k+1} - x \rangle \leq \frac{L}{p!}\|x_{k+1} - v_{k+1}\|^{p+1}. \tag{7.7}$$

Therefore, the solution of the subproblem in our framework is computationally tractable for the monotone setting. Other efficient numerical methods have been developed for the case of $p = 2$ and $\mathcal{X} = \mathbb{R}^d$ in the context of optimization [Grapiglia and Nesterov, 2021] and minimax optimization [Huang et al., 2022b, Adil et al., 2022, Lin et al., 2022d] and shown to be effective in practice.

In the nonmonotone setting, the VI in Eq. (7.6) is not necessarily monotone and computing a solution x_{k+1} satisfying Eq. (7.7) is intractable in general [Daskalakis et al., 2021]. However, $F_{v_{k+1}}$ is defined as the sum of a polynomial and a regularization term, and this special structure might lend itself to efficient numerical methods. For example, we consider the optimization setting where $F = \nabla f$ for a nonconvex function $f : \mathbb{R}^d \mapsto \mathbb{R}$ with a Lipschitz second-order derivative, $\mathcal{X} = \mathbb{R}^d$ and $p = 2$. Solving the VI in Eq. (7.6) is equivalent to solving cubic regularization subproblems in unconstrained optimization: finding a global solution of the regularized polynomial in the following form of

$$\langle \nabla f(v_{k+1}), x - v_{k+1} \rangle + \frac{1}{2}\langle x - v_{k+1}, \nabla^2 f(v_{k+1})(x - v_{k+1}) \rangle + \frac{L}{3}\|x - v_{k+1}\|^3.$$

This optimization problem is nonconvex but can be solved approximately in a provably efficient manner. Examples of cubic solvers include some generalized conjugate gradient methods with the Lanczos process [Gould et al., 1999, 2010] and a simple variant of gradient descent [Carmon and Duchi, 2019]. A recent textbook [Cartis et al., 2022] provides a detailed discussion of these techniques. The generalization of these techniques to handle the VI in Eq. (7.6) is challenging, however, and beyond the scope of this paper.

Comments on adaptive strategies. Our adaptive strategy for updating λ_{k+1} was inspired by an in-depth consideration of the reason a nontrivial binary search procedure is

needed in existing p^{th}-order methods. These methods compute a pair, $\lambda_{k+1} > 0, x_{k+1} \in \mathcal{X}$, that (approximately) solve the x-subproblem that contains λ and the λ-subproblem that contains x. In particular, the conditions can be written as follows:

$$\begin{aligned}&\alpha_- \leq \tfrac{\lambda_{k+1} L \|x_{k+1} - v_{k+1}\|^{p-1}}{p!} \leq \alpha_+ \text{ for proper choices of } \alpha_- \text{ and } \alpha_+,\\ &\langle F_{v_{k+1}}(x_{k+1}) + \tfrac{1}{\lambda_{k+1}}(x_{k+1} - v_{k+1}), x - x_{k+1}\rangle \geq 0 \text{ for all } x \in \mathcal{X},\end{aligned}$$

where

$$\begin{aligned}F_v(x) = F(v) + \langle \nabla F(v), x - v\rangle \\ + \ldots + \tfrac{1}{(p-1)!}\nabla^{(p-1)} F(v)[x-v]^{p-1} + \tfrac{L}{(p-1)!}\|x - v\|^{p-1}(x - v).\end{aligned} \tag{7.8}$$

A key observation is that there can be some x-subproblems that do not need to refer to λ; e.g., the one employed in Algorithm 20. Indeed, we compute $x_{k+1} \in \mathcal{X}$ that approximately satisfies the following condition:

$$\langle F_{v_{k+1}}(x_{k+1}), x - x_{k+1}\rangle \geq 0 \text{ for all } x \in \mathcal{X}.$$

It suffices to return $x_{k+1} \in \mathcal{X}$ with a sufficiently good quality to give us $\lambda_{k+1} > 0$ using a simple update rule. Intuitively, such an adaptive strategy makes sense since λ_{k+1} serves as the stepsize in the dual space and we need to be aggressive as the iterate x_{k+1} approaches the set of optimal solutions to the VI. Meanwhile, the quantity $\|x_{k+1} - v_{k+1}\|$ can be used to measure the distance between x_{k+1} and an optimal solution, and the order $p \in \{1, 2, 3, \ldots\}$ quantifies the relationship between the closeness and the exploitation of high-order derivative information. In summary, λ_{k+1} becomes larger for a better iterate $x_{k+1} \in \mathcal{X}$ and such a choice leads to a faster global rate of convergence.

Restart version of Perseus. We summarize the restarted version of our p^{th}-order method in Algorithm 21. This method, which we refer to as Perseus-restart(p, x_0, L, σ, D, T, opt), combines Algorithm 20 with a restart scheme; cf. Nemirovski and Nesterov [1985], Nesterov [2013a], O'donoghue and Candes [2015], Nesterov [2018].

Restart schemes stop an algorithm when a criterion is satisfied and then restart the algorithm with a new input. Originally studied in the setting of momentum-based methods, restarting has been recognized as an important tool for designing linearly convergent algorithms when the objective function is strongly/uniformly convex [Nemirovski and Nesterov, 1985, Nesterov, 2013a, Ghadimi and Lan, 2013a] or has some other structures [Freund and Lu, 2018, Necoara et al., 2019, Renegar and Grimmer, 2022]. Note that strong monotonicity is a generalization of such regularity conditions. As such, it is natural to consider a restarted version of our method, hoping to achieve linear convergence. Accordingly, at each iteration of Algorithm 21, we use $x_{k+1} = \text{Perseus}(p, x_k, L, t, \text{opt})$ as a subroutine. In other words, we simply restart Perseus every $t \geq 1$ iterations and take advantage of average iterates or best iterates to generate x_{k+1} from x_k. In addition, it is worth mentioning that the choice of t can

be specialized to different settings and/or different type of convergence guarantees. Indeed, we set $\mathsf{opt} = 0$ for the strong monotone setting and $\mathsf{opt} = 1$ to obtain a local convergence guarantee.

In the context of VI, restarting strategies have been used to extend high-order extragradient methods [Bullins and Lai, 2022, Adil et al., 2022] from the monotone setting to the strongly monotone setting [Ostroukhov et al., 2020, Huang and Zhang, 2022a]. Moreover, several papers focus on the investigation of adaptive restart schemes that speed up the convergence of classical first-order methods [Giselsson and Boyd, 2014, O'donoghue and Candes, 2015] and provide theoretical guarantees in a general setting where the objective function is smooth and has Hölderian growth [Roulet and d'Aspremont, 2017, Fercoq and Qu, 2019]. A drawback of these schemes is that they rely on knowing appropriately accurate approximations of problem parameters. The same issue arises for our method, given that Algorithm 21 needs to choose $T_{\text{inner}} \geq 1$. In the optimization setting, recent work by Renegar and Grimmer [2022] shows how to alleviate this problem via a simple restart scheme that makes no attempt to learn parameter values and only requires the information that is readily available in practice. It is an interesting open question as to whether such a scheme can be found in the VI setting for Perseus.

Main results. We provide our main results on the convergence rate for Algorithm 20 and 21 in terms of the number of calls of the subproblem solvers. Note that Assumption 7.2.3 will be made throughout and we impose the Minty condition (see Definition 7.2.5) for the nonmonotone setting.

Monotone setting. The following theorems give us the global convergence rate of Algorithm 20 and 21 for smooth and (strongly) monotone VIs.

Theorem 7.3.1 *Suppose that Assumption 7.2.3 holds and $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ is monotone and let $\epsilon \in (0, 1)$. Then, the required number of iterations is*

$$T = O\left(\left(\frac{LD^{p+1}}{\epsilon}\right)^{\frac{2}{p+1}}\right),$$

where $\hat{x} = \mathsf{Perseus}(p, x_0, L, T, 0)$ satisfies $\text{GAP}(\hat{x}) \leq \epsilon$ and the total number of calls of the subproblem solvers is equal to T. Here, $p \in \{1, 2, \ldots\}$ is an order, $L > 0$ is a Lipschitz constant for $(p-1)^{\text{th}}$-order smoothness of F and $D > 0$ is the diameter of $\mathcal{X}$.

Theorem 7.3.2 *Suppose that Assumption 7.2.3 holds and $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ is μ-strongly monotone and let $\epsilon \in (0, 1)$. Then, the required number of iterations is*

$$T = O\left(\log_2\left(\frac{D}{\epsilon}\right)\right),$$

such that $\hat{x} = \textsf{Perseus-restart}(p, x_0, L, \mu, D, T, 0)$ *satisfies* $\|\hat{x} - x^\star\| \leq \epsilon$ *and the total number of calls of the subproblem solvers is bounded by*

$$O\left((\kappa D^{p-1})^{\frac{2}{p+1}} \log_2\left(\frac{D}{\epsilon}\right)\right),$$

where $p \in \{1, 2, \ldots\}$ *is an order,* $\kappa = L/\mu > 0$ *is the condition number of* F*,* $D > 0$ *is the diameter of* $\mathcal{X}$ *and* $x^\star \in \mathcal{X}$ *is one weak solution.*

Remark 7.3.3 *For the first-order methods (i.e., the case of* $p = 1$*), the convergence guarantee in Theorem 7.3.1 recovers the global rate of* $O(L/\epsilon)$ *in Nesterov [2007, Theorem 2]. The same rate has been derived for other methods [Nemirovski, 2004, Monteiro and Svaiter, 2010, Mokhtari et al., 2020a, Kotsalis et al., 2022] and is known to match the established lower bound [Ouyang and Xu, 2021]. For the second-order and high-order methods (i.e., the case of* $p \geq 2$*), our results improve upon the state-of-the-art results [Monteiro and Svaiter, 2012, Bullins and Lai, 2022, Lin and Jordan, 2023, Jiang and Mokhtari, 2022] by shaving off the log factors.*

Remark 7.3.4 *For the first-order methods, Theorem 7.3.2 recovers the global linear convergence rate achieved by the dual extrapolation method and matches the established lower bound [Zhang et al., 2022a]. For the second-order and high-order methods, our results improve upon the results in Jiang and Mokhtari [2022] by shaving off the log factors. We believe that these bounds can not be improved although we do not know of lower bounds.*

Local convergence. We present the local convergence property of our methods for the strongly monotone VIs.

Theorem 7.3.5 *Suppose that Assumption 7.2.3 holds and* $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ *is* μ*-strongly monotone and let* $\{x_k\}_{k=0}^{T+1}$ *be generated by* $\textsf{Perseus-restart}(p, x_0, L, \mu, D, T, 1)$*. Then, the following statement holds true,*

$$\|x_{k+1} - x^\star\| \leq \sqrt{\frac{2^p(5p-2)\kappa}{p!}}\|x_k - x^\star\|^{\frac{p+1}{2}},$$

where $\kappa = L/\mu > 0$ *is the condition number of the VI,* $D > 0$ *is the diameter of* $\mathcal{X}$ *and* $x^\star$ *is the unique weak solution of the VI. As a consequence, if* $p \geq 2$ *and the following condition holds true,*

$$\|x_0 - x^\star\| \leq \frac{1}{2}\left(\frac{p!}{2^p(5p-2)\kappa}\right)^{\frac{1}{p-1}},$$

the iterates $\{x_k\}_{k=0}^{T+1}$ *converge to* $x^\star \in \mathcal{X}$ *in at least a superlinear rate.*

Remark 7.3.6 *The local convergence guarantee in Theorem 7.3.5 is derived for the second-order and high-order methods (i.e., the case of* $p \geq 2$*) and is posited as their advantage over first-order method if we hope to pursue high-accuracy solutions. In this context, Jiang and Mokhtari [2022] provided the same local convergence guarantee for the generalized optimistic gradient methods as our results in Theorem 7.3.5 but without counting the complexity bound of binary search procedure.*

Nonmonotone setting. We consider smooth and nonmonotone VIs satisfying the Minty condition and present the global rate of Algorithm 20 and 21 in terms of the number of calls of the subproblem solvers.

Theorem 7.3.7 *Suppose that Assumption 7.2.3 and the Minty condition hold true and let* $\epsilon \in (0,1)$. *Then, the required number of iterations is*

$$T = O\left(\left(\frac{LD^{p+1}}{\epsilon}\right)^{\frac{2}{p}}\right),$$

such that $\hat{x} = \textsf{Perseus}(p, x_0, L, T, 2)$ *satisfies* $\textsc{res}(\hat{x}) \leq \epsilon$ *and the total number of calls of the subproblem solvers is equal to* T. *Here,* $p \in \{1, 2, \ldots\}$ *is an order,* $L > 0$ *is the Lipschitz constant for* $(p-1)^{\text{th}}$*-order smoothness of* F *and* $D > 0$ *is the diameter of* $\mathcal{X}$.

Remark 7.3.8 *The convergence guarantee in Theorem 7.3.7 has been derived for other first-order methods [Dang and Lan, 2015, Song et al., 2020] for the case of* $p = 1$. *They are new for the case of* $p \geq 2$ *to the best of our knowledge.*

Remark 7.3.9 *For the smooth and nonmonotone VIs satisfying the strong Minty condition, we can obtain the same rate of convergence in Theorem 7.3.2 and 7.3.5 but to a weak solution rather than a unique strong solution. The proof would be the same as that used for strongly monotone setting.*

Lower bound. We provide the lower bound for the monotone setting under a linear span assumption. Our analysis and hard instance are largely inspired by the constructions and techniques from Nesterov [2021b] and Adil et al. [2022]. However, different from Adil et al. [2022], our lower bound is established for a class of p^{th}-order methods that include both our method and their method, rather than p^{th}-order methods restricted to solve the primal problem [Adil et al., 2022, Eq. (11)].

For constructing the problems that are difficult for p^{th}-order methods, it is convenient to consider the saddle point problem, $\min_{z\in\mathcal{Z}} \max_{y\in\mathcal{Y}} f(z,y)$, which is a special monotone VI defined via an operator F and a closed, convex and bounded set $\mathcal{X}$ as follows:

$$x = \begin{bmatrix} z \\ y \end{bmatrix}, \qquad F(x) = \begin{bmatrix} \nabla_z f(z,y) \\ -\nabla_y f(z,y) \end{bmatrix}, \qquad \mathcal{X} = \mathcal{Z} \times \mathcal{Y}.$$

Let us describe the abilities of p^{th}-order methods of degree $p \geq 2$ in generating the new iterates. More specifically, the output of oracle at a point $\bar{x} \in \mathcal{X}$ consists in the set of multi-linear forms given by

$$F(\bar{x}), \nabla F(\bar{x}), \ldots, \nabla^{(p-1)} F(\bar{x}).$$

Therefore, we assume that the p^{th}-order method in our algorithm class is able to compute the solution of the nonlinear equation $\Phi_{a,\gamma}(h) = 0$ where $\Phi_{a,\gamma,m}(h) = a_0 F(\bar{x}) + \sum_{i=1}^{p-1} a_i \nabla^{(i)} F(\bar{x})[h]^i + \gamma \|h\|^{m-1} h$ and the coefficients $a \in \mathbb{R}^p$, $\gamma > 0$ and $m \geq 2$. Following the notions defined in Adil et al. [2022], we denote by $\Gamma_{\bar{x},F}(a, \gamma, m)$ the set of all solutions of the aforementioned nonlinear equation. Then, we define the linear subspace given by

$$S_F(\bar{x}) = \text{Lin}\left(\Gamma_{\bar{x},F}(a, \gamma, m) : a \in \mathbb{R}^p, \gamma > 0, m \geq 2\right).$$

Our assumption about the form of p^{th}-order methods in our algorithm class is summarized as follows:

Assumption 7.3.10 *The p^{th}-order method generates a sequence of iterates $\{x_k\}_{k\geq 0}$ satisfying the recursive condition: $x_{k+1} \in x_0 + \sum_{i=0}^{k} S_F(x_i)$ for all $k \geq 0$.*

Note that Assumption 7.3.10 is well known as a linear span assumption [Nesterov, 2021b] and is satisfied for majority of high-order methods, including Algorithm 20. The same lower bound has been established in Adil et al. [2022] for a special class of p^{th}-order methods restricted to solve the primal problem under Assumption 7.3.10. Indeed, their construction is based on a saddle-point problem $\min_{z\in\mathcal{Z}} \max_{y\in\mathcal{Y}} f(z, y)$ and they assume that any method in their algorithm class not only satisfies Assumption 7.3.10 but has the access to $\nabla\phi(\bar{z}), \ldots, \nabla^{(p)}\phi(\bar{z})$ where $\phi(z) = \max_{y\in\mathcal{Y}} f(z, y)$ refers to the objective function of primal problem (see Adil et al. [2022, Lemma 4.3]).

Proving the lower bound for general p^{th}-order methods under Assumption 7.3.10 requires a new hard instance, which we shall see is a nonlinear generalization of the hard instance used in Adil et al. [2022]. The following theorem summarizes our main result.

Theorem 7.3.11 *Fixing $p \geq 2$, $L > 0$ and $T > 0$ and letting $d \geq 2T + 1$ be the problem dimension. There exists two closed, convex and bounded sets $\mathcal{Z}, \mathcal{Y} \subseteq \mathbb{R}^d$ and a function $f(z, y) : \mathcal{Z} \times \mathcal{Y} \mapsto \mathbb{R}$ that is convex-concave with an optimal saddle-point solution $(z_\star, y_\star) \in \mathcal{Z} \times \mathcal{Y}$ such that the iterates $\{(z_k, y_k)\}_{k\geq 0}$ generated by any p^{th}-order method under Assumption 7.3.10 must satisfy*

$$\min_{0\leq k\leq T} \left\{ \max_{y\in\mathcal{Y}} f(z_k, y) - \min_{z\in\mathcal{Z}} f(z, y_k) \right\} \geq \left(\tfrac{1}{4^{p+1}(p+1)!}\right) L D_{\mathcal{Z}} D_{\mathcal{Y}}^p T^{-\frac{p+1}{2}}.$$

Remark 7.3.12 *The lower bound in Theorem 7.3.11 shows that any p^{th}-order method satisfying Assumption 7.3.10 requires at least $\Omega((L D_{\mathcal{Z}} D_{\mathcal{Y}}^p)^{\frac{2}{p+1}} \epsilon^{-\frac{2}{p+1}})$ iterations to reach an ϵ-weak solution. Combined this result with Theorem 7.3.1 shows that Algorithm 20 is an optimal p^{th}-order method for solving smooth and monotone VIs. As mentioned before, we have improved the results in Adil et al. [2022] by constructing a new hard instance and deriving the same lower bound for a more broad class of p^{th}-order methods that include both Algorithm 20 and the high-order extragradient method in Adil et al. [2022].*

Remark 7.3.13 *For the lower bound for finding an ϵ-strong solution in monotone setting, the case for first-order VI methods have been investigated in Diakonikolas [2020]. The key idea is to use the lower bound for finding an ϵ-weak solution [Ouyang and Xu, 2021] and the algorithmic reductions to derive lower bounds. However, such a reduction is mostly based on the high-order generalization of Halpern iteration and is thus beyond the scope of the current manuscript. In particular, we have developed a simple and optimal p^{th}-order VI method for finding an ϵ-weak solution in the monotone setting. However, the optimal algorithm for finding an ϵ-strong solution in the monotone setting is likely to be different as evidenced by Diakonikolas [2020]. Computing an ϵ-strong solution and/or an ϵ-weak solution are complementary, yet different, and they indeed deserve separate study in their own right. Moreover, the lower bound for finding an ϵ-strong solution under the Minty condition is largely unexplored and missing even in the current literature for first-order VI methods.*

Remark 7.3.14 *It remains unclear whether or not Assumption 7.3.10 can be removed without sacrificing the lower bound on the iteration complexity. It may be possible to approach this issue by using the rotation technique [Carmon et al., 2020, Section 3.3] and [Arjevani et al., 2019]. Despite striking results on extending the rotation techniques from optimization to minimax optimization [Ouyang and Xu, 2021, Zhang et al., 2022a], this problem becomes challenging for p^{th}-order methods when $p \geq 2$ since the analysis from Ouyang and Xu [2021], Zhang et al. [2022a] cannot be directly extended from bilinear saddle point problems to general nonlinear saddle point problems. In our view, resolving it requires a new construction of "chain-style" hard function instances and rotation techniques.*

7.4 Convergence Analysis

We present the convergence analysis for our p^{th}-order method (Algorithm 20) and its restarted version (Algorithm 21). Indeed, we provide the global convergence guarantee (Theorems 7.3.1 and 7.3.2) and local convergence guarantee for the monotone setting (Theorems 7.3.5). We analyze the nonmonotone setting under the Minty condition (Theorems 7.3.7). Finally, we establish the lower bound under a linear span assumption (Theorem 7.3.11).

Technical lemmas. We define the following Lyapunov function for the iterates $\{x_k\}_{k\geq 0}$ that are generated by Algorithm 20:

$$\mathcal{E}_k = \max_{v\in\mathcal{X}} \langle s_k, v - x_0\rangle - \tfrac{1}{2}\|v - x_0\|^2. \tag{7.9}$$

This function is used to prove technical results that pertain to the dynamics of Algorithm 20.

Lemma 7.4.1 *Suppose that Assumption 7.2.3 holds true. For every integer $T \geq 1$, we have*

$$\begin{aligned}&\sum_{k=1}^{T} \lambda_k \langle F(x_k), x_k - x\rangle \\ &\quad\leq\ \mathcal{E}_0 - \mathcal{E}_T + \langle s_T, x - x_0\rangle - \tfrac{1}{10}\left(\sum_{k=1}^{T} \|x_k - v_k\|^2\right), \quad \text{for all } x \in \mathcal{X}.\end{aligned}$$

Proof. By combining Eq. (7.9) and the definition of v_{k+1}, we have

$$\mathcal{E}_k = \langle s_k, v_{k+1} - x_0\rangle - \tfrac{1}{2}\|v_{k+1} - x_0\|^2.$$

Then, we have

$$\begin{aligned}&\mathcal{E}_{k+1} - \mathcal{E}_k \\ &\quad=\ \langle s_{k+1}, v_{k+2} - x_0\rangle - \langle s_k, v_{k+1} - x_0\rangle - \tfrac{1}{2}\left(\|v_{k+2} - x_0\|^2 - \|v_{k+1} - x_0\|^2\right) \\ &\quad=\ \langle s_{k+1} - s_k, v_{k+1} - x_0\rangle + \langle s_{k+1}, v_{k+2} - v_{k+1}\rangle - \tfrac{1}{2}\left(\|v_{k+2} - x_0\|^2 - \|v_{k+1} - x_0\|^2\right).\end{aligned} \tag{7.10}$$

By using the update formula for v_{k+1} again, we have

$$\langle x - v_{k+1}, s_k - v_{k+1} + x_0\rangle \leq 0, \quad \text{for all } x \in \mathcal{X}.$$

Letting $x = v_{k+2}$ in this inequality and using $\langle a, b\rangle = \frac{1}{2}(\|a+b\|^2 - \|a\|^2 - \|b\|^2)$, we have

$$\begin{aligned}\langle s_k, v_{k+2} - v_{k+1}\rangle &\leq \langle v_{k+1} - x_0, v_{k+2} - v_{k+1}\rangle \\ &=\ \tfrac{1}{2}\left(\|v_{k+2} - x_0\|^2 - \|v_{k+1} - x_0\|^2 - \|v_{k+2} - v_{k+1}\|^2\right).\end{aligned} \tag{7.11}$$

Plugging Eq. (7.11) into Eq. (7.10) and using the update formula of s_{k+1}, we obtain:

$$\begin{aligned}&\mathcal{E}_{k+1} - \mathcal{E}_k \\ &\overset{\text{Eq. (7.11)}}{\leq} \langle s_{k+1} - s_k, v_{k+1} - x_0\rangle + \langle s_{k+1} - s_k, v_{k+2} - v_{k+1}\rangle - \tfrac{1}{2}\|v_{k+2} - v_{k+1}\|^2 \\ &= \langle s_{k+1} - s_k, v_{k+2} - x_0\rangle - \tfrac{1}{2}\|v_{k+2} - v_{k+1}\|^2 \\ &\leq \lambda_{k+1}\langle F(x_{k+1}), x_0 - v_{k+2}\rangle - \tfrac{1}{2}\|v_{k+2} - v_{k+1}\|^2 \\ &= \lambda_{k+1}\langle F(x_{k+1}), x_0 - x\rangle + \lambda_{k+1}\langle F(x_{k+1}), x - x_{k+1}\rangle \\ &\quad + \lambda_{k+1}\langle F(x_{k+1}), x_{k+1} - v_{k+2}\rangle - \tfrac{1}{2}\|v_{k+2} - v_{k+1}\|^2,\end{aligned}$$

for any $x \in \mathcal{X}$. Summing up this inequality over $k = 0, 1, \ldots, T-1$ and changing the counter $k+1$ to k yields that

$$\begin{aligned}\sum_{k=1}^{T} \lambda_k \langle F(x_k), x_k - x\rangle \leq \mathcal{E}_0 - \mathcal{E}_T + \underbrace{\sum_{k=1}^{T} \lambda_k \langle F(x_k), x_0 - x\rangle}_{\mathbf{I}} \\ + \underbrace{\sum_{k=1}^{T} \lambda_k \langle F(x_k), x_k - v_{k+1}\rangle - \tfrac{1}{2}\|v_k - v_{k+1}\|^2}_{\mathbf{II}}.\end{aligned} \tag{7.12}$$

Using the update formula for s_{k+1} and letting $s_0 = 0_d \in \mathbb{R}^d$, we have

$$\begin{aligned}\mathbf{I} &= \sum_{k=1}^{T} \langle \lambda_k F(x_k), x_0 - x \rangle \\ &= \sum_{k=1}^{T} \langle s_{k-1} - s_k, x_0 - x \rangle = \langle s_0 - s_T, x_0 - x \rangle = \langle s_T, x - x_0 \rangle.\end{aligned} \tag{7.13}$$

Since $x_{k+1} \in \mathcal{X}$ satisfies Eq. (7.7), we have

$$\langle F_{v_k}(x_k), x - x_k \rangle \geq -\tfrac{L}{p!}\|x_k - v_k\|^{p+1}, \quad \text{for all } x \in \mathcal{X}, \tag{7.14}$$

where $F_v(x) : \mathbb{R}^d \to \mathbb{R}^d$ is defined for any fixed $v \in \mathcal{X}$ as follows:

$$\begin{aligned}F_{v_k}(x) &= F(v_k) + \langle \nabla F(v_k), x - v_k \rangle + \ldots + \tfrac{1}{(p-1)!}\nabla^{(p-1)} F(v_k)[x - v_k]^{p-1} \\ &\quad + \tfrac{5L}{(p-1)!}\|x - v_k\|^{p-1}(x - v_k).\end{aligned}$$

Under Assumption 7.2.3, we obtain from Bullins and Lai [2022, Fact 2.5] or Jiang and Mokhtari [2022, Eq. (6)] that

$$\|F(x_k) - F_{v_k}(x_k) + \tfrac{5L}{(p-1)!}\|x_k - v_k\|^{p-1}(x_k - v_k)\| \leq \tfrac{L}{p!}\|x_k - v_k\|^p. \tag{7.15}$$

Letting $x = v_{k+1}$ in Eq. (7.14), we have

$$\langle F_{v_k}(x_k), x_k - v_{k+1} \rangle \leq \tfrac{L}{p!}\|x_k - v_k\|^{p+1}. \tag{7.16}$$

Inspired by Eq. (7.15) and Eq. (7.16), we decompose $\langle F(x_k), x_k - v_{k+1} \rangle$ as follows:

$$\begin{aligned}&\langle F(x_k), x_k - v_{k+1} \rangle \\ &= \langle F(x_k) - F_{v_k}(x_k) + \tfrac{5L}{(p-1)!}\|x_k - v_k\|^{p-1}(x_k - v_k), x_k - v_{k+1} \rangle \\ &\quad + \langle F_{v_k}(x_k), x_k - v_{k+1} \rangle - \tfrac{5L}{(p-1)!}\|x_k - v_k\|^{p-1}\langle x_k - v_k, x_k - v_{k+1} \rangle \\ &\leq \|F(x_k) - F_{v_k}(x_k) + \tfrac{5L}{(p-1)!}\|x_k - v_k\|^{p-1}(x_k - v_k)\| \cdot \|x_k - v_{k+1}\| \\ &\quad + \langle F_{v_k}(x_k), x_k - v_{k+1} \rangle - \tfrac{5L}{(p-1)!}\|x_k - v_k\|^{p-1}\langle x_k - v_k, x_k - v_{k+1} \rangle \\ &\overset{\text{Eq. (7.15) and Eq. (7.16)}}{\leq} \tfrac{L}{p!}\|x_k - v_k\|^p\|x_k - v_{k+1}\| + \tfrac{L}{p!}\|x_k - v_k\|^{p+1} \\ &\quad - \tfrac{5L}{(p-1)!}\|x_k - v_k\|^{p-1}\langle x_k - v_k, x_k - v_{k+1} \rangle \\ &\leq \tfrac{2L}{p!}\|x_k - v_k\|^{p+1} + \tfrac{L}{p!}\|x_k - v_k\|^p\|v_k - v_{k+1}\| \\ &\quad - \tfrac{5L}{(p-1)!}\|x_k - v_k\|^{p-1}\langle x_k - v_k, x_k - v_{k+1} \rangle.\end{aligned}$$

Note that we have

$$\begin{aligned}&\langle x_k - v_k, x_k - v_{k+1} \rangle \\ &= \|x_k - v_k\|^2 + \langle x_k - v_k, v_k - v_{k+1} \rangle \geq \|x_k - v_k\|^2 - \|x_k - v_k\|\|v_k - v_{k+1}\|.\end{aligned}$$

Putting these pieces together yields that

$$\langle F(x_k), x_k - v_{k+1}\rangle \le \tfrac{(5p+1)L}{p!}\|x_k - v_k\|^p\|v_k - v_{k+1}\| - \tfrac{(5p-2)L}{p!}\|x_k - v_k\|^{p+1}.$$

Since $\frac{1}{20p-8} \le \frac{\lambda_k L\|x_k - v_k\|^{p-1}}{p!} \le \frac{1}{10p+2}$ for all $k \ge 1$, we have

$$\begin{aligned}
\mathbf{II} &\le \sum_{k=1}^{T}\left(\tfrac{(5p+1)\lambda_k L}{p!}\|x_k - v_k\|^p\|v_k - v_{k+1}\| - \tfrac{1}{2}\|v_k - v_{k+1}\|^2 - \tfrac{(5p-2)\lambda_k L}{p!}\|x_k - v_k\|^{p+1}\right)\\
&\le \sum_{k=1}^{T}\left(\tfrac{1}{2}\|x_k - v_k\|\|v_k - v_{k+1}\| - \tfrac{1}{2}\|v_k - v_{k+1}\|^2 - \tfrac{1}{4}\|x_k - v_k\|^2\right)\\
&\le \sum_{k=1}^{T}\left(\max_{\eta\ge 0}\left\{\tfrac{1}{2}\|x_k - v_k\|\eta - \tfrac{1}{2}\eta^2\right\} - \tfrac{1}{4}\|x_k - v_k\|^2\right)\\
&= -\tfrac{1}{8}\left(\sum_{k=1}^{T}\|x_k - v_k\|^2\right). \qquad (7.17)
\end{aligned}$$

Plugging Eq. (7.13) and Eq. (7.17) into Eq. (7.12) yields that

$$\sum_{k=1}^{T}\lambda_k\langle F(x_k), x_k - x\rangle \le \mathcal{E}_0 - \mathcal{E}_T + \langle s_T, x - x_0\rangle - \tfrac{1}{8}\left(\sum_{k=1}^{T}\|x_k - v_k\|^2\right).$$

This completes the proof. □

Lemma 7.4.2 *Suppose that Assumption 7.2.3 and the Minty condition hold true and let $x \in \mathcal{X}$. For every integer $T \ge 1$, we have*

$$\sum_{k=1}^{T}\lambda_k\langle F(x_k), x_k - x\rangle \le \tfrac{1}{2}\|x - x_0\|^2, \qquad \sum_{k=1}^{T}\|x_k - v_k\|^2 \le 4\|x^\star - x_0\|^2,$$

where $x^\star \in \mathcal{X}$ denotes the weak solution to the VI.

Proof. For any $x \in \mathcal{X}$, we have

$$\mathcal{E}_0 - \mathcal{E}_T + \langle s_T, x - x_0\rangle = \mathcal{E}_0 - \left(\max_{v\in\mathcal{X}}\langle s_T, v - x_0\rangle - \tfrac{1}{2}\|v - x_0\|^2\right) + \langle s_T, x - x_0\rangle.$$

Since $s_0 = 0_d$, we have $\mathcal{E}_0 = 0$ and

$$\mathcal{E}_0 - \mathcal{E}_T + \langle s_T, x - x_0\rangle \le -\left(\langle s_T, x - x_0\rangle - \tfrac{1}{2}\|x - x_0\|^2\right) + \langle s_T, x - x_0\rangle = \tfrac{1}{2}\|x - x_0\|^2.$$

This together with Lemma 7.4.1 yields that

$$\sum_{k=1}^{T}\lambda_k\langle F(x_k), x_k - x\rangle + \tfrac{1}{8}\left(\sum_{k=1}^{T}\|x_k - v_k\|^2\right) \le \tfrac{1}{2}\|x - x_0\|^2, \quad \text{for all } x \in \mathcal{X},$$

which implies the first inequality. Since the VI satisfies the Minty condition (see Definition 7.2.5), there exists $x^\star \in \mathcal{X}$ such that $\langle F(x_k), x_k - x^\star \rangle \geq 0$ for all $k \geq 1$. Letting $x = x^\star$ in the above inequality yields the second inequality. □

We provide a technical lemma establishing a lower bound for $\sum_{k=1}^T \lambda_k$.

Lemma 7.4.3 *Suppose that Assumption 7.2.3 and the Minty condition hold true. For every integer $k \geq 1$, we have*

$$\sum_{k=1}^T \lambda_k \geq \tfrac{p!}{(20p-8)L} \left(\tfrac{1}{4\|x^\star - x_0\|^2} \right)^{\frac{p-1}{2}} T^{\frac{p+1}{2}},$$

where $x^\star \in \mathcal{X}$ denotes the weak solution to the VI.

Proof. Without loss of generality, we assume that $x_0 \neq x^\star$. For $p = 1$, we have $\lambda_k = \frac{1}{12L}$ for all $k \geq 1$. For $p \geq 2$, we have

$$\begin{aligned} &\sum_{k=1}^T (\lambda_k)^{-\frac{2}{p-1}} \left(\tfrac{p!}{(20p-8)L}\right)^{\frac{2}{p-1}} \\ &\leq \sum_{k=1}^T (\lambda_k)^{-\frac{2}{p-1}} (\lambda_k \|x_k - v_k\|^{p-1})^{\frac{2}{p-1}} = \sum_{k=1}^T \|x_k - v_k\|^2 \overset{\text{Lemma 7.4.2}}{\leq} 4\|x^\star - x_0\|^2. \end{aligned}$$

By the Hölder inequality, we have

$$\sum_{k=1}^T 1 = \sum_{k=1}^T \left((\lambda_k)^{-\frac{2}{p-1}} \right)^{\frac{p-1}{p+1}} (\lambda_k)^{\frac{2}{p+1}} \leq \left(\sum_{k=1}^T (\lambda_k)^{-\frac{2}{p-1}} \right)^{\frac{p-1}{p+1}} \left(\sum_{k=1}^T \lambda_k \right)^{\frac{2}{p+1}}.$$

Putting these pieces together yields that

$$T \leq (4\|x^\star - x_0\|^2)^{\frac{p-1}{p+1}} \left(\tfrac{(20p-8)L}{p!} \right)^{\frac{2}{p+1}} \left(\sum_{k=1}^T \lambda_k \right)^{\frac{2}{p+1}},$$

Plugging this into the above inequality yields that

$$\sum_{k=1}^T \lambda_k \geq \tfrac{p!}{(20p-8)L} \left(\tfrac{1}{4\|x^\star - x_0\|^2} \right)^{\frac{p-1}{2}} T^{\frac{p+1}{2}}.$$

This completes the proof. □

Proof of Theorem 7.3.1. We see from Harker and Pang [1990, Theorem 3.1] that at least one strong solution to the VI exists since F is continuous and $\mathcal{X}$ is convex, closed and bounded. Since any strong solution is a weak solution if F is further assumed to be monotone, we obtain that the VI satisfies the Minty condition.

Letting $x \in \mathcal{X}$, we derive from the monotonicity of F and the definition of $\tilde{x}_T$ (i.e., $\mathsf{opt} = 0$) that

$$\begin{aligned}&\langle F(x), \tilde{x}_T - x\rangle \\ &= \frac{1}{\sum_{k=1}^T \lambda_k}\left(\sum_{k=1}^T \lambda_k \langle F(x), x_k - x\rangle\right) \leq \frac{1}{\sum_{k=1}^T \lambda_k}\left(\sum_{k=1}^T \lambda_k \langle F(x_k), x_k - x\rangle\right).\end{aligned}$$

Combining this inequality with the first inequality in Lemma 7.4.2 yields that

$$\langle F(x), \tilde{x}_T - x\rangle \leq \frac{\|x - x_0\|^2}{2(\sum_{k=1}^T \lambda_k)}, \quad \text{for all } x \in \mathcal{X}.$$

Since $x_0 \in \mathcal{X}$, we have $\|x - x_0\| \leq D$ and hence

$$\langle F(x), \tilde{x}_T - x\rangle \leq \frac{D^2}{2(\sum_{k=1}^T \lambda_k)}, \quad \text{for all } x \in \mathcal{X}.$$

Then, we combine Lemma 7.4.3 and the fact that $\|x^\star - x_0\| \leq D$ to obtain that

$$\langle F(x), \tilde{x}_T - x\rangle \leq \frac{2^p(5p-2)}{p!} L D^{p+1} T^{-\frac{p+1}{2}}, \quad \text{for all } x \in \mathcal{X}.$$

By the definition of a gap function (see Eq. (7.4)), we have

$$\textsc{gap}(\tilde{x}_T) = \sup_{x \in \mathcal{X}} \langle F(x), \tilde{x}_T - x\rangle \leq \frac{2^p(5p-2)}{p!} L D^{p+1} T^{-\frac{p+1}{2}}. \tag{7.18}$$

Therefore, we conclude from Eq. (7.18) that we can set

$$T = O\left(\left(\frac{L D^{p+1}}{\epsilon}\right)^{\frac{2}{p+1}}\right),$$

such that $\hat{x} = \mathsf{Perseus}(p, x_0, L, T, 0)$ satisfies $\textsc{gap}(\hat{x}) \leq \epsilon$. The total number of calls of the subproblem solvers is equal to T since our algorithm calls the subproblem solvers once at each iteration. This completes the proof.

Proof of Theorem 7.3.2. In the strongly monotone setting with a convex, closed and bounded set, the solution $x^\star \in \mathcal{X}$ to the VI exists and is unique Facchinei and Pang [2007] and the VI satisfies the Minty condition.

We first consider the relationship between $\|\hat{x} - x^\star\|$ and $\|x_0 - x^\star\|$ where $\hat{x} = \mathsf{Perseus}(p, x_0, L, T_{\text{inner}}, 0)$. We derive from Jensen's inequality and the definition of $\tilde{x}_{T_{\text{inner}}}$ that

$$\|\tilde{x}_{T_{\text{inner}}} - x^\star\|^2 = \left\|\frac{1}{\sum_{k=1}^{T_{\text{inner}}} \lambda_k}\left(\sum_{k=1}^{T_{\text{inner}}} \lambda_k x_k\right) - x^\star\right\|^2 \leq \frac{1}{\sum_{k=1}^{T_{\text{inner}}} \lambda_k}\left(\sum_{k=1}^{T_{\text{inner}}} \lambda_k \|x_k - x^\star\|^2\right).$$

Since F is μ-strongly monotone, we have

$$\|x_k - x^\star\|^2 \leq \tfrac{1}{\mu}\langle F(x_k) - F(x^\star), x_k - x^\star\rangle \leq \tfrac{1}{\mu}\langle F(x_k), x_k - x^\star\rangle.$$

Putting these pieces together yields that

$$\|\tilde{x}_{T_{\text{inner}}} - x^\star\|^2 \leq \frac{1}{\mu(\sum_{k=1}^{T_{\text{inner}}} \lambda_k)} \left(\sum_{k=1}^{T_{\text{inner}}} \lambda_k \langle F(x_k), x_k - x^\star\rangle \right). \tag{7.19}$$

Combining the first inequality in Lemma 7.4.2 with Eq. (7.19) yields that

$$\|\tilde{x}_{T_{\text{inner}}} - x^\star\|^2 \leq \frac{1}{2\mu(\sum_{k=1}^{T_{\text{inner}}} \lambda_k)} \|x_0 - x^\star\|^2.$$

This together with Lemma 7.4.3 and the fact that $\hat{x} = \tilde{x}_{T_{\text{inner}}}$ yields that

$$\begin{aligned} \|\hat{x} - x^\star\|^2 &\leq \left((4\|x_0 - x^\star\|^2)^{\frac{p-1}{2}} \frac{10p-4}{p!} \frac{L}{\mu} t^{-\frac{p+1}{2}} \right) \|x_0 - x^\star\|^2 \\ &= \left(\frac{2^p(5p-2)}{p!} \frac{L}{\mu} t^{-\frac{p+1}{2}} \right) \|x_0 - x^\star\|^{p+1}. \end{aligned} \tag{7.20}$$

Since $x_{k+1} = \mathsf{Perseus}(p, x_k, L, T_{\text{inner}}, 0)$ in the scheme of Algorithm 21 and

$$T_{\text{inner}} = \left\lceil \left(\frac{2^{p+1}(5p-2)}{p!} \frac{LD^{p-1}}{\mu} \right)^{\frac{2}{p+1}} \right\rceil, \tag{7.21}$$

we have

$$\|x_{k+1} - x^\star\|^2 \leq \tfrac{1}{2}\|x_k - x^\star\|^2, \quad \text{for all } k = 0, 1, 2, \ldots, T. \tag{7.22}$$

Therefore, we conclude from Eq. (7.21) and Eq. (7.22) that we can set

$$T = O\left(\log_2 \left(\frac{D}{\epsilon} \right) \right),$$

such that $\hat{x} = \mathsf{Perseus\text{-}restart}(p, x_0, L, \mu, T, 0)$ satisfies $\|\hat{x} - x^\star\| \leq \epsilon$. The total number of calls of the subproblem solvers is bounded by

$$O\left(\left(\frac{LD^{p-1}}{\mu} \right)^{\frac{2}{p+1}} \log_2 \left(\frac{D}{\epsilon} \right) \right).$$

This completes the proof.

Proof of Theorem 7.3.5. We first consider the relationship between $\|\hat{x}-x^\star\|$ and $\|x_0-x^\star\|$ where $\hat{x} = \mathsf{Perseus}(p, x_0, L, T_{\text{inner}}, 1)$. By the same argument as used in Theorem 7.3.2, we have (see Eq. (7.20))

$$\|\hat{x} - x^\star\|^2 \leq \left(\tfrac{2^p(5p-2)}{p!}\tfrac{L}{\mu}t^{-\frac{p+1}{2}}\right)\|x_0 - x^\star\|^{p+1}.$$

Since $x_{k+1} = \mathsf{Perseus}(p, x_k, L, T_{\text{inner}}, 2)$ in the scheme of Algorithm 21 and $T_{\text{inner}} = 1$, we have

$$\|x_{k+1} - x^\star\|^2 \leq \left(\tfrac{2^p(5p-2)\kappa}{p!}\right)\|x_k - x^\star\|^{p+1},$$

which implies that

$$\|x_{k+1} - x^\star\| \leq \sqrt{\tfrac{2^p(5p-2)\kappa}{p!}}\|x_k - x^\star\|^{\frac{p+1}{2}}.$$

For the case of $p \geq 2$, we have $\frac{p+1}{2} \geq \frac{3}{2}$ and $p-1 \geq 1$. If the following condition holds true,

$$\|x_0 - x^\star\| \leq \tfrac{1}{2}\left(\tfrac{p!}{2^p(5p-2)\kappa}\right)^{\frac{1}{p-1}},$$

we have

$$\begin{aligned}
&\left(\tfrac{2^p(5p-2)\kappa}{p!}\right)^{\frac{1}{p-1}}\|x_{k+1} - x^\star\| \leq \left(\tfrac{2^p(5p-2)\kappa}{p!}\right)^{\frac{p+1}{2(p-1)}}\|x_k - x^\star\|^{\frac{p+1}{2}} \\
&= \left(\left(\tfrac{2^p(5p-2)\kappa}{p!}\right)^{\frac{1}{p-1}}\|x_k - x^\star\|\right)^{\frac{p+1}{2}} \leq \left(\left(\tfrac{2^p(5p-2)\kappa}{p!}\right)^{\frac{1}{p-1}}\|x_0 - x^\star\|\right)^{(\frac{p+1}{2})^{k+1}} \\
&\leq (\tfrac{1}{2})^{(\frac{p+1}{2})^{k+1}}.
\end{aligned}$$

This completes the proof.

Proof of Theorem 7.3.7. We see from the second inequality in Lemma 7.4.2 that

$$\min_{1\leq k\leq T}\|x_k - v_k\|^2 \leq \tfrac{1}{T}\sum_{k=1}^{T}\|x_k - v_k\|^2 \leq \tfrac{4\|x^\star - x_0\|^2}{T}.$$

By the definition of x_{k_T} (i.e., $\mathsf{opt} = 2$), we have

$$\|x_{k_T} - v_{k_T}\|^2 \leq \tfrac{4\|x^\star - x_0\|^2}{T}. \tag{7.23}$$

Recalling that $x_{k+1} \in \mathcal{X}$ satisfies Eq. (7.7), we have

$$\langle F_{v_k}(x_k), x - x_k\rangle \geq -\tfrac{L}{p!}\|x_k - v_k\|^{p+1}, \quad \text{for all } x \in \mathcal{X},$$

where $F_v(x) : \mathbb{R}^d \to \mathbb{R}^d$ is defined for any fixed $v \in \mathcal{X}$ as follows:

$$\begin{aligned}
F_{v_k}(x) \;=\; & F(v_k) + \langle \nabla F(v_k), x - v_k\rangle + \ldots + \tfrac{1}{(p-1)!}\nabla^{(p-1)}F(v_k)[x - v_k]^{p-1} \\
& + \tfrac{5L}{(p-1)!}\|x - v_k\|^{p-1}(x - v_k).
\end{aligned}$$

Under Assumption 7.2.3, we have Eq. (7.15) which further leads to

$$\|F(x_k) - F_{v_k}(x_k)\| \leq \tfrac{(5p+1)L}{p!}\|x_k - v_k\|^p.$$

Putting these pieces together yields that

$$\begin{aligned}\langle F(x_k), x_k - x\rangle &= \langle F(x_k) - F_{v_k}(x_k), x_k - x\rangle + \langle F_{v_k}(x_k), x_k - x\rangle \\ &\leq \|F(x_k) - F_{v_k}(x_k)\|\|x_k - x\| + \tfrac{L}{p!}\|x_k - v_k\|^{p+1} \\ &\leq \tfrac{L}{p!}\|x_k - v_k\|^p\left((5p+1)\|x_k - x\| + \|x_k - v_k\|\right), \quad \text{for all } x \in \mathcal{X}.\end{aligned}$$

This implies that (for all $x \in \mathcal{X}$)

$$\langle F(x_k), x_k - x\rangle \leq \tfrac{(5p+1)L}{p!}\|x_k - v_k\|^p\|x_k - x\| + \tfrac{L}{p!}\|x_k - v_k\|^{p+1}. \tag{7.24}$$

Then, we derive from the fact that $\|x_k - x\| \leq D$ and $\|x_k - v_k\| \leq D$ that

$$\langle F(x_k), x_k - x\rangle \leq \tfrac{(5p+2)LD}{p!}\|x_k - v_k\|^p, \quad \text{for all } x \in \mathcal{X}.$$

By the definition of a residue function (see Eq. (7.5)), we have

$$\begin{aligned}\textsc{res}(x_{k_T}) &= \sup_{x\in\mathcal{X}} \langle F(x_{k_T}), x_{k_T} - x\rangle \leq \tfrac{(5p+2)LD}{p!}\|x_{k_T} - v_{k_T}\|^p \\ &\overset{\text{Eq. (7.23)}}{\leq} \tfrac{(5p+2)LD}{p!}\left(\tfrac{4\|x^\star - x_0\|^2}{T}\right)^{\frac{p}{2}}.\end{aligned}$$

Since $x_0, x^\star \in \mathcal{X}$, we have $\|x^\star - x_0\| \leq D$ and hence

$$\textsc{res}(x_{k_T}) \leq \tfrac{2^p(5p+2)}{p!} LD^{p+1}T^{-\frac{p}{2}}. \tag{7.25}$$

Therefore, we conclude from Eq. (7.25)that we can set

$$T = O\left(\left(\frac{LD^{p+1}}{\epsilon}\right)^{\frac{2}{p}}\right),$$

such that $\hat{x} = \mathsf{Perseus}(p, x_0, L, T, 1)$ satisfies $\textsc{res}(\hat{x}) \leq \epsilon$. The total number of calls of the subproblem solvers is equal to T since our algorithm calls the subproblem solvers once at each iteration. This completes the proof.

Proof of Theorem 7.3.11. We first construct a hard function instance for any p^{th}-order method that satisfies Assumption 7.3.10. The basic function that we will use is as follows:

$$\eta(z, y) = \tfrac{1}{p}\sum_{i=1}^{d}(z^{(i)})^p \cdot y^{(i)}, \quad z \in \mathbb{R}^d_+,\ y \in \mathbb{R}^d_+.$$

Fixing $(z,y) \in \mathbb{R}^d_+ \times \mathbb{R}^d_+$, $(h_1,h_2) \in \mathbb{R}^d \times \mathbb{R}^d$ and $1 \le m+n \le p$, we have

$$\nabla^{(m,n)}\eta(z,y)[h_1]^m[h_2]^n = \tfrac{(p-1)!}{(p-m)!} \cdot \begin{cases} \sum_{i=1}^d (z^{(i)})^{p-m} y^{(i)} (h_1^{(i)})^m, & \text{if } n=0, \\ \sum_{i=1}^d (z^{(i)})^{p-m} (h_1^{(i)})^m h_2^{(i)}, & \text{if } n=1, \\ 0, & \text{otherwise.} \end{cases} \tag{7.26}$$

Note that $T \ge 1$ is an integer-valued parameter and $d \ge 2T+1$. We now define the following $2T \times 2T$ triangular matrix with two nonzero diagonals [Nesterov, 2021b]:

$$U = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ & \cdots & \cdots & \cdots & \\ 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \quad U^{-1} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 1 & \cdots & 1 \\ & \cdots & \cdots & \cdots & \\ 0 & 0 & \cdots & 1 & 1 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \quad U^\top = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 & 1 & \cdots & 0 & 0 \\ & \cdots & \cdots & \cdots & \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & -1 & 1 \end{bmatrix}.$$

Now, we introduce $d \times d$ upper triangular matrix A with the following structure:

$$A = \begin{bmatrix} U & 0 \\ 0 & I_{d-2T} \end{bmatrix}.$$

We are now ready to characterize a novel hard function and the corresponding two constraint sets:

$$f(z,y) = \tfrac{L}{2^{p+1}p!}\left(\eta(Az,y) - \tfrac{1}{p(p+1)}\sum_{i=2}^{2T}(y^{(i)})^{p+1} - (z^{(1)} - 2T + \tfrac{1}{p})\cdot y^{(1)}\right),$$

$$\mathcal{Z} = \left\{ z \in \mathbb{R}^d : \begin{array}{c} 0 \le z^{(i)} \le 2T-i+1 \text{ and } z^{(i+1)} \le z^{(i)} \text{ for all } 1 \le i \le 2T \\ \text{and } z^{(i)} = 0 \text{ for all } i > 2T \end{array} \right\},$$

$$\mathcal{Y} = \{y \in \mathbb{R}^d : 0 \le y^{(i)} \le 1 \text{ for all } 1 \le i \le 2T \text{ and } y^{(i)} = 0 \text{ for all } i > 2T\}.$$

We can see that the function $f : \mathcal{Z} \times \mathcal{Y} \mapsto \mathbb{R}$ is convex in z and concave in y. The computation of an optimal saddle-point solution of $f(z,y)$ is equivalent to solving a monotone VI with $x = \begin{bmatrix} z \\ y \end{bmatrix}$ and

$$\begin{aligned} F(x) &= \begin{bmatrix} \nabla^{(1,0)} f(z,y) \\ -\nabla^{(0,1)} f(z,y) \end{bmatrix} \\ &= \tfrac{L}{2^{p+1}p!} \cdot \begin{bmatrix} A^\top \nabla^{(1,0)} \eta(Az,y) - y^{(1)} \cdot e_d^{(1)} \\ -\nabla^{(0,1)} \eta(Az,y) + \frac{1}{p}\sum_{i=2}^{2T}(y^{(i)})^p \cdot e_d^{(i)} + (z^{(1)} - 2T + \frac{1}{p}) \cdot e_d^{(1)} \end{bmatrix}. \end{aligned}$$

Step 1. We show that $F : \mathcal{Z} \times \mathcal{Y} \mapsto \mathbb{R}^{2d}$ is $(p-1)^{\text{th}}$-order smooth with a Lipschitz constant $L > 0$. Indeed, we have

$$\begin{aligned} &\|\nabla^{(p-1)}F(x) - \nabla^{(p-1)}F(x')\|_{\text{op}} \\ &\leq \sum_{m+n=p} \tfrac{p!}{m!n!}\|\nabla^{(m,n)}f(z,y) - \nabla^{(m,n)}f(z',y')\|_{\text{op}}. \end{aligned} \tag{7.27}$$

We let $h = (h_1, h_2) \in \mathbb{R}^d \times \mathbb{R}^d$ and consider two cases. For the case of $p = 2$, we see from the definition of $f(z,y)$ that

$$\nabla^{(m,n)}f(z,y)[h_1]^m[h_2]^n = \tfrac{L}{2^{p+1}p!} \cdot \begin{cases} \nabla^{(1,1)}\eta(Az,y)[Ah_1][h_2] - h_1^{(1)}h_2^{(1)}, & \text{if } m=1 \text{ and } n=1, \\ -\sum_{i=2}^{2T}(y^{(i)})(h_2^{(i)})^p, & \text{if } n = p, \\ \nabla^{(m,n)}\eta(Az,y)[Ah_1]^m[h_2]^n, & \text{otherwise.} \end{cases}$$

For the case of $p \geq 3$, we see from the definition of $f(z,y)$ that

$$\nabla^{(m,n)}f(z,y)[h_1]^m[h_2]^n = \tfrac{L}{2^{p+1}p!} \cdot \begin{cases} -(p-1)!\sum_{i=2}^{2T}(y^{(i)})(h_2^{(i)})^p, & \text{if } n = p, \\ \nabla^{(m,n)}\eta(Az,y)[Ah_1]^m[h_2]^n, & \text{otherwise.} \end{cases}$$

Based on the above two equations, we have

$$\begin{aligned} &\sum_{m+n=p} \tfrac{p!}{m!n!}\|\nabla^{(m,n)}f(z,y)[h_1]^m[h_2]^n - \nabla^{(m,n)}f(z',y')[h_1]^m[h_2]^n\|_{\text{op}} \\ &\leq \tfrac{L}{2^{p+1}} \cdot \Bigg(\sum_{m+n=p, m\geq 1} \tfrac{1}{m!n!}\|\nabla^{(m,n)}\eta(Az,y)[Ah_1]^m[h_2]^n - \nabla^{(m,n)}\eta(Az',y')[Ah_1]^m[h_2]^n\|_{\text{op}} \\ &\quad + \tfrac{1}{p}\left|\sum_{i=2}^{2T}(y^{(i)} - (y')^{(i)})(h_2^{(i)})^p\right| \Bigg) \\ &\overset{\text{Eq. (7.26)}}{\leq} \tfrac{L}{2^{p+1}p} \cdot \Bigg(\left|\sum_{i=1}^{d}(y^{(i)} - (y')^{(i)})((Ah_1)^{(i)})^p\right| + p\left|\sum_{i=1}^{d}((Az)^{(i)} - (Az')^{(i)})((Ah_1)^{(i)})^{p-1}h_2^{(i)}\right| \\ &\quad + \left|\sum_{i=2}^{2T}(y^{(i)} - (y')^{(i)})(h_2^{(i)})^p\right| \Bigg). \end{aligned}$$

By the Cauchy-Schwartz inequality and $\|A\| \leq 2$ (see Nesterov [2021b, Eq. (4.2)]), we have

$$\begin{aligned} &\sum_{m+n=p} \tfrac{p!}{m!n!}\|\nabla^{(m,n)}f(z,y) - \nabla^{(m,n)}f(z',y')\|_{\text{op}} \\ &\leq \sup_{\|h\|=1}\left\{\sum_{m+n=p}\tfrac{p!}{m!n!}\left|\nabla^{(m,n)}f(z,y)[h_1]^m[h_2]^n - \nabla^{(m,n)}f(z',y')[h_1]^m[h_2]^n\right|\right\} \\ &\leq \sup_{\|h\|=1}\left\{\tfrac{L}{2^{p+1}p}\cdot\left(2^p\|y-y'\|\|h_1\|^p + 2^p p\|z-z'\|\|h_1\|^{p-1}\|h_2\| + \|y-y'\|\|h_2\|^p\right)\right\} \\ &\leq \sup_{\|h\|=1}\left\{\tfrac{L}{2^{p+1}p}\cdot(2^p + 2^p p + 1)\|x-x'\|\|h\|^p\right\} \;\leq\; L\|x-x'\|. \end{aligned}$$

Plugging the above equation into Eq. (7.27) yields the desired result.

Step 2. We show that there exists an optimal solution $x_\star = (z_\star, y_\star) \in \mathcal{Z} \times \mathcal{Y}$ such that $F(x_\star) = \mathbf{0}_{2d}$ and compute the optimal value of $f(z_\star, y_\star)$. By the definition, we have $F(x_\star) = \mathbf{0}_{2d}$ is equivalent to the following statement:

$$\begin{cases} A^\top \nabla^{(1,0)} \eta(Az_\star, y_\star) - y_\star^{(1)} \cdot e_d^{(1)} = \mathbf{0}_d, \\ \nabla^{(0,1)} \eta(Az_\star, y_\star) - \frac{1}{p} \sum_{i=2}^{2T} (y_\star^{(i)})^p \cdot e_d^{(i)} - (z_\star^{(1)} - 2T + \frac{1}{p}) \cdot e_d^{(1)} = \mathbf{0}_d. \end{cases} \tag{7.28}$$

Note that

$$\begin{aligned} \nabla^{(1,0)} \eta(Az_\star, y_\star) &= \sum_{i=1}^{d} ((Az_\star)^{(i)})^{p-1} y_\star^{(i)} e_d^{(i)}, \\ \nabla^{(0,1)} \eta(Az_\star, y_\star) &= \frac{1}{p} \left(\sum_{i=1}^{d} ((Az_\star^{(i)}))^p e_d^{(i)} \right). \end{aligned}$$

We claim that an optimal solution $x_\star = (z_\star, y_\star)$ is given by

$$z_\star^{(i)} = \begin{cases} 2T - i + 1, & \text{if } 1 \le i \le 2T, \\ 0 & \text{otherwise.} \end{cases} \qquad y_\star^{(i)} = \begin{cases} 1, & \text{if } 1 \le i \le 2T, \\ 0 & \text{otherwise.} \end{cases} \tag{7.29}$$

Indeed, we can see from the definition of $\mathcal{Z} \times \mathcal{Y}$ that $(z_\star, y_\star) \in \mathcal{Z} \times \mathcal{Y}$ and the definition of A that

$$(Az_\star)^{(i)} = \begin{cases} 1, & \text{if } 1 \le i \le 2T, \\ 0 & \text{otherwise.} \end{cases}$$

This implies that

$$\nabla^{(1,0)} \eta(Az_\star, y_\star) = \sum_{i=1}^{2T} e_d^{(i)}, \quad \nabla^{(0,1)} \eta(Az_\star, y_\star) = \frac{1}{p} \left(\sum_{i=1}^{2T} e_d^{(i)} \right).$$

By the definition of A, we have $A^\top \nabla^{(1,0)} \eta(Az_\star, y_\star) = e_d^{(1)}$. Thus, we can verify that Eq. (7.28) holds true. As such, we conclude that the optimal solution $x_\star = (z_\star, y_\star)$ defined in Eq. (7.29) belongs to $\mathcal{Z} \times \mathcal{Y}$ and the optimal value is

$$\begin{aligned} f(z_\star, y_\star) &= \frac{L}{2^{p+1} p!} \cdot \left(\eta(Az_\star, y_\star) - \frac{1}{p(p+1)} \sum_{i=2}^{2T} (y_\star^{(i)})^{p+1} - (z_\star^{(1)} - 2T + \frac{1}{p}) \cdot y_\star^{(1)} \right) \\ &= \frac{L}{2^{p+1}(p+1)!} (2T - 1). \end{aligned}$$

This implies the desired result.

Step 3. We now investigate the dynamics of any p^{th}-order method under Assumption 7.3.10. For simplicity, we denote

$$\mathbb{R}^d_k = \{z \in \mathbb{R}^d : z^{(i)} = 0 \text{ for all } i = k+1, k+2, \ldots, d\}, \quad \text{for all } 1 \le k \le d-1.$$

Without loss of generality, we assume that $x_0 = \mathbf{0}_{2d}$ is the initial iterate. Then, we show that the iterates $\{(z_k, y_k)\}_{k\ge 0}$ generated by any p^{th}-order method under Assumption 7.3.10 satisfy

$$z_k \in \mathbb{R}^d_k \cap \mathcal{Z}, \quad \text{for all } 1 \le k \le T. \tag{7.30}$$

It is clear that $z_k \in \mathcal{Z}$ for all $1 \le k \le T$ since the p^{th}-order method is applied to solve the optimization problem of $\min_{z\in\mathcal{Z}} \max_{y\in\mathcal{Y}} f(z,y)$. Thus, it suffices to show that $z_k \in \mathbb{R}^d_k$ for all $1 \le k \le T$.

The key ingredient of our proof is to show that the inclusion $x = (z,y) \in \mathbb{R}^d_k \times \mathbb{R}^d_k$ with $1 \le k \le T-1$ implies that $S_F(x) \subseteq \mathbb{R}^d_{k+1} \times \mathbb{R}^d_{k+1}$. Since A is an upper triangular matrix and $z \in \mathbb{R}^d_k$, we have $Az \in \mathbb{R}^d_k$. Also, we have $y \in \mathbb{R}^d_k$. Note that Eq. (7.26) implies that

$$\nabla^{(m,n)}\eta(z,y)[h_1]^m[h_2]^n = \tfrac{(p-1)!}{(p-m)!} \cdot \begin{cases} \sum_{i=1}^d (z^{(i)})^{p-m}(h_1^{(i)})^m h_2^{(i)}, & \text{if } n = 1, \\ \sum_{i=1}^d (z^{(i)})^{p-m} y^{(i)} (h_1^{(i)})^m, & \text{if } n = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, we have

$$\begin{aligned} \tfrac{\partial}{\partial h_1}(\nabla^{(m,n)}\eta(Az,y)[Ah_1]^m[h_2]^n) &= \sum_{i=1}^k c_i^{(m,n)} A^\top e_d^{(i)} \in \mathbb{R}^d_{k+1}, \\ \tfrac{\partial}{\partial h_2}(\nabla^{(m,n)}\eta(Az,y)[h_1]^m[h_2]^{n-1}) &= \sum_{i=1}^k d_i^{(m,n)} e_d^{(i)} \in \mathbb{R}^d_k \subseteq \mathbb{R}^d_{k+1}. \end{aligned}$$

Let us compute $F(x)$ and $\nabla F(x)[h]$ explicitly. We have

$$F(x) = \tfrac{L}{2^{p+1}p!} \cdot \begin{bmatrix} \tfrac{\partial}{\partial h_1}(\nabla^{(1,0)}\eta(Az,y)[Ah_1]) - y^{(1)} e_d^{(1)} \\ -\tfrac{\partial}{\partial h_2}(\nabla^{(0,1)}\eta(Az,y)[h_2]) + \tfrac{1}{p}\sum_{i=2}^{2T}(y^{(i)})^p \cdot e_d^{(i)} + (z^{(1)} - 2T + \tfrac{1}{p}) e_d^{(1)} \end{bmatrix},$$

and

$$\nabla F(x)[h] = \tfrac{L}{2^{p+1}p!} \cdot \begin{bmatrix} \tfrac{\partial}{\partial h_1}(\nabla^{(2,0)}\eta(Az,y)[Ah_1]^2) + \tfrac{\partial}{\partial h_2}(\nabla^{(1,1)}\eta(Az,y)[Ah_1][h_2]) - h_2^{(1)} e_d^{(1)} \\ -\tfrac{\partial}{\partial h_1}(\nabla^{(1,1)}\eta(Az,y)[Ah_1][h_2]) + h_1^{(1)} e_d^{(1)} - \tfrac{\partial}{\partial h_2}(\nabla^{(0,2)}\eta(Az,y)[h_2]^2) + \sum_{i=2}^{2T}(y^{(i)})^{p-1} h_2^{(i)} e_d^{(i)} \end{bmatrix}.$$

This together with

$$\tfrac{\partial}{\partial h_1}(\nabla^{(m,n)}\eta(Az,y)[Ah_1]^m[h_2]^n), \ \tfrac{\partial}{\partial h_2}(\nabla^{(m,n)}\eta(Az,y)[h_1]^m[h_2]^{n-1}) \in \mathbb{R}^d_{k+1}$$

yields $F(x), \nabla F(x)[h] \in \mathbb{R}^d_{k+1} \times \mathbb{R}^d_{k+1}$. By using a similar argument, we have

$$\nabla^{(j)} F(x)[h]^j \in \mathbb{R}^d_{k+1} \times \mathbb{R}^d_{k+1}, \quad \text{for all } 0 \leq j \leq p-1.$$

Thus, we conclude that all the solutions of $\Phi_{a,\gamma}(h) = 0$ belong to $\mathbb{R}^d_{k+1} \times \mathbb{R}^d_{k+1}$ where $\Phi_{a,\gamma,m}(h) = a_0 F(x) + \sum_{i=1}^{p-1} a_i \nabla^{(i)} F(x)[h]^i + \gamma \|h\|^{m-1} h$ and $a \in \mathbb{R}^p$, $\gamma > 0$ and $m \geq 2$. In other words, the inclusion $x = (z, y) \in \mathbb{R}^d_k \times \mathbb{R}^d_k$ with $k \geq 1$ implies that $S_F(x) \subseteq \mathbb{R}^d_{k+1} \times \mathbb{R}^d_{k+1}$.

We are now ready to prove the desired result in Eq. (7.30). We apply an inductive argument. For $k = 0$, we have $(z_0, y_0) = x_0 = \mathbf{0}_{2d}$. Thus, we have

$$F(x_0) = \tfrac{L}{2^{p+1} p!} \cdot \begin{bmatrix} \mathbf{0}_d \\ -(2T - \frac{1}{p}) e_d^{(1)} \end{bmatrix}, \ \nabla^{(j)} F(x_0)[h]^j = \mathbf{0}_{2d}, \quad \text{for all } 1 \leq j \leq p-1.$$

This implies that all the solutions of $\Phi_{a,\gamma}(h) = 0$ belong to $\mathbb{R}^d_1 \times \mathbb{R}^d_1$ where $\Phi_{a,\gamma,m}(h) = a_0 F(x_0) + \sum_{i=1}^{p-1} a_i \nabla^{(i)} F(x_0)[h]^i + \gamma \|h\|^{m-1} h$ and $a \in \mathbb{R}^p$, $\gamma > 0$ and $m \geq 2$. So $S_F(x_0) \subseteq \mathbb{R}^d_1 \times \mathbb{R}^d_1$. Assumption 7.3.10 ensures that $x_1 \in x_0 + S_F(x_0)$. Putting these pieces together yields that $x_1 \in \mathbb{R}^d_1 \times \mathbb{R}^d_1$.

We have already shown that the inclusion $x = (z, y) \in \mathbb{R}^d_k \times \mathbb{R}^d_k$ with $1 \leq k \leq T$ implies that $S_F(x) \subseteq \mathbb{R}^d_{k+1} \times \mathbb{R}^d_{k+1}$. Combining this result with $x_1 \in \mathbb{R}^d_1 \times \mathbb{R}^d_1$ and Assumption 7.3.10 guarantees that $x_k \in \mathbb{R}^d_k \times \mathbb{R}^d_k$ for all $1 \leq k \leq T$. Thus, $z_k \in \mathbb{R}^d_k \cap \mathcal{Z}$ for all $1 \leq k \leq T$.

Step 4. We now compute a lower bound on $\max_{y \in \mathcal{Y}} f(z_k, y)$. Eq. (7.30) in **Step 3** implies that

$$\max_{y \in \mathcal{Y}} f(z_k, y) \geq \min_{z \in \mathbb{R}^d_k \cap \mathcal{Z}} \max_{y \in \mathcal{Y}} f(z, y) \geq \min_{z \in \mathbb{R}^d_T \cap \mathcal{Z}} \max_{y \in \mathcal{Y}} f(z, y), \quad \text{for all } 0 \leq k \leq T.$$

We claim that $\min_{z \in \mathbb{R}^d_T \cap \mathcal{Z}} \max_{y \in \mathcal{Y}} f(z_k, y) \geq \frac{L}{2^{p+1} p!} (T + \frac{T-1}{p+1})$. We let $\mathcal{W} = \{w \in \mathbb{R}^d : w^{(i)} \geq 0 \text{ for all } 1 \leq i \leq 2T \text{ and } w^{(i)} = 0 \text{ for all } i > 2T\}$ and derive that

$$\begin{aligned}
& \min_{z \in \mathbb{R}^d_T \cap \mathcal{Z}} \max_{y \in \mathcal{Y}} f(z, y) \\
\overset{w=Az}{\geq} & \min_{w \in \mathbb{R}^d_T \cap \mathcal{W}} \max_{y \in \mathcal{Y}} \tfrac{L}{2^{p+1} p!} \left(\eta(w, y) - \tfrac{1}{p(p+1)} \sum_{i=2}^{2T} (y^{(i)})^{p+1} - \left(\sum_{i=1}^{d} w^{(i)} - 2T + \tfrac{1}{p} \right) \cdot y^{(1)} \right) \\
= & \min_{w \in \mathbb{R}^d_T \cap \mathcal{W}} \max_{y \in \mathcal{Y}} \tfrac{L}{2^{p+1} p!} \left(\tfrac{1}{p} \sum_{i=1}^{d} (w^{(i)})^p \cdot y^{(i)} - \tfrac{1}{p(p+1)} \sum_{i=2}^{2T} (y^{(i)})^{p+1} - \left(\sum_{i=1}^{d} w^{(i)} - 2T + \tfrac{1}{p} \right) \cdot y^{(1)} \right).
\end{aligned}$$

We see from $w \in \mathbb{R}^d_T \cap \mathcal{W}$ that $w^{(i)} \geq 0$ for all $1 \leq i \leq T$ and $w^{(i)} = 0$ for all $T+1 \leq i \leq 2T$. Fixing $w \in \mathbb{R}^d_T \cap \mathcal{W}$, we have

$$\begin{aligned} & \max_{y \in \mathcal{Y}} \left(\tfrac{1}{p} \sum_{i=1}^{d} (w^{(i)})^p \cdot y^{(i)} - \tfrac{1}{p(p+1)} \sum_{i=2}^{2T} (y^{(i)})^{p+1} - \left(\sum_{i=1}^{d} w^{(i)} - 2T + \tfrac{1}{p} \right) \cdot y^{(1)} \right) \\ = \; & \max \left\{ \tfrac{1}{p} (w^{(1)})^p - \left(\sum_{i=1}^{T} w^{(i)} - 2T + \tfrac{1}{p} \right), 0 \right\} + \tfrac{1}{p} \sum_{i=2}^{T} (w^{(i)})^p \cdot \min\{w^{(i)}, 1\} \\ & - \tfrac{1}{p(p+1)} \sum_{i=2}^{T} (\min\{w^{(i)}, 1\})^{p+1}. \end{aligned}$$

The key observation is that the second and third terms are independent of $w^{(1)}$ on the right-hand side. We also have

$$\begin{aligned} & \min_{w^{(1)} \geq 0} \max \left\{ \tfrac{1}{p} (w^{(1)})^p - \left(\sum_{i=1}^{T} w^{(i)} - 2T + \tfrac{1}{p} \right), 0 \right\} \\ = \; & \min_{w^{(1)} \geq 0} \max \left\{ \tfrac{1}{p} (w^{(1)})^p - w^{(1)} - \sum_{i=2}^{T} w^{(i)} + 2T - \tfrac{1}{p}, 0 \right\} \; \geq \; \max \left\{ 2T - 1 - \sum_{i=2}^{T} w^{(i)}, 0 \right\}. \end{aligned}$$

For simplicity, we define the function $g(w)$ as follows,

$$g(w) = \max \left\{ 2T - 1 - \sum_{i=2}^{T} w^{(i)}, 0 \right\} + \tfrac{1}{p} \sum_{i=2}^{T} (w^{(i)})^p \cdot \min\{w^{(i)}, 1\} - \tfrac{1}{p(p+1)} \sum_{i=2}^{T} (\min\{w^{(i)}, 1\})^{p+1}.$$

By the symmetry of the function g, we obtain that $\min_{w \in \mathbb{R}^d_T \cap \mathcal{W}} g(w)$ is achieved by the point with the same value of $w^{(i)}$ for all $2 \leq i \leq T$. Then, it suffices to solve the following one-dimensional optimization problem:

$$\min_{\eta \geq 0} \; h(\eta) = \max\{2T - 1 - \eta(T-1), 0\} + \tfrac{T-1}{p} \eta^p \cdot \min\{\eta, 1\} - \tfrac{T-1}{p(p+1)} (\min\{\eta, 1\})^{p+1}.$$

For the case of $0 \leq \eta \leq 1$, we have

$$h(\eta) = 2T - 1 - \eta(T-1) + \tfrac{T-1}{p+1} \eta^{p+1} \geq T + \tfrac{T-1}{p+1}.$$

For the case of $1 \leq \eta \leq \frac{2T-1}{T-1}$, we have

$$h(\eta) = 2T - 1 - \eta(T-1) + \tfrac{T-1}{p} \eta^{p+1} - \tfrac{T-1}{p(p+1)} \geq T + \tfrac{T-1}{p+1}.$$

For the case of $\eta \geq \frac{2T-1}{T-1}$, we have $\eta^{p+1} \geq 1 + (p+1)(\eta - 1) \geq 1 + \frac{pT}{T-1}$. Then, we have

$$h(\eta) = \tfrac{T-1}{p} \eta^{p+1} - \tfrac{T-1}{p(p+1)} \geq \tfrac{T-1}{p} \left(1 + \tfrac{pT}{T-1} \right) - \tfrac{T-1}{p(p+1)} \geq T + \tfrac{T-1}{p+1}.$$

Putting these pieces together yields that

$$\min_{z \in \mathbb{R}^d_T \cap \mathcal{Z}} \max_{y \in \mathcal{Y}} \; f(z, y) \geq \tfrac{L}{2^{p+1} p!} \left(T + \tfrac{T-1}{p+1} \right),$$

which implies the desired result.

Final Step. Since the point $(z_\star, y_\star) \in \mathcal{Z} \times \mathcal{Y}$ is an optimal saddle-point solution, we have

$$\max_{y\in\mathcal{Y}} f(z_k, y) - \min_{z\in\mathcal{Z}} f(z, y_k) \geq \max_{y\in\mathcal{Y}} f(z_k, y) - f(z_\star, y_\star).$$

Combining the results from **Step 2** and **Step 4**, we have

$$\min_{0\leq k\leq T} \left\{ \max_{y\in\mathcal{Y}} f(z_k, y) - \min_{z\in\mathcal{Z}} f(z, y_k) \right\} \geq \tfrac{L}{2^{p+1}p!}(T - \tfrac{T}{p+1}) \overset{p\geq 2}{\geq} \tfrac{LT}{2^p(p+1)!}.$$

Note that we set $D_\mathcal{Z} = 4T^{3/2}$ and $D_\mathcal{Y} = 2\sqrt{T}$ (cf. the definition of $\mathcal{Z}$ and $\mathcal{Y}$) and have $D_\mathcal{Z} D_\mathcal{Y}^p = 2^{p+2}T^{(p+3)/2}$. Then, we have

$$\min_{0\leq k\leq T} \left\{ \max_{y\in\mathcal{Y}} f(z_k, y) - \min_{z\in\mathcal{Z}} f(z, y_k) \right\} \geq \left(\tfrac{1}{4^{p+1}(p+1)!} \right) L D_\mathcal{Z} D_\mathcal{Y}^p T^{-\frac{p+1}{2}}.$$

This completes the proof.

7.5 Conclusion

We have proposed and analyzed a new p^{th}-order method—Perseus—for finding a weak solution of smooth and monotone variational inequalities (VIs) when F is $(p-1)^{\text{th}}$-order L-smooth. All of our theoretical results are based on the standard assumption that the subproblem arising from a $(p-1)^{\text{th}}$-order Taylor expansion of F can be computed approximately in an efficient manner. For the case of $p \geq 2$, the best existing p^{th}-order methods can achieve a global rate of $O(\epsilon^{-2/(p+1)} \log\log(1/\epsilon))$ [Bullins and Lai, 2022, Lin and Jordan, 2023, Jiang and Mokhtari, 2022] but require a nontrivial line-search procedure at each iteration. Notably, the open question has been whether it is possible to design a simple and optimal high-order method that achieves a global rate of $O(\epsilon^{-2/(p+1)})$ while dispensing with line search.

Our results settle this open problem. Indeed, our method converges to a weak solution with a global rate of $O(\epsilon^{-2/(p+1)})$ and the restarted version can attain global linear and local superlinear convergence for smooth and strongly monotone VIs. We also prove a lower bound for monotone VIs under a standard linear span assumption, showing that our method is *optimal* in the monotone setting. Moreover, we prove a global rate of $O(\epsilon^{-2/p})$ for solving a class of smooth and nonmonotone VIs satisfying the Minty condition and extend the results under the strong Minty condition. Future research include the investigation of lower bounds for the structured nonmonotone setting with the Minty condition and the comparative study of various lower-order methods in high-order smooth VI problems; see Nesterov [2021a,c,d] for recent examples of such comparisons in convex optimization.

Part III

Other Structured Problems

Chapter 8

Efficient Entropic Regularized Optimal Transport

We present several new complexity results for the entropic regularized algorithms that approximately solve the optimal transport (OT) problem between two discrete probability measures with at most n atoms. First, we improve the complexity bound of a greedy variant of Sinkhorn, known as *Greenkhorn*, from $\widetilde{O}(n^2\varepsilon^{-3})$ to $\widetilde{O}(n^2\varepsilon^{-2})$. Notably, our result can match the best known complexity bound of Sinkhorn and help clarify why Greenkhorn significantly outperforms Sinkhorn in practice in terms of row/column updates as observed by Altschuler et al. [2017]. Second, we propose a new algorithm, which we refer to as *APDAMD* and which generalizes an adaptive primal-dual accelerated gradient descent (APDAGD) algorithm [Dvurechensky et al., 2018] with a prespecified mirror mapping ϕ. We prove that APDAMD achieves the complexity bound of $\widetilde{O}(n^2\sqrt{\delta}\varepsilon^{-1})$ in which $\delta > 0$ stands for the regularity of ϕ. In addition, we show by a counterexample that the complexity bound of $\widetilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\})$ proved for APDAGD before is invalid and give a refined complexity bound of $\widetilde{O}(n^{5/2}\varepsilon^{-1})$. Further, we develop a *deterministic* accelerated variant of Sinkhorn via appeal to estimated sequence and prove the complexity bound of $\widetilde{O}(n^{7/3}\varepsilon^{-4/3})$. As such, we see that accelerated variant of Sinkhorn outperforms Sinkhorn and Greenkhorn in terms of $1/\varepsilon$ and APDAGD and accelerated alternating minimization (AAM) [Guminov et al., 2021] in terms of n. Finally, we conduct the experiments on synthetic and real data and the numerical results show the efficiency of entropic regularized algorithms in practice.

8.1 Introduction

From its origins in the seminal works by Monge [1781] and Kantorovich [1942] respectively in the eighteenth and twentieth centuries, and through to present day, the optimal transport (OT) problem has played a *determinative* role in the theory of mathematics [Villani, 2009]. It also has found a wide range of applications in problem domains beyond the original setting in logistics. In the current era, the strong and increasing linkage between optimization

and machine learning has brought new applications of OT to the fore; [see, e.g., Nguyen, 2013, Cuturi and Doucet, 2014, Srivastava et al., 2015, Rolet et al., 2016, Peyré et al., 2016, Nguyen, 2016, Carrière et al., 2017, Arjovsky et al., 2017, Gulrajani et al., 2017, Courty et al., 2017, Srivastava et al., 2018, Dvurechenskii et al., 2018, Tolstikhin et al., 2018, Sommerfeld et al., 2019, Lin et al., 2019b, Ho et al., 2019]. In these data-driven applications, the focus is on the probability distributions underlying the OT formulation; indeed, these distributions are either empirical distributions which are obtained by placing unit masses at data points, or are probability models of a putative underlying data-generating process. The OT problem accordingly often has a direct inferential meaning — as the definition of an estimator [Dudley, 1969, Fournier and Guillin, 2015, Weed and Bach, 2019, Lei, 2020], the definition of a likelihood [Sommerfeld and Munk, 2018, Bernton et al., 2019, Blanchet and Murthy, 2019], or as the robust variant of an estimator [Blanchet et al., 2019, Paty and Cuturi, 2019, Balaji et al., 2020]. The key challenge is computational [Peyré and Cuturi, 2019]. Indeed, the underlying distributions generally involve high-dimensional data and complex models in machine learning (ML) applications.

We study the OT problem in a discrete setting where we assume that the target and source probability distributions each have at most n atoms. In this setting, the OT problem can be solved exactly using linear programming (LP) solver based on specialized interior-point methods [Pele and Werman, 2009, Lee and Sidford, 2014, van den Brand et al., 2021], reflecting the LP formulation of the OT problem. In this context, van den Brand et al. [2021] have provided a bunch of randomized interior-point algorithms with improved runtimes for solving linear programs with two-sided constraints, leading to a new OT algorithm based on the Laplacian system solvers that achieved the best known complexity bounds of $\tilde{O}(n^2)$. However, it does not provide an effective solution to large-scale machine learning problems in practice since efficient implementations of Laplacian approach are yet unknown. Furthermore, many combinatorial techniques give exact algorithms for the OT problem. Indeed, the Hungarian algorithm [Kuhn, 1955, 1956, Munkres, 1957] solves the assignment problem in $O(n^3)$ time while there are several combinatorial algorithms that can solve the OT problem exactly in $\tilde{O}(n^{2.5})$ time [Gabow and Tarjan, 1991, Orlin and Ahuja, 1992]. Combined with the scaling technique, the network simplex algorithms [Orlin et al., 1993, Orlin, 1997] can be used to solve the OT problem exactly in $\tilde{O}(n^3)$ time and Lahn et al. [2019] have recently developed a faster approximation algorithm for the OT problem via appeal to the modification of the algorithm developed in Gabow and Tarjan [1991]. However, computing the OT problem exactly results in an output that is *not* differentiable with respect to measures' locations or weights [Bertsimas and Tsitsiklis, 1997]. Moreover, OT suffers from the curse of dimensionality [Dudley, 1969, Fournier and Guillin, 2015] and is thus likely to be meaningless when used on samples from high-dimensional densities.

An alternative to solve the OT problem is a class of approximation algorithms based on the entropy regularization which has been investigated in optimization and transportation science long before [Sinkhorn, 1974, Schneider and Zenios, 1990, Kalantari and Khachiyan, 1996, Knight, 2008, Kalantari et al., 2008, Chakrabarty and Khanna, 2021]. It was Cuturi [2013] that popularized the use of entropy regularization for OT in the machine learning

community and then initiated a productive line of research where an entropic regularization was imposed to approximate the non-negative constraints in the original OT problem. The resulting problem is referred to as *entropic regularized OT* and the corresponding class of approximation algorithms are called *entropic regularized algorithms*. It is worth mentioning that the entropic regularized OT has many favorable properties that the OT does not enjoy, motivating us to study the computational efficiency of entropic regularized algorithms in this paper. More specifically, from a statistical point of view, the entropic regularized OT enjoys significantly better sample complexity that is polynomial in the dimension [Genevay et al., 2019, Mena and Niles-Weed, 2019, Chizat et al., 2020], demonstrating that adding an entropy regularization will reduce the curse of dimensionality. Even from a computational point of view, such regularization in OT leads to *Sinkhorn* which attains a first near-linear time guarantee for the OT problem [Cuturi, 2013, Altschuler et al., 2017, Dvurechensky et al., 2018], and also makes the problem differentiable with regards to distributions [Feydy et al., 2019]; hence, the entropic regularized algorithms are more easily applicable to deep learning applications [Courty et al., 2017, Cuturi et al., 2019, Balaji et al., 2020] as opposed to combinatorial algorithms. This point was highlighted in Dong et al. [2020] and further necessitated the development of faster entropic regularized algorithms. In this regard, the greedy variant of Sinkhorn – Greenkhorn – was proposed and shown to outperform Sinkhorn empirically [Altschuler et al., 2017]. However, a sizable gap exists here since the best known complexity bound of $\widetilde{O}(n^2\varepsilon^{-3})$ for Greenkhorn [Altschuler et al., 2017] is worse than that of $\widetilde{O}(n^2\varepsilon^{-2})$ for Sinkhorn [Dvurechensky et al., 2018].

Further progress has been made by adapting first-order optimization algorithms for the OT problem [Cuturi and Peyré, 2016, Genevay et al., 2016, Blondel et al., 2018, Dvurechensky et al., 2018, Altschuler et al., 2019, Guo et al., 2020, Guminov et al., 2021]. Among these approaches, two of representatives are an adaptive primal-dual accelerated gradient descent (APDAGD) algorithm [Dvurechensky et al., 2018] with the claimed complexity bound of $\widetilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\})$ and an accelerated alternating minimization (AAM) algorithm [Guminov et al., 2021] with the complexity bound of $\widetilde{O}(n^{5/2}\varepsilon^{-1})$. Moreover, there are several second-order optimization algorithms [Allen-Zhu et al., 2017, Cohen et al., 2017] which are adapted for the OT problem [Blanchet et al., 2018, Quanrud, 2019] and guaranteed to achieve the improved complexity bound of $\widetilde{O}(n^2\varepsilon^{-1})$. However, the aforementioned second-order algorithms do not provide effective solutions to large-scale machine learning problems due to the lack of efficient implementations in practice.

Contributions. Given the advantages of entropic regularization in OT, we focus in his paper the computational efficiency of a class of entropic regularized algorithms for the OT problem and our theoretical analysis lead to several improvements over the state-of-the-art algorithms in the literature. We summarize the contributions as follows:

1. We improve the complexity bound of Greenkhorn from $\widetilde{O}(n^2\varepsilon^{-3})$ to $\widetilde{O}(n^2\varepsilon^{-2})$, which matches the best existing bound of Sinkhorn. The proof techniques are new and different from that used in Dvurechensky et al. [2018] for analyzing Sinkhorn. In particular,

Greenkhorn only updates a single row or column at each iteration and quantifying the per-iteration progress is more difficult than the measurement in Sinkhorn.

2. We propose an adaptive primal-dual accelerated mirror descent (APDAMD) algorithm which generalizes APDAGD with a prespecified mirror mapping ϕ and prove that APDAMD achieves the complexity bound of $\widetilde{O}(n^2\sqrt{\delta}\varepsilon^{-1})$ where $\delta > 0$ refers to the regularity of ϕ w.r.t. ℓ_∞ norm. We show by a counterexample that the complexity bound of $\widetilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\})$ proved for APDAGD [Dvurechensky et al., 2018] is invalid and give a refined complexity bound of $\widetilde{O}(n^{5/2}\varepsilon^{-1})$ which is worse than the claimed bound in terms of n.

3. We propose a deterministic accelerated variant of Sinkhorn via appeal to an estimated sequence and prove the complexity bound of $\widetilde{O}(n^{7/3}\varepsilon^{-4/3})$. In particular, accelerated Sinkhorn consists in an exact minimization for main iterates accompanied by another sequence of iterates based on coordinate gradient updates and monotone search. Our results show that accelerated Sinkhorn outperforms Sinkhorn and Greenkhorn in terms of $1/\varepsilon$ and APDAGD and AAM in terms of n.

We note that a preliminary version with only the analysis for Greenkhorn and APDAMD has been accepted by ICML [Lin et al., 2019a]. After our conference paper was published, some new algorithms were developed for solving the OT problem [Jambulapati et al., 2019, Lahn et al., 2019]. In particular, Jambulapati et al. [2019] developed a dual extrapolation algorithm with the complexity bound $\widetilde{O}\left(n^2\varepsilon^{-1}\right)$ using an area-convex mapping [Sherman, 2017]. Despite the theoretically sound complexity bound, the lack of simplicity and ease-of-implementation make this algorithm less competitive with Sinkhorn and Greenkhorn which remain the baseline solution methods in practice [Flamary and Courty, 2017].

Different from the algorithm in Jambulapati et al. [2019], the combinatorial algorithm in Lahn et al. [2019] is a practical solution method for the OT problem. It is worth mentioning that the algorithm in Lahn et al. [2019] and other combinatorial algorithms, e.g., the Hungarian algorithm, outperform our algorithms in practice. This is in consistence with the observation in Dong et al. [2020] who pointed out that combinatorial algorithms can outperform entropic regularized algorithms in speed even the latter ones are asymptotically faster for OT (i.e., the case of large n). However, we believe our results are still valuable due to the importance of entropic regularized algorithms as mentioned before.

Notation. For $n \geq 2$, we let $[n]$ be the set $\{1, 2, \ldots, n\}$ and $\mathbb{R}^n_+$ be the set of all vectors in $\mathbb{R}^n$ with non-negative coordinates. The notation $\Delta^n = \{v \in \mathbb{R}^n_+ : \sum_{i=1}^n v_i = 1\}$ stands for a probability simplex in $n-1$ dimensions. For a vector $x \in \mathbb{R}^n$ and let $1 \leq p < +\infty$, the notation $\|x\|_p$ stands for the ℓ_p-norm and $\|x\|$ indicates an ℓ_2-norm. $\mathrm{diag}(x)$ is a diagonal matrix which has the vector x on its diagonal. $\mathbf{1}_n$ and $\mathbf{0}_n$ are n-dimensional vector with all components being 1 and 0. For a matrix $A \in \mathbb{R}^{n\times n}$, we denote $\mathrm{vec}(A)$ as the vector in $\mathbb{R}^{n^2}$ obtained from concatenating the rows and columns of A. The notation $\|A\|_{1\to 1}$ stands

for $\sup_{\|x\|_1=1} \|Ax\|_1$ and the notations $r(A) = A\mathbf{1}_n$ and $c(A) = A^\top \mathbf{1}_n$ stand for the row and column sums respectively. For a function f, the notation $\nabla_x f$ denotes a partial derivative with respect to x. For the dimension n and tolerance $\varepsilon > 0$, the notations $a = O(b(n, \varepsilon))$ and $a = \Omega(b(n, \varepsilon))$ indicate that $a \leq C_1 \cdot b(n, \varepsilon)$ and $a \geq C_2 \cdot b(n, \varepsilon)$ respectively where C_1 and C_2 are independent of n and ε. We also denote $a = \Theta(b(n, \varepsilon))$ iff $a = O(b(n, \varepsilon)) = \Omega(b(n, \varepsilon))$. Similarly, we denote $a = \widetilde{O}(b(n, \varepsilon))$ to indicate the previous inequality where C_1 depends on some logarithmic function of n and ε.

8.2 Preliminaries

We first present the linear programming (LP) representation of the optimal transport (OT) problem as well as a specification of an approximate transportation plan. We also present an entropic regularized variant of the OT problem and derive the dual form where the objective function is in the form of the logarithm of sum of exponents. Finally, we establish several properties of that dual form which are useful for the subsequent analysis.

Linear programming representation. According to Kantorovich [1942], the problem of approximating the OT distance is equivalent to solving the following linear programming (LP) problem:

$$\min_{X \in \mathbb{R}^{n \times n}} \langle C, X \rangle \quad \text{s.t. } X\mathbf{1}_n = r, X^\top \mathbf{1}_n = c, X \geq 0. \tag{8.1}$$

In the above formulation, X refers to the *transportation plan*, $C = (C_{ij}) \in \mathbb{R}_+^{n \times n}$ stands for a cost matrix with non-negative components, and $r \in \mathbb{R}^n$ and $c \in \mathbb{R}^n$ are two probability distributions in the simplex Δ^n.

We see from Eq. (8.1), that the OT problem is a LP with $2n$ equality constraints and n^2 variables and can be solved by the interior-point method; however, this method performs poorly on large-scale problems due to its high per-iteration computational cost. In general, the solution that the algorithms aim at achieving is an ε-approximate transportation plan $\widehat{X} \in \mathbb{R}_+^{n \times n}$ satisfying the marginal distribution constraints $\widehat{X}\mathbf{1}_n = r$ and $\widehat{X}^\top \mathbf{1}_n = c$ and the inequality given by

$$\langle C, \widehat{X} \rangle \leq \langle C, X^\star \rangle + \varepsilon.$$

Here $X^\star$ is defined as an optimal transportation plan for the OT problem. For simplicity, we respectively denote $\langle C, \hat{X} \rangle$ an *ε-approximate transportation cost* and $\hat{X}$ an *ε-approximate transportation plan* for the original problem. Formally, we have the following definition of ε-approximate transportation plan.

Definition 8.2.1 *The matrix $\hat{X} \in \mathbb{R}_+^{n \times n}$ is called an ε-approximate transportation plan if $\hat{X}\mathbf{1}_n = r$ and $\hat{X}^\top \mathbf{1}_n = c$ and the following inequality holds true,*

$$\langle C, \widehat{X} \rangle \leq \langle C, X^\star \rangle + \varepsilon.$$

where $X^\star$ is defined as an optimal transportation plan for the OT problem.

With this definition in mind, the goal of this paper is to study the OT problem from a computational point of view. Indeed, we hope to derive an improved complexity bound of the current state-of-the-art algorithms and seek new practical algorithms whose running time required to obtain an ε-approximate transportation plan has better dependence on $1/\varepsilon$ than the benchmark algorithms in the literature. The aforementioned new algorithms are favorable in the machine learning applications where high precision (ε is small) is necessary.

Entropic regularized OT and its dual form. Seeking another formulation for OT distance that is more amenable to computationally efficient algorithms, Cuturi [2013] proposed to solve an entropic regularized version of the OT problem in Eq. (8.1), which is given by

$$\min_{X\in\mathbb{R}^{n\times n}} \langle C, X\rangle - \eta H(X), \quad \text{s.t. } X\mathbf{1}_n = r, X^\top \mathbf{1}_n = c, \tag{8.2}$$

where $\eta > 0$ denotes the regularization parameter and $H(X)$ denotes the entropic regularization term, which is given by:

$$H(X) := -\langle X, \log(X) - \mathbf{1}_{n\times n}\rangle.$$

Note that, the optimal solution of the entropic regularized OT problem exists since the objective function $\langle C, X\rangle - \eta H(X)$ is continuous and the feasible region $\{X \in \mathbb{R}^{n\times n} : X \geq 0, X\mathbf{1}_n = r, X^\top \mathbf{1}_n = c\}$ is compact. Furthermore, that optimal solution is also unique since the objective function $\langle C, X\rangle - \eta H(X)$ is strongly convex over the feasible region with respect to ℓ_1-norm. However, the optimal value of the entropic regularized OT problem (cf. Eq (8.2)) yields a poor approximation to the unregularized OT problem if η is large. An additional issue of entropic regularization is that the sparsity of the solution is lost. Even though an ε-approximate transportation plan can be found efficiently, it is not clear how different the sparsity pattern of this solution is with respect to the solution of the actual OT problem. In contrast, the actual OT distance suffers from the curse of dimensionality [Dudley, 1969, Fournier and Guillin, 2015, Weed and Bach, 2019] and is significantly worse than its entropic regularized version in terms of the sample complexity [Genevay et al., 2019, Mena and Niles-Weed, 2019, Chizat et al., 2020].

While there is an ongoing debate in the literature on the merits of solving the OT problem *v.s.* its entropic regularized version, we adopt here the viewpoint that reaching an additive approximation of the actual OT cost matters and therefore propose to scale η as a function of the desired accuracy of the approximation. Then, we proceed to derive the dual form of the entropic regularized OT problem in Eq. (8.2) and show that it remains an unconstrained smooth optimization problem. By introducing the dual variables $\alpha, \beta \in \mathbb{R}^n$, we define the Lagrangian function of the entropic regularized OT problem as follows:

$$\mathcal{L}(X, \lambda_1, \ldots, \lambda_m) = \langle C, X\rangle - \eta H(X) - \alpha^\top (X\mathbf{1}_n - r) - \beta^\top (X^\top \mathbf{1}_n - c). \tag{8.3}$$

In order to derive the smooth dual objective function, we consider the following minimization problem:

$$\min_{X:\|X\|_1=1} \langle C, X\rangle - \eta H(X) - \alpha^\top (X\mathbf{1}_n - r) - \beta^\top (X^\top \mathbf{1}_n - c).$$

The above objective function is strongly convex over the domain $\{X \in \mathbb{R}_+^{n\times n} \mid \|X\|_1 = 1\}$. Thus, the optimal solution is unique. After the simple calculations, the optimal solution $\bar{X} = X(\alpha, \beta)$ has the following form:

$$\bar{X}_{ij} = \frac{e^{\eta^{-1}(\alpha_i+\beta_j-C_{ij})}}{\sum_{1\le i,j\le n} e^{\eta^{-1}(\alpha_i+\beta_j-C_{ij})}}. \tag{8.4}$$

Plugging Eq. (8.4) into Eq. (8.3) yields that the dual form is:

$$\max_{\alpha,\beta} \left\{ -\eta \log \left(\sum_{1\le i,j\le n} e^{\eta^{-1}(\alpha_i+\beta_j-C_{ij})} \right) + \alpha^\top r + \beta^\top c \right\}.$$

In order to streamline our presentation, we perform a change of variables, $u = \eta^{-1}\alpha$ and $v = \eta^{-1}\beta$, and reformulate the above problem as

$$\min_{\alpha,\beta} \ \varphi(\alpha, \beta) := \log \left(\sum_{1\le i,j\le n} e^{u_i+v_j-\frac{C_{ij}}{\eta}} \right) - u^\top r - v^\top c.$$

To further simplify the notation, we define $B(u, v) := (B_{ij})_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$ by

$$B_{ij} = e^{u_i+v_j-\frac{C_{ij}}{\eta}}.$$

To this end, we obtain the *dual entropic regularized OT problem* defined by

$$\min_{u,v} \ \varphi(u, v) := \log(\|B(u, v)\|_1) - u^\top r - v^\top c. \tag{8.5}$$

Remark 8.2.2 *The first part of the objective function φ is in the form of the logarithm of sum of exponents while the second part is a linear function. This is different from the objective function used in previous dual entropic regularized OT problem [Cuturi, 2013, Altschuler et al., 2017, Dvurechensky et al., 2018, Lin et al., 2019a]. Notably, Eq. (8.5) is a special instance of a softmax minimization problem, and the objective function φ is known to be smooth [Nesterov, 2005]. Finally, we point out that the same formulation has been derived in Guminov et al. [2021] for analyzing AAM.*

In the remainder of the paper, we also denote $(u^\star, v^\star) \in \mathbb{R}^{2n}$ as an optimal solution of the dual entropic regularized OT problem in Eq. (8.5).

Properties of dual entropic regularized OT. We present several useful properties of the dual entropic regularized OT in Eq. (8.5). In particular, we show that there exists an optimal solution $(u^\star, v^\star) \in \mathbb{R}^{2n}$ such that it has an upper bound in terms of the ℓ_∞-norm.

Lemma 8.2.3 *For the dual entropic regularized OT problem in Eq. (8.5), there exists an optimal solution $(u^\star, v^\star)$ such that*

$$\|u^\star\|_\infty \le R, \qquad \|v^\star\|_\infty \le R,$$

where $R := \eta^{-1}\|C\|_\infty + \log(n) - \log(\min_{1\le i,j\le n}\{r_i, c_j\})$ depends on C, r and c.

Proof. First, we claim that there exists an optimal solution $(u^\star, v^\star)$ such that

$$\max_{1\le i\le n} u_i^\star \ge 0 \ge \min_{1\le i\le n} u_i^\star, \qquad \max_{1\le i\le n} v_i^\star \ge 0 \ge \min_{1\le i\le n} v_i^\star. \tag{8.6}$$

Indeed, letting $(\widehat{u}^\star, \widehat{v}^\star)$ be an optimal solution to Eq. (8.5), the claim holds true if $(\widehat{u}^\star, \widehat{v}^\star)$ satisfies Eq. (8.6). Otherwise, we define the shift term given by

$$\widehat{\Delta}_u = \frac{\max_{1\le i\le n}\widehat{u}_i^\star + \min_{1\le i\le n}\widehat{u}_i^\star}{2}, \quad \widehat{\Delta}_v = \frac{\max_{1\le i\le n}\widehat{v}_i^\star + \min_{1\le i\le n}\widehat{v}_i^\star}{2},$$

and define $(u^\star, v^\star)$ by

$$u^\star = \widehat{u}^\star - \widehat{\Delta}_u \mathbf{1}_n, \qquad v^\star = \widehat{v}^\star - \widehat{\Delta}_v \mathbf{1}_n.$$

By definition, we have $(u^\star, v^\star)$ satisfies Eq. (8.6). Since $\mathbf{1}_n^\top r = \mathbf{1}_n^\top c = 1$, we have $(u^\star)^\top r = (\widehat{u}^\star)^\top r - \widehat{\Delta}_u$ and $(v^\star)^\top c = (\widehat{v}^\star)^\top c - \widehat{\Delta}_v$. In addition, $\log(\|B(u^\star, v^\star)\|_1) = \log(\|B(\widehat{u}^\star, \widehat{v}^\star)\|_1) + \widehat{\Delta}_u + \widehat{\Delta}_v$. Putting these pieces together yields $\varphi(u^\star, v^\star) = \varphi(\widehat{u}^\star, \widehat{v}^\star)$. Therefore, $(u^\star, v^\star)$ is an optimal solution of the dual entropic regularized OT that satisfies Eq. (8.6).

Then, we show that

$$\max_{1\le i\le n} u_i^\star - \min_{1\le i\le n} u_i^\star \le \frac{\|C\|_\infty}{\eta} - \log\left(\min_{1\le i,j\le n}\{r_i, c_j\}\right), \tag{8.7}$$

$$\max_{1\le i\le n} v_i^\star - \min_{1\le i\le n} v_i^\star \le \frac{\|C\|_\infty}{\eta} - \log\left(\min_{1\le i,j\le n}\{r_i, c_j\}\right). \tag{8.8}$$

Indeed, for any $1 \le i \le n$, we derive from the optimality condition of $(u^\star, v^\star)$ that

$$\frac{e^{u_i^\star}\left(\sum_{j=1}^n e^{v_j^\star - \eta^{-1}C_{ij}}\right)}{\|B(u^\star, v^\star)\|_1} = r_i, \quad \text{for all } i \in [n].$$

Since $C_{ij} \ge 0$ for all $1 \le i, j \le n$ and $r_i \ge \min_{1\le i,j\le n}\{r_i, c_j\}$ for all $1 \le i \le n$, we have

$$u_i^\star \ge \log\left(\min_{1\le i,j\le n}\{r_i, c_j\}\right) - \log\left(\sum_{j=1}^n e^{v_j^\star}\right) + \log(\|B(u^\star, v^\star)\|_1), \quad \text{for all } i \in [n].$$

Since $0 < r_i \le 1$ and $C_{ij} \le \|C\|_\infty$, we have

$$u_i^\star \le \frac{\|C\|_\infty}{\eta} - \log\left(\sum_{j=1}^n e^{v_j^\star}\right) + \log(\|B(u^\star, v^\star)\|_1), \quad \text{for all } i \in [n].$$

Putting these pieces together yields Eq. (8.7). By the similar argument, we can prove Eq. (8.8).

Finally, we prove our main results. Indeed, Eq. (8.6) and Eq. (8.7) imply that

$$-\tfrac{\|C\|_\infty}{\eta} + \log\left(\min_{1\le i,j\le n}\{r_i, c_j\}\right) \le \min_{1\le i\le n} u_i^\star \le 0,$$

and

$$0 \le \max_{1\le i\le n} u_i^\star \le \tfrac{\|C\|_\infty}{\eta} - \log\left(\min_{1\le i,j\le n}\{r_i, c_j\}\right).$$

Combining the above two inequalities with the definition of R implies that $\|u^\star\|_\infty \le R$. By the similar argument, we can prove that $\|v^\star\|_\infty \le R$. As a consequence, we obtain the conclusion of the lemma. □

The upper bound for the ℓ_∞-norm of an optimal solution of dual entropic regularized OT in Lemma 8.2.3 directly leads to the following direct bound for the ℓ_2-norm.

Corollary 8.2.4 *For the dual entropic regularized OT problem in Eq. (8.5), there exists an optimal solution $(u^\star, v^\star)$ such that*

$$\|u^\star\| \le \sqrt{n}R, \qquad \|v^\star\| \le \sqrt{n}R,$$

where $R > 0$ is defined in Lemma 8.2.3.

Since the function $-H(X)$ is strongly convex with respect to the ℓ_1-norm on the probability simplex $Q \subseteq \mathbb{R}^{n\times n}$, the entropic regularized OT problem in Eq. (8.2) is a special case of the following linearly constrained convex optimization problem:

$$\min_{x\in Q}\ f(x), \quad \text{s.t. } Ax = b,$$

where f is strongly convex with respect to the ℓ_1-norm on the set Q:

$$f(x') - f(x) - (x'-x)^\top \nabla f(x) \ge \tfrac{\eta}{2}\|x'-x\|_1^2 \text{ for any } x', x \in Q.$$

By Nesterov [2005, Theorem 1] with the ℓ_2-norm for the dual space of the Lagrange multipliers, the dual objective function $\tilde{\varphi}$ satisfies the following inequality:

$$\widetilde{\varphi}(\alpha',\beta') - \widetilde{\varphi}(\alpha,\beta) - \begin{pmatrix}\alpha'-\alpha\\ \beta'-\beta\end{pmatrix}^\top \nabla\widetilde{\varphi}(\alpha,\beta) \le \tfrac{\|A\|_{1\to 2}^2}{2\eta}\left\|\begin{pmatrix}\alpha'-\alpha\\ \beta'-\beta\end{pmatrix}\right\|^2 \text{ for any } (\alpha',\beta'),(\alpha,\beta) \in \mathbb{R}^{2n}.$$

Recall that the function $\tilde{\varphi}$ is given by

$$\widetilde{\varphi}(\alpha,\beta) = -\eta\log\left(\sum_{1\le i,j\le n} e^{\eta^{-1}(\alpha_i+\beta_j-C_{ij})}\right) + \alpha^\top r + \beta^\top c. \tag{8.9}$$

Algorithm 22 GREENKHORN$(C, \eta, r, c, \varepsilon')$

Input: $t = 0$ and $u^0 = v^0 = \mathbf{0}_n$.
while $E_t > \varepsilon'$ **do**
  Compute $I = \operatorname{argmax}_{1 \le i \le n} \rho(r_i, r_i(B(u^t, v^t)))$ where $\rho(a, b) = b - a + a\log(a/b)$ and $(B(u^t, v^t))_{i'j'} = e^{u^t_{i'} + v^t_{j'} - \frac{C_{i'j'}}{\eta}}$ for all (i', j').
  Compute $J = \operatorname{argmax}_{1 \le j \le n} \rho(c_j, c_j(B(u^t, v^t)))$.
  if $\rho(r_i, r_i(B(u^t, v^t))) > \rho(c_j, c_j(B(u^t, v^t)))$ **then**
    $u^{t+1}_I = u^t_I + \log(r_I) - \log(r_I(B(u^t, v^t)))$.
  else
    $v^{t+1}_J = v^t_J + \log(c_J) - \log(c_J(B(u^t, v^t)))$.
  Increment by $t = t + 1$.
Output: $B(u^t, v^t)$.

We notice that the function φ in Eq. (8.5) satisfies that $\varphi(u, v) = -\eta^{-1}\widetilde{\varphi}(\eta u, \eta v)$. After some simple calculations, we have

$$\varphi(u', v') - \varphi(u, v) - \begin{pmatrix} u' - u \\ v' - v \end{pmatrix}^\top \nabla\varphi(u, v) \le \left(\frac{\|A\|^2_{1\to 2}}{2}\right) \left\| \begin{pmatrix} u' - u \\ v' - v \end{pmatrix} \right\|^2. \tag{8.10}$$

In the entropic regularized OT problem, each column of the matrix A contains no more than two nonzero elements which are equal to one. Since $\|A\|_{1\to 2}$ is equal to maximum ℓ_2-norm of the column of this matrix, we have $\|A\|_{1\to 2} = \sqrt{2}$. Thus, the dual objective function φ is 2-gradient Lipschitz with respect to the ℓ_2-norm.

8.3 Greenkhorn

We present a complexity analysis for Greenkhorn. In particular, we improve the existing best known complexity bound $O(n^2\|C\|^3_\infty \log(n)\varepsilon^{-3})$ [Altschuler et al., 2017] to $O(n^2\|C\|^2_\infty \log(n)\varepsilon^{-2})$, which matches the current state-of-the-art complexity bound for Sinkhorn [Dvurechensky et al., 2018].

To facilitate the subsequent discussion, we present the pseudocode of Greenkhorn in Algorithm 22 and its application to regularized OT in Algorithm 23. The function for quantifying the progress in the dual objective value between two consecutive iterates is given by $\rho(a, b) = b - a + a\log(a/b)$ and we recall that (u, v) is an optimal solution of the dual entropic regularized OT problem in Eq. (8.5) if $r(B(u, v)) - r = \mathbf{0}_n$ and $c(B(u, v)) - c = \mathbf{0}_n$. This leads to the quantity which measures the error of the t-th iterate in Algorithm 22:

$$E_t := \|r(B(u^t, v^t)) - r\|_1 + \|c(B(u^t, v^t)) - c\|_1.$$

Both Sinkhorn and Greenkhorn can be interpreted as coordinate descent for minimizing the following convex function [Cuturi, 2013, Altschuler et al., 2017, Dvurechensky et al., 2018,

Algorithm 23 Approximating OT by Algorithm 22

Input: $\eta = \frac{\varepsilon}{4\log(n)}$ and $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$.
Step 1: Let $\tilde{r} \in \Delta_n$ and $\tilde{c} \in \Delta_n$ be defined by $(\tilde{r}, \tilde{c}) = (1 - \frac{\varepsilon'}{8})(r, c) + \frac{\varepsilon'}{8n}(\mathbf{1}_n, \mathbf{1}_n)$.
Step 2: Compute $\tilde{X} = \textsc{Greenkhorn}(C, \eta, \tilde{r}, \tilde{c}, \frac{\varepsilon'}{2})$.
Step 3: Round $\tilde{X}$ to $\hat{X}$ using Altschuler et al. [2017, Algorithm 2] such that $\hat{X}\mathbf{1}_n = r$ and $\hat{X}^\top \mathbf{1}_n = c$.
Output: $\hat{X}$.

Lin et al., 2019a]:

$$f(u, v) := \|B(u, v)\|_1 - u^\top r - v^\top c. \tag{8.11}$$

Comparing to the scheme of Sinkhorn that consists in the updates of *all* rows and columns, Algorithm 22 updates only *one* row or column at each step. As such, Algorithm 22 updates only $O(n)$ entries per iteration rather than $O(n^2)$ in Sinkhorn. It is also worth mentioning that Algorithm 22 can be implemented such that each iteration runs in only $O(n)$ arithmetic operations [Altschuler et al., 2017].

Despite cheap per-iteration computational cost, it is difficult to quantify the per-iteration progress of Algorithm 22 and the proof techniques for Sinkhorn in Dvurechensky et al. [2018] are not applicable here. This motivates us to investigate another proof strategy which will be elaborated in the sequel.

Complexity analysis—bounding dual objective values. Given the definition of E_t, we first prove the following lemma which yields an upper bound for the objective values of the iterates.

Lemma 8.3.1 *Letting $\{(u^t, v^t)\}_{t\geq 0}$ be the iterates generated by Algorithm 22, we have*

$$f(u^t, v^t) - f(u^\star, v^\star) \leq 2E_t(\|u^\star\|_\infty + \|v^\star\|_\infty),$$

where (u^*, v^*) is a point that minimizes $f(u, v) = \|B(u, v)\|_1 - u^\top r - v^\top c$.

Proof. By the definition, we have

$$f(u, v) = \sum_{1\leq i,j\leq n} e^{u_i + v_j - \frac{C_{ij}}{\eta}} - \sum_{i=1}^n u_i r_i - \sum_{j=1}^n v_j c_j.$$

By definition, we have $\nabla_u f(u^t, v^t) = B(u^t, v^t)\mathbf{1}_n - r$ and $\nabla_v f(u^t, v^t) = B(u^t, v^t)^\top \mathbf{1}_n - c$. Thus, we have $E_t = \|\nabla_u f(u^t, v^t)\|_1 + \|\nabla_v f(u^t, v^t)\|_1$. Since f is convex and minimized at $(u^\star, v^\star)$, we have

$$f(u^t, v^t) - f(u^\star, v^\star) \leq (u^t - u^\star)^\top \nabla_u f(u^t, v^t) + (v^t - v^\star)^\top \nabla_v f(u^t, v^t).$$

Combining Hölder's inequality and the definition of E_t yields

$$f(u^t, v^t) - f(u^\star, v^\star) \leq E_t(\|u^t - u^\star\|_\infty + \|v^t - v^\star\|_\infty). \tag{8.12}$$

Thus, it suffices to show that

$$\|u^t - u^\star\|_\infty + \|v^t - v^\star\|_\infty \leq 2\|u^\star\|_\infty + 2\|v^\star\|_\infty.$$

The next result is the key observation that makes our analysis work for Greenkhorn. We use an induction argument to establish the following bound:

$$\max\{\|u^t - u^\star\|_\infty, \|v^t - v^\star\|_\infty\} \leq \max\{\|u^0 - u^\star\|_\infty, \|v^0 - v^\star\|_\infty\}. \tag{8.13}$$

It is clear that Eq. (8.13) holds true when $t = 0$. Suppose that the inequality holds true for $t \leq k_0$, we show that it also holds true for $t = k_0 + 1$. Without loss of generality, let I be the index chosen at the $(k_0 + 1)$-th iteration. Then

$$\|u^{k_0+1} - u^\star\|_\infty \leq \max\{\|u^{k_0} - u^\star\|_\infty, |u_I^{k_0+1} - u_I^\star|\}, \tag{8.14}$$
$$\|v^{k_0+1} - v^\star\|_\infty = \|v^{k_0} - v^\star\|_\infty. \tag{8.15}$$

By the updating formula for $u_I^{k_0+1}$ and the optimality condition for $u_I^\star$, we have

$$e^{u_I^{k_0+1}} = \frac{r_I}{\sum_{j=1}^n e^{-\frac{C_{ij}}{\eta} + v_j^{k_0}}}, \quad e^{u_I^\star} = \frac{r_I}{\sum_{j=1}^n e^{-\frac{C_{ij}}{\eta} + v_j^\star}}.$$

Putting these pieces together with the inequality that $\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \leq \max_{1\leq j\leq n} \frac{a_i}{b_i}$ for all $a_i, b_i > 0$ yields

$$|u_I^{k_0+1} - u_I^\star| = \left|\log\left(\frac{\sum_{j=1}^n e^{-\eta^{-1}C_{Ij} + v_j^{k_0}}}{\sum_{j=1}^n e^{-\eta^{-1}C_{Ij} + v_j^\star}}\right)\right| \leq \|v^{k_0} - v^\star\|_\infty. \tag{8.16}$$

Combining Eq. (8.14) and Eq. (8.16) yields

$$\|u^{k_0+1} - u^\star\|_\infty \leq \max\{\|u^{k_0} - u^\star\|_\infty, \|v^{k_0} - v^\star\|_\infty\}. \tag{8.17}$$

Combining Eq. (8.15) and Eq. (8.17) further implies Eq. (8.13). This together with $u^0 = v^0 = \mathbf{0}_n$ implies

$$\|u^t - u^\star\|_\infty + \|v^t - v^\star\|_\infty \leq 2(\|u^0 - u^\star\|_\infty + \|v^0 - v^\star\|_\infty) = 2\|u^\star\|_\infty + 2\|v^\star\|_\infty. \tag{8.18}$$

Putting Eq. (8.12) and Eq. (8.18) together yields the desired result. □

Our second lemma shows that at least one optimal solution $(u^\star, v^\star)$ of f has an upper bound of $\eta^{-1}\|C\|_\infty + \log(n) - 2\log(\min_{1\leq i,j\leq n}\{r_i, c_j\})$ in ℓ_∞-norm. This result is stronger than Dvurechensky et al. [2018, Lemma 1] and generalizes Blanchet et al. [2018, Lemma 10].

Lemma 8.3.2 *There exists an optimal solution $(u^\star, v^\star)$ of the function f defined in Eq. (8.11) such that the following inequality holds true,*

$$\|u^\star\|_\infty \leq R, \qquad \|v^\star\|_\infty \leq R,$$

where $R := \eta^{-1}\|C\|_\infty + \log(n) - 2\log(\min_{1\leq i,j\leq n}\{r_i, c_j\})$ depends on C, r and c.

Proof. By using the similar argument as in Lemma 8.2.3, we can first show that there exists an optimal solution pair $(u^\star, v^\star)$ such that (but not for $v^\star$ simultaneously)

$$\max_{1\leq i\leq n} u_i^\star \geq 0 \geq \min_{1\leq i\leq n} u_i^\star. \tag{8.19}$$

Then, we proceed to establish the bounds that are analogous to Eq. (8.7) and (8.8):

$$\max_{1\leq i\leq n} u_i^\star - \min_{1\leq i\leq n} u_i^\star \leq \tfrac{\|C\|_\infty}{\eta} - \log\left(\min_{1\leq i,j\leq n}\{r_i, c_j\}\right), \tag{8.20}$$

$$\max_{1\leq i\leq n} v_i^\star - \min_{1\leq i\leq n} v_i^\star \leq \tfrac{\|C\|_\infty}{\eta} - \log\left(\min_{1\leq i,j\leq n}\{r_i, c_j\}\right). \tag{8.21}$$

Indeed, for each $1 \leq i \leq n$, we have

$$e^{-\eta^{-1}\|C\|_\infty + u_i^\star}\left(\sum_{j=1}^n e^{v_j^\star}\right) \leq \sum_{j=1}^n e^{-\eta^{-1}C_{ij}+u_i^\star+v_j^\star} = [B(u^\star, v^\star)\mathbf{1}_n]_i = r_i \leq 1,$$

which implies $u_i^\star \leq \eta^{-1}\|C\|_\infty - \log(\sum_{j=1}^n e^{v_j^\star})$. Furthermore, we have

$$e^{u_i^\star}\left(\sum_{j=1}^n e^{v_j^\star}\right) \geq \sum_{j=1}^n e^{-\eta^{-1}C_{ij}+u_i^\star+v_j^\star} = [B(u^\star, v^\star)\mathbf{1}_n]_i = r_i \geq \min_{1\leq i,j\leq n}\{r_i, c_j\},$$

which implies $u_i^\star \geq \log(\min_{1\leq i,j\leq n}\{r_i, c_j\}) - \log(\sum_{j=1}^n e^{v_j^\star})$. Putting these pieces together yields Eq. (8.20). Using the similar argument, we can prove Eq. (8.21) holds true.

Finally, we prove our main results. Since $\max_{1\leq i\leq n} u_i^\star \geq 0 \geq \min_{1\leq i\leq n} u_i^\star$, we derive from Eq. (8.20) that

$$-\tfrac{\|C\|_\infty}{\eta} + \log\left(\min_{1\leq i,j\leq n}\{r_i, c_j\}\right) \leq \min_{1\leq i\leq n} u_i^\star \leq \max_{1\leq i\leq n} u_i^\star \leq \tfrac{\|C\|_\infty}{\eta} - \log\left(\min_{1\leq i,j\leq n}\{r_i, c_j\}\right).$$

This implies that $\|u^\star\|_\infty \leq R$. Then, we bound $\|v^\star\|_\infty$ by considering two different cases.

For the former case, we assume that $\max_{1\leq i\leq n} v_i^\star \geq 0$. Note that the optimality condition is $\sum_{i,j=1}^n e^{-\eta^{-1}C_{ij}+u_i^\star+v_j^\star} = 1$ and further implies that

$$\max_{1\leq i\leq n} u_i^\star + \max_{1\leq i\leq n} v_i^\star \leq \log\left(\max_{1\leq i,j\leq n} e^{\eta^{-1}C_{ij}}\right) = \tfrac{\|C\|_\infty}{\eta}.$$

Since $\max_{1\le i\le n} u_i^* \ge 0$ and $\max_{1\le i\le n} v_i^* \ge 0$, we have $0 \le \max_{1\le i\le n} v_i^\star \le \frac{\|C\|_\infty}{\eta}$. Combining $\max_{1\le i\le n} v_i^* \ge 0$ with Eq. (8.21) yields that

$$\min_{1\le i\le n} v_i^\star \ge -\tfrac{\|C\|_\infty}{\eta} + \log\left(\min_{1\le i,j\le n}\{r_i, c_j\}\right).$$

which implies that $\|v^\star\|_\infty \le R$.

For the latter case, we assume that $\max_{1\le i\le n} v_i^\star \le 0$. Then, we have

$$\min_{1\le i\le n} v_i^\star \ge \log\left(\min_{1\le i,j\le n}\{r_i, c_j\}\right) - \log\left(\sum_{i=1}^n e^{u_i^\star}\right).$$

This together with $\|u^\star\|_\infty \le \frac{\|C\|_\infty}{\eta} - \log(\min_{1\le i,j\le n}\{r_i, c_j\})$ yields that $\|v^\star\|_\infty \le R$. □

Putting Lemma 8.3.1 and 8.3.2 together, we have the following straightforward consequence:

Corollary 8.3.3 *Letting $\{(u^t, v^t)\}_{t\ge 0}$ be the iterates generated by Algorithm 22, we have*

$$f(u^t, v^t) - f(u^\star, v^\star) \le 4RE_t.$$

Remark 8.3.4 *The notation R is also used in Dvurechensky et al. [2018] and has the same order as ours since R in our paper only involves an term $\log(n) - \log(\min_{1\le i,j\le n}\{r_i, c_j\})$.*

Remark 8.3.5 *We further comment on the proof techniques in this paper and Dvurechensky et al. [2018]. Indeed, the proof for Dvurechensky et al. [2018, Lemma 2] depends on taking full advantage of the shift property of Sinkhorn; namely, either $B(\overline{u}^t, \overline{v}^t)\mathbf{1}_n = r$ or $B(\overline{u}^t, \overline{v}^t)^\top \mathbf{1}_n = c$ where $(\overline{u}^t, \overline{v}^t)$ stands for the iterate generated by Sinkhorn. Unfortunately, Greenkhorn does not enjoy such a shift property. We have thus proposed a different approach for bounding $f(u^t, v^t) - f(u^\star, v^\star)$ via appeal to the ℓ_∞-norm of the solution $(u^\star, v^\star)$.*

Complexity analysis—bounding the number of iterations. We proceed to provide an upper bound for the iteration number to achieve a desired tolerance ε' in Algorithm 22. First, we start with a lower bound for the difference of function values between two consecutive iterates of Algorithm 22:

Lemma 8.3.6 *Letting $\{(u^t, v^t)\}_{t\ge 0}$ be the iterates generated by Algorithm 22, we have*

$$f(u^t, v^t) - f(u^{t+1}, v^{t+1}) \ge \tfrac{(E_t)^2}{28n}.$$

Proof. Combining Altschuler et al. [2017, Lemma 5] and the fact that the row or column update is chosen in a greedy manner, we have

$$f(u^t, v^t) - f(u^{t+1}, v^{t+1}) \ge \tfrac{1}{2n}\left(\rho(r, r(B(u^t, v^t)) + \rho(c, c(B(u^t, v^t))\right).$$

Furthermore, Altschuler et al. [2017, Lemma 6] implies that

$$\rho(r, r(B(u^t, v^t)) + \rho(c, c(B(u^t, v^t)) \geq \tfrac{1}{7}\left(\|r - r(B(u^t, v^t))\|_1^2 + \|c - c(B(u^t, v^t))\|_1^2\right).$$

Putting these pieces together yields that

$$f(u^t, v^t) - f(u^{t+1}, v^{t+1}) \geq \tfrac{1}{14n}\left(\|r - r(B(u^t, v^t))\|_1^2 + \|c - c(B(u^t, v^t))\|_1^2\right).$$

Combining the above inequality with the definition of E_t implies the desired result. □

We are now able to derive the iteration complexity of Algorithm 22.

Theorem 8.3.7 *Letting $\{(u^t, v^t)\}_{t\geq 0}$ be the iterates generated by Algorithm 22, the number of iterations required to satisfy $E_t \leq \varepsilon'$ is upper bounded by $t \leq 2 + \frac{112nR}{\varepsilon'}$ where $R > 0$ is defined in Lemma 8.3.2.*

Proof. Letting $\delta_t = f(u^t, v^t) - f(u^\star, v^\star)$, we derive from Corollary 8.3.3 and Lemma 8.3.6 that

$$\delta_t - \delta_{t+1} \geq \max\left\{\tfrac{\delta_t^2}{448nR^2}, \tfrac{(\varepsilon')^2}{28n}\right\},$$

where $E_t \geq \varepsilon'$ as soon as the stopping criterion is not fulfilled. In the following step we apply a switching strategy introduced by Dvurechensky et al. [2018]. Given any $t \geq 1$, we have two estimates:

(i) Considering the process from the first iteration and the t-th iteration, we have

$$\tfrac{\delta_{t+1}}{448nR^2} \leq \tfrac{1}{t + 448nR^2\delta_1^{-2}} \Longrightarrow t \leq 1 + \tfrac{448nR^2}{\delta_t} - \tfrac{448nR^2}{\delta_1}.$$

(ii) Considering the process from the $(t+1)$-th iteration to the $(t+m)$-th iteration for any $m \geq 1$, we have

$$\delta_{t+m} \leq \delta_t - \tfrac{(\varepsilon')^2 m}{28n} \Longrightarrow m \leq \tfrac{28n(\delta_t - \delta_{t+m})}{(\varepsilon')^2}.$$

We then minimize the sum of two estimates by an optimal choice of $s \in (0, \delta_1]$:

$$t \leq \min_{0<s\leq\delta_1}\left(2 + \tfrac{448nR^2}{s} - \tfrac{448nR^2}{\delta_1} + \tfrac{28ns}{(\varepsilon')^2}\right) = \begin{cases} 2 + \frac{224nR}{\varepsilon'} - \frac{448nR^2}{\delta_1}, & \delta_1 \geq 4R\varepsilon', \\ 2 + \frac{28n\delta_1}{(\varepsilon')^2}, & \delta_1 \leq 4R\varepsilon'. \end{cases}$$

This implies that $t \leq 2 + \frac{112nR}{\varepsilon'}$ in both cases and completes the proof. □

Equipped with the result of Theorem 8.3.7 and the scheme of Algorithm 23, we are able to establish the following result for the complexity of Algorithm 23:

Theorem 8.3.8 *The Greenkhorn scheme for approximating optimal transport (Algorithm 23) returns an ε-approximate transportation plan (cf. Definition 8.2.1) in*

$$O\left(\frac{n^2\|C\|_\infty^2 \log(n)}{\varepsilon^2}\right)$$

arithmetic operations.

Proof. We follow the proof steps in [Altschuler et al., 2017, Theorem 1] and obtain that the transportation plan $\hat{X}$ returned by Algorithm 23 satisfies that

$$\begin{aligned}\langle C, \hat{X}\rangle - \langle C, X^\star\rangle &\leq 2\eta\log(n) + 4(\|\tilde{X}\mathbf{1}_n - r\|_1 + \|\tilde{X}^\top\mathbf{1}_n - c\|_1)\|C\|_\infty \\ &\leq \tfrac{\varepsilon}{2} + 4(\|\tilde{X}\mathbf{1}_n - r\|_1 + \|\tilde{X}^\top\mathbf{1}_n - c\|_1)\|C\|_\infty,\end{aligned}$$

where $X^\star$ is an optimal solution to the OT problem and $\tilde{X} = \textsc{Greenkhorn}(C, \eta, \tilde{r}, \tilde{c}, \frac{\varepsilon'}{2})$. The last inequality in the above display holds true since $\eta = \frac{\varepsilon}{4\log(n)}$. Furthermore,

$$\begin{aligned}\|\tilde{X}\mathbf{1}_n - r\|_1 + \|\tilde{X}^\top\mathbf{1}_n - c\|_1 &\leq \|\tilde{X}\mathbf{1}_n - \tilde{r}\|_1 + \|\tilde{X}^\top\mathbf{1}_n - \tilde{c}\|_1 + \|r - \tilde{r}\|_1 + \|c - \tilde{c}\|_1 \\ &\leq \tfrac{\varepsilon'}{2} + \tfrac{\varepsilon'}{4} + \tfrac{\varepsilon'}{4} = \varepsilon'.\end{aligned}$$

Putting these pieces together with $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$ yields that $\langle C, \hat{X}\rangle - \langle C, X^\star\rangle \leq \varepsilon$.

The remaining step is to analyze the complexity bound. It follows from Theorem 8.3.7 and the definition of $\tilde{r}$ and $\tilde{c}$ in Algorithm 23 that

$$\begin{aligned}t \leq 2 + \tfrac{112nR}{\varepsilon'} &\leq 2 + \tfrac{96n\|C\|_\infty}{\varepsilon}\left(\tfrac{\|C\|_\infty}{\eta} + \log(n) - 2\log\left(\min_{1\leq i,j\leq n}\{r_i, c_j\}\right)\right) \\ &\leq 2 + \tfrac{96n\|C\|_\infty}{\varepsilon}\left(\tfrac{4\|C\|_\infty\log(n)}{\varepsilon} + \log(n) - 2\log\left(\tfrac{\varepsilon}{64n\|C\|_\infty}\right)\right) \\ &= O\left(\tfrac{n\|C\|_\infty^2\log(n)}{\varepsilon^2}\right).\end{aligned}$$

The total iteration complexity in Step 2 of Algorithm 23 is bounded by $O(n\|C\|_\infty^2\log(n)\varepsilon^{-2})$. Each iteration of Algorithm 22 requires $O(n)$ arithmetic operations. Thus, the total number of arithmetic operations is $O(n^2\|C\|_\infty^2\log(n)\varepsilon^{-2})$. Moreover, $\tilde{r}$ and $\tilde{c}$ in Step 1 of Algorithm 23 can be found in $O(n)$ arithmetic operations and Altschuler et al. [2017, Algorithm 2] requires $O(n^2)$ arithmetic operations. Therefore, we conclude that the total number of arithmetic operations is $O(n^2\|C\|_\infty^2\log(n)\varepsilon^{-2})$. □

The complexity results presented in Theorem 8.3.8 improve the best known complexity bound $\widetilde{O}(n^2\epsilon^{-3})$ of Greenkhorn [Altschuler et al., 2017, Abid and Gower, 2018], Notably, it matches the best known complexity bound of Sinkhorn [Dvurechensky et al., 2018]. The key feature of our analysis is that the per-iteration progress of Greenkhorn can be lower bounded by a new quantity (cf. Lemmas 8.3.1 and 8.3.2). It allows us to apply the switching strategy in Theorem 8.3.7 to improve the complexity upper bound of Greenkhorn.

In practice, Greenkhorn has been reported to outperform Sinkhorn [Altschuler et al., 2017] in terms of row/column updates and our improved complexity bound can provide the theoretical justification for this phenomenon.

Algorithm 24 APDAMD$(\varphi, A, b, \varepsilon')$

Input: $t = 0$.
Initialization: $\bar{\alpha}^0 = \alpha^0 = 0$, $z^0 = \mu^0 = \lambda^0 = \mathbf{0}_{2n}$ and $L^0 = 1$.
repeat
  Set $M^t = \frac{L^t}{2}$.
  repeat
    Set $M^t = 2M^t$.
    Compute the stepsize: $\alpha^{t+1} = \frac{1+\sqrt{1+4\delta M^t \bar{\alpha}^t}}{2\delta M^t}$.
    Compute the average coefficient: $\bar{\alpha}^{t+1} = \bar{\alpha}^t + \alpha^{t+1}$.
    Compute the first average step: $\mu^{t+1} = \frac{\alpha^{t+1} z^t + \bar{\alpha}^t \lambda^t}{\bar{\alpha}^{t+1}}$.
    Compute the mirror descent: $z^{t+1} = \operatorname{argmin}_{z \in \mathbb{R}^n} \{(z - \mu^{t+1})^\top \nabla \varphi(\mu^{t+1}) + \frac{B_\phi(z, z^t)}{\alpha^{t+1}}\}$.
    Compute the second average step: $\lambda^{t+1} = \frac{\alpha^{t+1} z^{t+1} + \bar{\alpha}^t \lambda^t}{\bar{\alpha}^{t+1}}$.
  until $\varphi(\lambda^{t+1}) - \varphi(\mu^{t+1}) - (\lambda^{t+1} - \mu^{t+1})^\top \nabla \varphi(\mu^{t+1}) \le \frac{M^t}{2} \|\lambda^{t+1} - \mu^{t+1}\|_\infty^2$.
  Compute the main average step: $x^{t+1} = \frac{\alpha^{t+1} x(\mu^{t+1}) + \bar{\alpha}^t x^t}{\bar{\alpha}^{t+1}}$.
  Set $L^{t+1} = \frac{M^t}{2}$.
  Set $t = t + 1$.
until $\|Ax^t - b\|_1 \le \varepsilon'$.
Output: X^t where $x^t = \mathrm{vec}(X^t)$.

8.4 Adaptive Primal-Dual Accelerated Mirror Descent

We propose an adaptive primal-dual accelerated mirror descent (APDAMD) for solving the entropic regularized OT problem in Eq. (8.2). APDAMD and its application to the OT problem are presented in Algorithm 24 and 25. We prove the complexity bound of $O(n^2\sqrt{\delta}\|C\|_\infty \log(n)\varepsilon^{-1})$ where $\delta > 0$ stands for the regularity of the mirror mapping ϕ.

General setup. We consider the following generalization of the entropic regularized OT problem in Eq. (8.2):

$$\min_{x \in Q} \; f(x), \quad \text{s.t. } Ax = b, \tag{8.22}$$

where f is strongly convex with respect to the ℓ_1-norm on the set Q:

$$f(x') - f(x) - (x' - x)^\top \nabla f(x) \ge \tfrac{\eta}{2}\|x' - x\|_1^2 \text{ for any } x', x \in Q.$$

Note that, in the specific setting of the entropic regularized OT problem, the function $f(x) = \sum_{i,j} C_{ij} x_{j+n(i-1)} + \eta \cdot x_{j+n(i-1)} \log(x_{j+n(i-1)})$ where $x_{j+n(i-1)} = X_{ij}$ for any i, j where X is the transportation plan in equation (8.2), and the vector $b \in \mathbb{R}^{2n \times 1}$ is defined as: $b_i = r_i$ as $1 \le i \le n$ and $b_i = c_{i-n}$ when $n+1 \le i \le 2n$. Furthermore, the matrix $A = (A_{ij}) \in \mathbb{R}^{2n \times n^2}$ is defined as: When $1 \le i \le n$, we denote $A_{ij} = 1$ if $1 + n(i-1) \le j \le n \cdot i$ and 0 otherwise; When $n+1 \le i \le 2n$, we define $A_{ij} = 1$ if $j \in \{i - n + n(l-1) : 1 \le l \le n\}$ and 0 otherwise.

To be consistent with the notations in Algorithms 25 and 26, we specifically denote A_{ot} as the matrix A corresponding to the entropic regularized OT problem.

After some calculations with the general problem (8.22), we obtain that the dual problem is as follows:

$$\min_{\lambda \in \mathbb{R}^{2n}} \widetilde{\varphi}(\lambda) := \{\langle \lambda, b\rangle + \max_{x \in \mathbb{R}^{n^2}} \{-f(x) - \langle A^\top \lambda, x\rangle\}\}, \tag{8.23}$$

and $\nabla \widetilde{\varphi}(\lambda) = b - Ax(\lambda)$ where $x(\lambda) = \operatorname{argmax}_{x \in \mathbb{R}^n} \{-f(x) - \langle A^\top \lambda, x\rangle\}$; see the explicit form in Eq. (8.9) with $\lambda = (\alpha, \beta)$. By Nesterov [2005, Theorem 1] with ℓ_1-norm for the dual space of the Lagrange multipliers, the dual objective function $\tilde{\varphi}$ satisfies the following inequality:

$$\widetilde{\varphi}(\lambda') - \widetilde{\varphi}(\lambda) - (\lambda' - \lambda)^\top \nabla \widetilde{\varphi}(\lambda) \leq \tfrac{\|A\|_{1\to 1}^2}{2\eta} \|\lambda' - \lambda\|_\infty^2. \tag{8.24}$$

In the entropic regularized OT problem, each column of the matrix A_{ot} contains no more than two nonzero elements which are equal to one. Since $\|A_{\text{ot}}\|_{1\to 1}$ is equal to maximum ℓ_1-norm of the column of this matrix, we have $\|A_{\text{ot}}\|_{1\to 1} = 2$. Thus, the dual objective function $\widetilde{\varphi}$ is $\frac{4}{\eta}$-gradient Lipschitz with respect to the ℓ_∞-norm.

In addition, we define the Bregman divergence $B_\phi : \mathbb{R}^{2n} \times \mathbb{R}^{2n} \mapsto [0, +\infty)$ by

$$B_\phi(\lambda', \lambda) := \phi(\lambda') - \phi(\lambda) - (\lambda' - \lambda)^\top \nabla \phi(\lambda),$$

where the mirror mapping ϕ is $\frac{1}{\delta}$-strongly convex and 1-smooth on $\mathbb{R}^{2n}$ with respect to ℓ_∞-norm; that is,

$$\tfrac{1}{2\delta}\|\lambda' - \lambda\|_\infty^2 \leq \phi(\lambda') - \phi(\lambda) - (\lambda' - \lambda)^\top \nabla \phi(\lambda) \leq \tfrac{1}{2}\|\lambda' - \lambda\|_\infty^2.$$

For example, we can choose $\phi(\lambda) = \frac{1}{2n}\|\lambda\|^2$ and $B_\phi(\lambda', \lambda) = \frac{1}{2n}\|\lambda' - \lambda\|^2$ in APDAMD where $\delta = n$. As such, $\delta > 0$ is a function of n in general and it will appear in the complexity bound of APDAMD for approximating the OT problem (cf. Theorem 8.4.5). It is worth noting that our algorithm uses a regularizer that acts only in the dual and our complexity bound is the best existing one among this group of algorithms [Dvurechensky et al., 2018, Guo et al., 2020, Guminov et al., 2021]. A very recent work of Jambulapati et al. [2019] showed that the complexity bound can be improved to $\widetilde{O}(n^2\varepsilon^{-1})$ using a more advanced area-convex mirror mapping [Sherman, 2017].

Properties of APDAMD. We present several important properties of Algorithm 24 that can be used later for entropic regularized OT problems. First, we prove the following result regarding the number of line search iterations in Algorithm 24 for solving the entropic regularized OT problem:

Lemma 8.4.1 *The number of line search iterations in Algorithm 24 for solving the entropic OT problem is finite. Furthermore, the total number of gradient oracle calls after the t-th iteration is bounded as*

$$N_t \leq 4t + 4 + \frac{2\log(\frac{8}{\eta}) - 2\log(L^0)}{\log 2}.$$

Algorithm 25 Approximating OT by Algorithm 24

Input: $\eta = \frac{\varepsilon}{4\log(n)}$ and $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$.

Step 1: Let $\tilde{r} \in \Delta_n$ and $\tilde{c} \in \Delta_n$ be defined by $(\tilde{r}, \tilde{c}) = (1 - \frac{\varepsilon'}{8})(r, c) + \frac{\varepsilon'}{8n}(\mathbf{1}_n, \mathbf{1}_n)$.

Step 2: Let $A_{\mathrm{ot}} \in \mathbb{R}^{2n \times n^2}$ and $b \in \mathbb{R}^{2n}$ be defined by $A_{\mathrm{ot}}\mathrm{vec}(X) = \begin{pmatrix} X\mathbf{1}_n \\ X^\top \mathbf{1}_n \end{pmatrix}$ and $b = \begin{pmatrix} \tilde{r} \\ \tilde{c} \end{pmatrix}$.

Step 3: Compute $\tilde{X} = \textsc{Apdamd}(\widetilde{\varphi}, A_{\mathrm{ot}}, b, \frac{\varepsilon'}{2})$ where $\widetilde{\varphi}$ is defined by Eq. (8.23).

Step 4: Round $\tilde{X}$ to $\hat{X}$ using Altschuler et al. [2017, Algorithm 2] such that $\hat{X}\mathbf{1}_n = r$ and $\hat{X}^\top \mathbf{1}_n = c$.

Output: $\hat{X}$.

Proof. First, we observe that multiplying M^t by two will not stop until the line search stopping criterion is satisfied. Then, Eq. (8.24) implies that the number of line search iterations in the line search strategy is finite and $M^t \le \frac{2\|A_{\mathrm{ot}}\|_{1\to 1}^2}{\eta}$ holds true for all $t \ge 0$. Otherwise, the line search stopping criterion is satisfied with $\frac{M^t}{2}$ since $\frac{M^t}{2} \ge \frac{\|A_{\mathrm{ot}}\|_{1\to 1}^2}{\eta}$.

Letting i_j denote the total number of multiplication at the j-th iteration, we have

$$i_0 \le 1 + \frac{\log(\frac{M^0}{L^0})}{\log 2}, \qquad i_j \le 2 + \frac{\log(\frac{M^j}{M^{j-1}})}{\log 2}.$$

Then, the total number of line search iterations is bounded by

$$\sum_{j=0}^{t} i_j \le 1 + \frac{\log(\frac{M^0}{L^0})}{\log 2} + \sum_{j=1}^{t}\left(2 + \frac{\log(\frac{M^j}{M^{j-1}})}{\log 2}\right) \le 2t + 1 + \frac{\log(\frac{2\|A_{\mathrm{ot}}\|_{1\to 1}^2}{\eta}) - \log(L^0)}{\log 2}.$$

Since each line search contains two gradient oracle calls and $\|A_{\mathrm{ot}}\|_{1\to 1} = 2$, we conclude the desired upper bound for the total number of gradient oracle calls after the t-th iteration. □

The next lemma presents a property of the function $\widetilde{\varphi}$ in Algorithm 24.

Lemma 8.4.2 *For each iteration t of Algorithm 24 and any $z \in \mathbb{R}^{2n}$, we have*

$$\bar{\alpha}^t \widetilde{\varphi}(\lambda^t) \le \sum_{j=0}^{t} (\alpha^j(\widetilde{\varphi}(\mu^j) + (z - \mu^j)^\top \nabla\widetilde{\varphi}(\mu^j))) + \|z\|_\infty^2.$$

Proof. First, we claim that it holds for any $z \in \mathbb{R}^n$:

$$\alpha^{t+1}(z^t - z)^\top \nabla\widetilde{\varphi}(\mu^{t+1}) \le \bar{\alpha}^{t+1}(\widetilde{\varphi}(\mu^{t+1}) - \widetilde{\varphi}(\lambda^{t+1})) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}). \tag{8.25}$$

Indeed, the optimality condition in mirror descent implies that, for any $z \in \mathbb{R}^{2n}$, we have

$$(z - z^{t+1})^\top \left(\nabla\widetilde{\varphi}(\mu^{t+1}) + \frac{\nabla\phi(z^{t+1}) - \nabla\phi(z^t)}{\alpha^{t+1}}\right) \ge 0.$$

By definition, we have $B_\phi(z, z^t) - B_\phi(z, z^{t+1}) - B_\phi(z^{t+1}, z^t) = (z - z^{t+1})^\top(\nabla\phi(z^{t+1}) - \nabla\phi(z^t))$ and $B_\phi(z^{t+1}, z^t) \geq \frac{1}{2\delta}\|z^{t+1} - z^t\|_\infty^2$. Putting these pieces together yields that

$$\begin{aligned}
\alpha^{t+1}(z^t - z)^\top\nabla\widetilde{\varphi}(\mu^{t+1}) &= \alpha^{t+1}(z^t - z^{t+1})^\top\nabla\widetilde{\varphi}(\mu^{t+1}) + \alpha^{t+1}(z^{t+1} - z)^\top\nabla\widetilde{\varphi}(\mu^{t+1}) \\
&\leq \alpha^{t+1}(z^t - z^{t+1})^\top\nabla\widetilde{\varphi}(\mu^{t+1}) + (z - z^{t+1})^\top(\nabla\phi(z^{t+1}) - \nabla\phi(z^t)) \\
&= \alpha^{t+1}(z^t - z^{t+1})^\top\nabla\widetilde{\varphi}(\mu^{t+1}) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}) - B_\phi(z^{t+1}, z^t) \\
&\leq \alpha^{t+1}(z^t - z^{t+1})^\top\nabla\widetilde{\varphi}(\mu^{t+1}) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}) - \tfrac{\|z^{t+1} - z^t\|_\infty^2}{2\delta}. \qquad (8.26)
\end{aligned}$$

The update formulas of μ^{t+1}, λ^{t+1}, α^{t+1} and $\bar{\alpha}^{t+1}$ imply that

$$\lambda^{t+1} - \mu^{t+1} = \frac{\alpha^{t+1}}{\bar{\alpha}^{t+1}}(z^{t+1} - z^t), \qquad \delta M^t(\alpha^{t+1})^2 = \bar{\alpha}^{t+1}.$$

Therefore, we have

$$\alpha^{t+1}(z^t - z^{t+1})^\top\nabla\widetilde{\varphi}(\mu^{t+1}) = \bar{\alpha}^{t+1}(\mu^{t+1} - \lambda^{t+1})^\top\nabla\widetilde{\varphi}(\mu^{t+1}),$$

and

$$\|z^{t+1} - z^t\|_\infty^2 = (\tfrac{\bar{\alpha}^{t+1}}{\alpha^{t+1}})^2\|\mu^{t+1} - \lambda^{t+1}\|_\infty^2 = \delta M^t\bar{\alpha}^{t+1}\|\mu^{t+1} - \lambda^{t+1}\|_\infty^2.$$

Putting these pieces together with Eq. (8.26) yields that

$$\begin{aligned}
&\alpha^{t+1}(z^t - z)^\top\nabla\widetilde{\varphi}(\mu^{t+1}) \\
&\leq \bar{\alpha}^{t+1}(\mu^{t+1} - \lambda^{t+1})^\top\nabla\widetilde{\varphi}(\mu^{t+1}) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}) - \tfrac{\bar{\alpha}^{t+1}M^t}{2}\|\mu^{t+1} - \lambda^{t+1}\|_\infty^2 \\
&= \bar{\alpha}^{t+1}\left((\mu^{t+1} - \lambda^{t+1})^\top\nabla\widetilde{\varphi}(\mu^{t+1}) - \tfrac{M^t}{2}\|\mu^{t+1} - \lambda^{t+1}\|_\infty^2\right) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}) \\
&\leq \bar{\alpha}^{t+1}(\widetilde{\varphi}(\mu^{t+1}) - \widetilde{\varphi}(\lambda^{t+1})) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}),
\end{aligned}$$

where the last inequality comes from the stopping criterion in the line search. This implies that Eq. (8.25) holds true.

The next step is to bound the iterative objective gap given by

$$\begin{aligned}
&\bar{\alpha}^{t+1}\widetilde{\varphi}(\lambda^{t+1}) - \bar{\alpha}^t\widetilde{\varphi}(\lambda^t) \qquad (8.27) \\
&\leq \alpha^{t+1}(\widetilde{\varphi}(\mu^{t+1}) + (z - \mu^{t+1})^\top\nabla\widetilde{\varphi}(\mu^{t+1})) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}).
\end{aligned}$$

Indeed, by combining $\bar{\alpha}^{t+1} = \bar{\alpha}^t + \alpha^{t+1}$ and the update formula of μ^{t+1}, we have

$$\alpha^{t+1}(\mu^{t+1} - z^t) = (\bar{\alpha}^{t+1} - \bar{\alpha}^t)\mu^{t+1} - \alpha^{t+1}z^t = \alpha^{t+1}z^t + \bar{\alpha}^t\lambda^t - \bar{\alpha}^t\mu^{t+1} - \alpha^{t+1}z^t = \bar{\alpha}^t(\lambda^t - \mu^{t+1}).$$

This together with the convexity of $\widetilde{\varphi}$ implies that

$$\begin{aligned}
&\alpha^{t+1}(\mu^{t+1} - z)^\top\nabla\widetilde{\varphi}(\mu^{t+1}) \\
&= \alpha^{t+1}(\mu^{t+1} - z^t)^\top\nabla\widetilde{\varphi}(\mu^{t+1}) + \alpha^{t+1}(z^t - z)^\top\nabla\widetilde{\varphi}(\mu^{t+1}) \\
&= \bar{\alpha}^t(\lambda^t - \mu^{t+1})^\top\nabla\widetilde{\varphi}(\mu^{t+1}) + \alpha^{t+1}(z^t - z)^\top\nabla\widetilde{\varphi}(\mu^{t+1}) \\
&\leq \bar{\alpha}^t(\widetilde{\varphi}(\lambda^t) - \widetilde{\varphi}(\mu^{t+1})) + \alpha^{t+1}(z^t - z)^\top\nabla\widetilde{\varphi}(\mu^{t+1}).
\end{aligned}$$

Furthermore, we derive from Eq. (8.25) and $\bar{\alpha}^{t+1} = \bar{\alpha}^t + \alpha^{t+1}$ that

$$\begin{aligned}
&\bar{\alpha}^t(\widetilde{\varphi}(\lambda^t) - \widetilde{\varphi}(\mu^{t+1})) + \alpha^{t+1}(z^t - z)^\top \nabla\widetilde{\varphi}(\mu^{t+1}) \\
&\quad\leq\ \bar{\alpha}^t(\widetilde{\varphi}(\lambda^t) - \widetilde{\varphi}(\mu^{t+1})) + \bar{\alpha}^{t+1}(\widetilde{\varphi}(\mu^{t+1}) - \widetilde{\varphi}(\lambda^{t+1})) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}) \\
&\quad=\ \bar{\alpha}^t\widetilde{\varphi}(\lambda^t) - \bar{\alpha}^{t+1}\widetilde{\varphi}(\lambda^{t+1}) + \alpha^{t+1}\widetilde{\varphi}(\mu^{t+1}) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}).
\end{aligned}$$

Putting these pieces together yields that Eq. (8.27) holds true.

Finally, we prove our main results. By changing the index t to j in Eq. (8.27) and summing up the resulting inequality over $j = 0, 1, \ldots, t-1$, we have

$$\bar{\alpha}^t\widetilde{\varphi}(\lambda^t) - \bar{\alpha}^0\widetilde{\varphi}(\lambda^0) \leq \sum_{j=0}^{t-1}(\alpha^{j+1}(\widetilde{\varphi}(\mu^{j+1}) + (z - \mu^{j+1})^\top\nabla\widetilde{\varphi}(\mu^{j+1}))) + B_\phi(z, z^0) - B_\phi(z, z^t).$$

Since $\alpha^0 = \bar{\alpha}^0 = 0$, $B_\phi(z, z^t) \geq 0$ and ϕ is 1-smooth with respect to ℓ_∞-norm, we have

$$\begin{aligned}
\bar{\alpha}^t\widetilde{\varphi}(\lambda^t) &\leq \sum_{j=0}^{t}(\alpha^j(\widetilde{\varphi}(\mu^j) + (z - \mu^j)^\top\nabla\widetilde{\varphi}(\mu^j))) + B_\phi(z, z^0) \\
&\leq \sum_{j=0}^{t}(\alpha^j(\widetilde{\varphi}(\mu^j) + (z - \mu^j)^\top\nabla\widetilde{\varphi}(\mu^j))) + \|z - z^0\|_\infty^2 \\
&\overset{z^0=0}{=} \sum_{j=0}^{t}(\alpha^j(\widetilde{\varphi}(\mu^j) + (z - \mu^j)^\top\nabla\widetilde{\varphi}(\mu^j))) + \|z\|_\infty^2.
\end{aligned}$$

This completes the proof. □

The final lemma provides us with a key lower bound for the accumulating parameter.

Lemma 8.4.3 *For each iteration t of Algorithm 24, we have $\bar{\alpha}^t \geq \frac{\eta(t+1)^2}{32\delta}$.*

Proof. For $t = 1$, we have $\bar{\alpha}^1 = \alpha^1 = \frac{1}{\delta M^1} \geq \frac{\eta}{8\delta}$ since $M^1 \leq \frac{8}{\eta}$ was proven in Lemma 8.4.1. Thus, the desired result holds true when $t = 1$. Then we proceed to prove that it holds true for $t \geq 1$ using the induction. Indeed, we have

$$\begin{aligned}
\bar{\alpha}^{t+1} = \bar{\alpha}^t + \alpha^{t+1} &= \bar{\alpha}^t + \tfrac{1+\sqrt{1+4\delta M^t\bar{\alpha}^t}}{2\delta M^t} \\
&= \bar{\alpha}^t + \tfrac{1}{2\delta M^t} + \sqrt{\tfrac{1}{4(\delta M^t)^2} + \tfrac{\bar{\alpha}^t}{\delta M^t}} \\
&\geq \bar{\alpha}^t + \tfrac{1}{2\delta M^t} + \sqrt{\tfrac{\bar{\alpha}^t}{\delta M^t}} \\
&\geq \bar{\alpha}^t + \tfrac{\eta}{16\delta} + \sqrt{\tfrac{\eta\bar{\alpha}^t}{8\delta}},
\end{aligned}$$

where the last inequality comes from $M^t \leq \frac{8}{\eta}$ as shown in Lemma 8.4.1. Suppose that the desired result holds true for $t = k_0$, we have

$$\bar{\alpha}^{k_0+1} \geq \tfrac{\eta(k_0+1)^2}{32\delta} + \tfrac{\eta}{16\delta} + \sqrt{\tfrac{\eta^2(k_0+1)^2}{256\delta^2}} = \tfrac{\eta((k_0+1)^2+2+2(k_0+1))}{32\delta} \geq \tfrac{\eta(k_0+2)^2}{32\delta}.$$

This completes the proof. □

Complexity analysis for APDAMD. We are now ready to establish the complexity bound of APDAMD for solving the entropic regularized OT problem. Indeed, we recall that $\widetilde{\varphi}(\lambda)$ is defined with $\lambda = (\alpha, \beta)$ by

$$\widetilde{\varphi}(\alpha, \beta) = -\eta \log \left(\sum_{1 \leq i,j \leq n} e^{\eta^{-1}(\alpha_i + \beta_j - C_{ij})} \right) + \alpha^\top r + \beta^\top c.$$

Since (α, β) can be obtain by $\alpha_i = \eta u_i$ and $\beta_j = \eta v_j$, we derive from Lemma 8.2.3 that

$$\|\alpha^\star\|_\infty \leq \eta R, \quad \|\beta^\star\|_\infty \leq \eta R.$$

where R is defined accordingly. Then, we proceed to the following key result determining an upper bound for the number of iterations for Algorithm 24 to reach a desired accuracy ε':

Theorem 8.4.4 *Letting $\{X^t\}_{t \geq 0}$ be the iterates generated by Algorithm 24, the number of iterations required to satisfy $\|A_{ot}\text{vec}(X^t) - b\|_1 \leq \varepsilon'$ is upper bounded by*

$$t \leq 1 + \sqrt{\frac{128 \delta R}{\varepsilon'}},$$

where $R > 0$ is defined in Lemma 8.2.3.

Proof. From Lemma 8.4.2, we have

$$\bar{\alpha}^t \widetilde{\varphi}(\lambda^t) \leq \min_{z \in B_\infty(2\eta R)} \left\{ \sum_{j=0}^{t} (\alpha^j (\widetilde{\varphi}(\mu^j) + (z - \mu^j)^\top \nabla \widetilde{\varphi}(\mu^j))) + \|z\|_\infty^2 \right\},$$

where $B_\infty(r) := \{\lambda \in \mathbb{R}^n \mid \|\lambda\|_\infty \leq r\}$. This implies that

$$\bar{\alpha}^t \widetilde{\varphi}(\lambda^t) \leq \min_{z \in B_\infty(2\eta R)} \left\{ \sum_{j=0}^{t} (\alpha^j (\widetilde{\varphi}(\mu^j) + (z - \mu^j)^\top \nabla \widetilde{\varphi}(\mu^j))) \right\} + 4\eta^2 R^2.$$

Since $\widetilde{\varphi}$ is the objective function of dual entropic regularized OT problem, we have

$$\widetilde{\varphi}(\mu^j) + (z - \mu^j)^\top \nabla \widetilde{\varphi}(\mu^j) = -f(x(\mu^j)) + z^\top (b - A_{\text{ot}} x(\mu^j)).$$

Therefore, we conclude that

$$\begin{aligned}
\bar{\alpha}^t \widetilde{\varphi}(\lambda^t) &\leq \min_{z \in B_\infty(2\eta R)} \left\{ \sum_{j=0}^{t} (\alpha^j (\widetilde{\varphi}(\mu^j) + (z - \mu^j)^\top \nabla \widetilde{\varphi}(\mu^j))) \right\} + 4\eta^2 R^2 \\
&\leq 4\eta^2 R^2 - \bar{\alpha}^t f(x^t) + \min_{z \in B_\infty(2\eta R)} \{\bar{\alpha}^t z^\top (b - A_{\text{ot}} x^t)\} \\
&= 4\eta^2 R^2 - \bar{\alpha}^t f(x^t) - 2\bar{\alpha}^t \eta R \|A_{\text{ot}} x^t - b\|_1,
\end{aligned}$$

where the second inequality comes from the convexity of f and the last equality comes from the fact that ℓ_1-norm is the dual norm of ℓ_∞-norm. That is to say,

$$f(x^t) + \widetilde{\varphi}(\lambda^t) + 2\eta R\|A_{\rm ot}x^t - b\|_1 \leq \tfrac{4\eta^2R^2}{\bar{\alpha}^t}.$$

Suppose that $\lambda^\star$ is an optimal solution to dual entropic regularized OT problem satisfying $\|\lambda^\star\|_\infty \leq \eta R$, we have

$$\begin{aligned} f(x^t) + \widetilde{\varphi}(\lambda^t) &\geq f(x^t) + \widetilde{\varphi}(\lambda^\star) = f(x^t) + b^\top\lambda^\star + \max_{x\in\mathbb{R}^{n^2}} \left\{-f(x) - (\lambda^\star)^\top A_{\rm ot}x\right\} \\ &\geq f(x^t) + b^\top\lambda^\star - f(x^t) - (\lambda^\star)^\top A_{\rm ot}x^t = (b - A_{\rm ot}x^t)\lambda^\star \\ &\geq -\eta R\|A_{\rm ot}x^t - b\|_1, \end{aligned}$$

Therefore, we conclude that

$$\|A_{\rm ot}x^t - b\|_1 \leq \tfrac{4\eta R}{\bar{\alpha}^t} \leq \tfrac{128\delta R}{(t+1)^2}.$$

This completes the proof. □

We now present the complexity bound of Algorithm 25 for approximating the OT problem.

Theorem 8.4.5 *The APDAMD scheme for approximating optimal transport (Algorithm 25) returns an ε-approximate transportation plan (cf. Definition 8.2.1) in*

$$O\left(\frac{n^2\sqrt{\delta}\|C\|_\infty \log(n)}{\varepsilon}\right)$$

arithmetic operations.

Proof. Using the same argument as in Theorem 8.3.8, we have

$$\langle C, \hat{X}\rangle - \langle C, X^\star\rangle \leq \tfrac{\varepsilon}{2} + 4(\|\tilde{X}\mathbf{1}_n - r\|_1 + \|\tilde{X}^\top\mathbf{1}_n - c\|_1)\|C\|_\infty,$$

where $\hat{X}$ is returned by Algorithm 25, X^* is a solution to the OT problem and $\tilde{X} = \textsc{Apdamd}(\widetilde{\varphi}, A_{\rm ot}, b, \frac{\varepsilon'}{2})$. Since $\|\tilde{X}\mathbf{1}_n - r\|_1 + \|\tilde{X}^\top\mathbf{1}_n - c\|_1 \leq \varepsilon'$ and $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$, we have $\langle C, \hat{X}\rangle - \langle C, X^\star\rangle \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$.

The remaining step is to analyze the complexity bound. If follows from Lemma 8.4.1 and Theorem 8.4.4 that

$$\begin{aligned} N_t &\leq 4t + 4 + \tfrac{2\log(\frac{8}{\eta}) - 2\log(L^0)}{\log 2} \\ &\leq 8 + \sqrt{\tfrac{2048\delta R}{\varepsilon'}} + \tfrac{2\log(\frac{8}{\eta}) - 2\log(L^0)}{\log 2} \\ &= 8 + 256\sqrt{\tfrac{\delta R\|C\|_\infty\log(n)}{\varepsilon}} + \tfrac{2\log(\frac{32\log(n)}{\varepsilon}) - 2\log(L^0)}{\log 2}. \end{aligned}$$

Combining the definition of R in Lemma 8.2.3 with the definition of η, $\tilde{r}$ and $\tilde{c}$ in Algorithm 25, we have

$$R \leq \frac{4\|C\|_\infty \log(n)}{\varepsilon} + \log(n) - 2\log\left(\frac{\varepsilon}{64n\|C\|_\infty}\right).$$

Therefore, we conclude that

$$\begin{aligned} N_t \;\leq\; & 256\sqrt{\tfrac{\delta\|C\|_\infty \log(n)}{\varepsilon}}\sqrt{\tfrac{4\|C\|_\infty \log(n)}{\varepsilon} + \log(n) - 2\log\left(\tfrac{\varepsilon}{64n\|C\|_\infty}\right)} \\ & + \tfrac{2\log(\frac{32\log(n)}{\varepsilon}) - 2\log(L^0)}{\log 2} + 8 = O\left(\tfrac{\sqrt{\delta}\|C\|_\infty \log(n)}{\varepsilon}\right). \end{aligned}$$

The total iteration complexity in Step 3 of Algorithm 25 is bounded by $O(\sqrt{\delta}\|C\|_\infty \log(n)\varepsilon^{-1})$. Each iteration of Algorithm 24 requires $O(n^2)$ arithmetic operations. Therefore, the total number of arithmetic operations is $O(n^2\sqrt{\delta}\|C\|_\infty \log(n)\varepsilon^{-1})$. Moreover, $\tilde{r}$ and $\tilde{c}$ in Step 1 of Algorithm 25 can be found in $O(n)$ arithmetic operations and Altschuler et al. [2017, Algorithm 2] requires $O(n^2)$ arithmetic operations. Therefore, we conclude that the total number of arithmetic operations is $O(n^2\sqrt{\delta}\|C\|_\infty \log(n)\varepsilon^{-1})$. □

The complexity results in Theorem 8.4.5 suggests an interesting feature of the (regularized) OT problem. Indeed, the dependence of that bound on δ manifests the necessity of ℓ_∞-norm in the understanding of the complexity of the entropic regularized OT problem. This view is also in harmony with the proof technique of running time for Greenkhorn, where we rely on ℓ_∞-norm of optimal solutions of the dual entropic regularized OT problem to measure the progress in the objective value among the successive iterates.

Revisiting APDAGD. We revisit APDAGD [Dvurechensky et al., 2018] for the entropic regularized OT problem. First, we point out that the current complexity bound of $\widetilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\})$ is not valid by a simple counterexample. Then, we establish a new complexity bound of APDAGD. Despite the issue with entropic regularized OT, we wish to emphasize that APDAGD is still an interesting and efficient accelerated algorithm for general linearly constrained convex optimization problem with solid theoretical guarantee. More precisely, Dvurechensky et al. [2018, Theorem 3] is not applicable to entropic regularized OT since no dual solution exists with a constant bound in ℓ_2-norm. However, it can be used for analyzing other problems with bounded optimal dual solution.

To facilitate the ensuing discussion, we first present the complexity bound for entropic regularized OT in Dvurechensky et al. [2018] using our notation. Indeed, we recall that APDAGD is developed for solving the optimization problem with the objective function $\widehat{\varphi}$ defined as follows,

$$\min_{\alpha,\beta\in\mathbb{R}^n} \; \widehat{\varphi}(\alpha,\beta) := \eta\left(\sum_{i,j=1}^n e^{-\frac{C_{ij}-\alpha_i-\beta_j}{\eta}-1}\right) - \alpha^\top r - \beta^\top c. \tag{8.28}$$

Algorithm 26 Approximating OT by Dvurechensky et al. [2018, Algorithm 3]

Input: $\eta = \frac{\varepsilon}{4\log(n)}$ and $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$.

Step 1: Let $\tilde{r} \in \Delta_n$ and $\tilde{c} \in \Delta_n$ be defined by $(\tilde{r}, \tilde{c}) = (1 - \frac{\varepsilon'}{8})(r, c) + \frac{\varepsilon'}{8n}(\mathbf{1}_n, \mathbf{1}_n)$.

Step 2: Let $A_{\text{ot}} \in \mathbb{R}^{2n \times n^2}$ and $b \in \mathbb{R}^{2n}$ be defined by $A_{\text{ot}}\text{vec}(X) = \begin{pmatrix} X\mathbf{1}_n \\ X^\top \mathbf{1}_n \end{pmatrix}$ and $b = \begin{pmatrix} \tilde{r} \\ \tilde{c} \end{pmatrix}$.

Step 3: Compute $\tilde{X} = \textsc{Apdagd}(\widetilde{\varphi}, A_{\text{ot}}, b, \frac{\varepsilon'}{2})$ where $\widetilde{\varphi}$ is defined by Eq. (8.23).

Step 4: Round $\tilde{X}$ to $\hat{X}$ using Altschuler et al. [2017, Algorithm 2] such that $\hat{X}\mathbf{1}_n = r$ and $\hat{X}^\top \mathbf{1}_n = c$.

Theorem 8.4.6 (Theorem 4 in Dvurechensky et al. [2018]) *Let $\overline{R} > 0$ be defined such that there exists an optimal solution to the dual entropic regularized OT problem in Eq. (8.23), denoted by $(u^\star, v^\star)$, satisfying $\|(u^\star, v^\star)\| \leq \overline{R}$, the APDAGD scheme for approximating optimal transport (cf. Algorithm 26) returns an ε-approximate transportation plan (cf. Definition 8.2.1) in*

$$O\left(\min\left\{\frac{n^{9/4}\sqrt{\overline{R}\|C\|_\infty \log(n)}}{\varepsilon}, \frac{n^2 \overline{R}\|C\|_\infty \log(n)}{\varepsilon^2}\right\}\right),$$

arithmetic operations.

From the above theorem, Dvurechensky et al. [2018] claims that the complexity bound for APDAGD is $\widetilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\})$. However, there are two issues:

1. The upper bound $\overline{R}$ is assumed to be independent of n, which is not correct; see our counterexample in Proposition 8.4.7.

2. The known upper bound $\overline{R}$ for the optimal solution depends on $\min_{1 \leq i,j \leq n}\{r_i, c_j\}$ (cf. Dvurechensky et al. [2018, Lemma 1] or Lemma 8.2.3 in our paper). This implies that the valid algorithm needs to take the rounding error with r and c into account.

Corrected upper bound $\overline{R}$. Corollary 8.2.4 and Lemma 8.3.2 imply that a straightforward upper bound for $\overline{R}$ is $\widetilde{O}(\sqrt{n})$. Given a tolerance $\varepsilon \in (0, 1)$, we further show that $\overline{R}$ is indeed $\Omega(\sqrt{n})$ by using a specific entropic regularized OT problem as follows.

Proposition 8.4.7 *Suppose that $C = \mathbf{1}_n \mathbf{1}_n^\top$ and $r = c = \frac{1}{n}\mathbf{1}_n$. Given a tolerance $\varepsilon \in (0, 1)$ and the regularization term $\eta = \frac{\varepsilon}{4\log(n)}$, all the optimal solutions of the dual entropic regularized OT problem in Eq. (8.28) satisfy that $\|(\alpha^*, \beta^*)\| \gtrsim \sqrt{n}$.*

Proof. By the definition r, c and η, we rewrite the dual function $\widehat{\varphi}(\alpha, \beta)$ as follows:

$$\widehat{\varphi}(\alpha, \beta) = \frac{\varepsilon}{4e\log(n)} \sum_{1 \leq i,j \leq n} e^{-\frac{4\log(n)(1-\alpha_i-\beta_j)}{\varepsilon}} - \frac{\alpha^\top \mathbf{1}_n}{n} - \frac{\beta^\top \mathbf{1}_n}{n}.$$

Since $(\alpha^\star, \beta^\star)$ is an optimal solution of dual entropic regularized OT problem, we have

$$e^{\frac{4\log(n)\alpha_i^\star}{\varepsilon}} \sum_{j=1}^n e^{-\frac{4\log(n)(1-\beta_j^\star)}{\varepsilon}} = e^{\frac{4\log(n)\beta_i^\star}{\varepsilon}} \sum_{j=1}^n e^{-\frac{4\log(n)(1-\alpha_j^\star)}{\varepsilon}} = \frac{e}{n} \quad \text{for all } i \in [n]. \tag{8.29}$$

This implies $\alpha_i^\star = \alpha_j^\star$ and $\beta_i^\star = \beta_j^\star$ for all $i, j \in [n]$, and $\alpha_i^\star + \beta_i^\star$ are the same for all $i \in [n]$. Without loss of generality, we can let $\alpha_i^\star = 0$ in Eq. (8.29) and obtain that

$$\beta_i^\star = 1 + \frac{\varepsilon}{4\log(n)}(1 - 2\log(n)) = 1 + \frac{\varepsilon}{4\log(n)} - \frac{\varepsilon}{2}.$$

which implies that $\alpha_i^\star + \beta_i^\star = 1 + \frac{\varepsilon}{4\log(n)} - \frac{\varepsilon}{2} \geq \frac{1}{2}$ for all $i \in [n]$. Thus, we have

$$\|(\alpha^\star, \beta^\star)\| \geq \sqrt{\frac{\sum_{i=1}^n (\alpha_i^\star + \beta_i^\star)^2}{2}} = \frac{1}{2}\sqrt{\frac{n}{2}} \gtrsim \sqrt{n}.$$

As a consequence, we achieve the conclusion of the proposition. □

Approximation algorithm for OT by APDAGD. It is worth noting that the rounding procedure is missing in Dvurechensky et al. [2018, Algorithm 4] and we improve it to Algorithm 26. In particular, Dvurechensky et al. [2018, Algorithm 3] is used in Step 3 of Algorithm 26 for another function $\widetilde{\varphi}$ defined in Eq. (8.9). Given the corrected upper bound $\overline{R}$ and Algorithm 26 for approximating OT, we provide a new complexity bound of Algorithm 26 in the following proposition.

Proposition 8.4.8 *The APDAGD scheme for approximating optimal transport (Algorithm 26) returns an ε-approximate transportation plan (cf. Definition 8.2.1) in*

$$O\left(\frac{n^{5/2}\|C\|_\infty \sqrt{\log(n)}}{\varepsilon}\right)$$

arithmetic operations.

Proof. The proof is a simple modification of the proof for Dvurechensky et al. [2018, Theorem 4] and we only give a proof sketch here. In particular, we can obtain that the number of iterations for Algorithm 26 required to reach the tolerance ε is

$$t \leq O\left(\max\left\{\min\left\{\frac{n^{1/4}\sqrt{\overline{R}\|C\|_\infty \log(n)}}{\varepsilon}, \frac{\overline{R}\|C\|_\infty \log(n)}{\varepsilon^2}\right\}, \frac{\overline{R}\sqrt{\log n}}{\varepsilon}\right\}\right). \tag{8.30}$$

Moreover, we have $\overline{R} \leq \sqrt{n}\eta R$ where $R = \eta^{-1}\|C\|_\infty + \log(n) - 2\log(\min_{1\leq i,j\leq n}\{r_i, c_j\})$. Therefore, the total iteration complexity in Step 3 of Algorithm 26 is $O(\sqrt{n\log(n)}\|C\|_\infty \varepsilon^{-1})$. Each iteration of APDAGD requires $O(n^2)$ arithmetic operations. Therefore, the total number of arithmetic operations is $O(n^{5/2}\|C\|_\infty \sqrt{\log(n)}\varepsilon^{-1})$. Note that $\tilde{r}$ and $\tilde{c}$ in Step 1 of Algorithm 26 can be found in $O(n)$ arithmetic operations and Altschuler et al. [2017, Algorithm 2] requires $O(n^2)$ arithmetic operations. Therefore, we conclude that the total number of arithmetic operations is $O(n^{5/2}\|C\|_\infty \sqrt{\log(n)}\varepsilon^{-1})$. □

Remark 8.4.9 *As indicated in Proposition 8.4.8, the corrected complexity bound of APDAGD for the entropic regularized OT is similar to that of APDAMD when we choose* $\phi(\cdot) = \frac{1}{2n}\|\cdot\|^2$ *and have* $\delta = n$*. From this perspective, our algorithm can be viewed as a generalization of APDAGD. Since our algorithm utilizes* ℓ_∞*-norm in the line search criterion, it is more robust than APDAGD in practice.*

8.5 Accelerating Sinkhorn

We present an accelerated variant of Sinkhorn for solving the entropic regularized OT problem in Eq. (8.2). Combined with a rounding scheme, our algorithm can be used for solving the OT problem in Eq. (8.1) and achieves a complexity bound of $\widetilde{O}(n^{7/3}\varepsilon^{-4/3})$, which improves that of Sinkhorn in terms of $1/\varepsilon$ and APDAGD and AAM [Guminov et al., 2021] in terms of n. The idea comes from a novel combination of Nesterov's estimated sequence and the techniques for analyzing Sinkhorn.

Algorithmic procedure. We present the pseudocode of accelerated Sinkhorn in Algorithm 27. This algorithm achieves the acceleration by using Nesterov's estimate sequences [Nesterov, 2018]. While our algorithm can be interpreted as an accelerated block coordinate descent algorithm, it is worth mentioning that our algorithm is purely *deterministic* and thus differs from other accelerated randomized algorithms [Lin et al., 2015, Fercoq and Richtárik, 2015, Lu et al., 2018] in the optimization literature.

Algorithm 27 is a novel combination of Nesterov's estimate sequences, a monotone search step, the choice of greedy coordinate and two coordinate updates. It is applied to solve the dual entropic regularized OT problem in Eq. (8.5):

$$\min_{u,v} \ \varphi(u,v) := \log(\|B(u,v)\|_1) - u^\top r - v^\top c.$$

More specifically, Nesterov's estimate sequences are responsible for optimizing a dual objective function φ in a fast rate. The coordinate update guarantees that $\varphi(\widehat{u}^t, \widehat{v}^t) \le \varphi(\grave{u}^t, \grave{v}^t)$ and $\|B(\widehat{u}^t, \widehat{v}^t)\|_1 = 1$. The monotone search step guarantees that $\varphi(u^t, v^t) \le \varphi(\widehat{u}^t, \widehat{v}^t)$. The greedy coordinate update guarantees that $\varphi(\check{u}^{t+1}, \check{v}^{t+1}) \le \varphi(u^t, v^t)$ with sufficient progress.

Furthermore, we also use the same quantity as that in Greekhorn to measure the per-iteration residue of Algorithm 27:

$$E_t = \|r(B(u^t, v^t)) - r\|_1 + \|c(B(u^t, v^t)) - c\|_1. \tag{8.31}$$

The computationally expensive step is to compute $\frac{r(B(\bar{u}^t,\bar{v}^t))}{\|B(\bar{u}^t,\bar{v}^t)\|_1}$ and $\frac{c(B(\bar{u}^t,\bar{v}^t))}{\|B(\bar{u}^t,\bar{v}^t)\|_1}$. Since $B(\bar{u}^t, \bar{v}^t)$ does not have any special property, it is difficult to design some implementation trick to reduce the order of n. As such, the arithmetic operations for each iteration is $O(n^2)$ and is exactly the same as Sinkhorn [Cuturi, 2013], APDAGD [Dvurechensky et al., 2018] and AAM [Guminov et al., 2021]. Combining Algorithm 27 and Altschuler et al. [2017, Algorithm 2], we are ready to present the pseudocode of our main algorithm in Algorithm 28.

Algorithm 27 ACCELERATED SINKHORN$(C, \eta, r, c, \varepsilon')$

Input: $t = 0$, $\theta_0 = 1$ and $\breve{u}^0 = \tilde{u}^0 = \breve{v}^0 = \tilde{v}^0 = \mathbf{0}_n$.
while $E_t > \varepsilon'$ **do**
Compute $\begin{pmatrix} \bar{u}^t \\ \bar{v}^t \end{pmatrix} = (1-\theta_t)\begin{pmatrix} \breve{u}^t \\ \breve{v}^t \end{pmatrix} + \theta_t \begin{pmatrix} \tilde{u}^t \\ \tilde{v}^t \end{pmatrix}$.
Compute $\tilde{u}^{t+1}$ and $\tilde{v}^{t+1}$ by

$$\tilde{u}^{t+1} = \tilde{u}^t - \frac{1}{2\theta_t}\left(\frac{r(B(\bar{u}^t, \bar{v}^t))}{\|B(\bar{u}^t, \bar{v}^t)\|_1} - r\right), \qquad \tilde{v}^{t+1} = \tilde{v}^t - \frac{1}{2\theta_t}\left(\frac{c(B(\bar{u}^t, \bar{v}^t))}{\|B(\bar{u}^t, \bar{v}^t)\|_1} - c\right).$$

Compute $\begin{pmatrix} \grave{u}^t \\ \grave{v}^t \end{pmatrix} = \begin{pmatrix} \bar{u}^t \\ \bar{v}^t \end{pmatrix} + \theta_t \left(\begin{pmatrix} \tilde{u}^{t+1} \\ \tilde{v}^{t+1} \end{pmatrix} - \begin{pmatrix} \tilde{u}^t \\ \tilde{v}^t \end{pmatrix}\right)$.
if t is even **then**
$\widehat{u}^t = \grave{u}^t + \log(r) - \log(r(B(\grave{u}^t, \grave{v}^t)))$ and $\widehat{v}^t = \grave{v}^t$.
else
$\widehat{u}^t = \grave{u}^t$ and $\widehat{v}^t = \grave{v}^t + \log(c) - \log(c(B(\grave{u}^t, \grave{v}^t)))$.
Compute $\begin{pmatrix} u^t \\ v^t \end{pmatrix} = \operatorname{argmin}\left\{\varphi(u,v)\middle| \begin{pmatrix} u \\ v \end{pmatrix} \in \left\{\begin{pmatrix} \breve{u}^t \\ \breve{v}^t \end{pmatrix}, \begin{pmatrix} \widehat{u}^t \\ \widehat{v}^t \end{pmatrix}\right\}\right\}$.
if t is even **then**
$\breve{u}^{t+1} = u^t + \log(r) - \log(r(B(u^t, v^t)))$ and $\breve{v}^{t+1} = v^t$.
else
$\breve{u}^{t+1} = u^t$ and $\breve{v}^{t+1} = v^t + \log(c) - \log(c(B(u^t, v^t)))$.
Compute $\theta_{t+1} = \frac{\theta_t(\sqrt{\theta_t^2+4} - \theta_t)}{2}$.
Set $t = t + 1$.
Output: $B(u^t, v^t)$.

Algorithm 28 Approximating OT by Algorithm 27

Input: $\eta = \frac{\varepsilon}{4\log(n)}$ and $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$.

Step 1: Let $\tilde{r} \in \Delta_n$ and $\tilde{c} \in \Delta_n$ be defined by $(\tilde{r}, \tilde{c}) = (1 - \frac{\varepsilon'}{8})(r, c) + \frac{\varepsilon'}{8n}(\mathbf{1}_n, \mathbf{1}_n)$.
Step 2: Compute $\tilde{X} = $ ACCELERATED SINKHORN$(C, \eta, \tilde{r}, \tilde{c}, \frac{\varepsilon'}{2})$.
Step 3: Round $\tilde{X}$ to $\hat{X}$ using Altschuler et al. [2017, Algorithm 2] such that $\hat{X}\mathbf{1}_n = r$ and $\hat{X}^\top \mathbf{1}_n = c$.
Output: $\widehat{X}$.

The regularization parameter η is set as before, and Step 1 is necessary since accelerated Sinkhorn is not well behaved if the marginal distributions have sparse support.

Technical lemmas. We first present two technical lemmas which are essential in the analysis of Algorithm 27. The first lemma provides an inductive relationship on the quantity

$$\delta_t = \varphi(\check{u}^t, \check{v}^t) - \varphi(u^\star, v^\star), \tag{8.32}$$

where $(u^\star, v^\star)$ is an optimal solution of the dual entropic regularized OT problem in Eq. (8.5) that satisfies Lemma 8.2.4. To facilitate the discussion, we recall Eq. (8.10) with $\|A\|_{1\to 2} = \sqrt{2}$ as follows,

$$\varphi(u', v') - \varphi(u, v) - \begin{pmatrix} u' - u \\ v' - v \end{pmatrix}^\top \nabla\varphi(u, v) \le \left\| \begin{pmatrix} u' - u \\ v' - v \end{pmatrix} \right\|^2, \tag{8.33}$$

which will be used in the proof of the first lemma.

Lemma 8.5.1 *Let $\{(\check{u}^t, \check{v}^t)\}_{t\ge 0}$ be the iterates generated by Algorithm 27 and $(u^\star, v^\star)$ be an optimal solution of the dual entropic regularized OT problem. Then, we have*

$$\delta_{t+1} \le (1-\theta_t)\delta_t + \theta_t^2 \left(\left\| \begin{pmatrix} u^\star - \tilde{u}^t \\ v^\star - \tilde{v}^t \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^\star - \tilde{u}^{t+1} \\ v^\star - \tilde{v}^{t+1} \end{pmatrix} \right\|^2 \right).$$

Proof. Using Eq. (8.33) with $(u', v') = (\grave{u}^t, \grave{v}^t)$ and $(u, v) = (\bar{u}^t, \bar{v}^t)$, we have

$$\varphi(\grave{u}^t, \grave{v}^t) \le \varphi(\bar{u}^t, \bar{v}^t) + \theta_t \begin{pmatrix} \tilde{u}^{t+1} - \tilde{u}^t \\ \tilde{v}^{t+1} - \tilde{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) + \theta_t^2 \left\| \begin{pmatrix} \tilde{u}^{t+1} - \tilde{u}^t \\ \tilde{v}^{t+1} - \tilde{v}^t \end{pmatrix} \right\|^2.$$

After simple calculations, we find that

$$\begin{aligned} \varphi(\bar{u}^t, \bar{v}^t) &= (1-\theta_t)\varphi(\bar{u}^t, \bar{v}^t) + \theta_t \varphi(\bar{u}^t, \bar{v}^t), \\ \begin{pmatrix} \tilde{u}^{t+1} - \tilde{u}^t \\ \tilde{v}^{t+1} - \tilde{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) &= -\begin{pmatrix} \tilde{u}^t - \bar{u}^t \\ \tilde{v}^t - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) + \begin{pmatrix} \tilde{u}^{t+1} - \bar{u}^t \\ \tilde{v}^{t+1} - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t). \end{aligned}$$

Putting these pieces together yields that

$$\begin{aligned} \varphi(\grave{u}^t, \grave{v}^t) \le\ & \theta_t \Bigg(\underbrace{\varphi(\bar{u}^t, \bar{v}^t) + \begin{pmatrix} \tilde{u}^{t+1} - \bar{u}^t \\ \tilde{v}^{t+1} - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) + \theta_t \left\| \begin{pmatrix} \tilde{u}^{t+1} - \tilde{u}^t \\ \tilde{v}^{t+1} - \tilde{v}^t \end{pmatrix} \right\|^2}_{\mathbf{I}} \Bigg) \\ & \underbrace{+ (1-\theta_t)\varphi(\bar{u}^t, \bar{v}^t) - \theta_t \begin{pmatrix} \tilde{u}^t - \bar{u}^t \\ \tilde{v}^t - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t)}_{\mathbf{II}}. \end{aligned} \tag{8.34}$$

We first bound the term **I**. Indeed, by the update formula for $(\tilde{u}^{t+1}, \tilde{v}^{t+1})$ and the definition of $\nabla\varphi$, we have

$$\begin{pmatrix} u - \tilde{u}^{t+1} \\ v - \tilde{v}^{t+1} \end{pmatrix}^\top \left(\nabla\varphi(\bar{u}^t, \bar{v}^t) + 2\theta_t \begin{pmatrix} \tilde{u}^{t+1} - \tilde{u}^t \\ \tilde{v}^{t+1} - \tilde{v}^t \end{pmatrix} \right) = 0 \text{ for all } (u, v) \in \mathbb{R}^{2n}.$$

Letting $(u, v) = (u^\star, v^\star)$ and rearranging the resulting equation yields that

$$\begin{aligned} \begin{pmatrix} \tilde{u}^{t+1} - \bar{u}^t \\ \tilde{v}^{t+1} - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) &= \begin{pmatrix} u^\star - \bar{u}^t \\ v^\star - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) \\ &+\theta_t \left(\left\| \begin{pmatrix} u^\star - \tilde{u}^t \\ v^\star - \tilde{v}^t \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^\star - \tilde{u}^{t+1} \\ v^\star - \tilde{v}^{t+1} \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} \tilde{u}^{t+1} - \tilde{u}^t \\ \tilde{v}^{t+1} - \tilde{v}^t \end{pmatrix} \right\|^2 \right). \end{aligned}$$

Using the convexity of φ, we have

$$\begin{pmatrix} u^\star - \bar{u}^t \\ v^\star - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) \leq \varphi(u^\star, v^\star) - \varphi(\bar{u}^t, \bar{v}^t).$$

Putting these pieces together yields that

$$\mathbf{I} \leq \varphi(u^\star, v^\star) + \theta_t \left(\left\| \begin{pmatrix} u^\star - \tilde{u}^t \\ v^\star - \tilde{v}^t \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^\star - \tilde{u}^{t+1} \\ v^\star - \tilde{v}^{t+1} \end{pmatrix} \right\|^2 \right). \tag{8.35}$$

We then bound the term **II**. Indeed, we see from the definition of $(\bar{u}^t, \bar{v}^t)$ that

$$-\theta_t \begin{pmatrix} \tilde{u}^t - \bar{u}^t \\ \tilde{v}^t - \bar{v}^t \end{pmatrix} = \theta_t \begin{pmatrix} \bar{u}^t \\ \bar{v}^t \end{pmatrix} + (1 - \theta_t) \begin{pmatrix} \check{u}^t \\ \check{v}^t \end{pmatrix} - \begin{pmatrix} \bar{u}^t \\ \bar{v}^t \end{pmatrix} = (1 - \theta_t) \begin{pmatrix} \check{u}^t - \bar{u}^t \\ \check{v}^t - \bar{v}^t \end{pmatrix}.$$

Combining the above equation with the convexity of φ, we have

$$\mathbf{II} = (1 - \theta_t) \left(\varphi(\bar{u}^t, \bar{v}^t) + \begin{pmatrix} \check{u}^t - \bar{u}^t \\ \check{v}^t - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) \right) \leq (1 - \theta_t)\varphi(\check{u}^t, \check{v}^t). \tag{8.36}$$

Plugging Eq. (8.35) and Eq. (8.36) into Eq. (8.34) yields that

$$\varphi(\grave{u}^t, \grave{v}^t) \leq (1 - \theta_t)\varphi(\check{u}^t, \check{v}^t) + \theta_t\varphi(u^\star, v^\star) + \theta_t^2 \left(\left\| \begin{pmatrix} u^\star - \tilde{u}^t \\ v^\star - \tilde{v}^t \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^\star - \tilde{u}^{t+1} \\ v^\star - \tilde{v}^{t+1} \end{pmatrix} \right\|^2 \right).$$

Since $(\check{u}^{t+1}, \check{v}^{t+1})$ is obtained by a coordinate update from (u^t, v^t), we have $\varphi(u^t, v^t) \geq \varphi(\check{u}^{t+1}, \check{v}^{t+1})$. By the definition of (u^t, v^t), we have $\varphi(\widehat{u}^t, \widehat{v}^t) \geq \varphi(u^t, v^t)$. Since $(\widehat{u}^t, \widehat{v}^t)$ is obtained by a coordinate update from $(\grave{u}^t, \grave{v}^t)$, we have $\varphi(\grave{u}^t, \grave{v}^t) \geq \varphi(\widehat{u}^t, \widehat{v}^t)$. Collecting all of these results leads to

$$\varphi(\check{u}^{t+1}, \check{v}^{t+1}) - \varphi(u^\star, v^\star) \leq (1-\theta_t)(\varphi(\check{u}^t, \check{v}^t) - \varphi(u^\star, v^\star)) + \theta_t^2 \left(\left\| \begin{pmatrix} u^\star - \tilde{u}^t \\ v^\star - \tilde{v}^t \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^\star - \tilde{u}^{t+1} \\ v^\star - \tilde{v}^{t+1} \end{pmatrix} \right\|^2 \right).$$

This completes the proof. □

The second lemma provides an upper bound for δ_t defined by Eq. (8.32) where $\{(\check{u}^t, \check{v}^t)\}_{t\geq 0}$ are generated by Algorithm 27 and $(u^\star, v^\star)$ is an optimal solution defined by Corollary 8.2.4.

Lemma 8.5.2 *Let $\{(\check{u}^t, \check{v}^t)\}_{t\geq 0}$ be the iterates generated by Algorithm 27 and $(u^\star, v^\star)$ be an optimal solution of the dual entropic regularized OT problem satisfying that $\|(u^\star, v^\star)\| \leq \sqrt{2n}R$ where R is defined in Corollary 8.2.4. Then, we have*

$$\delta_t \leq \frac{8nR^2}{(t+1)^2}.$$

Proof. By simple calculations, we derive from the definition of θ_t that $\frac{\theta_{t+1}}{\theta_t} = \sqrt{1-\theta_{t+1}}$. Therefore, we conclude from Lemma 8.5.1 that

$$\left(\tfrac{1-\theta_{t+1}}{\theta_{t+1}^2}\right)\delta_{t+1} - \left(\tfrac{1-\theta_t}{\theta_t^2}\right)\delta_t \leq \left\| \begin{pmatrix} u^\star - \tilde{u}^t \\ v^\star - \tilde{v}^t \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^\star - \tilde{u}^{t+1} \\ v^\star - \tilde{v}^{t+1} \end{pmatrix} \right\|^2.$$

Equivalently, we have

$$\left(\tfrac{1-\theta_t}{\theta_t^2}\right)\delta_t + \left\| \begin{pmatrix} u^\star - \tilde{u}^t \\ v^\star - \tilde{v}^t \end{pmatrix} \right\|^2 \leq \left(\tfrac{1-\theta_0}{\theta_0^2}\right)\delta_0 + \left\| \begin{pmatrix} u^\star - \tilde{u}^0 \\ v^\star - \tilde{v}^0 \end{pmatrix} \right\|^2.$$

Since $\theta_0 = 1$ and $\tilde{u}^0 = \tilde{v}^0 = \mathbf{0}_n$, we have $\delta_t \leq \theta_{t-1}^2 \|(u^\star, v^\star)\|^2 \leq 2nR^2\theta_{t-1}^2$.

The remaining step is to show that $0 < \theta_t \leq \frac{2}{t+2}$. Indeed, the claim holds when $t = 0$ as we have $\theta_0 = 1$. Assume that the claim holds for $t \leq t_0$, i.e., $\theta_{t_0} \leq \frac{2}{t_0+2}$, we have

$$\theta_{t_0+1} = \frac{2}{1+\sqrt{1+\frac{4}{\theta_{t_0}^2}}} \leq \frac{2}{t_0+3}.$$

Putting these pieces together yields the desired inequality for δ_t. □

Main results. We present an upper bound for the number of iterations required by Algorithm 27. Note that the per-iteration progress of Algorithm 27 is measured by the function $\rho : \mathbb{R}^n_+ \times \mathbb{R}^n_+ \to \mathbb{R}_+$ given by: $\rho(a, b) := \mathbf{1}_n^\top (b - a) + \sum_{i=1}^n a_i \log(\frac{a_i}{b_i})$.

Theorem 8.5.3 *Let $\{(u^t, v^t)\}_{t\geq 0}$ be the iterates generated by Algorithm 27. The number of iterations required to reach the stopping criterion $E_t \leq \varepsilon'$ satisfies*

$$t \leq 1 + \left(\frac{16\sqrt{n}R}{\varepsilon'}\right)^{2/3},$$

where $R > 0$ is defined in Lemma 8.2.3.

Proof. We first claim that

$$\varphi(u^t, v^t) - \varphi(\breve{u}^{t+1}, \breve{v}^{t+1}) \geq \tfrac{1}{2}\left(\|r(B(u^t, v^t)) - r\|_1^2 + \|c(B(u^t, v^t)) - c\|_1^2\right). \tag{8.37}$$

By the definition of φ, we have

$$\begin{aligned}\varphi(u^t, v^t) - \varphi(\breve{u}^{t+1}, \breve{v}^{t+1}) &= \log(\|B(u^t, v^t)\|_1) \\ &\quad - \log(\|B(\breve{u}^{t+1}, \breve{v}^{t+1})\|_1) - (u^t - \breve{u}^{t+1})^\top r - (v^t - \breve{v}^{t+1})^\top c.\end{aligned} \tag{8.38}$$

From the update formula for $(\widehat{u}^t, \widehat{v}^t)$ and $(\breve{u}^{t+1}, \breve{v}^{t+1})$, it is clear that $\|B(\widehat{u}^t, \widehat{v}^t)\|_1 = 1$ and $\|B(\breve{u}^{t+1}, \breve{v}^{t+1})\|_1 = 1$ for all $t \geq 0$. Then, we derive from the update formula for (u^t, v^t) that $\|B(u^t, v^t)\|_1 = 1$ for all $t \geq 1$. Therefore, we have

$$\begin{aligned}\varphi(u^t, v^t) - \varphi(\breve{u}^{t+1}, \breve{v}^{t+1}) &= -(u^t - \breve{u}^{t+1})^\top r - (v^t - \breve{v}^{t+1})^\top c \\ &= (\log(r) - \log(r(B(u^t, v^t))))^\top r + (\log(c) - \log(c(B(u^t, v^t))))^\top c.\end{aligned}$$

Since $\mathbf{1}_n^\top r = \mathbf{1}_n^\top r(B(u^t, v^t)) = \mathbf{1}_n^\top c = \mathbf{1}_n^\top c(B(u^t, v^t)) = 1$, we have

$$\varphi(u^t, v^t) - \varphi(\breve{u}^{t+1}, \breve{v}^{t+1}) = \rho(r, r(B(u^t, v^t))) + \rho(c, c(B(u^t, v^t))).$$

Using Altschuler et al. [2017, Lemma 4], we derive Eq. (8.37) as desired.

By the definition of (u^t, v^t), we have $\varphi(\breve{u}^t, \breve{v}^t) \geq \varphi(u^t, v^t)$. Plugging this inequality into Eq. (8.37) together with the Cauchy-Schwarz inequality yields

$$\varphi(\breve{u}^t, \breve{v}^t) - \varphi(\breve{u}^{t+1}, \breve{v}^{t+1}) \geq \tfrac{1}{4} E_t^2.$$

Therefore, we conclude that

$$\varphi(\breve{u}^t, \breve{v}^t) - \varphi(\breve{u}^{t+1}, \breve{v}^{t+1}) \geq \tfrac{1}{4}\left(\sum_{i=j}^t E_i^2\right) \text{ for any } j \in \{1, 2, \ldots, t\}.$$

Since $\varphi(\breve{u}^{t+1}, \breve{v}^{t+1}) \geq \varphi(u^\star, v^\star)$ for all $t \geq 1$, we have $\varphi(\breve{u}^j, \breve{v}^j) - \varphi(\breve{u}^{t+1}, \breve{v}^{t+1}) \leq \delta_j$. Then, it follows from Lemma 8.5.2 that

$$\sum_{i=j}^t E_i^2 \leq \tfrac{32nR^2}{(j+1)^2}.$$

Putting these pieces together with the fact that $E_t \geq \varepsilon'$ as soon as the stopping criterion is not fulfilled yields

$$\frac{32nR^2}{(j+1)^2(t-j+1)} \geq (\varepsilon')^2.$$

Since this inequality holds true for all $j \in \{1, 2, \ldots, t\}$, we assume without loss of generality that t is even and let $j = t/2$. Then, we obtain that

$$t \leq 1 + \left(\frac{16\sqrt{n}R}{\varepsilon'}\right)^{2/3}.$$

This completes the proof of the theorem. □

We are ready to present the complexity bound of Algorithm 28 for solving the OT problem in Eq. (8.1). Note that $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$ is defined using the desired accuracy $\varepsilon > 0$.

Theorem 8.5.4 *The accelerated Sinkhorn scheme for approximating optimal transport (Algorithm 28) returns an ε-approximate transportation plan (cf. Definition 8.2.1) in*

$$O\left(\frac{n^{7/3}\|C\|_\infty^{4/3}(\log(n))^{1/3}}{\varepsilon^{4/3}}\right)$$

arithmetic operations.

Proof. Applying the same argument which is used in Theorem 8.3.8, we obtain that $\langle C, \widehat{X}\rangle - \langle C, X^\star\rangle \leq \varepsilon$ where $\tilde{X} = \textsc{Accelerated Sinkhorn}(C, \eta, \tilde{r}, \tilde{c}, \frac{\varepsilon'}{2})$ in Step 2 of Algorithm 28.

It remains to bound the number of iterations required by Algorithm 27 to reach the stopping criterion $E_t \leq \frac{\varepsilon'}{2}$. Using Theorem 8.5.3, we have

$$t \leq 1 + \left(\frac{32\sqrt{n}R}{\varepsilon'}\right)^{2/3}.$$

By the definition of R (cf. Lemma 8.2.3), $\eta = \frac{\varepsilon}{4\log(n)}$ and $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$, we have

$$\begin{aligned}
t &\leq 1 + \left(\frac{32\sqrt{n}R}{\varepsilon'}\right)^{2/3} \\
&\leq 1 + \left(\frac{256\sqrt{n}\|C\|_\infty}{\varepsilon}\left(\frac{\|C\|_\infty}{\eta} + \log(n) - \log\left(\min_{1\leq i,j\leq n}\{r_i, c_j\}\right)\right)\right)^{2/3} \\
&\leq 1 + \left(\frac{256\sqrt{n}\|C\|_\infty}{\varepsilon}\left(\frac{4\log(n)\|C\|_\infty}{\varepsilon} + \log(n) - \log\left(\frac{\varepsilon}{64n\|C\|_\infty}\right)\right)\right)^{2/3} \\
&= O\left(\frac{n^{1/3}\|C\|_\infty^{4/3}(\log(n))^{1/3}}{\varepsilon^{4/3}}\right).
\end{aligned}$$

Since each iteration of Algorithm 27 requires $O(n^2)$ arithmetic operations, the total number of arithmetic operations required by Step 2 of Algorithm 28 is $O(n^{7/3}\|C\|_\infty^{4/3}(\log(n))^{1/3}\varepsilon^{-4/3})$. Computing two vectors $\tilde{r}$ and $\tilde{c}$ in Step 1 of Algorithm 28 requires $O(n)$ arithmetic operations and Altschuler et al. [2017, Algorithm 2] requires $O(n^2)$ arithmetic operations. Therefore, the complexity bound of Algorithm 28 is $O(n^{7/3}\|C\|_\infty^{4/3}(\log(n))^{1/3}\varepsilon^{-4/3})$. □

Remark 8.5.5 *Theorem 8.5.4 shows that the complexity bound of accelerated Sinkhorn is better than that of Sinkhorn and Greenkhorn in terms of* $1/\varepsilon$ *but appears not to be near-linear in* n^2*. Thus, our algorithm is recommended when* $n \ll 1/\varepsilon$*. This occurs if the desired solution accuracy is relatively small, saying* 10^{-4}*, and the examples include the application problems from economics and operations research. In contrast, Sinkhorn and Greenkhorn are recommended when* $n \gg 1/\varepsilon$*. This occurs if the desired solution accuracy is relatively large, saying* 10^{-2}*, and the examples include the application problems from image processing.*

8.6 Experiments

We conduct the experiments to evaluate Greenkhorn, accelerated Sinkhorn and APDAMD on synthetic data and real images from the MNIST Digits dataset[1]. The baseline approaches include Sinkhorn [Cuturi, 2013], APDAGD [Dvurechensky et al., 2018] and GCPB[2] [Genevay et al., 2016] as the baseline approaches. Since the focus of this paper is the entropic regularized algorithms, we exclude the combinatorial algorithms from our experiment and refer to Dong et al. [2020] for an excellent comparative study.

In the literature, Greenkhorn and APDAGD were shown to outperform the Sinkhorn algorithm in terms of row/column updates [Altschuler et al., 2017, Dvurechensky et al., 2018] and we repeat the comparisons for the sake of completeness. For parameter tuning in the implementation of Greenkhorn, accelerated Sinkhorn and APDAMD, we follow most of the setups as shown in Algorithm 22, 24 and 27 but employ more aggressive choice of stepsize for the coordinate gradient updates in Algorithm 27. To obtain an optimal value of the OT problem, we employ the default LP solver in MATLAB.

Synthetic images. To generate the synthetic images, we adopt the process from Altschuler et al. [2017] and evaluate the performance of different algorithms on these synthetic images. The transportation distance is defined between two synthetic images while the cost matrix is defined based on the ℓ_1 distances among locations of pixel in the images. Each image is of size 20 by 20 pixels and generated by means of randomly placing a foreground square in a black background. Furthermore, a uniform distribution on $[0, 1]$ is used for the intensities of the pixels in the background while a uniform distribution on $[0, 50]$ is employed for the pixels in the foreground. We fix the proportion of the size of the foreground square as 10% of the whole images and implement all candidate algorithms.

We use the standard metrics to assess the performance of all the candidate algorithms. The first metric $d(.)$ is an ℓ_1 distance between the row, column outputs of some algorithm $\mathcal{A}$ and the corresponding transportation polytope of the probability measures, which is given by:

$$d(\mathcal{A}) := \|r(\mathcal{A}) - r\|_1 + \|c(\mathcal{A}) - c\|_1$$

[1] http://yann.lecun.com/exdb/mnist/

[2] GCPB is simply an application of stochastic averaged gradient [Schmidt et al., 2017] for solving the dual entropic regularized OT problem.

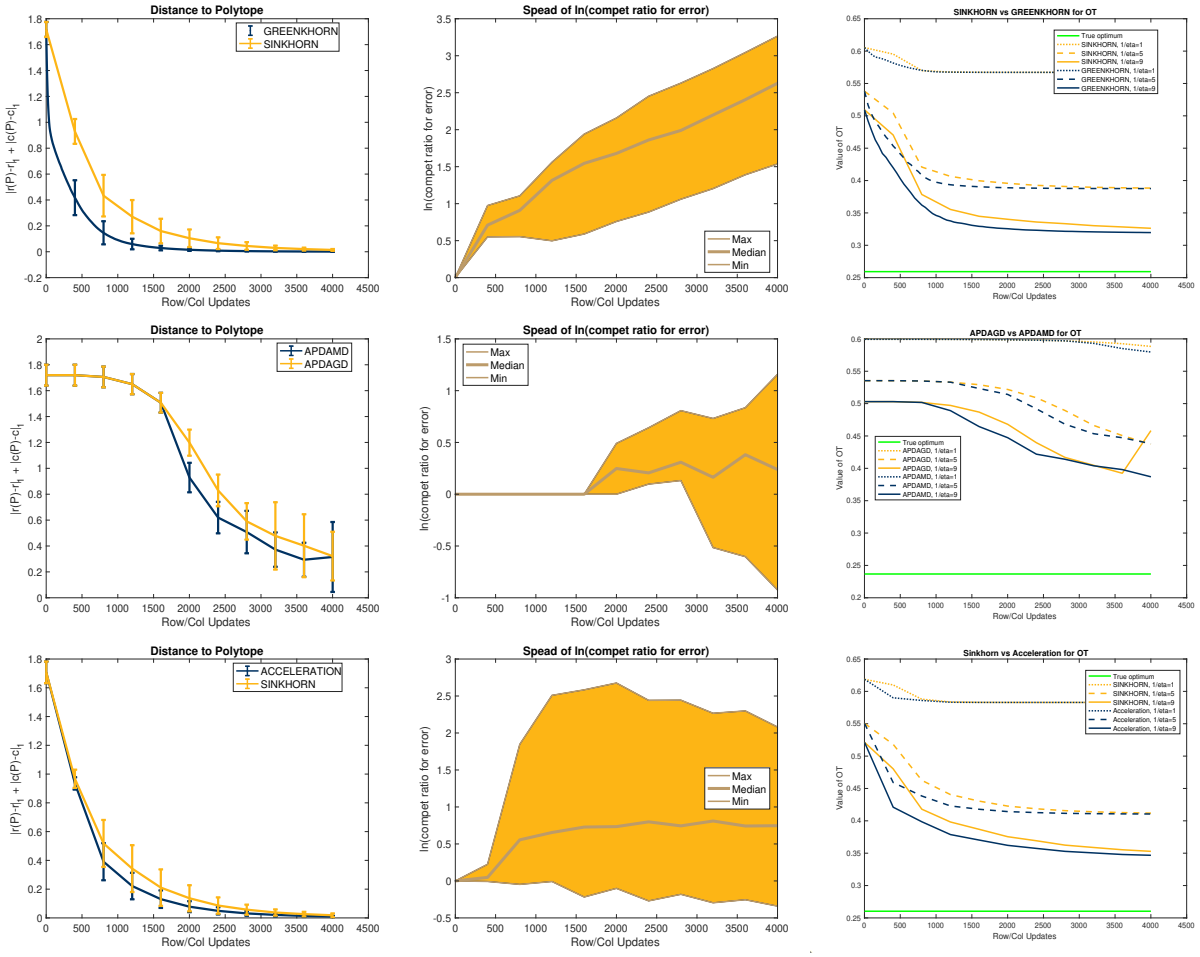


Figure 8.1: Performance of Sinkhorn v.s. Greenkhorn, APDAGD v.s. APDAMD and Sinkhorn v.s. accelerated Sinkhorn on synthetic images.

where $r(\mathcal{A})$ and $c(\mathcal{A})$ are the row and column obtained from the output of the algorithm $\mathcal{A}$ and r and c are row and column vectors of the original probability measures. The second metric is defined as competitive ratio $\log(d(\mathcal{A}_1)/d(\mathcal{A}_2))$ where $d(\mathcal{A}_1)$ and $d(\mathcal{A}_2)$ are the distances between the row, column outputs of algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ and the transportation polytope. We perform three pairwise comparative experiments on 10 randomly generated data: Sinkhorn v.s. Greekhorn, APDAGD v.s. APDAMD and Sinkhorn v.s. accelerated Sinkhorn. To further evaluate these algorithms, we compare their performance with respect to different choices of regularization parameter $\eta \in \{1, \frac{1}{5}, \frac{1}{9}\}$ while using the value of the OT problem as the baseline approach. The maximum number of iterations is $T = 5$. Figure 8.1 summarizes the experimental results. The images in the first row show the comparative performance of Sinkhorn and Greenkhorn in terms of the row/column updates. In the leftmost image, the comparison uses distance to transportation polytope $d(\mathcal{A})$ where $\mathcal{A}$ is either Sinkhorn or Greenkhorn. In the middle image, the maximum, median and min-

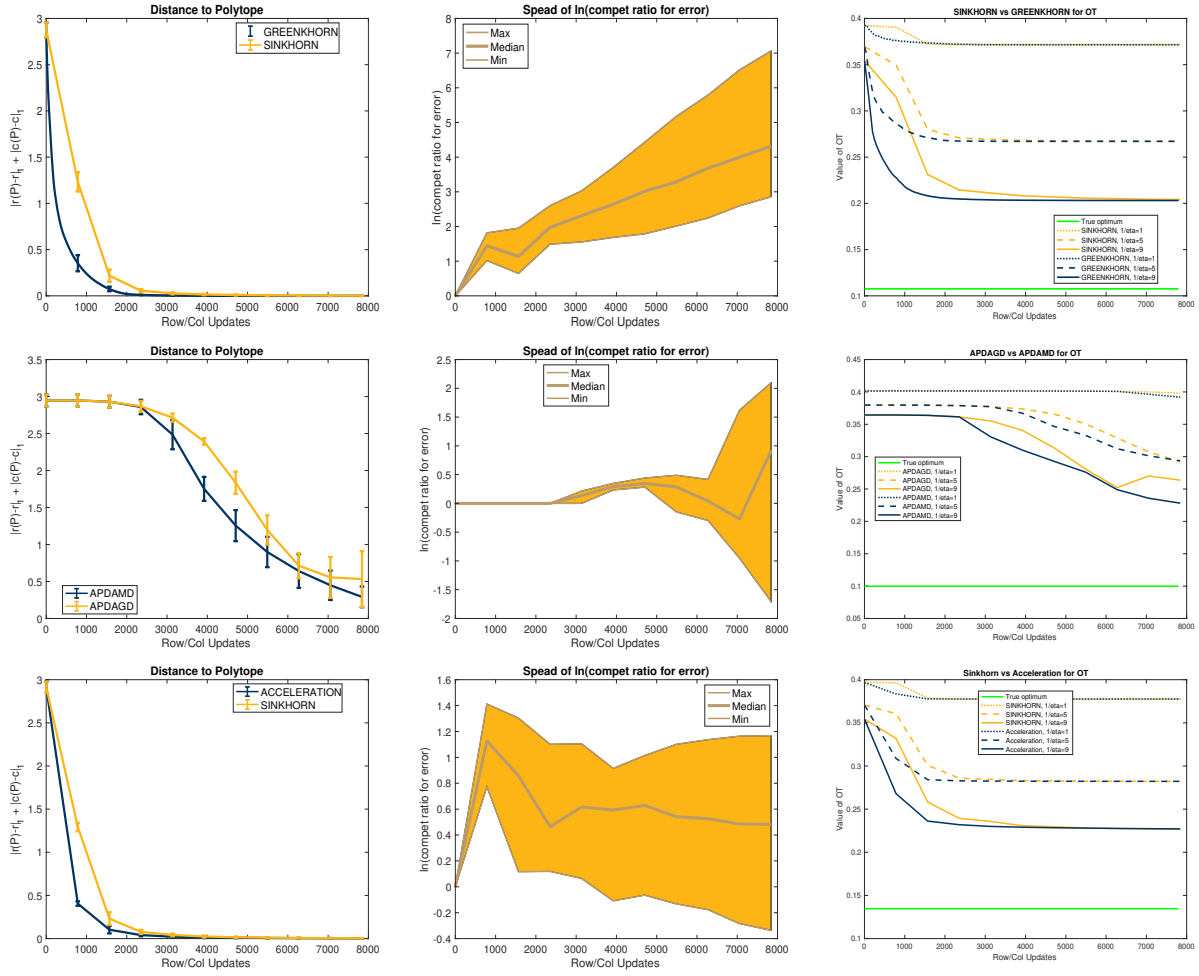


Figure 8.2: Performance of Sinkhorn v.s. Greenkhorn, APDAGD v.s. APDAMD and Sinkhorn v.s. accelerated Sinkhorn on the MNIST real images.

imum values of the competitive ratios $\log(d(\mathcal{A}_1)/d(\mathcal{A}_2))$ on 10 images are utilized for the comparison where $\mathcal{A}_1$ is Sinkhorn and $\mathcal{A}_2$ is Greenkhorn. In the rightmost image, we vary the regularization parameter $\eta \in \{1, \frac{1}{5}, \frac{1}{9}\}$ with these algorithms and using the value of the unregularized OT problem as the baseline. The other rows of images present comparative results for APDAGD v.s. APDAMD and Sinkhorn v.s. accelerated Sinkhorn. We find that (i) Greenkhorn outperforms Sinkhorn in terms of row/column updates, illustrating the improvement from *greedy coordinate descent*; (ii) APDAMD with $\delta = n$ and $\phi = (1/2n)\|\cdot\|^2$ is more robust than APDAGD, illustrating the advantage of using *mirror descent* and line search with $\|\cdot\|_\infty$; (iii) accelerated Sinkhorn outperforms Sinkhorn in terms of row/column updates, illustrating the improvement from *estimated sequence* and *monotone search*.

MNIST images. We proceed to the comparison between different algorithms on real images, using essentially the same evaluation metrics as in the synthetic images. The MNIST

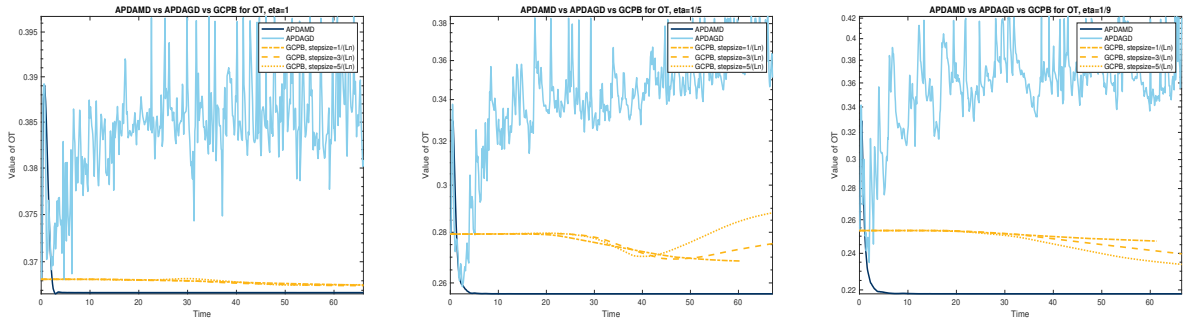


Figure 8.3: Performance of GCPB, APDAGD and APDAMD in term of time on the MNIST real images. These images specify the values of entropic regularized OT with varying regularization parameter $\eta \in \{1, \frac{1}{5}, \frac{1}{9}\}$, demonstrating the robustness of APDAMD.

dataset consists of 60,000 images of handwritten digits of size 28 by 28 pixels. To ensure that the masses of probability measures are dense, which leads to a tight dependence on n for our algorithms, we add a very small noise term (10^{-6}) to all zero elements in the measures and then normalize them so that their sum is 1. The maximum number of iterations is $T = 5$.

Figures 8.2 and 8.3 summarize the experimental results on MNIST. In the first row of Figure 8.2, we compare Sinkhorn and Greenkhorn in terms of row/column updates. The leftmost image specifies the distances $d(\mathcal{A})$ to the transportation polytope for the algorithm $\mathcal{A}$, which is either Sinkhorn or Greenkhorn; the middle image specifies the maximum, median and minimum of competitive ratios $\log(d(\mathcal{A}_1)/d(\mathcal{A}_2))$ on ten random pairs of MNIST images, where $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively correspond to Sinkhorn and Greenkhorn; the rightmost image specifies the values of the entropic regularized OT problem with varying regularization parameters $\eta \in \{1, \frac{1}{5}, \frac{1}{9}\}$. The remaining rows present comparative results for APDAGD v.s.APDAMD and Sinkhorn v.s.accelerated Sinkhorn. We observe that (i) the comparative performances of Sinkhorn v.s.Greenkhorn and APDAGD v.s.APDAMD are consistent with those on synthetic images; (ii) accelerated Sinkhorn deteriorates but remains better than Sinkhorn; (iii) APDAMD is more robust than APDAGD and GCPB.

8.7 Conclusion

We first show that the complexity bound of Greenkhorn can be improved to $\widetilde{O}(n^2\varepsilon^{-2})$, which matches the best known bound of Sinkhorn. Then, we propose APDAMD by generalizing APDAGD with a prespecified mirror mapping ϕ and show that it achieves the complexity bound of $\widetilde{O}(n^2\sqrt{\delta}\varepsilon^{-1})$ where $\delta > 0$ refers to the regularity of ϕ. We prove that the complexity bound of $\widetilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\})$ proved for APDAGD is invalid and prove a refined complexity bound of $\widetilde{O}(n^{5/2}\varepsilon^{-1})$. Moreover, we propose a *deterministic* accelerated variant of Sinkhorn via appeal to estimate sequence techniques and prove the complexity bound of $\widetilde{O}(n^{7/3}\varepsilon^{-4/3})$. As such, we see that accelerated Sinkhorn outperforms Sinkhorn

and Greenkhorn in terms of $1/\varepsilon$ and APDAGD and AAM in terms of n. Experiments on synthetic data and real images demonstrate the efficiency of our algorithms.

There are a few promising future directions arising from this work. First, it is important to develop fast algorithms to compute dimension-reduced versions of OT. Indeed, the OT suffers from the curse of dimensionality [Dudley, 1969, Fournier and Guillin, 2015], which means that a large amount of samples from two continuous measures is necessary to approximate the true OT between them. This can be mitigated when data lie on low-dimensional manifolds [Weed and Bach, 2019, Paty and Cuturi, 2019] but the sample complexity still remain pessimistic even in that case. This motivates recent works on efficient dimension-reduced OT, e.g., the sliced OT [Bonneel et al., 2015], generalized sliced OT [Kolouri et al., 2019], distributional sliced OT [Nguyen et al., 2021], further inspiring us to explore the application of our algorithms to these settings and eventually automatic differentiation schemes. Second, there have been several application problems arising from the interplay between OT and adversarial ML; see Bhagoji et al. [2019] and Pydi and Jog [2020] for example. However, it is known that OT has robustness issues when there are outliers in the supports of probability measures. Robust OT had been introduced to deal with these robustness issues [Balaji et al., 2020] where the idea is to relax the marginal constraints via certain probability divergences, such as KL divergence. It is to limit the amount of masses that the transportation plan will assign for the outliers in the supports of measures. Similar to OT, a key practical question with robust OT is computational. As such, we manage to develop efficient algorithms for the robust OT problem in the future work.

Chapter 9

Gradient-Free Nonconvex Nonsmooth Optimization

Nonsmooth nonconvex optimization problems broadly emerge in machine learning and business decision making, whereas two core challenges impede the development of efficient solution methods with finite-time convergence guarantee: the lack of computationally tractable optimality criterion and the lack of computationally powerful oracles. The contributions of this paper are two-fold. First, we establish the relationship between the celebrated Goldstein subdifferential [Goldstein, 1977] and uniform smoothing, thereby providing the basis and intuition for the design of gradient-free methods that guarantee the finite-time convergence to a set of Goldstein stationary points. Second, we propose the gradient-free method (GFM) and stochastic GFM for solving a class of nonsmooth nonconvex optimization problems and prove that both of them can return a (δ, ϵ)-Goldstein stationary point of a Lipschitz function f at an expected convergence rate at $O(d^{3/2}\delta^{-1}\epsilon^{-4})$ where d is the problem dimension. Two-phase versions of GFM and SGFM are also proposed and proven to achieve improved large-deviation results. Finally, we demonstrate the effectiveness of 2-SGFM on training ReLU neural networks with the MINST dataset.

9.1 Introduction

Many of the recent real-world success stories of machine learning have involved nonconvex optimization formulations, with the design of models and algorithms often being heuristic and intuitive. Thus a gap has arisen between theory and practice. Attempts have been made to fill this gap for different learning methodologies, including the training of multilayer neural networks [Choromanska et al., 2015], orthogonal tensor decomposition [Ge et al., 2015], M-estimators [Loh and Wainwright, 2015, Ma et al., 2020], synchronization and Max-Cut [Bandeira et al., 2016, Mei et al., 2017], smooth semidefinite programming [Boumal et al., 2016], matrix sensing and completion [Bhojanapalli et al., 2016, Ge et al., 2016], robust principal component analysis (RPCA) [Ge et al., 2017b] and phase retrieval [Wang

et al., 2017, Sun et al., 2018, Ma et al., 2020]. For an overview of nonconvex optimization formulations and the relevant machine learning applications, we refer to an excellent survey of Jain and Kar [2017].

It is intractable to compute an approximate global minimum [Nemirovski and Yudin, 1983] in general or to verify whether a point is a local minimum or a high-order saddle point [Murty and Kabadi, 1987]. Fortunately, the notion of *approximate stationary point* gives a reasonable optimality criterion when the objective function $f$is smooth; the goal here is to find a point $\mathbf{x} \in \mathbb{R}^d$ such that $\|\nabla f(\mathbf{x})\| \leq \epsilon$. Recent years have seen rapid algorithmic development through the lens of nonasymptotic convergence rates to ϵ-stationary points [Nesterov, 2013a, Ghadimi and Lan, 2013b, 2016, Carmon et al., 2017, 2018, Jin et al., 2021]. Another line of work establishes algorithm-independent lower bounds [Carmon et al., 2020, 2021, Arjevani et al., 2020, 2022].

Relative to its smooth counterpart, the investigation of nonsmooth optimization is relatively scarce, particularly in the nonconvex setting, both in terms of efficient algorithms and finite-time convergence guarantees. Yet, over several decades, nonsmooth nonconvex optimization formulations have found applications in many fields. A typical example is the training multi-layer neural networks with ReLU neurons, for which the piecewise linear activation functions induce nonsmoothness. Another example arises in controlling financial risk for asset portfolios or optimizing customer satisfaction in service systems or supply chain systems. Here, the nonsmoothness arises from the payoffs of financial derivatives and supply chain costs, e.g., options payoffs [Duffie, 2010] and supply chain overage/underage costs [Stadtler, 2008]. These applications make significant demands with respect to computational feasibility, and the design of efficient algorithms for solving nonsmooth nonconvex optimization problems has moved to the fore [Majewski et al., 2018, Davis et al., 2020, Daniilidis and Drusvyatskiy, 2020, Zhang et al., 2020a, Bolte and Pauwels, 2021, Davis et al., 2022, Tian et al., 2022].

The key challenges lie in two aspects: (i) the lack of a computationally tractable optimality criterion, and (ii) the lack of computationally powerful oracles. More specifically, in the classical setting where the function f is Lipschitz, we can define ϵ-stationary points based on the celebrated notion of Clarke stationarity [Clarke, 1990]. However, the value of such a criterion has been called into question by Zhang et al. [2020a], who show that no finite-time algorithm guarantees ϵ-stationarity when ϵ is less than a constant. Further, the computation of the gradient is impossible for many application problems and we only have access to a noisy function value at each point. This is a common issue in the context of simulation optimization [Nelson, 2010, Hong et al., 2015]; indeed, the objective function value is often achieved as the output of a black-box or complex simulator, for which the simulator does not have the infrastructure needed to effectively evaluate gradients; see also Ghadimi and Lan [2013b] and Nesterov and Spokoiny [2017] for the lack of gradient evaluation in practice.

Contribution. In this paper, we propose and analyze a class of deterministic and stochastic gradient-free methods for nonsmooth nonconvex optimization problems in which we only

assume that the function f is Lipschitz. Our contributions can be summarized as follows.

1. We establish a relationship between the Goldstein subdifferential and uniform smoothing via appeal to the hyperplane separation theorem. This result provides the basis for algorithmic design and finite-time convergence analysis of gradient-free methods to (δ, ϵ)-Goldstein stationary points.

2. We propose and analyze a gradient-free method (GFM) and stochastic GFM for solving a class of nonsmooth nonconvex optimization problems. Both of these methods are guaranteed to return a (δ, ϵ)-Goldstein stationary point of a Lipschitz function $f : \mathbb{R}^d \mapsto \mathbb{R}$ with an expected convergence rate of $O(d^{3/2}\delta^{-1}\epsilon^{-4})$ where $d \geq 1$ is the problem dimension. Further, we propose the two-phase versions of GFM and SGFM. As our goal is to return a (δ, ϵ)-Goldstein stationary point with user-specified high probability $1-\Lambda$, we prove that the two-phase version of GFM and SGFM can improve the dependence from $(1/\Lambda)^4$ to $\log(1/\Lambda)$ in the large-deviation regime.

Related works. Our work is related to a line of literature on gradient-based methods for nonsmooth and nonconvex optimization and gradient-free methods for smooth and nonconvex optimization. In the context of gradient-free methods, the basic idea is to approximate a full gradient using either a one-point estimator [Flaxman et al., 2005] or a two-point estimator [Agarwal et al., 2010, Ghadimi and Lan, 2013b, Duchi et al., 2015, Shamir, 2017, Nesterov and Spokoiny, 2017], where the latter approach achieves a better finite-time convergence guarantee. Despite the meteoric rise of two-point-based gradient-free methods, most of the work is restricted to convex optimization [Duchi et al., 2015, Shamir, 2017, Wang et al., 2018] and smooth and nonconvex optimization [Nesterov and Spokoiny, 2017, Ghadimi and Lan, 2013b, Lian et al., 2016, Liu et al., 2018, Chen et al., 2019, Ji et al., 2019, Huang et al., 2022a]. For nonsmooth and convex optimization, the best upper bound on the global rate of convergence is $O(d\epsilon^{-2})$ [Shamir, 2017] and this matches the lower bound [Duchi et al., 2015]. For smooth and nonconvex optimization, the best global rate of convergence is $O(d\epsilon^{-2})$ [Nesterov and Spokoiny, 2017] and $O(d\epsilon^{-4})$ if we only have access to noisy function value oracles [Ghadimi and Lan, 2013b]. Additional regularity conditions, e.g., a finite-sum structure, allow us to leverage variance-reduction techniques [Liu et al., 2018, Chen et al., 2019, Ji et al., 2019] and the best known result is $O(d^{3/4}\epsilon^{-3})$ [Huang et al., 2022a]. However, none of gradient-free methods have been developed for nonsmooth nonconvex optimization and the only gradient-free method we are aware of for the nonsmooth is summarized in Nesterov and Spokoiny [2017, Section 7].

9.2 Preliminaries

We provide the formal definitions for the class of Lipschitz functions considered in this paper, and the definitions for generalized gradients and the Goldstein subdifferential that lead to optimality conditions in nonsmooth nonconvex optimization.

Function classes. Imposing regularity on functions to be optimized is necessary for obtaining optimization algorithms with finite-time convergence guarantees [Nesterov, 2018]. In the context of nonsmooth optimization, there are two regularity conditions: Lipschitz properties of function values and bounds on function values.

We first list several equivalent definitions of Lipschitz continuity. A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is said to be *L-Lipschitz* if for every $\mathbf{x} \in \mathbb{R}^d$ and the direction $\mathbf{v} \in \mathbb{R}^d$ with $\|\mathbf{v}\| \leq 1$, the directional projection $f_{\mathbf{x},\mathbf{v}}(t) := f(\mathbf{x} + t\mathbf{v})$ defined for $t \in \mathbb{R}$ satisfies

$$|f_{\mathbf{x},\mathbf{v}}(t) - f_{\mathbf{x},\mathbf{v}}(t')| \leq L|t - t'|, \quad \text{for all } t, t' \in \mathbb{R}.$$

Equivalently, f is L-Lipschitz if for every $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, we have

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq L\|\mathbf{x} - \mathbf{x}'\|.$$

Further, the function value bound $f(\mathbf{x}^0) - \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ appears in complexity guarantees for smooth and nonconvex optimization problems [Nesterov, 2018] and is often assumed to be bounded by a positive constant $\Delta > 0$. Note that $\mathbf{x}^0$ is a prespecified point (i.e., an initial point for an algorithm) and we simply fix it for the remainder of this paper. We define the function class considered in this paper.

Definition 9.2.1 *Suppose that $\Delta > 0$ and $L > 0$ are both independent of the problem dimension $d \geq 1$. Then, we denote $\mathcal{F}_d(\Delta, L)$ as the set of L-Lipschitz functions $f : \mathbb{R}^d \mapsto \mathbb{R}$ with the bounded function value $f(\mathbf{x}^0) - \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \leq \Delta$.*

The function class $\mathcal{F}_d(\Delta, L)$ includes Lipschitz functions on $\mathbb{R}^d$ and is thus different from the nonconvex function class considered in the literature [Ghadimi and Lan, 2013b, Nesterov and Spokoiny, 2017]. First, we do not impose a smoothness condition on the function $f \in \mathcal{F}_d(\Delta, L)$, in contrast to the nonconvex functions studied in Ghadimi and Lan [2013b] which are assumed to have Lipschitz gradients. Second, Nesterov and Spokoiny [2017, Section 7] presented a complexity bound for a randomized optimization method for minimizing a nonsmooth nonconvex function. However, they did not clarify why the norm of the gradient of the approximate function $f_{\bar{\mu}}$ of the order δ (we use their notation) serves as a reasonable optimality criterion in nonsmooth nonconvex optimization. They also assume an exact function value oracle, ruling out many interesting application problems in simulation optimization and machine learning.

In contrast, our goal is to propose fast gradient-free methods for nonsmooth nonconvex optimization in the absence of an exact function value oracle. In general, the complexity bound of gradient-free methods will depend on the problem dimension $d \geq 1$ even when we assume that the function to be optimized is convex and smooth [Duchi et al., 2015, Shamir, 2017]. As such, we should consider a function class with a given dimension $d \geq 1$. In particular, we consider a optimality criterion based on the celebrated Goldstein subdifferential [Goldstein, 1977] and prove that the number of function value oracles required by our deterministic and stochastic gradient-free methods to find a (δ, ϵ)-Goldstein stationary point of $f \in \mathcal{F}_d(\Delta, L)$ is $O(\text{poly}(d, L, \Delta, 1/\epsilon, 1/\delta))$ when $\delta, \epsilon \in (0, 1)$ are constants.

It is worth mentioning that $\mathcal{F}_d(\Delta, L)$ contains a rather broad class of functions used in real-world application problems. Typical examples with additional regularity properties include Hadamard semi-differentiable functions [Shapiro, 1990, Delfour, 2019, Zhang et al., 2020a], Whitney-stratifiable functions [Bolte et al., 2007, Davis et al., 2020], o-minimally definable functions [Coste, 2000] and a class of semi-algebraic functions [Attouch et al., 2013a, Davis et al., 2020]. Thus, our gradient-free methods can be applied for solving these problems with finite-time convergence guarantees.

Generalized gradients and Goldstein subdifferential. We start with the definition of generalized gradients [Clarke, 1990] for nondifferentiable functions. This is perhaps the most standard extension of gradients to nonsmooth and nonconvex functions.

Definition 9.2.2 *Given a point $\mathbf{x} \in \mathbb{R}^d$ and a direction $\mathbf{v} \in \mathbb{R}^d$, the generalized directional derivative of a nondifferentiable function f is given by $Df(\mathbf{x}; \mathbf{v}) := \limsup_{\mathbf{y} \to \mathbf{x}, t \downarrow 0} \frac{f(\mathbf{y}+t\mathbf{v})-f(\mathbf{y})}{t}$. The generalized gradient of f is defined as $\partial f(\mathbf{x}) := \{\mathbf{g} \in \mathbb{R}^d : \mathbf{g}^\top \mathbf{v} \leq Df(\mathbf{x}; \mathbf{v}), \forall \mathbf{v} \in \mathbb{R}^d\}$.*

Rademacher's theorem guarantees that any Lipschitz function is almost everywhere differentiable. This implies that the generalized gradients of Lipschitz functions have additional properties and we can define them in a relatively simple way. The following proposition summarizes these results; we refer to Clarke [1990] for the proof details.

Proposition 9.2.3 *Suppose that f is L-Lipschitz for some $L > 0$, we have that $\partial f(\mathbf{x})$ is a nonempty, convex and compact set and $\|\mathbf{g}\| \leq L$ for all $\mathbf{g} \in \partial f(\mathbf{x})$. Further, $\partial f(\cdot)$ is an upper-semicontinuous set-valued map. Moreover, a generalization of mean-value theorem holds: for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$, there exist $\lambda \in (0, 1)$ and $\mathbf{g} \in \partial f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2)$ such that $f(\mathbf{x}_1) - f(\mathbf{x}_2) = \mathbf{g}^\top(\mathbf{x}_1 - \mathbf{x}_2)$. Finally, there is a simple way to represent the generalized gradient $\partial f(\mathbf{x})$:*

$$\partial f(\mathbf{x}) := \operatorname{conv}\left\{\mathbf{g} \in \mathbb{R}^d : \mathbf{g} = \lim_{\mathbf{x}_k \to \mathbf{x}} \nabla f(\mathbf{x}_k)\right\},$$

which is the convex hull of all limit points of $\nabla f(\mathbf{x}_k)$ over all sequences $\mathbf{x}_1, \mathbf{x}_2, \ldots$ of differentiable points of $f(\cdot)$ which converge to $\mathbf{x}$.

Given this definition of generalized gradients, a *Clarke stationary point* of f is a point $\mathbf{x}$ satisfying $\mathbf{0} \in \partial f(\mathbf{x})$. Then, it is natural to ask if an optimization algorithm can reach an ϵ-stationary point with a finite-time convergence guarantee. Here a point $\mathbf{x} \in \mathbb{R}^d$ is an ϵ-Clarke stationary point if

$$\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial f(\mathbf{x})\} \leq \epsilon.$$

This question has been addressed by [Zhang et al., 2020a, Theorem 1], who showed that finding an ϵ-Clarke stationary points in nonsmooth nonconvex optimization can not be achieved by any finite-time algorithm given a fixed tolerance $\epsilon \in [0, 1)$. One possible response is to consider a relaxation called a *near* ϵ-Clarke stationary point. Consider a point which is

δ-close to an ϵ-stationary point for some $\delta > 0$. A point $\mathbf{x} \in \mathbb{R}^d$ is near ϵ-stationary if the following statement holds true:

$$\min\left\{\|\mathbf{g}\| : \mathbf{g} \in \cup_{\mathbf{y}\in\mathbb{B}_\delta(\mathbf{x})}\partial f(\mathbf{y})\right\} \leq \epsilon.$$

Unfortunately, however, [Kornowski and Shamir, 2021, Theorem 1] demonstrated that it is impossible to obtain worst-case guarantees for finding a near ϵ-Clarke stationary point of $f \in \mathcal{F}_d(\Delta, L)$ when $\epsilon, \delta > 0$ are smaller than some certain constants unless the number of oracle calls has an exponential dependence on the problem dimension $d \geq 1$. These negative results suggest a need for rethinking the definition of targeted stationary points. We propose to consider the refined notion of Goldstein subdifferential.

Definition 9.2.4 *Given a point $\mathbf{x} \in \mathbb{R}^d$ and $\delta > 0$, the δ-Goldstein subdifferential of a Lipschitz function f at $\mathbf{x}$ is given by $\partial_\delta f(\mathbf{x}) := \mathrm{conv}(\cup_{\mathbf{y}\in\mathbb{B}_\delta(\mathbf{x})}\partial f(\mathbf{y}))$.*

The Goldstein subdifferential of f at $\mathbf{x}$ is the convex hull of the union of all generalized gradients at points in a δ-ball around $\mathbf{x}$. Accordingly, we can define the (δ, ϵ)-Goldstein stationary points; that is, a point $\mathbf{x} \in \mathbb{R}^d$ is a (δ, ϵ)-Goldstein stationary point if the following statement holds:

$$\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\mathbf{x})\} \leq \epsilon.$$

It is worth mentioning that (δ, ϵ)-Goldstein stationarity is a weaker notion than (near) ϵ-Clarke stationarity since any (near) ϵ-stationary point is a (δ, ϵ)-Goldstein stationary point but not vice versa. However, the converse holds true under a smoothness condition [Zhang et al., 2020a, Proposition 6] and $\lim_{\delta\downarrow 0}\partial_\delta f(\mathbf{x}) = \partial f(\mathbf{x})$ holds as shown in Zhang et al. [2020a, Lemma 7]. The latter result also enables an intuitive framework for transforming nonasymptotic analysis of convergence to (δ, ϵ)-Goldstein stationary points to classical asymptotic results for finding ϵ-Clarke stationary points. Thus, we conclude that finding a (δ, ϵ)-Goldstein stationary point is a reasonable optimality condition for general nonsmooth nonconvex optimization.

Remark 9.2.5 *Finding a (δ, ϵ)-Goldstein stationary point in nonsmooth nonconvex optimization has been formally shown to be computationally tractable in an oracle model [Zhang et al., 2020a, Davis et al., 2022, Tian et al., 2022]. Goldstein [1977] discovered that one can decrease the function value of a Lipschitz f by using the minimal-norm element of $\partial_\delta f(\mathbf{x})$ and this leads to a deterministic normalized subgradient method which finds a (δ, ϵ)-Goldstein stationary point within $O(\frac{\Delta}{\delta\epsilon})$ iterations. However, Goldstein's algorithm is only conceptual since it is computationally intractable to return an exact minimal-norm element of $\partial_\delta f(\mathbf{x})$. Recently, the randomized variants of Goldstein's algorithm have been proposed with a convergence guarantee of $O(\frac{\Delta L^2}{\delta\epsilon^3})$ [Zhang et al., 2020a, Davis et al., 2022, Tian et al., 2022]. However, it remains unknown if gradient-free methods find a (δ, ϵ)-Goldstein stationary point of a Lipschitz function f within $O(\mathrm{poly}(d, L, \Delta, 1/\epsilon, 1/\delta))$ iterations in the absence of an exact function value oracle. Note that the dependence on the problem dimension $d \geq 1$ is necessary for gradient-free methods.*

Randomized smoothing. The randomized smoothing approaches are simple and work equally well for convex and nonconvex functions. Formally, given the L-Lipschitz function f (possibly nonsmooth nonconvex) and a distribution $\mathbb{P}$, we define $f_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[f(\mathbf{x} + \delta\mathbf{u})]$. In particular, letting $\mathbb{P}$ be a standard Gaussian distribution, the function f_δ is a $\delta L\sqrt{d}$-approximation of $f(\cdot)$ and the gradient ∇f_δ is $\frac{L\sqrt{d}}{\delta}$-Lipschitz where $d \geq 1$ is the problem dimension; see Nesterov and Spokoiny [2017, Theorem 1 and Lemma 2]. Letting $\mathbb{P}$ be an uniform distribution on an unit ball in ℓ_2-norm, the resulting function f_δ is a δL-approximation of $f(\cdot)$ and ∇f_δ is also $\frac{cL\sqrt{d}}{\delta}$-Lipschitz where $d \geq 1$ is the problem dimension; see Yousefian et al. [2012, Lemma 8] and Duchi et al. [2012, Lemma E.2], rephrased as follows.

Proposition 9.2.6 *Let $f_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[f(\mathbf{x} + \delta\mathbf{u})]$ where $\mathbb{P}$ is an uniform distribution on an unit ball in ℓ_2-norm. Assuming that f is L-Lipschitz, we have (i) $|f_\delta(\mathbf{x}) - f(\mathbf{x})| \leq \delta L$, and (ii) f_δ is differentiable and L-Lipschitz with the $\frac{cL\sqrt{d}}{\delta}$-Lipschitz gradient where $c > 0$ is a constant. In addition, there exists a function f for which each of the above bounds are tight simultaneously.*

The randomized smoothing approaches form the basis for developing gradient-free methods [Flaxman et al., 2005, Agarwal et al., 2010, 2013, Ghadimi and Lan, 2013b, Nesterov and Spokoiny, 2017]. Given an access to function values of f, we can compute an unbiased estimate of the gradient of f_δ and plug them into stochastic gradient-based methods. Note that the Lipschitz constant of f_δ depends on the problem dimension $d \geq 1$ with at least a factor of $\sqrt{d}$ for many randomized smoothing approaches [Kornowski and Shamir, 2021, Theorem 2]. This is consistent with the lower bounds for all gradient-free methods in convex and strongly convex optimization [Duchi et al., 2015, Shamir, 2017].

9.3 Main Results

We establish a relationship between the Goldstein subdifferential and the uniform smoothing approach. We propose a gradient-free method (GFM), its stochastic variant (SGFM), and a two-phase version of GFM and SGFM. We analyze these algorithms using the Goldstein subdifferential; we provide the global rate and large-deviation estimates in terms of (δ, ϵ)-Goldstein stationarity.

Linking Goldstein subdifferential to uniform smoothing. Recall that $\partial_\delta f$ and f_δ are defined by $\partial_\delta f(\mathbf{x}) := \mathrm{conv}(\cup_{\mathbf{y}\in\mathbb{B}_\delta(\mathbf{x})}\partial f(\mathbf{y}))$ and $f_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[f(\mathbf{x} + \delta\mathbf{u})]$. It is clear that f is almost everywhere differentiable since f is L-Lipschitz. This implies that $\nabla f_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[\nabla f(\mathbf{x} + \delta\mathbf{u})]$ and demonstrates that $\nabla f_\delta(\mathbf{x})$ can be viewed intuitively as a convex combination of $\nabla f(\mathbf{z})$ over an infinite number of points $\mathbf{z} \in \mathbb{B}_\delta(\mathbf{x})$. As such, it is reasonable to conjecture that $\nabla f_\delta(\mathbf{x}) \in \partial_\delta f(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$. However, the above argument is not a rigorous proof; indeed, we need to justify why $\nabla f_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[\nabla f(\mathbf{x} + \delta\mathbf{u})]$ if f is almost

Algorithm 29 Gradient-Free Method (GFM)

Input: initial point $\mathbf{x}^0 \in \mathbb{R}^d$, stepsize $\eta > 0$, problem dimension $d \geq 1$, smoothing parameter δ and iteration number $T \geq 1$.
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Sample $\mathbf{w}^t \in \mathbb{R}^d$ uniformly from a unit sphere in $\mathbb{R}^d$.
  Compute $\mathbf{g}^t = \frac{d}{2\delta}(f(\mathbf{x}^t + \delta \mathbf{w}^t) - f(\mathbf{x}^t - \delta \mathbf{w}^t))\mathbf{w}^t$.
  Compute $\mathbf{x}^{t+1} = \mathbf{x}^t - \eta \mathbf{g}^t$.
Output: $\mathbf{x}^R$ where $R \in \{0, 1, 2, \ldots, T-1\}$ is uniformly sampled.

Algorithm 30 Two-Phase Gradient-Free Method (2-GFM)

Input: initial point $\mathbf{x}^0 \in \mathbb{R}^d$, stepsize $\eta > 0$, problem dimension $d \geq 1$, smoothing parameter δ, iteration number $T \geq 1$, number of rounds $S \geq 1$ and sample size B.
for $s = 0, 1, 2, \ldots, S-1$ **do**
  Call Algorithm 29 with $\mathbf{x}^0$, η, d, δ and T and let $\bar{\mathbf{x}}_s$ be an output.
for $s = 0, 1, 2, \ldots, S-1$ **do**
  for $k = 0, 1, 2, \ldots, B-1$ **do**
    Sample $\mathbf{w}^k \in \mathbb{R}^d$ uniformly from a unit sphere in $\mathbb{R}^d$.
    Compute $\mathbf{g}_s^k = \frac{d}{2\delta}(f(\bar{\mathbf{x}}_s + \delta \mathbf{w}^k) - f(\bar{\mathbf{x}}_s - \delta \mathbf{w}^k))\mathbf{w}^k$.
  Compute $\mathbf{g}_s = \frac{1}{B}\sum_{k=0}^{B-1} \mathbf{g}_s^k$.
Choose an index $s^\star \in \{0, 1, 2, \ldots, S-1\}$ such that $s^\star = \operatorname{argmin}_{s=0,1,2,\ldots,S-1} \|\mathbf{g}_s\|$.
Output: $\bar{\mathbf{x}}_{s^\star}$.

everywhere differentiable and generalize the idea of a convex combination to include infinite sums. To resolve these issues, we exploit a toolbox due to Rockafellar and Wets [2009].

In the following theorem, we summarize our result.

Theorem 9.3.1 *Suppose that f is L-Lipschitz and let $f_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[f(\mathbf{x} + \delta\mathbf{u})]$, where $\mathbb{P}$ is an uniform distribution on a unit ball in ℓ_2-norm and let $\partial_\delta f$ be a δ-Goldstein subdifferential of f (cf. Definition 9.2.4). Then, we have $\nabla f_\delta(\mathbf{x}) \in \partial_\delta f(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$.*

Theorem 9.3.1 resolves an important question and forms the basis for analyzing our gradient-free methods. Notably, our analysis can be extended to justify other randomized smoothing approaches in nonsmooth nonconvex optimization. For example, Nesterov and Spokoiny [2017] used Gaussian smoothing and estimated the number of iterations required by their methods to output $\hat{\mathbf{x}} \in \mathbb{R}^d$ satisfying $\|\nabla f_\delta(\hat{\mathbf{x}})\| \leq \epsilon$. By modifying the proof of Theorem 9.3.1 and Zhang et al. [2020a, Lemma 7], we can prove that ∇f_δ belongs to Goldstein subdifferential with Gaussian weights and this subdifferential converges to the Clarke subdifferential as $\delta \to 0$. Compared to uniform smoothing and the original Goldstein subdifferential, the proof for Gaussian smoothing is quite long and technical [Nesterov and Spokoiny, 2017, Page 554], and adding Gaussian weights seems unnatural in general.

Gradient-free methods. We analyze a gradient-free method (GFM) and its two-phase version (2-GFM) for optimizing a Lipschitz function. Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a L-Lipschitz function and the smooth version of f is then the function $f_\delta = \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[f(\mathbf{x} + \delta\mathbf{u})]$ where $\mathbb{P}$ is an uniform distribution on an unit ball in ℓ_2-norm. Equipped with Lemma 10 from Shamir [2017], we can compute an unbiased estimator for the gradient $\nabla f_\delta(\mathbf{x}^t)$ using function values.

This leads to the gradient-free method (GFM) in Algorithm 29 that simply performs a one-step gradient descent to obtain $\mathbf{x}^t$. It is worth mentioning that we use a random iteration count R to terminate the execution of Algorithm 29 and this will guarantee that GFM is valid. Indeed, we only derive that $\min_{t=1,2,\ldots,T} \|\nabla f_\delta(\mathbf{x}^t)\| \leq \epsilon$ in the theoretical analysis (see also Nesterov and Spokoiny [2017, Section 7]) and finding the best solution from $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^T\}$ is difficult since the quantity $\|\nabla f_\delta(\mathbf{x}^t)\|$ is unknown. To estimate them using Monte Carlo simulation would incur additional approximation errors and raise some reliability issues. The idea of random output is not new but has been used by Ghadimi and Lan [2013b] for smooth and nonconvex stochastic optimization. Such scheme also gives us a computational gain with a factor of two in expectation.

Theorem 9.3.2 *Suppose that f is L-Lipschitz and let $\delta > 0$ and $0 < \epsilon < 1$. Then, there exists some $T > 0$ such that the output of Algorithm 29 with $\eta = \frac{1}{10}\sqrt{\frac{\delta(\Delta+\delta L)}{cd^{3/2}L^3 T}}$ satisfies that $\mathbb{E}[\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\mathbf{x}^R)\}] \leq \epsilon$ and the total number of calls of the function value oracle is bounded by*

$$O\left(d^{\frac{3}{2}}\left(\frac{L^4}{\epsilon^4} + \frac{\Delta L^3}{\delta\epsilon^4}\right)\right),$$

where $d \geq 1$ is the problem dimension, $L > 0$ is the Lipschtiz parameter of f and $\Delta > 0$ is an upper bound for the initial objective function gap, $f(\mathbf{x}^0) - \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > 0$.

Remark 9.3.3 *Theorem 9.3.2 illustrates the difference between gradient-based and gradient-free methods in nonsmooth nonconvex optimization. Indeed, Davis et al. [2022] has recently proved the rate of $O(\delta^{-1}\epsilon^{-3})$ for a randomized gradient-based method in terms of (δ, ϵ)-Goldstein stationarity. Further, Theorem 9.3.2 demonstrates that nonsmooth nonconvex optimization is likely to be intrinsically harder than all other standard settings. More specifically, the state-of-the-art rate for gradient-free methods is $O(d\epsilon^{-2})$ for nonsmooth convex optimization in terms of objective function value gap [Duchi et al., 2015] and smooth nonconvex optimization in terms of gradient norm [Nesterov and Spokoiny, 2017]. Thus, the dependence on $d \geq 1$ is linear in their bounds yet $d^{\frac{3}{2}}$ in our bound. We believe it is promising to either improve the rate of gradient-free methods or show the impossibility by establishing a lower bound.*

While Theorem 9.3.2 establishes the expected convergence rate guarantee over many runs of Algorithm 29, we are also interested in the large-deviation properties for a single run. Indeed, we hope to establish a complexity bound for computing a $(\delta, \epsilon, \Lambda)$-*solution*; that is, a point $\mathbf{x} \in \mathbb{R}^d$ satisfying $\text{Prob}(\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\mathbf{x})\} \leq \epsilon) \geq 1 - \Lambda$ for some $\delta > 0$ and

$0 < \epsilon, \Lambda < 1$. By Theorem 9.3.2 and Markov's inequality,

$$\text{Prob}\left(\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\mathbf{x}^R)\} \geq \lambda \mathbb{E}[\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\mathbf{x}^R)\}]\right) \leq \tfrac{1}{\lambda}, \quad \text{for all } \lambda > 0,$$

we conclude that the total number of calls of the function value oracle is bounded by

$$O\left(d^{\frac{3}{2}}\left(\frac{L^4}{\Lambda^4\epsilon^4} + \frac{\Delta L^3}{\delta\Lambda^4\epsilon^4}\right)\right). \tag{9.1}$$

This complexity bound is rather pessimistic in terms of its dependence on Λ which is often set to be small in practice. To improve the bound, we combine Algorithm 29 with a post-optimization procedure [Ghadimi and Lan, 2013b], leading to a two-phase gradient-free method (2-GFM), shown in Algorithm 30.

Theorem 9.3.4 *Suppose that f is L-Lipschitz and let $\delta > 0$ and $0 < \epsilon, \Lambda < 1$. Then, there exists some $T, S, B > 0$ such that the output of Algorithm 30 with $\eta = \frac{1}{10}\sqrt{\frac{\delta(\Delta+\delta L)}{cd^{3/2}L^3T}}$ satisfies that $\text{Prob}(\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\bar{\mathbf{x}}_{s^\star})\}] \geq \epsilon) \leq \Lambda$ and the total number of calls of the function value oracle is bounded by*

$$O\left(d^{\frac{3}{2}}\left(\frac{L^4}{\epsilon^4} + \frac{\Delta L^3}{\delta\epsilon^4}\right)\log_2\left(\frac{1}{\Lambda}\right) + \frac{dL^2}{\Lambda\epsilon^2}\log_2\left(\frac{1}{\Lambda}\right)\right),$$

where $d \geq 1$ is the problem dimension, $L > 0$ is the Lipschtiz parameter of f and $\Delta > 0$ is an upper bound for the initial objective function gap, $f(\mathbf{x}^0) - \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > 0$.

Clearly, the bound in Theorem 9.3.4 is significantly smaller than the corresponding one in Eq. (9.1) in terms of the dependence on $1/\Lambda$, demonstrating the power of the post-optimization phase.

Stochastic gradient-free methods. We turn to the analysis of a stochastic gradient-free method (SGFM) and its two-phase version (2-SGFM) for optimizing a Lipschitz function $f(\cdot) = \mathbb{E}_{\xi\in\mathbb{P}_\mu}[F(\cdot, \xi)]$. In contrast to minimizing a deterministic function f, we only have access to the noisy function value $F(\mathbf{x}, \xi)$ at any point $\mathbf{x} \in \mathbb{R}^d$ where a data sample ξ is drawn from a distribution $\mathbb{P}_\mu$. Intuitively, this is a more challenging setup. It has been studied in the setting of optimizing a nonsmooth convex function [Duchi et al., 2015, Nesterov and Spokoiny, 2017] or a smooth nonconvex function [Ghadimi and Lan, 2013b]. As in these papers, we assume that (i) $F(\cdot, \xi)$ is $L(\xi)$-Lipschitz with $\mathbb{E}_{\xi\in\mathbb{P}_\mu}[L^2(\xi)] \leq G^2$ for some $G > 0$ and (ii) $\mathbb{E}[F(\mathbf{x}, \xi^t)] = f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$ where ξ^t is simulated from $\mathbb{P}_\mu$ at the t^{th} iteration.

Despite the noisy function value, we can compute an unbiased estimator of the gradient $\nabla f_\delta(\mathbf{x}^t)$, where $f_\delta = \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[f(\mathbf{x} + \delta\mathbf{u})] = \mathbb{E}_{\mathbf{u}\sim\mathbb{P},\xi\in\mathbb{P}_\mu}[F(\mathbf{x} + \delta\mathbf{u}, \xi)]$. In particular, we have $\hat{\mathbf{g}}^t = \frac{d}{2\delta}(F(\mathbf{x}^t + \delta\mathbf{w}^t, \xi^t) - F(\mathbf{x}^t - \delta\mathbf{w}^t, \xi^t))\mathbf{w}^t$. Clearly, under our assumption, we have

$$\mathbb{E}_{\mathbf{u}\sim\mathbb{P},\xi\in\mathbb{P}_\mu}[\hat{\mathbf{g}}^t] = \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[\mathbb{E}_{\xi\in\mathbb{P}_\mu}[\hat{\mathbf{g}}^t \mid \mathbf{u}]] = \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[\mathbf{g}^t] = \nabla f_\delta(\mathbf{x}^t),$$

Algorithm 31 Stochastic Gradient-Free Method (SGFM)

1: **Input:** initial point $\mathbf{x}^0 \in \mathbb{R}^d$, stepsize $\eta > 0$, problem dimension $d \geq 1$, smoothing parameter δ and iteration number $T \geq 1$.
2: **for** $t = 0, 1, 2, \ldots, T$ **do**
3: Simulate ξ^t from the distribution $\mathbb{P}_\mu$.
4: Sample $\mathbf{w}^t \in \mathbb{R}^d$ uniformly from a unit sphere in $\mathbb{R}^d$.
5: Compute $\hat{\mathbf{g}}^t = \frac{d}{2\delta}(F(\mathbf{x}^t + \delta \mathbf{w}^t, \xi^t) - F(\mathbf{x}^t - \delta \mathbf{w}^t, \xi^t))\mathbf{w}^t$.
6: Compute $\mathbf{x}^{t+1} = \mathbf{x}^t - \eta \mathbf{g}^t$.
7: **Output:** $\mathbf{x}^R$ where $R \in \{0, 1, 2, \ldots, T-1\}$ is uniformly sampled.

where $\mathbf{g}^t$ is defined in Algorithm 29. However, the variance of the estimator $\hat{\mathbf{g}}^t$ can be undesirably large since $F(\cdot, \xi)$ is $L(\xi)$-Lipschitz for a (possibly unbounded) random variable $L(\xi) > 0$. To resolve this issue, we revisit Shamir [2017, Lemma 10] and show that in deriving an upper bound for $\mathbb{E}_{\mathbf{u}\sim\mathbb{P},\xi\in\mathbb{P}_\mu}[\|\hat{\mathbf{g}}^t\|^2]$ it suffices to assume that $\mathbb{E}_{\xi\in\mathbb{P}_\mu}[L^2(\xi)] \leq G^2$ for some constant $G > 0$. The resulting bound achieves a linear dependence in the problem dimension $d > 0$ which is the same as in Shamir [2017, Lemma 10]. Note that the setup with *convex* and $L(\xi)$-Lipschitz functions $F(\cdot, \xi)$ has been considered in Duchi et al. [2015]. However, our estimator is different from their estimator of $\hat{\mathbf{g}}^t = \frac{d}{\delta}(F(\mathbf{x}^t + \delta \mathbf{w}^t, \xi^t) - F(\mathbf{x}^t, \xi^t))\mathbf{w}^t$ which essentially suffers from the quadratic dependence in $d > 0$. It is also necessary to employ a random iteration count R to terminate Algorithm 31.

Theorem 9.3.5 *Suppose that $F(\cdot, \xi)$ is $L(\xi)$-Lipschitz with $\mathbb{E}_{\xi\in\mathbb{P}_\mu}[L^2(\xi)] \leq G^2$ for some $G > 0$ and let $\delta > 0$ and $0 < \epsilon < 1$. Then, there exists some $T > 0$ such that the output of Algorithm 31 with $\eta = \frac{1}{10}\sqrt{\frac{\delta(\Delta+\delta G)}{cd^{3/2}G^3T}}$ satisfies that $\mathbb{E}[\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\mathbf{x}^R)\}] \leq \epsilon$ and the total number of calls of the noisy function value oracle is bounded by*

$$O\left(d^{\frac{3}{2}}\left(\frac{G^4}{\epsilon^4} + \frac{\Delta G^3}{\delta \epsilon^4}\right)\right),$$

where $d \geq 1$ is the problem dimension, $L > 0$ is the Lipschtiz parameter of f and $\Delta > 0$ is an upper bound for the initial objective function gap, $f(\mathbf{x}^0) - \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > 0$.

In the stochastic setting, the gradient-based method achieves the rate of $O(\delta^{-1}\epsilon^{-4})$ for a randomized gradient-based method in terms of (δ, ϵ)-Goldstein stationarity [Davis et al., 2022]. As such, our bound in Theorem 9.3.5 is tight up to the problem dimension $d \geq 1$. Further, the state-of-the-art rate for stochastic gradient-free methods is $O(d\epsilon^{-2})$ for nonsmooth convex optimization in terms of objective function value gap [Duchi et al., 2015] and $O(d\epsilon^{-4})$ for smooth nonconvex optimization in terms of gradient norm [Ghadimi and Lan, 2013b]. Thus, Theorem 9.3.5 demonstrates that nonsmooth nonconvex stochastic optimization is essentially the most difficult one among than all these standard settings.

As in the case of GFM, we hope to establish a complexity bound of SGFM for computing a $(\delta, \epsilon, \Lambda)$-solution. By Theorem 9.3.5 and Markov's inequality, we obtain that the total

Algorithm 32 Two-Phase Stochastic Gradient-Free Method (2-SGFM)

Input: initial point $\mathbf{x}^0 \in \mathbb{R}^d$, stepsize $\eta > 0$, problem dimension $d \geq 1$, smoothing parameter δ, iteration number $T \geq 1$, number of rounds $S \geq 1$ and sample size B.
for $s = 0, 1, 2, \ldots, S-1$ **do**
  Call Algorithm 31 with $\mathbf{x}^0$, η, d, δ and T and let $\bar{\mathbf{x}}_s$ be an output.
for $s = 0, 1, 2, \ldots, S-1$ **do**
  for $k = 0, 1, 2, \ldots, B-1$ **do**
    Simulate ξ^k from the distribution $\mathbb{P}_\mu$.
    Sample $\mathbf{w}^k \in \mathbb{R}^d$ uniformly from a unit sphere in $\mathbb{R}^d$.
    Compute $\hat{\mathbf{g}}_s^k = \frac{d}{2\delta}(F(\bar{\mathbf{x}}_s + \delta \mathbf{w}^k, \delta^k) - F(\bar{\mathbf{x}}_s - \delta \mathbf{w}^k, \delta^k))\mathbf{w}^k$.
  Compute $\hat{\mathbf{g}}_s = \frac{1}{B}\sum_{k=0}^{B-1} \hat{\mathbf{g}}_s^k$.
Choose an index $s^\star \in \{0, 1, 2, \ldots, S-1\}$ such that $s^\star = \operatorname{argmin}_{s=0,1,2,\ldots,S-1} \|\hat{\mathbf{g}}_s\|$.
Output: $\bar{\mathbf{x}}_{s^\star}$.

number of calls of the noisy function value oracle is bounded by

$$O\left(d^{\frac{3}{2}}\left(\frac{G^4}{\Lambda^4\epsilon^4} + \frac{\Delta G^3}{\delta\Lambda^4\epsilon^4}\right)\right). \tag{9.2}$$

We also propose a two-phase stochastic gradient-free method (2-SGFM) in Algorithm 32 by combining Algorithm 31 with a post-optimization procedure.

Theorem 9.3.6 *Suppose that $F(\cdot, \xi)$ is $L(\xi)$-Lipschitz with $\mathbb{E}_{\xi\in\mathbb{P}_\mu}[L^2(\xi)] \leq G^2$ for some $G > 0$ and let $\delta > 0$ and $0 < \epsilon, \Lambda < 1$. Then, there exists some $T, S, B > 0$ such that the output of Algorithm 32 with $\eta = \frac{1}{10}\sqrt{\frac{\delta(\Delta+\delta G)}{cd^{3/2}G^3T}}$ satisfies that $\text{Prob}(\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\bar{\mathbf{x}}_{s^\star})\}] \geq \epsilon) \leq \Lambda$ and the total number of calls of the noisy function value oracle is bounded by*

$$O\left(d^{\frac{3}{2}}\left(\frac{G^4}{\epsilon^4} + \frac{\Delta G^3}{\delta\epsilon^4}\right)\log_2\left(\frac{1}{\Lambda}\right) + \frac{dG^2}{\Lambda\epsilon^2}\log_2\left(\frac{1}{\Lambda}\right)\right),$$

where $d \geq 1$ is the problem dimension, $L > 0$ is the Lipschtiz parameter of f and $\Delta > 0$ is an upper bound for the initial objective function gap $f(\mathbf{x}^0) - \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > 0$.

Further discussions. We remark that the choice of stepsize η in all of our zeroth-order methods depend on Δ, whereas such dependence is not necessary in the first-order setting; see e.g., Zhang et al. [2020a]. Setting the stepsize without any prior knowledge of Δ, our methods can still achieve finite-time convergence guarantees but the order would become worse. This is possibly because the first-order information gives more characterization of the objective function than the zeroth-order information, so that for first-order methods the stepsize can be independent of more problem parameters without sacrificing the bound. A bit on the positive side is that, it suffices for our zeroth-order methods to know an estimate of the upper bound of $\Theta(\Delta)$, which can be done in certain application problems.

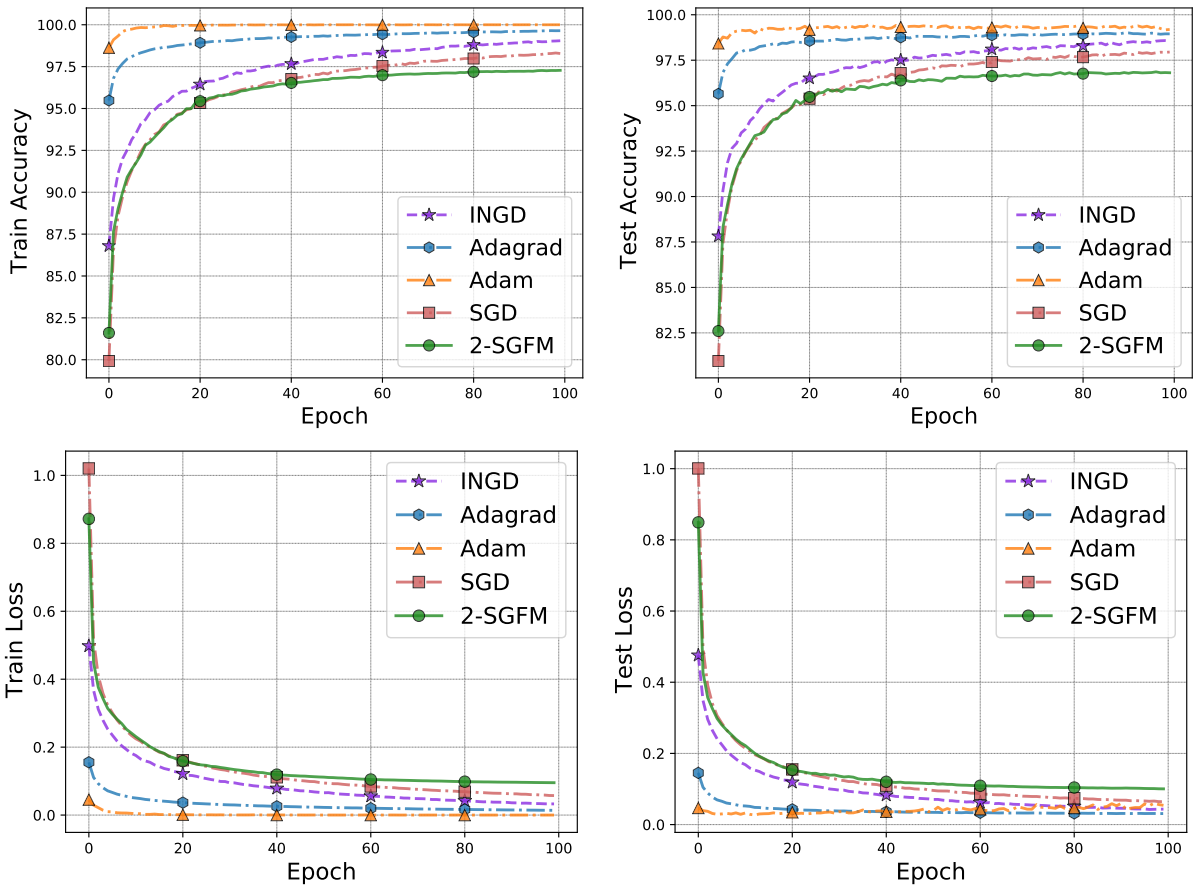


Figure 9.1: Performance of different methods on training CNNs with the MNIST dataset.

Moreover, we highlight that $\delta > 0$ is the desired tolerance in our setting. In fact, (δ, ϵ)-Goldstein stationarity (see Definition 9.2.4) relaxes ϵ-Clarke stationarity and our methods pursue an (δ, ϵ)-stationary point since finding an ϵ-Clarke point is intractable. This is different from smooth optimization where ϵ-Clarke stationarity reduces to $\nabla f(\mathbf{x}) \leq \epsilon$ and becomes tractable. In this context, the existing zeroth-order methods are designed to pursue an ϵ-stationary point. Notably, a (δ, ϵ)-Goldstein stationary point is provably an ϵ-stationary point in smooth optimization if we choose δ that relies on d and ϵ.

9.4 Experiments

We conduct numerical experiments to validate the effectiveness of our proposed methods. In particular, we evaluate the performance of our two-phase version of SGFM (Algorithm 32) on the task of image classification using convolutional neural networks (CNNs) with ReLU activations. The dataset we use is the MNIST dataset[1] [LeCun et al., 1998] and the CNN framework we use is: (i) we set two convolution layers and two fully connected layers where

[1] http://yann.lecun.com/exdb/mnist

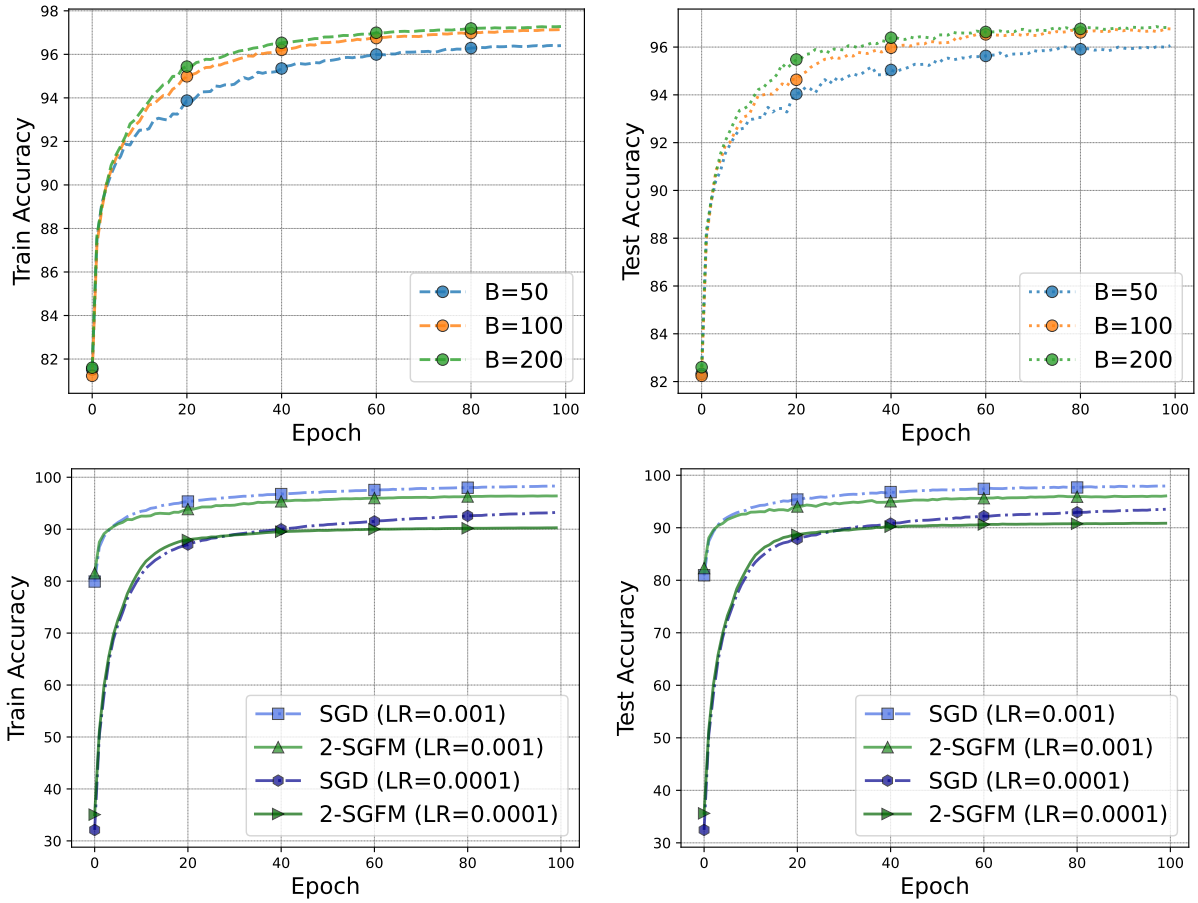


Figure 9.2: (Above) Performance of 2-SGFM with different choices of B. (Bottom) Performance of 2-SGFM and SGD with different choices of learning rates.

the dropout layers [Srivastava et al., 2014] are used before each fully connected layer, and (ii) two convolution layers and the first fully connected layer are associated with ReLU activation. It is worth mentioning that our setup follows the default one[2] and the similar setup was also consider in Zhang et al. [2020a] for evaluating the gradient-based methods.

The baseline approaches include three gradient-based methods: stochastic gradient descent (SGD), Adagrad [Duchi et al., 2011] and Adam [Kingma and Ba, 2015]. We compare these methods with 2-SGFM (cf. Algorithm 32) and set the learning rate η as 0.001. All the experiments are implemented using PyTorch [Paszke et al., 2019] on a workstation with a 2.6 GHz Intel Core i7 and 16GB memory.

Figure 9.1 summarizes the numerical results on the performance of SGD, Adagrad, Adagrad, Adam, INDG [Zhang et al., 2020a], and our method 2-SGFM with $\delta = 0.1$ and $B = 200$. Notably, 2-SGFM is comparable to other gradient-based methods in terms of training/test accuracy/loss even though it only use the function values. This demonstrates the potential value of our methods since the gradient-based methods are not applicable in

[2]https://github.com/pytorch/examples/tree/main/mnist

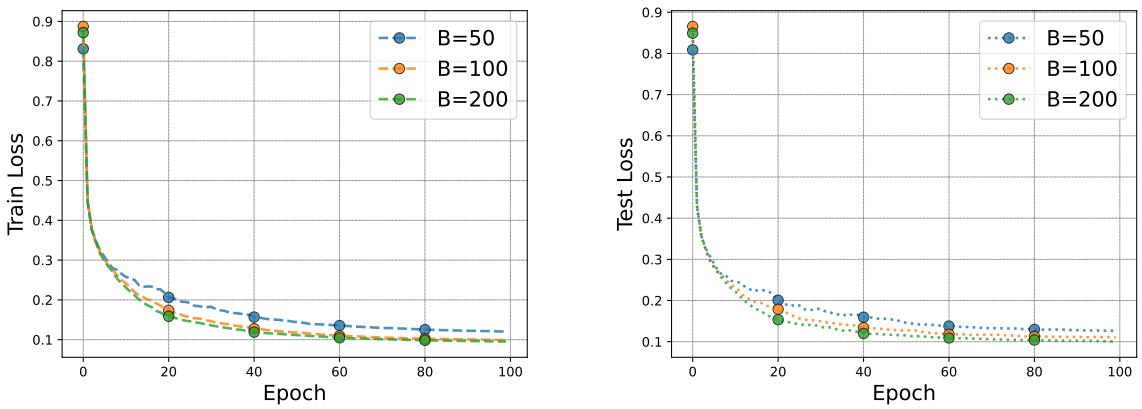


Figure 9.3: Performance of 2-SGFM with different choices of B.

many real-world application problems. Figure 9.2 (Above) presents the effect of batch size $B \geq 1$ in 2-SGFM; indeed, the larger value of B leads to better performance and this accords with Theorem 9.3.6. We compare the performance of SGD and 2-SGFM with different choices of η. From Figure 9.2 (Bottom), we see that SGD and 2-SGFM achieve similar performance in the early stage and converge to solutions with similar quality.

Figure 9.3 summarizes the experimental results on the effect of batch size B for 2-SGFM. Note that the evaluation metrics here are train loss and test loss. It is clear that the larger value of B leads to better performance and this is consistent with the results presented in the main context. Figure 9.4 summarizes the experimental results on the effect of learning rates for 2-SGFM. It is interesting to see that 2-SGFM can indeed benefit from a more aggressive choice of stepsize $\eta > 0$ in practice and the choice of $\eta = 0.0001$ seems to be too conservative.

9.5 Conclusion

We proposed and analyzed a class of deterministic and stochastic gradient-free methods for optimizing a Lipschitz function. Based on the relationship between the Goldstein subdifferential and uniform smoothing that we have established, the proposed GFM and SGFM are proved to return a (δ, ϵ)-Goldstein stationary point at an expected rate of $O(d^{3/2}\delta^{-1}\epsilon^{-4})$. We obtain a large-deviation guarantee and improve it by combining GFM and SGFM with a two-phase scheme. Experiments on training neural networks with the MNIST and CIFAR10 datasets demonstrate the effectiveness of our methods. Future directions include the theory for non-Lipschitz and nonconvex optimization [Bian et al., 2015] and applications of our methods to deep residual neural network (ResNet) [He et al., 2016] and deep dense convolutional network (DenseNet) [Huang et al., 2017].

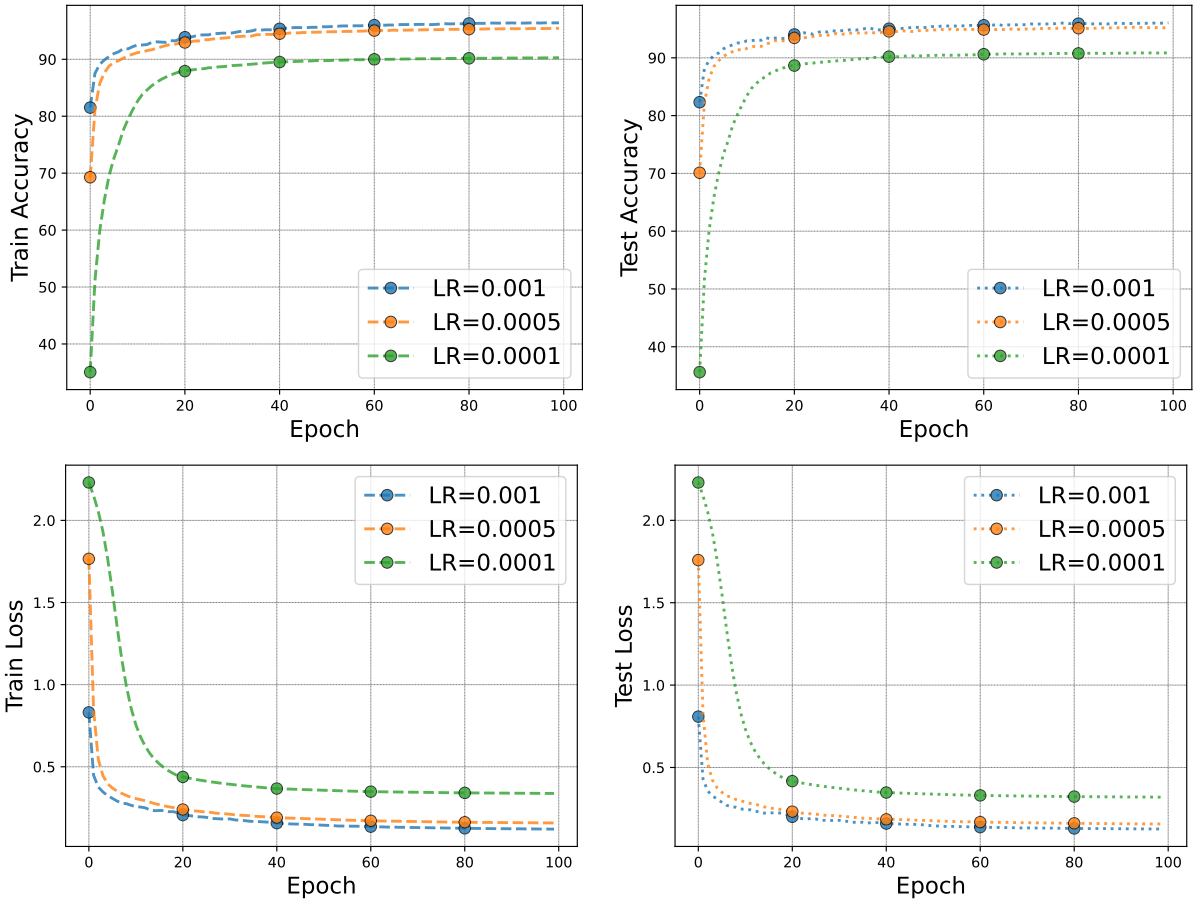


Figure 9.4: Performance of 2-SGFM with different choices of learning rates η.

9.6 Further Related Work on Nonsmooth Nonconvex Optimization

To appreciate the difficulty and the broad scope of the research agenda in nonsmooth nonconvex optimization, we start by describing the existing relevant literature. First, the existing work is mostly devoted to establishing the asymptotic convergence properties of various optimization algorithms, including gradient sampling (GS) methods [Burke et al., 2002a,b, 2005, Kiwiel, 2007, Burke et al., 2020], bundle methods [Kiwiel, 1996, Fuduli et al., 2004] and subgradient methods [Benaïm et al., 2005, Majewski et al., 2018, Davis et al., 2020, Daniilidis and Drusvyatskiy, 2020, Bolte and Pauwels, 2021]. More specifically, Burke et al. [2002a] provided a systematic investigation of approximating the Clarke subdifferential through random sampling and proposed a gradient bundle method [Burke et al., 2002b]—the precursor of GS methods—for optimizing a nonconvex, nonsmooth and non-Lipschitz function. Later, Burke et al. [2005] and Kiwiel [2007] proposed the GS methods by incorporating key modifications into the algorithmic scheme in Burke et al. [2002b] and proved that every cluster point of

the iterates generated by GS methods is a Clarke stationary point. For an overview of GS methods, we refer to Burke et al. [2020]. Another line of works extended the bundle methods to nonsmooth nonconvex optimization by considering either piece-wise linear models embedding possible downward shifting [Kiwiel, 1996] or a mixture of linear pieces that exhibit a convex or concave behavior [Fuduli et al., 2004]. There has been recent progress on analyzing subgradient methods for nonsmooth nonconvex optimization; indeed, the classical subgradient method on Lipschitz functions may fail to asymptotically find any stationary point due to the pathological examples [Daniilidis and Drusvyatskiy, 2020]. Under some additional regularity conditions, Benaïm et al. [2005] proved the asymptotic convergence of stochastic approximation methods from a continuous-time viewpoint and Majewski et al. [2018] generalized these results with proximal and implicit updates. Bolte and Pauwels [2021] justify the automatic differentiation schemes under the nonsmoothness conditions; Davis et al. [2020] proved the asymptotic convergence of classical subgradient methods for a class of Whitney stratifiable functions which include the functions studied in Majewski et al. [2018]. Recently, Zhang et al. [2020a] modified Goldstein's subgradient method [Goldstein, 1977] to optimize a class of Hadamard directionally differentiable function and proved nonasymptotic convergence guarantee. Davis et al. [2022] relaxed the assumption of Hadamard directionally differentiability and showed that another modification of Goldstein's subgradient method could achieve the same finite-time guarantee for any Lipschitz function. Concurrently, Tian et al. [2022] removed the subgradient selection oracle assumption in Zhang et al. [2020a, Assumption 1] and provided the third modification of Goldstein's subgradient method with the same finite-time convergence. Different from these gradient-based methods, we focus on the gradient-free methods in this paper.

We are also aware of many recent works on the algorithmic design in the structured nonsmooth nonconvex optimization. There are two primary settings where the proximal gradient methods is guaranteed to achieve nonasymptotic convergence if the proximal mapping can be efficiently evaluated. The former one considers the objective function with composition structure [Duchi and Ruan, 2018, Drusvyatskiy and Paquette, 2019, Davis and Drusvyatskiy, 2019], while the latter one focuses on composite objective functions with nonsmooth convex component [Bolte et al., 2018, Beck and Hallak, 2020]. However, both of these settings require the weak convexity of objective function and exclude many simple and important nonsmooth nonconvex functions used in the real-world application problems.

9.7 Proof of Proposition 9.2.6

We let $\mathbf{u} \in \mathbb{R}^d$ denote a random variable distributed uniformly on $\mathbb{B}_1(\mathbf{0})$ here. For the first statement, since f is L-Lipschitz, we have

$$|f_\delta(\mathbf{x}) - f(\mathbf{x})| = |\mathbb{E}[f(\mathbf{x} + \delta\mathbf{u}) - f(\mathbf{x})]| \leq \delta L \cdot \mathbb{E}[\|\mathbf{u}\|] \leq \delta L.$$

Then, we proceed to prove the second statement. Indeed, Bertsekas [1973, Proposition 2.4] guarantees that f_δ is everywhere differentiable. Since f is L-Lipschitz, we have

$$|f_\delta(\mathbf{x})-f_\delta(\mathbf{x}')| = |\mathbb{E}[f(\mathbf{x}+\delta\mathbf{u})-f(\mathbf{x}'+\delta\mathbf{u})]| \le L|\mathbb{E}[\|\mathbf{x}-\mathbf{x}'\|]| = L\|\mathbf{x}-\mathbf{x}'\|, \quad \text{for all } \mathbf{x},\mathbf{x}' \in \mathbb{R}^d.$$

It remains to prove that ∇f_δ is Lipschitz. Since f is L-Lipschitz, the Rademacher's theorem guarantees that f is almost everywhere differentiable. This implies that $\nabla f_\delta(\mathbf{x}) = \mathbb{E}[\nabla f(\mathbf{x}+\delta\mathbf{u})]$. Then, we have

$$\begin{aligned}
\|\nabla f_\delta(\mathbf{x}) - \nabla f_\delta(\mathbf{x}')\| &= \|\mathbb{E}[\nabla f(\mathbf{x}+\delta\mathbf{u})] - \mathbb{E}[\nabla f(\mathbf{x}'+\delta\mathbf{u})]\| \\
&= \tfrac{1}{\mathrm{Vol}(\mathbb{B}_1(\mathbf{0}))}\left\|\int_{\mathbf{u}\in\mathbb{B}_1(\mathbf{0})} \nabla f(\mathbf{x}+\delta\mathbf{u})\, d\mathbf{u} - \int_{\mathbf{u}\in\mathbb{B}_1(\mathbf{0})} \nabla f(\mathbf{x}'+\delta\mathbf{u})\, d\mathbf{u}\right\| \\
&= \tfrac{1}{\mathrm{Vol}(\mathbb{B}_\delta(\mathbf{0}))}\left\|\int_{\mathbf{y}\in\mathbb{B}_\delta(\mathbf{x})} \nabla f(\mathbf{y})\, d\mathbf{y} - \int_{\mathbf{y}\in\mathbb{B}_\delta(\mathbf{x}')} \nabla f(\mathbf{y})\, d\mathbf{y}\right\|.
\end{aligned}$$

Note that f is L-Lipschitz, we have $\|\nabla f(\mathbf{y})\| \le L$ for any $\mathbf{y} \in \mathbb{B}_\delta(\mathbf{x}) \cup \mathbb{B}_\delta(\mathbf{x}')$. Then, we turn to prove that $\|\nabla f_\delta(\mathbf{x}) - \nabla f_\delta(\mathbf{x}')\| \le \frac{L\sqrt{d}\|\mathbf{x}-\mathbf{x}'\|}{\delta}$ for two different cases one by one as follows.

Case I: $\|\mathbf{x}-\mathbf{x}'\| \ge 2\delta$**.** It is clear that

$$\|\nabla f_\delta(\mathbf{x}) - \nabla f_\delta(\mathbf{x}')\| \le 2L \le \tfrac{L\|\mathbf{x}-\mathbf{x}'\|}{\delta} \overset{d\ge 1}{\le} \tfrac{L\sqrt{d}\|\mathbf{x}-\mathbf{x}'\|}{\delta},$$

which implies the desired result.

Case II: $\|\mathbf{x}-\mathbf{x}'\| \le 2\delta$**.** It is clear that $\mathbb{B}_\delta(\mathbf{x}) \cap \mathbb{B}_\delta(\mathbf{x}') \ne \emptyset$. This implies that

$$\|\nabla f_\delta(\mathbf{x}) - \nabla f_\delta(\mathbf{x}')\| = \tfrac{1}{\mathrm{Vol}(\mathbb{B}_\delta(\mathbf{0}))}\left\|\int_{\mathbf{y}\in\mathbb{B}_\delta(\mathbf{x})\setminus\mathbb{B}_\delta(\mathbf{x}')} \nabla f(\mathbf{y})\, d\mathbf{y} - \int_{\mathbf{y}\in\mathbb{B}_\delta(\mathbf{x}')\setminus\mathbb{B}_\delta(\mathbf{x})} \nabla f(\mathbf{y})\, d\mathbf{y}\right\|.$$

Since $\|\nabla f(\mathbf{y})\| \le L$ for any $\mathbf{y} \in \mathbb{B}_\delta(\mathbf{x}) \cup \mathbb{B}_\delta(\mathbf{x}')$, we have

$$\|\nabla f_\delta(\mathbf{x}) - \nabla f_\delta(\mathbf{x}')\| \le \tfrac{L}{\mathrm{Vol}(\mathbb{B}_\delta(\mathbf{0}))}\left(\mathrm{Vol}(\mathbb{B}_\delta(\mathbf{x}) \setminus \mathbb{B}_\delta(\mathbf{x}')) + \mathrm{Vol}(\mathbb{B}_\delta(\mathbf{x}') \setminus \mathbb{B}_\delta(\mathbf{x}))\right).$$

By the symmetry from a geometrical point of view, we have $\mathrm{Vol}(\mathbb{B}_\delta(\mathbf{x})\setminus\mathbb{B}_\delta(\mathbf{x}')) = \mathrm{Vol}(\mathbb{B}_\delta(\mathbf{x}')\setminus\mathbb{B}_\delta(\mathbf{x}))$. For simplicity, we let $I = \mathbb{B}_\delta(\mathbf{x}) \setminus \mathbb{B}_\delta(\mathbf{x}')$ and obtain that

$$\|\nabla f_\delta(\mathbf{x}) - \nabla f_\delta(\mathbf{x}')\| \le \tfrac{2L}{\mathrm{Vol}(\mathbb{B}_\delta(\mathbf{0}))}\mathrm{Vol}(I) = \tfrac{2L}{c_d\delta^d}\mathrm{Vol}(I), \quad \text{where } c_d = \tfrac{\pi^{d/2}}{\Gamma(d/2+1)}.$$

It suffices to find an upper bound for $\mathrm{Vol}(I)$ in terms of $\|\mathbf{x}-\mathbf{x}'\|$. Let $V_{cap}(p)$ denote the volume of the spherical cap with the distance p from the center of the sphere, we have

$$\mathrm{Vol}(I) = \mathrm{Vol}(\mathbb{B}_\delta(\mathbf{0})) - 2V_{cap}(\tfrac{1}{2}\|\mathbf{x}-\mathbf{x}'\|) = c_d\delta^d - 2V_{cap}(\tfrac{1}{2}\|\mathbf{x}-\mathbf{x}'\|).$$

The volume of the d-dimensional spherical cap with distance p from the center of the sphere can be calculated in terms of the volumes of $(d-1)$-dimensional spheres as follows:

$$V_{cap}(p) = \int_p^\delta c_{d-1}(\delta^2 - \rho^2)^{\frac{d-1}{2}} d\rho, \quad \text{for all } p \in [0, \delta].$$

Since $V_{cap}(\cdot)$ is a convex function over $[0, \delta]$, we have $V_{cap}(p) \geq V_{cap}(0) + V'_{cap}(0)p$. By the definition, we have $V_{cap}(0) = \frac{1}{2}\text{Vol}(\mathbb{B}_\delta(\mathbf{0})) = \frac{1}{2}c_d\delta^d$ and $V'_{cap}(0) = -c_{d-1}\delta^{d-1}$. Thus, $V_{cap}(p) \geq \frac{1}{2}c_d\delta^d - c_{d-1}\delta^{d-1}p$. Furthermore, $\frac{1}{2}\|\mathbf{x} - \mathbf{x}'\| \in [0, \delta]$. Putting these pieces together yields that $\text{Vol}(I) \leq c_{d-1}\delta^{d-1}\|\mathbf{x} - \mathbf{x}'\|$. Therefore, we conclude that

$$\|\nabla f_\delta(\mathbf{x}) - \nabla f_\delta(\mathbf{x}')\| \leq \frac{2L}{c_d\delta^d}\text{Vol}(I) \leq \frac{2c_{d-1}}{c_d}\frac{L\|\mathbf{x}-\mathbf{x}'\|}{\delta}.$$

Since $c_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$, we have $\frac{2c_{d-1}}{c_d} = \begin{cases} \frac{d!!}{(d-1)!!} & \text{if } d \text{ is odd,} \\ \frac{2}{\pi}\frac{d!!}{(d-1)!!} & \text{otherwise.} \end{cases}$ and $\frac{1}{\sqrt{d}}\frac{2c_{d-1}}{c_d} \to \sqrt{\frac{\pi}{2}}$. Therefore, we conclude that the gradient ∇f_δ is $\frac{cL\sqrt{d}}{\delta}$-Lipschitz where $c > 0$ is a positive constant. In addition, for the construction of a function f in which each of the above bounds are tight, we consider a convex combination of "difficult" functions, in this case

$$f_1(\mathbf{x}) = L\|\mathbf{x}\|, \quad f_2(\mathbf{x}) = L|\langle \mathbf{x}, \frac{\mathbf{w}}{\|\mathbf{w}\|}\rangle - \frac{1}{2}|.$$

and choose $f(\mathbf{x}) = \frac{1}{2}(f_1(\mathbf{x}) + f_2(\mathbf{x}))$. Following up the same argument as in Duchi et al. [2012, Lemma 10], it is relatively straightforward to verify that the bounds in Proposition 9.2.6 cannot be improved by more than a constant factor. This completes the proof.

9.8 Proof of Theorem 9.3.1

We first show that $\nabla f_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[\nabla f(\mathbf{x} + \delta\mathbf{u})]$. Indeed, by the definition of f_δ, we have

$$f_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[f(\mathbf{x} + \delta\mathbf{u})] = \frac{1}{\text{Vol}(\mathbb{B}_1(\mathbf{0}))}\int_{\mathbf{u}\in\mathbb{B}_1(\mathbf{0})} f(\mathbf{x} + \delta\mathbf{u})\, d\mathbf{u} = \frac{1}{\text{Vol}(\mathbb{B}_\delta(\mathbf{0}))}\int_{\mathbf{v}\in\mathbb{B}_\delta(\mathbf{0})} f(\mathbf{x} + \mathbf{v})\, d\mathbf{v}.$$

Since f is L-Lipschitz, Bertsekas [1973, Proposition 2.3] guarantees that f_δ is everywhere differentiable. Thus, we have $\nabla f_\delta(\mathbf{x})$ exists for any $\mathbf{x} \in \mathbb{R}^d$ and satisfies that

$$\lim_{\|\mathbf{h}\|\to 0} \frac{|f_\delta(\mathbf{x}+\mathbf{h}) - f_\delta(\mathbf{x}) - \langle\nabla f_\delta(\mathbf{x}), \mathbf{h}\rangle|}{\|\mathbf{h}\|} = 0. \tag{9.3}$$

Further, we have

$$\frac{f_\delta(\mathbf{x}+\mathbf{h}) - f_\delta(\mathbf{x})}{\|\mathbf{h}\|} = \frac{1}{\text{Vol}(\mathbb{B}_\delta(\mathbf{0}))}\int_{\mathbf{v}\in\mathbb{B}_\delta(\mathbf{0})} \frac{f(\mathbf{x}+\mathbf{h}+\mathbf{v}) - f(\mathbf{x}+\mathbf{v})}{\|\mathbf{h}\|}\, d\mathbf{v}$$

Since f is L-Lipschitz, we have $\frac{f(\mathbf{x}+\mathbf{h}+\mathbf{v})-f(\mathbf{x}+\mathbf{v})}{\|\mathbf{h}\|} \leq L$. By the dominated convergence theorem, we have

$$\lim_{\|\mathbf{h}\|\to 0} \frac{f_\delta(\mathbf{x}+\mathbf{h})-f_\delta(\mathbf{x})}{\|\mathbf{h}\|} = \frac{1}{\mathrm{Vol}(\mathbb{B}_\delta(\mathbf{0}))} \int_{\mathbf{v}\in\mathbb{B}_\delta(\mathbf{0})} \left(\lim_{\|\mathbf{h}\|\to 0} \frac{f(\mathbf{x}+\mathbf{h}+\mathbf{v})-f(\mathbf{x}+\mathbf{v})}{\|\mathbf{h}\|} \right) d\mathbf{v}$$

Furthermore, Rademacher's theorem guarantees that f is almost everywhere differentiable. Letting $U \subseteq \mathbb{B}_\delta(\mathbf{0})$ such that $\mathrm{Vol}(U) = \mathrm{Vol}(\mathbb{B}_\delta(\mathbf{0}))$ and f is differentiable at $\mathbf{x}+\mathbf{v}$ for $\forall \mathbf{v} \in U$, we have

$$\lim_{\|\mathbf{h}\|\to 0} \frac{f_\delta(\mathbf{x}+\mathbf{h})-f_\delta(\mathbf{x})}{\|\mathbf{h}\|} = \frac{1}{\mathrm{Vol}(U)} \int_{\mathbf{v}\in U} \left(\lim_{\|\mathbf{h}\|\to 0} \frac{f(\mathbf{x}+\mathbf{h}+\mathbf{v})-f(\mathbf{x}+\mathbf{v})}{\|\mathbf{h}\|} \right) d\mathbf{v}, \tag{9.4}$$

and

$$\lim_{\|\mathbf{h}\|\to 0} \frac{|f(\mathbf{x}+\mathbf{h}+\mathbf{v})-f(\mathbf{x}+\mathbf{v})-\langle \nabla f(\mathbf{x}+\mathbf{v}), \mathbf{h}\rangle|}{\|\mathbf{h}\|} = 0. \tag{9.5}$$

Combining Eq. (9.3), Eq (9.4) and Eq. (9.5) together yields that

$$\lim_{\|\mathbf{h}\|\to 0} \frac{|\langle \nabla f_\delta(\mathbf{x}) - \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[\nabla f(\mathbf{x}+\delta\mathbf{u})], \mathbf{h}\rangle|}{\|\mathbf{h}\|} = 0.$$

Choosing $\mathbf{h} = t(\nabla f_\delta(\mathbf{x}) - \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[\nabla f(\mathbf{x}+\delta\mathbf{u})])$ with $t \to 0$, we have $\|\nabla f_\delta(\mathbf{x}) - \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[\nabla f(\mathbf{x}+\delta\mathbf{u})]\| = 0$.

It remains to show that $\nabla f_\delta(\mathbf{x}) \in \partial_\delta f(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$ using the proof argument by contradiction. In particular, we assume that there exists $\mathbf{x}_0 \in \mathbb{R}^d$ such that $\nabla f_\delta(\mathbf{x}_0) \notin \partial_\delta f(\mathbf{x}_0)$. Recall that

$$\partial_\delta f(\mathbf{x}_0) := \mathrm{conv}(\cup_{\mathbf{y}\in\mathbb{B}_\delta(\mathbf{x}_0)} \partial f(\mathbf{y})),$$

By the hyperplane separation theorem [Rockafellar and Wets, 2009], there exists a unit vector $\mathbf{g} \in \mathbb{R}^d$ such that $\langle \mathbf{g}, \nabla f_\delta(\mathbf{x}_0)\rangle > 0$ and

$$\langle \mathbf{g}, \xi\rangle \leq 0, \quad \text{for any } \xi \in \cup_{\mathbf{y}\in\mathbb{B}_\delta(\mathbf{x}_0)} \partial f(\mathbf{y}). \tag{9.6}$$

However, we already obtain that $\nabla f_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathbb{P}}[\nabla f(\mathbf{x}+\delta\mathbf{u})]$ which implies that

$$\nabla f_\delta(\mathbf{x}_0) = \frac{1}{\mathrm{Vol}(\mathbb{B}_1(\mathbf{0}))} \int_{\mathbf{u}\in\mathbb{B}_1(\mathbf{0})} \nabla f(\mathbf{x}_0+\delta\mathbf{u})\, d\mathbf{u} = \frac{1}{\mathrm{Vol}(\mathbb{B}_\delta(\mathbf{0}))} \int_{\mathbf{y}\in\mathbb{B}_\delta(\mathbf{x}_0)} \nabla f(\mathbf{y})\, d\mathbf{y}.$$

Thus, Eq. (9.6) implies that $\langle \mathbf{g}, \nabla f_\delta(\mathbf{x}_0)\rangle \leq 0$ which leads to a contradiction. Therefore, we conclude that $\nabla f_\delta(\mathbf{x}) \in \partial_\delta f(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$. This completes the proof.

9.9 Missing Proofs for Gradient-Free Methods

We present some lemmas for analyzing the convergence property of gradient-free method and its two-phase version. We also give the proofs of Theorem 9.3.2 and 9.3.4.

Technical lemmas. We provide two lemmas for analyzing Algorithm 29. The first lemma is a restatement of Shamir [2017, Lemma 10] which gives an upper bound on the quantity $\mathbb{E}[\|\mathbf{g}^t\|^2|\mathbf{x}^t]$ in terms of problem dimension $d \geq 1$ and the Lipschitz parameter $L > 0$. For the sake of completeness, we provide the proof details.

Lemma 9.9.1 *Suppose that f is L-Lipschitz and let $\{\mathbf{g}^t\}_{t=0}^{T-1}$ and $\{\mathbf{x}^t\}_{t=0}^{T-1}$ be generated by Algorithm 29. Then, we have $\mathbb{E}[\mathbf{g}^t|\mathbf{x}^t] = \nabla f_\delta(\mathbf{x}^t)$ and $\mathbb{E}[\|\mathbf{g}^t\|^2|\mathbf{x}^t] \leq 16\sqrt{2\pi}dL^2$.*

Proof. By the definition of $\mathbf{g}^t$ and the symmetry of the distribution of $\mathbf{w}^t$, we have

$$\begin{aligned}\mathbb{E}[\mathbf{g}^t \mid \mathbf{x}^t] &= \mathbb{E}\left[\tfrac{d}{2\delta}(f(\mathbf{x}^t+\delta\mathbf{w}^t) - f(\mathbf{x}^t-\delta\mathbf{w}^t))\mathbf{w}^t \mid \mathbf{x}^t\right] \\ &= \tfrac{1}{2}\left(\mathbb{E}\left[\tfrac{d}{\delta}f(\mathbf{x}^t+\delta\mathbf{w}^t)\mathbf{w}^t \mid \mathbf{x}^t\right] + \mathbb{E}\left[\tfrac{d}{\delta}f(\mathbf{x}^t+\delta(-\mathbf{w}^t))(-\mathbf{w}^t) \mid \mathbf{x}^t\right]\right) \\ &= \tfrac{1}{2}\left(\nabla f_\delta(\mathbf{x}^t) + \nabla f_\delta(\mathbf{x}^t)\right) \;=\; \nabla f_\delta(\mathbf{x}^t).\end{aligned}$$

It remains to show that $\mathbb{E}[\|\mathbf{g}^t\|^2 \mid \mathbf{x}^t] \leq 16\sqrt{2\pi}dL^2$. Indeed, since $\|\mathbf{w}^t\| = 1$, we have

$$\begin{aligned}&\mathbb{E}[\|\mathbf{g}^t\|^2 \mid \mathbf{x}^t] \\ &= \mathbb{E}\left[\tfrac{d^2}{4\delta^2}(f(\mathbf{x}^t+\delta\mathbf{w}^t) - f(\mathbf{x}^t-\delta\mathbf{w}^t))^2\|\mathbf{w}^t\|^2 \mid \mathbf{x}^t\right] \leq \mathbb{E}\left[\tfrac{d^2}{4\delta^2}(f(\mathbf{x}^t+\delta\mathbf{w}^t) - f(\mathbf{x}^t-\delta\mathbf{w}^t))^2 \mid \mathbf{x}^t\right].\end{aligned}$$

Using the elementary inequality $(a-b)^2 \leq 2a^2 + 2b^2$, we have

$$\begin{aligned}&\mathbb{E}[(f(\mathbf{x}^t+\delta\mathbf{w}^t) - f(\mathbf{x}^t-\delta\mathbf{w}^t))^2 \mid \mathbf{x}^t] \\ &= \mathbb{E}[(f(\mathbf{x}^t+\delta\mathbf{w}^t) - \mathbb{E}[f(\mathbf{x}^t+\delta\mathbf{w}^t) \mid \mathbf{x}^t] - (f(\mathbf{x}^t-\delta\mathbf{w}^t) - \mathbb{E}[f(\mathbf{x}^t+\delta\mathbf{w}^t) \mid \mathbf{x}^t]))^2 \mid \mathbf{x}^t] \\ &\leq 2\mathbb{E}[(f(\mathbf{x}^t+\delta\mathbf{w}^t) - \mathbb{E}[f(\mathbf{x}^t+\delta\mathbf{w}^t) \mid \mathbf{x}^t])^2 \mid \mathbf{x}^t] + 2\mathbb{E}[(f(\mathbf{x}^t-\delta\mathbf{w}^t) - \mathbb{E}[f(\mathbf{x}^t+\delta\mathbf{w}^t) \mid \mathbf{x}^t])^2 \mid \mathbf{x}^t].\end{aligned}$$

Since $\mathbf{w}^t$ has a symmetric distribution around the origin, we have

$$\mathbb{E}[(f(\mathbf{x}^t+\delta\mathbf{w}^t) - \mathbb{E}[f(\mathbf{x}^t+\delta\mathbf{w}^t) \mid \mathbf{x}^t])^2 \mid \mathbf{x}^t] = \mathbb{E}[(f(\mathbf{x}^t-\delta\mathbf{w}^t) - \mathbb{E}[f(\mathbf{x}^t+\delta\mathbf{w}^t) \mid \mathbf{x}^t])^2 \mid \mathbf{x}^t].$$

Putting these pieces together yields that

$$\mathbb{E}[\|\mathbf{g}^t\|^2 \mid \mathbf{x}^t] \leq \tfrac{d^2}{\delta^2}\mathbb{E}[(f(\mathbf{x}^t+\delta\mathbf{w}^t) - \mathbb{E}[f(\mathbf{x}^t+\delta\mathbf{w}^t) \mid \mathbf{x}^t])^2 \mid \mathbf{x}^t]. \tag{9.7}$$

For simplicity, we let $h(\mathbf{w}) = f(\mathbf{x}^t+\delta\mathbf{w})$. Since f is L-Lipschitz, this function h is δL-Lipschitz given a fixed $\mathbf{x}^t$. In addition, $\mathbf{w}^t \in \mathbb{R}^d$ is sampled uniformly from a unit sphere. Then, by Wainwright [2019, Proposition 3.11 and Example 3.12], we have

$$\mathbb{P}(|h(\mathbf{w}^t) - \mathbb{E}[h(\mathbf{w}^t)]| \geq \alpha) \leq 2\sqrt{2\pi}e^{-\frac{\alpha^2 d}{8\delta^2 L^2}}.$$

Then, we have

$$\begin{aligned}\mathbb{E}[(h(\mathbf{w}^t) - \mathbb{E}[h(\mathbf{w}^t)])^2] &= \int_0^{+\infty} \mathbb{P}((h(\mathbf{w}^t) - \mathbb{E}[h(\mathbf{w}^t)])^2 \geq \alpha)\, d\alpha \\ &= \int_0^{+\infty} \mathbb{P}(|h(\mathbf{w}^t) - \mathbb{E}[h(\mathbf{w}^t)]| \geq \sqrt{\alpha})\, d\alpha \leq 2\sqrt{2\pi}\int_0^{+\infty} e^{-\frac{\alpha d}{8\delta^2 L^2}}\, d\alpha \\ &= 2\sqrt{2\pi}\cdot\tfrac{8\delta^2L^2}{d} = \tfrac{16\sqrt{2\pi}\delta^2L^2}{d}.\end{aligned}$$

By the definition of h, we have

$$\mathbb{E}[(f(\mathbf{x}^t + \delta\mathbf{w}^t) - \mathbb{E}[f(\mathbf{x}^t + \delta\mathbf{w}^t) \mid \mathbf{x}^t])^2 \mid \mathbf{x}^t] \leq \tfrac{16\sqrt{2\pi}\delta^2 L^2}{d}. \tag{9.8}$$

Combining Eq. (9.7) and Eq. (9.8) yields the desired inequality. □

The second lemma gives a key descent inequality for analyzing Algorithm 29.

Lemma 9.9.2 *Suppose that f is L-Lipschitz and let $\{\mathbf{x}^t\}_{t=0}^{T-1}$ be generated by Algorithm 29. Then, we have*

$$\mathbb{E}[\|\nabla f_\delta(\mathbf{x}^t)\|^2] \leq \tfrac{\mathbb{E}[f_\delta(\mathbf{x}^t)] - \mathbb{E}[f_\delta(\mathbf{x}^{t+1})]}{\eta} + \eta \cdot \tfrac{(8\sqrt{2\pi})cd^{3/2}L^3}{\delta}, \quad \text{for all } 0 \leq t \leq T-1.$$

where $c > 0$ is a constant appearing in the smoothing parameter of f_δ (cf. Proposition 9.2.6).

Proof. By Proposition 9.2.6, we have f_δ is differentiable and L-Lipschitz with the $\frac{cL\sqrt{d}}{\delta}$-Lipschitz gradient where $c > 0$ is a constant. This implies that

$$f_\delta(\mathbf{x}^{t+1}) \leq f_\delta(\mathbf{x}^t) - \eta\langle \nabla f_\delta(\mathbf{x}^t), \mathbf{g}^t\rangle + \tfrac{c\eta^2 L\sqrt{d}}{2\delta}\|\mathbf{g}^t\|^2.$$

Taking the expectation of both sides conditioned on $\mathbf{x}^t$ and using Lemma 9.9.1, we have

$$\begin{aligned}
\mathbb{E}[f_\delta(\mathbf{x}^{t+1}) \mid \mathbf{x}^t] &\leq f_\delta(\mathbf{x}^t) - \eta\langle \nabla f_\delta(\mathbf{x}^t), \mathbb{E}[\mathbf{g}^t \mid \mathbf{x}^t]\rangle + \tfrac{c\eta^2 L\sqrt{d}}{2\delta}\mathbb{E}[\|\mathbf{g}^t\|^2 \mid \mathbf{x}^t] \\
&\leq f_\delta(\mathbf{x}^t) - \eta\|\nabla f_\delta(\mathbf{x}^t)\|^2 + \eta^2 \cdot \tfrac{cL\sqrt{d}}{2\delta} \cdot 16\sqrt{2\pi}dL^2 \\
&= f_\delta(\mathbf{x}^t) - \eta\|\nabla f_\delta(\mathbf{x}^t)\|^2 + \eta^2 \cdot (8\sqrt{2\pi})cd^{3/2}L^3\delta^{-1}.
\end{aligned}$$

Taking the expectation of both sides and rearranging yields that

$$\mathbb{E}[\|\nabla f_\delta(\mathbf{x}^t)\|^2] \leq \tfrac{\mathbb{E}[f_\delta(\mathbf{x}^t)] - \mathbb{E}[f_\delta(\mathbf{x}^{t+1})]}{\eta} + \eta \cdot \tfrac{(8\sqrt{2\pi})cd^{3/2}L^3}{\delta}.$$

This completes the proof. □

We present a proposition which is crucial to deriving the large deviation property.

Proposition 9.9.3 *Suppose that Ω is a polish space with a Borel probability measure $\mathbb{P}$ and let $\{\emptyset, \Omega\} = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots$ be a sequence of filtration. For an integer $N \geq 1$, we define a martingale-difference sequence of Borel functions $\{\zeta_k\}_{k=1}^N \subseteq \mathbb{R}^n$ such that ζ_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[\zeta_k \mid \mathcal{F}_{k-1}] = 0$. Then, if $\mathbb{E}[\|\zeta_k\|^2] \leq \sigma_k^2$ for all $k \geq 1$, we have $\mathbb{E}[\|\sum_{k=1}^N \zeta_k\|^2] \leq \sum_{k=1}^N \sigma_k^2$ and the following statement holds true,*

$$\text{Prob}\left(\left\|\sum_{k=1}^N \zeta_k\right\|^2 \geq \lambda \sum_{k=1}^N \sigma_k^2\right) \leq \tfrac{1}{\lambda}, \quad \text{for all } \lambda \geq 0.$$

This is a general result concerning about the large deviations of vector martingales; see, e.g., Juditsky and Nemirovski [2008, Theorem 2.1] or Ghadimi and Lan [2013b, Lemma 2.3].

Proof of Theorem 9.3.2. Summing up the inequality in Lemma 9.9.2 over $t = 0, 1, 2, \ldots, T-1$ yields that

$$\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_\delta(\mathbf{x}^t)\|^2] \leq \frac{f_\delta(\mathbf{x}^0) - \mathbb{E}[f_\delta(\mathbf{x}^T)]}{\eta} + \eta \cdot \frac{(8\sqrt{2\pi})cd^{3/2}L^3T}{\delta}.$$

By Proposition 9.2.6, we have $f(\mathbf{x}_0) \leq f_\delta(\mathbf{x}_0) \leq f(\mathbf{x}_0) + \delta L$. In addition, we see from the definition of f_δ that $f_\delta(\mathbf{x}) \geq \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$ and thus $\mathbb{E}[f_\delta(\mathbf{x}^T)] \geq \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. Putting these pieces together with $f \in \mathcal{F}_d(\Delta, L)$ yields that

$$f_\delta(\mathbf{x}^0) - \mathbb{E}[f_\delta(\mathbf{x}^T)] \leq f(\mathbf{x}_0) - \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + \delta L \leq \Delta + \delta L.$$

Therefore, we conclude that

$$\frac{1}{T}\left(\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_\delta(\mathbf{x}^t)\|^2]\right) \leq \frac{\Delta+\delta L}{\eta T} + \eta \cdot \frac{(8\sqrt{2\pi})cd^{3/2}L^3}{\delta} \leq \frac{\Delta+\delta L}{\eta T} + \eta \cdot \frac{100cd^{3/2}L^3}{\delta}.$$

Recalling that $\eta = \frac{1}{10}\sqrt{\frac{\delta(\Delta+\delta L)}{cd^{3/2}L^3T}}$, we have

$$\frac{1}{T}\left(\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_\delta(\mathbf{x}^t)\|^2]\right) \leq 20\sqrt{\frac{cd^{3/2}L^3}{T}(L + \tfrac{\Delta}{\delta})}.$$

Since the random count $R \in \{0, 1, 2, \ldots, T-1\}$ is uniformly sampled, we have

$$\mathbb{E}[\|\nabla f_\delta(\mathbf{x}^R)\|^2] = \frac{1}{T}\left(\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_\delta(\mathbf{x}^t)\|^2]\right) \leq 20\sqrt{\frac{cd^{3/2}L^3}{T}(L + \tfrac{\Delta}{\delta})}. \tag{9.9}$$

By Theorem 9.3.1, we have $\nabla f_\delta(\mathbf{x}^R) \in \partial_\delta f(\mathbf{x}^R)$. This together with the above inequality implies that

$$\mathbb{E}[\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\mathbf{x}^R)\}] \leq \mathbb{E}[\|\nabla f_\delta(\mathbf{x}^R)\|] \leq 5\left(\frac{cd^{3/2}L^3}{T}(L + \tfrac{\Delta}{\delta})\right)^{\frac{1}{4}}.$$

Therefore, we conclude that there exists some $T > 0$ such that the output of Algorithm 29 satisfies that $\mathbb{E}[\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\mathbf{x}^R)\}] \leq \epsilon$ and the total number of calling the function value oracles is bounded by

$$O\left(d^{\frac{3}{2}}\left(\frac{L^4}{\epsilon^4} + \frac{\Delta L^3}{\delta\epsilon^4}\right)\right).$$

This completes the proof.

Proof of Theorem 9.3.4. By the definition of $s^\star$ and using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\|\mathbf{g}_{s^\star}\|^2 &= \min_{s=0,1,2,\ldots,S-1} \|\mathbf{g}_s\|^2 \leq \min_{s=0,1,2,\ldots,S-1} \left\{2\|\nabla f_\delta(\bar{\mathbf{x}}_s)\|^2 + 2\|\mathbf{g}_s - \nabla f_\delta(\bar{\mathbf{x}}_s)\|^2\right\} \\ &\leq 2\left(\min_{s=0,1,2,\ldots,S-1} \|\nabla f_\delta(\bar{\mathbf{x}}_s)\|^2 + \max_{s=0,1,2,\ldots,S-1} \|\mathbf{g}_s - \nabla f_\delta(\bar{\mathbf{x}}_s)\|^2\right). \qquad (9.10)\end{aligned}$$

This implies that

$$\begin{aligned}\|\nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\|^2 &\leq 2\|\mathbf{g}_{s^\star}\|^2 + 2\|\mathbf{g}_{s^\star} - \nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\|^2 \qquad (9.11)\\ &\overset{\text{Eq. (9.10)}}{\leq} 4\left(\min_{s=0,1,2,\ldots,S-1} \|\nabla f_\delta(\bar{\mathbf{x}}_s)\|^2\right) + 4\left(\max_{s=0,1,2,\ldots,S-1} \|\mathbf{g}_s - \nabla f_\delta(\bar{\mathbf{x}}_s)\|^2\right) + 2\|\mathbf{g}_{s^\star} - \nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\|^2.\end{aligned}$$

The next step is to provide the probabilistic bounds on all the terms in the right-hand side of Eq. (9.11). In particular, for each $s = 0, 1, 2, \ldots, S-1$, we have $\bar{\mathbf{x}}_s$ is an output obtained by calling Algorithm 29 with $\mathbf{x}^0$, d, δ, T and $\eta = \frac{1}{10}\sqrt{\frac{\delta(\Delta+\delta L)}{cd^{3/2}L^3T}}$. Then, Eq. (9.9) in the proof of Theorem 9.3.2 implies that

$$\mathbb{E}[\|\nabla f_\delta(\bar{\mathbf{x}}_s)\|^2] \leq 20\sqrt{\tfrac{cd^{3/2}L^3}{T}(L + \tfrac{\Delta}{\delta})}.$$

Using the Markov's inequality, we have

$$\text{Prob}\left(\|\nabla f_\delta(\bar{\mathbf{x}}_s)\|^2 \geq 40\sqrt{\tfrac{cd^{3/2}L^3}{T}(L + \tfrac{\Delta}{\delta})}\right) \leq \tfrac{1}{2}.$$

Thus, we have

$$\text{Prob}\left(\min_{s=0,1,2,\ldots,S-1} \|\nabla f_\delta(\bar{\mathbf{x}}_s)\|^2 \geq 40\sqrt{\tfrac{cd^{3/2}L^3}{T}(L + \tfrac{\Delta}{\delta})}\right) \leq 2^{-S}. \qquad (9.12)$$

Further, for each $s = 0, 1, 2, \ldots, S-1$, we have

$$\mathbf{g}_s - \nabla f_\delta(\bar{\mathbf{x}}_s) = \tfrac{1}{B}\sum_{k=0}^{B-1}(\mathbf{g}_s^k - \nabla f_\delta(\bar{\mathbf{x}}_s)).$$

By Lemma 9.9.1, we have $\mathbb{E}[\mathbf{g}_s^t|\bar{\mathbf{x}}_s] = \nabla f_\delta(\bar{\mathbf{x}}_s)$ and $\mathbb{E}[\|\mathbf{g}_s^t\|^2|\bar{\mathbf{x}}_s] \leq 16\sqrt{2\pi}dL^2$. This implies that

$$\mathbb{E}[\mathbf{g}_s^t - \nabla f_\delta(\bar{\mathbf{x}}_s)|\bar{\mathbf{x}}_s] = 0, \quad \mathbb{E}[\|\mathbf{g}_s^t - \nabla f_\delta(\bar{\mathbf{x}}_s)\|^2] \leq 16\sqrt{2\pi}dL^2.$$

This together with Proposition 9.9.3 yields that

$$\text{Prob}\left(\|\mathbf{g}_s - \nabla f_\delta(\bar{\mathbf{x}}_s)\|^2 \geq \tfrac{\lambda(16\sqrt{2\pi}dL^2)}{B}\right) = \text{Prob}\left(\left\|\sum_{k=0}^{B-1}(\mathbf{g}_s^k - \nabla f_\delta(\bar{\mathbf{x}}_s))\right\|^2 \geq \lambda B(16\sqrt{2\pi}dL^2)\right) \leq \tfrac{1}{\lambda}.$$

Therefore, we conclude that

$$\text{Prob}\left(\max_{s=0,1,2,\ldots,S-1}\|\mathbf{g}_s-\nabla f_\delta(\bar{\mathbf{x}}_s)\|^2\geq\tfrac{\lambda(16\sqrt{2\pi}dL^2)}{B}\right)\leq\tfrac{S}{\lambda}. \tag{9.13}$$

By the similar argument, we have

$$\text{Prob}(\|\mathbf{g}_{s^\star}-\nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\|^2\geq\tfrac{\lambda(16\sqrt{2\pi}dL^2)}{B})\leq\tfrac{1}{\lambda}. \tag{9.14}$$

Combining Eq. (9.11), Eq. (9.12), Eq. (9.13) and Eq. (9.14) yields that

$$\text{Prob}\left(\|\nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\|^2\geq 160\sqrt{\tfrac{cd^{3/2}L^3}{T}(L+\tfrac{\Delta}{\delta})}+\tfrac{\lambda(96\sqrt{2\pi}dL^2)}{B}\right)\leq\tfrac{S+1}{\lambda}+2^{-S},\quad\text{for all }\lambda>0. \tag{9.15}$$

We set $\lambda=\frac{2(S+1)}{\Lambda}$ and the parameters (T,S,B) as follows,

$$T=cd^{3/2}L^3(L+\tfrac{\Delta}{\delta})(\tfrac{160}{\epsilon^2})^2,\quad S=\lceil\log_2(\tfrac{2}{\Lambda})\rceil,\quad B=\tfrac{(384\sqrt{2\pi}dL^2)(S+1)}{\Lambda\epsilon^2}.$$

Then, we have

$$\text{Prob}\left(\|\nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\|^2\geq\epsilon^2\right)\leq\text{Prob}\left(\|\nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\|^2\geq 160\sqrt{\tfrac{cd^{3/2}L^3}{T}(L+\tfrac{\Delta}{\delta})}+\tfrac{\lambda(96\sqrt{2\pi}dL^2)}{B}\right)\leq\Lambda.$$

By Theorem 9.3.1, we have $\nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\in\partial_\delta f(\bar{\mathbf{x}}_{s^\star})$. This together with the above inequality implies that there exists some $T,S,B>0$ such that the output of Algorithm 30 satisfies that $\mathbb{E}[\min\{\|\mathbf{g}\|:\mathbf{g}\in\partial_\delta f(\bar{\mathbf{x}}_{s^\star})\}]\leq\epsilon$ and the total number of calling the function value oracles is bounded by

$$O\left(d^{\frac{3}{2}}\left(\frac{L^4}{\epsilon^4}+\frac{\Delta L^3}{\delta\epsilon^4}\right)\log_2\left(\frac{1}{\Lambda}\right)+\frac{dL^2}{\Lambda\epsilon^2}\log_2\left(\frac{1}{\Lambda}\right)\right).$$

This completes the proof.

9.10 Missing Proofs for Stochastic Gradient-Free Methods

We present some lemmas for analyzing the convergence property of stochastic gradient-free method and its two-phase version. We also give the proofs of Theorem 9.3.5 and 9.3.6.

Technical lemmas. We provide two lemmas for analyzing Algorithm 31. The first lemma gives an upper bound on the quantity $\mathbb{E}[\|\hat{\mathbf{g}}^t\|^2|\mathbf{x}^t]$ in terms of problem dimension $d\geq 1$ and the constant $G>0$. The proof is based on a modification of the proof of Lemma 9.9.1.

Lemma 9.10.1 *Suppose that $\{\hat{\mathbf{g}}^t\}_{t=0}^{T-1}$ and $\{\mathbf{x}^t\}_{t=0}^{T-1}$ are generated by Algorithm 31. Then, we have $\mathbb{E}[\hat{\mathbf{g}}^t|\mathbf{x}^t]=\nabla f_\delta(\mathbf{x}^t)$ and $\mathbb{E}[\|\hat{\mathbf{g}}^t\|^2|\mathbf{x}^t]\leq 16\sqrt{2\pi}dG^2$.*

Proof. By the definition of $\hat{\mathbf{g}}^t$ and the symmetry of the distribution of $\mathbf{w}^t$, we have

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{g}}^t \mid \mathbf{x}^t] &= \mathbb{E}\left[\tfrac{d}{2\delta}(F(\mathbf{x}^t+\delta\mathbf{w}^t,\xi^t)-F(\mathbf{x}^t-\delta\mathbf{w}^t,\xi^t))\mathbf{w}^t \mid \mathbf{x}^t\right]\\ &= \tfrac{1}{2}\left(\mathbb{E}\left[\tfrac{d}{\delta}F(\mathbf{x}^t+\delta\mathbf{w}^t,\xi^t)\mathbf{w}^t \mid \mathbf{x}^t\right]+\mathbb{E}\left[\tfrac{d}{\delta}F(\mathbf{x}^t+\delta(-\mathbf{w}^t),\xi^t)(-\mathbf{w}^t) \mid \mathbf{x}^t\right]\right)\\ &= \mathbb{E}\left[\tfrac{d}{\delta}F(\mathbf{x}^t+\delta\mathbf{w}^t,\xi^t)\mathbf{w}^t \mid \mathbf{x}^t\right].\end{aligned}$$

By the tower property, we have

$$\mathbb{E}[\hat{\mathbf{g}}^t \mid \mathbf{x}^t] = \mathbb{E}\left[\tfrac{d}{\delta}\mathbb{E}[F(\mathbf{x}^t+\delta\mathbf{w}^t,\xi^t)\mathbf{w}^t \mid \mathbf{x}^t,\mathbf{w}^t] \mid \mathbf{x}^t\right] = \mathbb{E}\left[\tfrac{d}{\delta}f(\mathbf{x}^t+\delta\mathbf{w}^t)\mathbf{w}^t \mid \mathbf{x}^t\right] = \nabla f_\delta(\mathbf{x}^t).$$

It remains to show that $\mathbb{E}[\|\hat{\mathbf{g}}^t\|^2 \mid \mathbf{x}^t] \le 16\sqrt{2\pi}dG^2$. Indeed, by using the same argument as used in the proof of Lemma 9.9.1, we have

$$\mathbb{E}[\|\hat{\mathbf{g}}^t\|^2|\mathbf{x}^t] \le \tfrac{d^2}{\delta^2}\mathbb{E}[(F(\mathbf{x}^t+\delta\mathbf{w}^t,\xi^t)-\mathbb{E}[F(\mathbf{x}^t+\delta\mathbf{w}^t,\xi^t) \mid \mathbf{x}^t,\xi^t])^2 \mid \mathbf{x}^t]. \tag{9.16}$$

For simplicity, we let $h(\mathbf{w}) = F(\mathbf{x}^t+\delta\mathbf{w},\xi^t)$. Since $F(\cdot,\xi)$ is $L(\xi)$-Lipschitz, this function h is $\delta L(\xi^t)$-Lipschitz given a fixed $\mathbf{x}^t$ and ξ^t. In addition, $\mathbf{w}^t \in \mathbb{R}^d$ is sampled uniformly from a unit sphere. Then, by Wainwright [2019, Proposition 3.11 and Example 3.12], we have

$$\mathbb{P}(|h(\mathbf{w}^t)-\mathbb{E}[h(\mathbf{w}^t)]| \ge \alpha) \le 2\sqrt{2\pi}e^{-\frac{\alpha^2 d}{8\delta^2 L(\xi^t)^2}}.$$

Then, we have

$$\begin{aligned}\mathbb{E}[(h(\mathbf{w}^t)-\mathbb{E}[h(\mathbf{w}^t)])^2] &= \int_0^{+\infty}\mathbb{P}((h(\mathbf{w}^t)-\mathbb{E}[h(\mathbf{w}^t)])^2 \ge \alpha)\,d\alpha\\ &= \int_0^{+\infty}\mathbb{P}(|h(\mathbf{w}^t)-\mathbb{E}[h(\mathbf{w}^t)]| \ge \sqrt{\alpha})\,d\alpha \le 2\sqrt{2\pi}\int_0^{+\infty}e^{-\frac{\alpha d}{8\delta^2 L(\xi^t)^2}}\,d\alpha\\ &= 2\sqrt{2\pi}\cdot\tfrac{8\delta^2 L(\xi^t)^2}{d} = \tfrac{16\sqrt{2\pi}\delta^2 L(\xi^t)^2}{d}.\end{aligned}$$

By the definition of h, we have

$$\mathbb{E}[(F(\mathbf{x}^t+\delta\mathbf{w}^t,\xi^t)-\mathbb{E}[F(\mathbf{x}^t+\delta\mathbf{w}^t,\xi^t) \mid \mathbf{x}^t,\xi^t])^2 \mid \mathbf{x}^t] \le \tfrac{16\sqrt{2\pi}\delta^2}{d}\mathbb{E}[L(\xi^t)^2].$$

Since ξ^t is simulated from the distribution $\mathbb{P}_\mu$, we have $\mathbb{E}[L(\xi^t)^2] \le G^2$. Plugging this into the above inequality, we have

$$\mathbb{E}[(F(\mathbf{x}^t+\delta\mathbf{w}^t,\xi^t)-\mathbb{E}[F(\mathbf{x}^t+\delta\mathbf{w}^t,\xi^t) \mid \mathbf{x}^t,\xi^t])^2 \mid \mathbf{x}^t] \le \tfrac{16\sqrt{2\pi}\delta^2 G^2}{d} \tag{9.17}$$

Combining Eq. (9.16) and Eq. (9.17) yields the desired inequality. □

The second lemma gives a key descent inequality for analyzing Algorithm 31.

Lemma 9.10.2 *Suppose that $\{\mathbf{x}^t\}_{t=0}^{T-1}$ are generated by Algorithm 31. Then, we have*

$$\mathbb{E}[\|\nabla f_\delta(\mathbf{x}^t)\|^2] \le \tfrac{\mathbb{E}[f_\delta(\mathbf{x}^t)]-\mathbb{E}[f_\delta(\mathbf{x}^{t+1})]}{\eta} + \eta\cdot\tfrac{(8\sqrt{2\pi})cd^{3/2}G^3}{\delta}, \quad \text{for all } 0 \le t \le T-1.$$

Proof. Since $f(\cdot) = \mathbb{E}_{\xi\in\mathbb{P}_\mu}[F(\cdot,\xi)]$ and $F(\cdot,\xi)$ is $L(\xi)$-Lipschitz with $\mathbb{E}_{\xi\in\mathbb{P}_\mu}[L^2(\xi)] \le G^2$ for some $G>0$, we have f is G-Lipschitz. Then, by Proposition 9.2.6, we have f_δ is differentiable with the $\frac{cG\sqrt{d}}{\delta}$-Lipschitz gradient where $c>0$ is a constant. This implies that

$$f_\delta(\mathbf{x}^{t+1}) \le f_\delta(\mathbf{x}^t) - \eta\langle\nabla f_\delta(\mathbf{x}^t), \hat{\mathbf{g}}^t\rangle + \tfrac{c\eta^2 G\sqrt{d}}{2\delta}\|\hat{\mathbf{g}}^t\|^2.$$

Taking the expectation of both sides conditioned on $\mathbf{x}^t$ and using Lemma 9.10.1, we have

$$\begin{aligned}\mathbb{E}[f_\delta(\mathbf{x}^{t+1}) \mid \mathbf{x}^t] &\le f_\delta(\mathbf{x}^t) - \eta\langle\nabla f_\delta(\mathbf{x}^t), \mathbb{E}[\hat{\mathbf{g}}^t \mid \mathbf{x}^t]\rangle + \tfrac{c\eta^2 G\sqrt{d}}{2\delta}\mathbb{E}[\|\hat{\mathbf{g}}^t\|^2 \mid \mathbf{x}^t]\\ &\le f_\delta(\mathbf{x}^t) - \eta\|\nabla f_\delta(\mathbf{x}^t)\|^2 + \eta^2\cdot\tfrac{cG\sqrt{d}}{2\delta}\cdot 16\sqrt{2\pi}dG^2\\ &= f_\delta(\mathbf{x}^t) - \eta\|\nabla f_\delta(\mathbf{x}^t)\|^2 + \eta^2\cdot\tfrac{(8\sqrt{2\pi})cd^{3/2}G^3}{\delta}.\end{aligned}$$

Taking the expectation of both sides and rearranging yields that

$$\mathbb{E}[\|\nabla f_\delta(\mathbf{x}^t)\|^2] \le \tfrac{\mathbb{E}[f_\delta(\mathbf{x}^t)]-\mathbb{E}[f_\delta(\mathbf{x}^{t+1})]}{\eta} + \eta\cdot\tfrac{(8\sqrt{2\pi})cd^{3/2}G^3}{\delta}.$$

This completes the proof. □

Proof of Theorem 9.3.5. Summing up the inequality in Lemma 9.10.2 over $t=0,1,2,\ldots,T-1$ yields that

$$\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_\delta(\mathbf{x}^t)\|^2] \le \tfrac{f_\delta(\mathbf{x}^0)-\mathbb{E}[f_\delta(\mathbf{x}^T)]}{\eta} + \eta\cdot\tfrac{(8\sqrt{2\pi})cd^{3/2}G^3T}{\delta}.$$

Since $f(\cdot) = \mathbb{E}_{\xi\in\mathbb{P}_\mu}[F(\cdot,\xi)]$ and $F(\cdot,\xi)$ is $L(\xi)$-Lipschitz with $\mathbb{E}_{\xi\in\mathbb{P}_\mu}[L^2(\xi)] \le G^2$ for some $G>0$, we have f is G-Lipschitz. Thus, we have $f\in\mathcal{F}_d(\Delta,L)$. By using the same argument as used in the proof of Theorem 9.3.2, we have

$$\tfrac{1}{T}\left(\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_\delta(\mathbf{x}^t)\|^2]\right) \le \tfrac{\Delta+\delta G}{\eta T} + \eta\cdot\tfrac{(8\sqrt{2\pi})cd^{3/2}G^3}{\delta} \le \tfrac{\Delta+\delta G}{\eta T} + \eta\cdot\tfrac{100cd^{3/2}G^3}{\delta}.$$

Recalling that $\eta = \frac{1}{10}\sqrt{\frac{\delta(\Delta+\delta G)}{cd^{3/2}G^3T}}$, we have

$$\tfrac{1}{T}\left(\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_\delta(\mathbf{x}^t)\|^2]\right) \le 20\sqrt{\tfrac{cd^{3/2}G^3}{T}(G+\tfrac{\Delta}{\delta})}.$$

Since the random count $R\in\{0,1,2,\ldots,T-1\}$ is uniformly sampled, we have

$$\mathbb{E}[\|\nabla f_\delta(\mathbf{x}^R)\|^2] = \tfrac{1}{T}\left(\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_\delta(\mathbf{x}^t)\|^2]\right) \le 20\sqrt{\tfrac{cd^{3/2}G^3}{T}(G+\tfrac{\Delta}{\delta})}. \tag{9.18}$$

By Theorem 9.3.1, we have $\nabla f_\delta(\mathbf{x}^R) \in \partial_\delta f(\mathbf{x}^R)$. This together with the above inequality implies that

$$\mathbb{E}[\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\mathbf{x}^R)\}] \leq \mathbb{E}[\|\nabla f_\delta(\mathbf{x}^R)\|] \leq 5\left(\tfrac{cd^{3/2}G^3}{T}(G + \tfrac{\Delta}{\delta})\right)^{\frac{1}{4}}.$$

Therefore, we conclude that there exists some $T > 0$ such that the output of Algorithm 31 satisfies that $\mathbb{E}[\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\mathbf{x}^R)\}] \leq \epsilon$ and the total number of calling the function value oracles is bounded by

$$O\left(d^{\frac{3}{2}}\left(\frac{G^4}{\epsilon^4} + \frac{\Delta G^3}{\delta\epsilon^4}\right)\right).$$

This completes the proof.

Proof of Theorem 9.3.6. By the definition of $s^\star$ and using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\|\hat{\mathbf{g}}_{s^\star}\|^2 &= \min_{s=0,1,2,\ldots,S-1} \|\hat{\mathbf{g}}_s\|^2 \leq \min_{s=0,1,2,\ldots,S-1} \left\{2\|\nabla f_\delta(\bar{\mathbf{x}}_s)\|^2 + 2\|\hat{\mathbf{g}}_s - \nabla f_\delta(\bar{\mathbf{x}}_s)\|^2\right\} \quad (9.19)\\ &\leq 2\left(\min_{s=0,1,2,\ldots,S-1} \|\nabla f_\delta(\bar{\mathbf{x}}_s)\|^2 + \max_{s=0,1,2,\ldots,S-1} \|\hat{\mathbf{g}}_s - \nabla f_\delta(\bar{\mathbf{x}}_s)\|^2\right).\end{aligned}$$

This implies that

$$\begin{aligned}\|\nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\|^2 &\leq 2\|\hat{\mathbf{g}}_{s^\star}\|^2 + 2\|\hat{\mathbf{g}}_{s^\star} - \nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\|^2 \quad (9.20)\\ &\overset{\text{Eq. (9.19)}}{\leq} 4\left(\min_{s=0,1,2,\ldots,S-1} \|\nabla f_\delta(\bar{\mathbf{x}}_s)\|^2\right) + 4\left(\max_{s=0,1,2,\ldots,S-1} \|\hat{\mathbf{g}}_s - \nabla f_\delta(\bar{\mathbf{x}}_s)\|^2\right) + 2\|\hat{\mathbf{g}}_{s^\star} - \nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\|^2.\end{aligned}$$

The next step is to provide the probabilistic bounds on all the terms in the right-hand side of Eq. (9.20). In particular, for each $s = 0, 1, 2, \ldots, S-1$, we have $\bar{\mathbf{x}}_s$ is an output obtained by calling Algorithm 31 with $\mathbf{x}^0$, d, δ, T and $\eta = \frac{1}{10}\sqrt{\frac{\delta(\Delta+\delta G)}{cd^{3/2}G^3T}}$. Then, Eq. (9.18) in the proof of Theorem 9.3.5 implies that

$$\mathbb{E}[\|\nabla f_\delta(\bar{\mathbf{x}}_s)\|^2] \leq 20\sqrt{\tfrac{cd^{3/2}G^3}{T}(G + \tfrac{\Delta}{\delta})}.$$

Using the Markov's inequality, we have

$$\text{Prob}\left(\|\nabla f_\delta(\bar{\mathbf{x}}_s)\|^2 \geq 40\sqrt{\tfrac{cd^{3/2}G^3}{T}(G + \tfrac{\Delta}{\delta})}\right) \leq \tfrac{1}{2}.$$

Thus, we have

$$\text{Prob}\left(\min_{s=0,1,2,\ldots,S-1} \|\nabla f_\delta(\bar{\mathbf{x}}_s)\|^2 \geq 40\sqrt{\tfrac{cd^{3/2}G^3}{T}(G + \tfrac{\Delta}{\delta})}\right) \leq 2^{-S}. \quad (9.21)$$

Further, for each $s = 0, 1, 2, \ldots, S-1$, we have

$$\hat{\mathbf{g}}_s - \nabla f_\delta(\bar{\mathbf{x}}_s) = \frac{1}{B}\sum_{k=0}^{B-1}(\hat{\mathbf{g}}_s^k - \nabla f_\delta(\bar{\mathbf{x}}_s)).$$

By Lemma 9.10.1, we have $\mathbb{E}[\hat{\mathbf{g}}_s^t|\bar{\mathbf{x}}_s] = \nabla f_\delta(\bar{\mathbf{x}}_s)$ and $\mathbb{E}[\|\hat{\mathbf{g}}_s^t\|^2|\bar{\mathbf{x}}_s] \leq 16\sqrt{2\pi}dG^2$. This implies that

$$\mathbb{E}[\hat{\mathbf{g}}_s^t - \nabla f_\delta(\bar{\mathbf{x}}_s)|\bar{\mathbf{x}}_s] = 0, \quad \mathbb{E}[\|\hat{\mathbf{g}}_s^t - \nabla f_\delta(\bar{\mathbf{x}}_s)\|^2] \leq 16\sqrt{2\pi}dG^2.$$

This together with Proposition 9.9.3 yields that

$$\text{Prob}\left(\|\hat{\mathbf{g}}_s - \nabla f_\delta(\bar{\mathbf{x}}_s)\|^2 \geq \tfrac{\lambda(16\sqrt{2\pi}dG^2)}{B}\right) = \text{Prob}\left(\left\|\sum_{k=0}^{B-1}(\hat{\mathbf{g}}_s^k - \nabla f_\delta(\bar{\mathbf{x}}_s))\right\|^2 \geq \lambda B(16\sqrt{2\pi}dG^2)\right) \leq \tfrac{1}{\lambda}.$$

Therefore, we conclude that

$$\text{Prob}\left(\max_{s=0,1,2,\ldots,S-1}\|\hat{\mathbf{g}}_s - \nabla f_\delta(\bar{\mathbf{x}}_s)\|^2 \geq \tfrac{\lambda(16\sqrt{2\pi}dG^2)}{B}\right) \leq \tfrac{S}{\lambda}. \tag{9.22}$$

By the similar argument, we have

$$\text{Prob}(\|\hat{\mathbf{g}}_{s^\star} - \nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\|^2 \geq \tfrac{\lambda(16\sqrt{2\pi}dG^2)}{B}) \leq \tfrac{1}{\lambda}. \tag{9.23}$$

Combining Eq. (9.20), Eq. (9.21), Eq. (9.22) and Eq. (9.23) yields that

$$\text{Prob}\left(\|\nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\|^2 \geq 160\sqrt{\tfrac{cd^{3/2}G^3}{T}(G + \tfrac{\Delta}{\delta})} + \tfrac{\lambda(96\sqrt{2\pi}dG^2)}{B}\right) \leq \tfrac{S+1}{\lambda} + 2^{-S}, \quad \text{for all } \lambda > 0. \tag{9.24}$$

We set $\lambda = \frac{2(S+1)}{\Lambda}$ and the parameters (T, S, B) as follows,

$$T = cd^{3/2}G^3(G + \tfrac{\Delta}{\delta})(\tfrac{160}{\epsilon^2})^2, \quad S = \lceil \log_2(\tfrac{2}{\Lambda})\rceil, \quad B = \tfrac{(384\sqrt{2\pi}dG^2)(S+1)}{\Lambda\epsilon^2}.$$

Then, we have

$$\text{Prob}\left(\|\nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\|^2 \geq \epsilon^2\right) \leq \text{Prob}\left(\|\nabla f_\delta(\bar{\mathbf{x}}_{s^\star})\|^2 \geq 160\sqrt{\tfrac{cd^{3/2}G^3}{T}(G + \tfrac{\Delta}{\delta})} + \tfrac{\lambda(96\sqrt{2\pi}dG^2)}{B}\right) \leq \Lambda.$$

By Theorem 9.3.1, we have $\nabla f_\delta(\bar{\mathbf{x}}_{s^\star}) \in \partial_\delta f(\bar{\mathbf{x}}_{s^\star})$. This together with the above inequality implies that there exists some $T, S, B > 0$ such that the output of Algorithm 32 satisfies that $\mathbb{E}[\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\bar{\mathbf{x}}_{s^\star})\}] \leq \epsilon$ and the total number of calling the function value oracles is bounded by

$$O\left(d^{\frac{3}{2}}\left(\frac{G^4}{\epsilon^4} + \frac{\Delta G^3}{\delta\epsilon^4}\right)\log_2\left(\frac{1}{\Lambda}\right) + \frac{dG^2}{\Lambda\epsilon^2}\log_2\left(\frac{1}{\Lambda}\right)\right).$$

This completes the proof.

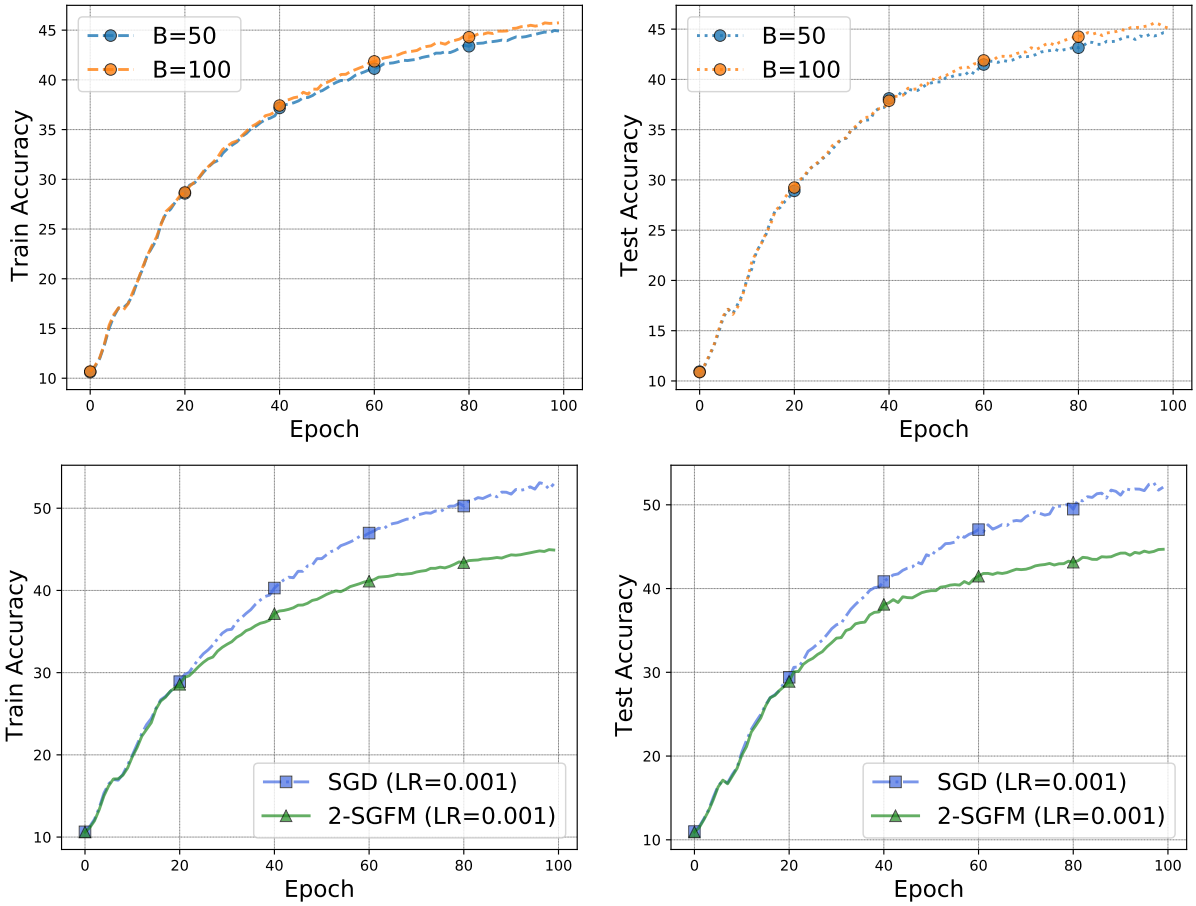


Figure 9.5: Addional experimental results on the CIFAR10 dataset [Krizhevsky and Hinton, 2009]. (Above) Performance of 2-SGFM with different choices of B. (Bottom) Performance of 2-SGFM and SGD.

9.11 Additional Experimental Results on Cifra10

We evaluate the performance of our two-phase SGFM (cf. Algorithm 32) on the CIFAR10 dataset [Krizhevsky and Hinton, 2009] using convolutional neural networks (CNNs) with ReLU activations. We provide the details about the network architecture as follows,

```
class CNN_CIFAR(nn.Module):
    def __init__(self):
        super(CNN_CIFAR, self).__init__()
        self.conv1 = nn.Conv2d(3, 6, 5)
        self.conv2 = nn.Conv2d(6, 16, 5)
        self.fc1 = nn.Linear(16*5*5, 120)
        self.fc2 = nn.Linear(120, 84)
        self.fc3 = nn.Linear(84, 10)

    def forward(self, x):
        out = F.relu(self.conv1(x))
        out = F.max_pool2d(out, 2)
        out = F.relu(self.conv2(out))
        out = F.max_pool2d(out, 2)
        out = out.view(out.size(0), -1)
        out = F.relu(self.fc1(out))
        out = F.relu(self.fc2(out))
        out = self.fc3(out)
        out = F.log_softmax(out, dim=1)
        return out
```

Moreover, we summarize the experimental results in Figure 9.5. In the above two figures, we study the effect of batch size $B \geq 1$ in 2-SGFM on the CIFAR10 dataset. In the bottom two figures, we compare the performance of SGD and 2-SGFM. Overall, these results show promising performance of our proposed gradient-free method on solving real-world complex image classification problems.

Chapter 10

Adaptive and Doubly Optimal Learning in Games

Online gradient descent (OGD) is well known to be doubly optimal under strong convexity or monotonicity assumptions: (1) in the single-agent setting, it achieves an optimal regret of $\Theta(\log T)$ for strongly convex cost functions; and (2) in the multi-agent setting of strongly monotone games, with each agent employing OGD, we obtain last-iterate convergence of the joint action to a unique Nash equilibrium at an optimal rate of $\Theta(\frac{1}{T})$. While these finite-time guarantees highlight its merits, OGD has the drawback that it requires knowing the strong convexity/monotonicity parameters. In this paper, we design a fully adaptive OGD algorithm, AdaOGD, that does not require a priori knowledge of these parameters. In the single-agent setting, our algorithm achieves $O(\log^2(T))$ regret under strong convexity, which is optimal up to a log factor. Further, if each agent employs AdaOGD in strongly monotone games, the joint action converges in a last-iterate sense to a unique Nash equilibrium at a rate of $O(\frac{\log^3 T}{T})$, again optimal up to log factors. We illustrate our algorithms in a learning version of the classical newsvendor problem, where due to lost sales, only (noisy) gradient feedback can be observed. Our results immediately yield the first feasible and near-optimal algorithm for both the single-retailer and multi-retailer settings. We also extend our results to the more general setting of exp-concave cost functions and games, using the online Newton step (ONS) algorithm.

10.1 Introduction

The problem of online learning with gradient feedback [Blum, 1998, Shalev-Shwartz, 2012, Hazan, 2016] can be described in its essential form by the following adaptive decision-making process:

1. An agent interfaces with the environment by choosing an *action* $x^t \in \mathcal{X}$ at period t where $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex and compact set. For example, the action is a route in a traffic network or an output quantity in an oligopoly. The agent chooses x^t through

an online learning algorithm, which makes its choice adaptively based on observable historical information.

2. The environment then returns a cost function $f_t(\cdot)$ so that the agent incurs cost $f_t(x^t)$ and receives $\nabla f_t(x^t)$ as feedback. The process then moves to the next period $t+1$ and repeats.

One appealing feature of the online learning framework is that one need not impose any statistical regularity assumption: $f_1(\cdot), \ldots, f_T(\cdot)$ can be an arbitrary fixed sequence of cost functions, hence accommodating a non-stationary or even adversarial environment. Further, the cost function $f_t(\cdot)$ does need not to be known by the agent (and indeed in many applications it is not known); only the gradient feedback is needed. In this general framework, the standard metric for judging the performance of an online learning algorithm is *regret* [Blum and Mansour, 2007]—the difference between the total cost incurred by the algorithm up to T and the total cost incurred by the best fixed action in hindsight:

$$\text{Regret}(T) = \sum_{t=1}^{T} f_t(x^t) - \min_{x \in \mathcal{X}} \left\{ \sum_{t=1}^{T} f_t(x) \right\}.$$

If the average regret (obtained by dividing by T) goes to zero, then the algorithm is referred to as a "no-regret learning algorithm."

A canonical example of a no-regret learning algorithm is online gradient descent (OGD), where the agent takes a gradient step (given the current action x_t) and performs a projection step onto $\mathcal{X}$ to obtain the next action x_{t+1}. Analyzing OGD in the standard setting where the cost function f_t is convex, Zinkevich [2003] proved that the algorithm with the stepsize rule $\eta_t = \frac{1}{\sqrt{T}}$ achieves a regret bound of $\Theta(\sqrt{T})$, which is known to be minimax optimal [Hazan, 2016]. If f_t is further assumed to be μ-strongly convex, Hazan et al. [2007] proved that the algorithm with the stepsize rule $\eta_t = \frac{1}{\mu t}$ achieves a regret bound of $\Theta(\log T)$; again this rate is minimax optimal. In fact, the $\Theta(\log T)$ regret bound is achievable even for a class of cost functions that are more general than strongly convex cost functions: if f_t is exp-concave—a class of functions properly subsuming strongly convex functions that has found widespread applications—then the online Newton step (ONS) [Hazan et al., 2007] achieves the minimax optimal regret bound of $\Theta(d \log T)$. In summary, OGD and ONS provide two of the most well-known optimal no-regret learning algorithms in the online learning/online convex optimization literature, with their algorithmic simplicity and theoretical elegance being matched by their broad applicability in practice.

Given their appealing theoretical and practical properties, the aforementioned no-regret learning algorithms such as OGD and ONS have also served as natural candidates for game-theoretic learning. In this setting, each agent makes online decisions in an environment consisting of other agents, each of whom are making adaptive decisions. Even if the game is fixed, the fact that all agents are adjusting their strategies simultaneously means that the stream of costs seen by any single agent is non-stationary and complex. It might be hoped

that no-regret learning—given its robustness to assumptions—can cope with the complexity of the multi-agent setting. An extensive literature has shown that this hope is borne out—under no-regret dynamics, the time average of the joint actions converges to equilibria in various classes of games [Cesa-Bianchi and Lugosi, 2006, Shoham and Leyton-Brown, 2008, Viossat and Zapechelnyuk, 2013, Bloembergen et al., 2015, Monnot and Piliouras, 2017].

Further progress in the online learning and optimization literature has yielded more refined statements regarding convergence; in particular, last-iterate rate guarantees have been established for many no-regret learning algorithms. This leads to an analogous question for the game-theoretic setting: *If each agent employs a no-regret learning algorithm to minimize its own regret, can the joint action converge to a Nash equilibrium at an optimal last-iterate rate?*

An affirmative answer to this question[1] would establish a remarkable "double optimality" for an online algorithm: while the algorithm itself is only designed for maximizing the (transient) performance for a finite time horizon T, the resulting long-run performance would also be optimal for all agents, in the sense that any agent, by unilaterally deviating from the action suggested by the algorithm, could only incur higher cost (by the definition of a Nash equilibrium). Without this double optimality, an agent could incur "regret" in the long term, since it may do better by not following such an algorithm.

To address this question, it is necessary to study the last-iterate convergence of algorithms (i.e., the convergence of the *actual* joint action), a problem that has been recognized to be considerably more difficult than the characterization of convergence of time averages [Krichene et al., 2015a, Balandat et al., 2016, Zhou et al., 2017, 2018, Mertikopoulos et al., 2019, Mertikopoulos and Zhou, 2019]. For instance, as pointed out by Mertikopoulos et al. [2018], there are situations where the time average of the iterate converges to a Nash equilibrium, but the last iterate cycles around the equilibrium point. Progress has been made on this problem during the past five years, but much of it only provides qualitative or asymptotic results, with only a few quantitative (finite-time, last-iterate convergence guarantees) results obtained, for games having special structures or using metrics other than the distance to Nash equilibria. In particular, Zhou et al. [2021] established the last-iterate convergence of multi-agent OGD to the unique Nash equilibrium in strongly monotone games[2] at an optimal rate of $\Theta(\frac{1}{T})$. This optimal convergence rate continues to hold even when the gradient feedback is corrupted by certain forms of noise, in which case we have $E[\|x_t - x^\star\|_2^2] = \Theta(\frac{1}{T})$, where x_t is the (random) joint action of all agents and $x^\star$ is the unique Nash equilibrium.[3] Thus, OGD is doubly optimal when a strong convexity structure is available (i.e., the game is

[1] For instance, if each vehicle in a traffic network employs an optimal no-regret learning algorithm (such as OGD) to choose their route adaptively over a certain horizon, would the system converge to a stable traffic distribution or devolve to perpetual congestion as users ping-pong between different routes? If it does converge to a stable distribution, is it Nash? Because if not, each agent is being irrational—by *not* following the no-regret learning algorithm, agents can do individually do better.

[2] When the strongly monotone games have Lipschitz gradients—a condition that does not hold in many games of interest—a classic result from the variational inequality literature implies that multi-agent OGD converges to the unique Nash equilibrium at a geometric rate due to a contraction.

[3] This result is further generalized in Loizou et al. [2021], who obtain the same last-iterate convergence

strongly monotone or the cost functions are strongly convex from a single-agent perspective), giving a compelling argument for its adoption in single-agent and multi-agent settings.

However, this argument suffers from a key, subtle weakness: the theoretical guarantees for OGD require choices of step size, in both single-agent [Hazan et al., 2007] and multi-agent [Zhou et al., 2021] settings, and these choices require prior knowledge of problem parameters. In particular, it is generally assumed that the strong convexity parameter of the cost functions (single-agent) or the strong monotonicity parameter of the game (multi-agent) are known. Thus, OGD's appealing guarantees are not feasible in practice if the choice of step sizes cannot be made *fully adaptive* to problem parameters. Further, the feasibility issue is more acute in the multi-agent setting: in addition to requiring prior knowledge of problem parameters for step-size designs, recent work on adaptive OGD has assumed that each agent determines their step size using global information from all agents [Lin et al., 2020e, Antonakopoulos et al., 2021, Hsieh et al., 2021]. This is a practical and theoretical conundrum—if the agents can achieve this level of coordination, learning would be unnecessary in the first place. The same issue occurs for ONS in the single-agent setting, where the exp-concave parameter is needed as an input to the algorithm.[4] Consequently, the feasibility considerations lead us to consider the following question: *Can we design a feasible and doubly optimal variant of OGD under strong convexity and strong monotonicity? What about ONS?*

Our answer is a "yes" in a strong sense. We present a single feasible OGD algorithm—and hence a single parameter-adaptive scheme—that simultaneously (up to log factors) achieves optimal regret in the single-agent setting and optimal last-iterate convergence rate to the unique Nash equilibrium in the multi-agent setting. This analysis is different from and more challenging than that involved in the design of feasible OGD algorithms separately for single-agent and multi-agent settings. In particular, it could be that an effective adaptive scheme for the strong convexity parameter in single-agent setting is different from that for the strong monotonicity parameter in multi-agent setting, in which case one has *at best* either a feasible algorithm with optimal regret or a feasible algorithm that has an optimal convergence-to-Nash guarantee, but not both. Such results would still be of considerable value but our results in this paper show that such intermediate results can be bypassed; indeed, the best of both worlds can be achieved. We also develop a single feasible variant of ONS that (up to log factors) achieves optimal regret in the single-agent setting with exp-concave loss functions and optimal time-average convergence rate to the unique Nash equilibrium in the multi-agent setting. For the latter result we introduce and analyze a new class of exp-concave games.

Our results can also be cast in the framework of variational inequalities (VIs). Indeed, they can be viewed as contributing to the VI literature by presenting a decentralized, feasible optimization algorithm for finding a solution of a strongly monotone VI. We prefer to emphasize, however, the online learning perspective, and the design of no-regret algorithms, given

rate for strongly variationally stable games, under weaker noise assumptions.

[4]The multi-agent ONS has not yet been explored, and even the time-average convergence of ONS still remains to be established.

the direct connection of those algorithms to game-theoretic settings. In multi-agent games, it is natural to focus on decentralized algorithms and on algorithms that make minimal assumptions about their environment, allowing that environment to consist of other agents that may be responding in complex ways to an agent's actions. Our double optimality contributions are best understood as a further weakening of these assumptions, allowing interacting agents to choose actions effectively in an unknown, possibly adversarial, environment.

Related works. In both single-agent online learning and offline optimization, considerable attention has been paid to the development of adaptive gradient-based schemes. In particular, Duchi et al. [2011] presented an adaptive gradient algorithm (known as AdaGrad) for online learning with convex cost functions that updates the step sizes without needing to know the problem parameters. This algorithm is guaranteed to achieve a minimax-optimal regret of $O(\sqrt{T})$. Subsequently, Adam was proposed in the offline optimization setting to further exploit geometric aspects of iterate trajectories, exhibiting better empirical convergence performance [Kingma and Ba, 2015]. Theoretical guarantees have been obtained for Adam and other related adaptive algorithms in both offline optimization and online learning settings [Reddi et al., 2018b, Zou et al., 2019]. In parallel, the norm version of AdaGrad was developed and theoretical guarantees were established for related convex and/or nonconvex optimization problems [Levy, 2017, Levy et al., 2018, Ward et al., 2019, Li and Orabona, 2019].

This adaptive family of algorithms has also been studied in the online learning literature under an assumption of strong convexity [Mukkamala and Hein, 2017, Wang et al., 2020]. The algorithms are guaranteed to achieve a minimax-optimal regret of $O(\log(T))$. However, unlike in the convex setting, these adaptive algorithms are not feasible in practice since they often require knowledge of the strong convexity parameter. In the offline optimization setting, the gradient-based methods can be made adaptive to the strong convexity parameter by exploiting the Polyak stepsize [Polyak, 1987, Hazan and Kakade, 2019]. However, these algorithms do not extend readily to online learning since the sub-optimality gap is not well-defined. A recent line of research has shown that adaptive algorithms can be designed in the finite-sum setting, where they exhibit adaptivity to the strong convexity parameter [Roux et al., 2012, Defazio et al., 2014, Xu et al., 2017, Lei and Jordan, 2017, 2020, Vaswani et al., 2019, Nguyen et al., 2022]. In particular, Lei and Jordan [2017, 2020] showed that the use of random, geometrically-distributed epoch length yields full adaptivity in variance-controlled stochastic optimization under an assumption of strong convexity. However, these algorithm are not no-regret and thus their strategies do not extend readily to the online setting.

In terms of the last-iterate convergence to Nash equilibria, due to the difficulties mentioned earlier, much of the existing literature provides only qualitative convergence guarantees for non-adaptive (and hence infeasible) no-regret learning algorithms for various games [Krichene et al., 2015a, Balandat et al., 2016, Zhou et al., 2017, 2018, Mertikopoulos et al., 2019, Mertikopoulos and Zhou, 2019]. More recently, finite-time last-iterate convergence rates have been obtained for specially structured games, such as strongly mono-

tone games [Zhou et al., 2021, Loizou et al., 2021], unconstrained cocoercive games [Lin et al., 2020e], unconstrained smooth games [Golowich et al., 2020a] and constrained smooth games [Cai et al., 2022]. Except for a class of strongly monotone games, the last-iterate convergence rate is measured in metrics other than $\|x^t - x^\star\|_2^2$. Further, among these results, only Lin et al. [2020e] provides an adaptive online learning algorithm that does not require knowing the cocoercivity parameter. However, their algorithm falls short in two respects: (i) it may not be no-regret; (ii) each agent needs to know all other agents' gradients, thus again rendering it infeasible in practice. Recently, Antonakopoulos et al. [2021] has developed an adaptive extragradient algorithm for strictly monotone games that converges asymptotically in a last-iterate sense to the unique Nash equilibrium. However, their algorithm also requires each agent to know all others' gradients and the no-regret property cannot be guaranteed; indeed, the original extragradient algorithm was shown to not be no-regret [Golowich et al., 2020a]. Hsieh et al. [2021] has proposed a no-regret adaptive online learning algorithm based on optimistic mirror descent and established a regret of $O(\sqrt{T})$ for convex cost functions. They also proved asymptotic last-iterate convergence to the unique Nash equilibrium for strictly variationally stable games (a superset of strictly monotone games).

Another line of relevant literature focuses on stochastic approximation methods for solving strongly monotone VIs. An early proposal using such an approach was presented by Jiang and Xu [2008], who proposed a stochastic projection method for solving strongly monotone VIs with an almost-sure convergence guarantee. Koshal et al. [2012] and Yousefian et al. [2013] proposed various regularized iterative stochastic approximation methods for solving monotone VIs and also established almost-sure convergence. A survey of these methods, as well as applications and the theory behind stochastic VI, can be found in Shanbhag [2013]. Juditsky et al. [2011] was among the first to establish an iteration complexity bound for stochastic VI methods by extending the mirror-prox method [Nemirovski, 2004] to stochastic setting. Yousefian et al. [2014] further extended the stochastic mirror-prox method with a more general step size choice and proved the same iteration complexity. They also proved an improved complexity bound for the stochastic extragradient method for solving strongly monotone VIs. Chen et al. [2017] studied a specific class of VIs and proposed a method that combines the stochastic mirror-prox method with Nesterov's acceleration [Nesterov, 2018], resulting in an optimal iteration complexity for such problem class. Kannan and Shanbhag [2019] analyzed a general variant of an extragradient method (which uses general distance-generating functions) and proved an optimal iteration bound under a slightly weaker assumption than strong monotonicity. Several other stochastic methods have also been shown to yield an optimal iteration bound for solving strongly monotone VIs [Kotsalis et al., 2022, Huang and Zhang, 2022b]. In recent years, there have been developments in variance-reduction-based methods [Balamurugan and Bach, 2016, Iusem et al., 2017, 2019, Jalilzadeh and Shanbhag, 2019, Yu et al., 2022, Alacaoglu and Malitsky, 2022, Jin et al., 2022, Huang et al., 2022a]. In the line of research aiming to model multistage stochastic VI (as compared to the single-stage VI considered in the aforementioned literature), the dynamics between the actions and the arrival of future information plays a central role. For further details regarding multistage stochastic VI, we refer to Rockafellar and Wets [2017]

and Rockafellar and Sun [2019].

We also note that there is a line of work focusing on parameter-free online learning [Foster et al., 2015, 2017, Orabona and Pál, 2016, Cutkosky and Orabona, 2018, Jun and Orabona, 2019, Cutkosky, 2020a,b]. This work considers an alternative form of regret: $\text{Regret}_T(x) = \sum_{t=1}^T f_t(x^t) - \sum_{t=1}^T f_t(x)$, where $x \in \mathcal{X}$ is an unknown *competitor*. The goal is to achieve expected regret bounds that have optimal dependency not only on T but also $\|x\|$. Notably, their framework provides a way to design algorithms that achieve minimax-optimal regret with respect to any competitor, without imposing a bounded set for the competitor nor any parameter to tune in online convex optimization. However, the strong convexity parameter and the exp-concavity parameter both characterize *the lower bound for curvature information*. Estimating these quantities will require new techniques. In summary, the possibility of designing an online algorithm that is both *doubly optimal* and *feasible* under strong convexity still remains open.

Contributions. We present a feasible variant of OGD that we refer to as AdaOGD that does not require knowing any problem parameter. It is guaranteed to achieve a minimax optimal (up to a log factor) regret bound of $O(\log^2(T))$ in the single-agent setting with strongly convex cost functions. Further, in a strongly monotone game, if each agent employs AdaOGD, we show that the joint action converges to the unique Nash equilibrium in a *last-iterate sense* at a rate of $O(\frac{\log^3(T)}{T})$, again optimal up to log factors. In comparison, the existing single-agent OGD [Hazan, 2016] and multi-agent OGD [Zhou et al., 2021] methods require prior knowledge of the strong convexity/monotonicity parameter (respectively) to perform the step-size design with theoretical guarantees. It is worth noting that if the game has only a single agent, the strong monotonicity parameter degenerates to the strong convexity parameter. However, when there are multiple agents, the strong monotonicity parameter depends on all agents' cost functions. As such, one would naturally think that these two settings would require two different adaptive schemes. Surprisingly, our AdaOGD algorithm, which is based on a single adaptive principle, works in both settings, achieving optimal regret in the single-agent setting and optimal last-iterate convergence in the multi-agent setting (up to log factors). A particularly important application of our results is the problem of learning to order in the classical newsvendor problem, where due to lost sales, only (noisy) gradient feedback can be observed. Our results immediately yield the first feasible near-optimal algorithm—both in the single-retailer setting [Huh and Rusmevichientong, 2009] and in the multi-retailer setting [Netessine and Rudi, 2003]. This is in contrast to previous work that requires problem parameters to be known. Indeed, the direct application of our results to the VI setting yields a decentralized, feasible optimization algorithm for finding a solution of a strongly monotone VI.

Additionally, we provide a feasible variant of ONS (that we refer to as AdaONS) that again does not require prior knowledge of any problem parameter. It is also guaranteed to achieve a minimax optimal regret bound of $O(d \log^2(T))$ (up to a log factor) in the single-agent setting with exp-concave cost functions. Further, we propose a new class of

exp-concave (EC) games and show that if each agents employs AdaONS, an optimal time-average convergence rate of $O(\frac{d\log^2(T)}{T})$ is obtained. Much like strongly monotone games that provide a multi-agent generalization of strongly convex cost functions, the EC games that we introduce are a natural generalization of exp-concave cost functions from the single-agent to a multi-agent setting. Again, in this case, a single adaptive scheme works for both of these two settings. One thing to note here is that we first establish an $O(\frac{d\log(T)}{T})$ time-average convergence rate for the multi-agent version of classical ONS that is non-adaptive (and hence not feasible). To the best of our knowledge, results of this kind have not appeared in the game-theoretic literature and they yield a decentralized, feasible optimization algorithm for finding a solution of a new class of VIs in that line of literature.

Perhaps the most surprising takeaway from our work is that both AdaOGD and AdaONS are based on a simple and unifying randomized strategy that selects the step size based on a set of independent and identically distributed geometric random variables.

10.2 Feasible Single-Agent Online Learning under Strongly Convex Costs

We present adaptive OGD (AdaOGD), a feasible single-agent online learning algorithm, and prove that AdaOGD achieves a near-optimal regret of $O(\log^2(T))$ for a class of strongly convex cost functions. We also show that our algorithm can be used to solve the problem of adaptive ordering in newsvendor problems with lost sales. To our knowledge, this is the first feasible no-regret learning algorithm for the newsvendor-with-lost-sales problem with strong regret guarantees.

Algorithmic scheme. We continue with the setup in the introduction, focusing on strongly convex cost functions f_t:

Definition 10.2.1 *A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is β-strongly convex if $f(\cdot) - 0.5\beta\|\cdot\|^2$ is convex.*

We work with a general (and relaxed) model of gradient feedback [Flaxman et al., 2005]:

1. At each round t, an unbiased and bounded gradient is observed. That is, the observed noisy gradient ξ^t satisfies $\mathbb{E}[\xi^t \mid x^t] = \nabla f_t(x^t)$ and $\mathbb{E}[\|\xi^t\|^2 \mid x^t] \leq G^2$ for all $t \geq 1$.
2. The action set $\mathcal{X}$ is bounded by a diameter $D > 0$, i.e., $\|x - x'\| \leq D$ for all $x, x' \in \mathcal{X}$.

A lower bound of $\Omega(\log(T))$ has been established in Hazan and Kale [2014, Theorem 18] under an assumption of perfect gradient feedback. However, even with noisy gradient feedback, OGD with a particular step size can achieve the minimax-optimal regret bound of $\Theta(\log(T))$ [Hazan et al., 2007]. In particular, we write OGD as $x^{t+1} \leftarrow \mathcal{P}_{\mathcal{X}}(x^t - \frac{1}{\beta(t+1)}\nabla f_t(x^t))$, which is equivalent to

$$\eta^{t+1} \leftarrow \beta(t+1), \qquad x^{t+1} \leftarrow \operatorname*{argmin}_{x\in\mathcal{X}}\{(x - x^t)^\top \nabla f_t(x^t) + \tfrac{\eta^{t+1}}{2}\|x - x^t\|^2\}.$$

Algorithm 33 AdaOGD(x^1, T)

Input: initial point $x^1 \in \mathcal{X}$ and the total number of rounds T.
Initialization: $p_0 = \frac{1}{\log(T+10)}$.
for $t = 1, 2, \ldots, T$ **do**
  sample $M^t \sim \mathsf{Geometric}(p_0)$.
  set $\eta^{t+1} \leftarrow \frac{t+1}{\sqrt{1+\max\{M^1,\ldots,M^t\}}}$.
  update $x^{t+1} \leftarrow \operatorname{argmin}_{x\in\mathcal{X}}\{(x - x^t)^\top \xi^t + \frac{\eta^{t+1}}{2}\|x - x^t\|^2\}$.

The value of $\beta(t+1)$ comes from the key inequality for β-strongly convex functions:

$$f(x') \geq f(x) + (x' - x)^\top \nabla f(x) + \tfrac{\beta}{2}\|x' - x\|^2. \tag{10.1}$$

Despite the elegance of OGD (and its optimal regret guarantee), however, it is inadequate since it requires knowledge of the problem parameter β. We can address this issue by a simple randomization strategy based on independent, identically distributed geometric random variables, $M^t \sim \mathsf{Geometric}(p_0)$, for $p_0 = \frac{1}{\log(T+10)}$; i.e., $\mathbb{P}(M^t = k) = (1 - p_0)^{k-1}p_0$ for $k \in \{1, 2, \ldots\}$. See Algorithm 33.

Remark 10.2.2 (Compared with doubling trick) *In the context of online learning, the doubling trick [Shalev-Shwartz, 2012] is commonly used to make OGD adaptive to* ***specific*** *unknown parameters under* ***convex*** *costs. In particular, for any algorithm that enjoys a regret bound of $O(\sqrt{T})$ but requires the knowledge of T (such as OGD under convex costs), the doubling trick converts such an algorithm into an algorithm that does not require the knowledge of T. The idea is to divide the time into periods of increasing size and run the original algorithm on each period: for $m = 0, 1, 2, \ldots$, we run the original algorithm on the 2^m rounds $t = 2^m, \ldots, 2^{m+1} - 1$. The resulting algorithm enjoys a regret bound of $O(\sqrt{T})$.*

However, it is nontrivial to apply the doubling trick to make OGD adaptive to the strongly convex parameter β under ***strongly convex*** *costs. In particular, a natural adaptation of the doubling trick under strongly convex costs is as follows: for $m = 0, 1, 2, \ldots$, we run OGD with $\eta^t = \frac{t}{2^m}$ on the rounds $t = 2^m, \ldots, 2^{m+1} - 1$. The analysis contains two parts: (i) for $0 \leq m \leq \lfloor \log_2(1/\beta) \rfloor$, the regret for each round is $O(2^m)$. This leads to a total constant regret of $O(1/\beta)$; (ii) for $\lceil \log_2(1/\beta) \rceil \leq m \leq \lceil \log_2(T) \rceil$, the regret for each round is $O(2^m)$. This unfortunately leads to a linear regret of $O(T)$. This argument of course does not eliminate the possibility that some variant of doubling could lead to doubly optimal learning algorithms for strongly monotone games. Our results do suggest that a fruitful way to search for such a procedure would be via some form of randomization.*

Remark 10.2.3 (Comparison with geometrization) *In offline finite-sum optimization, the geometrization trick [Lei and Jordan, 2017, 2020]—which sets the length of each epoch as a geometric random variable—has been used to make the stochastic variance-reduced gradient*

(SVRG) algorithm of Johnson and Zhang [2013] adaptive to both strong convexity parameter and target accuracy.

While similar in spirit, our approach is not a straightforward application of the geometrization trick. The key difference between OGD and SVRG is their dependence on the strongly convexity parameter. The former needs that parameter to set the stepsize while the latter algorithm needs it to set the length of each epoch. The intuition behind the geometrization trick is to randomly set the length of each epoch using geometric random variables, allowing terms to telescope across the outer and inner loops; such telescoping does not happen in SVRG, a fact which leads to the loss of adaptivity for SVRG (see Lei and Jordan [2020, Section 3.1] for more details). In contrast, our technique is designed to randomly set the stepsize using geometric random variables, implementing a trade-off for bounding the regret and last-iterate convergence rate. The telescoping from [Lei and Jordan, 2020] occurs due to the specific nature of SVRG (or more broadly, the setting of offline finite-sum optimization), and does not appear in our analysis for AdaOGD and its multi-agent generalization.

Regret guarantees. We present our result on the regret minimization property in the following theorem.

Theorem 10.2.4 *For an arbitrary fixed sequence of β-strongly convex functions $f_1, \ldots, f_T$, where each f_t satisfies $\mathbb{E}[\xi^t \mid x^t] = \nabla f_t(x^t)$ and $\mathbb{E}[\|\xi^t\|^2 \mid x^t] \leq G^2$ for all $t \geq 1$, and $\|x - x'\| \leq D$ for all $x, x' \in \mathcal{X}$. If the agent employs Algorithm 33, we have*

$$\mathbb{E}[\mathit{Regret}(T)] \leq \frac{D^2}{2}(1 + e^{\frac{1}{\beta^2 \log(T+10)}}) + \frac{G^2 \log(T+1)}{2}\sqrt{1 + \log(T+10) + \log(T)\log(T+10)}.$$

As a consequence, we have $\mathbb{E}[\mathit{Regret}(T)] = O(\log^2(T))$.

Remark 10.2.5 *Theorem 10.2.4 demonstrates that Algorithm 33 achieves a near-optimal regret since the upper bound matches the lower bound up to a log factor; indeed, Hazan and Kale [2014] proved the lower bound of $\Omega(\log(T))$ for this setting. Furthermore, Algorithm 33 dynamically adjusts η^{t+1} without any prior knowledge of problem parameters, only utilizing the noisy feedback $\{\xi^t\}_{t \geq 1}$.*

We provide a simple result on the maximum of independent identically distributed (i.i.d.) geometric random variables.

Proposition 10.2.6 *Let $X_1, \ldots, X_n$ be n i.i.d. geometric random variables: $X_i \sim \mathsf{Geometric}(p_0)$ for all $i = 1, 2, \ldots, n$ and with $p_0 \in (0, 1)$. Defining $\bar{X}_n = \max_{1 \leq i \leq n} X_i$, we have*

$$\sum_{n=1}^{+\infty} \mathbb{P}(\bar{X}_n \leq x) \leq e^{xp_0},$$

and

$$1 \leq \mathbb{E}[\bar{X}_n] \leq \frac{1 + \log(n)}{p_0}.$$

To prove Theorem 10.2.4, we also require a descent inequality for the iterates generated by Algorithm 33.

Lemma 10.2.7 *For an arbitrary fixed sequence of β-strongly convex functions $f_1, \ldots, f_T$, where each f_t satisfies $\mathbb{E}[\xi^t \mid x^t] = \nabla f_t(x^t)$ and $\mathbb{E}[\|\xi^t\|^2 \mid x^t] \le G^2$ for all $t \ge 1$, and $\|x - x'\| \le D$ for all $x, x' \in \mathcal{X}$. Letting the iterates $\{x^t\}_{t\ge 1}$ be generated by Algorithm 33, we have*

$$\sum_{t=1}^{T} \mathbb{E}[f_t(x^t) - f_t(x)] \le \tfrac{\eta^1}{2}\|x^1 - x\|^2 + \sum_{t=1}^{T} \mathbb{E}\left[\left(\tfrac{\eta^{t+1}-\eta^t}{2} - \tfrac{\beta}{2}\right)\|x^t - x\|^2\right] + \tfrac{G^2}{2}\left(\sum_{t=1}^{T} \mathbb{E}\left[\tfrac{1}{\eta^{t+1}}\right]\right), \text{ for all } x \in \mathcal{X}.$$

Proof of Theorem 10.2.4. Recall that $\mathcal{X}$ is convex and bounded with a diameter $D > 0$ and we have $\eta^{t+1} = \frac{t+1}{\sqrt{1+\max\{M^1,\ldots,M^t\}}}$ in Algorithm 33, we have

$$\tfrac{\eta^1}{2}\|x^1 - x\|^2 \le \tfrac{D^2}{2}, \qquad \eta^{t+1} - \eta^t \le \tfrac{1}{\sqrt{1+\max\{M^1,\ldots,M^t\}}}.$$

By Lemma 10.2.7, we have

$$\sum_{t=1}^{T} \mathbb{E}[f_t(x^t) - f_t(x)] \le \tfrac{D^2}{2} + \sum_{t=1}^{T} \mathbb{E}\left[\left(\tfrac{1}{2\sqrt{1+\max\{M^1,\ldots,M^t\}}} - \tfrac{\beta}{2}\right)\|x^t - x\|^2\right] + \tfrac{G^2}{2}\left(\sum_{t=1}^{T} \mathbb{E}\left[\tfrac{1}{\eta^{t+1}}\right]\right). \tag{10.2}$$

Further, we have

$$\sum_{t=1}^{T} \tfrac{1}{\eta^{t+1}} \le \sqrt{1+\max\{M^1,\ldots,M^T\}}\left(\sum_{t=1}^{T} \tfrac{1}{t+1}\right) \le \sqrt{1+\max\{M^1,\ldots,M^T\}}\log(T+1). \tag{10.3}$$

Plugging Eq. (10.3) into Eq. (10.2) yields that

$$\sum_{t=1}^{T} \mathbb{E}[f_t(x^t) - f_t(x)] \le \tfrac{D^2}{2} + \sum_{t=1}^{T} \mathbb{E}\left[\left(\tfrac{1}{2\sqrt{1+\max\{M^1,\ldots,M^t\}}} - \tfrac{\beta}{2}\right)\|x^t - x\|^2\right] + \tfrac{G^2\log(T+1)}{2}\mathbb{E}\left[\sqrt{1+\max\{M^1,\ldots,M^T\}}\right].$$

Since $\mathcal{X}$ is convex and bounded with a diameter $D > 0$, we have

$$\sum_{t=1}^{T}\left(\tfrac{1}{2\sqrt{1+\max\{M^1,\ldots,M^t\}}} - \tfrac{\beta}{2}\right)\|x^t - x\|^2 \le \tfrac{D^2}{2}\left(\sum_{t=1}^{T}\max\left\{0, \tfrac{1}{\sqrt{1+\max\{M^1,\ldots,M^t\}}} - \beta\right\}\right).$$

This implies that

$$\mathbb{E}[\text{Regret}(T)] \le \tfrac{D^2}{2} + \underbrace{\tfrac{D^2}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\max\left\{0, \tfrac{1}{\sqrt{1+\max\{M^1,\ldots,M^t\}}} - \beta\right\}\right]}_{\mathbf{I}} + \tfrac{G^2\log(T+1)}{2}\underbrace{\mathbb{E}\left[\sqrt{1+\max\{M^1,\ldots,M^T\}}\right]}_{\mathbf{II}}.$$

It remains to bound the terms **I** and **II** using Proposition 10.2.6. Indeed, we have

$$\begin{aligned}\mathbf{I} &= \sum_{t=1}^{T}\mathbb{E}\left[\max\left\{0, \frac{1}{\sqrt{1+\max\{M^1,\dots,M^t\}}}-\beta\right\}\right] \le \sum_{t=1}^{T}\mathbb{P}\left(\frac{1}{\sqrt{1+\max\{M^1,\dots,M^t\}}}-\beta\ge 0\right)\\ &= \sum_{t=1}^{T}\mathbb{P}\left(\sqrt{1+\max\{M^1,\dots,M^t\}}\le \tfrac{1}{\beta}\right) \le \sum_{t=1}^{T}\mathbb{P}\left(\max\{M^1,\dots,M^t\}\le \tfrac{1}{\beta^2}\right).\end{aligned}$$

Since $M^1,\dots,M^t$ are i.i.d. geometric random variables with $p_0=\frac{1}{\log(T+10)}$, Proposition 10.2.6 implies that

$$\sum_{t=1}^{T}\mathbb{P}\left(\max\{M^1,\dots,M^t\}\le \tfrac{1}{\beta^2}\right)\le e^{\frac{p_0}{\beta^2}} = e^{\frac{1}{\beta^2\log(T+10)}}.$$

Putting these pieces together yields that $\mathbf{I}\le e^{\frac{1}{\beta^2\log(T+10)}}$.

By using Jensen's inequality and the concavity of $g(x)=\sqrt{x}$, we have

$$\mathbf{II}\le\sqrt{1+\mathbb{E}\left[\max\{M^1,\dots,M^T\}\right]}.$$

Using Proposition 10.2.6 and $p_0=\frac{1}{\log(T+10)}$, we have

$$\mathbb{E}\left[\max\{M^1,\dots,M^T\}\right]\le\frac{1+\log(T)}{p_0}=\log(T+10)+\log(T)\log(T+10).$$

Putting these pieces together yields that $\mathbf{II}\le\sqrt{1+\log(T+10)+\log(T)\log(T+10)}$. Therefore, we conclude that

$$\begin{aligned}\mathbb{E}[\text{Regret}(T)] &\le \tfrac{D^2}{2}(1+e^{\frac{1}{\beta^2\log(T+10)}})+\tfrac{G^2\log(T+1)}{2}\sqrt{1+\log(T+10)+\log(T)\log(T+10)}\\ &= \tfrac{D^2}{2}(1+e^{\frac{1}{\beta^2\log(T+10)}})+\tfrac{G^2\log(T+1)}{2}\sqrt{1+\log(T+10)+\log(T)\log(T+10)}.\end{aligned}$$

This completes the proof.

Application: Feasible learning for Newsvendors with lost sales. The single-retailer version of the newsvendor problem is a famous model for perishable inventory control [Huh and Rusmevichientong, 2009]. The assumption in this model is that unsold inventory perishes at the end of each period. A retailer sells a product over a time horizon T and then makes inventory-ordering decisions $x^t\in[0,\bar{x}]$ at the beginning of each period t to maximize the profit. The unknown demand D^t is random and only realized with a value d^t after the retailer makes her decision. It is often assumed in the inventory control literature that the D^t are independent, corresponding to a stationary environment. Here, we do not need to make this assumption and we allow D^t to be arbitrary. Further, in the lost-sales setting, any unmet demand is lost and hence the retailer does *not* observe d^t; instead, she only observes the sales quantity $\min\{x^t,d^t\}$. The retailer's cost functions are defined by

$$f_t(x^t)=(p-c)\cdot\mathbb{E}[\max\{0,D^t-x^t\}]+c\cdot\mathbb{E}[\max\{0,x^t-D^t\}], \tag{10.4}$$

Algorithm 34 Newsvendor-AdaOGD(x^1, T)

Input: initial point $x^1 \in \mathcal{X}$ and the total number of rounds T.
Initialization: $p_0 = \frac{1}{\log(T+10)}$.
for $t = 1, 2, \ldots, T$ **do**
  sample $M^t \sim \mathsf{Geometric}(p_0)$.
  set $\eta^{t+1} \leftarrow \frac{t+1}{\sqrt{1+\max\{M^1,\ldots,M^t\}}}$.
  observe the sales quantity of $\min\{x^t, d^t\}$.
  update $x^{t+1} \leftarrow \begin{cases} \operatorname{argmin}_{x\in[0,\bar{x}]}\{(x-x^t)c + \frac{\eta^{t+1}}{2}(x-x^t)^2\}, & \text{if } x^t \geq \min\{x^t, d^t\}, \\ \operatorname{argmin}_{x\in[0,\bar{x}]}\{(x-x^t)(c-p) + \frac{\eta^{t+1}}{2}(x-x^t)^2\}, & \text{otherwise.} \end{cases}$

where the unit purchase cost is $c > 0$ and the unit selling price is $p \geq c$. It is known that minimizing this cost is equivalent to maximizing the profit, $\mathbb{E}[p \cdot \min\{x^t, D^t\} - c \cdot x^t]$, where:

$$\begin{aligned} &\mathbb{E}[p \cdot \min\{x^t, D^t\} - c \cdot x^t] \\ &= \mathbb{E}\left[p \cdot \left(D^t - \max\{0, D^t - x^t\}\right) - c \cdot \left(D^t - \max\{0, D^t - x^t\} + \max\{0, x^t - D^t\}\right)\right] \\ &= (p-c) \cdot \mathbb{E}[D^t] - (p-c) \cdot \mathbb{E}[\max\{0, D^t - x^t\}] - c \cdot \mathbb{E}[\max\{0, x^t - D^t\}] \\ &= (p-c) \cdot \mathbb{E}[D^t] - f_t(x^t). \end{aligned}$$

Note that the first term $(p-c) \cdot \mathbb{E}[D^t]$ is independent of x^t. Thus, the maximization of the profit $\mathbb{E}[p \cdot \min\{x^t, D^t\} - c \cdot x^t]$ is equivalent to minimizing the cost $f_t(\cdot)$ in Eq. (10.4).

In this context, Huh and Rusmevichientong [2009] have shown that the cost function $f_t(\cdot)$ is convex in general and αp-strongly convex if the demand is a random variable with a continuous density function q such that $\inf_{d\in[0,\bar{x}]} q(d) \geq \alpha > 0$. Moreover, the retailer only observes the sales quantity $\min\{x^t, d^t\}$ where d^t is a realization of D^t. Thus, the (noisy) bandit feedback is not observable. However, a noisy gradient feedback signal can be obtained:

$$\xi^t = \begin{cases} c, & \text{if } x^t \geq \min\{x^t, d^t\}, \\ c - p, & \text{otherwise,} \end{cases}$$

which is an unbiased and bounded gradient estimator. In this setting, the parameter $\alpha > 0$ is unavailable since the distribution of the demand is unknown. However, the retailer can apply our AdaOGD algorithm (cf. Algorithm 33) and obtain a near-optimal regret of $O(\log^2(T))$.

We specialize Algorithm 33 to the newsvendor problem in Algorithm 34 and we present the corresponding result on the regret minimization property in the following corollary.

Corollary 10.2.8 *In the single-retailer newsvendor problem, the retailer's cost functions are defined by Eq. (10.4) where the unit purchase cost is $c > 0$ and the unit selling price is $p \geq c$. Also, the demand is a random variable with a continuous density function q such that $\inf_{d\in[0,\bar{x}]} q(d) \geq \alpha > 0$. If the agent employs Algorithm 34, we have*

$$\mathbb{E}[\mathit{Regret}(T)] \leq \frac{\bar{x}^2}{2}\left(1 + e^{\frac{1}{(\alpha p)^2 \log(T+10)}}\right) + \frac{p^2 \log(T+1)}{2}\sqrt{1 + \log(T+10) + \log(T)\log(T+10)}.$$

As a consequence, we have $\mathbb{E}[\textit{Regret}(T)] = O(\log^2(T))$.

Proof. Recall that $f_t(x^t) = (p-c)\cdot\mathbb{E}[\max\{0, D^t - x^t\}] + c\cdot\mathbb{E}[\max\{0, x^t - D^t\}]$ and the noisy gradient feedback is given by

$$\xi^t = \begin{cases} c, & \text{if } x^t \geq \min\{x^t, d^t\}, \\ c-p, & \text{otherwise.} \end{cases}$$

So we have $\mathbb{E}[\xi^t \mid x^t] = \nabla f_t(x^t)$ and $\mathbb{E}[\|\xi^t\|^2 \mid x^t] \leq p^2$ for all $t \geq 1$. Since the demand has a continuous density function q such that $\inf_{d\in[0,\bar{x}]} q(d) \geq \alpha > 0$, we have $f_t(\cdot)$ is αp-strongly convex. In addition, $\mathcal{X} = [0, \bar{x}]$ implies that $D = \bar{x}$. Thus, Theorem 10.2.4 can be applied and implies the desired result. □

10.3 Feasible Multi-Agent Online Learning in Strongly Monotone Games

We consider feasible multi-agent learning in monotone games. Our main result is that if each agent applies AdaOGD in a strongly monotone game (the multi-agent generalization of strongly convex costs), the joint action of all agents converges in a last-iterate sense to the unique Nash equilibrium at a near-optimal rate. In contrast to previous work, our results provide the first feasible no-regret learning algorithm that is doubly optimal; in particular, in addition to not requiring any prior knowledge of the problem parameters, one does not need to adjust the step-size schedule based on whether an agent is in the single-agent setting or the multi-agent setting. It is important to note that these are two different merits, and our algorithm enjoys both of them.

Basic definitions and notations. We first review the definition of continuous games and consider a class of monotone games. In particular, we focus on continuous games played by a set of agents, $\mathcal{N} = \{1, 2, \ldots, N\}$. Each agent selects an *action* x_i from a convex and bounded $\mathcal{X}_i \subseteq \mathbb{R}^{d_i}$. The incurred cost for each agent is determined by the joint action $x = (x_i; x_{-i}) = (x_1, x_2, \ldots, x_N)$. We let $\|\cdot\|$ denote the Euclidean norm (Other norms can also be accommodated here and different $\mathcal{X}_i$'s can have different norms).

Definition 10.3.1 *A continuous game is a tuple* $\mathcal{G} = \{\mathcal{N}, \mathcal{X} = \Pi_{i=1}^N \mathcal{X}_i, \{u_i\}_{i=1}^N\}$, *where* $\mathcal{N}$ *is a set of* N *agents,* $\mathcal{X}_i \subseteq \mathbb{R}^{d_i}$ *is the* i^{th} *agent's action set that is both convex and bounded, and* $u_i : \mathcal{X} \to \mathbb{R}$ *is the* i^{th} *agent's cost function satisfying: (i)* $u_i(x_i; x_{-i})$ *is continuous in* x *and continuously differentiable in* x_i*; (ii)* $v_i(x) = \nabla_{x_i} u_i(x_i; x_{-i})$ *is continuous in* x. *For simplicity, we denote* $v(\cdot) = (v_1(\cdot), v_2(\cdot), \ldots, v_N(\cdot))$ *as the joint profile of all agents' individual gradients.*

We work with an analogous model of gradient feedback:

1. At each round t, an unbiased and bounded gradient is observed as a feedback signal. In particular, the observed noisy gradient ξ^t satisfies $\mathbb{E}[\xi^t \mid x^t] = v(x^t)$ and $\mathbb{E}[\|\xi^t\|^2 \mid x^t] \leq G^2$ for all $t \geq 1$.

2. The action set $\mathcal{X}$ is bounded by a diameter $D > 0$, i.e., $\|x - x'\| \leq D$ for all $x, x' \in \mathcal{X}$.

The study of monotone games dates to Rosen [1965] who considered a class of games that satisfy the *diagonal strict concavity* (DSC) condition.[5] Further contributions appeared in Sandholm [2015] and Sorin and Wan [2016], where games that satisfy DSC are called "contractive" and "dissipative." In this context, a game $\mathcal{G}$ is *monotone* if $(x' - x)^\top (v(x') - v(x)) \geq 0$ for any $x, x' \in \mathcal{X}$. Intuitively, the notion of monotonicity generalize the notion of convexity; indeed, the gradient operator of a convex function f satisfies that $(x' - x)^\top (\nabla f(x') - \nabla f(x)) \geq 0$ for any $x, x' \in \mathcal{X}$.

We now define the class of strongly monotone games:

Definition 10.3.2 *A continuous game $\mathcal{G}$ is said to be* β-strongly monotone *if we have $\langle x' - x, v(x') - v(x)\rangle \geq \beta\|x' - x\|^2$ for all $x, x' \in \mathcal{X}$.*

A standard solution concept for non-cooperative games is the *Nash equilibrium* (NE), where no agent has an incentive to deviate from her strategy [Osborne and Rubinstein, 1994]. For the continuous games considered in this paper, we are interested in pure-strategy Nash equilibria since the randomness introduced by mixed strategies is unnecessary when each action lives in a continuum.

Definition 10.3.3 *An action profile $x^\star \in \mathcal{X}$ is called a* Nash equilibrium *of $\mathcal{G}$ if it is resilient to unilateral deviations; that is, $u_i(x_i^\star; x_{-i}^\star) \leq u_i(x_i; x_{-i}^\star)$ for all $x_i \in \mathcal{X}_i$ and $i \in \mathcal{N}$.*

Debreu [1952] proved that any continuous game admits at least one Nash equilibrium if all action sets are convex and bounded, and all cost functions are individually convex (i.e., $u_i(x_i; x_{-i})$ is convex in x_i for a fixed x_{-i}). Moreover, there is a variational characterization that forms the basis of equilibrium computation under an individual convexity condition [Facchinei and Pang, 2007]. We summarize this characterization in the following proposition.

Proposition 10.3.4 *If all cost functions are in a continuous game $\mathcal{G}$ are individually convex, the joint action $x^\star \in \mathcal{X}$ is a Nash equilibrium if and only if $(x - x^\star)^\top v(x^\star) \geq 0$ for all $x \in \mathcal{X}$.*

The notion of strong monotonicity arises in various application domains. Examples include strongly-convex-strongly-concave zero-sum games, atomic splittable congestion games in networks with parallel links [Orda et al., 1993, Sorin and Wan, 2016, Mertikopoulos and Zhou,

[5] This condition is equivalent to the notion of strict monotonicity in convex analysis [Bauschke and Combettes, 2011]; see Facchinei and Pang [2007] for further discussion.

2019], wireless network optimization [Weeraddana et al., 2012, Tan, 2014, Zhou et al., 2021] and classical online decision-making problems [Cesa-Bianchi and Lugosi, 2006].

Strongly monotone games satisfy the individual convexity condition and hence the existence of at least one Nash equilibrium is ensured. Moreover, every strongly monotone game admits a unique Nash equilibrium [Zhou et al., 2021]. Thus, one appealing feature of strongly monotone games is that finite-time convergence can be derived in terms of $\|\hat{x}-x^\star\|^2$ where $x^\star \in \mathcal{X}$ is a unique Nash equilibrium. Despite some recent progress on last-iterate convergence rates for non-strongly monotone games [Lin et al., 2020e, Golowich et al., 2020a, Cai et al., 2022], last-iterate convergence rates in terms of $\|\hat{x} - x^\star\|^2$ are only available for strongly monotone games [Bravo et al., 2018, Zhou et al., 2021, Lin et al., 2021b]. An important gap in the literature is that there currently do not exist *doubly optimal* and *feasible* learning algorithms for strongly monotone games.

Algorithmic scheme. We review the multi-agent OGD method that is a generalization of single-agent OGD. Letting $x_i^1 \in \mathcal{X}_i$ for all $i \in \mathcal{N}$, the multi-agent version of OGD (cf. Definition 10.3.2) performs the following step at each round:

$$\eta_i^{t+1} \leftarrow \beta(t+1), \qquad x_i^{t+1} \leftarrow \operatorname*{argmin}_{x_i \in \mathcal{X}_i}\{(x_i - x_i^t)^\top \xi_i^t + \tfrac{\eta_i^{t+1}}{2}\|x_i - x_i^t\|^2\}. \tag{10.5}$$

In the following theorem, we summarize the results from Zhou et al. [2021] on the optimal last-iterate convergence rate of multi-agent OGD in Eq. (10.5) using a squared Euclidean distance function.

Theorem 10.3.5 *Suppose that a continuous game $\mathcal{G}$ is β-strongly monotone and let $G, D > 0$ be problem parameters satisfying $\mathbb{E}[\xi^t \mid x^t] = v(x^t)$ and let $\mathbb{E}[\|\xi^t\|^2 \mid x^t] \le G^2$ for all $t \ge 1$, and $\|x - x'\| \le D$ for all $x, x' \in \mathcal{X}$. If all agents employ multi-agent OGD in Eq. (10.5), we have*

$$\mathbb{E}[\|x^T - x^\star\|^2] \le \tfrac{4G^2}{\beta^2 T},$$

where $x^\star \in \mathcal{X}$ denotes the unique Nash equilibrium. As a consequence, we have $\mathbb{E}[\|x^T - x^\star\|^2] = O(\frac{1}{T})$.

Remark 10.3.6 *Theorem 10.3.5 demonstrates that multi-agent OGD can achieve a near-optimal convergence rate in strongly monotone games; indeed, the convergence rate of multi-agent OGD matches the lower bound of $\Omega(\frac{1}{T})$ proved in Nemirovski and Yudin [1983] for strongly convex optimization.*

A drawback of multi-agent OGD is that it requires knowledge of the strong monotonicity parameters in order to update η^{t+1} and is thus not feasible in practice. We are not aware of any research that addresses this key issue. This is possibly because existing adaptive techniques are specialized to *upper curvature information*, e.g., the Lipschitz constant of function values or gradients [Duchi et al., 2011, Kingma and Ba, 2015, Mukkamala and

Algorithm 35 MA-AdaOGD$(x_1^1, x_2^1, \ldots, x_N^1, T)$

Input: initial points $x_i^1 \in \mathcal{X}_i$ for all $i \in \mathcal{N}$ and the total number of rounds T.
Initialization: $p_0 = \frac{1}{\log(T+10)}$.
for $t = 1, 2, \ldots, T$ **do**
  for $i = 1, 2, \ldots, N$ **do**
    sample $M_i^t \sim \mathsf{Geometric}(p_0)$.
    set $\eta_i^{t+1} \leftarrow \frac{t+1}{\sqrt{1+\max\{M_i^1, \ldots, M_i^t\}}}$.
    update $x_i^{t+1} \leftarrow \operatorname{argmin}_{x_i \in \mathcal{X}_i}\{(x_i - x_i^t)^\top \xi_i^t + \frac{\eta_i^{t+1}}{2}\|x_i - x_i^t\|^2\}$.

Hein, 2017, Levy, 2017, Levy et al., 2018, Bach and Levy, 2019, Antonakopoulos et al., 2021, Hsieh et al., 2021], and are not suitable for estimating *lower curvature information* such as that encoded by the strong monotonicity parameter. The goal of this section is to extend Algorithm 33 to multi-agent learning in games, showing that our adaptive variant of OGD is both doubly optimal and feasible.

We again employ a randomization strategy such that the resulting algorithm is adaptive to the strong monotonicity parameter and other problem parameters. We again choose independently identical distributed geometric random variables, i.e., $M_i^t \sim \mathsf{Geometric}(p_0)$, for $p_0 = \frac{1}{\log(T+10)}$.[6] Note that all the other updates are analogous to those of Algorithm 33 and the resulting algorithm is decentralized. This defines our adaptive multi-agent variant of OGD, as detailed in Algorithm 35.

Finite-time last-iterate convergence. In the following theorem, we summarize our main results on the last-iterate convergence rate of Algorithm 35 using a distance function based on squared Euclidean norm.

Theorem 10.3.7 *Suppose that a continuous game $\mathcal{G}$ is β-strongly monotone and let $G, D > 0$ be problem parameters satisfying $\mathbb{E}[\xi^t \mid x^t] = v(x^t)$ and $\mathbb{E}[\|\xi^t\|^2 \mid x^t] \leq G^2$ for all $t \geq 1$, and $\|x - x'\| \leq D$ for all $x, x' \in \mathcal{X}$. If all agents agree to employ Algorithm 35, we have*

$$\begin{aligned}\mathbb{E}[\|x^T - x^\star\|^2] \;\leq\; & \frac{D^2}{T}\left(1 + e^{\frac{1}{4\beta^2 \log(T+10)}}\right)\sqrt{1 + \log(T+10) + \log(NT)\log(T+10)} \\ & + \frac{G^2}{T}\log(T+1)\left(1 + \log(T+10) + \log(NT)\log(T+10)\right),\end{aligned}$$

where $x^\star \in \mathcal{X}$ is the unique Nash equilibrium. Thus, we have $\mathbb{E}[\|x^T - x^\star\|^2] = O(\frac{\log^3(T)}{T})$.

Remark 10.3.8 *Theorem 10.3.7 demonstrates that Algorithm 35 achieves a near-optimal convergence rate since the upper bound matches the lower bound [Nemirovski and Yudin, 1983] up to a log factor. This result together with Theorem 10.2.4 shows that our adaptive variant of OGD is* doubly optimal *and* feasible *for strongly monotone games.*

[6] We can define agent-specific probabilities $p_0^i \in (0, 1)$ and prove the same finite-time convergence guarantee. For simplicity, we use an agent-independent probability $p_0 = \frac{1}{\log(T+10)}$.

Remark 10.3.9 (Importance of doubly optimality) *We add the remark to help better position and appreciate the concept of double optimality. Our starting point is single-agent online learning in an arbitrarily non-stationary and possibly adversarial environment. It is well-known that the simple and computationally efficient online gradient descent achieves the minimax optimal regret bounds ($\Theta(\log)T$ for strongly convex cost functions and $\Theta(\sqrt{T})$). As such, to an agent engaged in an online decision making process, it is natural to expect OGD to be adopted to achieve optimal finite-horizon performance against such an environment where statistical regularity is lacking. Now, the most common instantiation of such a non-stationary environment consists of other agents who are simultaneously engaged in the online decision making process and whose actions impact all others' costs/rewards. In other words, each agent is acting in an environment, whose cost/reward is determined by an opaque game: the cost/reward is determined by – and hence realized from – an underlying game, where that game – and even the number of agents that comprise of the game – may be unknown. Consequently, the immediate next question is: if each agent adopts OGD –an optimal no-regret online learning algorithm– to maximize its finite-horizon cumulative reward, would the system jointly converge to a Nash equilibrium, a multi-agent optimal outcome where no agent has any incentive to unilaterally deviate? If it does not converge to a Nash equilibrium, then that means in the long run, an agent would be able to do better by deviating from what is prescribed by the online learning algorithm, given what all the other agents are doing, thereby producing "regret". Consequently, an online learning algorithm that is doubly optimal – that is, single agent adopting it leads to optimal finite-time no-regret guarantees and all agents adopting it leads to convergence to a Nash equilibrium – effectively bridges optimal transient performance (i.e. finite-horizon performance) with optimal long-run performance (equilibrium outcome).*

To prove Theorem 10.3.7, we provide another descent inequality for the iterates generated by Algorithm 35.

Lemma 10.3.10 *Suppose that a continuous game $\mathcal{G}$ is β-strongly monotone and let $G, D > 0$ be problem parameters satisfying $\mathbb{E}[\xi^t \mid x^t] = v(x^t)$ and $\mathbb{E}[\|\xi^t\|^2 \mid x^t] \leq G^2$ for all $t \geq 1$, and $\|x - x'\| \leq D$ for all $x, x' \in \mathcal{X}$. Letting the iterates $\{x^t\}_{t\geq 1}$ be generated by Algorithm 35, we have*

$$\sum_{i=1}^{N} \eta_i^T \mathbb{E}\left[\|x_i^T - x_i^\star\|^2 \mid \{\eta_i^t\}_{1\leq i\leq N, 1\leq t\leq T}\right] \leq \sum_{i=1}^{N} \eta_i^1 \|x_i^1 - x_i^\star\|^2$$
$$+D^2\left(\sum_{t=1}^{T-1}\left(\max\left\{0, \max_{1\leq i\leq N}\left\{\eta_i^{t+1} - \eta_i^t\right\} - 2\beta\right\}\right)\right) + G^2\left(\sum_{t=1}^{T-1}\left(\max_{1\leq i\leq N}\left\{\tfrac{1}{\eta_i^{t+1}}\right\}\right)\right),$$

where $x^\star \in \mathcal{X}$ is the unique Nash equilibrium.

Proof of Theorem 10.3.7. Since $D > 0$ satisfies that $\|x - x'\| \le D$ for all $x, x' \in \mathcal{X}$ and $\eta_i^{t+1} = \frac{t+1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}$ in Algorithm 35, we have

$$\sum_{i=1}^{N} \eta_i^1 \|x_i^1 - x_i^\star\|^2 \le D^2, \qquad \eta_i^{t+1} - \eta_i^t \le \frac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}.$$

By Lemma 10.3.10, we have

$$\begin{aligned}\sum_{i=1}^{N} \eta_i^T \mathbb{E}\left[\|x_i^T - x_i^\star\|^2 \mid \{\eta_i^t\}_{1\le i\le N, 1\le t\le T}\right] &\le D^2 \\ + D^2\left(\sum_{t=1}^{T-1}\left(\max\left\{0, \max_{1\le i\le N}\left\{\frac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\} - 2\beta\right\}\right)\right) &+ G^2\left(\sum_{t=1}^{T-1}\left(\max_{1\le i\le N}\left\{\frac{1}{\eta_i^{t+1}}\right\}\right)\right).\end{aligned} \tag{10.6}$$

Further, we have

$$\sum_{t=1}^{T-1}\left(\max_{1\le i\le N}\left\{\frac{1}{\eta_i^{t+1}}\right\}\right) \le \sqrt{1 + \max_{1\le i\le N, 1\le t\le T}\{M_i^t\}}\left(\sum_{t=1}^{T-1}\frac{1}{t+1}\right) \le \log(T+1)\sqrt{1 + \max_{1\le i\le N, 1\le t\le T}\{M_i^t\}}. \tag{10.7}$$

Plugging Eq. (10.7) into Eq. (10.6) yields that

$$\begin{aligned}\sum_{i=1}^{N} \eta_i^T \mathbb{E}\left[\|x_i^T - x_i^\star\|^2 \mid \{\eta_i^t\}_{1\le i\le N, 1\le t\le T}\right] &\le D^2 \\ + D^2\left(\sum_{t=1}^{T-1}\left(\max\left\{0, \max_{1\le i\le N}\left\{\frac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\} - 2\beta\right\}\right)\right) &+ G^2\log(T+1)\sqrt{1 + \max_{1\le i\le N, 1\le t\le T}\{M_i^t\}}.\end{aligned}$$

By the definition of η_i^T, we have

$$\eta_i^T \ge \frac{T}{\sqrt{1+\max\{M_i^1,\ldots,M_i^T\}}} \ge \frac{T}{\sqrt{1+\max_{1\le i\le N, 1\le t\le T}\{M_i^t\}}},$$

which implies that

$$\sum_{i=1}^{N} \eta_i^T \mathbb{E}\left[\|x_i^T - x_i^\star\|^2 \mid \{\eta_i^t\}_{1\le i\le N, 1\le t\le T}\right] \ge \frac{T}{\sqrt{1+\max_{1\le i\le N, 1\le t\le T}\{M_i^t\}}} \cdot \mathbb{E}\left[\|x^T - x^\star\|^2 \mid \{\eta_i^t\}_{1\le i\le N, 1\le t\le T}\right].$$

Putting these pieces together yields that

$$\begin{aligned}\left(\frac{T}{\sqrt{1+\max_{1\le i\le N, 1\le t\le T}\{M_i^t\}}}\right)\mathbb{E}\left[\|x^T - x^\star\|^2 \mid \{\eta_i^t\}_{1\le i\le N, 1\le t\le T}\right] &\le D^2 \\ + D^2\left(\sum_{t=1}^{T-1}\left(\max\left\{0, \max_{1\le i\le N}\left\{\frac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\} - 2\beta\right\}\right)\right) &+ G^2\log(T+1)\sqrt{1 + \max_{1\le i\le N, 1\le t\le T}\{M_i^t\}}.\end{aligned}$$

Rearranging and taking the expectation of both sides, we have

$$T \cdot \mathbb{E}[\|x^T - x^\star\|^2] \leq D^2 \underbrace{\mathbb{E}\left[\sqrt{1 + \max_{1\leq i\leq N, 1\leq t\leq T}\{M_i^t\}}\right]}_{\mathbf{I}} + G^2 \log(T+1) \underbrace{\mathbb{E}\left[1 + \max_{1\leq i\leq N, 1\leq t\leq T}\{M_i^t\}\right]}_{\mathbf{II}}$$
$$+ D^2 \underbrace{\mathbb{E}\left[\sum_{t=1}^{T-1} \max\left\{0, \max_{1\leq i\leq N}\left\{\frac{\sqrt{1+\max_{1\leq i\leq N,1\leq t\leq T}\{M_i^t\}}}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\} - 2\beta\sqrt{1 + \max_{1\leq i\leq N,1\leq t\leq T}\{M_i^t\}}\right\}\right]}_{\mathbf{III}} \tag{10.8}$$

By using the previous argument with Proposition 10.2.6 and $p = \frac{1}{\log(T+10)}$, we have

$$\mathbf{I} \leq \sqrt{1 + \tfrac{1+\log(NT)}{p}} = \sqrt{1 + \log(T+10) + \log(NT)\log(T+10)}. \tag{10.9}$$

and

$$\mathbf{II} \leq 1 + \tfrac{1+\log(NT)}{p} = 1 + \log(T+10) + \log(NT)\log(T+10). \tag{10.10}$$

It remains to bound the term **III** using Proposition 10.2.6 and $p = \frac{1}{\log(T+10)}$. Indeed, we have

$$\begin{aligned}
\mathbf{III} &= \mathbb{E}\left[\sum_{t=1}^{T-1} \max\left\{0, \max_{1\leq i\leq N}\left\{\frac{\sqrt{1+\max_{1\leq i\leq N,1\leq t\leq T}\{M_i^t\}}}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\} - 2\beta\sqrt{1 + \max_{1\leq i\leq N,1\leq t\leq T}\{M_i^t\}}\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T-1} \max_{1\leq i\leq N}\left\{\frac{\sqrt{1+\max_{1\leq i\leq N,1\leq t\leq T}\{M_i^t\}}}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\} \cdot \mathbb{I}\left(\max_{1\leq i\leq N}\left\{\frac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\} - 2\beta \geq 0\right)\right].
\end{aligned}$$

Defining $i^t = \operatorname{argmax}_{1\leq i\leq N}\left\{\frac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\}$ as a random variable and recalling that $\{\max\{M_i^1,\ldots,M_i^t\}\}_{1\leq i\leq N}$ are independent and identically distributed, we have that $i^t \in \{1,\ldots,N\}$ is uniformly distributed. This implies that

$$\begin{aligned}
\mathbf{III} &\leq \tfrac{1}{N}\sum_{t=1}^{T-1}\sum_{j=1}^{N} \mathbb{E}\left[\max_{1\leq i\leq N}\left\{\frac{\sqrt{1+\max_{1\leq i\leq N,1\leq t\leq T}\{M_i^t\}}}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\} \cdot \mathbb{I}\left(\max_{1\leq i\leq N}\left\{\frac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\} - 2\beta \geq 0\right) \mid i^t = j\right] \\
&= \tfrac{1}{N}\sum_{t=1}^{T-1}\sum_{j=1}^{N} \mathbb{E}\left[\frac{\sqrt{1+\max_{1\leq i\leq N,1\leq t\leq T}\{M_i^t\}}}{\sqrt{1+\max\{M_j^1,\ldots,M_j^t\}}} \cdot \mathbb{I}\left(\frac{1}{\sqrt{1+\max\{M_j^1,\ldots,M_j^t\}}} - 2\beta \geq 0\right)\right] \\
&\leq \tfrac{1}{N}\sum_{t=1}^{T-1}\sum_{j=1}^{N} \mathbb{E}\left[\sqrt{1 + \max_{1\leq i\leq N,1\leq t\leq T,i\neq j}\{M_i^t\}} \cdot \mathbb{I}\left(\frac{1}{\sqrt{1+\max\{M_j^1,\ldots,M_j^t\}}} - 2\beta \geq 0\right)\right].
\end{aligned}$$

Since $\max_{1\leq i\leq N,1\leq t\leq T,i\neq j}\{M_i^t\}$ is independent of $\max\{M_j^1,\ldots,M_j^t\}$, we have

$$\mathbf{III} \leq \tfrac{1}{N}\sum_{t=1}^{T-1}\sum_{j=1}^{N} \mathbb{E}\left[\sqrt{1 + \max_{1\leq i\leq N,1\leq t\leq T,i\neq j}\{M_i^t\}}\right] \cdot \mathbb{P}\left(\frac{1}{\sqrt{1+\max\{M_j^1,\ldots,M_j^t\}}} - 2\beta \geq 0\right).$$

Since $\{M_i^t\}_{1\le i\le N, 1\le t\le T}$ is a sequence of independent and identically distributed geometric random variables with $p = \frac{1}{\log(T+10)}$, Proposition 10.2.6 implies that

$$\mathbb{E}\left[\max_{1\le i\le N, 1\le t\le T, i\neq j}\{M_i^t\}\right] \le \frac{1+\log(NT)}{p} = \log(T+10) + \log(NT)\log(T+10).$$

and

$$\sum_{t=1}^{T-1}\mathbb{P}\left(\frac{1}{\sqrt{1+\max\{M_j^1,\ldots,M_j^t\}}} - 2\beta \ge 0\right) \le e^{\frac{p}{4\beta^2}} = e^{\frac{1}{4\beta^2\log(T+10)}}.$$

Putting these pieces together with Jensen's inequality yields that

$$\textbf{III} \le e^{\frac{1}{4\beta^2\log(T+10)}}\sqrt{1+\log(T+10)+\log(NT)\log(T+10)}. \tag{10.11}$$

Plugging Eq. (10.9), Eq. (10.10) and Eq. (10.11) into Eq. (10.8) yields that

$$\begin{aligned}\mathbb{E}[\|x^T - x^\star\|^2] \;\le\; & \frac{D^2}{T}(1+e^{\frac{1}{4\beta^2\log(T+10)}})\sqrt{1+\log(T+10)+\log(NT)\log(T+10)} \\ & + \frac{G^2}{T}\log(T+1)(1+\log(T+10)+\log(NT)\log(T+10)).\end{aligned}$$

This completes the proof.

It is worth remarking that our proof does not use the structure of $v(\cdot)$ (i.e., $v_i(x) = \nabla_{x_i}u_i(x_i; x_{-i})$). Thus, Theorem 10.3.7, when extended to the VI setting, yields a decentralized, feasible optimization algorithm for finding a solution of a strongly monotone VI, contributing to that line of literature.

Corollary 10.3.11 *Suppose that the variational inequality defined by $v(\cdot)$ (without the structure that $v_i(x) = \nabla_{x_i}u_i(x_i; x_{-i})$) and $\mathcal{X}$ (any convex set) is β-strongly monotone and let $G, D > 0$ be problem parameters satisfying $\mathbb{E}[\xi^t \mid x^t] = v(x^t)$ and $\mathbb{E}[\|\xi^t\|^2 \mid x^t] \le G^2$ for all $t \ge 1$, and $\|x - x'\| \le D$ for all $x, x' \in \mathcal{X}$. Then, by employing Algorithm 35, we have*

$$\begin{aligned}\mathbb{E}[\|x^T - x^\star\|^2] \;\le\; & \frac{D^2}{T}(1+e^{\frac{1}{4\beta^2\log(T+10)}})\sqrt{1+\log(T+10)+\log(NT)\log(T+10)} \\ & + \frac{G^2}{T}\log(T+1)(1+\log(T+10)+\log(NT)\log(T+10)),\end{aligned}$$

where $x^\star \in \mathcal{X}$ is the unique solution of he VI. Thus, we have $\mathbb{E}[\|x^T - x^\star\|^2] = O(\frac{\log^3(T)}{T})$.

Applications: Feasible multi-agent learning for power management and Newsvendors with lost sales. We present two typical examples that satisfy the conditions in Definition 10.3.2, where only (noisy) gradient feedback is available.

Example 10.3.1 (Power Management in Wireless Networks) *Target-rate power management problems are well known in operations research and wireless communications [Rappaport, 2001, Goldsmith, 2005]. We consider a wireless network of N communication links, each link consisting of a transmitter (e.g., phone, tablet and sensor) and an intended receiver*

(each transmitter consumes power to send signals to their intended receivers). Assume further that the transmitter in the i^{th} link transmits with power $a_i \geq 0$ and let $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$ denote the joint power profile of all transmitters in the network. In this context, the quality-of-service rate of the i^{th} link depends not only on how much power its transmitter is employing but also on how much power all the other transmitters are concurrently employing. Formally, the i^{th} link's quality-of-service rate is given by $r_i(a) = \frac{G_{ii}a_i}{\sum_{j\neq i} G_{ij}a_j + \eta_i}$ where η_i is the thermal noise associated with the receiver of the i^{th} link and $G_{ij} \geq 0$ is the unit power gain between the transmitter in the j^{th} link and the receiver in the i^{th} link, which is determined by the network topology but is unknown in practice. Note that $\sum_{j\neq i} G_{ij}a_j$ is the interference caused by other links to link i—all else being equal, the larger the powers of transmitters in other links, the lower the service rate the i^{th} link. Intuitively, the transmitter in the i^{th} link aims to balance between two objectives: maintaining a target service rate $r_i^\star$ while consuming as little power as possible. This consideration leads to the following standard cost function [Rappaport, 2001, Goldsmith, 2005]:

$$u_i(a) = \frac{a_i^2}{2}\left(1 - \frac{r_i^\star}{r_i(a)}\right)^2 = \frac{1}{2}\left(a_i - \frac{r_i^\star(\sum_{j\neq i} G_{ij}a_j + \eta_i)}{G_{ii}}\right)^2.$$

Notably, $u_i(a) = 0$ if the realized service rate $r_i(a)$ is equal to the target rate $r_i^\star$. Otherwise, there will be a cost either due to not meeting $r_i^\star$ or due to consuming unnecessary power. Our prior work has shown that this is a β-strongly monotone game [Zhou et al., 2021] with $\beta = \lambda_{\min}(I - \frac{1}{2}(W + W^\top)) > 0$, where $W_{ii} = 0$ for all $1 \leq i \leq N$ and $W_{ij} = \frac{r_i^\star G_{ij}}{G_{ii}}$ for $i \neq j$.

Finally, gradient feedback is available for the aforementioned class of strongly monotone games; indeed, the i^{th} link's quality-of-service rate $r_i(\mathbf{a})$ is available to the transmitter in the i^{th} link, who can use it to compute the gradient of $u_i(\mathbf{a})$ with respect to a_i.

In this setting, the parameter $\beta > 0$ is not available in practice since finding the minimal eigenvalue is computationally prohibitive. However, the retailer can apply our MA-AdaOGD algorithm (cf. Algorithm 35) and obtain a near-optimal last-iterate convergence rate of $O(\frac{\log^3(T)}{T})$.

Example 10.3.2 (Newsvendor with Lost Sales) *We consider the multi-retailer generalization of newsvendor problem [Netessine and Rudi, 2003]. For simplicity, we focus on a single product with same per-unit price $p > 0$ and perishable inventory control. Given the set of retailers $\mathcal{N} = \{1, 2, \ldots, N\}$, the i^{th} retailer's action x_i is assumed to lie in the interval $[0, \bar{x}_i]$. For the i^{th} retailer, the demand is random and depends on the inventory levels of other retailers, therefore we denote it as $D_i(x_{-i})$ and let $d_i \geq 0$ denote a realization of this random variable. The i^{th} retailer only observes the sales quantity $\min\{x_i, d_i\}$. Using the previous argument, the i^{th} retailer's cost function is defined by*

$$u_i(x) = (p - c_i) \cdot \mathbb{E}[\max\{0, D_i(x_{-i}) - x_i\}] + c_i \cdot \mathbb{E}[\max\{0, x_i - D_i(x_{-i})\}], \tag{10.12}$$

and the noisy gradient feedback signal is given by

$$\xi_i = \begin{cases} c_i, & \text{if } x_i \geq \min\{x_i, d_i\}, \\ c_i - p, & \text{otherwise.} \end{cases}$$

It remains to extend the analysis from Huh and Rusmevichientong [2009, Section 3.5] and provide a condition on the distribution of $D_i(x_{-i})$ that can guarantee that the multi-retailer newsvendor problem is β-strongly monotone for some constant $\beta > 0$. Indeed, after simple calculations, we have

$$v_i(x) = \nabla_{x_i} u_i(x) = p \cdot \mathbb{P}(D_i(x_{-i}) \leq x_i) - p + c_i = p \cdot F_i(x) - p + c_i.$$

Letting $F = (F_1, F_2, \ldots, F_N)$ be an operator from $\prod_{i=1}^N [0, \bar{x}_i]$ to $[0,1]^N$, if F is α-strongly monotone, we have

$$\langle x' - x, v(x') - v(x) \rangle = p \cdot \langle x' - x, F(x') - F(x) \rangle \geq \alpha p \|x' - x\|^2, \quad \text{for all } x, x' \in \prod_{i=1}^N [0, \bar{x}_i].$$

As a concrete example, we can let the distribution of a demand $D_i(x_{-i})$ be given by

$$\mathbb{P}(D_i(x_{-i}) \leq z) = 1 - \frac{1 + \sum_{j \neq i} x_j}{(1 + z + \sum_{j \neq i} x_j)^2}.$$

It is clear that $\lim_{z \to +\infty} \mathbb{P}(D_i(x_{-i}) \leq z) = 1$ for all $x_{-i} \in \prod_{j \neq i} [0, \bar{x}_j]$. Also, we have

$$\mathbb{P}(D_i(x_{-i}) = 0) = 1 - \frac{1}{1 + \sum_{j \neq i} x_j},$$

which characterizes the dependence of the distribution of the demand for the i^{th} retailer on other retailers' actions x_{-i}. Indeed, the demand $D_i(x_{-i})$ is likely to be small if the total inventory provided by other retailers $\sum_{j \neq i} x_j$ is large. Two extreme cases are (i) $\mathbb{P}(D_i(x_{-i}) = 0) \to 1$ as $\sum_{j \neq i} x_j \to +\infty$ and (ii) $\mathbb{P}(D_i(x_{-i}) = 0) = 0$ as $\sum_{j \neq i} x_j = 0$. Finally, we can prove that this is a β-strongly monotone game with $\beta = \frac{p}{(1 + \sum_{k=1}^N \bar{x}_k)^3} > 0$.

Thus, if all the retailers agree to apply our MA-AdaOGD algorithm (cf. Algorithm 35), we can obtain a near-optimal last-iterate convergence rate of $O(\frac{\log^3(T)}{T})$.

We summarize Algorithm 35 specialized to the multi-retailer generalization of the newsvendor problem in Algorithm 36 and we present the corresponding result on the last-iterate convergence rate guarantee in the following corollary.

Corollary 10.3.12 *In the multi-retailer generalization of the newsvendor problem, each retailer's cost functions are defined by Eq. (10.12) where the unit purchase cost is $c_i > 0$ and the unit selling price is $p \geq c$. Also, the demand is a random variable with a continuous density function defined in Example 10.3.2. If the agent employs Algorithm 36, we have*

$$\begin{aligned} \mathbb{E}[\|x^T - x^\star\|^2] \quad \leq \quad & \frac{\sum_{i=1}^N \bar{x}_i^2}{T} (1 + e^{\frac{1}{4\beta^2 \log(T+10)}}) \sqrt{1 + \log(T+10) + \log(NT)\log(T+10)} \\ & + \frac{Np^2}{T} \log(T+1)(1 + \log(T+10) + \log(NT)\log(T+10)), \end{aligned}$$

As a consequence, we have $\mathbb{E}[\|x^T - x^\star\|^2] = O(\frac{\log^3(T)}{T})$.

Algorithm 36 Newsvendor-MA-AdaOGD$(x_1^1, x_2^1, \ldots, x_N^1, T)$

Input: initial point $x_i^1 \in \mathcal{X}_i$ for all $i \in \mathcal{N}$ and the total number of rounds T.
Initialization: $p_0 = \frac{1}{\log(T+10)}$.
for $t = 1, 2, \ldots, T$ **do**
  for $i = 1, 2, \ldots, N$ **do**
    sample $M_i^t \sim \mathsf{Geometric}(p_0)$.
    set $\eta_i^{t+1} \leftarrow \frac{t+1}{\sqrt{1+\max\{M_i^1, \ldots, M_i^t\}}}$.
    observe the sales quantity of $\min\{x^t, d^t\}$.
    update $x_i^{t+1} \leftarrow \begin{cases} \operatorname{argmin}_{x_i \in [0, \bar{x}_i]} \{(x_i - x_i^t)c_i + \frac{\eta_i^{t+1}}{2}(x_i - x_i^t)^2\}, & \text{if } x_i^t \geq \min\{x_i^t, d_i^t\}, \\ \operatorname{argmin}_{x_i \in [0, \bar{x}_i]} \{(x_i - x_i^t)(c_i - p) + \frac{\eta_i^{t+1}}{2}(x_i - x_i^t)^2\}, & \text{otherwise.} \end{cases}$

Proof. Recall that $u_i(x^t) = (p - c_i) \cdot \mathbb{E}[\max\{0, D_i(x_{-i}^t) - x_i^t\}] + c_i \cdot \mathbb{E}[\max\{0, x_i^t - D_i(x_{-i}^t)\}]$ and the noisy gradient feedback is given by

$$\xi_i^t = \begin{cases} c_i, & \text{if } x_i^t \geq \min\{x_i^t, d_i^t\}, \\ c_i - p, & \text{otherwise.} \end{cases}$$

So we have $\mathbb{E}[\xi^t \mid x^t] = v(x^t)$ and $\mathbb{E}[\|\xi^t\|^2 \mid x^t] \leq Np^2$ for all $t \geq 1$. We later show that this is a β-strongly monotone game with $\beta = \frac{p}{(1+\sum_{k=1}^N \bar{x}_k)^3} > 0$. In addition, $\mathcal{X}_i = [0, \bar{x}_i]$ implies that $D^2 = \sum_{i=1}^N \bar{x}_i^2$. Thus, Theorem 10.3.7 can be applied and implies the desired result. □

10.4 Extensions to Exp-Concave Cost Functions and Games

We extend our results—for both single-agent and multi-agent learning—to a broader class of cost functions and game structures defined in terms of exp-concavity. For exp-concave cost functions, our adaptive variant of an online Newton step (AdaONS) can achieve a near-optimal regret of $O(d \log^2(T))$. For exp-concave games, we propose a multi-agent online Newton step with a near-optimal time-average convergence rate of $O(\frac{d \log(T)}{T})$. We also extend it to a multi-agent adaptive online Newton step (MA-AdaONS) that achieves a near-optimal rate of $O(\frac{d \log^2(T)}{T})$.

Single-agent learning with exp-concave cost. We start by recalling the definition of exp-concave functions.

Definition 10.4.1 *A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is α-exp-concave if $e^{-\alpha f(\cdot)}$ is concave.*

Exp-Concave (EC) functions, as a strict subclass of convex functions and as a strict superclass of strongly convex functions, have found applications in many fields, including information

theory [Cover, 1999], stochastic portfolio theory [Fernholz, 2002], optimal transport [Villani, 2009], optimization [Hazan et al., 2014, Mahdavi et al., 2015], probability theory [Pal et al., 2018] and statistics [Juditsky et al., 2008, Koren and Levy, 2015, Yang et al., 2018].

We work with the following model of perfect gradient feedback:[7]:

1. At each round t, an exact gradient is observed. That is, we have $\nabla f_t(x^t)$ for all $t \geq 1$.
2. The action set $\mathcal{X}$ is bounded by a diameter $D > 0$, i.e., $\|x - x'\| \leq D$ for all $x, x' \in \mathcal{X}$.

For this class of cost functions, a lower bound on regret of $\Omega(d \log(T))$ is known [see Section 3 in Abernethy et al., 2008]. Two typical examples which admit EC cost functions are online linear regression [Kivinen and Warmuth, 1999] and universal portfolio management [Cover, 1991].

1. **In universal portfolio management**, we have $f_t(x) = -\log(a_t^\top x)$ where $x \in \Delta^d$ is a probability vector and $a^t \in \mathbb{R}^d_+$ is the d-dimensional vector of all d assets' relative prices between period t and period $t-1$. Here, f_t is EC and Ordentlich and Cover [1998] established a lower bound of $\Omega(d \log(T))$ for the cumulative regret.
2. **In online linear regression**, we have $f_t(x) = (a_t^\top x - b_t)^2$, which is in general not a strongly convex function, but is indeed EC for any a_t, b_t. Formally, Vovk [1997] proved that there exists a randomized strategy of the adversary for choosing the vectors a_t, b_t such that the expected cumulative regret scales as $\Omega(d \log(T))$.

This lower bound was established without making the connectio to the general EC function class. Note also that the $\Omega(d \log(T))$ lower bound is interesting given that $d \geq 1$ does not enter the lower bound for strongly convex functions [Hazan and Kale, 2014].

On the upper-bound side, there are many algorithms that achieve the minimax-optimal regret, such as the online Newton step (ONS) of Hazan et al. [2007]. Formally, let $\{f_t\}_{t \geq 1}$ be α-exp-concave cost functions and let $G > 0$ be problem parameters satisfying $\|\nabla f_t(x)\| \leq G$ for all $t \geq 1$. Then, the ONS algorithm is given by

$$\begin{aligned} &\eta \leftarrow \tfrac{1}{2} \min\left\{\tfrac{1}{4GD}, \alpha\right\}, \qquad A^{t+1} \leftarrow A^t + \nabla f_t(x^t) \nabla f_t(x^t)^\top, \\ &x^{t+1} \leftarrow \operatorname{argmin}_{x \in \mathcal{X}} \{(x - x^t)^\top \nabla f_t(x^t) + \tfrac{\eta}{2}(x - x^t)^\top A^{t+1}(x - x^t)\}. \end{aligned} \tag{10.13}$$

The choice $\frac{1}{2}\min\{\frac{1}{4GD}, \alpha\}$ comes from Hazan et al. [2007, Lemma 3]: if $\|\nabla f(x)\| \leq G$ and $\|x - x'\| \leq D$ for all $x, x' \in \mathcal{X}$, the α-exp-concave function f satisfies

$$f(x') \geq f(x) + (x' - x)^\top \nabla f(x) + \tfrac{1}{4}\min\{\tfrac{1}{4GD}, \alpha\}(x' - x)^\top (\nabla f(x) \nabla f(x)^\top)(x' - x). \tag{10.14}$$

Intuitively, this equation implies that any α-exp-concave function can be approximated by a local quadratic function with a matrix $\eta \nabla f(x) \nabla f(x)^\top$ and moreover ONS can exploit such

[7] Can we generalize the aforementioned results to the setting with noisy gradient feedback? We leave the answer to this question to future work.

Algorithm 37 AdaONS(x^1, T)

Input: initial point $x^1 \in \mathcal{X}$ and the total number of rounds T.
Initialization: $A^1 = I_d$ where $I_d \in \mathbb{R}^{d\times d}$ is an identity matrix and $p_0 = \frac{1}{\log(T+10)}$.
for $t = 1, 2, \ldots, T$ **do**
  sample $M^t \sim \mathsf{Geometric}(p_0)$.
  set $\eta^{t+1} \leftarrow \frac{1}{\sqrt{1+\max\{M^1,\ldots,M^t\}}}$.
  update $A^{t+1} \leftarrow A^t + \nabla f_t(x^t)\nabla f_t(x^t)^\top$.
  update $x^{t+1} \leftarrow \operatorname{argmin}_{x\in\mathcal{X}}\{(x - x^t)^\top \nabla f_t(x^t) + \frac{\eta^{t+1}}{2}(x - x^t)^\top A^{t+1}(x - x^t)\}$.

structure. Although ONS suffers from a quadratic dependence on the dimension in terms of per-iteration cost, there has been progress in alleviating this complexity. For example, Luo et al. [2016] proposed a variant of ONS that attains linear dependence on the dimension using sketching techniques.

Feasible single-agent online learning under EC cost. The scheme of ONS in Eq. (10.13) is infeasible in practice since it requires prior knowledge of the problem parameters α, G and D. To address this issue, we invoke the same randomization strategy as described in Algorithm 33. This results in the single-agent adaptive variant of ONS (cf. Algorithm 37).

We summarize our main results on the regret minimization properties of the algorithm in the following theorem.

Theorem 10.4.2 *Consider an arbitrary fixed sequence of α-exp-concave functions $f_1, \ldots, f_T$, where each f_t satisfies $\|\nabla f_t(x)\| \le G$ for all $t \ge 1$ and $\|x - x'\| \le D$ for all $x, x' \in \mathcal{X}$. If the agent employs Algorithm 37, we have*

$$\mathbb{E}[\mathit{Regret}(T)] \le \frac{D^2}{2}(1+G^2 e^{\frac{(\max\{8GD,2\alpha^{-1}\})^2}{\log(T+10)}}) + \frac{d\log(TG^2+1)}{2}\sqrt{1+\log(T+10)+\log(T)\log(T+10)}.$$

Thus, we have $\mathbb{E}[\mathit{Regret}(T)] = O(d\log^2(T))$.

Remark 10.4.3 *Theorem 10.4.2 demonstrates that Algorithm 37 achieves a near-optimal regret since the upper bound matches the lower bound up to a log factor; indeed, the lower bound of $\Omega(d\log(T))$ has been established for online linear regression and universal portfolio management. Further, Algorithm 37 dynamically adjusts η^{t+1} without any prior knowledge of problem parameters.*

To prove Theorem 10.4.2, we again make use of a key descent inequality.

Lemma 10.4.4 *Consider an arbitrary fixed sequence of α-exp-concave functions $f_1, \ldots, f_T$, where each f_t satisfies $\|\nabla f_t(x)\| \le G$ for all $t \ge 1$ and $\|x - x'\| \le D$ for all $x, x' \in \mathcal{X}$.*

Letting the iterates $\{x^t\}_{t\geq 1}$ *be generated by Algorithm 37, we have*

$$\sum_{t=1}^{T} f_t(x^t) - \sum_{t=1}^{T} f_t(x) \leq \tfrac{\eta^1}{2}(x^1 - x)^\top A^1 (x^1 - x) + \tfrac{1}{2}\left(\sum_{t=1}^{T} \tfrac{1}{\eta^{t+1}} \nabla f_t(x^t)^\top (A^{t+1})^{-1} \nabla f_t(x^t)\right)$$
$$+ \sum_{t=1}^{T} (x^t - x)^\top \left(\tfrac{\eta^{t+1}}{2} A^{t+1} - \tfrac{\eta^t}{2} A^t - \tfrac{1}{4}\min\{\tfrac{1}{4GD}, \alpha\} \nabla f_t(x^t) \nabla f_t(x^t)^\top\right)(x^t - x), \quad \text{for all } x \in \mathcal{X}.$$

Exp-concave games. We define a class of exp-concave (EC) games via a game-theoretic generalization of the optimization framework in the previous section. Letting f be α-exp-concave with $\|\nabla f(x)\| \leq G$ and $\|x - x'\| \leq D$ for all $x, x' \in \mathcal{X}$, Hazan et al. [2007, Lemma 3] guarantees that

$$f(x') \geq f(x) + (x' - x)^\top \nabla f(x) + \tfrac{1}{4}\min\{\tfrac{1}{4GD}, \alpha\}(x' - x)^\top (\nabla f(x) \nabla f(x)^\top)(x' - x).$$

Equivalently, we have

$$(x' - x)^\top (\nabla f(x') - \nabla f(x)) \geq \tfrac{1}{4}\min\{\tfrac{1}{4GD}, \alpha\}(x' - x)^\top (\nabla f(x') \nabla f(x')^\top + \nabla f(x) \nabla f(x)^\top)(x' - x).$$

This leads to the following formal definition in which $v(\cdot)$ replaces $\nabla f(\cdot)$.

Definition 10.4.5 *A continuous game* $\mathcal{G}$ *is said to be* α*-exp-concave if we have* $\langle x' - x, v(x') - v(x)\rangle \geq \frac{1}{4}\min\{\frac{1}{4GD}, \alpha\}(\sum_{i=1}^{N}(x_i' - x_i)^\top (v_i(x') v_i(x')^\top + v_i(x) v_i(x)^\top)(x_i' - x_i))$ *for all* $x, x' \in \mathcal{X}$.

Since EC games satisfy the individual convexity condition, we have the existence of at least one Nash equilibrium. However, multiple Nash equilibria might exist for EC games. For example, letting $d \geq 2$, we consider a single-agent game with the cost function $f(x) = \frac{1}{2}(a^\top x)^2$ and $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$. It is easy to verify that this is a $(2/\|a\|^2)$-exp-concave game and any $x \in \mathcal{X}$ satisfying $a^\top x = 0$ will be a Nash equilibrium. It is thus natural to ask how to measure the optimality of a point $\hat{x} \in \mathcal{X}$. Proposition 10.3.4 inspires us to use a gap function, $\textsc{gap}(\cdot) : \mathcal{X} \mapsto \mathbb{R}_+$, given by

$$\textsc{gap}(\hat{x}) = \sup_{x \in \mathcal{X}} (\hat{x} - x)^\top v(x). \tag{10.15}$$

The above gap function is well defined and is nonnegative for all $\hat{x} \in \mathcal{X}$, given that at least one Nash equilibrium exists in EC games. Note that such a function has long been a standard example in the literature [Facchinei and Pang, 2007, Mertikopoulos and Zhou, 2019].

All strongly monotone games are clearly EC games if $\|v(x)\| \leq G$ for all $x \in \mathcal{X}$. Indeed, we have $\sum_{i=1}^{N}(x_i' - x_i)^\top (v_i(x') v_i(x')^\top + v_i(x) v_i(x)^\top)(x_i' - x_i) \leq 2G^2 \|x' - x\|^2$ and

$$\langle x' - x, v(x') - v(x)\rangle \geq \beta \|x' - x\|^2 \geq \tfrac{\beta}{2G^2}\left(\sum_{i=1}^{N}(x_i' - x_i)^\top (v_i(x') v_i(x')^\top + v_i(x) v_i(x)^\top)(x_i' - x_i)\right).$$

This implies that a β-strongly monotone game is α-exp-concave if we have $\alpha \leq \frac{2\beta}{G^2}$. More generally, we remark that strongly monotone games are a rich class of games that arise in many real-world application problems [Bravo et al., 2018, Mertikopoulos and Zhou, 2019, Lin et al., 2021b]. Moreover, there are also applications that can be cast as EC games rather than strongly monotone games, such as empirical risk minimization [Koren and Levy, 2015, Yang et al., 2018] and universal portfolio management [Cover, 1991, Fernholz, 2002].

Multi-agent online learning in EC games. We start by extending the single-agent ONS algorithm in Eq. (10.13) to the multi-agent online learning in EC games (cf. Definition 10.4.5).

We again work with a model of exact gradient feedback:

1. At each round t, an exact gradient is observed. That is, we have $v(x^t)$ for all $t \geq 1$.
2. The action set $\mathcal{X} = \Pi_{i=1}^N \mathcal{X}_i$ is bounded by a diameter $D > 0$, i.e., $\|x - x'\| \leq D$ for all $x, x' \in \mathcal{X}$.

Letting $x_i^1 \in \mathcal{X}_i$ for all $i \in \mathcal{N}$, the multi-agent version of ONS performs the following step at each round ($A_i^1 = I_{d_i}$ for all $i \in \mathcal{N}$):

$$\begin{aligned} &\eta_i \leftarrow \tfrac{1}{2} \min\left\{\tfrac{1}{4GD}, \alpha\right\}, \quad A_i^{t+1} \leftarrow A_i^t + v_i(x^t) v_i(x^t)^\top, \\ &x_i^{t+1} \leftarrow \operatorname{argmin}_{x_i \in \mathcal{X}_i} \{(x_i - x_i^t)^\top v_i(x^t) + \tfrac{\eta_i}{2}(x_i - x_i^t)^\top A_i^{t+1}(x_i - x_i^t)\}. \end{aligned} \tag{10.16}$$

We see from Eq. (10.16) that our multi-agent learning algorithm is a generalization of single-agent ONS in Eq. (10.13). Notably, it is a decentralized algorithm since each agent does not need to know any other agents' gradients. In the following theorem, we summarize our main results on the time-average convergence rate using the gap function in Eq. (10.15).

Theorem 10.4.6 *Suppose that a continuous game $\mathcal{G}$ is α-exp-concave and let $G, D > 0$ be problem parameters satisfying $\|v(x)\| \leq G$ and $\|x - x'\| \leq D$ for all $x, x' \in \mathcal{X}$. If all agents agree to employ multi-agent ONS in Eq. (10.16), we have*

$$\textsc{gap}(\bar{x}^T) \leq \tfrac{\alpha D^2}{4T} + \tfrac{d \log(TG^2+1)}{T} \max\left\{4GD, \tfrac{1}{\alpha}\right\},$$

where $\bar{x}^T = \frac{1}{T}\sum_{t=1}^T x^t$ denotes a time-average iterate. Thus, we have $\textsc{gap}(\bar{x}^T) = O(\frac{d \log(T)}{T})$.

Remark 10.4.7 *Theorem 10.4.6 shows that the multi-agent ONS algorithm can achieve a near-optimal convergence rate in EC games; indeed, the convergence rate of multi-agent ONS matches up to a log factor the lower bound of $\Omega(\frac{d}{T})$ proved in Mahdavi et al. [2015] for exp-concave optimization.*

The proof of Theorem 10.4.6 is again based on a descent inequality.

Algorithm 38 MA-AdaONS$(x_1^1, x_2^1, \ldots, x_N^1, T)$

Input: initial points $x_i^1 \in \mathcal{X}_i$ for all $i \in \mathcal{N}$ and the total number of rounds T.
Initialization: $A_i^1 = I_{d_i}$ where $I_{d_i} \in \mathbb{R}^{d_i \times d_i}$ is an identity matrix and $p_0 = \frac{1}{\log(T+10)}$.
for $t = 1, 2, \ldots, T$ **do**
  for $i = 1, 2, \ldots, N$ **do**
    sample $M_i^t \sim \mathsf{Geometric}(p_0)$.
    set $\eta_i^{t+1} \leftarrow \frac{1}{\sqrt{1+\max\{M_i^1, \ldots, M_i^t\}}}$.
    update $A_i^{t+1} \leftarrow A_i^t + v_i(x^t) v_i(x^t)^\top$.
    update $x_i^{t+1} \leftarrow \operatorname{argmin}_{x_i \in \mathcal{X}_i} \{(x_i - x_i^t)^\top v_i(x^t) + \frac{\eta_i^{t+1}}{2}(x_i - x_i^t)^\top A_i^{t+1}(x_i - x_i^t)\}$.

Lemma 10.4.8 *Suppose that a continuous game $\mathcal{G}$ is α-exp-concave and let $G, D > 0$ be problem parameters satisfying $\|v(x)\| \le G$ and $\|x - x'\| \le D$ for all $x, x' \in \mathcal{X}$. Letting the iterates $\{x^t\}_{t \ge 1}$ be generated by the multi-agent ONS in Eq. (10.16), we have*

$$\sum_{t=1}^{T} (x^t - x)^\top v(x) \le \sum_{i=1}^{N} \tfrac{\eta_i}{2} (x_i^1 - x_i)^\top A_i^1 (x_i^1 - x_i) + \tfrac{1}{2} \left(\sum_{t=1}^{T} \sum_{i=1}^{N} \tfrac{1}{\eta_i} v_i(x^t)^\top (A_i^{t+1})^{-1} v_i(x^t) \right)$$
$$+ \sum_{t=1}^{T} \sum_{i=1}^{N} (x_i^t - x_i)^\top \left(\tfrac{\eta_i}{2} A_i^{t+1} - \tfrac{\eta_i}{2} A_i^t - \tfrac{1}{4} \min\{\tfrac{1}{4GD}, \alpha\} v_i(x^t) v_i(x^t)^\top \right) (x_i^t - x_i).$$

Feasible multi-agent online learning in EC games. We extend Algorithm 37 to multi-agent learning in EC games, showing that our adaptive variant of ONS is not only feasible but achieves a near-optimal convergence rate in terms of a gap function.

By employing a randomization strategy, we obtain an algorithm that is adaptive to exp-concavity parameter and other problem parameters. All the other updates are analogous to Algorithm 37 and the overall multi-agent framework, shown in Algorithm 38, is decentralized. we summarize our main results on the time-average convergence rate of Algorithm 38 using the gap function in Eq. (10.15).

Theorem 10.4.9 *Suppose that a continuous game $\mathcal{G}$ is α-exp-concave and let $G, D > 0$ be problem parameters satisfying $\|v(x)\| \le G$ and $\|x - x'\| \le D$ for all $x, x' \in \mathcal{X}$. If all agents agree to employ Algorithm 38, we have*

$$\mathbb{E}[\mathrm{GAP}(\bar{x}^T)] \le \tfrac{D^2}{2T}(1 + G^2 e^{\frac{(\max\{8GD, 2\alpha^{-1}\})^2}{\log(T+10)}}) + \tfrac{d \log(TG^2+1)}{2T} \sqrt{1 + \log(T+10) + \log(T)\log(T+10)},$$

where $\bar{x}^T = \frac{1}{T}\sum_{t=1}^{T} x^t$ denotes a time-average iterate. Thus, we have $\mathbb{E}[\mathrm{GAP}(\bar{x}^T)] = O(\frac{d \log^2(T)}{T})$.

Directly extending Lemma 10.4.8 to multi-agent learning in EC games, we prove that the iterates $\{x^t\}_{t\geq 1}$ generated by Algorithm 38 satisfy

$$\sum_{t=1}^{T}(x^t-x)^\top v(x) \leq \sum_{i=1}^{N}\tfrac{\eta_i^1}{2}(x_i^1-x_i)^\top A_i^1(x_i^1-x_i)+\tfrac{1}{2}\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\tfrac{1}{\eta_i^{t+1}}v_i(x^t)^\top(A_i^{t+1})^{-1}v_i(x^t)\right)$$
$$+\sum_{t=1}^{T}\sum_{i=1}^{N}(x_i^t-x_i)^\top\left(\tfrac{\eta_i^{t+1}}{2}A_i^{t+1}-\tfrac{\eta_i^t}{2}A_i^t-\tfrac{1}{4}\min\{\tfrac{1}{4GD},\alpha\}v_i(x^t)v_i(x^t)^\top\right)(x_i^t-x_i). \tag{10.17}$$

This inequality is crucial to the proof of Theorem 10.4.9 and its proof can be found later.

10.5 Discussion

Our results open up several directions for further research. First, we have assumed that the gradient feedback is always received at the end of each period. In practice, there may be delays. For instance, in the power-control example, the signal-to-noise ratio sent by the receiver to the transmitter through the feedback channel, from which the gradient can be recovered, is often received with a delay. As such, it is important to understand how delays impact the performance as well as to design the delay-adaptive algorithms that can operate robustly in such environments. Second, our paper has focused on (noisy) gradient feedback while another important yet more challenging type of feedback is bandit feedback, where we only observe (noisy) function values at the chosen action. This problem domain has been less explored than that of learning with (noisy) gradient feedback. For instance, it remains unknown what the minimax optimal regret bound for convex cost functions is; the optimal dependence on T is $\sqrt{T}$ [Bubeck et al., 2021] but the optimal dependence on the dimension d is unknown (note that the OGD regret is dimension-independent if gradient feedback can be observed). Thus, it is desirable to develop the feasible and optimal bandit learning algorithms in both single-agent and multi-agent settings. Can our randomization technique be applicable in that setting and improve the existing algorithms [Bravo et al., 2018, Lin et al., 2021b]? This is a natural question for future work.

10.6 Missing Proofs for Single-Agent Setting

Proof of Proposition 10.2.6. Let $q_0 = 1-p_0$. Fixing a constant $x \geq 1$, we have

$$\mathbb{P}(\bar{X}_n \leq x) = \mathbb{P}(\max\{X_1, X_2, \ldots, X_n\} \leq x) = \mathbb{P}(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x).$$

Since $X_1, X_2, \ldots, X_n$ are n independently distributed random variables, we have

$$\mathbb{P}(\bar{X}_n \leq x) = \mathbb{P}(X_1 \leq x)\mathbb{P}(X_2 \leq x)\cdots\mathbb{P}(X_n \leq x).$$

Since $X_i \sim \mathsf{Geometric}(p_0)$ for all $i = 1, 2, \ldots, n$, we have

$$\mathbb{P}(X_i \le x) = \sum_{k=1}^{\lfloor x \rfloor} \mathbb{P}(X_i = k) = \sum_{k=1}^{\lfloor x \rfloor} (1-p_0)^{k-1} p_0 = p_0 \cdot \tfrac{1-(1-p_0)^{\lfloor x \rfloor}}{p_0} = 1 - q_0^{\lfloor x \rfloor},$$

where $\lfloor x \rfloor$ is the largest integer that is smaller than x. As such, we have $\mathbb{P}(\bar{X}_n \le x) = (1 - q_0^{\lfloor x \rfloor})^n$. Given that $q_0 \in (0,1)$, we have

$$\sum_{n=1}^{+\infty} \mathbb{P}(\bar{X}_n \le x) = \sum_{n=1}^{+\infty} (1 - q_0^{\lfloor x \rfloor})^n = (1 - q_0^{\lfloor x \rfloor}) \cdot \frac{1}{q_0^{\lfloor x \rfloor}} = \frac{1}{q_0^{\lfloor x \rfloor}} - 1 \le \frac{1}{q_0^x} = (1-p_0)^{-x}.$$

Since $1 + x \le e^x$ for all $x \in \mathbb{R}$, we have $1 - p_0 \le e^{-p_0}$. Putting these pieces together yields the first inequality.

Moreover, it follows from the definition of $\bar{X}_n$ that $\bar{X}_n \ge 1$ and hence $\mathbb{E}[\bar{X}_n] \ge 1$. We also have

$$\mathbb{E}[\bar{X}_n] = \sum_{k=0}^{+\infty} \mathbb{P}(\bar{X}_n > k) = \sum_{k=0}^{+\infty} (1 - (1 - q_0^k)^n).$$

By viewing the infinite sum as an Riemann sum approximation of an integral, we obtain

$$\sum_{k=0}^{+\infty} (1 - (1 - q_0^k)^n) \le 1 + \int_0^{+\infty} (1 - (1 - q_0^x)^n)\, dx.$$

We perform the change of variables $u = 1 - q_0^x$ and obtain

$$\begin{aligned}
\int_0^{+\infty} (1 - (1 - q_0^x)^n) dx &= -\tfrac{1}{\log(q_0)} \int_0^1 \tfrac{1-u^n}{1-u}\, du = -\tfrac{1}{\log(q_0)} \int_0^1 (1 + u + \ldots + u^{n-1})\, du \\
&= -\tfrac{1}{\log(q_0)} (1 + \tfrac{1}{2} + \ldots + \tfrac{1}{n}) \le -\tfrac{1}{\log(q_0)} \left(1 + \int_1^n \tfrac{1}{x} dx\right) \\
&= -\tfrac{1+\log(n)}{\log(1-p_0)}.
\end{aligned}$$

Recalling again that $1 + x \le e^x$ for all $x \in \mathbb{R}$, we have $\log(1 - p_0) \le -p_0$, which implies $-\frac{1}{\log(1-p_0)} \le \frac{1}{p_0}$. Putting these pieces together yields the second inequality.

Proof of Lemma 10.2.7. Recall that the update formula of x^{t+1} in Algorithm 33 is

$$x^{t+1} \leftarrow \operatorname*{argmin}_{x \in \mathcal{X}} \{(x - x^t)^\top \xi^t + \tfrac{\eta^{t+1}}{2} \|x - x^t\|^2\}.$$

The first-order optimality condition implies that

$$(x - x^{t+1})^\top \xi^t + \eta^{t+1} (x - x^{t+1})^\top (x^{t+1} - x^t) \ge 0, \quad \text{for all } x \in \mathcal{X}.$$

Equivalently, we have

$$\begin{aligned}\tfrac{\eta^{t+1}}{2}(\|x^{t+1}-x\|^2-\|x^t-x\|^2) &\le (x-x^{t+1})^\top\xi^t-\tfrac{\eta^{t+1}}{2}\|x^{t+1}-x^t\|^2\\ &= (x-x^t)^\top\xi^t+(x^t-x^{t+1})^\top\xi^t-\tfrac{\eta^{t+1}}{2}\|x^{t+1}-x^t\|^2 \le (x-x^t)^\top\xi^t+\tfrac{1}{2\eta^{t+1}}\|\xi^t\|^2.\end{aligned}$$

Since $\mathbb{E}[\xi^t \mid x^t]=\nabla f_t(x^t)$ and $\mathbb{E}[\|\xi^t\|^2 \mid x^t]\le G^2$ for all $t\ge 1$, we have

$$\mathbb{E}\left[\tfrac{\eta^{t+1}}{2}\left(\|x^{t+1}-x\|^2-\|x^t-x\|^2\right) \mid x^t\right] \le (x-x^t)^\top\nabla f_t(x^t)+\mathbb{E}\left[\tfrac{G^2}{2\eta^{t+1}}\right].$$

Since f_t is β-strongly convex, we have

$$\mathbb{E}\left[\tfrac{\eta^{t+1}}{2}\left(\|x^{t+1}-x\|^2-\|x^t-x\|^2\right) \mid x^t\right] \le f_t(x)-f_t(x^t)-\tfrac{\beta}{2}\|x^t-x\|^2+\mathbb{E}\left[\tfrac{G^2}{2\eta^{t+1}}\right],$$

Taking the expectation of both sides and rearranging the resulting inequality yields

$$\mathbb{E}\left[f_t(x^t)+\tfrac{\eta^{t+1}}{2}\|x^{t+1}-x\|^2-\tfrac{\eta^t}{2}\|x^t-x\|^2\right]-f_t(x) \le \mathbb{E}\left[\left(\tfrac{\eta^{t+1}-\eta^t}{2}-\tfrac{\beta}{2}\right)\|x^t-x\|^2+\tfrac{G^2}{2\eta^{t+1}}\right].$$

Summing up the above inequality over $t=1,2,\dots,T$ yields the desired inequality.

10.7 Missing Proofs for Multi-Agent Setting

Proofs for Example 10.3.1 and 10.3.2. We show that the games in Example 10.3.1 and 10.3.2 are β-strongly monotone (cf. Definition 10.3.2) for some $\beta>0$.

Power management in wireless networks. Example 10.3.1 satisfies Definition 10.3.2 with $\beta=\lambda_{\min}(I-\frac{1}{2}(W+W^\top))>0$ where $W_{ii}=0$ for all $1\le i\le N$ and $W_{ij}=\frac{r_i^\star G_{ij}}{G_{ii}}$ for $i\ne j$. An analysis of this setting has been given in Zhou et al. [2021]; we provide the details for completeness. The cost function is given by

$$u_i(a)=\tfrac{1}{2}\left(a_i-\tfrac{r_i^\star(\sum_{j\ne i}G_{ij}a_j+\eta_i)}{G_{ii}}\right)^2.$$

Taking the derivative of $u_i(a)$ with respect to a_i yields that $v_i(a)=a_i-\frac{r_i^\star(\sum_{j\ne i}G_{ij}a_j+\eta_i)}{G_{ii}}$. Consequently, by the definition of W_{ij}, we have

$$\begin{aligned}\langle a'-a, v(a')-v(a)\rangle &= \|a'-a\|^2-\sum_{i=1}^N\tfrac{r_i^\star}{G_{ii}}\sum_{j\ne i}G_{ij}\langle a_i'-a_i, a_j'-a_j\rangle\\ &= \|a'-a\|^2-\sum_{i=1}^N\sum_{j=1}^N W_{ij}\langle a_i'-a_i, a_j'-a_j\rangle\\ &= \|a'-a\|^2-\langle a'-a, W(a'-a)\rangle\\ &= \langle a'-a, (I-\tfrac{1}{2}(W+W^\top))(a'-a)\rangle\\ &\ge \lambda_{\min}(I-\tfrac{1}{2}(W+W^\top))\|a'-a\|^2=\beta\|a'-a\|^2.\end{aligned}$$

This yields the desired result.

Newsvendor with lost sales. Example 10.3.2 satisfies Definition 10.3.2 with $\beta = \alpha p$. Indeed, each players' reward function is given by

$$u_i(x) = (p - c_i) \cdot \mathbb{E}[\max\{0, D_i(x_{-i}) - x_i\}] + c_i \cdot \mathbb{E}[\max\{0, x_i - D_i(x_{-i})\}],$$

Equivalently, we have

$$u_i(x) = (p - c_i) \cdot \int_{x_i}^{+\infty} (k - x_i) \cdot d\mathbb{P}(D_i(x_{-i}) \leq k) + c_i \cdot \int_{-\infty}^{x_i} (x_i - k) \cdot d\mathbb{P}(D_i(x_{-i}) \leq k).$$

Using the Leibniz integral rule, we have

$$v_i(x) = \nabla_{x_i} u_i(x) = p \cdot \mathbb{P}(D_i(x_{-i}) \leq x_i) - p + c_i = p \cdot F_i(x) - p + c_i.$$

Let $F = (F_1, F_2, \ldots, F_N)$ be an operator from $\prod_{i=1}^N [0, \bar{x}_i]$ to $[0,1]^N$. Then, if F is α-strongly monotone, we have

$$\langle x' - x, v(x') - v(x) \rangle = p \cdot \langle x' - x, F(x') - F(x) \rangle \geq \alpha p \|x' - x\|^2, \quad \text{for all } x, x' \in \prod_{i=1}^N [0, \bar{x}_i].$$

Considering the example where the distribution of a demand $D_i(x_{-i})$ is given by

$$\mathbb{P}(D_i(x_{-i}) \leq z) = 1 - \tfrac{1 + \sum_{j \neq i} x_j}{(1 + z + \sum_{j \neq i} x_j)^2}.$$

Then, we have

$$v_i(x) = p \cdot \left(1 - \tfrac{1 + \sum_{j \neq i} x_j}{(1 + \sum_{i=1}^N x_i)^2}\right) - p + c_i.$$

The following proposition, a modification of Rosen [1965, Theorem 6], plays an important role in the subsequent analysis.

Proposition 10.7.1 *Given a continuous game $\mathcal{G} = (\mathcal{N}, \mathcal{X} = \prod_{i=1}^N \mathcal{X}_i, \{u_i\}_{i=1}^N)$, where each u_i is twice continuously differentiable. For each $x \in \mathcal{X}$, we define the Hessian matrix $H(x)$ as follows:*

$$H_{ij}(x) = \tfrac{1}{2} \nabla_j v_i(x) + \tfrac{1}{2} (\nabla_i v_j(x))^\top.$$

If $H(x)$ is positive definite for every $x \in \mathcal{X}$, we have $\langle x' - x, v(x') - v(x) \rangle \geq 0$ for all $x, x' \in \mathcal{X}$ where the equality holds true if and only if $x = x'$.

As a consequence of Proposition 10.7.1, we have $\langle x' - x, v(x) - v(x) \rangle \geq \beta \|x' - x\|^2$ for all $x, x' \in \mathcal{X}$ if $H(x) \succeq \beta I_N$ for all $x \in \mathcal{X}$. For our example, it suffices to show that

$$H(x) \succeq \left(\tfrac{p}{(1 + \sum_{i=1}^N \bar{x}_i)^3}\right) I_N, \quad \text{for all } x \in \prod_{i=1}^N [0, \bar{x}_i]. \tag{10.18}$$

Indeed, after a straightforward calculation, we have

$$\nabla_i v_i(x) = \tfrac{2p(1+\sum_{k\neq i} x_k)}{(1+\sum_{k=1}^N x_k)^3}, \qquad \nabla_j v_i(x) = \tfrac{p(1+\sum_{k\neq i} x_k - x_i)}{(1+\sum_{k=1}^N x_k)^3}, \qquad \nabla_i v_j(x) = \tfrac{p(1+\sum_{k\neq j} x_k - x_j)}{(1+\sum_{k=1}^N x_k)^3}.$$

This implies that

$$H(x) = \tfrac{p}{(1+\sum_{i=1}^N x_i)^3} \cdot \begin{pmatrix} 2+2\sum_{i\neq 1} x_i & 1+\sum_{i\neq 1,2} x_i & 1+\sum_{i\neq 1,3} x_i & \cdots & 1+\sum_{i\neq 1,N} x_i \\ 1+\sum_{i\neq 1,2} x_i & 2+2\sum_{i\neq 2} x_i & 1+\sum_{i\neq 2,3} x_i & \cdots & 1+\sum_{i\neq 2,N} x_i \\ 1+\sum_{i\neq 1,3} x_i & 1+\sum_{i\neq 2,3} x_i & 2+2\sum_{i\neq 3} x_i & \cdots & 1+\sum_{i\neq 3,N} x_i \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1+\sum_{i\neq 1,N} x_i & 1+\sum_{i\neq 2,N} x_i & 1+\sum_{i\neq 3,N} x_i & \cdots & 2+2\sum_{i\neq N} x_i \end{pmatrix}.$$

Equivalently, we have

$$\begin{aligned} & H(x) \\ = \; & \tfrac{p}{(1+\sum_{i=1}^N x_i)^3} \cdot \left\{ \left(1+\sum_{i=1}^N x_i\right) \cdot \begin{pmatrix} 2 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 1 & \cdots & 1 \\ 1 & 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 2 \end{pmatrix} - \begin{pmatrix} 2x_1 & x_1+x_2 & x_1+x_3 & \cdots & x_1+x_N \\ x_1+x_2 & 2x_2 & x_2+x_3 & \cdots & x_2+x_N \\ x_1+x_3 & x_2+x_3 & 2x_3 & \cdots & x_3+x_N \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1+x_N & x_2+x_N & x_3+x_N & \cdots & 2x_N \end{pmatrix} \right\} \\ = \; & \tfrac{p}{(1+\sum_{i=1}^N x_i)^3} \cdot \left\{ \left(1+\sum_{i=1}^N x_i\right) \cdot (I_N + \mathbf{1}_N \mathbf{1}_N^\top) - x\mathbf{1}_N^\top - \mathbf{1}_N x^\top \right\}. \end{aligned}$$

Note that $x \in \prod_{i=1}^N [0, \bar{x}_i]$. If $x = \mathbf{0}_N$, we have $H(x) = p \cdot (I_N + \mathbf{1}_N \mathbf{1}_N^\top)$ and thus satisfy Eq. (10.18). Otherwise, we can let $y = \frac{x}{\sum_{i=1}^N x_k}$ and obtain that $\|y\| \leq 1$. It is clear that $(y - \mathbf{1}_N)(y - \mathbf{1}_N)^\top$ is positive semidefinite. Thus, we have

$$yy^\top + \mathbf{1}_N \mathbf{1}_N^\top \succeq y\mathbf{1}_N^\top + \mathbf{1}_N y^\top.$$

Using the definition of y, we have $yy^\top \preceq I_N$. This together with the above inequality implies that

$$\left(\sum_{i=1}^N x_i\right) \cdot (I_N + \mathbf{1}_N \mathbf{1}_N^\top) \succeq x\mathbf{1}_N^\top + \mathbf{1}_N x^\top.$$

Putting these pieces together yields that

$$H(x) \succeq \tfrac{p}{(1+\sum_{i=1}^N x_i)^3} \cdot (I_N + \mathbf{1}_N \mathbf{1}_N^\top) \succeq \left(\tfrac{p}{(1+\sum_{i=1}^N \bar{x}_i)^3}\right) I_N, \quad \text{for all } x \in \prod_{i=1}^N [0, \bar{x}_i].$$

This completes the proof.

Proof of Lemma 10.3.10. Recall that the update formula of x_i^{t+1} in Algorithm 35 is

$$x_i^{t+1} \leftarrow \operatorname*{argmin}_{x_i \in \mathcal{X}_i} \{(x_i - x_i^t)^\top \xi_i^t + \tfrac{\eta_i^{t+1}}{2}\|x_i - x_i^t\|^2\}.$$

The first-order optimality condition implies that

$$(x_i - x_i^{t+1})^\top \xi_i^t + \eta_i^{t+1}(x_i - x_i^{t+1})^\top (x_i^{t+1} - x_i^t) \geq 0, \quad \text{for all } x_i \in \mathcal{X}_i.$$

Equivalently, we have

$$\begin{aligned} \tfrac{\eta_i^{t+1}}{2}(\|x_i^{t+1} - x_i\|^2 - \|x_i^t - x_i\|^2) &\leq (x_i - x_i^{t+1})^\top \xi_i^t - \tfrac{\eta_i^{t+1}}{2}\|x_i^{t+1} - x_i^t\|^2 \\ &= (x_i - x_i^t)^\top \xi_i^t + (x_i^t - x_i^{t+1})^\top \xi_i^t - \tfrac{\eta_i^{t+1}}{2}\|x_i^{t+1} - x_i^t\|^2 \\ &\leq (x_i - x_i^t)^\top \xi_i^t + \tfrac{1}{2\eta_i^{t+1}}\|\xi_i^t\|^2. \end{aligned}$$

Letting $x = x^\star$ be a unique Nash equilibrium and rearranging the above inequality, we have

$$\tfrac{\eta_i^{t+1}}{2}\|x_i^{t+1} - x_i^\star\|^2 - \tfrac{\eta_i^t}{2}\|x_i^t - x_i^\star\|^2 \leq (x_i^\star - x_i^t)^\top \xi_i^t + \left(\tfrac{\eta_i^{t+1}}{2} - \tfrac{\eta_i^t}{2}\right)\|x_i^t - x_i^\star\|^2 + \tfrac{1}{2\eta_i^{t+1}}\|\xi_i^t\|^2.$$

Summing up the above inequality over $i = 1, 2, \ldots, N$ and rearranging, we have

$$\begin{aligned} &\sum_{i=1}^N \left(\eta_i^{t+1}\|x_i^{t+1} - x_i^\star\|^2 - \eta_i^t\|x_i^t - x_i^\star\|^2\right) \\ &\quad \leq 2(x^\star - x^t)^\top \xi^t + \left(\sum_{i=1}^N (\eta_i^{t+1} - \eta_i^t)\|x_i^t - x_i^\star\|^2\right) + \left(\max_{1 \leq i \leq N}\left\{\tfrac{1}{\eta_i^{t+1}}\right\}\right)\|\xi^t\|^2. \end{aligned}$$

Note that $\{\eta_i^t\}_{1 \leq i \leq N, 1 \leq t \leq T}$ are generated independently of any noisy gradient feedback, we have

$$\begin{aligned} \mathbb{E}[(x^\star - x^t)^\top \xi^t \mid x^t, \{\eta_i^t\}_{1 \leq i \leq N, 1 \leq t \leq T}] &= \mathbb{E}[(x^\star - x^t)^\top v(x^t) \mid \{\eta_i^t\}_{1 \leq i \leq N, 1 \leq t \leq T}], \\ \mathbb{E}[\|\xi^t\|^2 \mid x^t, \{\eta_i^t\}_{1 \leq i \leq N, 1 \leq t \leq T}] &\leq G^2. \end{aligned}$$

Thus, we have

$$\begin{aligned} &\sum_{i=1}^N \left(\eta_i^{t+1}\mathbb{E}\left[\|x_i^{t+1} - x_i^\star\|^2 \mid \{\eta_i^t\}_{1 \leq i \leq N, 1 \leq t \leq T}\right] - \eta_i^t \mathbb{E}\left[\|x_i^t - x_i^\star\|^2 \mid \{\eta_i^t\}_{1 \leq i \leq N, 1 \leq t \leq T}\right]\right) \\ &\leq 2\mathbb{E}\left[(x^\star - x^t)^\top v(x^t) \mid \{\eta_i^t\}_{1 \leq i \leq N, 1 \leq t \leq T}\right] + \sum_{i=1}^N \left(\eta_i^{t+1} - \eta_i^t\right)\mathbb{E}\left[\|x_i^t - x_i^\star\|^2 \mid \{\eta_i^t\}_{1 \leq i \leq N, 1 \leq t \leq T}\right] \\ &\quad + G^2\left(\max_{1 \leq i \leq N}\left\{\tfrac{1}{\eta_i^{t+1}}\right\}\right). \end{aligned}$$

Since a continuous game $\mathcal{G}$ is β-strongly monotone and $x^\star \in \mathcal{X}$ is a unique Nash equilibrium, we have

$$(x^\star - x^t)^\top v(x^t) \le (x^\star - x^t)^\top v(x^\star) - \beta\|x^\star - x^t\|^2 \le -\beta\|x^\star - x^t\|^2.$$

Putting these pieces together yields that

$$\begin{aligned}
&\sum_{i=1}^N \left(\eta_i^{t+1}\mathbb{E}\left[\|x_i^{t+1} - x_i^\star\|^2 \mid \{\eta_i^t\}_{1\le i\le N, 1\le t\le T}\right] - \eta_i^t\mathbb{E}\left[\|x_i^t - x_i^\star\|^2 \mid \{\eta_i^t\}_{1\le i\le N, 1\le t\le T}\right]\right) \\
&\quad\le \sum_{i=1}^N \left(\eta_i^{t+1} - \eta_i^t - 2\beta\right)\mathbb{E}\left[\|x_i^t - x_i^\star\|^2 \mid \{\eta_i^t\}_{1\le i\le N, 1\le t\le T}\right] + G^2\left(\max_{1\le i\le N}\left\{\tfrac{1}{\eta_i^{t+1}}\right\}\right).
\end{aligned}$$

Since $D > 0$ satisfies that $\|x - x'\| \le D$ for all $x, x' \in \mathcal{X}$, we have

$$\begin{aligned}
&\sum_{i=1}^N \left(\eta_i^{t+1}\mathbb{E}\left[\|x_i^{t+1} - x_i^\star\|^2 \mid \{\eta_i^t\}_{1\le i\le N, 1\le t\le T}\right] - \eta_i^t\mathbb{E}\left[\|x_i^t - x_i^\star\|^2 \mid \{\eta_i^t\}_{1\le i\le N, 1\le t\le T}\right]\right) \\
&\quad\le \left(\max_{1\le i\le N}\left\{\eta_i^{t+1} - \eta_i^t\right\} - 2\beta\right)\mathbb{E}\left[\|x^t - x^\star\|^2 \mid \{\eta_i^t\}_{1\le i\le N, 1\le t\le T}\right] + G^2\left(\max_{1\le i\le N}\left\{\tfrac{1}{\eta_i^{t+1}}\right\}\right) \\
&\quad\le D^2\left(\max\left\{0, \max_{1\le i\le N}\left\{\eta_i^{t+1} - \eta_i^t\right\} - 2\beta\right\}\right) + G^2\left(\max_{1\le i\le N}\left\{\tfrac{1}{\eta_i^{t+1}}\right\}\right).
\end{aligned}$$

Summing over $t = 1, 2, \ldots, T-1$ yields the desired inequality.

Proof of Lemma 10.4.4. Recall that the update formula of x^{t+1} in Algorithm 37 is

$$x^{t+1} \leftarrow \operatorname*{argmin}_{x\in\mathcal{X}}\{(x - x^t)^\top \nabla f_t(x^t) + \tfrac{\eta^{t+1}}{2}(x - x^t)^\top A^{t+1}(x - x^t)\}.$$

The first-order optimality condition implies that

$$(x - x^{t+1})^\top \nabla f_t(x^t) + \eta^{t+1}(x - x^{t+1})^\top A^{t+1}(x^{t+1} - x^t) \ge 0, \quad \text{for all } x \in \mathcal{X}.$$

Equivalently, we have

$$\begin{aligned}
&\tfrac{\eta^{t+1}}{2}((x^{t+1} - x)^\top A^{t+1}(x^{t+1} - x) - (x^t - x)^\top A^{t+1}(x^t - x)) \\
&\quad\le (x - x^{t+1})^\top \nabla f_t(x^t) - \tfrac{\eta^{t+1}}{2}(x^{t+1} - x^t)^\top A^{t+1}(x^{t+1} - x^t) \\
&\quad= (x - x^t)^\top \nabla f_t(x^t) + (x^t - x^{t+1})^\top \nabla f_t(x^t) - \tfrac{\eta^{t+1}}{2}(x^{t+1} - x^t)^\top A^{t+1}(x^{t+1} - x^t) \\
&\quad\le (x - x^t)^\top \nabla f_t(x^t) + \tfrac{1}{2\eta^{t+1}}\nabla f_t(x^t)^\top (A^{t+1})^{-1}\nabla f_t(x^t).
\end{aligned}$$

Since f_t is α-exp-concave and satisfies that $\|\nabla f_t(x)\| \le G$ and $\|x - x'\| \le D$ for all $x, x' \in \mathcal{X}$, we derive from Hazan et al. [2007, Lemma 3] that

$$f_t(x) \ge f_t(x^t) + (x - x^t)^\top \nabla f_t(x^t) + \tfrac{1}{4}\min\{\tfrac{1}{4GD}, \alpha\}(x - x^t)^\top (\nabla f_t(x^t)\nabla f_t(x^t)^\top)(x - x^t).$$

For simplicity, we let $\gamma = \frac{1}{4}\min\{\frac{1}{4GD}, \alpha\}$. Putting these pieces together yields that

$$\begin{aligned}\tfrac{\eta^{t+1}}{2}((x^{t+1}-x)^\top A^{t+1}(x^{t+1}-x) - (x^t-x)^\top A^{t+1}(x^t-x)) &\le f_t(x) - f_t(x^t)\\ -\gamma(x-x^t)^\top(\nabla f_t(x^t)\nabla f_t(x^t)^\top)(x-x^t) &+ \tfrac{1}{2\eta^{t+1}}\nabla f_t(x^t)^\top(A^{t+1})^{-1}\nabla f_t(x^t).\end{aligned}$$

Rearranging the above inequality yields that

$$\begin{aligned}&f_t(x^t) - f_t(x) + \tfrac{\eta^{t+1}}{2}(x^{t+1}-x)^\top A^{t+1}(x^{t+1}-x) - \tfrac{\eta^t}{2}(x^t-x)^\top A^t(x^t-x)\\ &\le\ (x-x^t)^\top\left(\tfrac{\eta^{t+1}}{2}A^{t+1} - \tfrac{\eta^t}{2}A^t - \gamma\nabla f_t(x^t)\nabla f_t(x^t)^\top\right)(x-x^t) + \tfrac{1}{2\eta^{t+1}}\nabla f_t(x^t)^\top(A^{t+1})^{-1}\nabla f_t(x^t).\end{aligned}$$

Summing over $t = 1, 2, \ldots, T$ yields the desired inequality.

Proof of Theorem 10.4.2. By the update formula of η^{t+1} in Algorithm 37, we have $\eta^{t+1} = \frac{1}{\sqrt{1+\max\{M^1,\ldots,M^t\}}}$. By the update formula of A^{t+1} in Algorithm 37, we have $A^1 = I_d$ where $I_d \in \mathbb{R}^{d\times d}$ is an identity matrix and $A^{t+1} = A^t + \nabla f_t(x^t)\nabla f_t(x^t)^\top$. Since $\mathcal{X}$ is convex and bounded with a diameter $D > 0$, we have

$$\tfrac{\eta^1}{2}(x^1-x)^\top A^1(x^1-x) \le \tfrac{D^2}{2}, \qquad \eta^{t+1}A^{t+1} - \eta^t A^t \preceq \tfrac{1}{\sqrt{1+\max\{M^1,\ldots,M^t\}}}\nabla f_t(x^t)\nabla f_t(x^t)^\top.$$

By Lemma 10.4.4, we have

$$\begin{aligned}\sum_{t=1}^T f_t(x^t) - \sum_{t=1}^T f_t(x) &\le \tfrac{D^2}{2} + \tfrac{1}{2}\left(\sum_{t=1}^T \tfrac{1}{\eta^{t+1}}\nabla f_t(x^t)^\top(A^{t+1})^{-1}\nabla f_t(x^t)\right) \qquad (10.19)\\ &+ \sum_{t=1}^T\left(\tfrac{1}{2\sqrt{1+\max\{M^1,\ldots,M^t\}}} - \tfrac{1}{4}\min\{\tfrac{1}{4GD}, \alpha\}\right)(x^t-x)^\top\nabla f_t(x^t)\nabla f_t(x^t)^\top(x^t-x).\end{aligned}$$

Since $\|\nabla f_t(x)\| \le G$, $A^1 = I_d$ where $I_d \in \mathbb{R}^{d\times d}$ is an identity matrix and $A^{t+1} = A^t + \nabla f_t(x^t)\nabla f_t(x^t)^\top$, we derive from Hazan et al. [2007, Lemma 11] that

$$\begin{aligned}\sum_{t=1}^T \tfrac{1}{\eta^{t+1}}\nabla f_t(x^t)^\top(A^{t+1})^{-1}\nabla f_t(x^t) &\le \sqrt{1+\max\{M^1,\ldots,M^T\}}\left(\sum_{t=1}^T \nabla f_t(x^t)^\top(A^{t+1})^{-1}\nabla f_t(x^t)\right)\\ &\le \sqrt{1+\max\{M^1,\ldots,M^T\}}(d\log(TG^2+1)). \qquad (10.20)\end{aligned}$$

Plugging Eq. (10.20) into Eq. (10.19) yields that

$$\begin{aligned}\sum_{t=1}^T f_t(x^t) - \sum_{t=1}^T f_t(x) &\le \tfrac{D^2}{2} + \tfrac{d\sqrt{1+\max\{M^1,\ldots,M^T\}}}{2}\log(TG^2+1)\\ &+ \sum_{t=1}^T\left(\tfrac{1}{2\sqrt{1+\max\{M^1,\ldots,M^t\}}} - \tfrac{1}{4}\min\{\tfrac{1}{4GD}, \alpha\}\right)(x^t-x)^\top\nabla f_t(x^t)\nabla f_t(x^t)^\top(x^t-x).\end{aligned}$$

Since $\|\nabla f_t(x)\| \leq G$ and $\mathcal{X}$ is convex and bounded with a diameter $D > 0$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \left(\frac{1}{2\sqrt{1+\max\{M^1,\ldots,M^t\}}} - \frac{1}{4}\min\{\frac{1}{4GD}, \alpha\} \right) (x^t - x)^\top \nabla f_t(x^t) \nabla f_t(x^t)^\top (x^t - x) \\
&\leq \frac{G^2 D^2}{2} \left(\sum_{t=1}^{T} \max\left\{ 0, \frac{1}{\sqrt{1+\max\{M^1,\ldots,M^t\}}} - \frac{1}{2}\min\{\frac{1}{4GD}, \alpha\} \right\} \right).
\end{aligned}$$

Putting these pieces together yields that

$$\begin{aligned}
&\text{Regret}(T) \\
&\leq \frac{D^2}{2} + \frac{G^2 D^2}{2} \left(\sum_{t=1}^{T} \max\left\{ 0, \frac{1}{\sqrt{1+\max\{M^1,\ldots,M^t\}}} - \frac{1}{2}\min\{\frac{1}{4GD}, \alpha\} \right\} \right) + \frac{d\sqrt{1+\max\{M^1,\ldots,M^T\}}}{2} \log(TG^2+1).
\end{aligned}$$

For simplicity, we let $\gamma = \frac{1}{2}\min\{\frac{1}{4GD}, \alpha\}$. Taking the expectation of both sides, we have

$$\begin{aligned}
&\mathbb{E}[\text{Regret}(T)] \\
&\leq \frac{D^2}{2} + \frac{G^2 D^2}{2} \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \max\left\{ 0, \frac{1}{\sqrt{1+\max\{M^1,\ldots,M^t\}}} - \gamma \right\} \right]}_{\mathbf{I}} + \frac{d\log(TG^2+1)}{2} \underbrace{\mathbb{E}\left[\sqrt{1+\max\{M^1,\ldots,M^T\}} \right]}_{\mathbf{II}}.
\end{aligned}$$

It remains to bound the terms **I** and **II** using Proposition 10.2.6. By using the same argument as applied in the proof of Theorem 10.2.4, we have

$$\mathbf{I} \leq e^{\frac{1}{\gamma^2 \log(T+10)}},$$

and

$$\mathbf{II} \leq \sqrt{1 + \log(T+10) + \log(T)\log(T+10)}.$$

Therefore, we conclude that

$$\mathbb{E}[\text{Regret}(T)] \leq \frac{D^2}{2}(1 + e^{\frac{1}{\gamma^2 \log(T+10)}}) + \frac{d\log(TG^2+1)}{2}\sqrt{1 + \log(T+10) + \log(T)\log(T+10)}.$$

This completes the proof.

Proof of Lemma 10.4.8. Recall that the update formula of x_i^{t+1} in either the multi-agent ONS is

$$x_i^{t+1} \leftarrow \operatorname*{argmin}_{x_i \in \mathcal{X}_i} \{(x_i - x_i^t)^\top v_i(x^t) + \tfrac{\eta_i}{2}(x_i - x_i^t)^\top A_i^{t+1}(x_i - x_i^t)\}.$$

The first-order optimality condition implies that

$$(x_i - x_i^{t+1})^\top v_i(x^t) + \eta_i (x_i - x_i^{t+1})^\top A_i^{t+1}(x_i^{t+1} - x_i^t) \geq 0, \quad \text{for all } x_i \in \mathcal{X}_i.$$

Equivalently, we have

$$\begin{aligned}
&\tfrac{\eta_i}{2}((x_i^{t+1}-x_i)^\top A_i^{t+1}(x_i^{t+1}-x_i)-(x_i^t-x_i)^\top A_i^{t+1}(x_i^t-x_i))\\
&\quad\le\ (x_i-x_i^{t+1})^\top v_i(x^t)-\tfrac{\eta_i}{2}(x_i^{t+1}-x_i^t)^\top A_i^{t+1}(x_i^{t+1}-x_i^t)\\
&\quad=\ (x_i-x_i^t)^\top v_i(x^t)+(x_i^t-x_i^{t+1})^\top v_i(x^t)-\tfrac{\eta_i}{2}(x_i^{t+1}-x_i^t)^\top A_i^{t+1}(x_i^{t+1}-x_i^t)\\
&\quad\le\ (x_i-x_i^t)^\top v_i(x^t)+\tfrac{1}{2\eta_i}v_i(x^t)^\top(A_i^{t+1})^{-1}v_i(x^t).
\end{aligned}$$

Rearranging this inequality, we have

$$\begin{aligned}
&\tfrac{\eta_i}{2}(x_i^{t+1}-x_i)^\top A_i^{t+1}(x_i^{t+1}-x_i)-\tfrac{\eta_i}{2}(x_i^t-x_i)^\top A_i^t(x_i^t-x_i)\\
&\quad\le\ (x_i-x_i^t)^\top v_i(x^t)+(x_i^t-x_i)^\top\left(\tfrac{\eta_i}{2}A_i^{t+1}-\tfrac{\eta_i}{2}A_i^t\right)(x_i^t-x_i)+\tfrac{1}{2\eta_i}v_i(x^t)^\top(A_i^{t+1})^{-1}v_i(x^t).
\end{aligned}$$

Summing over $i=1,2,\dots,N$, we have

$$\begin{aligned}
&\sum_{i=1}^N\left(\tfrac{\eta_i}{2}(x_i^{t+1}-x_i)^\top A_i^{t+1}(x_i^{t+1}-x_i)-\tfrac{\eta_i}{2}(x_i^t-x_i)^\top A_i^t(x_i^t-x_i)\right)\\
&\quad\le\ (x-x^t)^\top v(x^t)+\tfrac12\left(\sum_{i=1}^N(x_i^t-x_i)^\top\left(\eta_iA_i^{t+1}-\eta_iA_i^t\right)(x_i^t-x_i)\right)+\tfrac12\left(\sum_{i=1}^N\tfrac{1}{\eta_i}v_i(x^t)^\top(A_i^{t+1})^{-1}v_i(x^t)\right).
\end{aligned}$$

Since $\mathcal{G}$ is α-exp-concave and satisfies that $\|v(x)\|\le G_i$ and $\|x-x'\|\le D_i$ for all $x_i,x_i'\in\mathcal{X}_i$, we have

$$\langle x-x^t,v(x)-v(x^t)\rangle\ge\tfrac14\left(\sum_{i=1}^N\min\{\tfrac{1}{4G_iD_i},\alpha\}(x_i^t-x_i)^\top(v_i(x)v_i(x)^\top+v_i(x^t)v_i(x^t)^\top)(x_i^t-x_i)\right).$$

Putting these pieces together yields that

$$\begin{aligned}
&\sum_{i=1}^N\left(\tfrac{\eta_i}{2}(x_i^{t+1}-x_i)^\top A_i^{t+1}(x_i^{t+1}-x_i)-\tfrac{\eta_i}{2}(x_i^t-x_i)^\top A_i^t(x_i^t-x_i)\right)\\
&\quad\le\ (x-x^t)^\top v(x)+\sum_{i=1}^N(x_i^t-x_i)^\top(\tfrac{\eta_i}{2}A_i^{t+1}-\tfrac{\eta_i}{2}A_i^t-\tfrac14\min\{\tfrac{1}{4GD},\alpha\}v_i(x^t)v_i(x^t)^\top)(x_i^t-x_i)\\
&\qquad+\tfrac12\left(\sum_{i=1}^N\tfrac{1}{\eta_i}v_i(x^t)^\top(A_i^{t+1})^{-1}v_i(x^t)\right).
\end{aligned}$$

Summing over $t=1,2,\dots,T$ yields the desired inequality.

Proof of Theorem 10.4.6. We can see from the multi-agent ONS algorithm that $\eta_i=\frac12\min\{\frac{1}{4GD},\alpha\}$, $A_i^1=I_{d_i}$ where $I_{d_i}\in\mathbb{R}^{d_i\times d_i}$ is an identity matrix, and $A_i^{t+1}=A_i^t+v_i(x^t)v_i(x^t)^\top$. Since $\mathcal{X}$ is convex and bounded with a diameter $D>0$, we have

$$\sum_{i=1}^N\tfrac{\eta_i}{2}(x_i^1-x_i)^\top A_i^1(x_i^1-x_i)\le\tfrac{\alpha D^2}{4},\qquad\eta_iA_i^{t+1}-\eta_iA_i=\tfrac12\min\{\tfrac{1}{4GD},\alpha\}v_i(x^t)v_i(x^t)^\top.$$

By Lemma 10.4.8, we have

$$\sum_{t=1}^{T}(x^t - x)^\top v(x) \le \tfrac{\alpha D^2}{4} + \tfrac{1}{2}\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\tfrac{1}{\eta_i}v_i(x^t)^\top (A_i^{t+1})^{-1}v_i(x^t)\right). \tag{10.21}$$

Since $\|v(x)\| \le G$, $A_i^1 = I_{d_i}$ and $A_i^{t+1} = A_i^t + v_i(x^t)v_i(x^t)^\top$, Hazan et al. [2007, Lemma 11] guarantees that

$$\sum_{t=1}^{T} v_i(x^t)^\top (A_i^{t+1})^{-1}v_i(x^t) \le d_i \log(TG^2+1),$$

which implies that

$$\sum_{t=1}^{T}\sum_{i=1}^{N}\tfrac{1}{\eta_i}v_i(x^t)^\top (A_i^{t+1})^{-1}v_i(x^t) \le \max\left\{8GD, \tfrac{2}{\alpha}\right\}(d\log(TG^2+1)). \tag{10.22}$$

Plugging Eq. (10.22) into Eq. (10.21) yields that

$$\sum_{t=1}^{T}(x^t - x)^\top v(x) \le \tfrac{\alpha D^2}{4} + \max\left\{4GD, \tfrac{1}{\alpha}\right\}(d\log(TG^2+1)).$$

By the definition of $\textsc{gap}(\cdot)$ and $\bar{x}^T$ (i.e., $\bar{x}^T = \frac{1}{T}\sum_{t=1}^{T} x^t$), we have

$$\textsc{gap}(\bar{x}^T) \le \tfrac{\alpha D^2}{4T} + \tfrac{d\log(TG^2+1)}{T}\max\left\{4GD, \tfrac{1}{\alpha}\right\}.$$

This completes the proof.

Proof of Theorem 10.4.9. We can see from Algorithm 38 that $\eta_i^{t+1} = \frac{1}{\sqrt{1+\max\{M_i^1,\dots,M_i^t\}}}$, $A_i^1 = I_{d_i}$ where $I_{d_i} \in \mathbb{R}^{d_i\times d_i}$ is an identity matrix, and $A_i^{t+1} = A_i^t + v_i(x^t)v_i(x^t)^\top$. Since $\mathcal{X}$ is convex and bounded with a diameter $D > 0$, we have

$$\sum_{i=1}^{N}\tfrac{\eta_i^1}{2}(x_i^1 - x_i)^\top A_i^1(x_i^1 - x_i) \le \tfrac{D^2}{2}, \qquad \eta_i^{t+1}A_i^{t+1} - \eta_i^t A_i \preceq \tfrac{1}{\sqrt{1+\max\{M_i^1,\dots,M_i^t\}}}v_i(x^t)v_i(x^t)^\top.$$

We can see from Eq. (10.17) that

$$\begin{aligned}\sum_{t=1}^{T}(x^t - x)^\top v(x) \le \tfrac{D^2}{2} + \tfrac{1}{2}\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\tfrac{1}{\eta_i^{t+1}}v_i(x^t)^\top (A_i^{t+1})^{-1}v_i(x^t)\right) \\ + \sum_{t=1}^{T}\sum_{i=1}^{N}\left(\tfrac{1}{2\sqrt{1+\max\{M_i^1,\dots,M_i^t\}}} - \tfrac{1}{4}\min\{\tfrac{1}{4GD}, \alpha\}\right)(x_i^t - x_i)^\top v_i(x^t)v_i(x^t)^\top (x_i^t - x_i).\end{aligned} \tag{10.23}$$

Since $\|v(x)\| \leq G$, $A_i^1 = I_{d_i}$ and $A_i^{t+1} = A_i^t + v_i(x^t)v_i(x^t)^\top$, Hazan et al. [2007, Lemma 11] guarantees that

$$\begin{aligned}\sum_{t=1}^{T} \tfrac{1}{\eta_i^{t+1}} v_i(x^t)^\top (A_i^{t+1})^{-1} v_i(x^t) &\leq \sqrt{1+\max\{M_i^1,\ldots,M_i^T\}} \left(\sum_{t=1}^{T} v_i(x^t)^\top (A_i^{t+1})^{-1} v_i(x^t)\right) \\ &\leq \sqrt{1+\max\{M_i^1,\ldots,M_i^T\}}(d_i \log(TG^2+1)),\end{aligned}$$

which implies that

$$\sum_{t=1}^{T}\sum_{i=1}^{N} \tfrac{1}{\eta_i^{t+1}} v_i(x^t)^\top (A_i^{t+1})^{-1} v_i(x^t) \leq \sqrt{1+\max_{1\leq i\leq N, 1\leq t\leq T}\{M_i^t\}}(d\log(TG^2+1)). \tag{10.24}$$

Plugging Eq. (10.24) into Eq. (10.23) yields that

$$\begin{aligned}\sum_{t=1}^{T}(x^t - x)^\top v(x) &\leq \tfrac{D^2}{2} + \tfrac{d\log(TG^2+1)}{2}\sqrt{1+\max_{1\leq i\leq N, 1\leq t\leq T}\{M_i^t\}} \\ &\quad + \sum_{t=1}^{T}\sum_{i=1}^{N}\left(\tfrac{1}{2\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}} - \tfrac{1}{4}\min\{\tfrac{1}{4GD},\alpha\}\right)(x_i^t - x_i)^\top v_i(x^t)v_i(x^t)^\top (x_i^t - x_i).\end{aligned}$$

Since $\|v(x)\| \leq G$ and $\mathcal{X}$ is convex and bounded with a diameter $D > 0$, we have

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\tfrac{1}{2\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}} - \tfrac{1}{4}\min\{\tfrac{1}{4GD},\alpha\}\right)(x_i^t - x_i)^\top v_i(x^t)v_i(x^t)^\top (x_i^t - x_i) \\ &\leq \tfrac{G^2D^2}{2}\left(\sum_{t=1}^{T}\max\left\{0, \max_{1\leq i\leq N}\left\{\tfrac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\} - \tfrac{1}{2}\min\{\tfrac{1}{4GD},\alpha\}\right\}\right).\end{aligned}$$

Putting these pieces together yields that

$$\begin{aligned}\sum_{t=1}^{T}(x^t - x)^\top v(x) &\leq \tfrac{D^2}{2} + \tfrac{d\log(TG^2+1)}{2}\sqrt{1+\max_{1\leq i\leq N, 1\leq t\leq T}\{M_i^t\}} \\ &\quad + \tfrac{G^2D^2}{2}\left(\sum_{t=1}^{T}\max\left\{0, \max_{1\leq i\leq N}\left\{\tfrac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\} - \tfrac{1}{2}\min\{\tfrac{1}{4GD},\alpha\}\right\}\right).\end{aligned}$$

For simplicity, we let $\gamma = \frac{1}{2}\min\{\frac{1}{4GD},\alpha\}$. By the definition of $\bar{x}^T$ (i.e., $\bar{x}^T = \frac{1}{T}\sum_{t=1}^{T} x^t$), we have

$$\begin{aligned}&\textsc{gap}(\bar{x}^T) \\ &\leq \tfrac{D^2}{2T} + \tfrac{G^2D^2}{2T}\left(\sum_{t=1}^{T}\max\left\{0, \max_{1\leq i\leq N}\left\{\tfrac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\} - \gamma\right\}\right) + \tfrac{d\log(TG^2+1)}{2T}\sqrt{1+\max_{1\leq i\leq N, 1\leq t\leq T}\{M_i^t\}}.\end{aligned}$$

Taking the expectation of both sides, we have

$$\mathbb{E}[\textsc{gap}(\bar{x}^T)] \leq \frac{D^2}{2T} + \frac{G^2D^2}{2T}\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\max\left\{0, \max_{1\leq i\leq N}\left\{\frac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\}-\gamma\right\}\right]}_{\mathbf{I}} \tag{10.25}$$
$$+\frac{d\log(TG^2+1)}{2T}\underbrace{\mathbb{E}\left[\sqrt{1+\max_{1\leq i\leq N,1\leq t\leq T}\{M_i^t\}}\right]}_{\mathbf{II}}.$$

It remains to bound the terms **I** and **II** using Proposition 10.2.6. Indeed, we have

$$\mathbf{I} = \sum_{t=1}^{T}\mathbb{E}\left[\max\left\{0, \max_{1\leq i\leq N}\left\{\frac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\}-\gamma\right\}\right] \leq \sum_{t=1}^{T}\mathbb{P}\left(\max_{1\leq i\leq N}\left\{\frac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\}-\gamma\geq 0\right).$$

Considering $i^t = \operatorname{argmax}_{1\leq i\leq N}\left\{\frac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\}$ that is a random variable and then recalling that $\{\max\{M_i^1,\ldots,M_i^t\}\}_{1\leq i\leq N}$ are i.i.d., we have $i^t \in \{1,\ldots,N\}$ is uniformly distributed. This implies that

$$\mathbf{I} \leq \frac{1}{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\mathbb{P}\left(\max_{1\leq i\leq N}\left\{\frac{1}{\sqrt{1+\max\{M_i^1,\ldots,M_i^t\}}}\right\}-\gamma\geq 0 \mid i^t=j\right) = \frac{1}{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\mathbb{P}\left(\frac{1}{\sqrt{1+\max\{M_j^1,\ldots,M_j^t\}}}-\gamma\geq 0\right).$$

Since $\{M_i^t\}_{1\leq i\leq N,1\leq t\leq T}$ are i.i.d. geometric random variables with $p_0 = \frac{1}{\log(T+10)}$, Proposition 10.2.6 implies that

$$\sum_{t=1}^{T}\mathbb{P}\left(\frac{1}{\sqrt{1+\max\{M_j^1,\ldots,M_j^t\}}}-\gamma\geq 0\right) \leq e^{\frac{p}{\gamma^2}} = e^{\frac{1}{\gamma^2\log(T+10)}}.$$

Putting these pieces together yields that

$$\mathbf{I} \leq e^{\frac{1}{\gamma^2\log(T+10)}} = e^{\frac{(\max\{8GD,2\alpha^{-1}\})^2}{\log(T+10)}}. \tag{10.26}$$

By using the similar argument with Proposition 10.2.6 and $p_0 = \frac{1}{\log(T+10)}$, we have

$$\mathbf{II} \leq \sqrt{1+\frac{1+\log(NT)}{p_0}} = \sqrt{1+\log(T+10)+\log(NT)\log(T+10)}. \tag{10.27}$$

Plugging Eq. (10.26) and Eq. (10.27) into Eq. (10.25) yields that

$$\mathbb{E}[\textsc{gap}(\bar{x}^T)] \leq \frac{D^2}{2T}(1+G^2e^{\frac{(\max\{8GD,2\alpha^{-1}\})^2}{\log(T+10)}})+\frac{d\log(TG^2+1)}{2T}\sqrt{1+\log(T+10)+\log(T)\log(T+10)}.$$

This completes the proof.

Bibliography

S. S. Abadeh, P. M. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *NIPS*, pages 1576–1584, 2015. (Cited on page 12.)

B. Abbas, H. Attouch, and B. F. Svaiter. Newton-like dynamics and forward-backward methods for structured monotone inclusions in Hilbert spaces. *Journal of Optimization Theory and Applications*, 161(2):331–360, 2014. (Cited on pages 126, 133, 139, 141, 164, 167, 176, 187, and 188.)

J. Abernethy, P. L. Bartlett, A. Rakhlin, and A. Tewari. Optimal strategies and minimax lower bounds for online convex games. In *COLT*, page 414–424. Omnipress, 2008. (Cited on page 333.)

J. Abernethy, K. A. Lai, and A. Wibisono. Last-iterate convergence rates for min-max optimization: Convergence of hamiltonian gradient descent and consensus optimization. In *ALT*, pages 3–47. PMLR, 2021. (Cited on pages 53 and 89.)

B. K. Abid and R. M. Gower. Stochastic algorithms for entropy-regularized optimal transport problems. In *AISTATS*, pages 1505–1512. PMLR, 2018. (Cited on page 255.)

P-A. Absil and S. Hosseini. A collection of nonsmooth Riemannian optimization problems. In *Nonsmooth Optimization and Its Applications*, pages 1–15. Springer, 2019. (Cited on page 93.)

P-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009. (Cited on pages 93 and 95.)

D. Adil, B. Bullins, A. Jambulapati, and S. Sachdeva. Optimal methods for higher-order smooth monotone variational inequalities. *ArXiv Preprint: 2205.06167*, 2022. (Cited on pages 209, 213, 217, 219, 221, and 222.)

S. Adly and H. Attouch. Finite convergence of proximal-gradient inertial algorithms combining dry friction with hessian-driven damping. *SIAM Journal on Optimization*, 30(3): 2134–2162, 2020. (Cited on pages 126 and 188.)

S. Adly and H. Attouch. First-order inertial algorithms involving dry friction damping. *Mathematical Programming*, 193(1):405–445, 2022. (Cited on pages 126, 129, and 188.)

L. Adolphs, H. Daneshmand, A. Lucchi, and T. Hofmann. Local saddle point optimization: A curvature exploitation approach. In *AISTATS*, pages 486–495. PMLR, 2019. (Cited on pages 13, 16, and 89.)

A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, pages 28–40. PMLR, 2010. (Cited on pages 280 and 284.)

A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240, 2013. (Cited on page 284.)

R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, Inc., 1993. (Cited on page 1.)

A. Alacaoglu and Y. Malitsky. Stochastic variance reduction for variational inequality methods. In *COLT*, pages 778–816. PMLR, 2022. (Cited on page 314.)

F. Alimisis, A. Orvieto, G. Bécigneul, and A. Lucchi. A continuous-time perspective for modeling acceleration in Riemannian optimization. In *AISTATS*, pages 1297–1307. PMLR, 2020. (Cited on pages 97, 99, and 108.)

M. S. Alkousa, A. V. Gasnikov, D. M. Dvinskikh, D. A. Kovalev, and F. S. Stonyakin. Accelerated methods for saddle-point problem. *Computational Mathematics and Mathematical Physics*, 60(11):1787–1809, 2020. (Cited on pages 49, 50, 51, and 54.)

Z. Allen-Zhu, Y. Li, R. Oliveira, and A. Wigderson. Much faster algorithms for matrix scaling. In *FOCS*, pages 890–901. IEEE, 2017. (Cited on page 242.)

J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *NIPS*, pages 1964–1974, 2017. (Cited on pages 240, 242, 246, 249, 250, 253, 254, 255, 258, 263, 264, 265, 266, 267, 271, 272, and 273.)

J. Altschuler, F. Bach, A. Rudi, and J. Niles-Weed. Massively scalable Sinkhorn distances via the Nyström method. In *NeurIPS*, pages 4429–4439, 2019. (Cited on page 242.)

F. Alvarez. On the minimizing property of a second order dissipative system in Hilbert spaces. *SIAM Journal on Control and Optimization*, 38(4):1102–1119, 2000. (Cited on pages 126 and 179.)

F. Alvarez and H. Attouch. An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. *Set-Valued Analysis*, 9(1):3–11, 2001. (Cited on pages 126, 164, 166, and 176.)

F. Alvarez and J. M. Pérez C. A dynamical system associated with Newton's method for parametric approximations of convex minimization problems. *Applied Mathematics and Optimization*, 38:193–217, 1998. (Cited on pages 126, 139, 167, and 175.)

F. Alvarez, H. Attouch, J. Bolte, and P. Redont. A second-order gradient-like dissipative dynamical system with Hessian-driven damping: Application to optimization and mechanics. *Journal de mathématiques pures et appliquées*, 81(8):747–779, 2002. (Cited on pages 126, 128, 133, 138, 139, 167, 175, and 176.)

V. S. Amaral, R. Andreani, E. G. Birgin, D. S. Marcondes, and J. M. Martínez. On complexity and convergence of high-order coordinate descent algorithms for smooth nonconvex box-constrained minimization. *Journal of Global Optimization*, 84(3):527–561, 2022. (Cited on page 130.)

A. Anandkumar and R. Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. In *COLT*, pages 81–102. PMLR, 2016. (Cited on page 89.)

A. S. Antipin. Method of convex programming using a symmetric modification of Lagrange function. *Matekon*, 14(2):23–38, 1978. (Cited on page 207.)

A. S. Antipin. Minimization of convex functions on convex sets by means of differential equations. *Differential Equations*, 30(9):1365–1375, 1994. (Cited on page 126.)

K. Antonakopoulos, E. V. Belmega, and P. Mertikopoulos. Online and stochastic optimization beyond Lipschitz continuity: A Riemannian approach. In *ICLR*, 2020. URL https://openreview.net/forum?id=rkxZyaNtwB. (Cited on pages 88 and 89.)

K. Antonakopoulos, V. Belmega, and P. Mertikopoulos. Adaptive extra-gradient methods for min-max optimization and games. In *ICLR*, pages 1–12, 2021. URL https://openreview.net/forum?id=R0a0kFI3dJx. (Cited on pages 312, 314, and 325.)

V. Apidopoulos, J-F. Aujol, and C. Dossal. Convergence rate of inertial Forward-Backward algorithm beyond Nesterov's rule. *Mathematical Programming*, 180(1):137–156, 2020. (Cited on page 175.)

Y. Arjevani, O. Shamir, and R. Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, 178(1):327–360, 2019. (Cited on pages 127, 130, 209, and 223.)

Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, A. Sekhari, and K. Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *COLT*, pages 242–299. PMLR, 2020. (Cited on page 279.)

Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pages 1–50, 2022. (Cited on page 279.)

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223. PMLR, 2017. (Cited on page 241.)

R. D. Armstrong and Z. Jin. A new strongly polynomial dual network simplex algorithm. *Mathematical Programming*, 78:131–148, 1997. (Cited on page 1.)

H. Attouch and J-B. Baillon. Weak versus strong convergence of a regularized Newton dynamic for maximal monotone operators. *Vietnam Journal of Mathematics*, 46(1):177–195, 2018. (Cited on page 188.)

H. Attouch and A. Cabot. Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity. *Journal of Differential Equations*, 263(9):5412–5458, 2017. (Cited on pages 126 and 175.)

H. Attouch and A. Cabot. Convergence of damped inertial dynamics governed by regularized maximally monotone operators. *Journal of Differential Equations*, 264(12):7138–7182, 2018. (Cited on pages 126, 133, 164, 176, and 177.)

H. Attouch and A. Cabot. Convergence of a relaxed inertial proximal algorithm for maximally monotone operators. *Mathematical Programming*, 184(1):243–287, 2020. (Cited on pages 126, 129, 133, 164, 166, 176, 177, and 188.)

H. Attouch and R. Cominetti. A dynamical approach to convex minimization coupling approximation with the steepest descent method. *Journal of Differential Equations*, 128 (2):519–540, 1996. (Cited on page 126.)

H. Attouch and S. C. László. Continuous Newton-like inertial dynamics for monotone inclusions. *Set-Valued and Variational Analysis*, pages 1–27, 2020a. (Cited on pages 126, 139, and 164.)

H. Attouch and S. C. László. Newton-like inertial dynamics and proximal algorithms governed by maximally monotone operators. *SIAM Journal on Optimization*, 30(4):3252–3283, 2020b. (Cited on pages 126, 133, 139, 164, 167, 176, 177, and 188.)

H. Attouch and S. C. László. Continuous Newton-like inertial dynamics for monotone inclusions. *Set-Valued and Variational Analysis*, 29(3):555–581, 2021. (Cited on pages 167, 176, 177, and 188.)

H. Attouch and P-E. Maingé. Asymptotic behavior of second-order dissipative evolution equations combining potential with non-potential effects. *ESAIM: Control, Optimisation and Calculus of Variations*, 17(3):836–857, 2011. (Cited on pages 176 and 187.)

H. Attouch and J. Peypouquet. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than 1/k^2. *SIAM Journal on Optimization*, 26(3): 1824–1834, 2016. (Cited on pages 126, 129, and 175.)

H. Attouch and J. Peypouquet. Convergence rate of proximal inertial algorithms associated with Moreau envelopes of convex functions. In *Splitting Algorithms, Modern Operator*

Theory, and Applications, pages 1–44. Springer, 2019. (Cited on pages 126, 129, 166, 172, 176, 177, and 188.)

H. Attouch and P. Redont. The second-order in time continuous Newton method. In *Approximation, optimization and mathematical economics*, pages 25–36. Springer, 2001. (Cited on pages 126, 139, 167, 176, and 179.)

H. Attouch and B. F. Svaiter. A continuous dynamical Newton-like approach to solving monotone inclusions. *SIAM Journal on Control and Optimization*, 49(2):574–598, 2011. (Cited on pages 126, 133, 139, 141, 147, 164, 167, 173, 174, 176, 187, and 188.)

H. Attouch, X. Goudou, and P. Redont. The heavy ball with friction method, I. the continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. *Communications in Contemporary Mathematics*, 2(01):1–34, 2000. (Cited on pages 126 and 179.)

H. Attouch, P-E. Maingé, and P. Redont. A second-order differential system with Hessian-driven damping: Application to non-elastic shock laws. *Differential Equations & Applications*, 4(1):27–65, 2012. (Cited on pages 126, 128, 133, 139, and 167.)

H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming*, 137(1):91–129, 2013a. (Cited on page 282.)

H. Attouch, P. Redont, and B. F. Svaiter. Global convergence of a closed-loop regularized Newton method for solving monotone inclusions in Hilbert spaces. *Journal of Optimization Theory and Applications*, 157(3):624–650, 2013b. (Cited on pages 126, 127, 133, 139, 140, 141, 148, 164, 167, 173, 176, 177, 187, and 188.)

H. Attouch, M. M. Alves, and B. F. Svaiter. A dynamic approach to a proximal-Newton method for monotone inclusions in Hilbert spaces, with complexity o(1/n^ 2). *Journal of Convex Analysis*, 23(1):139–180, 2016a. (Cited on pages 126, 127, 128, 133, 139, 140, 141, 148, 164, 167, 169, 173, 176, 177, 179, 187, and 188.)

H. Attouch, J. Peypouquet, and P. Redont. Fast convex optimization via inertial dynamics with Hessian driven damping. *Journal of Differential Equations*, 261(10):5734–5783, 2016b. (Cited on pages 126, 127, 128, 133, 138, 141, 148, 175, 176, and 179.)

H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168(1-2): 123–175, 2018. (Cited on pages 126, 129, and 175.)

H. Attouch, Z. Chbani, and H. Riahi. Fast convex optimization via time scaling of damped inertial gradient dynamics. *Pure and Applied Functional Analysis*, To appear, 2019a. (Cited on pages 126, 129, 147, 148, and 153.)

H. Attouch, Z. Chbani, and H. Riahi. Rate of convergence of the nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:2, 2019b. (Cited on pages 129 and 148.)

H. Attouch, Z. Chbani, and H. Riahi. Fast proximal methods via time scaling of damped inertial dynamics. *SIAM Journal on Optimization*, 29(3):2227–2256, 2019c. (Cited on pages 126, 147, 148, 153, 177, and 188.)

H. Attouch, A. Balhag, Z. Chbani, and H. Riahi. Fast convex optimization via inertial dynamics combining viscous and Hessian-driven damping with time rescaling. *Evolution Equations and Control Theory*, 11(2):487–514, 2022a. (Cited on pages 126, 127, 128, 129, 139, 147, 148, 153, 175, 176, 188, and 197.)

H. Attouch, R. I. Bot, and E. R. Csetnek. Fast optimization via inertial dynamics with closed-loop damping. *Journal of the European Mathematical Society*, To appear, 2022b. (Cited on pages 126, 127, 128, 129, 133, 139, 140, 147, 149, 164, and 177.)

H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. Fast convergence of dynamical ADMM via time scaling of damped inertial dynamics. *Journal of Optimization Theory and Applications*, 193(1-3):704–736, 2022c. (Cited on pages 126, 147, and 148.)

H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. First-order optimization algorithms via inertial systems with Hessian driven damping. *Mathematical Programming*, 193(1):113–155, 2022d. (Cited on pages 126, 129, and 175.)

A. Auslender and M. Teboulle. Interior projection-like methods for monotone variational inequalities. *Mathematical programming*, 104(1):39–68, 2005. (Cited on page 53.)

W. Azizian, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *AISTATS*, pages 2863–2873. PMLR, 2020a. (Cited on pages 15 and 89.)

W. Azizian, D. Scieur, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel. Accelerating smooth games by manipulating spectral shapes. In *AISTATS*, pages 1705–1715. PMLR, 2020b. (Cited on page 53.)

M. Bacak. *Convex Analysis and Optimization in Hadamard Spaces*, volume 22. Walter de Gruyter GmbH & Co KG, 2014. (Cited on page 107.)

F. Bach and K. Y. Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *COLT*, pages 164–194. PMLR, 2019. (Cited on page 325.)

M. Baes. Estimate sequence methods: extensions and approximations. *Institute for Operations Research, ETH, Zürich, Switzerland*, 2009. (Cited on pages 130, 161, 207, and 209.)

J. P. Bailey and G. Piliouras. Multiplicative weights update in zero-sum games. In *EC*, pages 321–338, 2018. (Cited on page 16.)

J. B. Baillon. Un exemple concernant le comportement asymptotique de la solution du problème du dt+ $\partial\vartheta$ $(\mu) \in 0$. *Journal of Functional Analysis*, 28(3):369–376, 1978. (Cited on page 188.)

Y. Balaji, R. Chellappa, and S. Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. In *NeurIPS*, pages 12934–12944, 2020. (Cited on pages 241, 242, and 277.)

P. Balamurugan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In *NIPS*, pages 1416–1424, 2016. (Cited on page 314.)

M. Balandat, W. Krichene, C. Tomlin, and A. Bayen. Minimizing regret on reflexive Banach spaces and Nash equilibria in continuous zero-sum games. In *NIPS*, pages 154–162, 2016. (Cited on pages 311 and 313.)

D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel. The mechanics of n-player differentiable games. In *ICML*, pages 354–363. PMLR, 2018. (Cited on pages 16 and 89.)

A. S. Bandeira, N. Boumal, and V. Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. In *COLT*, pages 361–382. PMLR, 2016. (Cited on page 278.)

T. Bárta and E. Fašangová. Convergence to equilibrium for solutions of an abstract wave equation with general damping function. *Journal of Differential Equations*, 260(3):2259–2274, 2016. (Cited on pages 126 and 149.)

T. Bárta, R. Chill, and E. Fašangová. Every ordinary differential equation with a strict Lyapunov function is a gradient system. *Monatshefte für Mathematik*, 166(1):57–72, 2012. (Cited on pages 126 and 149.)

T. Basar and G. J. Olsder. *Dynamic Noncooperative Game Theory*, volume 23. SIAM, 1999. (Cited on pages 12 and 49.)

H. H. Bauschke and P. L. Combettes. A weak-to-strong convergence principle for Fejér-monotone methods in Hilbert spaces. *Mathematics of Operations Research*, 26(2):248–264, 2001. (Cited on page 181.)

H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, volume 408. Springer, 2011. (Cited on pages 94 and 323.)

G. Becigneul and O-E. Ganea. Riemannian adaptive optimization methods. In *ICLR*, 2019. URL https://openreview.net/forum?id=r1eiqi09K7. (Cited on page 93.)

A. Beck and N. Hallak. On the convergence to stationary points of deterministic and randomized feasible descent directions methods. *SIAM Journal on Optimization*, 30(1):56–79, 2020. (Cited on page 294.)

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Science*, 2(1):183–202, 2009. (Cited on pages 59 and 126.)

P. Bégout, J. Bolte, and M. A. Jendoubi. On damped second-order gradient systems. *Journal of Differential Equations*, 259(7):3115–3143, 2015. (Cited on pages 126, 147, and 149.)

R. Bellman. The stability of solutions of linear differential equations. *Duke Mathematical Journal*, 10(4):643–647, 1943. (Cited on page 187.)

R. Bellman. *Dynamic Programming*. Courier Corporation, 2013. (Cited on page 1.)

A. Ben-Tal, L. EL Ghaoui, and A. Nemirovski. *Robust Optimization*, volume 28. Princeton University Press, 2009. (Cited on pages 5 and 95.)

M. Benaım and M. W. Hirsch. Mixed equilibria and dynamical systems arising from fictitious play in perturbed games. *Games and Economic Behavior*, 29(1-2):36–72, 1999. (Cited on page 13.)

M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005. (Cited on pages 293 and 294.)

G. C. Bento, O. P. Ferreira, and J. G. Melo. Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 173(2):548–562, 2017. (Cited on pages 93 and 94.)

J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, 2013. (Cited on page 49.)

R. Bergmann and R. Herzog. Intrinsic formulation of KKT conditions and constraint qualifications on smooth manifolds. *SIAM Journal on Optimization*, 29(4):2423–2444, 2019. (Cited on page 95.)

E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019. (Cited on page 241.)

D. P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973. (Cited on pages 295 and 296.)

D. Bertsimas and J. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997. (Cited on page 241.)

D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011. (Cited on page 5.)

D. Bertsimas, E. Litvinov, X. A. Sun, J. Zhao, and T. Zheng. Adaptive robust optimization for the security constrained unit commitment problem. *IEEE Transactions on Power Systems*, 28(1):52–63, 2012. (Cited on page 5.)

D. Bertsimas, X. Boix, K. V. Carballo, and D. Hertog. A robust optimization approach to deep learning. *ArXiv Preprint: 2112.09279*, 2021. (Cited on page 6.)

M. Betancourt, M. I. Jordan, and A. C. Wilson. On symplectic optimization. *ArXiv Preprint: 1802.03653*, 2018. (Cited on pages 126 and 129.)

A. N. Bhagoji, D. Cullina, and P. Mittal. Lower bounds on adversarial robustness from optimal transport. In *NeurIPS*, pages 7498–7510, 2019. (Cited on page 277.)

S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *NIPS*, pages 3880–3888, 2016. (Cited on pages 2 and 278.)

W. Bian, X. Chen, and Y. Ye. Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization. *Mathematical Programming*, 149(1):301–327, 2015. (Cited on page 292.)

I. Bihari. A generalization of a lemma of Bellman and its application to uniqueness problems of differential equations. *Acta Mathematica Hungarica*, 7(1):81–94, 1956. (Cited on page 146.)

E. G. Birgin, J. L. Gardenghi, J. M. Martinez, S. A. Santos, and P. L. Toint. Evaluation complexity for nonlinear constrained optimization using unscaled KKT conditions and high-order models. *SIAM Journal on Optimization*, 26(2):951–967, 2016. (Cited on pages 130, 161, and 209.)

E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and P. L. Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, 163(1-2):359–368, 2017. (Cited on pages 130, 161, 207, and 209.)

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. (Cited on page 2.)

J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019. (Cited on page 241.)

J. Blanchet, A. Jambulapati, C. Kent, and A. Sidford. Towards optimal running times for optimal transport. *ArXiv Preprint: 1810.07717*, 2018. (Cited on pages 242 and 251.)

J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019. (Cited on page 241.)

D. Bloembergen, K. Tuyls, D. Hennes, and M. Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53:659–697, 2015. (Cited on page 311.)

M. Blondel, V. Seguy, and A. Rolet. Smooth and sparse optimal transport. In *AISTATS*, pages 880–889. PMLR, 2018. (Cited on page 242.)

A. Blum and Y. Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(6):1307–1324, 2007. (Cited on page 310.)

Avrim Blum. Online algorithms in machine learning. In *Online Algorithms*, pages 306–325. Springer, 1998. (Cited on page 309.)

J. Bolte and E. Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188(1):19–51, 2021. (Cited on pages 279, 293, and 294.)

J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007. (Cited on page 282.)

J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010. (Cited on page 149.)

J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018. (Cited on page 294.)

I. M. Bomze, P. Mertikopoulos, W. Schachinger, and M. Staudigl. Hessian barrier algorithms for linearly constrained optimization problems. *SIAM Journal on Optimization*, 29(3): 2100–2127, 2019. (Cited on page 89.)

S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. (Cited on pages 89 and 93.)

N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015. (Cited on page 277.)

J. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer Science & Business Media, 2010. (Cited on page 182.)

T. Bose and F. Meyer. *Digital Signal and Image Processing*. John Wiley & Sons, Inc., 2003. (Cited on page 166.)

R. I. Bot and E. R. Csetnek. Second order forward-backward dynamical systems for monotone inclusion problems. *SIAM Journal on Control and Optimization*, 54(3):1423–1443, 2016. (Cited on pages 126, 133, 164, 176, and 188.)

R. I. Boţ, E. R. Csetnek, and S. C. László. Tikhonov regularization of a second order dynamical system with Hessian driven damping. *Mathematical Programming*, 189:151–186, 2021. (Cited on page 126.)

N. Boumal and P-A. Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *NIPS*, pages 406–414, 2011. (Cited on pages 88 and 96.)

N. Boumal, B. Mishra, P-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(1):1455–1459, 2014. (Cited on page 104.)

N. Boumal, V. Voroninski, and A. S. Bandeira. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *NIPS*, pages 2765–2773, 2016. (Cited on page 278.)

N. Boumal, P-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019. (Cited on page 93.)

M. Bravo, D. Leslie, and P. Mertikopoulos. Bandit learning in concave N-person games. In *NIPS*, pages 5661–5671, 2018. (Cited on pages 324, 336, and 338.)

H. Brézis. *Operateurs Maximaux Monotones: Et semi-groupes de contractions dans les espaces de Hilbert*. Elsevier, 1973. (Cited on page 187.)

H. Brézis. Asymptotic behavior of some evolution systems. In *Nonlinear Evolution Equations*, pages 141–154. Elsevier, 1978. (Cited on page 187.)

L. M. Briceno-Arias and P. L. Combettes. A monotone+ skew splitting model for composite monotone inclusions in duality. *SIAM Journal on Optimization*, 21(4):1230–1250, 2011. (Cited on page 166.)

L. M. Briceno-Arias and D. Davis. Forward-backward-half forward algorithm for solving monotone inclusions. *SIAM Journal on Optimization*, 28(4):2839–2871, 2018. (Cited on page 166.)

L. Brighi and R. John. Characterizations of pseudomonotone maps and economic equilibrium. *Journal of Statistics and Management Systems*, 5(1-3):253–273, 2002. (Cited on page 213.)

L. E. J. Brouwer. Über abbildung von mannigfaltigkeiten. *Mathematische Annalen*, 71(1): 97–115, 1911. (Cited on page 89.)

G. W. Brown. Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation*, 13(1):374–376, 1951. (Cited on page 53.)

R. E. Bruck Jr. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 61(1):159–164, 1977. (Cited on page 183.)

S. Bubeck, Q. Jiang, Y. T. Lee, Y. Li, and A. Sidford. Near-optimal method for highly smooth convex optimization. In *COLT*, pages 492–507. PMLR, 2019. (Cited on pages 125, 127, 129, 130, 149, 150, 156, 161, 162, and 163.)

S. Bubeck, R. Eldan, and Y. T. Lee. Kernel-based methods for bandit convex optimization. *Journal of the ACM (JACM)*, 68(4):1–35, 2021. (Cited on page 338.)

B. Bullins. Highly smooth minimization of nonsmooth problems. In *COLT*, pages 988–1030. PMLR, 2020. (Cited on pages 130 and 209.)

B. Bullins and K. A. Lai. Higher-order methods for convex-concave min-max optimization and monotone variational inequalities. *SIAM Journal on Optimization*, 32(3):2208–2229, 2022. (Cited on pages 130, 165, 168, 177, 178, 183, 189, 194, 197, 205, 207, 219, 220, 225, and 238.)

D. Burago, I. D. Burago, Y. Burago, S. Ivanov, S. V. Ivanov, and S. A. Ivanov. *A Course in Metric Geometry*, volume 33. American Mathematical Soc., 2001. (Cited on pages 92 and 106.)

Y. Burago, M. Gromov, and G. Perel'man. A. D. Alexandrov spaces with curvature bounded below. *Russian Mathematical Surveys*, 47(2):1, 1992. (Cited on page 97.)

J. V. Burke, A. S. Lewis, and M. L. Overton. Approximating subdifferentials by random sampling of gradients. *Mathematics of Operations Research*, 27(3):567–584, 2002a. (Cited on page 293.)

J. V. Burke, A. S. Lewis, and M. L. Overton. Two numerical methods for optimizing matrix stability. *Linear Algebra and its Applications*, 351:117–145, 2002b. (Cited on page 293.)

J. V. Burke, A. S. Lewis, and M. L. Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005. (Cited on page 293.)

J. V. Burke, F. E. Curtis, A. S. Lewis, M. L. Overton, and L. E. A. Simões. Gradient sampling methods for nonsmooth optimization. *Numerical Nonsmooth Optimization: State of the Art Algorithms*, pages 201–225, 2020. (Cited on pages 293 and 294.)

Y. Cai, A. Oikonomou, and W. Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *NeurIPS*, pages 33904–33919, 2022. (Cited on pages 89, 106, 314, and 324.)

E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008. (Cited on page 88.)

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011. (Cited on page 104.)

Y. Carmon and J. Duchi. Gradient descent finds the cubic-regularized nonconvex Newton step. *SIAM Journal on Optimization*, 29(3):2146–2178, 2019. (Cited on page 217.)

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *ICML*, pages 654–663. PMLR, 2017. (Cited on page 279.)

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018. (Cited on page 279.)

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1-2):71–120, 2020. (Cited on pages 63, 64, 223, and 279.)

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points II: first-order methods. *Mathematical Programming*, 185(1-2):315–355, 2021. (Cited on pages 64 and 279.)

Y. Carmon, D. Hausler, A. Jambulapati, Y. Jin, and A. Sidford. Optimal and adaptive Monteiro-Svaiter acceleration. In *NeurIPS*, 2022. URL https://openreview.net/forum?id=n3lr7GdcbyD. (Cited on page 209.)

M. Carrière, M. Cuturi, and S. Oudot. Sliced Wasserstein kernel for persistence diagrams. In *ICML*, pages 1–10. PMLR, 2017. (Cited on page 241.)

C. Cartis, N. I. M. Gould, and P. L. Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010. (Cited on page 209.)

C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011a. (Cited on page 209.)

C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function-and derivative-evaluation complexity. *Mathematical Programming*, 130(2):295–319, 2011b. (Cited on page 209.)

C. Cartis, N. I. M. Gould, and P. L. Toint. Second-order optimality and beyond: Characterization and evaluation complexity in convexly constrained nonlinear optimization. *Foundations of Computational Mathematics*, 18(5):1073–1107, 2018. (Cited on page 130.)

C. Cartis, N. I. Gould, and P. L. Toint. Universal regularization methods: Varying the power, the smoothness and the accuracy. *SIAM Journal on Optimization*, 29(1):595–615, 2019. (Cited on pages 130 and 209.)

C. Cartis, N. I. M. Gould, and P. L. Toint. *Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives*. SIAM, 2022. (Cited on page 217.)

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. (Cited on pages 6, 12, 50, 166, 206, 311, and 324.)

D. Chakrabarty and S. Khanna. Better and simpler error analysis of the Sinkhorn-Knopp algorithm for matrix scaling. *Mathematical Programming*, 188(1):395–407, 2021. (Cited on page 241.)

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011. (Cited on page 53.)

A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016. (Cited on pages 50 and 54.)

T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *NeurIPS*, pages 393–403, 2019. (Cited on page 53.)

J. Cheeger and D. G. Ebin. *Comparison Theorems in Riemannian Geometry*, volume 9. North-Holland Amsterdam, 1975. (Cited on page 97.)

G. H. G. Chen and R. T. Rockafellar. Convergence rates in forward-backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997. (Cited on page 13.)

N. Chen, A. Klushyn, R. Kurle, X. Jiang, J. Bayer, and P. Smagt. Metrics for deep generative models. In *AISTATS*, pages 1540–1550. PMLR, 2018. (Cited on page 89.)

S. Chen, S. Ma, A. M-C. So, and T. Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM Journal on Optimization*, 30(1):210–239, 2020. (Cited on page 93.)

X. Chen, X. Deng, and S-H. Teng. Settling the complexity of computing two-player Nash equilibria. *Journal of the ACM (JACM)*, 56(3):1–57, 2009. (Cited on page 5.)

X. Chen, S. Liu, K. Xu, X. Li, X. Lin, M. Hong, and D. Cox. ZO-AdaMM: zeroth-order adaptive momentum method for black-box optimization. In *NeurIPS*, pages 7204–7215, 2019. (Cited on page 280.)

Y. Chen, G. Lan, and Y. Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014. (Cited on pages 50 and 53.)

Y. Chen, G. Lan, and Y. Ouyang. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165(1):113–149, 2017. (Cited on pages 53, 206, 211, and 314.)

L. Chergui. Convergence of global and bounded solutions of a second order gradient like system with nonlinear dissipation and analytic nonlinearity. *Journal of Dynamics and Differential Equations*, 3(20):643–652, 2008. (Cited on pages 126 and 149.)

A. Cherukuri, B. Gharesifard, and J. Cortes. Saddle-point dynamics: conditions for asymptotic stability of saddle points. *SIAM Journal on Control and Optimization*, 55(1):486–511, 2017. (Cited on pages 13 and 15.)

R. Chill and E. Fašangová. Gradient systems. In *Lecture Notes of the 13th International Internet Seminar, Matfyzpress, Prague*, 2010. (Cited on pages 126 and 149.)

L. Chizat, P. Roussillon, F. Léger, F-X. Vialard, and G. Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. In *NeurIPS*, pages 2257–2269, 2020. (Cited on pages 242 and 245.)

S. C. Choi, W. S. DeSarbo, and P. T. Harker. Product positioning under price competition. *Management Science*, 36(2):175–199, 1990. (Cited on page 213.)

A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *AISTATS*, pages 192–204. PMLR, 2015. (Cited on pages 2 and 278.)

K. L. Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, pages 463–483, 1954. (Cited on page 120.)

F. H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, 1990. (Cited on pages 279 and 282.)

E. A. Coddington and N. Levinson. *Theory of Ordinary Differential Equations*. Tata McGraw-Hill Education, 1955. (Cited on pages 135 and 171.)

M. B. Cohen, A. Madry, D. Tsipras, and A. Vladu. Matrix scaling and balancing via box constrained Newton's method and interior point methods. In *FOCS*, pages 902–913. IEEE, 2017. (Cited on page 242.)

P. L. Combettes. Systems of structured monotone inclusions: duality, algorithms, and applications. *SIAM Journal on Optimization*, 23(4):2420–2447, 2013. (Cited on page 166.)

P. L. Combettes and J. Eckstein. Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions. *Mathematical Programming*, 168(1):645–672, 2018. (Cited on page 166.)

M. Coste. *An Introduction to o-Minimal Geometry*. Istituti Editoriali E Poligrafici Internazionali Pisa, 2000. (Cited on page 282.)

R. Cottle, F. Giannessi, and J-L. Lions. *Variational Inequalities and Complementarity Problems: Theory and Applications*. John Wiley & Sons, 1980. (Cited on page 206.)

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9): 1853–1865, 2017. (Cited on pages 241 and 242.)

T. M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, 1991. (Cited on pages 333 and 336.)

T. M. Cover. *Elements of Information Theory*. John Wiley & Sons, 1999. (Cited on page 333.)

C. Criscitiello and N. Boumal. Efficiently escaping saddle points on manifolds. In *NeurIPS*, pages 5987–5997, 2019. (Cited on pages 89 and 93.)

C. Criscitiello and N. Boumal. An accelerated first-order method for non-convex optimization on manifolds. *Foundations of Computational Mathematics*, pages 1–77, 2022. (Cited on page 104.)

E. R. Csetnek, A. Eberhard, and M. K. Tam. Convergence rates for boundedly regular systems. *Advances in Computational Mathematics*, 47(5):1–18, 2021. (Cited on page 184.)

S. Cui, U. Shanbhag, M. Staudigl, and P. Vuong. Stochastic relaxed inertial forward-backward-forward splitting for monotone inclusions in Hilbert spaces. *Computational Optimization and Applications*, 83(2):465–524, 2022. (Cited on page 182.)

A. Cutkosky. Better full-matrix regret via parameter-free online learning. In *NeurIPS*, pages 8836–8846, 2020a. (Cited on page 315.)

A. Cutkosky. Parameter-free, dynamic, and strongly-adaptive online learning. In *ICML*, pages 2250–2259. PMLR, 2020b. (Cited on page 315.)

A. Cutkosky and F. Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *COLT*, pages 1493–1529. PMLR, 2018. (Cited on page 315.)

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, pages 2292–2300, 2013. (Cited on pages 241, 242, 245, 246, 249, 266, and 273.)

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *ICML*, pages 685–693, 2014. (Cited on page 241.)

M. Cuturi and G. Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016. (Cited on page 242.)

M. Cuturi, O. Teboul, and J-P. Vert. Differentiable ranks and sorting using optimal transport. In *NeurIPS*, pages 6861–6871, 2019. (Cited on page 242.)

C. D. Dang and G. Lan. On the convergence properties of non-Euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and Applications*, 60(2):277–310, 2015. (Cited on pages 212 and 221.)

A. Daniilidis and D. Drusvyatskiy. Pathological subgradient dynamics. *SIAM Journal on Optimization*, 30(2):1327–1338, 2020. (Cited on pages 279, 293, and 294.)

G. B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, 1998. (Cited on page 53.)

C. Daskalakis and I. Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *ArXiv Preprint: 1807.04252*, 2018a. (Cited on pages 16 and 19.)

C. Daskalakis and I. Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *NIPS*, pages 9236–9246, 2018b. (Cited on pages 16 and 89.)

C. Daskalakis and I. Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *ITCS*, 2019. (Cited on page 89.)

C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou. The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, 39(1):195, 2009. (Cited on page 5.)

C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *ICLR*, 2018. URL https://openreview.net/forum?id=SJJySbbAZ. (Cited on pages 13, 19, and 53.)

C. Daskalakis, S. Skoulakis, and M. Zampetakis. The complexity of constrained min-max optimization. In *STOC*, pages 1466–1478, 2021. (Cited on pages 5, 212, and 217.)

D. Davis. Convergence rate analysis of primal-dual splitting schemes. *SIAM Journal on Optimization*, 25(3):1912–1943, 2015. (Cited on page 166.)

D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019. (Cited on pages 17, 23, 26, 65, 81, 86, and 294.)

D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1):119–154, 2020. (Cited on pages 279, 282, 293, and 294.)

D. Davis, D. Drusvyatskiy, Y. T. Lee, S. Padmanabhan, and G. Ye. A gradient sampling method with complexity guarantees for Lipschitz functions in high and low dimensions. In *NeurIPS*, pages 6692–6703, 2022. (Cited on pages 279, 283, 286, 288, and 294.)

G. Debreu. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences*, 38(10):886–893, 1952. (Cited on page 323.)

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pages 1646–1654, 2014. (Cited on page 313.)

M. C. Delfour. *Introduction to Optimization and Hadamard Semidifferential Calculus*. SIAM, 2019. (Cited on page 282.)

J. Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *COLT*, pages 1428–1451. PMLR, 2020. (Cited on pages 89, 167, 213, and 223.)

J. Diakonikolas and M. I. Jordan. Generalized momentum-based methods: A Hamiltonian perspective. *SIAM Journal on Optimization*, 31(1):915–944, 2021. (Cited on pages 126, 164, and 197.)

J. Diakonikolas and L. Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 29(1):660–689, 2019. (Cited on pages 126, 129, and 175.)

J. Diakonikolas, C. Daskalakis, and M. I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *AISTATS*, pages 2746–2754. PMLR, 2021. (Cited on page 212.)

N. Doikov and Y. Nesterov. Local convergence of tensor methods. *Mathematical Programming*, 193(1):315–336, 2022. (Cited on pages 130 and 209.)

Y. Dong, Y. Gao, R. Peng, I. Razenshteyn, and S. Sawlani. A study of performance of optimal transport. *ArXiv Preprint: 2005.01182*, 2020. (Cited on pages 242, 243, and 273.)

D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018. (Cited on pages 28 and 183.)

D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1):503–558, 2019. (Cited on page 294.)

D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. Nonsmooth optimization using taylor-like models: error bounds, convergence, and termination criteria. *Mathematical Programming*, 185(1):357–383, 2021. (Cited on page 183.)

S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, B. Póczos, and A. Singh. Gradient descent can take exponential time to escape saddle points. In *NIPS*, pages 1067–1077, 2017. (Cited on page 89.)

Simon S Du and Wei Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *AISTATS*, pages 196–205. PMLR, 2019. (Cited on page 13.)

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011. (Cited on pages 291, 313, and 324.)

J. C. Duchi and F. Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018. (Cited on page 294.)

J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012. (Cited on pages 284 and 296.)

J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015. (Cited on pages 280, 281, 284, 286, 287, and 288.)

R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969. (Cited on pages 241, 245, and 277.)

D. Duffie. *Dynamic Asset Pricing Theory*. Princeton University Press, 2010. (Cited on page 279.)

P. Dvurechenskii, D. Dvinskikh, A. Gasnikov, C. Uribe, and A. Nedich. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In *NIPS*, pages 10783–10793, 2018. (Cited on page 241.)

P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *ICML*, pages 1367–1376. PMLR, 2018. (Cited on pages 240, 242, 243, 246, 249, 250, 251, 253, 254, 255, 257, 263, 264, 265, 266, and 273.)

J. Eckstein and B. F. Svaiter. General projective splitting methods for sums of maximal monotone operators. *SIAM Journal on Control and Optimization*, 48(2):787–811, 2009. (Cited on page 166.)

E. Esser, X. Zhang, and T. F. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010. (Cited on page 54.)

C. Ewerhart. Cournot games with biconcave demand. *Games and Economic Behavior*, 85: 37–47, 2014. (Cited on page 213.)

F. Facchinei and J-S. Pang. *Finite-dimensional Variational Inequalities and Complementarity Problems*. Springer Science & Business Media, 2007. (Cited on pages 5, 49, 94, 166, 167, 206, 211, 228, 323, and 335.)

A. Fallah, A. Ozdaglar, and S. Pattathil. An optimal multistage stochastic gradient method for minimax problems. In *CDC*, pages 3573–3579. IEEE, 2020. (Cited on page 103.)

R. Z. Farahani and M. Hekmatfar. *Facility Location: Concepts, Models, Algorithms and Case Studies*. Springer Science & Business Media, 2009. (Cited on page 166.)

M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado. Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. *SIAM Journal on Optimization*, 28(3):2654–2689, 2018. (Cited on page 126.)

J. Fearnley, P. W. Goldberg, A. Hollender, and R. Savani. The complexity of gradient descent: CLS = PPAD∩ PLS. In *STOC*, pages 46–59, 2021. (Cited on page 89.)

O. Fercoq and Z. Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *IMA Journal of Numerical Analysis*, 39(4):2069–2095, 2019. (Cited on page 219.)

Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015. (Cited on page 266.)

E. R. Fernholz. *Stochastic Portfolio Theory*, volume 48. Springer Science & Business Media, 2002. (Cited on pages 333 and 336.)

O. P. Ferreira and P. R. Oliveira. Subgradient algorithm on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 97(1):93–104, 1998. (Cited on page 93.)

O. P. Ferreira and P. R. Oliveira. Proximal point algorithm on Riemannian manifolds. *Optimization*, 51(2):257–270, 2002. (Cited on page 93.)

O. P. Ferreira, L. R. Pérez, and S. Z. Németh. Singularities of monotone vector fields and an extragradient-type algorithm. *Journal of Global Optimization*, 31(1):133–151, 2005. (Cited on page 94.)

J. Feydy, T. Séjourné, F-X. Vialard, S-I. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *AISTATS*, pages 2681–2690. PMLR, 2019. (Cited on page 242.)

R. Flamary and N. Courty. POT: Python optimal transport library, 2017. URL `https://github.com/rflamary/POT`. (Cited on page 243.)

A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *SODA*, pages 385–394. SIAM, 2005. (Cited on pages 280, 284, and 316.)

P. T. Fletcher and S. Joshi. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87(2):250–262, 2007. (Cited on page 95.)

M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM Journal on Optimization*, 21(4):1614–1640, 2011. (Cited on page 88.)

D. J. Foster, A. Rakhlin, and K. Sridharan. Adaptive online learning. In *NIPS*, pages 3375–3383, 2015. (Cited on page 315.)

D. J. Foster, S. Kale, M. Mohri, and K. Sridharan. Parameter-free online learning via model selection. In *NIPS*, pages 6022–6032, 2017. (Cited on page 315.)

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015. (Cited on pages 241, 245, and 277.)

G. França, J. Sulam, D. P. Robinson, and R. Vidal. Conformal symplectic and relativistic optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124008, 2020. (Cited on pages 126 and 129.)

G. França, M. I. Jordan, and R. Vidal. On dissipative symplectic integration with applications to gradient-based optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(4):043402, 2021. (Cited on pages 126, 129, 164, and 197.)

R. M. Freund and P. Grigas. New analysis and results for the Frank-Wolfe method. *Mathematical Programming*, 155(1-2):199–230, 2016. (Cited on page 1.)

R. M. Freund and H. Lu. New computational guarantees for solving convex optimization problems with first order methods, via a function growth condition measure. *Mathematical Programming*, 170(2):445–477, 2018. (Cited on page 218.)

R. M. Freund, P. Grigas, and R. Mazumder. An extended frank-wolfe method with "in-face" directions, and its application to low-rank matrix completion. *SIAM Journal on Optimization*, 27(1):319–346, 2017. (Cited on page 1.)

A. Fuduli, M. Gaudioso, and G. Giallombardo. Minimizing nonconvex nonsmooth functions via cutting planes and proximity control. *SIAM Journal on Optimization*, 14(3):743–756, 2004. (Cited on pages 293 and 294.)

M. Fukushima. Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems. *Mathematical Programming*, 53:99–110, 1992. (Cited on page 206.)

H. N. Gabow and R. E. Tarjan. Faster scaling algorithms for general graph matching problems. *Journal of the ACM (JACM)*, 38(4):815–853, 1991. (Cited on page 241.)

G. Gallego and M. Hu. Dynamic pricing of perishable assets under competition. *Management Science*, 60(5):1241–1259, 2014. (Cited on page 213.)

B. Gao, X. Liu, X. Chen, and Y. Yuan. A new first-order algorithmic framework for optimization problems with orthogonality constraints. *SIAM Journal on Optimization*, 28(1): 302–332, 2018. (Cited on page 93.)

A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, and C. A. Uribe. Optimal tensor methods in smooth convex and uniformly convex optimization. In *COLT*, pages 1374–1391. PMLR, 2019a. (Cited on pages 125, 127, 129, 130, 149, 150, 161, 162, and 163.)

A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, C. A. Uribe, B. Jiang, H. Wang, S. Zhang, S. Bubeck, Q. Jiang, Y. T. Lee, Y. Li, and A. Sidford. Near optimal methods for minimizing convex functions with Lipschitz p-th derivatives. In *COLT*, pages 1392–1393. PMLR, 2019b. (Cited on pages 176, 178, 197, and 209.)

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *COLT*, pages 797–842. PMLR, 2015. (Cited on pages 89 and 278.)

R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *NIPS*, pages 2981–2989, 2016. (Cited on pages 2 and 278.)

R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *ICML*, pages 1233–1242. PMLR, 2017a. (Cited on page 89.)

R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *ICML*, pages 1233–1242. PMLR, 2017b. (Cited on page 278.)

A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *NIPS*, pages 3440–3448, 2016. (Cited on pages 242 and 273.)

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. In *AISTATS*, pages 1574–1583. PMLR, 2019. (Cited on pages 242 and 245.)

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013a. (Cited on page 218.)

S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013b. (Cited on pages 279, 280, 281, 284, 286, 287, 288, and 299.)

S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016. (Cited on page 279.)

A. Giannou, E. V. Vlatakis-Gkaragkounis, and P. Mertikopoulos. On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. In *NeurIPS*, pages 22655–22666, 2021a. (Cited on page 88.)

A. Giannou, E. V. Vlatakis-Gkaragkounis, and P. Mertikopoulos. Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information. In *COLT*, pages 2147–2148. PMLR, 2021b. (Cited on page 88.)

G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR*, 2019a. (Cited on pages 50, 51, and 54.)

G. Gidel, R. A. Hemmat, M. Pezeshki, R. Le Priol, G. Huang, S. Lacoste-Julien, and I. Mitliagkas. Negative momentum for improved game dynamics. In *AISTATS*, pages 1802–1811. PMLR, 2019b. (Cited on page 89.)

P. Giselsson and S. Boyd. Monotonicity and restart in fast gradient methods. In *CDC*, pages 5058–5063. IEEE, 2014. (Cited on page 219.)

K. Goebel and W. A. Kirk. *Topics in Metric Fixed Point Theory*. Cambridge University Press, 1990. (Cited on page 180.)

Donald Goldfarb and Jianxiu Hao. Polynomial-time primal simplex algorithms for the minimum cost network flow problem. *Algorithmica*, 8(1-6):145–160, 1992. (Cited on page 1.)

A. Goldsmith. *Wireless Communications*. Cambridge University Press, 2005. (Cited on pages 329 and 330.)

A. Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13(1):14–22, 1977. (Cited on pages 278, 281, 283, and 294.)

T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014. (Cited on page 54.)

N. Golowich, S. Pattathil, and C. Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *NeurIPS*, pages 20766–20778, 2020a. (Cited on pages 183, 314, and 324.)

N. Golowich, S. Pattathil, C. Daskalakis, and A. Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *COLT*, pages 1758–1784. PMLR, 2020b. (Cited on page 89.)

E. G. Golshtein. Generalized gradient method for finding saddle points. *Matekon*, 10(3): 36–52, 1974. (Cited on page 15.)

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. (Cited on pages 5, 12, 16, 50, 88, 166, and 206.)

N. I. M Gould, S. Lucidi, M. Roma, and P. L. Toint. Solving the trust-region subproblem using the Lanczos method. *SIAM Journal on Optimization*, 9(2):504–525, 1999. (Cited on page 217.)

N. I. M. Gould, D. P. Robinson, and H. S. Thorne. On solving trust-region and other regularised subproblems in optimization. *Mathematical Programming Computation*, 2(1): 21–57, 2010. (Cited on page 217.)

A. Granas and J. Dugundji. *Fixed Point Theory*. Springer Science & Business Media, 2013. (Cited on pages 135 and 136.)

G. N. Grapiglia and Y. Nesterov. Regularized Newton methods for minimizing functions with Hölder continuous Hessians. *SIAM Journal on Optimization*, 27(1):478–506, 2017. (Cited on pages 130 and 209.)

G. N. Grapiglia and Y. Nesterov. Accelerated regularized Newton methods for minimizing composite convex functions. *SIAM Journal on Optimization*, 29(1):77–99, 2019. (Cited on pages 130 and 209.)

G. N. Grapiglia and Y. Nesterov. Tensor methods for minimizing convex functions with Hölder continuous higher-order derivatives. *SIAM Journal on Optimization*, 30(4):2750–2779, 2020. (Cited on pages 130 and 209.)

G. N. Grapiglia and Y. Nesterov. On inexact solution of auxiliary problems in tensor methods for convex optimization. *Optimization Methods and Software*, 36(1):145–170, 2021. (Cited on page 217.)

G. N. Grapiglia and Y. Nesterov. Adaptive third-order methods for composite convex optimization. *ArXiv Preprint: 2202.12730*, 2022a. (Cited on page 209.)

G. N. Grapiglia and Y. Nesterov. Tensor methods for finding approximate stationary points of convex functions. *Optimization Methods and Software*, 37(2):605–638, 2022b. (Cited on pages 130 and 144.)

P. Grnarova, K. Y. Levy, A. Lucchi, T. Hofmann, and A. Krause. An online learning approach to generative adversarial networks. In *ICLR*, 2018. (Cited on pages 15 and 55.)

T. H. Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, pages 292–296, 1919. (Cited on page 187.)

O. Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, 1991. (Cited on pages 166 and 181.)

O. Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992. (Cited on pages 126, 153, 161, 166, and 175.)

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *NIPS*, pages 5767–5777, 2017. (Cited on page 241.)

S. Guminov, P. Dvurechensky, N. Tupitsa, and A. Gasnikov. On a combination of alternating minimization and Nesterov's momentum. In *ICML*, pages 3886–3898. PMLR, 2021. (Cited on pages 240, 242, 246, 257, and 266.)

W. Guo, N. Ho, and M. Jordan. Fast algorithms for computational optimal transport and Wasserstein barycenter. In *AISTATS*, pages 2088–2097. PMLR, 2020. (Cited on pages 242 and 257.)

E. Y. Hamedani and N. S. Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021. (Cited on pages 50, 51, 54, and 89.)

J. H. Hammond and T. L. Magnanti. Generalized descent methods for asymmetric systems of equations. *Mathematics of Operations Research*, 12(4):678–699, 1987. (Cited on page 206.)

A. Han, B. Mishra, P. K. Jawanpuria, and J. Gao. On Riemannian optimization over positive definite matrices with the Bures-Wasserstein geometry. In *NeurIPS*, pages 8940–8953, 2021. (Cited on page 88.)

A. Han, B. Mishra, P. Jawanpuria, P. Kumar, and J. Gao. Riemannian Hamiltonian methods for min-max optimization on manifolds. *ArXiv Preprint: 2204.11418*, 2022. (Cited on pages 91, 94, and 105.)

M. Harandi, M. Salzmann, and R. Hartley. Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):48–62, 2017. (Cited on page 104.)

P. T. Harker and J-S. Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications. *Mathematical Programming*, 48(1):161–220, 1990. (Cited on pages 206, 212, 216, and 228.)

P. Hartman and G. Stampacchia. On some non-linear elliptic differential-functional equations. *Acta Mathematica*, 115:271–310, 1966. (Cited on page 206.)

T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009. (Cited on page 2.)

E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016. (Cited on pages 309, 310, and 315.)

E. Hazan and S. Kakade. Revisiting the Polyak step size. *ArXiv Preprint: 1905.00313*, 2019. (Cited on page 313.)

E. Hazan and S. Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(1): 2489–2512, 2014. (Cited on pages 316, 318, and 333.)

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007. (Cited on pages 310, 312, 316, 333, 335, 344, 345, 348, and 349.)

E. Hazan, T. Koren, and K. Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *COLT*, pages 197–209. PMLR, 2014. (Cited on page 333.)

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. (Cited on page 292.)

Y. He and R. D. C. Monteiro. An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems. *SIAM Journal on Optimization*, 26(1):29–56, 2016. (Cited on page 53.)

E. D. Helly. Über Mengen konvexer körper mit gemeinschaftlichen Punkte. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 32:175–176, 1923. (Cited on page 90.)

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NIPS*, pages 6626–6637, 2017. (Cited on pages 13, 16, 19, and 89.)

N. Ho, V. Huynh, D. Phung, and M. I. Jordan. Probabilistic multilevel clustering via composite transportation distance. In *AISTATS*, pages 3149–3157. PMLR, 2019. (Cited on page 241.)

C. H. Hommes and M. I. Ochea. Multiple equilibria and limit cycles in evolutionary games with Logit dynamics. *Games and Economic Behavior*, 74(1):434–441, 2012. (Cited on page 13.)

L. J. Hong, B. L. Nelson, and J. Xu. Discrete optimization via simulation. In *Handbook of Simulation Optimization*, pages 9–44. Springer, 2015. (Cited on page 279.)

R. Hosseini and S. Sra. Matrix manifold optimization for Gaussian mixtures. In *NIPS*, pages 910–918, 2015. (Cited on page 89.)

Y-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS*, pages 6936–6946, 2019. (Cited on page 53.)

Y-G. Hsieh, K. Antonakopoulos, and P. Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to Nash equilibrium. In *COLT*, pages 2388–2422. PMLR, 2021. (Cited on pages 312, 314, and 325.)

B. Hu and L. Lessard. Dissipativity theory for Nesterov's accelerated method. In *ICML*, pages 1549–1557. PMLR, 2017. (Cited on pages 126 and 164.)

J. Hu, A. Milzarek, Z. Wen, and Y. Yuan. Adaptive quadratically regularized Newton method for Riemannian optimization. *SIAM Journal on Matrix Analysis and Applications*, 39(3): 1181–1207, 2018. (Cited on page 93.)

J. Hu, B. Jiang, L. Lin, Z. Wen, and Y. Yuan. Structured quasi-Newton methods for optimization with orthogonality constraints. *SIAM Journal on Scientific Computing*, 41 (4):A2239–A2269, 2019. (Cited on page 93.)

J. Hu, X. Liu, Z-W. Wen, and Y-X. Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8(2):199–248, 2020. (Cited on page 95.)

F. Huang and S. Gao. Gradient descent ascent for minimax problems on Riemannian manifolds. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pages 1–11, 2023. (Cited on pages 90 and 94.)

F. Huang, S. Gao, J. Pei, and H. Huang. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *Journal of Machine Learning Research*, 23 (36):1–70, 2022a. (Cited on pages 280 and 314.)

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. (Cited on page 292.)

K. Huang and S. Zhang. An approximation-based regularized extra-gradient method for monotone variational inequalities. *ArXiv Preprint: 2210.04440*, 2022a. (Cited on pages 209 and 219.)

K. Huang and S. Zhang. New first-order algorithms for stochastic variational inequalities. *SIAM Journal on Optimization*, 32(4):2745–2772, 2022b. (Cited on page 314.)

K. Huang, J. Zhang, and S. Zhang. Cubic regularized Newton method for the saddle point models: A global and local convergence analysis. *Journal of Scientific Computing*, 91(2): 1–31, 2022b. (Cited on page 217.)

L. Huang, X. Liu, B. Lang, A. Yu, Y. Wang, and B. Li. Orthogonal weight normalization: solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *AAAI*, pages 3271–3278, 2018. (Cited on pages 88 and 96.)

S-Z. Huang. *Gradient Inequalities: with Applications to Asymptotic Behavior and Stability of Gradient-like Systems*, volume 126. American Mathematical Soc., 2006. (Cited on pages 126 and 149.)

W. T. Huh and P. Rusmevichientong. A nonparametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research*, 34(1):103–123, 2009. (Cited on pages 315, 320, 321, and 331.)

A. Ibrahim, W. Azizian, G. Gidel, and I. Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *ICML*, pages 4583–4593. PMLR, 2020. (Cited on pages 49, 50, 51, 54, 60, 61, and 64.)

A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2): 686–724, 2017. (Cited on pages 212 and 314.)

A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Variance-based extragradient methods with line search for stochastic variational inequalities. *SIAM Journal on Optimization*, 29(1):175–206, 2019. (Cited on page 314.)

S. Ivanov. On Helly's theorem in geodesic spaces. *Electronic Research Announcements*, 21: 109, 2014. (Cited on page 90.)

M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML*, pages 427–435. PMLR, 2013. (Cited on page 1.)

P. Jain and P. Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–363, 2017. (Cited on page 279.)

A. Jalilzadeh and U. V. Shanbhag. A proximal-point algorithm with variable sample-sizes (PPAWSS) for monotone stochastic variational inequality problems. In *WSC*, pages 3551–3562. IEEE, 2019. (Cited on page 314.)

A. Jambulapati, A. Sidford, and K. Tian. A direct tilde {O}(1/epsilon) iteration parallel algorithm for optimal transport. In *NeurIPS*, pages 11355–11366, 2019. (Cited on pages 243 and 257.)

P. Jawanpuria and B. Mishra. A unified framework for structured low-rank matrix learning. In *ICML*, pages 2254–2263. PMLR, 2018. (Cited on page 97.)

K. Ji, Z. Wang, Y. Zhou, and Y. Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *ICML*, pages 3100–3109. PMLR, 2019. (Cited on page 280.)

B. Jiang, H. Wang, and S. Zhang. An optimal high-order tensor method for convex optimization. In *COLT*, pages 1799–1801. PMLR, 2019. (Cited on pages 125, 127, 129, 130, 149, 150, 156, 161, 162, and 163.)

B. Jiang, T. Lin, and S. Zhang. A unified adaptive tensor approximation scheme to accelerate composite convex optimization. *SIAM Journal on Optimization*, 30(4):2897–2926, 2020. (Cited on pages 130, 161, and 209.)

H. Jiang and H. Xu. Stochastic approximation approaches to the stochastic variational inequality problem. *IEEE Transactions on Automatic Control*, 53(6):1462–1475, 2008. (Cited on page 314.)

R. Jiang and A. Mokhtari. Generalized optimistic methods for convex-concave saddle point problems. *ArXiv Preprint: 2202.09674*, 2022. (Cited on pages 205, 207, 216, 220, 225, and 238.)

C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *ICML*, pages 1724–1732. PMLR, 2017. (Cited on page 89.)

C. Jin, P. Netrapalli, and M. I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *ICML*, pages 4880–4889. PMLR, 2020. (Cited on pages 13, 14, 15, 42, 52, 54, and 89.)

C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021. (Cited on pages 2, 89, and 279.)

Y. Jin, A. Sidford, and K. Tian. Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods. In *COLT*, pages 4362–4415. PMLR, 2022. (Cited on page 314.)

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013. (Cited on page 318.)

M. I. Jordan. Artificial intelligence – the revolution hasn't happened yet. *Medium. Vgl. Ders.(2018): Perspectives and Challenges. Presentation SysML*, 2018. (Cited on pages 6, 13, and 50.)

M. I. Jordan, T. Lin, and E-V. Vlatakis-Gkaragkounis. First-order algorithms for min-max optimization in geodesic metric spaces. In *NeurIPS*, pages 6557–6574, 2022a. (Cited on pages 6, 7, and 9.)

M. I. Jordan, T. Lin, and M. Zampetakis. On the complexity of deterministic nonsmooth and nonconvex optimization. *ArXiv Preprint: 2209.12463*, 2022b. (Cited on pages 6 and 9.)

M. I. Jordan, T. Lin, and Z. Zhou. Adaptive, doubly optimal no-regret learning in games with gradient feedback. *Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4212851*, 2022c. (Cited on pages 6, 8, and 10.)

A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. *Optimization for Machine Learning*, 30(9): 149–183, 2011. (Cited on pages 50, 51, and 54.)

A. Juditsky and A. S. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *ArXiv Preprint: 0809.0813*, 2008. (Cited on page 299.)

A. Juditsky, P. Rigollet, A. B. Tsybakov, et al. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008. (Cited on page 333.)

A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011. (Cited on pages 28, 103, and 314.)

K-S. Jun and F. Orabona. Parameter-free online convex optimization with sub-exponential noise. In *COLT*, pages 1802–1823. PMLR, 2019. (Cited on page 315.)

S. Kakutani. A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal*, 8(3):457–459, 1941. (Cited on page 89.)

B. Kalantari and L. Khachiyan. On the complexity of nonnegative matrix scaling. *Linear Algebra and its Applications*, 240:87–103, 1996. (Cited on page 241.)

B. Kalantari, I. Lari, F. Ricca, and B. Simeone. On the complexity of general matrix scaling and entropy minimization via the RAS algorithm. *Mathematical Programming*, 112(2): 371–401, 2008. (Cited on page 241.)

D. Kamzolov. Near-optimal hyperfast second-order method for convex optimization. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 167–178. Springer, 2020. (Cited on page 130.)

A. Kannan and U. V. Shanbhag. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Computational Optimization and Applications*, 74(3):779–820, 2019. (Cited on pages 212 and 314.)

L. V. Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942. (Cited on pages 240 and 244.)

S. Karamardian. The complementarity problem. *Mathematical Programming*, 2(1):107–129, 1972. (Cited on page 166.)

H. Kasai and B. Mishra. Inexact trust-region algorithms on Riemannian manifolds. In *NIPS*, pages 4249–4260, 2018. (Cited on page 93.)

H. Kasai, P. Jawanpuria, and B. Mishra. Riemannian adaptive stochastic gradient algorithms on matrix manifolds. In *ICML*, pages 3262–3271. PMLR, 2019. (Cited on page 93.)

C. T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*. SIAM, 1995. (Cited on page 166.)

D. Kim. Accelerated proximal point method for maximally monotone operators. *Mathematical Programming*, 190(1):57–87, 2021. (Cited on page 167.)

D. Kinderlehrer and G. Stampacchia. *An Introduction to Variational Inequalities and Their Applications*. SIAM, 2000. (Cited on page 206.)

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. URL https://openreview.net/forum?id=8gmWwjFyLj. (Cited on pages 105, 291, 313, and 324.)

J. Kivinen and M. K. Warmuth. Averaging expert predictions. In *EuroCOLT*, pages 153–167. Springer, 1999. (Cited on page 333.)

K. C. Kiwiel. Restricted step and Levenberg-Marquardt techniques in proximal bundle methods for nonconvex nondifferentiable optimization. *SIAM Journal on Optimization*, 6 (1):227–249, 1996. (Cited on pages 293 and 294.)

K. C. Kiwiel. Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 18(2):379–388, 2007. (Cited on page 293.)

B. Kleinberg, Y. Li, and Y. Yuan. An alternative view: When does SGD escape local minima? In *ICML*, pages 2698–2707. PMLR, 2018. (Cited on page 213.)

B. Knaster, C. Kuratowski, and S. Mazurkiewicz. Ein Beweis des Fixpunktsatzes für n-dimensionale Simplexe. *Fundamenta Mathematicae*, 14(1):132–137, 1929. (Cited on page 89.)

P. A. Knight. The Sinkhorn-Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008. (Cited on page 241.)

O. Kolossoski and R. D. C. Monteiro. An accelerated non-Euclidean hybrid proximal extragradient-type algorithm for convex-concave saddle-point problems. *Optimization Methods and Software*, 32(6):1244–1272, 2017. (Cited on page 53.)

S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde. Generalized sliced Wasserstein distances. In *NeurIPS*, pages 261–272, 2019. (Cited on page 277.)

H. Komiya. Elementary proof for Sion's minimax theorem. *Kodai Mathematical Journal*, 11 (1):5–7, 1988. (Cited on pages 90 and 94.)

W. Kong and R. D. C. Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *SIAM Journal on Optimization*, 31(4):2558–2585, 2021. (Cited on pages 14, 15, 22, 52, 54, 55, and 89.)

T. Koren and K. Levy. Fast rates for exp-concave empirical risk minimization. In *NIPS*, pages 1477–1485, 2015. (Cited on pages 333 and 336.)

G. Kornowski and O. Shamir. Oracle complexity in nonsmooth nonconvex optimization. In *NeurIPS*, pages 324–334, 2021. (Cited on pages 283 and 284.)

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976. (Cited on pages 13, 15, 50, 53, 94, and 207.)

T. Kose. Solutions of saddle value problems by differential equations. *Econometrica, Journal of the Econometric Society*, pages 59–70, 1956. (Cited on page 15.)

J. Koshal, A. Nedić, and U. V. Shanbhag. Regularized iterative stochastic approximation methods for stochastic variational inequality problems. *IEEE Transactions on Automatic Control*, 58(3):594–609, 2012. (Cited on page 314.)

G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, I: operator extrapolation. *SIAM Journal on Optimization*, 32(3):2041–2073, 2022. (Cited on pages 92, 103, 207, 220, and 314.)

D. Kovalev and A. Gasnikov. The first optimal acceleration of high-order methods in smooth convex optimization. In *NeurIPS*, 2022. URL https://openreview.net/forum?id=YgmiL2UrO1P. (Cited on page 209.)

S. Krichene, W. Krichene, R. Dong, and A. Bayen. Convergence of heterogeneous distributed learning in stochastic routing games. In *Allerton*, pages 480–487. IEEE, 2015a. (Cited on pages 311 and 313.)

W. Krichene, A. Bayen, and P. L. Bartlett. Accelerated mirror descent in continuous and discrete time. In *NIPS*, pages 2845–2853, 2015b. (Cited on pages 126 and 129.)

A. Kristály. Nash-type equilibria on Riemannian manifolds: A variational approach. *Journal de Mathématiques Pures et Appliquées*, 101(5):660–688, 2014. (Cited on pages 90 and 94.)

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. (Cited on pages vi and 307.)

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. (Cited on page 241.)

H. W. Kuhn. Variants of the Hungarian method for assignment problems. *Naval Research Logistics Quarterly*, 3(4):253–258, 1956. (Cited on page 241.)

A. Kumar, P. Sattigeri, and P. T. Fletcher. Semi-supervised learning with GANs: Manifold invariance with improved inference. In *NIPS*, pages 5540–5550, 2017. (Cited on page 89.)

K. Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pages 769–783, 1998. (Cited on page 149.)

N. Lahn, D. Mulchandani, and S. Raghvendra. A graph theoretic additive approximation of optimal transport. In *NeurIPS*, pages 13813–13823, 2019. (Cited on pages 241 and 243.)

G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order augmented Lagrangian methods for convex programming. *Mathematical Programming*, 155(1-2):511–547, 2016. (Cited on page 54.)

G. Lan and Y. Zhou. An optimal randomized incremental gradient method. *Mathematical Programming*, 171(1):167–215, 2018a. (Cited on page 208.)

G. Lan and Y. Zhou. Random gradient extrapolation for distributed and stochastic optimization. *SIAM Journal on Optimization*, 28(4):2753–2782, 2018b. (Cited on page 208.)

J. LaSalle. Uniqueness theorems and successive approximations. *Annals of Mathematics*, pages 722–730, 1949. (Cited on page 146.)

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. (Cited on page 290.)

J. Lee. *Introduction to Smooth Manifolds*, volume 218. Springer Science & Business Media, 2012. (Cited on page 92.)

J. Lee, G. Kim, M. Olfat, M. Hasegawa-Johnson, and C. D. Yoo. Fast and efficient MMD-based fair PCA via optimization over Stiefel manifold. In *AAAI*, pages 7363–7371, 2022. (Cited on page 90.)

J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176(1-2): 311–337, 2019. (Cited on page 2.)

Y. T. Lee and A. Sidford. Path finding methods for linear programming: Solving linear programs in $\widetilde{O}(\text{sqrt(rank)})$ iterations and faster algorithms for maximum flow. In *FOCS*, pages 424–433. IEEE, 2014. (Cited on page 241.)

J. Lei. Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020. (Cited on page 241.)

L. Lei and M. I. Jordan. Less than a single pass: Stochastically controlled stochastic gradient. In *AISTATS*, pages 148–156. PMLR, 2017. (Cited on pages 313 and 317.)

L. Lei and M. I. Jordan. On the adaptivity of stochastic gradient-based optimization. *SIAM Journal on Optimization*, 30(2):1473–1500, 2020. (Cited on pages 313, 317, and 318.)

B. Lemaire. An asymptotical variational principle associated with the steepest descent method for a convex function. *Journal of Convex Analysis*, 3:63–70, 1996. (Cited on page 188.)

C. E. Lemke and J. T. Howson. Equilibrium points of bimatrix games. *Journal of the Society for Industrial and Applied Mathematics*, 12(2):413–423, 1964. (Cited on page 206.)

L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016. (Cited on pages 126 and 164.)

K. Y. Levy. Online to offline conversions, universality and adaptive minibatch sizes. In *NIPS*, pages 1612–1621, 2017. (Cited on pages 313 and 325.)

K. Y. Levy, A. Yurtsever, and V. Cevher. Online adaptive methods, universality and acceleration. In *NIPS*, pages 6501–6510, 2018. (Cited on pages 313 and 325.)

A. S. Lewis and J-S. Pang. Error bounds for convex inequality systems. In *Generalized Convexity, Generalized Monotonicity: Recent Results*, pages 75–110. Springer, 1998. (Cited on page 183.)

C. Li, G. López, and V. Martín-Márquez. Monotone vector fields and the proximal point algorithm on Hadamard manifolds. *Journal of the London Mathematical Society*, 79(3): 663–683, 2009. (Cited on page 94.)

X. Li and F. Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *AISTATS*, pages 983–992. PMLR, 2019. (Cited on page 313.)

X. Li, S. Chen, Z. Deng, Q. Qu, Z. Zhu, and A. Man-Cho So. Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. *SIAM Journal on Optimization*, 31(3):1605–1634, 2021. (Cited on page 93.)

Y. Li and Y. Yuan. Convergence analysis of two-layer neural networks with ReLU activation. In *NIPS*, pages 597–607, 2017. (Cited on page 213.)

X. Lian, H. Zhang, C-J. Hsieh, Y. Huang, and J. Liu. A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. In *NIPS*, pages 3062–3070, 2016. (Cited on page 280.)

T. Liang and J. Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *AISTATS*, pages 907–915. PMLR, 2019. (Cited on pages 13, 19, 53, 89, and 102.)

Q. Lin, Z. Lu, and L. Xiao. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization*, 25(4):2244–2273, 2015. (Cited on page 266.)

T. Lin and M. I. Jordan. A continuous-time perspective on monotone equation problems. *ArXiv Preprint: 2206.04770*, 2022a. (Cited on pages 6 and 7.)

T. Lin and M. I. Jordan. A control-theoretic perspective on optimal high-order optimization. *Mathematical Programming*, 195(1):929–975, 2022b. (Cited on pages 6, 7, 9, 173, 175, 176, 177, 178, and 210.)

T. Lin and M. I. Jordan. Perseus: A simple high-order regularization method for variational inequalities. *ArXiv Preprint: 2205.03202*, 2022c. (Cited on pages 6, 7, and 9.)

T. Lin and M. I. Jordan. Monotone inclusions, acceleration, and closed-loop control. *Mathematics of Operations Research*, To appear, 2023. (Cited on pages 6, 7, 9, 205, 207, 220, and 238.)

T. Lin, N. Ho, and M. Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *ICML*, pages 3982–3991. PMLR, 2019a. (Cited on pages 6, 8, 243, 246, and 250.)

T. Lin, Z. Hu, and X. Guo. Sparsemax and relaxed Wasserstein for topic sparsity. In *WSDM*, pages 141–149. ACM, 2019b. (Cited on page 241.)

T. Lin, C. Fan, N. Ho, M. Cuturi, and M. I. Jordan. Projection robust Wasserstein distance and Riemannian optimization. In *NeurIPS*, pages 9383–9397, 2020a. (Cited on pages 6, 89, and 95.)

T. Lin, N. Ho, X. Chen, M. Cuturi, and M. I. Jordan. Fixed-support Wasserstein barycenters: Computational hardness and fast algorithm. In *NeurIPS*, pages 5368–5380, 2020b. (Cited on pages 6 and 8.)

T. Lin, C. Jin, and M. I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *ICML*, pages 6083–6093. PMLR, 2020c. (Cited on pages 6, 7, 9, 52, 54, 57, and 89.)

T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. In *COLT*, pages 2738–2779. PMLR, 2020d. (Cited on pages 6, 7, 9, and 89.)

T. Lin, Z. Zhou, P. Mertikopoulos, and M. I. Jordan. Finite-time last-iterate convergence for multi-agent learning in games. In *ICML*, pages 6161–6171. PMLR, 2020e. (Cited on pages 6, 8, 312, 314, and 324.)

T. Lin, Z. Zheng, E. Chen, M. Cuturi, and M. I. Jordan. On projection robust optimal transport: Sample complexity and model misspecification. In *AISTATS*, pages 262–270. PMLR, 2021a. (Cited on pages 6, 8, and 96.)

T. Lin, Z. Zhou, W. Ba, and J. Zhang. Optimal no-regret learning in strongly monotone games with bandit feedback. *ArXiv Preprint: 2112.02856*, 2021b. (Cited on pages 324, 336, and 338.)

T. Lin, N. Ho, M. Cuturi, and M. I. Jordan. On the complexity of approximating multi-marginal optimal transport. *Journal of Machine Learning Research*, 23(65):1–43, 2022a. (Cited on pages 6 and 8.)

T. Lin, N. Ho, and M. I. Jordan. On the efficiency of entropic regularized algorithms for optimal transport. *Journal of Machine Learning Research*, 23(137):1–42, 2022b. (Cited on pages 6, 8, and 10.)

T. Lin, C. Jin, and M. I. Jordan. A nonasymptotic analysis of gradient descent ascent for nonconvex-concave minimax problems. *Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4181867*, 2022c. (Cited on pages 6, 7, and 9.)

T. Lin, P. Mertikopoulos, and M. I. Jordan. Explicit second-order min-max optimization methods with optimal convergence guarantee. *ArXiv Preprint: 2210.12860*, 2022d. (Cited on pages 6, 7, and 217.)

T. Lin, A. Pacchiano, Y. Yu, and M. I. Jordan. Online nonsubmodular minimization with delayed costs: From full information to bandit feedback. In *ICML*, pages 13441–13467. PMLR, 2022e. (Cited on page 6.)

T. Lin, Z. Zheng, and M. I. Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. In *NeurIPS*, pages 26160–26175, 2022f. (Cited on pages 6, 8, and 10.)

P. L. Lions. Une méthode itérative de résolution d'une inéquation variationnelle. *Israel Journal of Mathematics*, 31(2):204–208, 1978. (Cited on page 183.)

H. Liu, A. M-C. So, and W. Wu. Quadratic optimization with orthogonality constraint: explicit łojasiewicz exponent and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods. *Mathematical Programming*, 178(1-2):215–262, 2019. (Cited on page 93.)

M. Liu, Y. Mroueh, J. Ross, W. Zhang, X. Cui, P. Das, and T. Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. In *ICLR*, 2020. URL https://openreview.net/forum?id=SJxIm0VtwH. (Cited on page 89.)

M. Liu, H. Rafique, Q. Lin, and T. Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *Journal of Machine Learning Research*, 22(1):7651–7684, 2021. (Cited on pages 16, 89, and 212.)

S. Liu, B. Kailkhura, P-Y. Chen, P. Ting, S. Chang, and L. Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. In *NIPS*, pages 3731–3741, 2018. (Cited on page 280.)

P-L. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16(19): 559–616, 2015. (Cited on page 278.)

N. Loizou, H. Berard, G. Gidel, I. Mitliagkas, and S. Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. In *NeurIPS*, pages 19095–19108, 2021. (Cited on pages 311 and 314.)

S. Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963. (Cited on page 89.)

H. Lu, R. Freund, and V. Mirrokni. Accelerating greedy coordinate descent methods. In *ICML*, pages 3263–3272. PMLR, 2018. (Cited on page 266.)

S. Lu, I. Tsaknakis, M. Hong, and Y. Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020. (Cited on pages 14, 15, 18, 21, 22, 52, 55, and 89.)

H. Luo, A. Agarwal, N. Cesa-Bianchi, and J. Langford. Efficient second order online learning by sketching. In *NIPS*, pages 902–910, 2016. (Cited on page 334.)

C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20(3):451–632, 2020. (Cited on pages 278 and 279.)

C. J. Maddison, D. Paulin, Y. W. Teh, B. O'Donoghue, and A. Doucet. Hamiltonian descent methods. *ArXiv Preprint: 1809.05042*, 2018. (Cited on pages 126, 129, and 130.)

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb. (Cited on pages 5, 6, 13, 16, and 206.)

T. L. Magnanti and G. Perakis. A unifying geometric solution framework and complexity analysis for variational inequalities. *Mathematical Programming*, 71(3):327–351, 1995. (Cited on page 206.)

T. L. Magnanti and G. Perakis. Averaging schemes for variational inequalities and systems of equations. *Mathematics of Operations Research*, 22(3):568–587, 1997a. (Cited on page 213.)

T. L. Magnanti and G. Perakis. The orthogonality theorem and the strong-f-monotonicity condition for variational inequality algorithms. *SIAM Journal on Optimization*, 7(1):248–273, 1997b. (Cited on page 206.)

T. L. Magnanti and G. Perakis. Solving variational inequality and fixed point problems by line searches and potential optimization. *Mathematical Programming*, 101(3):435–461, 2004. (Cited on page 207.)

M. Mahdavi, L. Zhang, and R. Jin. Lower and upper bounds on the generalization of stochastic exponentially concave optimization. In *COLT*, pages 1305–1320. PMLR, 2015. (Cited on pages 333 and 336.)

P-E. Maingé. First-order continuous Newton-like systems for monotone inclusions. *SIAM Journal on Control and Optimization*, 51(2):1615–1638, 2013. (Cited on pages 126, 133, 139, 164, 167, and 176.)

S. Majewski, B. Miasojedow, and E. Moulines. Analysis of nonsmooth stochastic approximation: the differential inclusion approach. *ArXiv Preprint: 1805.01916*, 2018. (Cited on pages 279, 293, and 294.)

Y. Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015. (Cited on page 53.)

M. Marques Alves. Variants of the A-HPE and large-step A-HPE algorithms for strongly convex problems with applications to accelerated high-order tensor methods. *Optimization Methods and Software*, 37(6):2021–2051, 2022. (Cited on pages 130 and 209.)

B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. rev. française informat. *Recherche Opérationnelle*, 4:154–158, 1970. (Cited on pages 93, 153, and 166.)

B. Martinet. Détermination approchée d'un point fixe d'une application pseudo-contractante. *CR Acad. Sci. Paris*, 274(2):163–165, 1972. (Cited on pages 153 and 166.)

J. Martínez. On high-order model regularization for constrained optimization. *SIAM Journal on Optimization*, 27(4):2447–2458, 2017. (Cited on pages 130, 161, and 209.)

G. Mateos, J. A. Bazerque, and G. B. Giannakis. Distributed sparse linear regression. *IEEE Transactions on Signal Processing*, 58(10):5262–5276, 2010. (Cited on pages 12 and 50.)

R. May. Asymptotic for a second-order evolution equation with convex potential and vanishing damping term. *Turkish Journal of Mathematics*, 41(3):681–685, 2017. (Cited on pages 126 and 129.)

E. Mazumdar, L. J. Ratliff, and S. S. Sastry. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020. (Cited on page 89.)

E. V. Mazumdar, M. I. Jordan, and S. S. Sastry. On finding local Nash equilibria (and only local Nash equilibria) in zero-sum games. *ArXiv Preprint: 1901.00838*, 2019. (Cited on pages 13 and 16.)

J. Mei, Y. Gao, B. Dai, C. Szepesvari, and D. Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *ICML*, pages 7555–7564. PMLR, 2021. (Cited on page 89.)

S. Mei, T. Misiakiewicz, A. Montanari, and R. I. Oliveira. Solving SDPs for synchronization and MaxCut problems via the Grothendieck inequality. In *COLT*, pages 1476–1515. PMLR, 2017. (Cited on page 278.)

G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *NeurIPS*, pages 4541–4551, 2019. (Cited on pages 242 and 245.)

P. Mertikopoulos and W. H. Sandholm. Riemannian game dynamics. *Journal of Economic Theory*, 177:315–364, 2018. (Cited on page 89.)

P. Mertikopoulos and Z. Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1):465–507, 2019. (Cited on pages 206, 311, 313, 323, 335, and 336.)

P. Mertikopoulos, C. Papadimitriou, and G. Piliouras. Cycles in adversarial regularized learning. In *SODA*, pages 2703–2717. SIAM, 2018. (Cited on pages 13, 19, and 311.)

P. Mertikopoulos, B. Lecouat, H. Zenati, C-S Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *ICLR*, 2019. URL https://openreview.net/forum?id=Bkg8jjC9KQ. (Cited on pages 13, 16, 53, 89, 311, and 313.)

G. J. Minty. Monotone (nonlinear) operators in Hilbert space. *Duke Mathematical Journal*, 29(3):341–346, 1962. (Cited on page 206.)

K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik, and Y. Malitsky. Revisiting stochastic extragradient. In *AISTATS*, pages 4573–4582. PMLR, 2020. (Cited on page 53.)

A. Mokhtari, A. Ozdaglar, and S. Pattathil. Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020a. (Cited on pages 15, 28, 53, 207, 208, 211, and 220.)

A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *AISTATS*, pages 1497–1507. PMLR, 2020b. (Cited on pages 15, 50, 51, 53, 54, and 89.)

G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781. (Cited on page 240.)

B. Monnot and G. Piliouras. Limits and limitations of no-regret learning in games. *The Knowledge Engineering Review*, 32, 2017. (Cited on page 311.)

R. D. C. Monteiro and B. F. Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6):2755–2787, 2010. (Cited on pages 53, 151, 173, 183, 190, 191, 197, 206, 211, and 220.)

R. D. C. Monteiro and B. F. Svaiter. Complexity of variants of Tseng's modified FB splitting and Korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. *SIAM Journal on Optimization*, 21(4):1688–1720, 2011. (Cited on pages 53, 183, 206, and 211.)

R. D. C. Monteiro and B. F. Svaiter. Iteration-complexity of a Newton proximal extragradient method for monotone variational inequalities and inclusion problems. *SIAM Journal on Optimization*, 22(3):914–935, 2012. (Cited on pages 165, 167, 168, 169, 173, 177, 183, 189, 190, 191, 194, 197, 207, 211, 215, and 220.)

R. D. C. Monteiro and B. F. Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013. (Cited on pages 125, 127, 129, 130, 139, 149, 150, 151, 154, 155, 156, 158, 161, 163, 176, 178, and 209.)

M. Muehlebach and M. I. Jordan. A dynamical systems perspective on Nesterov acceleration. In *ICML*, pages 4656–4662. PMLR, 2019. (Cited on pages 126 and 129.)

M. Muehlebach and M. I. Jordan. Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives. *Journal of Machine Learning Research*, 22(1): 3407–3456, 2021. (Cited on pages 126, 129, 164, 175, and 197.)

M. C. Mukkamala and M. Hein. Variants of RMSProp and Adagrad with logarithmic regret bounds. In *ICML*, pages 2545–2553. PMLR, 2017. (Cited on pages 313 and 324.)

J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957. (Cited on page 241.)

K. G. Murty and S. N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming: Series A and B*, 39(2):117–129, 1987. (Cited on pages 2, 56, 65, and 279.)

H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NIPS*, pages 2208–2216, 2016. (Cited on pages 15, 16, and 55.)

I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019. (Cited on page 218.)

A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009. (Cited on pages 13, 15, and 53.)

B. L. Nelson. Optimization via simulation over discrete decision variables. In *Risk and Optimization in an Uncertain World*, pages 193–207. INFORMS, 2010. (Cited on page 279.)

A. Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004. (Cited on pages 13, 15, 50, 51, 53, 94, 183, 197, 207, 211, 220, and 314.)

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. (Cited on page 94.)

A. S. Nemirovski. Effective iterative methods for solving equations with monotone operators. *Ekonom. i Mat. Metody*, 17(2):344–359, 1981. (Cited on page 183.)

A. S. Nemirovski and Y. E. Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–30, 1985. (Cited on page 218.)

A. S. Nemirovski and D. B. Yudin. Cesari convergence of the gradient method of approximating saddle points of convex-concave functions. *Doklady Akademii Nauk*, 239(5):1056–1059, 1978. (Cited on page 183.)

A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. J. Wiley, 1983. (Cited on pages 279, 324, and 325.)

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005. (Cited on pages 50, 53, 54, 246, 248, and 257.)

Y. Nesterov. Cubic regularization of Newton's method for convex problems with constraints. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2006. (Cited on pages 205, 207, 214, and 215.)

Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007. (Cited on pages 50, 51, 53, 207, 211, 213, 214, 215, and 220.)

Y. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008. (Cited on pages 130, 209, and 214.)

Y. Nesterov. How to make the gradients small. *Optima*, 88:10–11, 2012. (Cited on page 144.)

Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013a. (Cited on pages 57, 126, 218, and 279.)

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013b. (Cited on page 26.)

Y. Nesterov. *Lectures on Convex Optimization*, volume 137. Springer, 2018. (Cited on pages 57, 67, 68, 74, 75, 78, 81, 127, 129, 164, 205, 207, 212, 218, 266, 281, and 314.)

Y. Nesterov. Inexact high-order proximal-point methods with auxiliary search procedure. *SIAM Journal on Optimization*, 31(4):2807–2828, 2021a. (Cited on pages 209 and 238.)

Y. Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, 186(1-2):157–183, 2021b. (Cited on pages 130, 150, 161, 194, 209, 217, 221, 222, 232, and 233.)

Y. Nesterov. Inexact accelerated high-order proximal-point methods. *Mathematical Programming*, pages 1–26, 2021c. (Cited on pages 209 and 238.)

Y. Nesterov. Superfast second-order methods for unconstrained convex optimization. *Journal of Optimization Theory and Applications*, 191(1):1–30, 2021d. (Cited on pages 130, 209, and 238.)

Y. Nesterov. Inexact accelerated high-order proximal-point methods. *Mathematical Programming*, 197(1):1–26, 2023. (Cited on page 130.)

Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006. (Cited on pages 209 and 214.)

Y. Nesterov and L. Scrimali. Solving strongly monotone variational and quasi-variational inequalities. *Available at SSRN 970903*, 2006. (Cited on pages 51 and 54.)

Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017. (Cited on pages 279, 280, 281, 284, 285, 286, and 287.)

Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Dokl. Akad. Nauk Sssr*, volume 269, pages 543–547, 1983. (Cited on pages 57, 126, 129, 153, 175, and 214.)

S. Netessine and N. Rudi. Centralized and competitive inventory models with demand substitution. *Operations Research*, 51(2):329–335, 2003. (Cited on pages 315 and 330.)

J. V. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928. (Cited on pages 12, 14, and 90.)

K. Nguyen, N. Ho, T. Pham, and H. Bui. Distributional sliced-Wasserstein and applications to generative modeling. In *ICLR*, 2021. URL https://openreview.net/forum?id=QYjO70ACDK. (Cited on page 277.)

L. M. Nguyen, M. van Dijk, D. T. Phan, P. H. Nguyen, T-W. Weng, and J. R. Kalagnanam. Finite-sum smooth optimization with SARAH. *Computational Optimization and Applications*, 82(3):561–593, 2022. (Cited on page 313.)

X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 4(1):370–400, 2013. (Cited on page 241.)

X. Nguyen. Borrowing strength in hierarchical Bayes: posterior concentration of the Dirichlet base measure. *Bernoulli*, 22(3):1535–1571, 2016. (Cited on page 241.)

N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007. (Cited on pages 12 and 49.)

M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *NeurIPS*, pages 14905–14916, 2019. (Cited on pages 13, 14, 15, 18, 21, 22, 42, 52, 55, and 89.)

B. O'Donoghue and C. J. Maddison. Hamiltonian descent for composite objectives. In *NeurIPS*, pages 14470–14480, 2019. (Cited on page 126.)

Z. Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bulletin of the American Mathematical Society*, 73(4):591–597, 1967. (Cited on page 179.)

F. Orabona and D. Pál. Coin betting and parameter-free online learning. In *NIPS*, pages 577–585, 2016. (Cited on page 315.)

A. Orda, R. Rom, and N. Shimkin. Competitive routing in multiuser communication networks. *IEEE/ACM Transactions on Networking*, 1(5):510–521, 1993. (Cited on page 323.)

E. Ordentlich and T. M. Cover. The cost of achieving the best portfolio in hindsight. *Mathematics of Operations Research*, 23(4):960–982, 1998. (Cited on page 333.)

J. B. Orlin. A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78:109–129, 1997. (Cited on pages 1 and 241.)

J. B. Orlin and R. K. Ahuja. New scaling algorithms for the assignment and minimum mean cycle problems. *Mathematical Programming*, 54(1):41–56, 1992. (Cited on page 241.)

J. B. Orlin, S. A. Plotkin, and E. Tardos. Polynomial dual network simplex algorithms. *Mathematical Programming*, 60(1-3):255–276, 1993. (Cited on pages 1 and 241.)

M. J. Osborne. *An Introduction to Game Theory*. Oxford University Press, New York, 2004. (Cited on page 166.)

M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1994. (Cited on page 323.)

P. Ostroukhov, R. Kamalov, P. Dvurechensky, and A. Gasnikov. Tensor methods for strongly convex strongly concave saddle point problems and strongly monotone variational inequalities. *ArXiv Preprint: 2012.15595*, 2020. (Cited on pages 130 and 219.)

D. M. Ostrovskii, A. Lowy, and M. Razaviyayn. Efficient search of first-order Nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4): 2508–2538, 2021. (Cited on pages 52, 55, and 89.)

Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1-2):1–35, 2021. (Cited on pages 50, 51, 54, 64, 183, 207, 214, 220, and 223.)

Y. Ouyang, Y. Chen, G. Lan, and E. Pasiliao Jr. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, 8(1):644–681, 2015. (Cited on pages 50 and 54.)

B. O'donoghue and E. Candes. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015. (Cited on pages 218 and 219.)

S. Pal, T-K. L. Wong, et al. Exponentially concave functions and a new information geometry. *The Annals of Probability*, 46(2):1070–1113, 2018. (Cited on page 333.)

S. Park. Riemannian manifolds are KKM spaces. *Advances in the Theory of Nonlinear Analysis and its Application*, 3(2):64–73, 2019. (Cited on pages 90 and 94.)

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga. Pytorch: an imperative style, high-performance deep learning library. In *NeurIPS*, pages 8026–8037, 2019. (Cited on page 291.)

F-P. Paty and M. Cuturi. Subspace robust Wasserstein distances. In *ICML*, pages 5072–5081. PMLR, 2019. (Cited on pages 241 and 277.)

O. Pele and M. Werman. Fast and robust earth mover's distance. In *ICCV*. IEEE, 2009. (Cited on page 241.)

X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006. (Cited on page 95.)

P. Petersen. *Riemannian Geometry*, volume 171. Springer, 2006. (Cited on page 108.)

G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. (Cited on page 241.)

G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *ICML*, pages 2664–2672, 2016. (Cited on page 241.)

B. T. Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963. (Cited on page 89.)

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. (Cited on page 175.)

B. T. Polyak. *Introduction to Optimization*. Optimization Software, Inc, New York, 1987. (Cited on pages 126 and 313.)

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. (Cited on page 93.)

A. D. Polyanin and V. F. Zaitsev. *Handbook of Nonlinear Partial Differential Equations: Exact Solutions, Methods, and Problems*. Chapman and Hall/CRC, 2003. (Cited on page 166.)

L. D. Popov. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980. (Cited on page 207.)

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014. (Cited on page 1.)

M. S. Pydi and V. Jog. Adversarial risk via optimal transport and optimal couplings. In *ICML*, pages 7814–7823. PMLR, 2020. (Cited on page 277.)

K. Quanrud. Approximating optimal transport with linear programs. In *SOSA*, pages 61–69, 2019. (Cited on page 242.)

H. Rafique, M. Liu, Q. Lin, and T. Yang. Weakly-convex-concave min-max optimization: Provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022. (Cited on pages 14, 15, 52, and 54.)

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *COLT*, pages 993–1019, 2013a. (Cited on page 53.)

S. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. In *NIPS*, pages 3066–3074, 2013b. (Cited on page 53.)

D. Ralph and S. J. Wright. Superlinear convergence of an interior-point method for monotone variational inequalities. *Complementarity and Variational Problems: State of the Art*, pages 345–385, 1997. (Cited on page 206.)

T. Rappaport. *Wireless Communications: Principles and Practice*. Prentice Hall PTR, 2001. (Cited on pages 329 and 330.)

G. Raskutti and S. Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015. (Cited on page 88.)

S. Reddi, M. Zaheer, S. Sra, B. Poczos, F. Bach, R. Salakhutdinov, and A. Smola. A generic approach for escaping saddle points. In *AISTATS*, pages 1233–1242. PMLR, 2018a. (Cited on page 89.)

S. J. Reddi, S. Kale, and S. Kumar. On the convergence of Adam and beyond. In *ICLR*, pages 1–23, 2018b. URL https://openreview.net/forum?id=ryQu7f-RZ. (Cited on page 313.)

J. Renegar and B. Grimmer. A simple nearly optimal restart scheme for speeding up first-order methods. *Foundations of Computational Mathematics*, 22(1):211–256, 2022. (Cited on pages 218 and 219.)

J. Robinson. An iterative method of solving a game. *Annals of Mathematics*, pages 296–301, 1951. (Cited on page 15.)

S. M. Robinson. Local structure of feasible sets in nonlinear programming, Part III: Stability and sensitivity. In *Nonlinear Analysis and Optimization*, pages 45–66. Springer, 1987. (Cited on page 197.)

R. T. Rockafellar. *Convex Analysis*, volume 28. Princeton University Press, 1970. (Cited on pages 87, 165, 181, 182, 198, and 199.)

R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976. (Cited on pages 93, 161, and 166.)

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 2015. (Cited on pages 25 and 26.)

R. T. Rockafellar and J. Sun. Solving monotone stochastic variational inequalities and complementarity problems by progressive hedging. *Mathematical Programming*, 174(1-2): 453–471, 2019. (Cited on page 315.)

R. T. Rockafellar and R. J-B. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009. (Cited on pages 216, 285, and 297.)

R. T. Rockafellar and R. J. B. Wets. Stochastic variational inequalities: Single-stage to multistage. *Mathematical Programming*, 165(1):331–360, 2017. (Cited on page 314.)

A. Rolet, M. Cuturi, and G. Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In *AISTATS*, pages 630–638, 2016. (Cited on page 241.)

J. B. Rosen. Existence and uniqueness of equilibrium points for concave N-person games. *Econometrica: Journal of the Econometric Society*, pages 520–534, 1965. (Cited on pages 323 and 341.)

V. Roulet and A. d'Aspremont. Sharpness, restart and acceleration. In *NeurIPS*, pages 1119–1129, 2017. (Cited on page 219.)

N. L. Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2663–2671, 2012. (Cited on page 313.)

D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988. (Cited on page 93.)

W. H. Sandholm. Population games and deterministic evolutionary dynamics. In *Handbook of Game Theory with Economic Applications*, volume 4, pages 703–778. Elsevier, 2015. (Cited on page 323.)

M. Sanjabi, J. Ba, M. Razaviyayn, and J. D. Lee. On the convergence and robustness of training gans with regularized optimal transport. In *NIPS*, pages 7091–7101, 2018. (Cited on page 15.)

H. Scarf. The approximation of fixed points of a continuous mapping. *SIAM Journal on Applied Mathematics*, 15(5):1328–1343, 1967. (Cited on page 206.)

M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017. (Cited on page 273.)

M. H. Schneider and S. A. Zenios. A comparative study of algorithms for matrix balancing. *Operations Research*, 38(3):439–455, 1990. (Cited on page 241.)

D. Scieur, V. Roulet, F. Bach, and A. d'Aspremont. Integration methods and optimization algorithms. In *NIPS*, pages 1109–1118, 2017. (Cited on page 129.)

O. Sebbouh, C. Dossal, and A. Rondepierre. Convergence rates of damped inertial dynamics under geometric conditions and perturbations. *SIAM Journal on Optimization*, 30(3): 1850–1877, 2020. (Cited on pages 126 and 129.)

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012. (Cited on pages 309 and 317.)

O. Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(1):1703–1713, 2017. (Cited on pages 280, 281, 284, 286, 288, and 298.)

J. Shamma. *Cooperative Control of Distributed Multi-agent Systems*. John Wiley & Sons, 2008. (Cited on pages 12, 50, and 166.)

U. V. Shanbhag. Stochastic variational inequality problems: Applications, analysis, and algorithms. In *Theory Driven by Influential Applications*, pages 71–107. INFORMS, Catonsville, MD, 2013. (Cited on page 314.)

A. Shapiro. On concepts of directional differentiability. *Journal of Optimization Theory and Applications*, 66(3):477–487, 1990. (Cited on page 282.)

J. Sherman. Area-convexity, ℓ_∞ regularization, and undirected multicommodity flow. In *STOC*, pages 452–460. ACM, 2017. (Cited on pages 243 and 257.)

B. Shi, S. S. Du, W. J. Su, and M. I. Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In *NeurIPS*, pages 5744–5752, 2019. (Cited on page 129.)

B. Shi, S. S. Du, M. I. Jordan, and W. J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195(1-2):79–148, 2022. (Cited on pages 126, 127, 128, 129, 138, 141, 144, 161, 163, and 176.)

Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2008. (Cited on page 311.)

A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *ICLR*, 2018. URL https://openreview.net/forum?id=Hk6kPgZA-. (Cited on pages 5, 6, 12, 13, 15, 16, 24, 25, 50, 55, 166, and 206.)

R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *Proceedings of the American Mathematical Society*, 45(2):195–198, 1974. (Cited on page 241.)

M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. (Cited on pages 15, 56, 90, 94, and 99.)

M. V. Solodov. Convergence rate analysis of iteractive algorithms for solving variational inequality problems. *Mathematical Programming*, 96(3):513–528, 2003. (Cited on page 184.)

M. V. Solodov and B. F. Svaiter. A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7(4): 323–345, 1999a. (Cited on page 151.)

M. V. Solodov and B. F. Svaiter. A new projection method for variational inequality problems. *SIAM Journal on Control and Optimization*, 37(3):765–776, 1999b. (Cited on page 212.)

M. V. Solodov and B. F. Svaiter. Forcing strong convergence of proximal point iterations in a Hilbert space. *Mathematical Programming*, 87(1):189–202, 2000. (Cited on page 181.)

M. Sommerfeld and A. Munk. Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238, 2018. (Cited on page 241.)

M. Sommerfeld, Y. Zemel, and A. Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20(105):1–23, 2019. (Cited on page 241.)

C. Song, Z. Zhou, Y. Zhou, Y. Jiang, and Y. Ma. Optimistic dual extrapolation for coherent non-monotone variational inequalities. In *NeurIPS*, pages 14303–14314, 2020. (Cited on pages 212 and 221.)

C. Song, Y. Jiang, and Y. Ma. Unified acceleration of high-order algorithms under general Hölder continuity. *SIAM Journal on Optimization*, 31(3):1797–1826, 2021. (Cited on pages 127, 130, 140, 178, and 209.)

S. Sorin and C. Wan. Finite composite games: Equilibria and dynamics. *Journal of Dynamics and Games*, 3(1):101–120, 2016. (Cited on page 323.)

S. Sra and R. Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015. (Cited on page 89.)

S. Sra and R. Hosseini. Geometric optimization in machine learning. In *Algorithmic Advances in Riemannian Geometry and Applications*, pages 73–91. Springer, 2016. (Cited on page 89.)

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. (Cited on page 291.)

S. Srivastava, V. Cevher, Q. Dinh, and D. Dunson. WASP: Scalable Bayes via barycenters of subset posteriors. In *AISTATS*, pages 912–920, 2015. (Cited on page 241.)

S. Srivastava, C. Li, and D. Dunson. Scalable Bayes via barycenter in Wasserstein space. *Journal of Machine Learning Research*, 19(8):1–35, 2018. (Cited on page 241.)

H. Stadtler. Supply chain management — an overview. *Supply Chain Management and Advanced Planning*, pages 9–36, 2008. (Cited on page 279.)

W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Journal of Machine Learning Research*, 17(1):5312–5354, 2016. (Cited on pages 126, 127, 129, 141, and 175.)

J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016a. (Cited on pages 88 and 96.)

J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2): 885–914, 2016b. (Cited on pages 88 and 96.)

J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018. (Cited on page 279.)

X. A. Sun and A. J. Conejo. *Robust Optimization in Electric Energy Systems*. Springer, 2021. (Cited on page 5.)

Y. Sun, N. Flammarion, and M. Fazel. Escaping from saddle points on Riemannian manifolds. In *NeurIPS*, pages 7274–7284, 2019. (Cited on page 93.)

W. A. Sutherland. *Introduction to Metric and Topological Spaces*. Oxford University Press, 2009. (Cited on page 136.)

C. W. Tan. Wireless network optimization by Perron-Frobenius theory. In *CISS*, pages 1–6. IEEE, 2014. (Cited on page 324.)

K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. In *NeurIPS*, pages 12659–12670, 2019. (Cited on pages 14, 15, 18, 22, 50, 51, 52, 54, and 89.)

L. Tian, K. Zhou, and A. M-C. So. On the finite-time complexity and practical computation of approximate stationarity concepts of Lipschitz functions. In *ICML*, pages 21360–21379. PMLR, 2022. (Cited on pages 279, 283, and 294.)

A. A. Titov, S. S. Ablaev, M. S. Alkousa, F. S. Stonyakin, and A. V. Gasnikov. Some adaptive first-order methods for variational inequalities with relatively strongly monotone operators and generalized smoothness. *ArXiv Preprint: 2207.09544*, 2022. (Cited on page 217.)

M. J. Todd. *The Computation of Fixed Points and Applications*. Springer, 2013. (Cited on page 206.)

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. In *ICLR*, 2018. URL https://openreview.net/forum?id=HkL7n1-0b. (Cited on page 241.)

R. Trémolières, J-L. Lions, and R. Glowinski. *Numerical Analysis of Variational Inequalities*. Elsevier, 2011. (Cited on page 206.)

N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan. Averaging stochastic gradient descent on Riemannian manifolds. In *COLT*, pages 650–687, 2018. (Cited on page 93.)

R. Tron, B. Afsari, and R. Vidal. Riemannian consensus for manifolds with bounded curvature. *IEEE Transactions on Automatic Control*, 58(4):921–934, 2012. (Cited on page 89.)

P. Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995. (Cited on pages 49, 50, 51, 53, 54, 183, and 197.)

P. Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000. (Cited on pages 207 and 211.)

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2:3, 2008. (Cited on pages 50, 51, and 53.)

P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125(2):263–295, 2010. (Cited on page 126.)

H. Uzawa. Iterative methods for concave programming. *Studies in Linear and Nonlinear Programming*, 6:154–165, 1958. (Cited on page 15.)

J. van den Brand, Y. T. Lee, Y. P. Liu, T. Saranurak, A. Sidford, Z. Song, and D. Wang. Minimum cost flows, MDPs, and ℓ_1-regression in nearly linear time for dense instances. In *STOC*, pages 859–869, 2021. (Cited on page 241.)

B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. (Cited on page 97.)

A. Vassilis, A. Jean-François, and D. Charles. The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case b $\leq$ 3. *SIAM Journal on Optimization*, 28(1):551–574, 2018. (Cited on pages 126, 129, and 164.)

S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien. Painless stochastic gradient: interpolation, line-search, and convergence rates. In *NeurIPS*, pages 3732–3745, 2019. (Cited on page 313.)

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009. (Cited on pages 240 and 333.)

Y. Viossat and A. Zapechelnyuk. No-regret dynamics and fictitious play. *Journal of Economic Theory*, 148(2):825–842, 2013. (Cited on page 311.)

E. V. Vlatakis-Gkaragkounis, L. Flokas, and G. Piliouras. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. In *NeurIPS*, pages 10450–10461, 2019. (Cited on page 89.)

E. V. Vlatakis-Gkaragkounis, L. Flokas, T. Lianeas, P. Mertikopoulos, and G. Piliouras. No-regret learning and mixed Nash equilibria: They do not mix. In *NeurIPS*, pages 1380–1391, 2020. (Cited on page 88.)

E. V. Vlatakis-Gkaragkounis, L. Flokas, and G. Piliouras. Solving min-max optimization with hidden structure via gradient descent ascent. In *NeurIPS*, pages 2373–2386, 2021. (Cited on page 89.)

J. Von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior (Commemorative Edition)*. Princeton University Press, 2007. (Cited on pages 12 and 49.)

V. Vovk. Competitive online linear regression. In *NIPS*, pages 364–370, 1997. (Cited on page 333.)

M. J. Wainwright. *High-Dimensional Statistics: A Non-asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019. (Cited on pages 298 and 303.)

G. Wang, G. B. Giannakis, and Y. C. Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794, 2017. (Cited on page 278.)

G. Wang, S. Lu, Q. Cheng, W. Tu, and L. Zhang. SAdam: A variant of Adam for strongly convex functions. In *ICLR*, pages 1–21, 2020. URL https://openreview.net/forum?id=rye5YaEtPr. (Cited on page 313.)

J. H. Wang, G. López, V. Martín-Márquez, and C. Li. Monotone and accretive vector fields on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 146 (3):691–708, 2010. (Cited on page 94.)

Y. Wang, S. Du, S. Balakrishnan, and A. Singh. Stochastic zeroth-order optimization in high dimensions. In *AISTATS*, pages 1356–1365. PMLR, 2018. (Cited on page 280.)

R. Ward, X. Wu, and L. Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In *ICML*, pages 6677–6686. PMLR, 2019. (Cited on page 313.)

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019. (Cited on pages 241, 245, and 277.)

P. C. Weeraddana, M. Codreanu, M. Latva-aho, A. Ephremides, and C. Fischione. *Weighted Sum-Rate Maximization in Wireless Networks: A Review*. Now Foundations and Trends, 2012. (Cited on page 324.)

Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013. (Cited on page 93.)

A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016. (Cited on pages 126, 129, 140, 164, 178, 197, and 210.)

A. Wiesel. Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing*, 60(12):6182–6189, 2012. (Cited on page 89.)

A. C. Wilson, L. Mackey, and A. Wibisono. Accelerating rescaled gradient descent: Fast optimization of smooth functions. In *NeurIPS*, pages 13555–13565, 2019. (Cited on pages 129 and 130.)

A. C. Wilson, B. Recht, and M. I. Jordan. A Lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(1):5040–5073, 2021. (Cited on pages 141 and 187.)

Z. Xie and J. Shi. Accelerated primal dual method for a class of saddle point problem with strongly convex component. *ArXiv Preprint: 1906.07691*, 2019. (Cited on page 54.)

H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(7):1485–1510, 2009. (Cited on pages 12 and 50.)

Y. Xu. Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. *Mathematical Programming*, 185:199–244, 2021. (Cited on page 54.)

Y. Xu and S. Zhang. Accelerated primal-dual proximal block coordinate updating methods for constrained convex optimization. *Computational Optimization and Applications*, 70 (1):91–128, 2018. (Cited on page 54.)

Y. Xu, Q. Lin, and T. Yang. Adaptive SVRG methods under error bound conditions with unknown growth parameter. In *NIPS*, pages 3279–3289, 2017. (Cited on page 313.)

A. Yadav, S. Shah, Z. Xu, D. Jacobs, and T. Goldstein. Stabilizing adversarial nets with prediction methods. In *ICLR*, 2018. URL https://openreview.net/forum?id=Skj8Kag0Z. (Cited on page 53.)

T. Yang, Z. Li, and L. Zhang. A simple analysis for exp-concave empirical minimization with arbitrary convex regularizer. In *AISTATS*, pages 445–453, 2018. (Cited on pages 333 and 336.)

Y. Yazıcı, C-S. Foo, S. Winkler, K-H. Yap, G. Piliouras, and V. Chandrasekhar. The unusual effectiveness of averaging in GAN training. In *ICLR*, 2019. URL https://openreview.net/forum?id=SJgw_sRqFQ. (Cited on page 105.)

Y. Ye. The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, 36(4): 593–603, 2011. (Cited on page 1.)

F. Yousefian, A. Nedić, and U. V. Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012. (Cited on page 284.)

F. Yousefian, A. Nedić, and U. V. Shanbhag. A regularized smoothing stochastic approximation (RSSA) algorithm for stochastic variational inequality problems. In *WSC*, pages 933–944. IEEE, 2013. (Cited on page 314.)

F. Yousefian, A. Nedić, and U. V. Shanbhag. Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems. In *CDC*, pages 5831–5836. IEEE, 2014. (Cited on page 314.)

Y. Yu, T. Lin, E. Mazumdar, and M. I. Jordan. Fast distributionally robust learning with variance-reduced min-max optimization. In *AISTATS*, pages 1219–1250. PMLR, 2022. (Cited on page 314.)

J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018. (Cited on page 4.)

H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *COLT*, pages 1617–1638. PMLR, 2016. (Cited on pages 93, 97, 99, and 108.)

H. Zhang, S. J. Reddi, and S. Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *NIPS*, pages 4592–4600, 2016. (Cited on page 93.)

J. Zhang, A. Mokhtari, S. Sra, and A. Jadbabaie. Direct Runge-Kutta discretization achieves acceleration. In *NIPS*, pages 3900–3909, 2018. (Cited on pages 129 and 130.)

J. Zhang, H. Lin, S. Jegelka, S. Sra, and A. Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *ICML*, pages 11173–11182. PMLR, 2020a. (Cited on pages 279, 282, 283, 285, 289, 291, and 294.)

J. Zhang, S. Ma, and S. Zhang. Primal-dual optimization algorithms over Riemannian manifolds: An iteration complexity analysis. *Mathematical Programming*, 184(1):445–490, 2020b. (Cited on page 93.)

J. Zhang, M. Hong, and S. Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, 194(1-2):901–935, 2022a. (Cited on pages 49, 50, 51, 54, 60, 61, 64, 102, 220, and 223.)

P. Zhang, J. Zhang, and S. Sra. Minimax in geodesic metric spaces: Sion's theorem and algorithms. *ArXiv Preprint: 2202.06950*, 2022b. (Cited on pages 88, 90, 91, 92, 94, 99, 100, 102, 104, 105, 106, 112, 117, and 121.)

R. Zhao. Accelerated stochastic algorithms for convex-concave saddle-point problems. *Mathematics of Operations Research*, 47(2):1443–1473, 2022. (Cited on pages 51 and 54.)

R. Zhao. A primal-dual smoothing framework for max-structured nonconvex optimization. *Mathematics of Operations Research*, To appear, 2023. (Cited on pages 52 and 55.)

Z. Zhou, P. Mertikopoulos, A. L. Moustakas, N. Bambos, and P. Glynn. Mirror descent learning in continuous games. In *CDC*, pages 5776–5783. IEEE, 2017. (Cited on pages 311 and 313.)

Z. Zhou, P. Mertikopoulos, S. Athey, N. Bambos, P. Glynn, and Y. Ye. Learning in games with lossy feedback. In *NIPS*, pages 1–11, 2018. (Cited on pages 311 and 313.)

Z. Zhou, P. Mertikopoulos, A. L. Moustakas, N. Bambos, and P. Glynn. Robust power management via learning and game design. *Operations Research*, 69(1):331–345, 2021. (Cited on pages 4, 311, 312, 314, 315, 324, 330, and 340.)

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003. (Cited on page 310.)

F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu. A sufficient condition for convergences of Adam and RMSProp. In *CVPR*, pages 11127–11135, 2019. (Cited on page 313.)