Decision-Making in multi-agent systems : delays, adaptivity, and learning in games

Yu-Guan Hsieh

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Mathématiques et Informatique
Unité de recherche : Laboratoire Jean Kuntzmann

Prise de décision dans les systèmes multi-agents : délais, adaptabilité et apprentissage dans les jeux

Decision-Making in multi-agent systems: delays, adaptivity, and learning in games

Présentée par :

Yu-Guan HSIEH

Direction de thèse :

Jerome MALICK DIRECTEUR DE RECHERCHE, CNRS DELEGATION ALPES	Directeur de thèse
Panayotis MERTIKOPOULOS CHARGE DE RECHERCHE HDR, CNRS DELEGATION ALPES	Co-directeur de thèse
Franck IUTZELER PROFESSEUR DES UNIVERSITES, UNIVERSITE TOULOUSE III	Co-encadrant de thèse

Rapporteurs :

CONSTANTINOS DASKALAKIS
FULL PROFESSOR, MASSACHUSETTS INSTITUTE OF TECHNOLOGY
SYLVAIN SORIN
PROFESSEUR DES UNIVERSITES EMERITE, Sorbonne Université

Thèse soutenue publiquement le **7 novembre 2023**, devant le jury composé de :

JERÔME MALICK DIRECTEUR DE RECHERCHE, CNRS DELEGATION ALPES	Directeur de thèse
CONSTANTINOS DASKALAKIS FULL PROFESSOR, MASSACHUSETTS INSTITUTE OF TECHNOLOGY	Rapporteur
SYLVAIN SORIN PROFESSEUR DES UNIVERSITES EMERITE, Sorbonne Université	Rapporteur
NICOLO CESA-BIANCHI FULL PROFESSOR, UNIVERSITA DEGLI STUDI DI MILANO	Examinateur
MARYAM KAMGARPOUR ASSISTANT PROFESSOR, ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE	Examinatrice
ALEXANDRE D'ASPREMONT DIRECTEUR DE RECHERCHE, CNRS DELEGATION PARIS CENTRE	Examinateur
ANATOLI JUDITSKY PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES	Président

Invités :

FRANCK IUTZELER
PROFESSEUR DES UNIVERSITES, UNIVERSITE TOULOUSE III
PANAYOTIS MERTIKOPOULOS
CHARGE DE RECHERCHE HDR, CNRS DELEGATION ALPES

DECISION-MAKING IN MULTI-AGENT SYSTEMS: DELAYS, ADAPTIVITY, AND LEARNING IN GAMES

YU-GUAN HSIEH

Ph.D. Thesis

JURY

RAPPORTEUR	Constantinos Daskalakis Massachusetts Institute of Technology
RAPPORTEUR	Sylvain Sorin Sorbonne Université
EXAMINATEUR	Nicolò Cesa-Bianchi Università degli Studi di Milano
EXAMINATEUR	Alexandre d'Aspremont CNRS & École Normale Supérieure
EXAMINATEUR	Anatoli Juditsky Université Grenoble Alpes
EXAMINATRICE	Maryam Kamgarpour École Polytechnique Fédérale de Lausanne

DIRECTION DE THÈSE

DIRECTEUR DE THÈSE	Jérôme Malick CNRS
CO-DIRECTEUR	Franck Iutzeler Université Toulouse III - Paul Sabatier
CO-DIRECTEUR	Panayotis Mertikopoulos CNRS

Université Grenoble Alpes
École doctorale MSTII
Spécialité: *Mathématiques-informatique*

Decision-making in multi-agent systems, a journey toward the unknown

Depicted by Stable Diffusion

PREFACE

With the increasing deployment of decision-making and learning algorithms in multi-agent systems, it becomes imperative to understand their efficiency and improve their performance. The design and analysis of these systems, however, confront significant challenges. These range from practical implementation issues to the intrinsic complexity of multi-agent dynamics, where agent interactions can be cooperative, competitive, or a mix of the two. On top of this is the presence of non-stationarity, driven by either the unpredictable character of nature or interaction with other strategic entities.

This thesis represents a targeted attempt to navigate this complex landscape, investigating separately two critical aspects: the impact of delays and the interactions among agents with non-aligned interests. This dual focus is due to the relevance of these issues to practical deployment and the inherent difficulty of learning in such systems, aiming to reveal fundamental insights about how information flow and strategic interactions influence the overall system's learning and decision-making processes. Our approaches are grounded in decentralized optimization and game theory, using online learning as a principal methodology to address the non-stationarity of the environment.

Our first series of contributions concerns the study of a dual averaging algorithm in a cooperative online learning setup. This setup features asynchronicity and delays, which pose a significant obstacle to conventional regret analysis. To address this difficulty, we introduce the key concepts of virtual iterates and faithful permutations, which enable us to establish a universal regret bound for this setting. Our results further extend to an optimistic version of dual averaging, which leverages slow variation in the sequence of losses encountered by the agents.

Moving forward, we investigate convergence to Nash equilibrium and individual performance guarantees, as measured by the agents' regrets, when the agents' interactions are governed by a general, non-cooperative game. Our algorithms are again based on the principle of optimism, incorporating a "lookahead" step that reuses the most recent information.

Importantly, across both contexts, we put emphasis on the "adaptivity" of our algorithms and their resilience in handling "uncertainty" during interactions. Our methods work without any coordination among agents, and can be implemented even when the agents are completely oblivious of their environment (and/or the game that they are involved in). A significant aspiration of our approach is to provide adaptive guarantees, robust to the dynamic nature of the environments, where uncertainty can stem from a lack of knowledge or be modeled as we do with a noisy oracle in the learning-in-games setup.

RÉSUMÉ

FACE au déploiement croissant d'algorithmes de décision et d'apprentissage dans les systèmes multi-agents, il devient impératif de comprendre leur efficacité et d'améliorer leurs performances. Cependant, la conception et l'analyse de ces systèmes se heurtent à défis importants, qui s'étendent des problèmes pratiques d'implémentation jusqu'à la complexité intrinsèque des dynamiques multi-agents, avec des interactions entre les agents qui peuvent être coopératives, compétitives ou un mélange des deux.

Cette thèse vise à naviguer dans ce paysage complexe, en examinant séparément deux aspects critiques : l'impact du délai et des interactions entre agents aux intérêts contradictoires. L'objectif ici est d'établir des connaissances fondamentales sur la façon dont le flux d'informations et les interactions stratégiques influencent les processus d'apprentissage et de prise de décision. Nos méthodes s'inscrivent dans le cadre de l'optimisation décentralisée et de la théorie des jeux, en utilisant une approche d'apprentissage en ligne pour gérer la non-stationnarité de l'environnement.

Concrètement, nos premières contributions concernent l'étude d'un algorithme du type "dual averaging" dans l'apprentissage en ligne coopératif. Nous considérons pour ceci une configuration qui comporte de l'asynchronicité et des délais, présentant des obstacles à l'analyse classique du regret. Malgré cela, nous introduisons plusieurs concepts clés, dont les itérés virtuels et la permutation fidèle, qui nous permettent d'établir des bornes sur les regrets dans ce contexte. Nos résultats s'étendent également à une version optimiste du dual averaging, qui exploite la variation lente de la perte subie par les agents.

Ensuite, nous étudions la convergence vers les équilibres et la garantie de performance individuelle, mesurée par le regret, dans l'apprentissage dans les jeux. Le comportement ou la décision de chaque agent peut influencer les résultats des autres, créant une dynamique complexe qui doit être soigneusement analysée. Nos algorithmes sont à nouveau basés sur le principe optimiste, incorporant une étape de prévision qui réutilise l'information la plus récente.

Il est important de souligner que, dans les deux contextes, nous mettons l'accent sur l'adaptabilité de nos algorithmes et leur résilience face à l'incertitude lors des interactions. Nos méthodes fonctionnent sans aucune coordination entre les agents et peuvent être implémentées même par une entité qui ignore l'environnement avec laquelle elle interagit. Une particularité de notre approche est qu'elle fournit des garanties adaptatives, robustes face à la nature dynamique des environnements, où l'incertitude peut découler d'un manque de connaissance ou être modélisée, comme nous le faisons, avec un bruit dans la cadre de l'apprentissage dans les jeux.

ACKNOWLEDGMENTS

To begin with, I am profoundly grateful to my three Ph.D. advisors, Jérôme, Franck, and Panayotis. Their contant support, on personal, scientific, and administrative levels, has been truly invaluable to me over the past four and a half years. They granted me great flexibility during my thesis, allowing for a distinct Ph.D. experience with a year of "pré-thèse", two internships abroad, and six months of stay in Taiwan during the initial outbreak of COVID-19. They also afforded me the freedom to carve out my research direction while furnishing comprehensive backing from various perspectives.

Interestingly, my three advisors somehow assumed different roles during this journey. In a sense, Jérôme acted as a barrier between my academic pursuits and external distractions. From securing Ph.D. funding to acting as a guarantor for my apartment, his consistent support and kind advice were foundational pillars to the success of my Ph.D. Franck, on the other hand, felt more like a friend. His office at the end of the hallway was ever accessible, which I frequently passed by to discuss anything from conference reviews to weekend activities. On the scientific side, the content of this thesis is much related to the research direction of Panayotis. I have learned a lot from our discussion, his publications, and from the two game theory conferences he co-organized that I had the privilege to take part in. Of course, this brief overview hardly covers the multifaceted roles each of them played throughout this journey, but I shall better stop here for both the sake of space and for my inability to put all these in words.

Next, I extend my heartfelt appreciation to the two reviewers, Constantinos Daskalakis and Sylvain Sorin for having accepted to review my thesis. Their insightful feedback not only enhanced the quality of my manuscript but also deepened my understanding in several related subjects. Following that, I would like to express my sincere gratitude to the other esteemed members of the jury, Nicolò Cesa-Bianchi, Alexandre d'Aspremont, and Maryam Kamgarpour, as well as the president of the jury, Anatoli Juditsky, for dedicating their time and expertise to this pivotal process in my academic journey. Their individual works have been, and will undoubtedly continue to be, sources of profound inspiration for me. A special mention to Alexandre d'Aspremont, without whom this thesis would not have started; his introduction to my advisors for the Master 2 internship marked the very beginning of this adventure.

A particularity of my thesis is the inclusion of two enriching internships at Amazon. I wish to acknowledge my internship manager as well as other collaborators and colleagues that I met during these internships. Even though our research interests did not directly align, Shiva graciously brought me on board as an intern. This allowed us to delve into two fascinating research topics that I might not have explored otherwise. Brano, as an expert of the field, contributed greatly to these projects. From our initial interactions, Brano exhibited a genuine interest in our collaborations, even when there was not an

apparent obligation for him to be involved. I shall also mention my mentors Dominik Janzing and Patrick Bloebaum, the team's leader Yasser Jadidi, and fellow team members and interns including Jeff, Elke, Remi, Sergio, Luigi, and many others that I do not have the space to list here. I greatly benefited from their kindness and wisdom during these two internships.

Apart from the internships, I have been fortunate to interact and collaborate with several esteemed professors and researchers in other occasions over the past few years. First and foremost, I am sincerely grateful to Claire Vernade for her warm reception when I first reached out and for guiding me towards the internship opportunities at Amazon. Although our paths did not converge directly for collaboration, it was her initial suggestion that opened the door to the invaluable experiences and insights that I gathered in these internships.

During my stay in Taiwan in June 2022, I had the privilege of engaging with several great researchers in the fields including Yen-Huan Li, Ching-Pei Lee, and Prof. Jein-Shan Chen (let me also mention Prof. Hsuan-Tien Lin and Dr. Hsiang-Fu Yu which I met later in California during my second internship at Amazon). The insights and perspectives I received during these exchanges were truly enlightening. I am particularly indebted to Ching-Pei Lee and Prof. Jein-Shan Chen who extended the honor of hosting me as an invited researcher and offered me the opportunity to present my work at NTNU. Furthermore, interactions with other distinguished researchers, including Prof. Volkan Cevher, Prof. Georgios Piliouras, and Prof. Simon Lacoste-Julien have also helped me a lot in refining my academic perspective.

Equally important are the colleagues and friends who have journeyed alongside me during these years. First, I would like to express my gratitude to my fellow co-authors, Kimon and Yassine, for our collaborative efforts. I was fortunate to share my daily workspace with Anatole, Carlos, and Hubert, whose company made the routine more enjoyable. I truly appreciate it. I also want to thank the DAO members, including notably Dima, Florian, Gilles, Hamza, Mytia, Nils, Pierre-Louis, Roland, Sergeï, Sélim, Sylvain, Victor, Waïss, Yannis, as well as other colleagues in LJK, Alexandre, Alexis, Benji, David, Flora, Kliment, Manon, Margaux, Qiao, Rishabh, Thibault, Yunjiao and all those I inadvertently miss here, for creating such a positive atmosphere in the laboratory that was an integral part of this experience. There is another long list of friends which I met in different stages of my life and in different occasions that I will not enumerate here. From driving me around, letting me stay at their apartments during vacations, providing space for my belongings, to hiking and enjoying delightful meals together and having in-depth discussions in conferences, I am deeply grateful to all who recognize themselves in these words.

I am equally grateful for the myriad of ways I have been supported. My special thanks go to the administrative team of LJK, to Cultural Vista for facilitating my internship in the US, and notably to my landlords in France, Germany, and California, who eased my transitions in these places.

I have been intensively engaged in the Stable Diffusion community during this final year. I cherish the collaborations, discussions, and the projects that stemmed from our interactions. In a less conventional vein, I also want to thank some no-human entities, including Stable Diffusion for bridging my journey from theory to application, and ChatGPT for its help in many different tasks. I owe much to light novels / mangas / animes and the people behind them, as they have been a constant escape and relaxation for me.

I reserve the last words for my family, whom I dedicate this thesis to. Their unwavering love and support are paramount, and words cannot capture my gratitude. This achievement is as much yours as it is mine.

CONTENTS

PREFACE v

RÉSUMÉ vii

ACKNOWLEDGMENTS ix

1 INTRODUCTION 1
- 1.1 Philosophical Context and Scientific Positioning 1
- 1.2 Diagrammatic outline 4
  - 1.2.1 Part I: Learning in the Presence of Delays and Asynchronicities 5
  - 1.2.2 Part II: Adaptive Learning in Games 6
- 1.3 Works Not Included in This Thesis 9

PART I LEARNING IN THE PRESENCE OF DELAYS & ASYNCHRONICITIES 13

2 FUNDAMENTALS OF ONLINE OPTIMIZATION 15
- 2.1 Online Learning and Regret 15
- 2.2 Mirror Descent and Dual Averaging 17
  - 2.2.1 Regularizers, Bregman Divergences, and Mirror Maps 17
  - 2.2.2 Mirror Descent 20
  - 2.2.3 Dual Averaging 23
- 2.3 Adaptive Learning Rate 26

3 MULTI-AGENT ONLINE OPTIMIZATION WITH DELAYS 31
- 3.1 A Framework for Asynchronous Online Optimization 32
  - 3.1.1 Problem Setup 32
  - 3.1.2 Non-Monotonicity of Feedback Sequence and Lack of Synchronization 34
- 3.2 Delayed Dual Averaging and Faithful Permutations 35
  - 3.2.1 Delayed Dual Averaging 35
  - 3.2.2 Dependencies and Faithful Permutations 36
  - 3.2.3 Bounding the Regret of Delayed Dual Averaging 38
  - 3.2.4 Constant Learning Rate and Lags 39
- 3.3 Tuning the Learning Rate in the Presence of Delays 42
  - 3.3.1 Pessimistic Non-Adaptive Learning Rate 42
  - 3.3.2 Adaptation to Delays in Distributed Systems 44
  - 3.3.3 Adaptation to Unbounded Delays in the Single-Agent Setting 48
- 3.4 Multi-Agent Online Learning for Minimization of Global Losses 50
  - 3.4.1 From Effective Regret to Collective Regret 51
  - 3.4.2 Decentralized Delayed Dual Averaging 52
  - 3.4.3 A More Practical Learning Rate 55
- 3.5 Simulations in Static and Open Networks 57
  - 3.5.1 Problem Description 57
  - 3.5.2 Static networks 58
  - 3.5.3 Open networks 60

4 SLOW VARIATION AND THE ROLE OF OPTIMISM 63
- 4.1 Optimistic Gradient Descent 63

4.2 Delayed Optimistic Dual Averaging 66
4.2.1 Algorithmic Template and Regret Analysis 66
4.2.2 Necessity of Scale Separation for Robustness to Delays 69
4.3 Delayed Online Learning with Slow Variation 73
4.3.1 Full Information Setup and Choices of Guess Vectors 73
4.3.2 Adaptive Learning Rate 74

PART II ADAPTIVE LEARNING IN GAMES 79

5 FROM ONLINE LEARNING TO LEARNING IN GAMES 81
5.1 Learning in Games 81
5.1.1 Interaction Model, Regret, and Equlibrium 82
5.1.2 Adversarial Opponents and Self-Play 84
5.1.3 Assumptions on the Underlying Game 85
5.2 Variational Inequalities 86
5.2.1 Problem Formulation 86
5.2.2 Merit Functions 87
5.2.3 Variational Inequalities and Learning in Games 89
5.3 Algorithms 90
5.3.1 Optimistic Mirror Descent 90
5.3.2 Optimistic Dual Averaging 91
5.3.3 Optimistic Gradient as Approximate Projection onto Separating Hyperplane 92

6 LEARNING RATE ADAPTATION FOR GAMES WITH PERFECT FEEDBACK 95
6.1 Adaptive Learning Rate 96
6.2 A Family of Optimistic Methods 97
6.2.1 Compatibility with Dynamic Learning Rate 97
6.2.2 Example Algorithms 99
6.3 Optimal Regret Bounds 101
6.3.1 No-Regret Against Adversarial Opponents 101
6.3.2 Bound on Social Regret in Self-Play 102
6.3.3 Bound on Individual Regret in Self-Play 104
6.4 Trajectory Convergence 105
6.4.1 Reciprocity Conditions 105
6.4.2 Convergence to Best Response 106
6.4.3 Convergence to Nash Equilibrium 109
6.4.4 Adaptive OMWU Converges in Finite Two-Player Zero-Sum Games 111
6.5 Numerical Illustrations 115

7 DEALING WITH STOCHASTIC FEEDBACK I: TRAJECTORY CONVERGENCE 117
7.1 Feedback Model and Failure of Optimistic Methods 118
7.1.1 Noise Model 118
7.1.2 Non-convergence of EG and OG with Stochastic Feedback 119
7.2 Learning Rate Separation and Energy Inequalities 120
7.2.1 Learning Rate Separation as a Remedy: EG+ and OG+ 120
7.2.2 Generalized OG+ 122
7.2.3 Inequalities for EG+ 122
7.2.4 Inequalities for OG+ 124
7.3 Global Convergence 128
7.3.1 Asymptotic Convergence 128

7.3.2 Convergence Rate 135
7.4 Local Convergence 140
7.4.1 Stability of Equilibrium 140
7.4.2 Asymptotic Convergence 146
7.4.3 Convergence Rate 148
8 DEALING WITH STOCHASTIC FEEDBACK II: NO-REGRET AND ADAPTIVE LEARNING 151
8.1 Preliminary Inequalities 152
8.1.1 Generalized OptDA+ 153
8.1.2 Inequalities for OptDA+ 154
8.2 OptDA+ with Predetermined Learning Rates 159
8.2.1 No-Regret Against Adversarial Opponents 159
8.2.2 Fast Convergence of Pseudo-Gradient in Self-Play 161
8.2.3 Improved Regret in Self-Play 163
8.2.4 Convergence to Equilibrium under Multiplicative Noise 165
8.3 OptDA+ with Adaptive Learning Rates 167
8.3.1 Adaptivity in the Face of Noise 167
8.3.2 Preliminary Lemmas 169
8.3.3 No-Regret Against Adversarial Opponents 170
8.3.4 Fast Convergence of Pseudo-Gradient in Self-Play 171
8.3.5 Improved Regret in Self-Play 177
8.3.6 Convergence to Equilibrium under Multiplicative Noise 179
8.4 Numerical Illustrations 181
8.4.1 A Bilinear Zero-sum Game 181
8.4.2 Linear Quadratic Gaussian GAN 182

PART DISCUSSION AND CONCLUDING REMARKS 185
9 CONCLUSION AND PERSPECTIVES 187

BIBLIOGRAPHY 193
APPENDIX 209
A BREGMAN DIVERGENCES, MIRROR MAPS, AND FENCHEL COUPLINGS 211
B TECHNICAL LEMMAS ON NUMERICAL AND STOCHASTIC SEQUENCES 215
C FROM PSEUDO-REGRET TO EXPECTED REGRET 221
D LIST OF PUBLICATIONS 223
INDEX 225

1

INTRODUCTION

Existentialists posit that every decision we make contributes to defining our humanity. But what about the decisions made by machines? This philosophical contemplation largely bypasses the scope of this thesis, but one thing is certain. When Sartre wrote down his famous quote "Hell is other people" in "No exit", it is unlikely he could have envisaged that it might one day apply to a machine as well.[1] This complexity of a machine making decisions in the context of interaction with others forms the core subject of this thesis, where we approach it from a mathematical standpoint.

1.1 PHILOSOPHICAL CONTEXT AND SCIENTIFIC POSITIONING

As we stand at the dawn of a new era, the world around us is changing at an unprecedented pace. News of advancements in artificial intelligence flood the media, sketching a vivid picture of a future society teeming with intelligent agents. Soon, factories will become places where robots with unique intelligence collaborate on complex tasks. Non-player characters in games will operate without manual programming, and autonomous vehicles will navigate our city streets. Looking ahead, we can envision a future where every moving object will be infused with artificial intelligence.

This perspective is a thrilling one, but it is not without its apprehensions. The deployment of such intelligent agents at a large scale could bring about unforeseen impacts, from mass unemployment to an exacerbated wealth disparity. Some have even gone so far as to characterize the risk posed by these developments as "existential". Amid such uncertainties, it becomes crucial to approach the design of these entities with caution and strategic forethought. Given the multi-agent nature of the problem, understanding the interplay and mutual influence between individual agents is of paramount importance, as it is these interactions that often shape the overall dynamics and outcomes of the entire system.

Taking one step back, the study of such complex systems is deeply rooted in various disciplines, where the interacting entities could represent individuals in society or populations in nature. Mathematical models have been developed to explain phenomena observed in these contexts. On top of this is the advent of the Internet of Things (IoT), which marks the first time in human history that so many people can interact with each other in a common place. This massive scale of interaction is nonetheless heavily influenced by algorithms. In online ad auctions, bidders rely on automated bidding agents to place their bids. On social media platforms, the content shown to us is largely determined by underlying machine learning models.

1 "L'enfer, c'est les autres" from "Huis clos" [241] in French. Note that here we are neither using it is popular (mis)interpretation nor adhering to Sartre's original meaning of the phrase.

In parallel, there are systems like sensor networks and robots that may be physically dispersed yet require cooperation to function effectively. Server farms spread across geographical distances need to work in unison, as do devices in a smart home environment. In either case, the challenge is not just explaining the outcome of interactions, but also designing new entities—algorithms or AI agents—to efficiently solve specific tasks or handle more general situations.

This thesis delves into this crucial matter, focusing on algorithm design in *decentralized* multi-agent systems. Decentralization, in this context, is not just a technical challenge or an inevitability, but also a philosophical choice. In this era of AI, ensuring that the power of intelligence is within everyone's reach is vital, as this fosters data independence, a concept as crucial for digital consumers as energy independence is for sovereign countries. The ultimate goal is to design algorithms that operate from an individual's perspective but contribute to both the benefit of the individual and the harmony of the whole system.

Keeping this in mind, this thesis is built on three critical pillars: *multi-agent systems*, *uncertainty*, and *adaptivity*. We aim to study systems where multiple agents interact under uncertain conditions and adapt their behaviors in response to their observations, as we detail below.

MULTI-AGENT SYSTEMS. At the heart of the thesis is the investigation of decision-making, or learning, within multi-agent systems. This brings to light a multitude of challenges, both practical and conceptual. In this manuscript, we focus on mathematical frameworks that capture certain aspects of these challenges. Particularly, we delve into two distinct yet complementary topics: the practical challenges that stem from delays and the conceptual challenges related to strategic interactions in competitive environments.

Asynchronicity and delayed feedback

Such challenges often arise when attempting to deploy multi-agent systems in real-world environments: notably, we have to contend here with asynchronicity and delayed feedback. Asynchronicity can pertain to the timing of agent activations, inter-agent communications, local computations, and more. It is an unavoidable characteristic, often resulting from factors like heterogeneous computational resources, varying communication channels, or simply the unplanned nature of real-world environments.

Meanwhile, delayed feedback refers to the situation where the feedback received by an agent regarding some event occurring in the network arrives after a certain delay, a natural characteristic of multi-agent systems due to network latency. These challenges necessitate the development of learning algorithms that can effectively handle asynchronicity and delay with minimal performance sacrifice. For some examples of such algorithms, we direct readers to [12, 135, 231, 281] and references therein.

Cooperative learning and distributed optimization

In Part I of this thesis, we investigate this problem in the realm of cooperative learning. Here, the agents share a common objective. A practical example of this could be a fleet of robots in factories or a cluster of servers in cloud computing. In this context, a seemingly elementary but already challenging problem to address is distributed optimization of the following form

$$\min_{x \in \mathcal{X}} \frac{1}{N} \sum_{i=1}^{N} \ell^i(x). \tag{1.1}$$

In the above, each ℓ^i is an agent's local loss function and the agents aim to find a single variable $x_\star$ from the shared constraint set $\mathcal{X}$ that minimizes the sum of these loss functions. The study of this problem can be traced back at least

to the works [64, 268, 269] in the context of parallel and distributed numerical methods; see also the textbook [19] and the surveys [206, 285]. The key obstacle here comes from the need for communication between agents: in the absence of communication, the best that each agent can achieve is the minimization of their own loss functions. This, however, does not in general translate into a solution for (1.1). Conversely, if all agents could communicate instantaneously, the problem would reduce to a single-agent optimization one where we can directly execute operations with respect to the sum of the functions. In practice, the situation is usually somewhere in between. Agents communicate in a constrained manner, with communications being local and delayed. It is within these circumstances that the complexities of asynchronicities and delays, as outlined earlier, become significantly pronounced and warrant the need for a specialized treatment.

More concretely, we examine in Part I an extension of the above problem, where neither the number of agents nor their loss functions are fixed. Instead, they arrive in an *online* fashion, and we seek algorithms with minimal *regret*, generally defined as the difference between the total loss experienced by the agents and that incurred by some comparator strategy.

Strategic Interactions and Interdependencies

On the other hand, the very nature of multi-agent systems can fundamentally change the decision-making problem. Consider for instance the bidding dynamics in auctions or the path selection process in network routing. In these scenarios, the loss experienced by an agent is a function of both its own actions and the actions of others. As a consequence, the complexity in these situations does not merely originate from the communication process, but also stems from the intricate interplay of the agents' actions in shaping their losses. The question that arises is: can we still draw any insights about the agents' behaviors amidst such layered intricacies?

This is where game theory enters the scene. From its early attempts to elucidate human rationality to its modern adaptations for analyzing evolving populations in biology and computer networks, game theory offers a mathematical toolbox tailored to the complex interactions between agents [90, 166, 217].

Learning in games

What is particularly relevant to our study is the dynamics in games, an area explored in various subfields of game theory including evolutionary game theory [239, 278], repeated games [190], and algorithmic game theory [214]. Our concerns predominantly lie within the paradigm of learning in games, which we review and study in Part II of this manuscript. Our aim is to identify algorithms that, in this context, can ensure stability of the system while also providing at least some minimal performance guarantee for each agent, as measured by their regrets.

UNCERTAINTY. From the perspective of each individual agent, the landscape of multi-agent systems is riddled with uncertainties. These uncertainties stem from a variety of sources. For instance, the inherent unpredictability of the environment, the volatile nature of other agents' behaviors, and/or the lack of complete knowledge about the system, all contribute to an uncertain landscape. While some of these uncertainties are hard-coded in the multi-agent nature of the model, others require a dedicated treatment. Among these, we single out two types of uncertainties: those brought about by *delays*, and those modeled by *stochastic feedback*.

Uncertainty related to delays

As previously noted, the delayed and asynchronous nature of communication among agents presents a unique set of challenges. In Chapters 3 and 4, we dive into the complexities caused by these delays. Uncertainty in this context manifests in several forms—from an inability to ascertain the number

of decisions or updates that have been enacted within the system, to a lack of precise knowledge about when feedback was generated.[2] These aspects add an extra layer of complexity, over the inherent challenge posed by agents having to operate with outdated information.

Uncertainty related to stochastic feedback

On another front, we examine uncertainty related to the value of feedback in Chapters 7 and 8. This uncertainty can be intrinsic to the problem, be caused by measurement or transmission errors, or be deliberately injected into the algorithm for efficiency or privacy reasons. To represent these, we incorporate an unbiased noise component to the true, noiseless feedback. This leads to recursive approaches commonly studied in stochastic approximation [136, 233, 276]. The integration of multi-agent dynamics with stochastic feedback, however, presents new challenges that are not encountered when dealing with each of them separately.

ADAPTIVITY. One crucial, yet challenging, element for effective algorithms in multi-agent systems lies in the need for these algorithms to be practically implementable. What qualifies as such largely depends on the problem at hand, but a universal criterion we strive for is that the algorithm should operate with minimal knowledge and little to no coordination between agents. Furthermore, an ideal algorithm should leverage the distinct facets of the interactions to improve its performance.

Adaptivity from a theoretical viewpoint

From a theoretical perspective, we can either capture this with bounds that reflect some fine-grained characterization of the interaction process, or by establishing a collection of optimal guarantees for different situations. In this context, "adaptive" functions as an umbrella term signifying that an algorithm meets at least some of the aforementioned criteria.

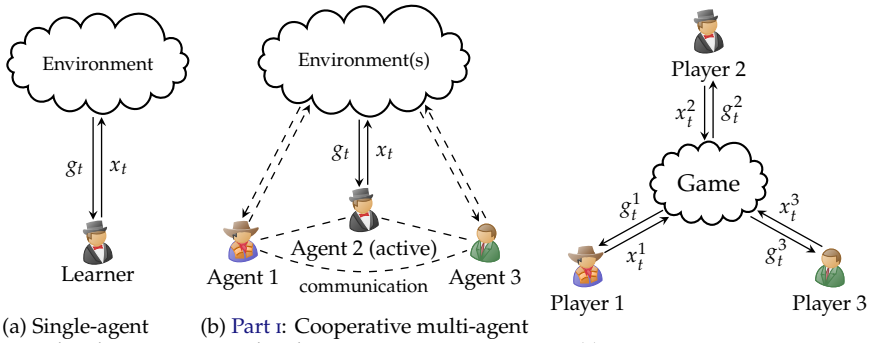
A significant part of this thesis is devoted to the study of such adaptive algorithms. We investigate their application in various settings in Chapters 3, 4, 6 and 8. While the algorithms that we consider follow a similar guiding philosophy, as elaborated in Section 2.3, their actual adaptation and the accompanying analysis is far from straightforward. This exploration results in a rich body of results that furthers our understanding of the effectiveness of adaptive learning within multi-agent systems.

1.2 DIAGRAMMATIC OUTLINE

Following the above discussion, the body of this thesis is divided into two parts, each of them dedicated to a different model for multi-agent decision-making. In Part I, we extend the distributed optimization problem described in Eq. (1.1) to an asynchronous, online setup and devise adaptive algorithms to address this situation. In Part II, we place ourselves in the framework of learning in games and investigate how we can make the so-called *optimistic* algorithms adaptive and robust to noise. For the sake of illustration, we depict these two frameworks along with the basic single-agent online learning setup in Fig. 1.1.

Each of the two parts of this thesis begins with a chapter that lays out the necessary technical preliminaries. The subsequent chapters of each part build upon these foundations and present novel contributions to the field. Finally, in Chapter 9, we conclude the manuscript by exploring some potential avenues for future research.

2 While timestamping each piece of feedback is plausible, it still leaves us unsure of how to order it relative to the feedback elements that have not yet reached the agent.



(a) Single-agent online learning

(b) Part I: Cooperative multi-agent online learning

(c) Part II: Learning in games

Figure 1.1: Illustrations of the learning setups considered in this thesis. The two multi-agent setups are natural extensions of single-agent online learning and depict respectively the cooperative and the competitive scenarios. Note that only one agent is active per round in b and hence the dashed arrows between the inactive agents and the environment(s).

1.2.1 *Part I: Learning in the Presence of Delays and Asynchronicities*

We use the framework of online learning for studying sequential decision making (see Fig. 1.1a). This first part looks into the difficulties that arise from delayed and asynchronous communications. The agents have a shared objective: the minimization of their *joint regret*. Simply put, this is the cumulative regret incurred over the entirety of the system, thereby serving as a benchmark for the overall performance of the decision-making process. However, achieving this objective is far from straightforward: it requires the design and analysis of robust, adaptive algorithms that can navigate the complexities of delays and asynchronicities, while still delivering optimal performance guarantees within these multifaceted settings.

The contributions presented in this part draw from the following publications.

> Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Optimization in Open Networks via Dual Averaging. In *IEEE Conference on Decision and Control*, 2021.
>
> Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Multi-agent Online Optimization with Delays: Asynchronicity, Adaptivity, and Optimism. *Journal of Machine Learning Research*, 2022.

Fundamentals of Online Optimization

CHAPTER 2. In this introductory chapter, we establish the mathematical groundwork for the understanding of online learning, with a particular emphasis on online convex optimization. In our pursuit, we spotlight two central algorithms—mirror descent and dual averaging—discussing their difference and stressing instances where the latter might prove more advantageous than the former. Our main objective here is to build an understanding of the basics, set up the notations, and underscore important elements of the regret analysis. This is why we include proofs of several propositions, even though the results presented in this chapter are rather standard. As the chapter concludes, we present AdaGrad-norm, an adaptive learning rate strategy well studied in the literature. This strategy sets the basis for the design of our adaptive algorithms, which are discussed in later parts of the thesis.

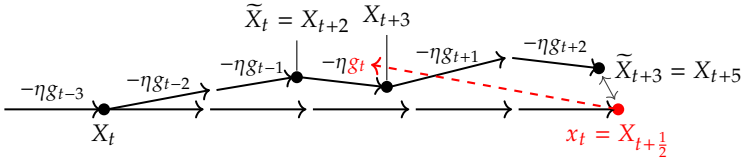


Figure 1.2: Schematic representation of (DOptDA). This algorithm introduced in Chapter 4 addresses delays with an optimistic step that extrapolates the base state X_t to the actual prediction x_t. The extrapolation is meant to compensate missing feedback and allows the algorithm to "look into the future" when the feedback sequence varies slowly. More detailed explanation of this schema is provided in Section 4.2.1

Multi-Agent Online Optimization with Delays

Chapter 3. Formulating a mathematical framework that captures simultaneously the online, multi-agent, and asynchronous aspects is not an easy endeavor, and this is what we aim to provide in this chapter. Our proposed framework models the cooperation of asynchronously communicating agents within a potentially time-varying environment (cf. Fig. 1.1b). To accommodate this challenging situation, we extend dual averaging to cope with delayed feedback, and provide in Theorem 3.1 a general regret bound for learning rate sequences that are non-increasing along a *faithful permutation*, a permutation of the time indices that is compatible with the feedback structure (Definition 3.1).

Going beyond this general result, we design adaptive learning rates that are practically implementable by the agents under different constraints. These algorithms mostly satisfy the minimum prior knowledge and no coordination desiderata outlined earlier, all the while enjoy guarantees that reflect both the magnitude of feedback and the actual delays. We conclude this chapter by drawing a connection with distributed online learning through alternative approaches to problem formulation and regret definition. Additionally, we provide experimental results for the solution of a distributed least absolute deviation problem in static and open networks.

Slow Variation & The Role of Optimism

Chapter 4. We transition our focus in this chapter from the harsh adversarial landscapes of loss sequences to more benevolent environments where identifiable patterns and slower changes offer opportunities for an improved regret. Optimistic algorithms, originally developed for such conditions, play a significant role in our strategy. After revisiting these algorithms, we introduce adaptations to address the impact of delays. In particular, we find it necessary to place more weight on the optimistic step to offset the effects of delayed feedback, as we illustrate in Fig. 1.2. These adaptations are underpinned by lower bounds, which formally validate the need for such a "learning rate separation" and certify the optimality of our regret bounds. In a practical vein, we develop adaptive algorithms specifically designed for the multi-agent setup introduced in Chapter 3, and provide theoretical guarantees that demonstrate their efficacy in this context. Lastly, it is important to highlight that this chapter also serves as a prelude to Part II, where we rely heavily on optimistic algorithms.

1.2.2 *Part II: Adaptive Learning in Games*

In this part, we shift our focus to learning in games, focusing particularly on learning in continuous games with first-order feedback. While each agent, or player, is still engaged in an online learning problem, their loss functions are now dictated by the actions of other players, rather than some ambiguous external "nature". This transition opens the door for us to devise algorithms

that provide enhanced performance, especially when players exhibit rational and non-adversarial behavior.

Notably, unlike the previous part of the manuscript, which explored the complexities introduced by delays, we consciously put aside this aspect here to fully immerse ourselves in the challenges and opportunities unique to the learning-in-games context. Furthermore, this choice also allows us to focus on the refinement of the algorithms, improving their adaptability across various scenarios. We particularly look into the adversarial, self-play, and stochastic feedback settings. This investigation would have been hard to achieve had we further attempted to take the delays into account.

The development of this part lays down on the following works.

> Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore Aggressively, Update Conservatively: Stochastic Extragradient Methods with Variable Stepsize Scaling. In *Advances in Neural Information Processing Systems*, 2020.
>
> Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive Learning in Continuous Games: Optimal Regret Bounds and Convergence to Nash Equilibrium. In *Conference on Learning Theory*, 2021.
>
> Yu-Guan Hsieh, Kimon Antonakopoulos, Volkan Cevher, and Panayotis Mertikopoulos. No-Regret Learning in Games with Noisy Feedback: Faster Rates and Adaptivity via Learning Rate Separation. In *Advances in Neural Information Processing Systems*, 2022.

From Online Learning to Learning in Games

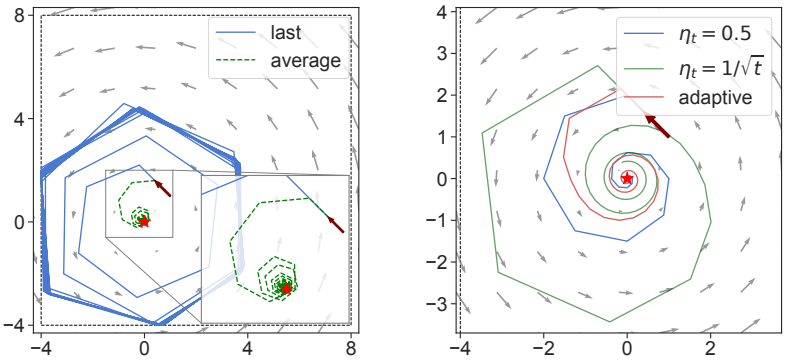
CHAPTER 5. This chapter serves as a gentle introduction to the learning-in-continuous-games setup that we study. After presenting the basic definitions and underlying assumptions, we relate it to variational inequalities to showcase the potential of our approach. The chapter concludes with a more in-depth examination of optimistic algorithms, highlighting their improved regret guarantees, convergence properties, and a geometric intuition behind the methods.

Learning Rate Adaptation for Games with Perfect Feedback

CHAPTER 6. The first challenge we aim to address is the need for delicately tuning the learning rate for the algorithms to work properly. To account for this, we propose in this chapter a set of no-regret policies based on optimistic mirror descent and optimistic dual averaging, bearing the following advantageous properties:

1. They do not necessitate any prior tuning or knowledge of the game.
2. They all deliver $O(\sqrt{T})$ regret against arbitrary, adversarial opponents.
3. They converge to the best response against convergent opponents.
4. If employed by all players, they guarantee $O(1)$ *social* regret, while the resulting sequence of play converges to a Nash equilibrium with $O(1)$ *individual* regret in all *variationally stable* games.

These guarantees shed light on the adaptivity of the method and mark our initial attempt at enhancing the practical applicability of these algorithms in real-world settings. Importantly, the convergence of the algorithm, as formally proved in Theorem 6.9, demonstrates the benefit of adaptive learning rates beyond regret minimization. Such result is less common in the literature. In Fig. 1.3, we illustrate this convergence, and showcase how the algorithm performs under various choices of the learning rate in a bilinear game.



(a) (OptDA) with constant learning rate $\eta_t \equiv 0.6$ causes played iterate to diverge and average iterate to converge to spurious point.

(b) Convergence of (OptDA) with suitable learning rate. Adaptive method converges faster without requiring knowledge about the game.

Figure 1.3: The trajectories of play (and the time-average of one of these trajectories) obtained by running (OptDA) with a quadratic regularizer on the game $\min_{\theta\in[-4,8]} \max_{\phi\in[-4,8]} \theta\phi$ using different learning rates. The algorithm is presented in Chapter 5 and the adaptive learning rate for this setup is introduced in Chapter 6.

Dealing with Stochastic Feedback I: Trajectory Convergence

Chapter 7. In this chapter and the subsequent one, we tackle the additional complexity brought about by stochastic feedback. To do so, we consider a unified noise model that captures both additive and multiplicative noise components. The primary objective of our study is to extend the theoretical guarantees of the methods to these cases where the feedback is subject to such unpredictable perturbations. Convergence and regret of the algorithms that we examine in these two chapters are illustrated in Fig. 1.4.

This chapter, in particular, delves into the convergence of the sequence of play in such scenarios. We focus our attention here on two distinct algorithms: extragradient and optimistic gradient. Through a counterexample, we first show that these methods fail to converge in the presence of noise. To overcome this limitation, we propose modifications that, echoing our approach in Chapter 4, employ two distinct learning rate sequences, wherein the update step is smaller relative to the optimistic step. This helps to mitigate the effects of noisy feedback and recover the benefits of the optimistic step. Formally, we establish the almost sure last-iterate convergence of these revised algorithms in variationally stable games, complete with convergence rates under an additional error bound condition. We also introduce localized versions of these results, enabling us to bypass the stringent global assumptions. These findings all together indicate the resilience of our methods in stochastic environments.

Dealing with Stochastic Feedback II: No-Regret and Adaptive Learning

Chapter 8. Having introduced the necessary modifications for adaptive algorithm design in Chapter 6 and for dealing with noisy feedback in Chapter 7, we are naturally led to a question: can these two elements be seamlessly combined? We set out to answer this question in this final chapter of Part II. To this end, we introduce OptDA+, a double-learning-rate variant of optimistic dual averaging, and analyze its regret and convergence behavior.

A particular focus of this chapter is the scenarios that involve multiplicative noise exclusively. In this situation, the noise scales with the feedback, thus offering an opportunity for improved performance. Specifically, we establish constant bounds on both regret and the sum of squared-gradient norms, both of which serve as measures for the players' performances. Compared to the

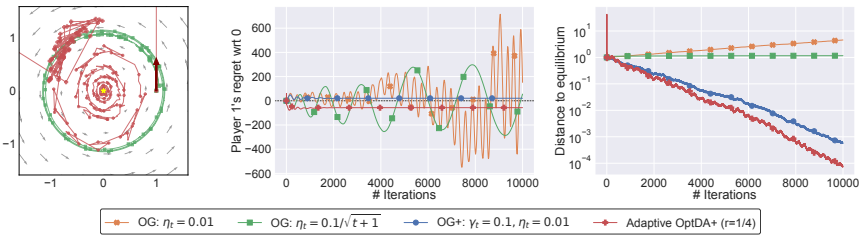


Figure 1.4: The behaviors of different algorithms on $\min_{\theta \in \mathbb{R}} \max_{\phi \in \mathbb{R}} \theta\phi$ when the feedback is corrupted by (multiplicative) noise. Left: trajectories of play. Center: regret of Player 1 with respect to 0. Right: distance to equilibrium. (OG+) and adaptive (OptDA+) are respectively introduced in Chapters 7 and 8. These algorithms achieve convergence in stochastic environments through "learning rate separation".

algorithms studied in Chapter 7, OptDA+ allows each player to use different learning rates, but we also note that we only manage to prove its almost sure last-iterate convergence when noise is multiplicative. Pushing this frontier even further, we develop an adaptive version of the algorithm that achieves these guarantees automatically. This does not necessitate player coordination or prior knowledge of the game or noise profile, thereby culminating our journey toward robust, adaptive learning in games.

1.3 WORKS NOT INCLUDED IN THIS THESIS

Throughout my Ph.D., I have had the privilege to work on various research projects. However, for the sake of coherence in this manuscript, I have chosen to leave out four of these projects, even though they were developed within the context of my thesis. Brief summaries of these projects are provided below, and a complete list of my publications is included in Appendix D.

On the convergence of single-call stochastic extra-gradient methods

SINGLE-CALL STOCHASTIC EXTRA-GRADIENT. In this project, we developed a synthetic view of *single-call extra-gradient methods* and provided analyses thereof. First, we demonstrated these methods retain the $O(1/t)$ ergodic convergence rate of the two-call methods in smooth, deterministic problems. Subsequently, we showed that this rate is also achieved by the last iterate of the algorithms in stochastic variational inequalities with strongly monotone operators provided that the optimizer has access to an oracle with bounded variance. Finally, we derived a high-probability $O(1/t)$ local convergence rate to solutions of non-monotone variational inequalities that satisfy a second-order sufficient condition.

This project was carried out with my Ph.D. advisors in an internship prior to my Ph.D. It can be regarded as a prelude to the problems that we address in Part II, though the distinction between the different single-call methods as outlined in this work is not present in the current manuscript.

The publication associated to the project is:

> Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the Convergence of Single-Call Stochastic Extra-Gradient Methods. In *Advances in Neural Information Processing Systems*, 2019.

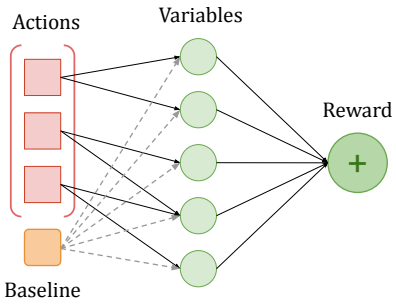


Figure 1.5: Illustration of the uplifting bandit model. This example has 3 arms, 5 variables, and each arm affects 2 variables. Dash lines indicate the variables' payoffs follow the baseline distribution by default.

Push–pull with device sampling

PUSH–PULL WITH DEVICE SAMPLING. In this project, we considered decentralized optimization with the objective of minimizing the average of the agents' local functions, as expressed by (1.1). We were interested in an asynchronous model where a random portion of agents performed computations at each iteration, while the information exchange could be conducted between all the nodes and in an asymmetric fashion. For this setting, we proposed an algorithm that combined gradient tracking and variance reduction over the entire network, enabling each node to track the average of the gradients of the objective functions. The effectiveness of the algorithm was validated both theoretically, through a dedicated analysis for the strongly convex setup under mild connectivity conditions on the network, and numerically, through experiments on synthetic and real-world datasets.

Importantly, this project offers a unique complement to the results presented in Part I of the manuscript, in both the design of the algorithms (gossip-based as opposed to simple aggregation of feedback) and the setup that we operate with (offline as opposed to online). Nevertheless, both situations feature challenges related to asynchronicity and communication that are specific to multi-agent systems. In the realization of this project, I had the pleasure to work with the former Ph.D. student Yassine Laguel from our team. The fruits of our collaboration were published in the following journal paper:

> Yu-Guan Hsieh, Yassine Laguel, Franck Iutzeler, and Jérôme Malick. Push–Pull with Device Sampling. *IEEE Transactions on Automatic Control*, 2023.

Uplifting bandits

UPLIFTING BANDITS. In this project, we developed a new stochastic multi-armed bandit model, where the reward is a sum of multiple random variables, and each arm only alters the distributions of some of them. We visualize this in Fig. 1.5. After each action, the agent observes the realizations of all the variables. Our model finds applications in areas like marketing campaigns and recommender systems, where the variables could represent outcomes on individual customers, such as clicks. We explored several variations of the problem, including scenarios where the baseline and affected variables are unknown. We designed UCB-style algorithms that estimate the uplifts of the actions over a baseline, and proved sublinear regret bounds for all variations. We also established lower bounds to justify the necessity of our modeling assumptions. Furthermore, we conducted numerical simulations to underscore the benefit of our approach.

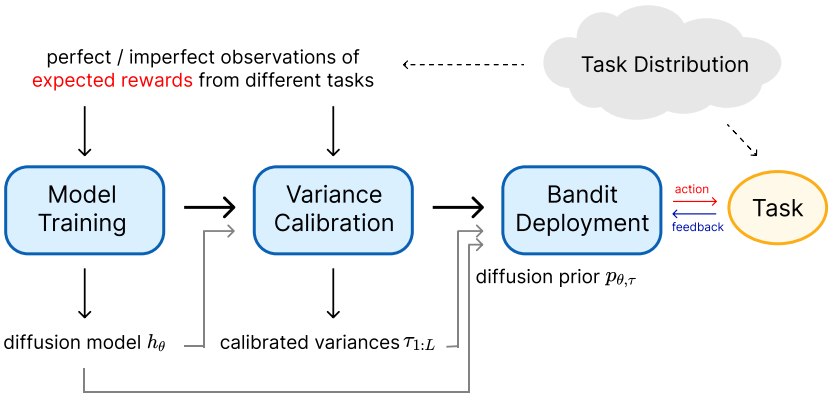


Figure 1.6: Overview of the meta-learning for bandits with diffusion prior framework.

This project was conducted during my internship at Amazon AWS in Tübingen, Germany. The focus on stochastic bandits here offers an interesting contrast to this thesis, which exclusively tackles adversarial scenarios and learning in games with first-order feedback. A natural question that arises is: Can we leverage the structure of the problem in a similar manner under adversarial settings? This project culminated in the publication of the following article:

> Yu-Guan Hsieh, Shiva Kasiviswanathan, and Branislav Kveton. Upliting Bandits. In *Advances in Neural Information Processing Systems*, 2022.

Thompson sampling with diffusion generative prior

THOMPSON SAMPLING WITH DIFFUSION GENERATIVE PRIOR. In this project, we initiated the idea of using denoising diffusion models to learn priors for online decision-making problems, with a particular focus on the meta-learning for bandit framework. The entire procedure is illustrated in Fig. 1.6. Precisely, our goal was to learn a strategy that consistently performs well across bandit tasks within the same class. To accomplish this, we trained a diffusion model to learn the underlying task distribution and combined this with Thompson sampling to deal with new tasks at test time. Our posterior sampling algorithm was meticulously designed to strike a balance between the learned prior and the noisy observations resulting from the learner's interaction with the environment. In order to accommodate realistic bandit scenarios, we also introduced a novel diffusion model training procedure that can learn from incomplete and/or noisy data. The potential of our proposed approach is confirmed by our experimental evaluations on both synthetic and real-world datasets.

This project was realized during my internship at Amazon AWS in Santa Clara, USA. The experimental focus and algorithm design emphasis of this project, alongside the incorporation of cutting-edge deep generative models, allowed me to view decision-making problems through a different lens. This viewpoint, while distinct, provides a valuable counterpoint to the more theoretical explorations of decision-making in this thesis. The research outcomes of this project are published at:

> Yu-Guan Hsieh, Shiva Kasiviswanathan, Branislav Kveton, and Patrick Bloebaum. Thompson Sampling with Diffusion Generative Prior. In *International Conference on Machine Learning*, 2023.

Part I

LEARNING IN THE PRESENCE OF DELAYS & ASYNCHRONICITIES

2

FUNDAMENTALS OF ONLINE OPTIMIZATION

We embark on our journey by delving into the realm of online learning, a versatile paradigm for sequential decision-making that finds applications in fields as diverse as portfolio selection, online auctions, and recommender systems, among others [30, 112, 246]. This approach embraces the dynamic and time-varying nature of the decision-making process, operating within environments where cost functions change over time. Furthermore, online learning encapsulates the view of learning as a continuous process, and pushes the boundary of distribution-free results.

Following this introductory perspective, we zoom in on the area of online convex optimization in this chapter. While the scope of this field is vast, we aim to illuminate its essential elements that are relevant to our study, providing a compact yet comprehensive overview. The specific framework and the corresponding algorithms we introduce here lays the groundwork for the multi-agent settings discussed in the later chapters of this thesis.

OUTLINE OF THIS CHAPTER. This chapter unfolds as follows. In Section 2.1, we formulate the framework of online optimization and elaborate on the concept of regret. Proceeding to Section 2.2, we turn our attention to two central algorithms in online convex optimization: *mirror descent* and *dual averaging*. We present their regret guarantees and furnish the necessary tools for proving these results. Finally, in Section 2.3, we look into adaptive learning rates. They hold a special place in online learning, as they offer anytime and data-dependent regret bounds, thereby enhancing the robustness and adaptability of the learning algorithms.

2.1 ONLINE LEARNING AND REGRET

Online learning

In its most general form, online learning is characterized as a series of repeated interactions between a learner and the environment. This process is outlined in Fig. 2.1. At each round t, the learner selects an *action* x_t from their *action set* $\mathcal{X}$. Subsequently, the environment reveals a *loss function* $\ell_t : \mathcal{X} \to \mathbb{R}$, and the learner suffers *loss* $\ell_t(x_t)$. It is important to note that the exact form of the loss function is generally unknown to the learner before the action selection and may even be chosen in an adversarial manner.

Adversarial setup

Specifically, in what is termed the *adversarial setup*, we do not make any statistical assumptions about these loss functions. In other words, we do not assume any probabilistic model for the generation or selection of these functions. This framework simulates the hardest case scenario for the learner, where the sequence of loss functions could be strategically chosen to maximize the learner's loss.

Regret

In this challenging setting, the standard measure of performance is *regret*, which offers a comparison between the learner's accumulated loss and the total

At each round $t = 1, 2, \ldots$, the learner

- Plays an action $x_t \in \mathcal{X}$
- Suffers loss $\ell_t(x_t)$ and receives corresponding feedback

Figure 2.1: The online learning framework.

loss that the learner would have incurred if a specific fixed action had been consistently chosen throughout the process.

Definition 2.1 (Regret)**.** For any comparator action $z \in \mathcal{X}$ and time horizon $T \in \mathbb{N}$, the *regret* of the learner relative to z after T rounds is defined as

$$\mathrm{Reg}_T(z) = \sum_{t=1}^{T} [\ell_t(x_t) - \ell_t(z)].$$

The notion of regret has its roots in the seminal work of Blackwell [20] and Hannan [108]. In a wider sense, regret is just a metric that compares the losses suffered by the learner against the losses that would have been incurred by a certain baseline. This conceptual framework has given rise to a plethora of performance measures, including but not limited to internal regret [21, 86], where the base line is a mapping of the actions, dynamic regret [105, 298], where the baseline is a sequence of actions that evolves over time, and adaptive regret [113], where we consider all the time intervals with baselines as fixed actions over these time intervals. In this context, the specific form of regret defined in Definition 2.1 is often called "external regret". This is the principal variant of regret that we employ in this thesis. For the sake of brevity, we will henceforth refer to it simply as "regret".

Online convex optimization

ONLINE CONVEX OPTIMIZATION. For the scope of this thesis, we narrow down our attention to the framework of online convex optimization. In this setting, $\mathcal{X}$ is a closed convex subset of an ambient finite-dimensional vector space $\mathbb{R}^d$, and the loss functions are always convex and subdifferentiable, as we state below.

Convexity of action set and losses

Assumption 2.1 (Convexity, closeness, and subdifferentiability)**.** The action set $\mathcal{X} \in \mathbb{R}^d$ is convex and close. Moreover, for all $t \in \mathbb{N}$, the loss function ℓ_t is convex and subdifferentiable.

First-order feedback

In terms of feedback, we consider that the learner receives at the end of each round t a subgradient vector $g_t \in \partial \ell_t(x_t)$ evaluated at the selected point.[1] This is possible as long as Assumption 2.1 is satisfied. With this, we may consider the "linearized" loss $\tilde{\ell}_t \colon x \to \ell(x_t) + \langle g_t, x - x_t \rangle$. When g_t is the gradient of ℓ_t at x_t, this is exactly the first-order approximation of ℓ_t at x_t. We then define the *linearized regret* as the regret with respect to this sequence of losses.

Linearized regret

Definition 2.2 (Linearized regret)**.** For any comparator action $z \in \mathcal{X}$ and time horizon $T \in \mathbb{N}$, the *linearized regret* of the learner relative to z after T rounds is defined as

$$\mathrm{LinReg}_T(z) = \sum_{t=1}^{T} \langle g_t, x_t - z \rangle.$$

1 In a slight abuse of terminology, the terms gradient and subgradient will be used interchangeably in the sequel.

The importance of the convexity assumption lies in that it ensures the linearized regret to be an upper bound on the actual regret. In fact, by the definition of subgradient, it holds that

$$\ell_t(x_t) - \ell_t(z) \leq \langle g_t, x_t - z \rangle.$$

Therefore, summing this up over t, we get the following lemma.

Regret upper bounded by linearized regret

Lemma 2.1. *Suppose that Assumption 2.1 holds. Then, for any sequence of actions $(x_t)_{t\in\mathbb{N}}$ and any comparator action $z \in \mathcal{X}$, we have*

$$\mathrm{Reg}_T(z) \leq \mathrm{LinReg}_T(z).$$

Lemma 2.1 significantly simplifies the analysis as we essentially reduce the problem of online learning with *convex* losses to that of online learning with *linear* losses. In particular, the algorithms and results that we present in the remaining of this chapter hold for any (bounded) sequence of feedback $(g_t)_{t\in\mathbb{N}}$ as long as we replace the regret by the linearized regret. More generally, this lemma is used throughout the thesis whenever we present a result on regret bounds. Therefore, to avoid repetition, we will *not* refer to Lemma 2.1 when it is used, and our proof often ends up with a bound on the linearized regret.

Full-information and bandit feedback

Remark 2.1 (Types of feedback in online learning). Besides the first-order feedback that we study in this thesis, there exist various other types of feedback in online learning. Full-information feedback, for example, provides the learner with the entire loss function. Such feedback typically arises in learning system used for prediction tasks like classification or regression, where actions represent model parameters, and all relevant losses can be computed upon observation of the data [246]. We will only need this more stringent assumption in Section 4.3. On the other extreme, bandit feedback simply returns the loss value evaluated at the executed action [30, 168], presenting additional challenges due to the minimal information made accessible to the learner. We discuss potential extensions of our results to this more demanding scenario in the perspectives (Chapter 9) at the end of this thesis.

Online non-convex optimization

Remark 2.2 (Online non-convex optimization). Although there are a few works on online non-convex optimization, these works either drastically relax the definition of the regret [107, 115], or consider randomized algorithms that require the solution of a non-convex optimization or a sampling problem in each round [162, 258]. In fact, there is generally no hope to achieve sublinear regret with a deterministic algorithm when the loss functions are non-convex [258].

2.2 MIRROR DESCENT AND DUAL AVERAGING

In this section, we present two core algorithms for online convex optimization: *mirror descent* (MD) and *dual averaging* (DA). We recall their regret guarantees and provide accompanying proofs, which serve as fundamental building blocks for more complex results presented later in this thesis. For a more comprehensive understanding of this topic, we recommend readers to consult [29, 143, 189, 191] and references therein.

2.2.1 *Regularizers, Bregman Divergences, and Mirror Maps*

Projected gradient descent

In the online convex optimization setup that we described in Section 2.1, the

most natural algorithm to consider is the (projected) gradient descent method

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta_{t+1} g_t).$$

In the above, $\Pi_{\mathcal{X}} \colon \mathbb{R}^d \to \mathcal{X}$ is the Euclidean projector

$$\Pi_{\mathcal{X}}(z) = \arg\min_{x' \in \mathcal{X}} \|x' - x\| \tag{2.1}$$

and $\eta_{t+1} > 0$ is the *learning rate* of the algorithm for the update of x_{t+1}. Nonetheless, in certain situations, this method may lead to computational inefficiency and high regret due to the limitations of the Euclidean distance in capturing the underlying problem geometry.

In this regard, MD and DA extend these methods to take into account non-Euclidean geometries, thereby allowing for more flexibility and better performance in a wide range of problems. To define them, we need a *regularizer* defined in the following sense.

Regularizer

Definition 2.3 (Regularizer)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$. We define a *regularizer* on $\mathcal{X}$ as a function $h \colon \mathcal{X} \to \mathbb{R}$ that has the following properties

- h is continuous and 1-strongly convex relative to a norm $\|\cdot\|$ on $\mathbb{R}^d$.
- The subdifferential of the function ∂h admits a continuous selection ∇h. That is, $\nabla h \colon \operatorname{dom} \partial h \to \mathbb{R}$ is a continuous function such that $\nabla h(x) \in \partial h(x)$ for all $x \in \operatorname{dom} \partial h$.

Remark 2.3. The function h defined above has several different names in the literature, ranging from distance generating function [209], Bregman function [41], to mirror map [29], among others. We use the term "regualrizer"for the sake of simplicity and to emphasize its role in preventing excessive variation in the learner's actions from one round to the next.

Notations on norms

We note that unlike in (2.1) where $\|\cdot\|$ stands for the L2 norm, in Definition 2.3 $\|\cdot\|$ can be any norm on $\mathbb{R}^d$. Therefore, to avoid confusion, we will use $\|\cdot\|_2$ for the L2 norm hereinafter, and similarly, we respectively use $\|\cdot\|_1$ and $\|\cdot\|_\infty$ to denote the L1 and the L-infinity norms. As for the notation $\|\cdot\|$, we reserve it for the norm associated to h, i.e., the norm for which h is 1-strongly convex, unless otherwise stated. Its dual norm is written as $\|\cdot\|_*$ and is defined by $\|y\|_* = \max_{\|x\| \leq 1} \langle y, x \rangle$.

Moving forward, with the help of a regularizer h, we can then define a sort of "distance" on the action set $\mathcal{X}$.

Bregman divergence

Definition 2.4 (Bregman divergence)**.** The Bregman divergence associated with a regularizer h between two points $z \in \mathcal{X}$ and $x \in \operatorname{dom} \partial h$ is defined as

$$D_h(z, x) = h(z) - h(x) - \langle \nabla h(x), z - x \rangle.$$

We drop the subscript h from D_h whenever the choice of the regularizer is clear from the context.

Although Bregman divergence is not a distance in the strict mathematical sense (since it lacks symmetry and the triangle inequality), it serves as a measure of dissimilarity that captures how far the points are from each other in the geometry specified by the regularizer. This aspect is both important for the design of the algorithms and the analyses thereof.

The last element that we would like to introduce is the mirror map, which maps a vector in the "dual space" to a point in the "primal space".

Definition 2.5 (Mirror map)**.** The mirror map associated with a regularizer h is a function $Q_h \colon \mathbb{R}^d \to \mathcal{X}$ such that for all $y \in \mathbb{R}^d$ *Mirror map*

$$Q_h(y) = \arg\min_{x \in \mathcal{X}} \langle -y, x \rangle + h(x).$$

Similarly, we drop the subscript h from Q_h when there is no confusion.

Remark 2.4. As noted in Remark 2.3, the term mirror map is sometimes used in the literature to refer to the regularizer h.

Primal-dual perspective

As mentioned, the mirror map's role is to bridge the dual space with the primal space. In fact, for the algorithms that we are going to consider, it is convenient to regard the actions taken by the learner as "primal points" and the received feedback as "dual vectors". This primal-dual perspective is tied intrinsically to the definition of a gradient as a linear functional, and is indispensable when the ambient space is just a Banach space without additional structure. Even in our setup where both the actions and gradients are vectors in $\mathbb{R}^d$, this viewpoint is still useful in conceptualizing MD and DA as algorithms that navigate between the primal and the dual spaces via the gradient of the regularizer ∇h and the mirror map Q.

Importantly, this passage between the primal and the dual spaces is possible particularly because the mirror map always produces a value in $\operatorname{dom} \partial h$, i.e., $\operatorname{im} Q_h \subseteq \operatorname{dom} \partial h$. In Appendix A, we provide more technical details on this fact and other aspects of these elements. While these details are essential for our analysis, they may not be crucial for a broad understanding of the topic. Consequently, we have chosen to relegate them to the appendix and instead focus here on offering concrete examples to illustrate these concepts.

Regularizer examples

Example 2.1 (Quadratic regularizer)**.** Consider $h(x) = \|x\|_2^2/2$. This is known as the quadratic regularizer and it is 1-strongly convex with respect to the L2 norm $\|\cdot\|_2$. For any action set $\mathcal{X} \subseteq \mathbb{R}^d$, the gradient map $\nabla h \colon x \to x$ is a valid continuous selection of the subgradient, the associated Bregman divergence is the squared Euclidean distance, $D_h(z, x) = \|z - x\|_2^2/2$, and the associated mirror map is the Euclidean projector onto $\mathcal{X}$ defined in (2.1), i.e., $Q_h = \Pi_{\mathcal{X}}$.

Example 2.2 (Negentropy regularizer for probability simplex)**.** Consider $h(x) = \sum_{k=1}^d x_k \log(x_k)$ for $x \in \mathcal{X}$, where $\mathcal{X} = \Delta_d$ is the probability simplex and x_k is the k-th coordinate of x. This is called the negentropy regularizer as $h(x)$ is the negative entropy of the distribution described by x. One can show that h is 1-strongly convex with respect to the L1 norm $\|\cdot\|_1$, and a possible continuous selection of the subgradient is $\nabla h \colon x \to (\log(x_1), \ldots, \log(x_d))$.[2]

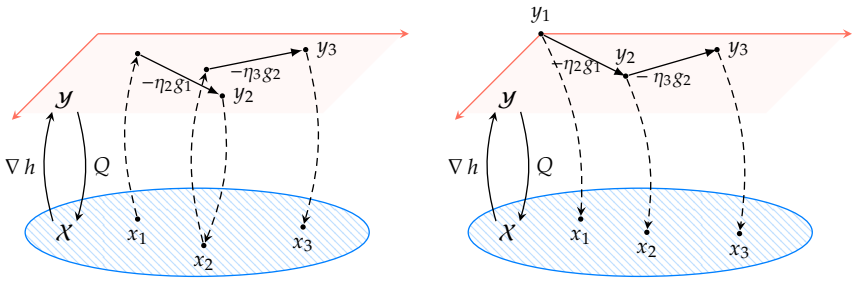
Regardless of the specific choice of ∇h, the Bregman divergence associated with this regularizer is the Kullback–Leibler (KL) divergence,

$$D_h(z, x) = \sum_{k=1}^d z_k \log\left(\frac{z_k}{x_k}\right),$$

where it is understood that $0 \log(0) = 0$ and $z_k \log(z_k/0) = \infty$ if $z_k > 0$. Note that, however, in Definition 2.4 we only define Bregman divergence for $x \in \operatorname{dom} \partial h$, which means that x would never have a coordinate that equals 0.

Finally, the mirror map in this case is often referred to as the *softmax function*. It takes a vector $y \in \mathbb{R}^d$ and returns a point in the simplex $\mathcal{X}$, with the k-th coordinate given by $(Q_h(y))_k = \exp(y_k)/(\sum_{l=1}^d \exp(y_l))$.

2 Note that this is *not* the gradient of h when viewed as a function defined over the entire $\mathbb{R}^d$.



(a) Agile mirror descent (b) Lazy mirror descent

Figure 2.2: Schematic representations of mirror descent. $\mathcal{Y}$ denotes the dual space and the y_ts are dual variables.

2.2.2 *Mirror Descent*

Mirror descent

The origin of the *mirror descent* (MD) algorithm traces back to the work of Nemirovski and Yudin [208]. It uses a regularizer h and its associated Bregman divergence D. During each iteration, the algorithm updates the current point x_t to a new point x_{t+1}, using a subgradient $g_t \in \partial \ell_t(x_t)$ as:

$$x_{t+1} = \operatorname*{arg\,min}_{x \in \mathcal{X}} \langle g_t, x \rangle + \frac{D(x, x_t)}{\eta_{t+1}}. \tag{MD}$$

The above minimization allows to find a balance between moving along the direction of the gradient and staying close to the current point in terms of the Bregman divergence. Alternatively, the update (MD) can be formulated as

$$x_{t+1} = Q(\nabla h(x_t) - \eta_{t+1} g_t).$$

This formulation illuminates the "mirror" aspect of the algorithm. To compute x_{t+1}, it first maps the learner's action x_t to the dual space via ∇h, performs an update in the dual space in the direction of the opposite of the gradient, and then maps it back to the primal space using the mirror map Q (see Fig. 2.2a for an illustration).

Lazy mirror descent

Besides this "eager" version that moves back and forth between the primal and the dual space, it is often considered in the literature a "lazy" version of the algorithm (see e.g., [194, 246]) that, as shown in Fig. 2.2b, outputs directly the projection of the aggregated gradient vector

$$x_{t+1} = Q\left(-\sum_{s=1}^{t} \eta_{s+1} g_s\right). \tag{LMD}$$

The two variants of MD coincide when h is infinitely "steep" at the boundary of $\mathcal{X}$—that is, when $\operatorname{dom} \partial h \cap \mathcal{X} = \operatorname{ri} \mathcal{X}$; otherwise, they yield different sequences of actions [163]. For the sake of illustration, we present below two examples that employ different regularizers and thus represent different instantiations of the MD algorithm.

Example: projected gradient descent

Example 2.3 (Projected gradient descent)**.** Consider the quadratic regularizer $h = \|\cdot\|_2^2/2$ as discussed in Example 2.1. In this case, ∇h is the identity function, and Q is the Euclidean projector onto $\mathcal{X}$. The eager version of MD thus corresponds to projected gradient descent

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta_{t+1} g_t).$$

On the other hand, the lazy version of MD gives rise to *lazy gradient descent*, which updates according to

$$x_{t+1} = \Pi_{\mathcal{X}} \left(x_1 - \sum_{s=1}^{t} \eta_{s+1} g_s \right).$$

Note that the above is obtained by adjusting the quadratic regularizer to be centered at x_1, i.e., $h = \|\cdot - x_1\|_2^2/2$. Apparently, projected gradient descent and lazy gradient descent coincide when $\mathcal{X}$ is an affine subspace of $\mathbb{R}^d$, in which case $\operatorname{dom} \partial h \cap \mathcal{X} = \operatorname{ri} \mathcal{X}$, but they are otherwise two different algorithms.

Example: multiplicative weights update

Example 2.4 (Multiplicative weights update)**.** Consider the negentropy regularizer $h(x) = \sum_{k=1}^{d} x_k \log(x_k)$ on the probability simplex $\mathcal{X} = \Delta_d$ as discussed in Example 2.2.[3] The mapping via ∇h turns a probability vector x into a log-probability vector, and the mapping via Q transforms a vector of scores into a probability vector by applying the softmax function. This leads to the multiplicative weights update (MWU) algorithm, whose coordinate-wise update is

$$x_{t+1,k} = \frac{x_{t,k} \exp(-\eta_{t+1} g_{t,k})}{\sum_{l=1}^{d} x_{t,l} \exp(-\eta_{t+1} g_{t,l})}.$$

If we consider the lazy formulation, we have

$$x_{t+1,k} = \frac{x_{1,k} \exp(-\sum_{s=1}^{t} \eta_{s+1} g_{s,k})}{\sum_{l=1}^{d} x_{1,l} \exp(-\sum_{s=1}^{t} \eta_{s+1} g_{s,l})}.$$

It is straightforward to see that these two update rules are equivalent. As a side note, this algorithm has various names across different fields in the literature, including *exponential weights* in multi-armed bandits [13], *entropic mirror descent* in optimization [18], *Hedge* in game theory [88], and *weighted majority* in machine learning [178]. While their exact description may differ in detail, they all share the same underlying principle of modifying the components in a multiplicative way as determined by the feedback.

REGRET BOUND. We next shift our focus to the regret guarantee of MD. We focus here on the eager version of the algorithm, as described by (MD), but a similar result can also be proved for the lazy version [246] (in our statement below, it is sufficient to replace the Bregman divergences by some "Fenchel couplings" that we introduce in Section 2.2.3).

Regret of MD

Proposition 2.2. *Suppose that Assumption 2.1 holds and that* (MD) *is run with learning rates* $(\eta_t)_{t \in \mathbb{N}}$*. Then, for any* $z \in \mathcal{X}$ *and* $T \in \mathbb{N}$*, the regret of the learner relative to* z *after* T *rounds is bounded as*

$$\operatorname{Reg}_T(z) \leq \frac{D(z, x_1)}{\eta_2} + \sum_{t=2}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) D(z, x_t) + \frac{1}{2} \sum_{t=1}^{T} \eta_{t+1} \|g_t\|_*^2.$$

3 In this example, a subscript could denote either a coordinate index or a time index, but this should cause no confusion. Specifically, coordinate indices are consistently denoted by k or l, while time indices are denoted by t or s.

Proof. We proceed to bound the linearized regret of the algorithm. For this, we decompose it as follows

$$\mathrm{LinReg}_T(z) = \sum_{t=1}^{T} \langle g_t, x_t - z \rangle = \underbrace{\sum_{t=1}^{T} \langle g_t, x_t - x_{t+1} \rangle}_{A} + \underbrace{\sum_{t=1}^{T} \langle g_t, x_{t+1} - z \rangle}_{B}. \tag{2.2}$$

Sum A in (2.2) is sometimes referred to as the *prediction drift* and it quantifies how much the learner changes their predictions between consecutive rounds. We bound it by Young's inequality as

$$\sum_{t=1}^{T} \langle g_t, x_t - x_{t+1} \rangle \leq \sum_{t=1}^{T} \left(\frac{\eta_{t+1} \|g_t\|_*^2}{2} + \frac{\|x_t - x_{t+1}\|^2}{2\eta_{t+1}} \right) \tag{2.3}$$

As for sum B in (2.2), it may be referred to as the *forward regret* and it measures the linearized regret that the learner would have incurred if they had chosen the more informed prediction x_{t+1} in the place of x_t (note that computing x_{t+1} requires the the use of g_t which is only received in round t, so this is just a conceptualized algorithm that can not be implemented in practice). To bound this term, we resort to the optimality condition (precisely, Lemma A.1) of the update rule $x_{t+1} = Q(\nabla h(x_t) - \eta_{t+1} g_t)$. This gives

$$\langle \nabla h(x_{t+1}), x_{t+1} - z \rangle \leq \langle \nabla h(x_t) - \eta_{t+1} g_t, x_{t+1} - z \rangle$$

Rearranging and applying the three-point identity for Bregman divergence (Lemma A.2, Eq. A.1), we get

$$\begin{aligned} \langle g_t, x_{t+1} - z \rangle &\leq \frac{1}{\eta_{t+1}} \langle \nabla h(x_t) - \nabla h(x_{t+1}), x_{t+1} - z \rangle \\ &\leq \frac{1}{\eta_{t+1}} (D(z, x_t) - D(z, x_{t+1}) - D(x_{t+1}, x_t)). \end{aligned} \tag{2.4}$$

The regularizer h being 1-strongly convex relative to the norm $\|\cdot\|$, we have

$$D(x_{t+1}, x_t) \geq \frac{\|x_t - x_{t+1}\|^2}{2}. \tag{2.5}$$

Combining (2.2)–(2.5), we then obtain

$$\mathrm{LinReg}_T(z) \leq \frac{D(z, x_1)}{\eta_2} + \sum_{t=2}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) D(z, x_t) + \sum_{t=1}^{T} \frac{\eta_{t+1} \|g_t\|_*^2}{2}.$$

This concludes the proof. □

As an immediate consequence of Proposition 2.2, we see that (MD) with properly chosen learning rates guarantees sublinear regret when the feedback is bounded. Concretely, we assume the following.

Boundedness of feedback

Assumption 2.2 (Boundedness)**.** There exists $G > 0$ such that for all $t \in \mathbb{N}$, we have $\|g_t\|_* \leq G$.

This boundedness assumption is crucial for achieving sublinear regret. Otherwise, the loss function of each round can be adversarially chosen so that the difference between the loss incurred by the selected action and the loss

incurred by another action is arbitrarily large, making any meaningful control over the regret unachievable.

With this additional assumption, we are now ready to state a regret bound for constant learning rate (MD) that is standard across the literature.

$O(\sqrt{T})$ regret of constant learning rate MD

Corollary 2.3. *Suppose that Assumptions 2.1 and 2.2 hold and that (MD) is run with constant learning rate*

$$\eta = \frac{R}{G}\sqrt{\frac{2}{T}}.$$

Then, for any $z \in \mathcal{X}$ such that $D(z, x_1) \leq R^2$, the regret of the learner relative to z after T rounds is bounded as

$$\mathrm{Reg}_T(z) \leq RG\sqrt{2T}.$$

Remark 2.5. Although Assumption 2.2 requires $\|g_t\|_* \leq G$ to hold for all t, it is clear that for any finite-horizon result of this type it is sufficient to have $G \geq \max_{1\leq t\leq T}\|g_t\|_*$.

Minimax optimality of $O(\sqrt{T})$ regret bound

As we can see from the statement, the regret bound involves three quantities: the time horizon T, a bound R on the distance from the initial point to the comparator action (measured by the Bregman divergence), and a bound G on the magnitude of feedback (measured by the dual norm). This dependence on R, G, and $\sqrt{T}$ is minimax optimal [1, 114, 245]. Formally, it is shown in [1, Th. 4.1] that for any online learning algorithm $\mathfrak{A}$ and any R, G, and T, there exists an action set $\mathcal{X}$ of diameter R (i.e., for any $x, x' \in \mathcal{X}$, we have $\|x - x'\|_2 \leq R$), and a sequence of convex loss functions $(\ell_t)_{1\leq t\leq T}$ such that when the learner executes $\mathfrak{A}$ against this sequence, they receive feedback vectors $(g_t)_{1\leq t\leq T}$ satisfying $\|g_t\|_2 \leq G$, and incur regret that is lower bounded as

$$\max_{z\in\mathcal{X}} \mathrm{Reg}_T(z) \geq \frac{RG\sqrt{T}}{2\sqrt{2}}.$$

In this sense, Corollary 2.3 guarantees, up to a constant factor, the *best* performance we can expect in the *worst* case, encapsulating the essence of the *minimax* optimality.

Anytime algorithm and doubling trick

We however note that the learning rate of Corollary 2.3 depends on the time horizon T. This dependency is not ideal in real-world scenarios because we typically desire an *anytime* algorithm—one that works well irrespective of the specific time horizon. This issue can be addressed via the well-known *doubling trick* [13, 246], which involves halving the learning rate and restarting the algorithm every time the number of iterations doubles.

Limitation of MD with decreasing learning rate

Alternatively, we could adopt a simpler solution: a decreasing learning rate of the form $\eta_t = 1/\sqrt{t}$. Yet, as indicated by Proposition 2.2, this strategy guarantees sublinear regret only when the Bregman divergence is bounded on $\mathcal{X}$. In fact, it is formally shown by Orabona and Pál [216] that (MD) with such learning rate can induce linear or even superlinear regret in the setup considered in Corollary 2.3. This motivates us to introduce the dual averaging algorithm in the next subsection, aimed at addressing this shortcoming of MD.

2.2.3 *Dual Averaging*

Viewed abstractly, the issue with MD is that with decreasing learning rates, new information enters the algorithm with a diminishing weight. This is problematic from a learning viewpoint because it gives more weight to earlier, less informed updates, and less weight to more recent, more relevant ones. An adversary

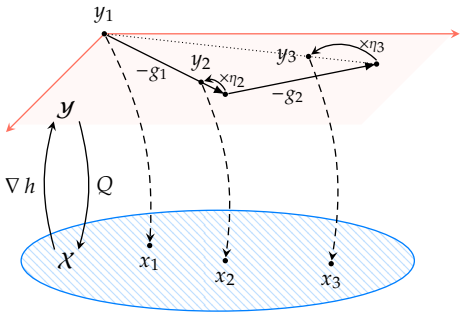


Figure 2.3: Schematic representations of dual averaging. $\mathcal{Y}$ denotes the dual space and the y_ts are dual variables defined by $y_t = -\eta_t \sum_{s=1}^{t-1} g_s$.

could exploit this characteristic and push the algorithm far from an optimal point in the starting iterations, leading to suboptimal performance of the learner.

Dual averaging

The *dual averaging* (DA) algorithm, introduced by Nesterov [212], provides a solution to this problem. It employs the same weighting for all gradients in the algorithm. The update rule of DA is given by

$$x_t = \operatorname*{arg\,min}_{x \in \mathcal{X}} \sum_{s=1}^{t-1} \langle g_s, x \rangle + \frac{h(x)}{\eta_t}. \tag{DA}$$

DA as linearized FTRL

The above form reflects the relation between DA and the *follow the regularized leader* (FTRL) algorithm [247]. In fact, DA may be viewed as a "linearized" version of FTRL, and it aligns with FTRL when the encountered loss functions are linear.

Feedback averaging in dual space

As for the terminology "dual averaging", it pays homage to the process of averaging gradients in the dual space before mirroring them back into the action set $\mathcal{X}$. This is made explicit by the following alternative form of the update rule (see also Fig. 2.3)

$$x_t = Q\left(-\eta_t \sum_{s=1}^{t-1} g_s\right).$$

DA versus Lazy MD

The difference between (DA) and (LMD) also becomes apparent thanks to the above formula. Unlike (LMD), where the learning rate is applied before the sum of feedback, (DA) applies it after summing the feedback. This seemingly small difference significantly impacts the performance of the algorithm. As we will see in the next theorem, it helps derive a regret bound that circumvents the "finite Bregman diameter" limitation which restrains MD's applicability.

Regret of DA

Proposition 2.4. *Suppose that Assumption 2.1 holds and that* (DA) *is run with non-increasing learning rates* $(\eta_t)_{t \in \mathbb{N}}$. *Then, for any* $z \in \mathcal{X}$ *and* $T \in \mathbb{N}$, *the regret of the learner relative to* z *after* T *rounds is bounded as*

$$\mathrm{Reg}_T(z) \leq \frac{h(z) - \min h}{\eta_T} + \frac{1}{2} \sum_{t=1}^{T} \eta_t \|g_t\|_*^2.$$

Compared to Proposition 2.2, all the Bregman divergence terms $D(z, x_t)$ are replaced by the difference $h(z) - \min h$ (this equals to $D_{h'}(z, x_1)$ if we set h to be $D_{h'}(\cdot, x_1)$ for some regularizer h'). From this we deduce immediately the following regret bound that applies to (DA) with $\eta_t = \Theta(1/\sqrt{t})$ learning rate.

$O(\sqrt{T})$ regret of decreasing learning rate DA

Corollary 2.5. *Suppose that Assumptions 2.1 and 2.2 hold and that (DA) is run with decreasing learning rate*

$$\eta_t = \frac{R}{G}\frac{1}{\sqrt{t}}.$$

Then, for any $T \in \mathbb{N}$ and any $z \in \mathcal{X}$ such that $h(z) - \min h \le R^2$, the regret of the learner relative to z after T rounds is bounded as

$$\mathrm{Reg}_T(z) \le 2RG\sqrt{T}.$$

Corollary 2.5 demonstrates that the minimax optimal regret bound can be achieved using (DA) with a simple decreasing learning rate schedule.

Another notable distinction between Proposition 2.4 and Proposition 2.2 lies in the coefficient preceding $\|g_t\|_\star^2$. In Proposition 2.2, we have the learning rate η_{t+1} used to compute x_{t+1}, whereas in Proposition 2.4 we have η_t. This shift in index is a necessary trade-off for the anytime property of (DA), and it significantly affects our analysis of the adaptive variants of these algorithms, as we will discuss in Section 2.3.

Let us now turn to the proof of Proposition 2.4. Our analysis is based on the use of *Fenchel coupling*, an elegant distance measure introduced by Mertikopoulos and Sandholm [192] (see also [26, 193]).

Fenchel coupling

Definition 2.6 (Fenchel coupling)**.** The Fenchel coupling associated with a regularizer h between a primal point $z \in \mathcal{X}$ and a dual vector $y \in \mathbb{R}^d$ is defined as

$$F_h(z, y) = h(z) + h^*(y) - \langle y, z\rangle.$$

In the above, $h^*\colon y \to \max_{x\in\mathcal{X}}\langle y, x\rangle - h(x)$ is the Fenchel conjugate of h. Again, we drop the subscript h from F_h whenever the choice of the regularizer is clear from the context.

Fenchel coupling may be regarded as a "primal-dual" version of Bregman divergence. By the definition of the mirror map, we deduce immediately that

$$F(z, y) = h(z) - h(Q(y)) - \langle y, z - Q(y)\rangle.$$

Provided that $y \in \partial h(Q(y))$ (see Lemma A.1), Fenchel coupling is also closely related to a generalized version of Bregman divergence which is defined for $z \in \mathcal{X}$, $x \in \operatorname{dom}\partial h$, and $g \in \partial h(x)$ as $D(z, x; g) = h(z) - h(x) - \langle g, z - x\rangle$. This definition is formally introduced by Juditsky et al. [143], but its use in the literature can be traced back to much earlier works such as [156].

In the analysis that follows, we leverage Fenchel coupling to provide an alternative formulation for the standard analysis of (DA), as described in [112, 282]. This formulation not only accentuates the link with the proof of Proposition 2.2, but it also highlights the role of maintaining uniform weight for all the feedback in eliminating the undesirable factors in the analysis.

Proof of Proposition 2.4. Following the decomposition of the linearized regret in (2.2), we start by bounding the forward regret. Let us define the dual variable $y_t = -\eta_t \sum_{s=1}^{t-1} g_s$ so that $x_t = Q(y_t)$. Applying the three-point identity for Fenchel coupling (Lemma A.2, Eq. A.2) to the update of x_{t+1} gives

$$\begin{aligned}\langle g_t, x_{t+1} - z\rangle &= \left\langle \frac{y_t}{\eta_t} - \frac{y_{t+1}}{\eta_{t+1}}, x_{t+1} - z\right\rangle \\ &= \frac{1}{\eta_t}\langle y_t - y_{t+1}, x_{t+1} - z\rangle + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\langle 0 - y_{t+1}, x_{t+1} - z\rangle\end{aligned}$$

$$= \frac{1}{\eta_t}(F(z, y_t) - F(z, y_{t+1}) - F(x_{t+1}, y_t)) + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)(F(z, 0) - F(z, y_{t+1}) - F(X_{t+1}, 0)).$$

Since $\eta_t \geq \eta_{t+1}$ and $F(x_{t+1}, 0) \geq 0$, the last term in the above equation can be dropped. With $F(z, 0) = h(z) - h(Q(0)) = h(z) - \min h$, we then deduce

$$\langle g_t, x_{t+1} - z\rangle \leq \frac{F(z, y_t)}{\eta_t} - \frac{F(z, y_{t+1})}{\eta_{t+1}} - \frac{F(x_{t+1}, y_t)}{\eta_t} + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)(h(z) - \min h). \tag{2.6}$$

To proceed, we bound the prediction drift with the help of Young's inequality.

$$\langle g_t, x_t - x_{t+1}\rangle \leq \left(\frac{\eta_t \|g_t\|_*^2}{2} + \frac{\|x_t - x_{t+1}\|^2}{2\eta_t}\right) \tag{2.7}$$

Moreover, by Lemma A.3, the following inequality holds given that $x_t = Q(y_t)$.

$$F(x_{t+1}, y_t) \geq \frac{\|x_{t+1} - x_t\|^2}{2} \tag{2.8}$$

Putting (2.2) and (2.6)–(2.8) together readily leads to

$$\mathrm{Reg}_T(z) \leq \frac{h(z) - \min h}{\eta_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \eta_t \|g_t\|_*^2.$$

To see that it is possible to have η_T instead of η_{T+1} in the denominator, just note that we may define $y'_{t+1} = -\eta_t \sum_{s=1}^{t} g_s$ and $x'_{t+1} = Q(y'_{t+1})$. Then, (2.6) still holds after replacing η_{t+1}, x_{t+1}, and y_{t+1} respectively by η_t, y'_{t+1}, and x'_{t+1}. We can thus use this version of inequality for $t = T$. □

Remark 2.6. As can be seen from the proof of Proposition 2.4, for (DA) there is actually no need to assume the existence of the continuous selection ∇h. The strong convexity itself is sufficient. The same remark applies to (LMD).

Remark 2.7. One may be wondering if there is an "eager" variant of (DA) whose update is closer to (MD) but applies the same weight to all the feedback. One possibility to achieve this is via the use of a "stabilization" step [76]. We formally introduce this in Chapter 6.

2.3 ADAPTIVE LEARNING RATE

Data-dependent regret bound

In Section 2.2, we mainly focus on the dependence of the regret on the number of rounds T. In particular, we simply bound the magnitude of feedback by an upper bound G in Corollaries 2.3 and 2.5. Nonetheless, if we look closely at Propositions 2.2 and 2.4, we may notice that these regret bounds actually depend on the norm of each piece of feedback. By taking a constant learning rate of the form

$$\eta = \frac{R}{\sqrt{\sum_{t=1}^{T} \|g_t\|_*^2}}, \tag{2.9}$$

we obtain an optimal $O\left(\sqrt{\sum_{t=1}^{T}\|g_t\|_*^2}\right)$ *data-dependent* bound (For optimality of this regret bound, see [216, Th. 5]). This bound is highly advantageous when dealing with feedback of uneven distribution or scale.

AdaGrad-norm

Unfortunately, the learning rate (2.9) cannot be computed in advance as it requires knowledge of all the feedback norms for the entire course of the interaction. Instead, the best we can hope for is to use the historical feedback information collected up to the current instant. This gives rise to the following learning rate policy.

$$\eta_t = \frac{R}{\sqrt{\beta + \sum_{s=1}^{t-1}\|g_s\|_*^2}}, \qquad \text{(AdaGrad-norm)}$$

where $R > 0$ and $\beta \geq 0$ are constants chosen by the learner at the beginning of the learning process. The name *AdaGrad-norm* is to distinguish it from the original *AdaGrad* algorithm introduced by Duchi et al. [68], which uses either coordinate-wise (diagonal) or "full-matrix" learning rate. In our notation, this requires to have a different regularizer h_t in each round that is tuned adaptively according to the feedback.

Generally speaking, both AdaGrad-norm and AdaGrad adjust their learning rates responsively to the fluctuating landscape of feedback. This flexibility, as opposed to being confined to a single fixed learning rate, indeed results in data-dependent regret bounds that are optimal up to a multiplicative constant.

To demonstrate this benefit of the adaptive methods, we rely on the following standard lemma which provides an upper bound for the sum of a sequence of non-negative real numbers, each weighted by the inverse of the square root of their cumulative sum.

The "inverse sqaure root of sum" lemma

Lemma 2.6 (Auer et al. [14, Lem 3.5])**.** *Let $T \in \mathbb{N}$ and $\varepsilon \in \mathbb{R}_+$. For any sequence of non-negative real numbers $a_1, \ldots, a_T$, it holds*

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt{\varepsilon + \sum_{s=1}^{t} a_s}} \leq 2\sqrt{\sum_{t=1}^{T} a_t},$$

where it is understood that $0/0 = 0$.

With this in mind, the regret bound for adaptive (MD) comes from a straightforward combination of Proposition 2.2 and Lemma 2.6. We state it below.

Regret of adaptive MD

Proposition 2.7. *Suppose that Assumption 2.1 holds and that (MD) is run with adaptive learning rate (AdaGrad-norm). Then, for any $z \in \mathcal{X}$ and $T \in \mathbb{N}$, the regret of the learner relative to z after T rounds is bounded as*

$$\mathrm{Reg}_T(z) \leq \left(\frac{\sup_{x\in\mathcal{X}} D(z,x)}{R} + R\right)\sqrt{\beta + \sum_{t=1}^{T}\|g_t\|_*^2}$$

Proof. On on hand, we have

$$\frac{D(z,x_1)}{\eta_2} + \sum_{t=2}^{T}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)D(z,x_t) \leq \frac{\sup_{x\in\mathcal{X}} D(z,x)}{\eta_{T+1}}$$

On the other hand, by Lemma 2.6 we know that

$$\frac{1}{2}\sum_{t=1}^{T}\eta_{t+1}\|g_t\|_*^2 = \frac{R}{2}\sum_{t=1^T}\frac{\|g_t\|_*^2}{\sqrt{\beta+\sum_{s=1}^{t}\|g_s\|_*^2}} \leq R\sqrt{\sum_{s=1}^{t}\|g_t\|_*^2}.$$

Plugging the above two inequalities into the bound of Proposition 2.2 gives the desired result. □

In a similar vein, we have the following regret bound for adaptive (DA).

Regret of adaptive DA

Proposition 2.8. *Suppose that Assumptions 2.1 and 2.2 hold and that* (DA) *is run with adaptive learning rate* (AdaGrad-norm)*. Then, for any $z \in \mathcal{X}$ and $T \in \mathbb{N}$, the regret of the learner relative to z after T rounds is bounded as*

$$\mathrm{Reg}_T(z) \leq \left(\frac{h(z)}{R}+R\right)\sqrt{\beta+\sum_{t=1}^{T}\|g_t\|_*^2}+\frac{RG^2}{\sqrt{\beta}}$$

Proof. The main challenge here is that $\|g_t\|_*^2$ only scales in g_t in Proposition 2.4. To overcome this, we use the following trick

$$\eta_t\|g_t\|_*^2 = \eta_{t+1}\|g_t\|_*^2 + (\eta_t-\eta_{t+1})\|g_t\|_*^2 \leq \eta_{t+1}\|g_t\|_*^2 + (\eta_t-\eta_{t+1})G^2.$$

Putting this with Proposition 2.4 and Lemma 2.6 completes the proof. □

Proposition 2.7 along with Proposition 2.8 suggest that both (MD) and (DA) with adaptive learning rate (AdaGrad-norm) can guarantee the optimal $\mathcal{O}\left(\sqrt{\sum_{t=1}^{T}\|g_t\|_*^2}\right)$ data-dependent regret bound. Yet, these results come with a few caveats.

For (MD), the regret bound is only meaningful when the Bregman diameter $\sup_{x\in\mathcal{X}} D(z,x)$ is finite. This requirement is unavoidable as discussed in Section 2.2.2. Regarding (DA), we additionally assume Assumption 2.2 in Proposition 2.8. This is not crucial because we can replace G by $\max_{1\leq t\leq T}\|g_t\|_*$ in the bound as noted in Remark 2.5. What is more problematic is the square dependence on G and the the inverse square root dependence on the initialization parameter β. This implies that β must be set proportional to G^2 for this term to scale linearly with G, otherwise, the bound could be excessively large, or even become vacuous when $\beta = 0$. Fortunately, this issue is just an artifact of our analysis rather than an inherent drawback of the algorithm. It is indeed possible to refine Proposition 2.4 and subsequently Proposition 2.8 to yield a regret bound for the case $\beta = 0$, with an additional term that is only in the order of $G\sqrt{T}$. We refer the readers to [216] for more details on this result.

Overall, the dual averaging method and its adaptive variants exhibit numerous advantages and require less stringent assumptions, making them the primary focus of our investigation in this thesis.

Adaptive methods beyond online learning

To complete the picture, we note that the value of these adaptive methods is not limited to the regret guarantees presented here. Specifically, these methods have also been shown to converge and yield optimal convergence rates in (non-convex) optimization [174, 275, 294]. In this context, the term "adaptive" alludes to the algorithm's capacity to achieve the desired guarantees under various conditions. This characteristic is further investigated in the context of learning in games in Chapters 6 and 8.

Finally, it is important to note that the reach of adaptive methods extends to its numerous variants. In particular, Adam [155] and AdamW [180], among others [63, 232, 250, 265, 287], have made significant contributions to current advancements in deep learning and artificial intelligence. These variants continue to be at the forefront of this rapidly evolving field, demonstrating their versatility and effectiveness in handling a wide range of problems.

3

MULTI-AGENT ONLINE OPTIMIZATION WITH DELAYS

 # This chapter incorporates material from Hsieh et al. [128, 130]

HAVING established the basics of online convex optimization in the previous chapter, we now turn to our main contribution in the cooperative multi-agent setup: a multi-agent online learning framework with a focus on *delays* and *asynchronicities*.

Multi-agent online learning

Imagine deploying an online learning algorithm across a network of agents—it could be a group of autonomous vehicles navigating a city [248], a distributed energy management system optimizing power use across a smart grid [200], or a recommender system running on servers scattered around the globe [296]. This multi-agent context is not just a mere extension of the single-agent setting but presents a new frontier with unique challenges and opportunities.

Delay

In this multi-agent setting, one particular challenge that becomes increasingly relevant is delays. It is not uncommon for a significant lag to occur between an agent taking action and receiving the corresponding feedback. This can be due to a myriad of factors such as computational overheads [50, 188], communication latencies between agents (learners) [186, 268], or the inherent complexity of predicting long-term effects [280]. This delay introduces an additional layer of complexity and uncertainty into the learning process.

Lack of centralized control

Compounding these difficulties, the multi-agent setting is often devoid of a centralized control mechanism. In particular, agents may not have access to a global counter to use as a reference point, which represents a substantial deviation from single-agent situations.

Our framework aims to tackle these challenges by generalizing the methods and results discussed in the previous chapter, fitting them into this multi-agent setup with delays. Importantly, we focus on the information available for producing each action rather than the actual delay associated with each feedback. The aim is to provide a comprehensive theory for understanding and dealing with asynchronicities and delays in online learning, not just for specific applications, but from a broader, more universal standpoint.

CONTRIBUTIONS AND OUTLINE. There are three major underlying themes in our analysis. As we discussed above, the first has to do with *delays*: due to this lag between "action" and "reaction", agents may have to update their actions based on feedback that is potentially stale and obsolete. The second has to do with *multi-agent* systems: in a network setting, learners may have to take decisions with very different information at their disposal, and with no realistic means of coordinating their decision-making mechanisms. Expanding further on this point, the third has to do with *adaptivity*: we are interested in learning algorithms that can be run with minimal information prerequisites at the agent end, while still achieving optimal regret bounds.

A framework for asynchronous online optimization

To take all this into account, we introduce in Section 3.1 a flexible framework that unifies several models of online learning in the presence of delays—including both single- and multi-agent setups.

Delayed dual averaging

Building upon this, we extend the (DA) template to account for delays in Section 3.2. At the core of our analysis are the notions of *dependency graph* and *faithful permutation*. These allow us to reorder time indices in a way that is compatible with the decision-making process, effectively helping us tackle unique challenges that arise from the multi-agent setting.

Adaptive methods

Our results are concretized in Section 3.3, where we design and analyze adaptive algorithms that achieve optimal data- and delay-dependent regret bounds in this completely decentralized setting. In addition to this, we also develop a data- and delay-adaptive algorithm for the single-agent scenario that bypasses the "bounded delay" assumption.

Effective regret versus collective regret

In an endeavor to showcase the versatility of our framework, we proceed to present an alternative problem formulation in Section 3.4. This formulation closely resembles those commonly found in distributed online optimization. Following this, we provide bounds for the agents' *effective* and *collective* regrets to account for two distinct objectives of the learning system: in the former, the goal is to perform well on every upcoming request; while in the latter, the goal is to enhance the collaborative task undertaken by the entire group of agents.

Simulations on static and open networks

We close up this chapter with numerical simulations in static and open networks in Section 3.5. In our experiments, we address a decentralized least absolute deviation regression problem and compare our methods with decentralized gradient descent. This concluding analysis provides practical perspectives, complementing our theoretical findings and highlighting the broad applicability of our framework.

3.1 A FRAMEWORK FOR ASYNCHRONOUS ONLINE OPTIMIZATION

In this section, we lay out the general asynchronous online optimization framework that we study throughout this chapter. We also highlight the two challenges that arise in our framework due to its multi-agent nature.

3.1.1 *Problem Setup*

Problem setup

Consider a set of agents $\mathcal{N} = \{1, \dots, N\}$ playing against a sequence of time-varying loss functions, with the goal of minimizing their (joint) regret. Formally, at each time slot t, one of the agents becomes *active*, they select an action x_t from the action set $\mathcal{X}$, and they incur a loss $\ell_t(x_t)$.[1] The performance of the agents is measured by the regret defined in Definition 2.1.

$$\mathrm{Reg}_T(z) = \sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(z)$$

As in Section 2.1, $\mathcal{X}$ is assumed to be a closed convex subset of $\mathbb{R}^d$, each $\ell_t : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex and subdifferentiable on $\mathcal{X}$ (Assumption 2.1), and the agents (learners) receive first-order feedback $g_t \in \partial \ell_t(x_t)$. Irrespective of the nature of the problem, we will refer to x_t interchangeably as the *prediction* made by the active agent or the *action* played by the active agent at time t, and we will write $i(t)$ for the agent that is active at time t.

1 We discuss the case where multiple agents are active at each time step in Section 3.4.

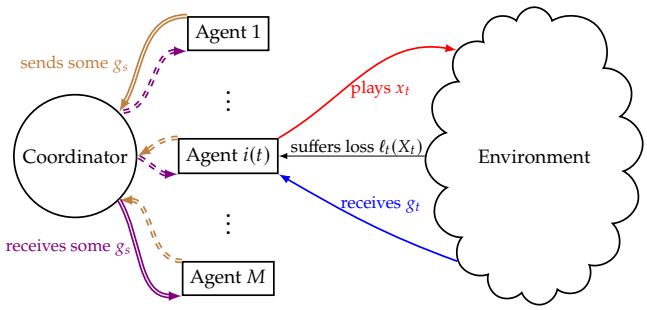


Figure 3.1: Illustration of the considered multi-agent online-learning setup: the case of coordinator-worker architecture.

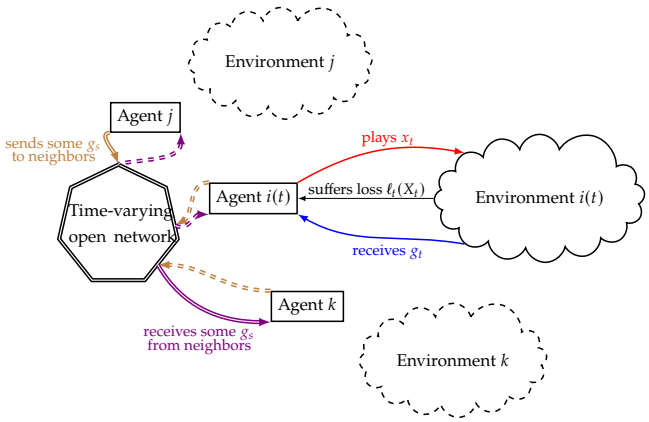


Figure 3.2: Illustration of the considered multi-agent online-learning setup: the case of decentralized open network.

Illustrative examples

For visualization purposes, the above setup is illustrated in Figs. 3.1 and 3.2. We highlight in these two figures the fact that we do not put any restriction on the communication architecture. In particular, the network may be either *centralized*, where agents exchange information through a single coordinator (Fig. 3.1), or *decentralized*, and even with agents that join and leave freely (Fig. 3.2).

Delayed feedback

THE DELAY MODEL. In environments with *delayed feedback*, g_t is only received by all the agents $i \in \mathcal{N}$ a certain amount of time after the generating action x_t is played. To express this formally, we write $[t] = \{1, ..., t\}$ for any $t \in \mathbb{N}$ and denote the set of gradient timestamps that are available to agent i at time t as $\mathcal{S}_t^i \subseteq [t-1]$ for ; in other words, at time t, the i-th agent only has $\{g_s : s \in \mathcal{S}_t^i\}$ at their disposal. Clearly, at each stage t, the active agent $i(t)$ can only compute x_t based on $\{g_s : s \in \mathcal{S}_t^{i(t)}\}$, the set of subgradients available for it at time t. This quantity is of utmost importance in our framework. We thus define

$$\mathcal{S}_t = \mathcal{S}_t^{i(t)} \qquad \text{and} \qquad \mathcal{U}_t = [t-1] \setminus \mathcal{S}_t \tag{3.1}$$

for the set of timestamps that are available (resp. unavailable) to the active agent at time t.

In a slight abuse of terminology, we will refer to both $(\mathcal{S}_t^i)_{t\in[T]}$ and $(\mathcal{S}_t)_{t\in[T]}$ as feedback sequences although, strictly speaking, they only contain the *timestamps* of the corresponding feedback. Clearly, the non-delayed setting corresponds to the case $\mathcal{S}_t = \mathcal{S}_t^i = [t-1]$ and $\mathcal{U}_t = \varnothing$.

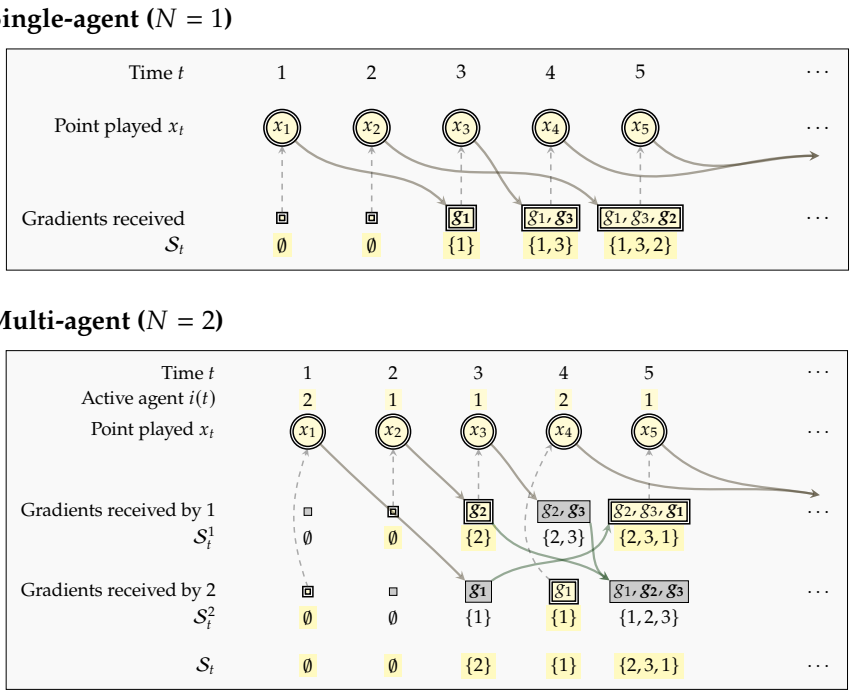


Figure 3.3: Illustration of the type of feedback sequences that may occur in a multi-agent setting. In the standard single-agent case, the feedback sequence $(\mathcal{S}_t)_{t\in\mathbb{N}}$ is necessarily non-decreasing: even though the feedback may not arrive with the same order as the corresponding actions, the *number* of available gradients can only grow. This no longer holds when multiple agents are involved in the optimization process.

3.1.2 *Non-Monotonicity of Feedback Sequence and Lack of Synchronization*

Non-monotonicity of feedback

We now highlight two prominent features of our asynchronous online optimization framework that distinguish it from the large corpus of literature on *single-agent* online learning with delays. First, from the point of view of *any* single agent i, the feedback sequence $(\mathcal{S}_t^i)_{t\in[T]}$ is non-decreasing by definition, i.e., $\mathcal{S}_t^i \subseteq \mathcal{S}_{t+1}^i$ for all $t \in \mathbb{N}$. However, this may not be the case for the *active* feedback sequence $(\mathcal{S}_t)_{t\in[T]}$ which is in general *non-monotone*. In fact, due to communication delays, the same element of feedback may not arrive at each node at the same time. Thus, as the active agent differs from one time slot to another, a timestamp contained in $\mathcal{S}_t$ may not belong to $\mathcal{S}_{t+1}$ (see Fig. 3.3 for an illustration). This leads to the first challenge we seek to overcome:

> Challenge I. Design learning algorithms capable of handling non-monotone feedback sequences.

Remark 3.1. We stress here that this issue is inextricably tied to the multi-agent character of our model. In the single-agent case, $\mathcal{S}_t$ is *de facto* monotone, so this problem does not arise.

Lack of global information

Second, in our model the agents only communicate when they exchange the received feedback. Without additional coordination, the network does not maintain any global information about the evolution of the learning process. In particular, for reasons of privacy and information security, we do not assume that agents have access to a global counter that indicates how many actions have been played at any given stage (as this could carry sensitive, identification-prone information). Similarly, other quantities of interest, such as the current

cumulative unavailability D_t defined below, are also unavailable to each agent. This leads to our second challenge:

Challenge II. Dispense of the need to know t or other non-local information.

As shown above, the lack of network synchronization, along with the non-monotonicity of the active feedback sequence, poses crucial challenges to both the design of the algorithms and the accompanying analysis. In face of these, we introduce in Section 3.2.2 an appropriate reordering of time that enables us to go beyond the algorithms developed for the single-agent setting.

QUANTIFYING THE IMPACT OF DELAYS. As illustrated in Fig. 3.3, having multiple agents also means that we can no longer associate a single delay to each individual feedback element. This explains our choice of focusing on the available subgradients instead of the actual delays, which largely simplifies the description of the framework. The delays, in turn, are still implicitly encoded in the sets $(\mathcal{S}_t^i)$. To quantify their effect, it will be convenient to consider the following measures, defined over a time horizon of T:

- The *maximum delay* τ is the longest wait to receive an element of feedback: $\tau = \min\{s : [t-s-1] \subseteq \mathcal{S}_t \text{ for all } t \in [T]\}$. *Maximum delay*
- The *maximum unavailability* ν of the feedback is $\nu = \max_{t\in[T]} \operatorname{card}(\mathcal{U}_t)$. This is the maximum number of subgradients that could have—but otherwise *haven't*—been communicated to an active agent at activation time. It is straightforward to see that $\nu \leq \tau$.[2] *Maximum unavailability*
- The *cumulative unavailability* D_t is given by $D_t = \sum_{s=1}^{t} \operatorname{card}(\mathcal{U}_s)$. This generalizes the sum of delays to the multi-agent case; clearly, $D_t \leq \nu t$. *Cumulative unavailability*

3.2 DELAYED DUAL AVERAGING AND FAITHFUL PERMUTATIONS

In this section, we present *delayed dual averaging*, our main algorithmic template that we use to address the limitations identified in the previous section. We also introduce the notion of *faithful permuation*, which plays a major role in the analysis to come, as illustrated by the template regret bound in Theorem 3.1.

3.2.1 *Delayed Dual Averaging*

To begin, recall that at each time t, the active agent $i(t)$ has access to a collection of *previously received subgradients* $\{g_s : s \in \mathcal{S}_t\}$ where $\mathcal{S}_t \subseteq [t-1]$ represents the set of timestamps corresponding to the subgradients that can be used by the agent to produce x_t. Put differently, if $s \in \mathcal{S}_t$, then $g_s \in \partial \ell_s(x_s)$ could be used in the computation leading to playing x_t at time t.

Our candidate algorithm for this asynchronous setup builds on the (DA) master template. Of course, the formulation (DA) stated previously is not a practical algorithm here since the active agent $i(t)$ only has at its disposal the subgradients $\{g_s : s \in \mathcal{S}_t\}$ at time t. To resolve this, we propose a natural *Delayed dual averaging*

2 For any $t \in [T]$, we have $[t-\tau-1] \subseteq \mathcal{S}_t$ and thus $\mathcal{U}_t = [t-1] \setminus \mathcal{S}_t \subseteq \{t-\tau-1, \ldots, t-1\}$ which consists of τ elements. On the other hand, if, for some reason, one feedback is *lost*, say the first one, then, the maximum delay is $\tau = T-1$ while the maximum unavailability is $\nu = 1$ (suppose all other feedback arrives with no delays), in which case $\nu \ll \tau$.

Algorithm 3.1: (DDA) – from the point of view of agent i

Initialize: $\mathcal{G}_i \leftarrow \emptyset$, $t \leftarrow 1$.
while not stopped **do**
  asynchronously receive feedback g_s from time s
    $\mathcal{G}_i \leftarrow \mathcal{G}_i \cup \{s\}$
    Relay g_s if necessary
  if the agent becomes active, i.e., $i(t) = i$ **then**
    $\mathcal{S}_t \leftarrow \mathcal{G}_i$
    Update η_t and play $x_t = \arg\min_{x \in \mathcal{X}} \sum_{s \in \mathcal{S}_t} \langle g_s, x \rangle + \frac{h(x)}{\eta_t}$
  end if
end while

adaptation of DA that is suitable for this framework, which we refer to as *delayed dual averaging* (DDA).

$$x_t = \arg\min_{x \in \mathcal{X}} \left\{ \sum_{s \in \mathcal{S}_t} \langle g_s, x \rangle + \frac{h(x)}{\eta_t} \right\} = Q\left(-\eta_t \sum_{s \in \mathcal{S}_t} g_s \right). \tag{DDA}$$

In words, (DDA) simply averages all the received feedback in the dual space and maps it back to the action set via the mirror map to form the prediction. Clearly, as long as η_t can be computed locally, (DDA) can indeed be implemented independently by each agent of the network without requiring a global clock; for a pseudo-code implementation, see Algorithm 3.1.

MD and DA are not equally robust to delays

Remark 3.2. Naturally, one might also consider extending (MD) to accommodate this multi-agent setup. However, apart from the finite Bregman diameter limitation that could constrain the algorithm when using non-constant learning rates, the fact that the feedback may arrive in different order to each agent complicates the adoption of the algorithm in this context. In fact, if feedback from different rounds arrives out-of-order, the natural extension of the method would be to incorporate them sequentially following (MD) in the order of arrival. This process, however, would lead to a final output that varies from agent to agent once all the feedback has arrived. This is in stark contrast to (DA), where all gradients contribute to the model with equal weight, ensuring the actions taken by the agents do not deviate too much from each other when the delays are small. In particular, the final output will be identical for all agents once they've received all the feedback.

3.2.2 *Dependencies and Faithful Permutations*

A crucial challenge in (DDA) is the choice of η_t. Indeed, the standard analysis of DA requires the learning rate sequence to be non-increasing, a property that can hardly be ensured in our situation due to the non-monotonicity of the active feedback sequence and the lack of network synchronization. To sidestep this issue, we need to rethink what "time", or the ordering of the timestamps means to (DDA), and how this can be leveraged to construct a valid algorithm.

Dependency graph

Our starting point will be to redefine the algorithm's internal clock (and corresponding learning rate) based exclusively on the active timestamp sets $(\mathcal{S}_t)_{t \in [T]}$. To that end, we will start by viewing each timestamp as a node in a

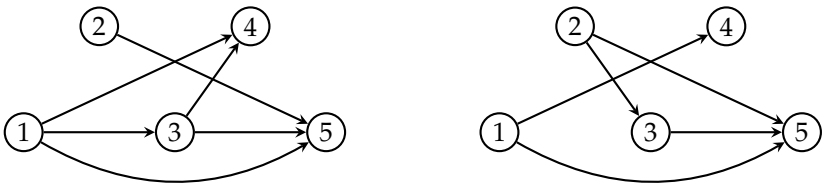


Figure 3.4: The dependency graphs for the two examples of Fig. 3.3. The left and right graphs correspond respectively to the single- and multi-agent examples presented therein. The active feedback at time t is exactly the set of in-neighbors of vertex t.

"causal graph", and we will include a directed edge from s to t if and only if $s \in \mathcal{S}_t$: this represents a "causal dependency" of t on s in the sense that the gradient g_s has been used to define x_t (cf. Fig. 3.4). We will refer to this graph as the *dependency graph* associated to the active feedback sequence $(\mathcal{S}_t)_{t\in[T]}$, and we will denote it by $\mathfrak{G}$; for clarity, we also stress here that we do not assume that this structure is known to the agents.

A first important observation is that the default time ordering $t = 1, 2, \dots$ represents a topological sort of $\mathfrak{G}$, i.e., a linear ordering of its vertices such that $s < t$ for any directed edge $s \rightsquigarrow t$ in $\mathfrak{G}$.[3] Second, since the update structure of (DDA) is determined entirely by $\mathfrak{G}$ and the value of η_t at each vertex of $\mathfrak{G}$, it follows that any reshuffling of time that respects the causal structure of $\mathfrak{G}$ should be an equally viable alternative for the algorithm. We formalize this idea below via the notion of a *faithful permutation*.

Faithful permutation

Definition 3.1 (Faithful permutation)**.** A permutation π of $[T]$ is *faithful* if and only if, for all $s, t \in [T]$, we have

$$s \in \mathcal{S}_t \implies \pi^{-1}(s) < \pi^{-1}(t). \tag{3.2}$$

Equivalently, π is faithful if and only if $\pi(1), \dots, \pi(T)$ is a topological ordering of $\mathfrak{G}$.

Definition 3.1 means that the feedback used at time $\pi(t)$ (whose time indices are in $\mathcal{S}_{\pi(t)}$) form a subset of $\{g_{\pi(1)}, \dots, g_{\pi(t-1)}\}$. To show this, note that if $\pi(s) \in \mathcal{S}_{\pi(t)}$, then $s = \pi^{-1}(\pi(s)) < \pi^{-1}(\pi(t)) = t$, i.e., $s \in [t-1]$. Thus, a faithful permutation can be seen as a reordering of the time that would still be compatible with the feedback used by each agent at every time. We illustrate this notion with two examples below.

Identity permutation

Example 3.1. Clearly, the identity permutation $t \mapsto t$ is always faithful.

Ordering by arrival

Example 3.2. In the single-agent setting, we can define the *ordering by arrival* as follows: if the k-th received subgradient originates from round t—i.e., $g_t \in \partial \ell_t(x_t)$—we set $\pi(k) = t$, so g_t is the $\pi^{-1}(t)$-th received gradient.[4] In this notation, the timestamps of all feedback received *before* g_t can be written as $\mathcal{F}_t := \{\pi(1), \dots, \pi(\pi^{-1}(t) - 1)\}$ for that g_t is the $\pi^{-1}(t)$-th feedback. Along with the inclusion $\mathcal{S}_t \subseteq \mathcal{F}_t$, which holds because g_t is necessarily computed with gradients arriving before itself, we see that π is indeed a faithful permutation.

Remark 3.3. A similar notion was considered by Zimmert and Seldin [297], but for a completely different purpose. There, the authors aimed to provide optimal algorithms for *single-agent* adversarial bandits with delays. They defined a "dependency-preserving permutation" exactly as the inverse of what we call a

3 In particular, this property implies that $\mathfrak{G}$ is a directed acyclic graph (DAG).

4 If multiple gradients arrive at a given round, we resolve ties arbitrarily; this ambiguity in the definition of π plays no role in the analysis.

faithful permutation, and they used this notion to analyze an algorithm that can "skip" certain rounds of feedback when tuning the algorithm's learning rate. Our definition is motivated by—and tailored to—the multi-agent setting, where the non-monotonicity of the active feedback sequence $\mathcal{S}_t$ plays a major role (we recall that this phenomenon cannot arise in the single-agent case). These elements are altogether absent in the single-agent considerations of Zimmert and Seldin [297].

3.2.3 *Bounding the Regret of Delayed Dual Averaging*

We are now in a position to state and prove our main, data-dependent regret guarantee for (DDA) when run with learning rates that are non-increasing *along a faithful permutation*. For simplicity, we assume throughout the sequel that h is non-negative. This is possible because h is strongly convex and we can thus always replace h by the non-negative function $h - \min h$ without affecting our algorithms.

Similar to $[t]$ and $\mathcal{U}_t$, for a faithful permutation π, we also define the set of the first t elements under the new ordering and the set of unavailable elements induced by this ordering as

$$[t]^\pi = \{\pi(1), \dots, \pi(t)\} \quad \text{and} \quad \mathcal{U}_t^\pi = [t-1]^\pi \setminus \mathcal{S}_{\pi(t)}.$$

We have the following theorem concerning the regret of (DDA).

Template regret bound for delayed dual averaging

Theorem 3.1. *Suppose that Assumption 2.1 holds, π is a faithful permutation of $[T]$, and (DDA) is run with learning rates $(\eta_t)_{t\in[T]}$ such that $\eta_{\pi(t+1)} \leq \eta_{\pi(t)}$ for all $t \in [T]$. Then, the algorithm enjoys the regret bound*

$$\mathrm{Reg}_T(z) \leq \frac{h(z)}{\eta_{\pi(T)}} + \frac{1}{2}\sum_{t=1}^{T} \eta_{\pi(t)} \left(\|g_{\pi(t)}\|_*^2 + 2\|g_{\pi(t)}\|_* \sum_{s\in\mathcal{U}_t^\pi} \|g_s\|_* \right). \tag{3.3}$$

Proof. Our analysis leverages the so-called "perturbed iterate" framework for analyzing asynchronous algorithms in the spirit of [185] and [141]. Formally, we define the following virtual iterate sequence

$$\widetilde{x}_t = \operatorname*{arg\,min}_{x\in\mathcal{X}} \sum_{s=1}^{t-1} \langle g_{\pi(s)}, x\rangle + \frac{h(x)}{\eta_{\pi(t)}}.$$

and decompose the sum as:

$$\sum_{t=1}^{T} \langle g_t, x_t - z\rangle = \underbrace{\sum_{t=1}^{T} \langle g_t, \widetilde{x}_{\pi^{-1}(t)} - z\rangle}_{A} + \underbrace{\sum_{t=1}^{T} \langle g_t, x_t - \widetilde{x}_{\pi^{-1}(t)}\rangle}_{B}. \tag{3.4}$$

We now proceed to bound each term separately.

Term A. The first term is exactly the linearized regret of the iterates $\widetilde{x}_1, \dots, \widetilde{x}_T$ that is constructed with the feedback $g_{\pi(1)}, \dots, g_{\pi(T)}$. Thus, as we have shown in the proof of Proposition 2.4, this term can be bounded as

$$\sum_{t=1}^{T} \langle g_t, \widetilde{x}_{\pi^{-1}(t)} - z\rangle = \sum_{t=1}^{T} \langle g_{\pi(t)}, \widetilde{x}_t - z\rangle \leq \frac{h(z)}{\eta_{\pi(T)}} + \frac{1}{2}\sum_{t=1}^{T} \eta_{\pi(t)} \|g_{\pi(t)}\|_*^2. \tag{3.5}$$

Note that the assumption $\eta_{\pi(t+1)} \leq \eta_{\pi(t)}$ on the learning rate sequence is crucial for the derivation of this bound.

Term B. For the second term, we would like to bound the distance between x_t and $\widetilde{x}_{\pi^{-1}(t)}$, or equivalently, the distance between $x_{\pi(t)}$ and $\widetilde{x}_t$ (as we consider all the $t \in \{1, \dots, T\}$). To that end, we write

$$x_{\pi(t)} = Q\left(-\eta_{\pi(t)} \sum_{s \in \mathcal{S}_{\pi(t)}} g_s\right) \quad \text{and} \quad \widetilde{x}_t = Q\left(-\eta_{\pi(t)} \sum_{s \in [t-1]^\pi} g_s\right).$$

Since the permutation π is faithful, we have $\mathcal{S}_{\pi(t)} \subseteq \{\pi(1), .., \pi(t-1)\} = [t-1]^\pi$. We can then use the non-expansivity of the mirror map (Lemma A.4 in Appendix A) to get

$$\|x_{\pi(t)} - \widetilde{x}_t\| \leq \|\eta_{\pi(t)} \sum_{s \in \mathcal{U}_t^\pi} g_s\|_* \leq \eta_{\pi(t)} \sum_{s \in \mathcal{U}_t^\pi} \|g_s\|_*.$$

Subsequently,

$$\begin{aligned} \sum_{t=1}^{T} \langle g_t, x_t - \widetilde{x}_{\pi^{-1}(t)} \rangle &= \sum_{t=1}^{T} \langle g_{\pi(t)}, x_{\pi(t)} - \widetilde{x}_t \rangle \\ &\leq \sum_{t=1}^{T} \|g_{\pi(t)}\|_* \|x_{\pi(t)} - \widetilde{x}_t\| \\ &\leq \sum_{t=1}^{T} \eta_{\pi(t)} \|g_{\pi(t)}\|_* \sum_{s \in \mathcal{U}_t^\pi} \|g_s\|_*. \end{aligned} \tag{3.6}$$

Combining (3.4), (3.5) and (3.6), we obtain the desired result. □

Theorem 3.1 provides a template regret bound that forms the basis of all the upcoming analysis of this chapter. To begin, we note that the bound (3.3) consists of the usual bound for DA (cf. Proposition 2.4) plus a term containing $\sum_{s \in \mathcal{U}_t^\pi} \|g_s\|_*$ that reflects the impact of delay. Similar decompositions have been proven by McMahan and Streeter [188], Joulani et al. [139] and Joulani et al. [141] respectively for online gradient descent, online mirror descent, and dual averaging.[5] These papers focused on the single-agent (shared-memory) setting and conducted the analysis by either choosing π as the identity or the ordering by arrival. Theorem 3.1 thus extends these results by providing a larger class of possible learning rate policies, which enables us to devise efficient and truly implementable learning rate update schemes for the fully decentralized setting in Section 3.3.

3.2.4 *Constant Learning Rate and Lags*

To get an idea of the optimal regret that the algorithm can achieve, we fix a *Cumulative lag*

5 In [188], the authors work with the specific setting of coordinate-wise unconstrained gradient methods. Therefore, instead of products of norms they have products of scalars in their analysis.

constant learning rate $\eta_t \equiv \eta$, which we subsequently optimize to minimize the upper-bound on the regret. To proceed, we define the *cumulative lag* as

$$\Lambda_t^\pi = \sum_{s=1}^{t} \left(\|g_{\pi(s)}\|_*^2 + 2\|g_{\pi(s)}\|_* \sum_{l \in \mathcal{U}_s^\pi} \|g_l\|_* \right) = \sum_{s \in [t]^\pi} \|g_s\|_*^2 + 2 \sum_{\{s,l\} \in \mathcal{D}_t^\pi} \|g_s\|_* \|g_l\|_*, \tag{3.7}$$

where

$$\mathcal{D}_t^\pi = \{\{\pi(s), l\} : s \in [t], l \in \mathcal{U}_s^\pi\}.$$

In words, $\{s', l\} \in \mathcal{D}_t^\pi$ if (*i*) g_l is not used to define $x_{s'}$; and (*ii*) after reordering by π, l comes before s' and s' comes before $\pi(t)$. We also write $\Lambda_t = \Lambda_t^{\mathrm{id}}$ for the lag associated to the standard time ordering and define $D_t^\pi = \mathrm{card}(\mathcal{D}_t^\pi) = \sum_{s=1}^T \mathrm{card}(\mathcal{U}_s^\pi)$ for the cumulative unavailability under the order induced by π.

Compared to D_t^π, the cumulative lag Λ_t^π regroups the actual errors caused by the inability of the learners to compensate the missing feedback, and gives the most fine-grained characterization of the effect of delayed feedback on the regret. In the single-agent setting, Joulani et al. [139] and McMahan and Streeter [188] also considered the same quantity but in the special case where π is the ordering-by-arrival permutation discussed in Section 3.2.2. In general, it is clear that $\Lambda_t^\pi \leq (t + 2D_t^\pi)G^2$ provided that all subgradients are bounded in norm by G (Assumption 2.2); moreover, if π is the identity permutation, we further have $D_t^\pi = D_t \leq \nu t$. With all this in mind, a direct application of Theorem 3.1 gives the following series of more explicit bounds.

Regret bounds with constant learning rate

Corollary 3.2. *Suppose that Assumption 2.1 holds and* (DDA) *is run with a constant learning rate $\eta > 0$. Then, for any faithful permutation π, we have*

- *If $\|g_t\|_*$ is uniformly bounded (Assumption 2.2) and $\eta = \Theta\left(1/\sqrt{\max(1,\nu)T}\right)$, then $\mathrm{Reg}_T(z) = O\left(\sqrt{\max(1,\nu)T}\right)$.*
- *If $\|g_t\|_*$ is uniformly bounded (Assumption 2.2) and $\eta = \Theta\left(1/\sqrt{\max(T, D_T^\pi)}\right)$, then $\mathrm{Reg}_T(z) = O\left(\sqrt{\max(T, D_T^\pi)}\right)$.*
- *If $\eta = \Theta\left(1/\sqrt{\Lambda_T^\pi}\right)$, then $\mathrm{Reg}_T(z) = O\left(\sqrt{\Lambda_T^\pi}\right)$.*

Corollary 3.2 recapitulates several types of regret bounds that we can expect from (DDA), depending on the tuning of η_t (either by using a pessimistic upper bound on the delays and the norms of the gradients, or using the actual delays and/or received gradients). Specifically, if we focus on the standard time ordering $\pi = \mathrm{id}$, Corollary 3.2 allows us to recover the optimal data-dependent bound of $O(\sqrt{\Lambda_T})$ that was previously obtained for the single-agent setting by Joulani et al. [139] and McMahan and Streeter [188].

Going further, if we assume that $\|g_t\|_* \leq G$ for all $t \in [T]$, we have $\Lambda_T \leq (T + 2D_T)G^2$, which leads to the well-known $O(\sqrt{D_T})$ bound on the regret (see e.g., Quanrud and Khashabi [226]). Finally, if we only tune our learning rate based the maximum unavailability ν (it can also be an upper bound thereof), we get a regret in $O(\sqrt{\nu T})$. In the single-agent case, this is equivalent to the $O(\sqrt{\tau T})$ regret bound shown by the pioneering works of Langford et al. [165] and Weinberger and Ordentlich [279].

On the downside, Corollary 3.2 would seem to suggest that the derived regret bounds depend on the choice of the permutation π, a concept that is relevant for the analysis, but which is otherwise devoid of physical meaning (at

least, relative to the sequence of events as it unfolds in real time). Because of this, the computation of the optimal learning rates required by Corollary 3.2 seems beyond reach in practice—even if we assume that the various quantities involved are somehow known to the agents. However, as we show below, *this is not the case:* the values of both D_T^π and Λ_T^π are independent of π, and hence, so are the bounds of Corollary 3.2. To prove this, we first provide a new characterization of the set $\mathcal{D}_t^\pi$ which is of independent interest:

Alternative characterization of the set $\mathcal{D}_t^\pi$

Proposition 3.3. *Let π be a faithful permutation. Then*

$$\mathcal{D}_t^\pi = \{\{s,l\} \subseteq [t]^\pi : s \text{ and } l \text{ are not adjacent in } \mathfrak{G}\}. \tag{3.8}$$

Proof. By definition of the dependency graph, s and l are not adjacent in $\mathfrak{G}$ if and only if $\{s \notin \mathcal{S}_l, l \notin \mathcal{S}_s\}$. We will thus show that

$$\mathcal{D}_t^\pi = \{\{s,l\} \subseteq [t]^\pi : s \notin \mathcal{S}_l, l \notin \mathcal{S}_s\}.$$

This relies on a two-way inclusion argument.

Inclusion ("$\subseteq$"). Let $s \in [t]$ and $l \in \mathcal{U}_s^\pi = [s-1]^\pi \setminus \mathcal{S}_{\pi(s)}$. By definition of $[t]^\pi$ we have $\pi(s) \in [t]^\pi$ and $l \in [s-1]^\pi \subseteq [t]^\pi$. It remains to prove that $\pi(s) \notin \mathcal{S}_l$. We exploit the equivalence

$$\begin{aligned} l \in [s-1]^\pi &\iff \pi^{-1}(l) \leq s-1 \\ &\iff \pi^{-1}(l) < \pi^{-1}(\pi(s)) \\ &\iff \pi(s) \notin [\pi^{-1}(l)]^\pi. \end{aligned} \tag{3.9}$$

To conclude, we use the fact that π is a faithful permutation and accordingly $\mathcal{S}_l \subseteq [\pi^{-1}(l)-1]^\pi \subseteq [\pi^{-1}(l)]^\pi$. Along with (3.9) we deduce that $\pi(s) \notin \mathcal{S}_l$.

Containment ("$\supseteq$"). Let $\{s,l\} \subset [t]^\pi$ such that $s \notin \mathcal{S}_l$ and $l \notin \mathcal{S}_s$. We assume without loss of generality $\pi^{-1}(l) < \pi^{-1}(s)$. This is equivalent to $l \in [\pi^{-1}(s)-1]^\pi$ and therefore $l \in \mathcal{U}_{\pi^{-1}(s)}^\pi$. We complete the proof by noting that $s \in [t]^\pi$ if and only if $\pi^{-1}(s) \in [t]$. □

In contrast to the original definition of $\mathcal{D}_t^\pi$, the characterization of Proposition 3.3—i.e., the non-adjacency of the vertices—is independent of the ordering of the timestamps. By defining $\mathfrak{G}_t^\pi$ as the subgraph of $\mathfrak{G}$ spanned by the vertices of $[t]^\pi$ in $\mathfrak{G}$, the proposition says that $\mathcal{D}_t^\pi$ contains exactly the non-adjacent vertex pairs of $\mathfrak{G}_t^\pi$. With this in mind, we readily obtain the following important corollary:

Independence of the regret bounds in relation to π

Corollary 3.4. *For any two faithful permutations π and π', we have $\mathcal{D}_T^\pi = \mathcal{D}_T^{\pi'}$, and, a fortiori, $D_T^\pi = D_T^{\pi'}$ and $\Lambda_T^\pi = \Lambda_T^{\pi'}$. In other words, the regret bounds of Corollary 3.2 are independent of π.*

Proof. Simply note that $[T]^\pi = [T]^{\pi'} = [T]$. □

Corollary 3.4 shows that the regret bounds of Corollary 3.2 are indeed meaningful, as they do not depend on any "virtual" reordering of time by a faithful permutation. However, given that the quantities Λ_T and D_T cannot be assumed known beforehand, the agents might need to employ a much more conservative learning rate of the order of $\Theta(1/\sqrt{vT})$ to minimize their regret. We address this important issue via the design of suitable adaptive learning methods in the next section.

3.3 TUNING THE LEARNING RATE IN THE PRESENCE OF DELAYS

Requirements for our adaptive algorithms

In this section, we exploit the template bound of Theorem 3.1 to design efficient leaning rates that provably achieve low regret. To clarify our objective, we begin by identifying the main desiderata that we seek to achieve:

Anytime

- **Anytime / Restart-free:** the algorithm should not require the knowledge of the horizon T and/or include a restart schedule where previous information is discarded.

Coordination-free

- **Coordination-free:** the learning rate of each agent must be computable based *exclusively* on local information without any need for coordination.

Data-dependent

- **Data-dependent bounds:** the algorithm's regret guarantees should feature the actual gradients observed instead of an upper bound thereof.

Delay-dependent

- **Adaptivity to delays:** the algorithm's regret should depend on the observed delays and not only on a pessimistic, worst-case estimate thereof.

The naive adaptation of AdaGrad is not implementable

To derive a learning rate with the above properties, we draw inspiration from the (AdaGrad-norm) policy. This is perhaps the easiest to illustrate in the case $\pi = \mathrm{id}$: here, to obtain an $O(\sqrt{\Lambda_T})$ regret, we could employ the policy $\eta_t = 1/\sqrt{\Lambda_t} = 1/\sqrt{\sum_{s=1}^{t} \lambda_s}$ where

$$\lambda_s = \|g_s\|_*^2 + 2\|g_s\|_* \sum_{l \in \mathcal{U}_s} \|g_l\|_*.$$

The key in the analysis of this policy is provided by Lemma 2.6. Based on this lemma, it is straightforward to show that (DDA) with learning rate $\eta_t = 1/\sqrt{\Lambda_t}$ incurs at most $O(\sqrt{\Lambda_T})$ regret. However, this policy is not implementable because it involves unobserved feedback—and hence violates one of our principal desiderata. In the rest of this section, we show how this difficulty can be circumvented in many relevant scenarios.

3.3.1 *Pessimistic Non-Adaptive Learning Rate*

Bounded delay and bounded feedback

To set the stage for the analysis to come, we begin by assuming that the agents know τ an upper bound on the maximum delay and G an upper bound on the norms of the observed gradients. This leads to $\lambda_s \leq G^2(1+2\nu) \leq G^2(1+2\tau)$, and subsequently $\Lambda_t \leq G^2(1+2\tau)t$. Given this preliminary result, it is tempting to choose $\eta_t = \Theta(1/G\sqrt{t(1+2\tau)})$. This is however still unrealistic as the agents do not know the exact value of t, and may only estimate it by using $t \leq \mathrm{card}(\mathcal{S}_t) + \tau + 1$. To justify this strategy, we need to prove that the corresponding learning rate is indeed non-increasing along some faithful permutation in order to apply Theorem 3.1. For this, we make the following assumption.

Assumption on number of received feedback elements

Assumption 3.1. If $s \in \mathcal{S}_t$, then $\mathrm{card}(\mathcal{S}_s) < \mathrm{card}(\mathcal{S}_t)$.

In words, the assumption requires that if g_s is used to compute x_t, then x_s is computed with fewer gradients than x_t. This is a fairly mild requirement which is in turn implied by the upcoming Assumption 3.2 (see the accompanying discussion). In particular, if the agents relay the information $\mathrm{card}(\mathcal{S}_t)$ as well, Assumption 3.1 can be ensured by delaying the actual usage of a received

feedback when necessary.[6] Then, when the actual delays are bounded by τ, the gradients $\{g_1, \dots, g_{t-\tau-1}\}$ can always be used for computing x_t. Therefore, introducing this extra delay will not increase the maximum delay and has no effect on the regret bound of the following proposition.

Regret for anytime but non-adaptive learning rate

Proposition 3.5. *Suppose that Assumptions 2.1, 2.2 and 3.1 hold and the maximum delay is bounded by τ. Assume further that (DDA) is run with the learning rate*

$$\eta_t = \frac{R}{G\sqrt{(1+2\tau)(\operatorname{card}(\mathcal{S}_t)+\tau+1)}}.$$

Then, for any z such that $h(z) \le R^2$, the algorithm enjoys the regret bound

$$\operatorname{Reg}_T(z) \le 2RG\sqrt{(T+\tau)(1+2\tau)}.$$

Proof. We will in fact prove a stronger variant for which it is sufficient to assume that τ is an upper bound on the maximum unavailability (denoted by ν previously). Let $\overline{\Lambda}_t = G^2(1+2\tau)(\operatorname{card}(\mathcal{S}_t)+\tau+1)$ so that $\eta_t = R/\sqrt{\overline{\Lambda}_t}$. We choose a permutation π that satisfies if $\overline{\Lambda}_s < \overline{\Lambda}_t$ then $\pi^{-1}(s) < \pi^{-1}(t)$ (we just need to sort the time indices using $\overline{\Lambda}_t$ and map to this new order). From Assumption 3.1 and the definition of $\overline{\Lambda}_t$ we know that π is a faithful permutation. Moreover, $(\overline{\Lambda}_t)_t$ is non-decreasing along π: indeed, if this were not the case—that is, if $\overline{\Lambda}_{\pi(t+1)} < \overline{\Lambda}_{\pi(t)}$ for some t—we would have $t+1 = \pi^{-1}(\pi(t+1)) < \pi^{-1}(\pi(t)) = t$, a contradiction.

We now proceed to prove $\operatorname{card}(\mathcal{U}_t^\pi) \le \tau$, or equivalently $\operatorname{card}(\mathcal{S}_{\pi(t)}) \ge t-1-\tau$. For this we show $[t]^\pi \subseteq [\operatorname{card}(\mathcal{S}_{\pi(t)})+\tau+1]$, which implies $t \le \operatorname{card}(\mathcal{S}_{\pi(t)})+\tau+1$ and thus the above inequality. Provided that $\overline{\Lambda}_t$ is non-decreasing along π, for $s \le t$ we have $\operatorname{card}(\mathcal{S}_{\pi(s)}) \le \operatorname{card}(\mathcal{S}_{\pi(t)})$. Using the bounded unavailability assumption we get $\operatorname{card}([\pi(s)-1] \setminus \mathcal{S}_{\pi(s)}) \le \tau$ so that $\pi(s)-1-\operatorname{card}(\mathcal{S}_{\pi(s)}) \le \tau$ and subsequently $\pi(s) \le \operatorname{card}(\mathcal{S}_{\pi(t)})+\tau+1$. This proves $[t]^\pi \subseteq [\operatorname{card}(\mathcal{S}_{\pi(t)})+\tau+1]$.

From $\operatorname{card}(\mathcal{U}_t^\pi) \le \tau$ it follows immediately that for all t

$$\lambda_t^\pi := \|g_{\pi(t)}\|_*^2 + 2\|g_{\pi(t)}\|_* \sum_{s\in\mathcal{U}_t^\pi} \|g_s\|_* \le G^2(1+2\tau).$$

Along with $t \le \operatorname{card}(\mathcal{S}_{\pi(t)})+\tau+1$ we deduce

$$\Lambda_t^\pi \le G^2(1+2\tau)(\operatorname{card}(\mathcal{S}_{\pi(t)})+\tau+1) = \overline{\Lambda}_{\pi(t)}.$$

6 In this case, $\mathcal{S}_t$ refers to the timestamps of the gradients that are used for the computation of x_t; however, this does not necessarily contain all the gradients that the active agent $i(t)$ has received by time t.

Applying Theorem 3.1 and Lemma 2.6, we then obtain

$$\begin{aligned}\mathrm{Reg}_T(z) &\le \frac{h(z)}{\eta_{\pi(T)}} + \frac{1}{2}\sum_{t=1}^{T}\eta_{\pi(t)}\left(\|g_{\pi(t)}\|_*^2 + 2\|g_{\pi(t)}\|_*\sum_{s\in\mathcal{U}_t^\pi}\|g_s\|_*\right)\\ &\le R\sqrt{\Lambda_{\pi(T)}} + \frac{R}{2}\sum_{t=1}^{T}\frac{\lambda_t^\pi}{\sqrt{\Lambda_{\pi(t)}}}\\ &\le R\sqrt{\Lambda_{\pi(T)}} + \frac{R}{2}\sum_{t=1}^{T}\frac{\lambda_t^\pi}{\sqrt{\Lambda_t^\pi}}\\ &\le R\sqrt{\Lambda_{\pi(T)}} + R\sqrt{\Lambda_T^\pi} \le 2R\sqrt{\Lambda_{\pi(T)}}.\end{aligned}$$

Our assertion follows by noting that $\mathrm{card}(\mathcal{S}_{\pi(T)}) \le \pi(T) - 1 \le T - 1$. □

Proposition 3.5 shows that, even in the fully decentralized case where no global counter is available, it is *still* possible to design implementable algorithms that retain the optimal $O(\sqrt{\tau T})$ regret bound. Our next step is to further improve the algorithm so that it can adapt to both the *data* and the *delay* of the feedback. The aforementioned characterization of delay will turn out to be crucial for this.

3.3.2 *Adaptation to Delays in Distributed Systems*

Ordering by arrival at each agent

To design a learning rate policy that adapts to both data and delays, we have to find a way to estimate Λ_t by only using local information of each agent. To that end, define for each agent i the *individual* ordering by arrival as a permutation π_i of $[T]$ such that the k-th received feedback of i comes from $x_{\pi_i(k)}$ (played by i or another player), i.e., the k-th received feedback of i is $g_{\pi_i(k)} \in \partial\ell_{\pi_i(k)}(x_{\pi_i(k)})$. With this notation, we can define the set of all feedback received *before* g_t *by agent* i; since g_t is the $\pi_i^{-1}(t)$-th feedback, this set is defined as $\mathcal{F}_t^i := \{\pi_i(1), \pi_i(2), \dots, \pi_i(\pi_i^{-1}(t) - 1)\}$.

Approximation of cummulative lag

Using these definitions and looking closely at the definition of the lag (3.7), we notice that:

1. The quantity $\sum_{s=1}^{t}\|g_{\pi(s)}\|_*^2$ cannot be known at instant $\pi(t)$ since the set of gradients available at that time is $\mathcal{S}_{\pi(t)}$. It is thus natural to consider approximating it by $\sum_{s\in\mathcal{S}_{\pi(t)}}\|g_s\|_*^2$.

2. For each t the quantity $\sum_{\{s,l\}\in\mathcal{D}_t^\pi}\|g_s\|_*\|g_l\|_*$, gathering the pairs of feedback of $[t]^\pi$ satisfying the relation $\{s \notin \mathcal{S}_l, l \notin \mathcal{S}_s\}$ (Proposition 3.3), is generally unknown. Building on the works of Joulani et al. [139] and McMahan and Streeter [188], this sum can be approximated by $\sum_{s\in\mathcal{S}_{\pi(t)}}(\|g_s\|_*\sum_{l\in\mathcal{F}_s^{i(\pi(t))}\setminus\mathcal{S}_s}\|g_l\|_*)$. In words, for all $s \in \mathcal{S}_{\pi(t)}$, the worker $i(\pi(t))$ aggregates the feedback received before g_s but was not used to generate g_s.

Putting these two points together, a reasonable surrogate for Λ_t^π would be $\Gamma_{\pi(t)}$, where for all $t \in [T]$, we define

$$\Gamma_t = \sum_{s\in\mathcal{S}_t}\left(\|g_s\|_*^2 + 2\|g_s\|_*\sum_{l\in\mathcal{F}_s^{i(t)}\setminus\mathcal{S}_s}\|g_l\|_*\right).$$

To make Γ_t a valid approximation, we would need s to satisfy $s \notin \mathcal{S}_l$ whenever $l \in \mathcal{F}_s^{i(t)} \setminus \mathcal{S}_s$ given the characterization of Proposition 3.3. In particular, this is true if g_s is not used to generate x_l whenever g_l arrives before g_s at node $i(t)$. This leads to the following mild assumption: when an agent receives a gradient g_t, they must have already received all the feedback used to compute it.

Assumption on composition of feedback elements

Assumption 3.2. For every worker $i \in \mathcal{N}$ and all $t = 1, 2, \dots$, we have $\mathcal{S}_t \subseteq \mathcal{F}_t^i$.

The above assumption is notably verified in the following scenarios: (*i*) a coordinator-worker scheme in which the transmission of the gradients occurs *in order*, in first-come, first-serve manner; (*ii*) broadcasting of newly received and computed gradient over a fixed communication network; (*iii*) whenever two agents communicate their gradient pools are synchronized and the gradients are exchanged in the order they become available to the agents. As a consequence, Assumption 3.2 is satisfied in many relevant setups and can otherwise be enforced by imposing *iii)* at the price of a slightly higher communication cost.

Now, since the active agent $i(t)$ at time t knows $\mathcal{S}_t$ (by definition) and $\mathcal{F}_s^{i(t)}$ for $s \in \mathcal{S}_t$ (by construction), the quantity Γ_t is indeed computable with purely local information. The agents can thus run (DDA) with a learning rate of the form $\eta_t = \Theta(1/\sqrt{\Gamma_t})$. The obtained algorithm, which we call AdaDelay-Dist, is detailed in Algorithm 3.2; its principal regret guarantee is given below:

Regret for adaptive learning rate

Theorem 3.6. *Suppose that Assumptions 2.1, 2.2 and 3.2 hold and the maximum delay is bounded by τ. Assume further that* (DDA) *is run with the learning rate*

$$\eta_t = \frac{R}{\sqrt{\Gamma_t + \beta}}, \tag{AdaDelay–Dist}$$

where $\beta > 0$ is a positive constant. Then, for all z such that $h(z) \leq R^2$, the algorithm enjoys the regret bound

$$\mathrm{Reg}_T(z) \leq 2R\sqrt{\Lambda_T} + 2R\sqrt{\beta} + \frac{R}{\sqrt{\beta}} G^2 (2\tau + 1)^2.$$

The bound of Theorem 3.6 differs from the optimal data-dependent bound by at most a time-independent constant, and this is achieved at the worst-case cost of transmitting an additional scalar (i.e., $\sum_{s \in \mathcal{S}_t} \|g_s\|_*$) per element of feedback sent. Moreover, we should also stress that, similar to the learning rate in Proposition 3.5, (AdaDelay–Dist) does not use the global time. Time indices are present in Algorithm 3.2 only for ease of comprehension, notably to highlight the fact that a worker knows (and keeps track) of the feedback used to produce past points (i.e., $\sum_{s \in \mathcal{S}_t} \|g_s\|_*$ for each point x_t played by the worker). Finally, notice that although the theorem assumes the gradients and delays to be bounded, the algorithm itself does *not* require any knowledge of these bounds. A bad estimate of these quantities would only cause the method to suffer from higher regret at the first iterations.

We now proceed to prove Theorem 3.6. For this, let $\mathcal{A}_t^i := \{\{s, l\} : s \in \mathcal{S}_t, l \in \mathcal{F}_s^i \setminus \mathcal{S}_s\}$ so that

$$\Gamma_t = \sum_{s \in \mathcal{S}_t} \left(\|g_s\|_*^2 + 2\|g_s\|_* \sum_{l \in \mathcal{F}_s^{i(t)} \setminus \mathcal{S}_s} \|g_l\|_* \right) = \sum_{s \in \mathcal{S}_t} \|g_s\|_*^2 + 2 \sum_{\{s,l\} \in \mathcal{A}_t^{i(t)}} \|g_s\|_* \|g_l\|_* \tag{3.10}$$

Algorithm 3.2: AdaDelay–Dist – from the point of view of agent i

Initialize: $\mathcal{G}_i \leftarrow \emptyset$, $\Gamma^i \leftarrow \beta > 0$
while not stopped **do**
  asynchronously receive g_t along with $\sum_{s\in\mathcal{S}_t}\|g_s\|_*$ from other agents
    $\Gamma^i \leftarrow \Gamma^i + \|g_t\|_*^2 + 2\|g_t\|_*(\sum_{s\in\mathcal{G}^i}\|g_s\|_* - \sum_{s\in\mathcal{S}_t}\|g_s\|_*)$
    $\mathcal{G}^i \leftarrow \mathcal{G}^i \cup \{t\}$
    Relay the information if necessary
  asynchronously receive g_t as a feedback
    Retrieve $\sum_{s\in\mathcal{S}_t}\|g_s\|_*$ from the memory
    $\Gamma^i \leftarrow \Gamma^i + \|g_t\|_*^2 + 2\|g_t\|_*(\sum_{s\in\mathcal{G}^i}\|g_s\|_* - \sum_{s\in\mathcal{S}_t}\|g_s\|_*)$
    $\mathcal{G}^i \leftarrow \mathcal{G}^i \cup \{t\}$
    Send g_t and $\sum_{s\in\mathcal{S}_t}\|g_s\|_*$ to other agents
  if the agent becomes active, i.e., $i(t) = i$ **then**
    $\mathcal{S}_t \leftarrow \mathcal{G}_i$
    $\eta_t \leftarrow R/\sqrt{\Gamma^i}$
    Play $x_t = \arg\min_{x\in\mathcal{X}} \sum_{s\in\mathcal{S}_t}\langle g_s, x\rangle + \frac{h(x)}{\eta_t}$
  end if
end while

To simplify the notation, we will write $\mathcal{A}_t = \mathcal{A}_t^{i(t)}$. In the following proposition, we show that $\mathcal{A}_t$ can be characterized in the same way as $\mathcal{D}_t^\pi$.

Alternative characterization of the set $\mathcal{A}_t$

Proposition 3.7. *Let π be a faithful permutation and let Assumption 3.2 hold. Then*

$$\mathcal{A}_t = \{\{s,l\} \subseteq \mathcal{S}_t : s \text{ and } l \text{ are not adjacent in } \mathfrak{G}\}$$

Proof. The proof is similar to that of Proposition 3.3. We prove

$$\mathcal{A}_t = \{\{s,l\} \subseteq \mathcal{S}_t : s \notin \mathcal{S}_l, l \notin \mathcal{S}_s\}$$

by a two-way inclusion argument.

Inclusion ("$\subseteq$"). Let $s \in \mathcal{S}_t$ and $l \in \mathcal{F}_s^{i(t)} \setminus \mathcal{S}_s$. The inclusion $l \in \mathcal{F}_s^{i(t)}$ means that g_l arrives earlier than g_s on node $i(t)$. As all the available gradients are used when playing x_t and $s \in \mathcal{S}_t$, we deduce $l \in \mathcal{S}_t$. On the other hand, $l \in \mathcal{F}_s^{i(t)}$ also implies $s \notin \mathcal{F}_l^{i(t)}$. Using Assumption 3.2 we know that $\mathcal{S}_l \subseteq \mathcal{F}_l^{i(t)}$, and consequently $s \notin \mathcal{S}_l$.

Containment ("$\supseteq$") Let $\{s,l\} \subseteq \mathcal{S}_t$ such that $s \notin \mathcal{S}_l$ and $l \notin \mathcal{S}_s$. Since either $l \in \mathcal{F}_s^{i(t)}$ or $s \in \mathcal{F}_l^{i(t)}$ (but not both) we conclude immediately $\{s,l\} \in \mathcal{A}_t$. □

Thanks to Proposition 3.3 and Proposition 3.7, comparing $\mathcal{D}_t^\pi$ with $\mathcal{A}_{\pi(t)}$ amounts to comparing $[t]^\pi$ with $\mathcal{S}_{\pi(t)}$. Using the bounded delay assumption, we can prove the following properties on a faithful permutation.

Properties of a faithful permutation

Proposition 3.8. *Let π be a faithful permutation and assume that the maximum delay is bounded by τ. We have*

(a) $[t]^\pi \subseteq [\pi(t) + \tau]$.

(b) $[t]^\pi \setminus \mathcal{S}_{\pi(t)} \subseteq \{\pi(t) - \tau, ..., \pi(t) + \tau\}$.

(c) $|\pi(t) - t| \leq \tau$.

Proof. (*a*) Let $s, t \in [T]$ such that $s \leq t$. We claim that $\pi(s) \leq \pi(t) + \tau$. Assume the opposite, that is, $\pi(s) > \pi(t) + \tau$. Then, from the bounded delay assumption, $\pi(t) \in \mathcal{S}_{\pi(s)}$. π being a faithful permutation, this implies $t = \pi^{-1}(\pi(t)) < \pi^{-1}(\pi(s)) = s$, a contradiction, and hence the claim. To prove the inclusion, note that $[t]^\pi = \{\pi(1), \dots, \pi(t)\} = \{\pi(s) : s \leq t\}$ and we thus have $[t]^\pi \subseteq [\pi(t) + \tau]$ with the aforementioned claim.

(*b*) This is immediate from (*a*) and the inclusion $[\pi(t) - \tau - 1] \subseteq \mathcal{S}_{\pi(t)}$ which holds since the maximum delay is assumed to be bounded by τ.

(*c*) Fix $t \in [T]$. For all $s \leq t$, we have $\pi(s) \leq \pi(t) + \tau$ and therefore $\max_{s \leq t} \pi(s) \leq \pi(t) + \tau$. π being a permutation of $[T]$, it holds $\max_{s \leq t} \pi(s) \geq t$ and subsequently $t \leq \pi(t) + \tau$. Similarly, we also have $\pi(t) - \tau \leq \min_{t \leq s} \pi(s)$ and $\min_{t \leq s} \pi(s) \leq t$. This implies $\pi(t) - \tau \leq t$. Combining the two we conclude $|\pi(t) - t| \leq \tau$. □

Interestingly, Proposition 3.8(*c*) shows that when the delays are bounded by τ, a faithful permutation can at most move an element τ steps away from its original position. We are now ready to provide the complete proof of Theorem 3.6.

Proof of Theorem 3.6. Let $\overline{\Lambda}_t = \Gamma_t + \beta$ so that $\eta_t = R/\sqrt{\overline{\Lambda}_t}$ and π be a permutation such that (*i*) if $\overline{\Lambda}_s < \overline{\Lambda}_t$ then $\pi^{-1}(s) < \pi^{-1}(t)$; (*ii*) if $\overline{\Lambda}_s = \overline{\Lambda}_t$ and $s \in \mathcal{S}_t$ then $\pi^{-1}(s) < \pi^{-1}(t)$. $(\overline{\Lambda}_t)_t$ is obviously non-decreasing along π (see proof of Proposition 3.5). We claim that this is a faithful permutation. For this, let $s \in \mathcal{S}_t$ and we would like to show $\pi^{-1}(s) < \pi^{-1}(t)$. By Assumption 3.2 we have $\mathcal{S}_s \subseteq \mathcal{F}_s^{i(t)}$ and from $s \in \mathcal{S}_t$ it holds $\mathcal{F}_s^{i(t)} \subseteq \mathcal{S}_t$; accordingly, $\mathcal{S}_s \subseteq \mathcal{S}_t$. Invoking Proposition 3.7 we deduce $\mathcal{A}_s \subseteq \mathcal{A}_t$. Using (3.10) we then get $\overline{\Lambda}_s \leq \overline{\Lambda}_t$. This inequality along with $s \in \mathcal{S}_t$ imply $\pi^{-1}(s) < \pi^{-1}(t)$.

In the remainder of the proof, we will use the notation $\Gamma_t = \Lambda_T = \Lambda_T^\pi$ for $t > T$. Let us prove that $\Gamma_{\pi(t)+2\tau+1} \geq \Lambda_t^\pi$ for $t \in [T]$. This is the case when $\pi(t) + 2\tau + 1 > T$ by the previous definition. Otherwise, with (3.7), (3.10), Propositions 3.3 and 3.7, this is equivalent to proving that $[t]^\pi \subseteq \mathcal{S}_{\pi(t)+2\tau+1}$. The inclusion holds since on one hand, by Proposition 3.8(a) we have $[t]^\pi \subseteq [\pi(t) + \tau]$ and on the other hand $[\pi(t) + \tau] \subseteq \mathcal{S}_{\pi(t)+2\tau+1}$ by the definition of maximum delay.

As we have proved that π is a faithful permutation, it holds $\mathcal{S}_{\pi(t)} \subseteq [t]^\pi$. The above hence also implies $\mathcal{S}_{\pi(t)} \subseteq \mathcal{S}_{\pi(t)+2\tau+1}$, and accordingly, $\overline{\Lambda}_{\pi(t)+2\tau+1} \geq \overline{\Lambda}_{\pi(t)}$. The inequality is still true when $\pi(t) + 2\tau + 1 > T$ as $\Gamma_t \leq \Lambda_T$ always holds by Propositions 3.3 and 3.7 and $\mathcal{S}_t \subseteq [T]$. Applying Theorem 3.1 gives

$$\begin{aligned}
\mathrm{Reg}_T(z) &\leq \frac{h(z)}{\eta_{\pi(T)}} + \frac{1}{2}\sum_{t=1}^{T} \eta_{\pi(t)} \left(\|g_{\pi(t)}\|_*^2 + 2\|g_{\pi(t)}\|_* \sum_{s \in \mathcal{U}_t^\pi} \|g_s\|_* \right) \\
&\leq R\sqrt{\overline{\Lambda}_{\pi(T)}} + \frac{R}{2}\sum_{t=1}^{T} \frac{\lambda_t^\pi}{\sqrt{\overline{\Lambda}_{\pi(t)}}} \\
&= R\sqrt{\overline{\Lambda}_{\pi(T)}} + \frac{R}{2}\sum_{t=1}^{T} \left(\frac{1}{\sqrt{\overline{\Lambda}_{\pi(t)+2\tau+1}}} + \frac{1}{\sqrt{\overline{\Lambda}_{\pi(t)}}} - \frac{1}{\sqrt{\overline{\Lambda}_{\pi(t)+2\tau+1}}} \right) \lambda_t^\pi,
\end{aligned}$$

where as in the proof of Proposition 3.5 we write

$$\lambda_t^\pi = \|g_{\pi(t)}\|_*^2 + 2\|g_{\pi(t)}\|_* \sum_{s\in\mathcal{U}_t^\pi} \|g_s\|_*.$$

From Proposition 3.8(b) we know that $[t]^\pi \setminus \mathcal{S}_{\pi(t)} \subseteq \{\pi(t)-\tau, ..., \pi(t)+\tau\}$. Since $[t-1]^\pi = [t]^\pi \setminus \{\pi(t)\}$ and $\pi(t) \notin \mathcal{S}_{\pi(t)}$, we deduce that $\mathrm{card}(\mathcal{U}_t^\pi) \leq 2\tau$ and hence $\lambda_t^\pi \leq G^2(1+4\tau)$. With the non-negativity of $1/\sqrt{\overline{\Lambda}_{\pi(t)}} - 1/\sqrt{\overline{\Lambda}_{\pi(t)+2\tau+1}}$ and the fact that $\Lambda_t^\pi \leq \Gamma_{\pi(t)+2\tau+1} < \overline{\Lambda}_{\pi(t)+2\tau+1}$ we then get

$$\begin{aligned}
\mathrm{Reg}_T(z) &\leq R\sqrt{\overline{\Lambda}_{\pi(T)}} + \frac{R}{2}\sum_{t=1}^T \frac{\lambda_t^\pi}{\sqrt{\Lambda_t^\pi}} + \frac{R}{2}\sum_{t=1}^T \left(\frac{1}{\sqrt{\overline{\Lambda}_{\pi(t)}}} - \frac{1}{\sqrt{\overline{\Lambda}_{\pi(t)+2\tau+1}}}\right) G^2(1+4\tau) \\
&\leq R\sqrt{\overline{\Lambda}_{\pi(T)}} + R\sqrt{\Lambda_T^\pi} + \frac{R}{2}\sum_{t=1}^T \left(\frac{1}{\sqrt{\overline{\Lambda}_t}} - \frac{1}{\sqrt{\overline{\Lambda}_{t+2\tau+1}}}\right) G^2(1+4\tau) \\
&\leq 2R\sqrt{\Lambda_T+\beta} + \frac{R}{2\sqrt{\beta}}(2\tau+1)(4\tau+1)G^2 \\
&\leq 2R\sqrt{\Lambda_T} + 2R\sqrt{\beta} + \frac{R}{\sqrt{\beta}}(2\tau+1)^2G^2
\end{aligned}$$

The second inequality uses Lemma 2.6 and reorders the timestamps of the sum; the third inequality upper bounds both $\overline{\Lambda}_{\pi(T)}$ and $\Lambda_T^\pi = \Lambda_T$ by $\Lambda_T + \beta$ for the first term, and uses telescoping and lower bounds $\overline{\Lambda}_t$ by β for the second term; in the last inequality we employ the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b \geq 0$. This concludes the proof. □

3.3.3 *Adaptation to Unbounded Delays in the Single-Agent Setting*

In this part, we will show that when there is only one agent (i.e., $N = 1$), we can extend the ideas developed in the previous section to cope even with *unbounded* delays. In fact, in this situation the agent knows exactly the delay of each feedback and how each iterate is computed, so they can tune their learning rate accordingly. This is in sharp contrast with the decentralized case in which the agents are in general unable to estimate the number of actions that have been played in the network but for which they have not received the corresponding feedback (i.e., $\mathrm{card}(\mathcal{U}_t)$).

Approximation of cumulative lag (single-agent and unbounded delay)

To put all this in motion, let G be an upper bound on the norms of gradients that we assume to be known by the agent, and let $\mathcal{F}_t = \mathcal{F}_t^1$ denotes the set of feedback (represented by their timestamps) received before g_t. Our goal is to provide an upper bound of $\Lambda_t = \Lambda_t^{\mathrm{id}}$ that is as tight as possible. As in Section 3.3.2, this is done in two steps (we write below $\mathcal{D}_t = \mathcal{D}_t^{\mathrm{id}}$ for simplicity)

1. The quantity $\sum_{s=1}^t \|g_s\|_*^2$ can be approximated by $\sum_{s\in\mathcal{S}_t} \|g_s\|_*^2$. Clearly,

$$\sum_{s=1}^t \|g_s\|_*^2 \leq \sum_{s\in\mathcal{S}_t} \|g_s\|_*^2 + G^2(\mathrm{card}(\mathcal{U}_t)+1);$$

2. A proxy for $\sum_{\{s,l\}\in\mathcal{D}_t}\|g_s\|_*\|g_l\|_*$, is $\sum_{s\in\mathcal{S}_t}(\|g_s\|_*\sum_{l\in\mathcal{F}_s\setminus\mathcal{S}_s}\|g_l\|_*)$. Thanks to Proposition 3.3 and Proposition 3.7, we have indeed

$$\sum_{\{s,l\}\in\mathcal{D}_t}\|g_s\|_*\|g_l\|_* \le \sum_{s\in\mathcal{S}_t}\left(\|g_s\|_*\sum_{l\in\mathcal{F}_s\setminus\mathcal{S}_s}\|g_l\|_*\right) + G^2(\operatorname{card}(\mathcal{D}_t)-\operatorname{card}(\mathcal{A}_t)).$$

In summary, we have shown that $\Lambda_t \le \Gamma_t + G^2\tilde{\tau}_t$ where $\tilde{\tau}_t := t + 2D_t - \operatorname{card}(\mathcal{S}_t) - 2\operatorname{card}(\mathcal{A}_t)$. This has the following immediate consequences.

Regret for adaptive learning rate (single-agent and unbounded delay)

Theorem 3.9. *Suppose that Assumptions 2.1 and 2.2 hold and the sequence of active feedback is non-decreasing, i.e., $\mathcal{S}_t \subseteq \mathcal{S}_{t+1}$. Assume further that (DDA) is run with the learning rate sequence*

$$\eta_t = \min\left(\eta_{t-1}, \frac{R}{\sqrt{\Gamma_t + G^2\tilde{\tau}_t}}\right) \tag{AdaDelay+}$$

where $\tilde{\tau}_t = t + 2D_t - \operatorname{card}(\mathcal{S}_t) - 2\operatorname{card}(\mathcal{A}_t)$. Then, for any z such that $h(z) \le R^2$, the algorithm enjoys the regret bound

$$\operatorname{Reg}_T(z) \le 2R\max_{1\le t\le T}\sqrt{\Gamma_t + G^2\tilde{\tau}_t} \le 2R\min\left(\max_{1\le t\le T}\sqrt{\Lambda_t + G^2\tilde{\tau}_t}, G\sqrt{T+2D_T}\right).$$

Proof. Let $\overline{\Lambda}_t = R^2/\eta_t^2$ so that $\eta_t = R/\sqrt{\overline{\Lambda}_t}$. It holds that $\overline{\Lambda}_t \ge \Gamma_t + \tilde{\tau}_t G^2 \ge \Lambda_t$. The first inequality comes from the definition of η_t and the second inequality was shown just above. Applying Theorem 3.1 with $\pi = \mathrm{id}$ and Lemma 2.6 yields

$$\begin{aligned}
\operatorname{Reg}_T(z) &\le \frac{h(z)}{\eta_T} + \frac{1}{2}\sum_{t=1}^{T}\eta_t\left(\|g_t\|_*^2 + 2\|g_t\|_*\sum_{s\in\mathcal{U}_t}\|g_s\|_*\right)\\
&\le R\sqrt{\overline{\Lambda}_T} + \frac{R}{2}\sum_{t=1}^{T}\frac{1}{\sqrt{\Lambda_t}}\left(\|g_t\|_*^2 + 2\|g_t\|_*\sum_{s\in\mathcal{U}_t}\|g_s\|_*\right)\\
&\le R\sqrt{\overline{\Lambda}_T} + R\sqrt{\Lambda_T} \le 2R\sqrt{\overline{\Lambda}_T}.
\end{aligned}$$

Since $\overline{\Lambda}_T = \max_{1\le t\le T}\Gamma_t + \tilde{\tau}_t G^2$, we have already proved the first inequality. For the second inequality, we use both $\Gamma_t \le \Lambda_t$ and $\Gamma_t \le (\operatorname{card}(\mathcal{S}_t) + 2\operatorname{card}(\mathcal{A}_t))G^2$ (cf. (3.10)). □

We refer to this new adaptive scheme as AdaDelay+ and we provide one possible pseudo-code implementation as Algorithm 3.3. Notice that we do not use directly $\eta_t = R/\sqrt{\Gamma_t + G^2\tilde{\tau}_t}$ since we want the learning rate to be non-increasing.

According to Theorem 3.9, AdaDelay+ enjoys a regret bound that is both data- *and* delay-dependent, all the while bypassing the bounded delay assumption. To provide a better comparison between the bounds of Theorems 3.6 and 3.9, we show below that $\tilde{\tau}_t$ can be further bounded from above if delays are bounded by a constant.

Proposition 3.10. *Assume that the maximum delay is bounded by τ. Then $\tilde{\tau}_t \le 2\tau^2 + 3\tau + 1$.*

Proof. To begin, we have $t - \operatorname{card}(\mathcal{S}_t) \le \tau + 1$ as $[t - \tau - 1] \subseteq \mathcal{S}_t$. Next, let us consider a pair $\{s,l\} \in \mathcal{D}_t \setminus \mathcal{A}_t$. From Propositions 3.3 and 3.7 we know that

Algorithm 3.3: AdaDelay+

Initialize: $\mathcal{G} \leftarrow \emptyset$, $t \leftarrow 1$, $\tilde{\tau} \leftarrow 0$, $\Gamma \leftarrow 0$.
while not stopped **do**
  if receive feedback g_t **then**
    $\tilde{\tau} \leftarrow \tilde{\tau} - 1 - 2(\mathrm{card}(\mathcal{G}) - \mathrm{card}(\mathcal{S}_t))$
    $\Gamma \leftarrow \Gamma + \|g_t\|_*^2 + 2\|g_t\|_*(\sum_{s\in\mathcal{G}}\|g_s\|_* - \sum_{s\in\mathcal{S}_t}\|g_s\|_*)$
    $\mathcal{G} \leftarrow \mathcal{G} \cup \{t\}$
  else if requested to play an action x_t **then**
    $\mathcal{S}_t \leftarrow \mathcal{G}$
    $\tilde{\tau} \leftarrow \tilde{\tau} + 1 + 2((t-1) - \mathrm{card}(\mathcal{S}_t))$
    $\overline{\Lambda} \leftarrow \max(\Gamma, \Gamma + G^2\tilde{\tau})$
    $x_t \leftarrow \arg\min_{x\in\mathcal{X}} \sum_{s\in\mathcal{S}_t}\langle g_s, x\rangle + (\sqrt{\overline{\Lambda}}/R)h(x)$
    $t \leftarrow t+1$
  end if
end while

$\{s,l\} \not\subseteq \mathcal{S}_t$, so we have either $s \in \{t-\tau,...,t\}$ or $l \in \{t-\tau,...,t\}$. Without loss of generality, we suppose $s < l$, then $l \in \{t-\tau,...,t\}$. By Proposition 3.3 we have $s \notin \mathcal{S}_l$, and thus $s \in \{l-\tau,...,l-1\}$. This shows $\mathrm{card}(\mathcal{D}_t \setminus \mathcal{A}_t) \leq \tau(\tau+1)$. We can therefore conclude $\tilde{\tau}_t \leq 2\tau(\tau+1) + \tau + 1 = 2\tau^2 + 3\tau + 1$. □

Theorem 3.9 along with Proposition 3.10 shows that the regret bound of AdaDelay+ achieves the best of both worlds:

- When the delays are bounded by τ, we have $\tilde{\tau} \leq 2\tau^2 + 3\tau + 1$, so this worst-case bound still outperforms (by an additive constant) the data-dependent bound of Theorem 3.6. In the same setting, Joulani et al. [139] also proposed an adaptive algorithm based on FTRL-Prox with a regret bound of the same order.
- It also achieves the optimal square-root dependence on the cumulative unavailability D_T no matter whether the delays are bounded or not.

In summary, our analysis suggests that AdaDelay+ could offer improved performance under various conditions, relative to existing methods for the single-agent setup. This potential advantage underscores the value of the insights we've presented for this general framework.

3.4 MULTI-AGENT ONLINE LEARNING FOR MINIMIZATION OF GLOBAL LOSSES

Thus far in this chapter, our analysis has focused on the agents' *individual* losses (ℓ_t being the loss of the active agent $i = i(t)$), and thus *leads to regret bounds that characterize how much the whole network actually pays.* While these bounds have an interest, networks of agents may also want to monitor *global* losses over the agents. This is typically the case of distributed online optimization, where the agents cooperate to solve a time-varying global problem.

In this section, we demonstrate the flexibility of our framework by showing that the aforementioned algorithms and analyses can be easily extended to this setup. This, on one hand, bridges the gap between our work and the broad corpus of literature on distributed online optimization [123, 244, 284], and, on the other hand, provides the occasion to directly address the case of open networks where agents can join and depart the optimization process freely [87, 119, 120].

3.4.1 *From Effective Regret to Collective Regret*

Anonymous and simultaneously active agents

In distributed optimization, it is often assumed that multiple predictions are made in a same time slot. Formally, we denote by N_t the number of active agents at time t and identify these agents from 1 to N_t instead of identifying each agent independently. This notation clarifies the fact that the agents are anonymous with respect to the algorithm and each other. The functions and the played points at time t are respectively denoted by $\ell_t^1, \dots, \ell_t^{N_t}$ and $x_t^1, \dots, x_t^{N_t}$.

Effective regret

By directly extending the regret defined in Definition 2.1 to our current setup, we obtain the following:

$$\mathrm{Reg}_T^{\ell}(z) = \sum_{t=1}^{T} \sum_{i=1}^{N_t} \ell_t^i(x_t^i) - \sum_{t=1}^{T} \sum_{i=1}^{N_t} \ell_t^i(z), \qquad \text{(Effective Regret)}$$

where the superscript ℓ means that the regret sums over the *local* costs of the learners. Each agent only pays for the function it serves and the ultimate goal for a single agent is to perform well on the functions that it encounters. As an example, on-device machine learning aims to equip users' personal devices with intelligent machine features such as conversational understanding and image recognition, for the purposes of providing a satisfying user experience to each individual [251, 274].

Collective regret

In contrast, we can also define *global* loss functions $\ell_t = \sum_{i=1}^{N_t} \ell_t^i$ at every instant t and evaluate each active agents' action with respect to this function. This leads to the following regret formulation:

$$\mathrm{Reg}_T^{g}(z) = \sum_{t=1}^{T} \sum_{i=1}^{N_t} \ell_t^i(x_t^1) - \sum_{t=1}^{T} \sum_{i=1}^{N_t} \ell_t^i(z), \qquad \text{(Collective Regret)}$$

where, instead of evaluating ℓ_t^i at the point x_t^i played by learner i, we now evaluate all the ℓ_t^i at a single point x_t^1 independently of the worker i. The choice of the *reference agent* can vary with time; it is however possible to fix its index to 1 in advance given that the attribution of the worker indices at each t is arbitrary.

When the number of agents are fixed, *collective regret* reduces to the usual regret formulation employed in the distributed online optimization literature [123, 244, 284]. This performance measure suits better the applications related to wireless sensor networks such as distributed estimation [227] and data fusion [202, 230]. In fact, sensor networks are mostly deployed for a common objective shared by all the sensors. To attain this objective, the sensor nodes may need to cooperate to track some unknown variable or to collaborate to learn a global assessment of the situation. The collective regret then measures each agent's performance with respect to this *collective* mission, hence the name thereof.

Lipschitz continuity of the loss functions

To relate these two different measures, we introduce a stronger version of Assumption 2.2 that places constraints on both the loss functions and the feedback, written as $g_t^i \in \partial \ell_t^i(x_t^i)$.

Assumption 3.3. There exists $G > 0$ such that for all $t \in \mathbb{N}$, $i \in [N_t]$, ℓ_t^i is G-Lipschitz with respect to the norm $\|\cdot\|$, and $\|g_t^i\|_* \leq G$.

Remark 3.4. It is obvious that the boundedness of subgradients would be implied by the Lipschitz continuity of losses if ℓ_t^i were defined over the entire space $\mathbb{R}^d$. However, provided that we have defined the loss functions as functions on $\mathcal{X}$, the subgradients do not need to be bounded even if the functions are Lipschitz. This is why we state the two conditions separately in the assumption.

Inequality relating collective regret to effective regret

Under this strengthened assumption, we have the following relation between Reg_T^g and Reg_T^ℓ.

Lemma 3.11. *Suppose that Assumption 3.3 holds. Then,*

$$\mathrm{Reg}_T^g(z) \leq \mathrm{Reg}_T^\ell(z) + \sum_{t=1}^{T} \sum_{i=1}^{N_t} G\|x_t^i - x_t^1\|.$$

Proof. This is immediate from the definition of the regrets and the Lipschitz continuity of the losses. □

A note on open multi-agent systems

Finally, let us highlight that our formulation also admits the additional flexibility of involving different sets and numbers of agents at each iteration. This is of particularly interest for open multi-agent systems where the agents can join and leave the system at any moment [87, 119, 120]. Some examples of such systems include volunteer computing [70], vehicular ad-hoc networks [92], and elastic distributed training of machine learning models [203]. For the sake of illustration, we conduct experiments for this setup in Section 3.5.3.

3.4.2 *Decentralized Delayed Dual Averaging*

Thanks to Lemma 3.11, a bound on the effective regret can be directly translated into one on the collective regret as long as the distances between the agents' predictions for a same moment can be controlled. To illustrate this idea, we adapt (DDA) to the current setup and bound its induced collective regret for appropriately chosen learning rates. Let us first slightly extend the previously introduced notations and concepts to the current framework: The set of available gradients at time t for a worker i, $\mathcal{S}_t^i$, now represents the set of the (learner, time) indices of the feedback available for playing x_t^i so that if $(j,s) \in \mathcal{S}_t^i$ then necessarily $s \in [t-1]$. The maximum delay τ is to be understood with respect to the global time index t. That is, for every $s \in [t-\tau-1]$ and $j \in [N_s]$ we must have $(j,s) \in \mathcal{S}_t^i$. We also introduce the quadratic mean of number of active agents by $\overline{N} = \sqrt{(1/T)\sum_{t=1}^T N_t^2}$.

Decentralized delayed dual averaging

With these notations, the update of *decentralized delayed dual averaging* (D-DDA) writes at time t for an agent i as[7]

$$x_t^i = \operatorname*{arg\,min}_{x \in \mathcal{X}} \sum_{(j,s) \in \mathcal{S}_t^i} \langle g_s^j, x \rangle + \frac{h(x)}{\eta_t^i}. \tag{D-DDA}$$

In order to understand the mechanics of collective regret in our setup, we first consider the case of a fixed learning rate $\eta_t^i \equiv \eta$. To bound the collective regret, three elements come into play.

- **Effective regret:** For this part, we change the time indices to have exactly one point played at each time. We define $M_t = \sum_{s=1}^t N_s$ and $M = M_T$; then, the index of worker i at time t is changed to $\phi(i,t) = M_{t-1} + i$ (so

7 It is worth noticing that the original (DDA) is already a decentralized algorithm. We add the term "decentralized" here to emphasize the similarity of the underlying framework to that of the more classic setup of decentralized online optimization.

that only one action is performed at that time). This maps our problem to the setting of Theorem 3.1 with $\eta_t \equiv \eta$. Taking $\pi = \mathrm{id}$, we then get

$$\mathrm{Reg}_T^{\ell}(z) \leq \frac{h(z)}{\eta} + \frac{1}{2}\sum_{m=1}^{M} \eta \left(\|g'_m\|_*^2 + 2\|g'_m\|_* \sum_{l \in [m-1]\setminus \mathcal{S}'_m} \|g'_l\|_* \right) \tag{3.11}$$

where $g'_{\phi(i,t)} = g_t^i$ and $\mathcal{S}'_{\phi(i,t)} = \{\phi(j,s) : (j,s) \in \mathcal{S}_t^i\}$.

- **Maximum delay τ:** Bounding from above the number of unavailable gradients for a (learner, time) pair and translating this condition to a bound on $\mathrm{card}([m-1]\setminus \mathcal{S}'_m)$, we get

$$\mathrm{Reg}_T^{\ell}(z) \leq \frac{h(z)}{\eta} + \eta(\tau+1)G^2 \sum_{t=1}^{T} N_t^2. \tag{3.12}$$

- **Non-expansiveness of the mirror map (Lemma A.4):** This part enables us to go from the effective regret to the collective regret using Lemma 3.11.

Putting together these points we manage to show the following bound on the collective regret.

Regret for D-DDA with constant learning rate

Theorem 3.12. *Suppose that Assumptions 2.1 and 3.3 hold and that the maximum delay is bounded by τ. Then, for any z such that $h(z) \leq R^2$, running (D-DDA) with constant learning rate*

$$\eta_t^i \equiv \eta = \frac{R}{G\overline{N}\sqrt{(2\tau+1)T}}$$

Then, for any z such that $h(z) \leq R^2$, guarantees the following upper bound on the collective regret

$$\mathrm{Reg}_T^{g}(z) \leq 2RG\overline{N}\sqrt{(2\tau+1)T} = O(\overline{N}\sqrt{\tau T}).$$

Proof. Let us start with (3.11). Thanks to the boundedness of the feedback (Assumption 3.3), we have

$$\begin{aligned}\mathrm{Reg}_T^{\ell}(z) &\leq \frac{h(z)}{\eta} + \frac{1}{2}\sum_{m=1}^{M} \eta \left(\|g'_m\|_*^2 + 2\|g'_m\|_* \sum_{l \in [m-1]\setminus \mathcal{S}'_m} \|g'_l\|_* \right) \\ &\leq \frac{h(z)}{\eta} + \frac{\eta}{2}\sum_{m=1}^{M} (1 + 2\,\mathrm{card}([m-1]\setminus \mathcal{S}'_m))G^2. \end{aligned} \tag{3.13}$$

To bound $\mathrm{card}([m-1]\setminus \mathcal{S}'_m)$, we write $m = \phi(i,t)$. On one hand, the subgradients

$$\{g_t^{i-1}, \dots, g_t^1\} = \{g'_{m-1}, \dots, g'_{m-i+1}\}$$

of instant t are necessarily unavailable when making the prediction $x_t^i = x'_m$. On the other hand, the maximum delay assumption guarantees that all the subgradients received before time $t-\tau$ are used in the computation of x_t^i. This leads to the inequality

$$\mathrm{card}([m-1]\setminus \mathcal{S}'_m) \leq i - 1 + \sum_{s=1}^{\tau} N_{t-s},$$

with the convention $N_l = 0$ if $l \leq 0$. Subsequently, for any $t \in [T]$,

$$\sum_{m=M_{t-1}+1}^{M_t} \operatorname{card}([m-1] \setminus \mathcal{S}'_m) \leq \frac{N_t(N_t-1)}{2} + N_t \sum_{s=1}^{\tau} N_{t-s} \leq \frac{(\tau+1)}{2} N_t^2 + \frac{1}{2} \sum_{s=1}^{\tau} N_{t-s}^2. \tag{3.14}$$

In the above, we used Young's inequality to bound the second term in the second inequality. Substituting (3.14) in (3.13) then yields

$$\operatorname{Reg}_T^{\ell}(z) \leq \frac{h(z)}{\eta} + \eta(\tau+1)G^2 \sum_{t=1}^{T} N_t^2. \tag{3.15}$$

At this point, we have proved an upper bound on the effective regret. To go from this to the collective regret, we need to bound the difference $\|x_t^i - x_t^j\|$ for all $t \in [T]$ and $i, j \in [N_t]$. To that end, we write $x_t^i = Q(-y_t^i)$ and $x_t^j = Q(-y_t^j)$ where $y_t^i = \eta \sum_{(k,s)\in\mathcal{S}_t^i} g_s^k$ and $y_t^j = \eta \sum_{(k,s)\in\mathcal{S}_t^j} g_s^k$. From the maximum delay assumption we know that $\mathcal{S}_t^i$ and $\mathcal{S}_t^j$ differ by at most $\sum_{s=1}^{\tau} N_{t-s}$ samples. Using the boundedness of the feedback and the non-expansiveness of the mirror map (Lemma A.4), we obtain

$$\sum_{i=1}^{N_t} G\|x_t^i - x_t^j\| \leq \eta G^2 N_t \sum_{s=1}^{\tau} N_{t-s} \leq \eta G^2 \left(\frac{\tau N_t^2}{2} + \frac{1}{2} \sum_{s=1}^{\tau} N_{t-s}^2 \right). \tag{3.16}$$

With (3.15) and (3.16), invoking Lemma 3.11 gives

$$\operatorname{Reg}_T^{g}(z) \leq \frac{h(z)}{\eta} + \eta(2\tau+1)G^2 \sum_{t=1}^{T} N_t^2.$$

The theorem follows immediately. □

Interestingly, our regret bound features the quadratic mean of number of active agents $\overline{N}$. If we fix the number of total agents M across T rounds, then a larger value of $\overline{N}$ indicates a more important variation in the number of active agents across iterations. The fact that this causes a larger regret is expected because the algorithm would need more time to accommodate the change in such scenarios.

The case of fixed communication network

To provide a comparison with the existing literature, let us now zoom in on the case of a fixed communication network of $N_t \equiv N$ agents over an underlying graph $\mathfrak{C}$. Theorem 3.12 yields an $O(N\sqrt{\tau T})$ bound in this situation. Notably, the maximum delay τ is typically in the order of the diameter of the graph $\operatorname{diag}(\mathfrak{C})$ if (D-DDA) is implemented by broadcasting the receives gradients to all the neighbors. On the other hand, algorithms that are based on the gossip protocol [25] often exhibit a regret bound expressed as $O(N\sqrt{T/(1-\varsigma)})$, where $1-\varsigma$ is the spectral gap of the gossip matrix [69, 123].

Given that we have $\operatorname{diag}(\mathfrak{C}) = O(1/(1-\varsigma))$ for many gossip matrices that are used in practice, see e.g., [206], this shows (D-DDA) offers an improvement in regret over gossip-based methods, at the price of a higher communication cost (we also refer the readers to [60] for a collection of results on the relationship between a graph's diameter and its algebraic connectivity).

Regarding the more general case of asynchronous distributed online optimization, we are only aware of the work of Jiang et al. [137] that analyzes a push-sum strategy. It is difficult to compare our result with theirs due to the difference in assumptions. On the other extreme, when there is no delay $\tau = 0$ and the number of agents is fixed at $N_t \equiv N$, we recover a regret in $O(N\sqrt{T})$. This matches the regret bound of running (DA) on the loss sequences defined by $\ell_t = \sum_{i=1}^{N} \ell_t^i$, as each single loss is NG-Lipschitz (see Corollary 2.5).

3.4.3 *A More Practical Learning Rate*

The learning rate of Theorem 3.12 requires knowledge of the quadratic mean of number of agents $\overline{N}$. Nonetheless, since the network may be evolving, this average number may often not be available in advance; neither is the time horizon T nor the current time index t. A first solution can be taking learning rates of the form $\eta_t^i = \eta_t = \Theta(1/N_{\max}\sqrt{\tau t})$. However, this still requires the knowledge of the global time t which is typically out of reach in the setup we are considering; in addition, it can be overly pessimistic with the dependence in $N_{\max}\sqrt{\tau}$. To overcome these issues, we exploit the ideas of Section 3.3 and show that a learning rate scheme similar to the one considered in Section 3.3.1 equally guarantees low collective regret. To begin with, we rewrite Assumption 3.1 to accommodate the new notation.

Assumption 3.4. If $(j,s) \in \mathcal{S}_t^i$ then $\operatorname{card}(\mathcal{S}_s^j) < \operatorname{card}(\mathcal{S}_t^i)$.

Assumption on number of received feedback element

Under this assumption, we prove the following theorem which extends Proposition 3.5 to provide a bound on the collective regret for our current setup.

Theorem 3.13. *Suppose that Assumptions 2.1, 3.3 and 3.4 hold and that the maximum delay is bounded by τ. Then, for any z satisfying $h(z) \le R^2$, running (D-DDA) with learning rates*

$$\eta_t^i = \frac{R}{G\sqrt{(5\tau+3)(\operatorname{card}(\mathcal{S}_t^i) + (\tau+1)N_{\max})N_{\max}}} \tag{3.17}$$

Regret for D-DDA with anytime, partially adaptive learning rate

guarantees a collective regret in

$$\operatorname{Reg}_T^g(z) = O(\sqrt{\tau M N_{\max}}).$$

Proof. With a slight abuse of notation, we only work with the (worker, time) index pair in this proof, but it should be understood that the change of index ϕ intervenes implicitly when we apply the arguments of the previous sections (notably when we compare the indices). Compared to Theorem 3.12, the two additional difficulties are: (*i*) the non-monotonicity of learning rates which are solved by the introduction of a suitable faithful permutation; (*ii*) the predictions of a time instant are not generated by the same learning rate, but we still manage to control the deviation since these learning rates are close enough.

To begin, we consider a permutation π satisfying $\pi^{-1}(j,s) < \pi^{-1}(i,t)$ if $\operatorname{card}(\mathcal{S}_s^j) < \operatorname{card}(\mathcal{S}_t^i)$. Such a π is necessarily faithful according to Assumption 3.4. We claim that $\operatorname{card}(\mathcal{U}^{\pi}_{\pi^{-1}(i,t)}) \le (\tau+1)N_{\max}$ (where $\mathcal{U}^{\pi}_{\pi^{-1}(i,t)} = [\pi^{-1}(i,t)-1]^{\pi} \setminus \mathcal{S}_t^i$). Let $s \in [\tau]$ such that $M_{t+s-\tau} > \operatorname{card}(\mathcal{S}_t^i) \ge M_{t+s-\tau-1}$. Then for any $j \in [N_{t+l}]$ with $l > s$, it holds

$$\operatorname{card}(\mathcal{S}_{t+l}^j) \ge M_{t+l-\tau-1} \ge M_{t+s-\tau} > \operatorname{card}(\mathcal{S}_t^i)$$

Accordingly, $\pi^{-1}(i,t) < \pi^{-1}(j,t+l)$. This shows that if $\pi^{-1}(j,l) < \pi^{-1}(i,t)$ for some $l \in [T]$ and $j \in [N_l]$, then $l \leq t+s$, and subsequently $\mathrm{card}([\pi^{-1}(i,t)-1]^\pi) \leq M_{t+s}$. We have therefore

$$\mathrm{card}([\pi^{-1}(i,t)-1]^\pi \setminus \mathcal{S}_t^i) \leq M_{t+s} - M_{t+s-\tau-1} = \sum_{l=0}^{\tau} N_{t+s-l} \leq (\tau+1)N_{\max}.$$

Since $\eta_t^i \leq \eta_s^j$ if and only if $\mathrm{card}(\mathcal{S}_t^i) \geq \mathrm{card}(\mathcal{S}_s^j)$, we have indeed $\eta_{\pi((i,t)+1)} \leq \eta_{\pi(i,t)}$ (here and below we use the notation $\eta_{j,s} = \eta_s^j$). Invoking Theorem 3.1, one has[8]

$$\begin{aligned}\mathrm{Reg}_T^\ell(z) &\leq \frac{h(z)}{\eta_{\pi(N_T,T)}} + \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\eta_t^i\left(\|g_t^i\|_*^2 + 2\|g_t^i\|_* \sum_{s\in\mathcal{U}^\pi_{\pi^{-1}(i,t)}} \|g_s\|_*\right)\\ &\leq \frac{h(z)}{\min_{t\in[T],i\in[N_t]}\eta_t^i} + \frac{1}{2}\sum_{t=1}^{T}\left(\max_{i\in[N_t]}\eta_t^i\right) G^2(2\tau+3)N_t N_{\max}. \qquad (3.18)\end{aligned}$$

We next bound the difference $\|x_t^i - x_t^j\|$ for all $t \in [T]$ and all $i,j \in [N_t]$. Similar to the proof of Theorem 3.12, we write $x_t^i = Q(-y_t^i)$ and $x_t^j = Q(-y_t^j)$ where $y_t^i = \eta_t^i \sum_{(k,s)\in\mathcal{S}_t^i} g_s^k$ and $y_t^j = \eta_t^j \sum_{(k,s)\in\mathcal{S}_t^j} g_s^k$. By the non-expansiveness of the mirror map (Lemma A.4), it is then sufficient to bound $\|y_t^i - y_t^j\|$. For ease of notation, in the rest of the proof we denote by $\mathcal{S}_\cap$ the intersection of $\mathcal{S}_t^i$ and $\mathcal{S}_t^j$, i.e., $\mathcal{S}_\cap = \mathcal{S}_t^i \cap \mathcal{S}_t^j$, and we write $\mathcal{S}_t^i \triangle \mathcal{S}_t^j$ for the symmetric difference of these two sets. It follows

$$\begin{aligned}\|y_t^i - y_t^j\| &= \|(\eta_t^i - \eta_t^j)\sum_{(k,s)\in\mathcal{S}_\cap} g_s^k + \eta_t^i \sum_{(k,s)\in\mathcal{S}_t^i\setminus\mathcal{S}_\cap} g_s^k - \eta_t^j \sum_{(k,s)\in\mathcal{S}_t^j\setminus\mathcal{S}_\cap} g_s^k\|\\ &\leq |\eta_t^i - \eta_t^j| \sum_{(k,s)\in\mathcal{S}_\cap}\|g_s^k\| + \eta_t^i \sum_{(k,s)\in\mathcal{S}_t^i\setminus\mathcal{S}_\cap}\|g_s^k\| + \eta_t^j \sum_{(k,s)\in\mathcal{S}_t^j\setminus\mathcal{S}_\cap}\|g_s^k\|\\ &\leq G(|\eta_t^i - \eta_t^j|\,\mathrm{card}(\mathcal{S}_\cap) + \max(\eta_t^i,\eta_t^j)\,\mathrm{card}(\mathcal{S}_t^i\triangle\mathcal{S}_t^j))\\ &\leq G(|\eta_t^i - \eta_t^j| M_{t-1} + \max(\eta_t^i,\eta_t^j)\tau N_{\max}). \qquad (3.19)\end{aligned}$$

In the last inequality we use the fact that if one element belongs to one set but not the other then it must come from the last τ time steps to bound $\mathrm{card}(\mathcal{S}_t^i\triangle\mathcal{S}_t^j)$.

To control $|\eta_t^i - \eta_t^j|$, we note that for any $b > a > 0$, it holds

$$\frac{1}{\sqrt{a}} - \frac{1}{\sqrt{b}} = \frac{b-a}{\sqrt{ab}(\sqrt{a}+\sqrt{b})} \leq \frac{b-a}{2a\sqrt{a}}.$$

For every $k \in [N_t]$, we have $\mathrm{card}(\mathcal{S}_t^k) + (\tau+1)N_{\max} \geq M_t > M_{t-1}$. Therefore, with the learning rate rule (3.17), we get

$$|\eta_t^i - \eta_t^j| \leq \frac{R\,|\mathrm{card}(\mathcal{S}_t^i) - \mathrm{card}(\mathcal{S}_t^j)|}{2GM_{t-1}\sqrt{(5\tau+3)M_t N_{\max}}} \leq \frac{R\tau N_{\max}}{2GM_{t-1}\sqrt{(5\tau+3)M_t N_{\max}}}. \qquad (3.20)$$

8 Note that the sum is ordered differently as stated in the theorem.

Let us denote $\eta_t = R/(G\sqrt{(5\tau+3)M_t N_{\max}})$; then $\eta_t^i \le \eta_t$ for all $i \in [N_t]$. We also take

$$\underline{\eta} = \frac{R}{G\sqrt{(5\tau+3)(MN_{\max}+(\tau+1)N_{\max}^2)}}$$

so that $\eta_t^i \ge \underline{\eta}$ for all $t \in [T]$, $i \in [N_t]$. We conclude with the help of Lemmas 2.6, 3.11 and A.4, and the inequalities (3.18), (3.19) and (3.20):

$$\begin{aligned}
\mathrm{Reg}_T^g(z) &\le \frac{h(z)}{\underline{\eta}} + \frac{1}{2}\sum_{t=1}^{T}\left(\eta_t G^2(4\tau+3)N_t N_{\max} + \frac{RG\tau N_t N_{\max}}{\sqrt{(5\tau+3)M_t N_{\max}}}\right) \\
&= \frac{h(z)}{\underline{\eta}} + \frac{1}{2}\sum_{t=1}^{T}\frac{RG(5\tau+3)N_t N_{\max}}{\sqrt{(5\tau+3)M_t N_{\max}}} \\
&\le RG\sqrt{(5\tau+3)(MN_{\max}+(\tau+1)N_{\max}^2)} + RG\sqrt{(5\tau+3)MN_{\max}}.
\end{aligned}$$

Accordingly, $\mathrm{Reg}_T^g(z) = O(\sqrt{\tau M N_{\max}})$. □

The bound of Theorem 3.13 is worse than the one shown in Theorem 3.12 since

$$\overline{N}^2 T = \sum_{t=1}^{T} N_t^2 \le \sum_{t=1}^{T} N_t N_{\max} = MN_{\max}.$$

This deterioration seems to be unavoidable if the number of active agents of each round N_t is not known by the agents. In spite of this, having the total number of actions taken in the full process M in the bound still suggests that the algorithm is at least partially adaptive to the number of agents. More importantly, since $\mathrm{card}(\mathcal{S}_t^i)$ is obviously available to each agent at time t, the learning rate (3.17) is indeed implementable by every single agent as long as the constants G, τ, and $N_{\max}$ are known.

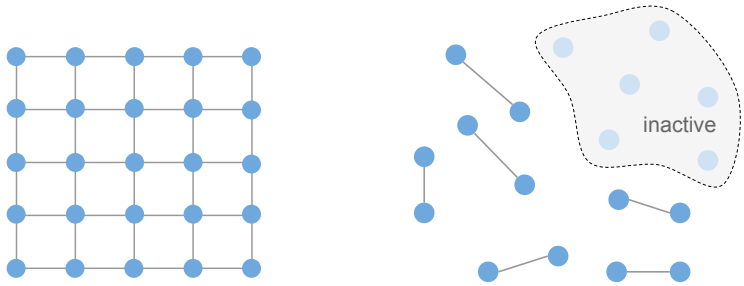
Remark 3.5. From our analysis, we notice that all ℓ_t^i may not happen exactly at the same time. More generally, the time index t can stand for a time interval in a physical sense. In this case, it is possible to have instantaneous feedback (i.e., $g_t^i \in \mathcal{S}_t^j$ for some i, j) and a single physical agent can play several times during the period corresponding to t. In such situations, the same proof template can be readily applied.

3.5 SIMULATIONS IN STATIC AND OPEN NETWORKS

In this section, we carry out numerical simulations based on the formulation presented in Section 3.4. We focus on agents whose loss functions are fixed, operating within either a static or an open network, and we use constant learning rates instead of adaptive ones. Therefore, the findings of these experiments are rather complementary to the theoretical contributions that we made in preceding sections, with the online aspect solely coming from the change in the composition of the network.

3.5.1 *Problem Description*

Let us consider a decentralized *least absolute deviation* (LAD) regression model. *Least absolute deviation regression*



(a) In the static network experiment, we use a fixed 2d grid graph as the underlying communication graph.

(b) In the open network experiment, the active agents of each round are paired with each other to exchange information.

Figure 3.5: Schematic representation of the communication graphs used in our experiments. These diagrams are simplified representations of the actual graphs, which contain 64 nodes, and mainly depict the general structure and connectivity.

Given a data set evenly distributed on N nodes $(a_{ik}, b_{ik})_{i,k\in[N]\times[K]}$ with a_{ik} in $\mathbb{R}^d$ and $b_{ik} \in \mathbb{R}$, it consists in solving

$$\min_{x\in\mathbb{R}^d} \left\{ \ell(x) := \frac{1}{N}\sum_{i=1}^{N} \frac{1}{K}\sum_{k=1}^{K} |a_{ik}^\top x - b_{ik}| \right\}. \tag{3.21}$$

Compared to least square regression, LAD is known to be more resistant to the presence of outliers [157]. Although the use of absolute value makes the problem non-differentiable, (D-DDA) can be run with subgradients as suggested by our analysis. For the experiments, we generate synthetic data as follows:

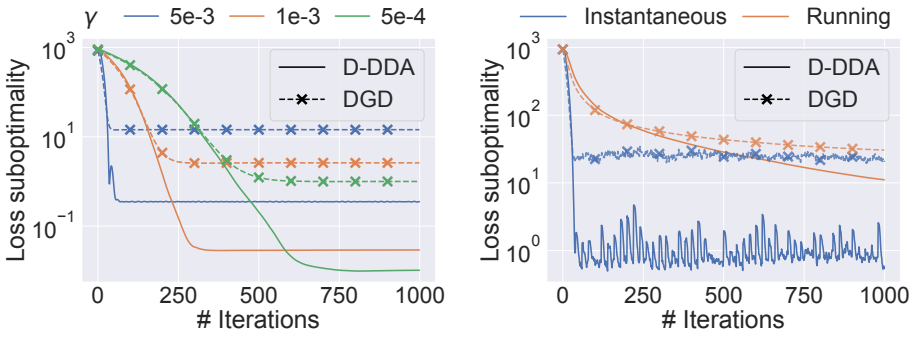
1. The ground truth model $x_\star$ is uniformly drawn from $\in [-5, 5]^d$.
2. The local model $x_\star^i$ of the node i is obtained by perturbing $x_\star$ with a Gaussian noise, i.e., $x_\star^i = x_\star + \varepsilon^i$ where $\varepsilon^i \sim \mathcal{N}(0, I_d)$.
3. We sample $a_{ik} \sim \mathcal{N}(0, I_d)$ and compute $b_{ik} = a_{ik}^\top x_\star^i + \varepsilon_{ik}$ with $\varepsilon_{ik} \sim \mathcal{N}(0, 1)$.
4. On each node, a random portion of samples are corrupted. For these samples, we replace b_{ik} by a random value generated from a Gaussian distribution.

In the above, we introduce the second and the fourth steps mainly for two reasons. First, it makes the problem more heterogeneous, and thus more difficult. Second, it makes the communication between agents more important for finding a good approximation of $x_\star$. In the following, we will take $N = 64$ nodes, $K = 200$ samples per node, and dimension $d = 20$. On each node, the number of corrupted samples is random in $\{0, ..., 120\}$. We also verify that the solution $\hat{x}$ of (3.21) is not too far from $x_\star$.[9]

3.5.2 *Static networks*

We first investigate the performance of the algorithm on a static network. The nodes are arranged in a 2d grid of size 8×8 (a simplified version is illustrated in Fig. 3.5a). Adjacent nodes exchange gradients at each iteration. Communication-

9 Since (3.21) does not admit a close-form solution, we solve it numerically using the python library `statsmodels`. We do the same when evaluating a minimum in the remaining of the experiments.



(a) Static network (b) Open network

Figure 3.6: Comparison of (D-DDA) and (DGD). For a static network we plot in (a) the averaged suboptimality. For an open network we plot in (b) the averaged instantaneous suboptimality (3.22) and the averaged running loss (3.23).

computation overlap is allowed for better efficiency. Then, with a constant learning rate η and the use of quadratic regularizer, the (D-DDA) update writes

$$x_{t+1}^i = x_t^i - \eta \sum_{j=1}^N g_{t-\tau_{j,i}}^j,$$

where $\tau_{j,i}$ is the distance between the nodes j and i. For illustration purposes, we also compare with a decentralized gradient descent (DGD) method [205] with constant learning rate γ and a mixing matrix $W = (w_{i,j})$. Its update is

Decentralized gradient desent

$$x_{t+1}^i = \sum_{j=1}^N w_{i,j} x_t^j - \gamma g_t^i \qquad \text{(DGD)}$$

and we take W as the Metropolis matrix of the graph in our experiments:

$$w_{i,j} = \begin{cases} 1/(\max(\deg(i), \deg(j)) + 1) & \text{if } \{i,j\} \in \mathcal{E}, \\ 1 - \sum_{k=1}^N w_{i,k} & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{E}$ stands for the edges of the communication graph and $\deg(i)$ denotes the degree of the node i.

For a proper comparison of the two algorithms, it is important to notice that a subgradient is sent to all the N nodes in (D-DDA) while it is averaged out in (DGD). Therefore, we will take $\gamma = N\eta$ and refer to it as the *effective learning rate* of both methods. With this in mind, in Fig. 3.6a we plot the convergence of the averaged optimality gap

Effective learning rate

$$\frac{1}{N} \sum_{i=1}^N \ell(x_t^i) - \min_{x \in \mathbb{R}^d} \ell(x)$$

for (D-DDA) and (DGD) with different choices of γ.

Interestingly, we observe that when using the same effective learning rate, the two algorithms establish similar convergence behavior until reaching their respective fixed points. However, (D-DDA) is able to converge to a point with higher accuracy. This is in line with our discussion in Section 3.4.2. In fact, with

D-DDA converges to a point with higher accuracy

Algorithm 3.4: D-DDA at each node i as implemented in Section 3.5.3

Initialize: Index set of received subgradeints $\mathcal{S}_1^i \leftarrow \emptyset$, activation status $\zeta^i \in \{0,1\}$, network parameters $J \in \mathbb{N}$, $p \in [0,1]$
for $t = 1, 2, \ldots$ **do**
  `Agent update`
  if $\zeta^i = 1$ **then**
    Get randomly paired with another active agent j
    Predict x_t^i with (D-DDA) and compute $g_t^i \in \partial \ell^i(x_t^i)$
    Update $\mathcal{S}_{t+1}^i \leftarrow \mathcal{S}_t^i \cup \mathcal{S}_t^j \cup \{(i,t)\}$
  end if
  `Network evolution`
  if $(t+1) \equiv 0 \mod J$ **then**
    Draw a Bernoulli random variable $z^i \sim \mathcal{B}(p)$
    $\zeta^i \leftarrow \zeta^i + z^i \mod 2$
    if $z^i = 1$ and $\zeta^i = 1$ **then**
      Pick randomly $j \in \mathcal{N}_t \cap \mathcal{N}_{t+1}$
      Update $\mathcal{S}_{t+1}^i \leftarrow \mathcal{S}_{t+1}^j$
    end if
  end if
end for

$\operatorname{diag}(\mathfrak{C})$ and $1-\varsigma$ respectively the diameter of the communication graph and the spectral gap of the mixing matrix, for (D-DDA) we can roughly bound $\|x_t^i - x_t^j\|$ by $N\eta\tau G = \gamma \operatorname{diag}(\mathfrak{C}) G$ as we have shown in the proof of Theorem 3.12. As for (DGD), we have asymptotically $\|x_t^i - x_t^j\| \lesssim \gamma G/(1-\lambda)$ [206, Lem. 11]. Hence, when $\operatorname{diag}(\mathfrak{C}) \leq 1/(1-\varsigma)$, which is effectively the case here, we expect the variables $(x_t^i)_{i\in\mathcal{N}}$ to be closer to each other in (D-DDA), and this translates into a higher convergence accuracy.

3.5.3 *Open networks*

Now that we have shown that (D-DDA) performs at least comparably to standard decentralized optimization methods in a static network, we proceed to study its behavior in an open multi-agent system (Algorithm 3.4). Following [270], we model the arrivals and departures of the agents by a Bernoulli process. Initially, only half of the 64 nodes are active. Then, every $J = 20$ iterations, each agent may change its activation status (i.e., from active to inactive or vice-versa) with probability $p = 0.05$. At each iteration, the active nodes are randomly paired with each other, then each pair synchronizes their gradients.[10] Formally, if nodes i and j are paired at time t, then $\mathcal{S}_{t+1}^i \setminus \{g_t^i\} = \mathcal{S}_{t+1}^j \setminus \{g_t^j\} = \mathcal{S}_t^i \cup \mathcal{S}_t^j$. In the spirit of (DGD), we also implement an algorithm which directly updates the primal variables as

Decentralized gradient descent for open network

$$x_{t+1}^i = \frac{x_t^i + x_t^j}{2} - \gamma g_t^i,$$
$$x_{t+1}^j = \frac{x_t^i + x_t^j}{2} - \gamma g_t^j.$$

In both cases, an agent that becomes active at the end of round $t-1$ is assigned the state variable (i.e., $\mathcal{S}_t^i$ or x_t^i) of a random node in $\mathcal{N}_{t-1} \cap \mathcal{N}_t$, where $\mathcal{N}_{t-1}$ and

10 If there is an odd number of nodes, one node is ignored in this process.

$\mathcal{N}_t$ are respectively the agents that are active at round $t-1$ and t. We take $\gamma = N\eta/2$ in our experiment since on average $N/2$ nodes are active.

Instantaneous loss and running loss

As for the performance measure, let us recall that the total number of active workers over t rounds is denoted by $M_t = \sum_{s=1}^{t} N_s$, Moreover, we define the instantaneous loss function as $\ell_t = (1/N_t)\sum_{i=1}^{N_t} \ell_t^i$ (note this is smaller than the ℓ_t of Section 3.4.1 by a factor of $1/N_t$). We then consider the averaged instantaneous optimality gap $\bar{\ell}^{\text{inst}}(t)$ and the averaged running loss $\bar{\ell}^{\text{run}}(t)$ defined by

$$\bar{\ell}^{\text{inst}}(t) = \frac{1}{N_t}\sum_{i\in\mathcal{N}_t} \ell_t(x_t^i) - \min_{x\in\mathbb{R}^d} \ell_t(x) \tag{3.22}$$

$$\bar{\ell}^{\text{run}}(t) = \frac{1}{M_t}\sum_{s=1}^{t}\sum_{i\in\mathcal{N}_s} \ell_s(x_s^i) - \min_{x\in\mathbb{R}^d} \frac{1}{M_t}\sum_{s=1}^{t}\sum_{i=1}^{N} \ell_t^i(x). \tag{3.23}$$

Here, the averaged running loss $\bar{\ell}^{\text{run}}(t)$ is essentially the collective regret divided by the total number of function evaluations and averaged across agents. In fact, if we replace x_s^i with $x_s^{j(s)}$ where $j(s)$ is an agent that is active at round s independent of the index i, we get exactly the collective regret divided by M_t. On the other hand, the averaged instantaneous optimality gap $\bar{\ell}^{\text{inst}}(t)$ can be regarded as a "dynamic" counterpart to $\bar{\ell}^{\text{run}}(t)$. However, it is more challenging to minimize as the minimum of the instantaneous loss can change abruptly from one round to another.

D-DDA is able to track the solution in dynamic environments

In Fig. 3.6b, we plot the evolution of these two measures for $\gamma = 0.005$, which roughly corresponds to the largest learning rate that leads to the decrease of the losses. We see that both algorithms converge to an area where potential solutions are located whereas (D-DDA) gets much closer to the optimum of ℓ_t. This leads to a smooth decrease in the running loss which would eventually stabilize due to the use of constant learning rate. In contrast, the instantaneous loss for (D-DDA) experiences sharp increases when the set of active agents changes, but subsequently decreases, indicating the algorithm's ability to track the instantaneous solution.

4

SLOW VARIATION AND THE ROLE OF OPTIMISM

 # This chapter incorporates material from Hsieh et al. [130]

In the preceding chapter, we have established regret guarantees with respect to the worst-case scenarios. In particular, the losses that we encounter can be arbitrary or even adversarial in nature. Nonetheless, the environment we operate within is not always so unforgiving. Often, it presents a softer, more predictable landscape where patterns in loss functions may emerge, offering us opportunities to achieve smaller regret. For instance, loss sequences may vary slowly or losses may be generated via a *game* mechanism.

In this chapter, we turn our focus toward the first scenario—loss sequences that change slowly over time, while reserving the exploration of the second scenario for Part II of this thesis. Central to our studies here are *optimistic* algorithms. These algorithms are able to exploit the predictability in loss sequences to yield improved performance guarantees.

This chapter serves a dual purpose. On one hand, it provides a preliminary introduction to optimistic algorithms. On the other, it investigates their utility in the multi-agent setup introduced in Chapter 3. For simplicity, our discussion in this chapter concerns exclusively the unconstrained Euclidean setup, i.e., $\mathcal{X} = \mathbb{R}^d$, $h = 1/2\|\cdot\|_2^2$. We thus adopt the notation $\|\cdot\| = \|\cdot\|_2$ throughout this chapter. A more comprehensive introduction to optimistic algorithms, including their use in the general setup with arbitrary (closed convex) action set and regularizer, is deferred to Chapter 5.

CONTRIBUTIONS AND OUTLINE. We start this chapter by recalling in Section 4.1 the *optimistic gradient* algorithm, focusing here on the advantage and the limitation of the method in vanilla online convex optimization. Following this, we extend it to the multi-agent setup of Chapter 3 with the introduction of *delayed optimistic dual averaging* in Section 4.2. The key ingredient of this algorithm is the use of two learning rate sequences whose ratio is adjusted according to the maximum delay. We provide both an upper bound and a matching lower bound on the method's regret, while also presenting examples that justify the necessity of this "learning rate separation". As we conclude the chapter, we delve into a more practical examination of the algorithm's implementation in a multi-agent environment. To that end, we suggest several viable choices for the "guess" vector and explore the use of adaptive learning rates in Section 4.3.

4.1 OPTIMISTIC GRADIENT DESCENT

As as can be easily deduced from the regret analysis of (MD) and (DA) (cf. proofs of Propositions 2.2 and 2.4), the *forward regret* of these algorithms can be bounded by a constant if the learning rate η_t is taken constant. However, this

necessitates playing x_{t+1} in place of x_t, which is not feasible because g_t is only revealed to the learner at the end of round t, and moreover, it is evaluated at x_t.

Optimistic gradient

To circumvent this limitation, optimistic algorithms estimate the gradient of ℓ_t by designing a gradient *guess* $\widetilde{g}_t = \widetilde{g}_t(g_1, \dots, g_{t-1})$. For concreteness, let us consider *optimistic gradient* (OG), one of the most well-known variants of these algorithms. For the unconstrained setup that we study here, its update can be simplified as (we only consider delayed feedback starting from the next section)

$$X_t = X_{t-1} - \eta_t g_{t-1}, \quad X_{t+\frac{1}{2}} = X_t - \eta_t \widetilde{g}_t. \tag{OG}$$

In the above formulation, (OG) operates with two states per round. The *base state*, denoted as X_t, is updated following the classical online gradient step $X_t = X_{t-1} - \eta_t g_{t-1}$. However, for optimistic methods, the point X_t *is not played* at time t; instead, the learner plays the *leading state* $X_{t+\frac{1}{2}}$, which is updated using the estimated gradient $\widetilde{g}_t$. This is the *optimistic step* (also known as the extrapolation step or the exploration step in the literature), and it is expressed as $X_{t+\frac{1}{2}} = X_t - \eta_t \widetilde{g}_t$ in the case of (OG). In other words, the played point is $x_t = X_{t+\frac{1}{2}}$.[1] The update of the algorithm can thus be alternatively written as

$$x_{t+1} = x_t - \eta_{t+1}(g_t + \widetilde{g}_{t+1}) + \eta_t \widetilde{g}_t.$$

While this formulation relates directly the actions taken by the learner, thereby obviating us from the need for introducing the additional variables, we will systematically work with the (OG) formulation as it facilities both the analysis and the generalization of the algorithm. Furthermore, the (OG) formulation clearly highlights the relation between optimistic algorithms and the idealized strategy of playing x_{t+1} instead of x_t in MD and DA. Indeed, we have the following regret guarantee for (OG).

Regret bound for OG

Proposition 4.1. *Suppose that Assumption 2.1 holds. Then, for any $z \in \mathcal{X}$ and $T \in \mathbb{N}$, the regret induced by (OG) relative to z after T rounds is bounded as*

$$\mathrm{Reg}_T(z) \le \frac{\|z - X_1\|^2}{2\eta_{T+1}} + \sum_{t=1}^{T} \frac{\eta_t}{2} \|g_t - \widetilde{g}_t\|^2. \tag{4.1}$$

Proof. This result is by now standard in the literature, see e.g., [45, 140, 198, 228] and references therein. □

Proposition 4.1 suggests that we can get much smaller regret if the guess $\widetilde{g}_t$ is an approximation of g_t, a result that also holds for the more general version of the algorithm presented in Section 5.3. By optimally choosing the learning rates $(\eta_t)_{t\in\mathbb{N}}$, we attain a regret in $O\left(\sqrt{\sum_{t=1}^{T} \|g_t - \widetilde{g}_t\|^2}\right)$. We are thus *optimistic* in the sense that we hope that this sum of squares of differences is small. On the other hand, we recover the regret of vanilla online gradient descent for $\widetilde{g}_t = 0$ (no optimistic guess). In practice, one sensible choice is to use the last received feedback as the guess, i.e., $\widetilde{g}_t = g_{t-1}$. This choice can lead to favorable guarantees when the loss functions are smooth and demonstrate slow changes according to specific measures over time, as evidenced by [45, 140].

1 Note that we use capital X to denote the iterates of an optimistic algorithm while the lowercase x represents the action taken by the learner. This convention is adopted throughout the manuscript.

FAILURE OF OG WITH DECREASING LEARNING RATE. Despite the positive result presented above, (OG) faces the same problem as (MD). Specifically, it cannot guarantee sublinear regret if used with decreasing learning rates (note that the Bregman diameter is clearly unbounded for the unconstrained setup). We illustrate this with the following proposition.

OG with decreasing learning rate induces superlinear regret

Proposition 4.2. (OG) *with learning rate* $\eta_t = 1/\sqrt{t}$ *and guess* $\widetilde{g}_t = g_{t-1}$ *cannot guarantee* $o(T)$ *regret against linear loss functions with bounded loss vectors.*

Proof. We will actually show that the algorithm cannot guarantee regret in $o(T^{\frac{3}{2}})$. For that, let $T \geq 2$ and consider the loss sequence $g_t = (-1)^{\lfloor (2t-1)/T \rfloor}$. In other words,

$$g_1 \dots g_T = \underbrace{+1 \dots +1}_{\lfloor T/2 \rfloor} \; \underbrace{-1 \dots -1}_{\lceil T/2 \rceil}.$$

By the update rule of (OG), we have

$$x_t = \begin{cases} x_1 - \sum_{s=1}^{t-1} \frac{1}{\sqrt{s}} - \frac{1}{\sqrt{t-1}} & \text{if } 2 \leq t \leq \lfloor \frac{T}{2} \rfloor + 1 \\ x_1 - \sum_{s=1}^{\lfloor T/2 \rfloor} \frac{1}{\sqrt{s}} + \sum_{s=\lfloor T/2 \rfloor + 1}^{t-1} \frac{1}{\sqrt{s}} + \frac{1}{\sqrt{t-1}} & \text{if } t > \lfloor \frac{T}{2} \rfloor + 1 \end{cases}$$

Therefore, the regret with respect to the initialization point x_1 is

$$\begin{aligned} \mathrm{Reg}_T(x_1) &= \sum_{t=2}^{\lfloor T/2 \rfloor} \left(- \sum_{s=1}^{t-1} \frac{1}{\sqrt{s}} - \frac{1}{\sqrt{t-1}} \right) - \left(- \sum_{s=1}^{\lfloor T/2 \rfloor} \frac{1}{\sqrt{s}} - \frac{1}{\sqrt{\lfloor T/2 \rfloor}} \right) \\ &\quad - \sum_{t=\lfloor T/2 \rfloor + 2}^{T} \left(- \sum_{s=1}^{\lfloor T/2 \rfloor} \frac{1}{\sqrt{s}} + \sum_{s=\lfloor T/2 \rfloor + 1}^{t-1} \frac{1}{\sqrt{s}} + \frac{1}{\sqrt{t-1}} \right) \\ &= \sum_{s=1}^{\lfloor T/2 \rfloor - 1} \frac{-\lfloor T/2 \rfloor + s - 1}{\sqrt{s}} + \frac{1}{\sqrt{\lfloor T/2 \rfloor}} + \left\lceil \frac{T}{2} \right\rceil \sum_{s=1}^{\lfloor T/2 \rfloor} \frac{1}{\sqrt{s}} - \sum_{s=\lfloor T/2 \rfloor + 1}^{T-1} \frac{T-s+1}{\sqrt{s}} \\ &= \sum_{s=1}^{\lfloor T/2 \rfloor} \frac{\lceil T/2 \rceil - \lfloor T/2 \rfloor}{\sqrt{s}} + \frac{1}{\sqrt{\lfloor T/2 \rfloor}} \\ &\quad - \sum_{s=1}^{\lfloor T/2 \rfloor - 1} \frac{1}{\sqrt{s}} + \sum_{s=1}^{T-1} \sqrt{s} - (T+1) \sum_{s=\lfloor T/2 \rfloor + 1}^{T-1} \frac{1}{\sqrt{s}} \end{aligned}$$

We next drop the two non-negative terms in the second to last line and bound the terms of the last line from below with integrals. This gives

$$\begin{aligned} \mathrm{Reg}_T(x_1) &\geq -1 - \int_{u=1}^{\lfloor T/2 \rfloor - 1} \frac{1}{\sqrt{u}} \, du + \int_{u=0}^{T-1} \sqrt{u} \, du - (T+1) \int_{u=\lfloor T/2 \rfloor}^{T-1} \frac{1}{\sqrt{u}} \, du \\ &= 1 - 2\sqrt{\left\lfloor \frac{T}{2} \right\rfloor - 1} + \frac{2}{3}(T-1)^{\frac{3}{2}} - 2(T+1)\left(\sqrt{T-1} - \sqrt{\left\lfloor \frac{T}{2} \right\rfloor} \right) \\ &\geq -2\sqrt{\left\lfloor \frac{T}{2} \right\rfloor} - 4\sqrt{T-1} + \left(\sqrt{2} + \frac{2}{3} - 2 \right) \frac{2}{3}(T-1)^{\frac{3}{2}}. \end{aligned}$$

Since $\sqrt{2} + 2/3 - 2 \geq 0$, the last line grows in $\Theta(T^{\frac{3}{2}})$ when we increase T, showing that it is impossible for the algorithm to guarantee a regret in $o(T^{\frac{3}{2}})$. □

Remark 4.1. Stating a lower bound is often delicate. For $\mathcal{J}$ some function that maps any finite sequence of d-dimensional vectors to a real number, in our statements (see also Theorems 4.5 and 4.6), we simply say that an online learning algorithm that satisfies a certain criterion cannot guarantee a regret in $o(U)$ against feedback $g_1, \dots, g_T$ such that $\mathcal{J}(g_1, \dots, g_T) \le U$. Concretely, we show the existence of absolute constants $G, c > 0$ such that for any algorithm $\mathfrak{A}$ satisfying the criterion and any $N \in \mathbb{N}$, we can find $T \in \mathbb{N}$, $U \ge N$, and a sequence $\ell_1, \dots, \ell_T$ ensuring the following: if the learner uses algorithm $\mathfrak{A}$ against this sequence, then

1. The learner receives feedback $g_1, \dots, g_T$ with $\|g_t\| \le G$ for all t and $\mathcal{J}(g_1, \dots, g_T) \le U$.
2. The learner incurs regret larger than cU with respect to some comparator point z that is close to x_1 (e.g., $\|x_1 - z\| \le 1$).

As a matter of fact, the above result effectively indicates that for any online learning algorithm satisfying the criterion, the supremum of regret taken over all the close enough comparator points, and all the losses such that $\|g_t\| \le G$ for all t and $\mathcal{J}(g_1, \dots, g_T) \le U$ is not $o(U)$.

In light of the above negative result and the suitability of using DA in the asynchronous multi-agent setup as discussed in Remark 3.2, our focus of the next section will be an optimistic version of the (DDA) algorithm.

4.2 DELAYED OPTIMISTIC DUAL AVERAGING

In this section and the next, we places ourselves in the framework of Chapter 3 and present how delayed dual averaging can be extended to incorporate an optimistic step in the unconstrained Euclidean setup. Importantly, we show that the dual averaging step has to be done with a smaller learning rate than the optimistic step.

4.2.1 *Algorithmic Template and Regret Analysis*

Delayed optimistic dual averaging

Optimistic methods are able to leverage the slow variation of predictable sequences, thereby offering improved regret guarantees. However, when gradients arrive out of order, the predictability of a loss sequence may be compromised. To account for this, we introduce below a "separation of timescales" between the sensing and updating steps of the *delayed optimistic dual averaging* (DOptDA) method.[2]

$$\begin{aligned} X_t &= \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \sum_{s \in \mathcal{S}_t} \langle g_s, x \rangle + \frac{\|x - X_1\|^2}{2\eta_t} = X_1 - \eta_t \sum_{s \in \mathcal{S}_t} g_s, \\ X_{t+\frac{1}{2}} &= \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \langle \widetilde{g}_t, x \rangle + \frac{\|x - X_t\|^2}{2\gamma_t} = X_t - \gamma_t \widetilde{g}_t. \end{aligned} \tag{DOptDA}$$

Following our delay framework, X_t is computed using gradients from time moments $\mathcal{S}_t$. Similarly, $\widetilde{g}_t$ must be derived solely based on information available to the active agent $i(t)$ at time t. In addition to these natural restrictions, the algorithm uses two learning rates at each round: the update learning rate η_t

2 Another related algorithm is optimistic FTRL [140]. It coincides with optimistic dual averaging in the setup considered here.

and the optimistic learning rate γ_t. This additional flexibility allows us to compensate the missing information that have not arrived due to delays, giving rise to the following regret bound.

Template regret bound for DOptDA

Theorem 4.3. *Suppose that Assumption 2.1 holds and that the maximum delay is bounded by τ. Assume further that (DOptDA) is run with learning rate sequences $(\eta_t)_{t\in[T]}$, $(\gamma_t)_{t\in[T]}$ satisfying $\eta_{t+1} \le \eta_t$ and $(2\tau+1)\eta_t \le \gamma_t$ for all $t \in [T]$. Then, the regret of the algorithm (evaluated at the points $X_{\frac{3}{2}}, \dots, X_{T+\frac{1}{2}}$) satisfies*

$$\mathrm{Reg}_T(z) \le \frac{\|z - X_1\|^2}{2\eta_T} + \sum_{t=1}^{T} \frac{\gamma_t}{2} \left(\|g_t - \widetilde{g}_t\|^2 - \|\widetilde{g}_t\|^2 \right).$$

Proof. Let us consider the virtual iterates and the corresponding dual vectors

$$\widetilde{X}_t = X_1 - \eta_t \sum_{s=1}^{t-1} g_s, \quad \widetilde{Y}_t = -\eta_t \sum_{s=1}^{t-1} g_s.$$

Notice that the regret is measured with respect to the leading states

$$\langle g_t, X_{t+\frac{1}{2}} - z \rangle = \langle g_t, X_{t+\frac{1}{2}} - \widetilde{X}_{t+1} \rangle + \langle g_t, \widetilde{X}_{t+1} - z \rangle \tag{4.2}$$

For the second term, as in the proof of Proposition 2.4 we can show (see Eq. (2.6))

$$\begin{aligned} \langle g_t, \widetilde{X}_{t+1} - z \rangle \le{}& \frac{F(z, \widetilde{Y}_t)}{\eta_t} - \frac{F(z, \widetilde{Y}_{t+1})}{\eta_{t+1}} - \frac{\|\widetilde{X}_{t+1} - \widetilde{X}_t\|^2}{2\eta_t} \\ &+ \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) (h(z) - \min h). \end{aligned}$$

Since $h = \|\cdot - X_1\|^2/2$, the above equals to

$$\begin{aligned} \langle g_t, \widetilde{X}_{t+1} - z \rangle \le{}& \frac{\|z - \widetilde{X}_t\|^2}{2\eta_t} - \frac{\|z - \widetilde{X}_{t+1}\|^2}{2\eta_{t+1}} - \frac{\|\widetilde{X}_{t+1} - \widetilde{X}_t\|^2}{2\eta_t} \\ &+ \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \frac{\|z - X_1\|^2}{2}. \end{aligned} \tag{4.3}$$

For the other term, we recall the definition $\mathcal{U}_t = [t-1] \setminus \mathcal{S}_t$ and define $\nu_t = \mathrm{card}(\mathcal{U}_t)$. Then,

$$\begin{aligned} \langle g_t, X_{t+\frac{1}{2}} - \widetilde{X}_{t+1} \rangle &= \langle g_t, X_{t+\frac{1}{2}} - X_t \rangle + \langle g_t, X_t - \widetilde{X}_t \rangle + \langle g_t, \widetilde{X}_t - \widetilde{X}_{t+1} \rangle \\ &= \langle g_t, -\gamma_t \widetilde{g}_t \rangle + \langle g_t, \eta_t \sum_{s\in\mathcal{U}_t} g_s \rangle + \langle g_t, \widetilde{X}_t - \widetilde{X}_{t+1} \rangle \\ &= \frac{\gamma_t}{2} \left(\|g_t - \widetilde{g}_t\|^2 - \|g_t\|^2 - \|\widetilde{g}_t\|^2 \right) \\ &\quad + \eta_t \sum_{s\in\mathcal{U}_t} \langle g_t, g_s \rangle + \langle g_t, \widetilde{X}_t - \widetilde{X}_{t+1} \rangle \\ &\le \frac{\gamma_t}{2} \left(\|g_t - \widetilde{g}_t\|^2 - \|g_t\|^2 - \|\widetilde{g}_t\|^2 \right) \\ &\quad + \frac{\eta_t}{2} \|g_t\|^2 + \frac{1}{2\eta_t} \|\widetilde{X}_t - \widetilde{X}_{t+1}\|^2 + \frac{\nu_t \eta_t}{2} \|g_t\|^2 + \frac{\eta_t}{2} \sum_{s\in\mathcal{U}_t} \|g_s\|^2. \end{aligned} \tag{4.4}$$

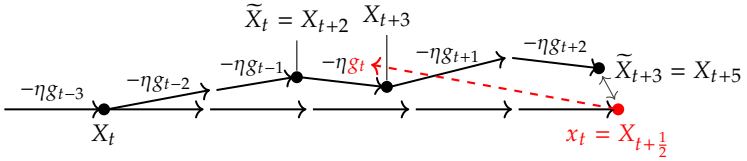


Figure 4.1: Schematic representation of (DOptDA): The delay is fixed at $\tau = 2$. We use constant learning η and guess vector $g_t = g_{t-\tau-1}$. Using a larger optimistic step helps mitigate the effect of delay.

Combining (4.2), (4.3), (4.4) and summing from $t = 1$ to T yields

$$\mathrm{Reg}_T(z) \le \frac{\|z - X_1\|^2}{2\eta_{T+1}} + \sum_{t=1}^{T} \frac{\gamma_t}{2}\left(\|g_t - \widetilde{g}_t\|^2 - \|\widetilde{g}_t\|^2\right) + \left(-\frac{\gamma_t}{2} + \frac{(\nu_t+1)\eta_t}{2} + \sum_{t\in\mathcal{U}_l}\frac{\eta_l}{2}\right)\|g_t\|^2. \tag{4.5}$$

Given that the maximum delay is τ, we have $\nu_t \le \nu \le \tau$ and if $t \in \mathcal{U}_l$ it holds $l > t \ge l - \tau$ and thus $\mathrm{card}(\{l : t \in \mathcal{U}_l\}) \le \tau$. Moreover, as $(\eta_t)_{t\in[T]}$ is a decreasing sequence, $t \in \mathcal{U}_l$ also implies $\eta_l \le \eta_t$. The last term of (4.5) can thus be bounded as following

$$\left(-\frac{\gamma_t}{2} + \frac{(\nu_t+1)\eta_t}{2} + \sum_{t\in\mathcal{U}_l}\frac{\eta_l}{2}\right)\|g_t\|^2 \le \frac{1}{2}((2\tau+1)\eta_t - \gamma_t)\|g_t\|^2 \le 0, \tag{4.6}$$

where the second inequality leverages the condition $\gamma_t \ge (2\tau+1)\eta_t$. Combining (4.5) and (4.6) and setting $\eta_{t+1} = \eta_t$ gives the desired result. □

In Theorem 4.3, we successfully show that (DOptDA) retains the desired property of undelayed optimistic gradient descent: the regret of the algorithm is solely determined by the distance between g_t and $\widetilde{g}_t$ (see Proposition 4.1). Precisely, the theorem guarantees a regret in $O\left(\sqrt{\tau\sum_{t=1}^{T}\|g_t - \widetilde{g}_t\|^2}\right)$ for fix learning rate sequences $\eta_t \equiv \eta$, $\gamma_t \equiv (2\tau+1)\eta$ that are optimally chosen. Similar to the case of delayed mirror descent and delayed dual averaging, an additional factor of $\sqrt{\tau}$ appears in the regret bound, and their regret is recovered tightly by setting $\widetilde{g}_t = 0$.

Philosophy behind the factor $2\tau+1$

To understand why we take γ_t to be $2\tau+1$ times larger than η_t, we visualize the optimistic update in Fig. 4.1. Here, we use constant learning rate η and the delay is fixed at $\tau = 2$. The latter indicates that g_t generated at x_t is used to compute X_{t+3}. On the other hand, by playing $x_t = X_t - 5\eta g_{t-3}$, we have $x_t \approx \widetilde{X}_{t+3}$ if the feedback sequence varies slowly, where $\widetilde{X}_{t+3} = X_{t+5}$ is the iterate that the learner would have played with (DA) at time $t+3$ if there were no delay. In this way, the more aggressive optimistic step makes sure that we update each X_t with a feedback element that is computed at a point close to $\widetilde{X}_t$. This mimics the behavior of the idealized algorithm that updates x_t with g_t.

Alternative approach to address delays with optimism

Provided that $\widetilde{g}_t$ can be arbitrarily chosen, in terms of algorithm design we can also set $\gamma_t = \eta_t$ and use a guess that is much larger (in the order of τ times larger than g_t as explained above). This strategy was adopted by Flaspohler et al. [80], and they also offer regret bounds that are presented in a different way. These bounds are much less comparable to the standard result presented in Proposition 4.1, but provide different insights into the algorithm. We thus

refer the readers to [80] for an alternative perspective on on how optimistic algorithms can help mitigate the delays.

More refined regret bound

Remark 4.2. The dependence on the maximum delay in Theorem 4.3 can actually be dropped. Nonetheless, we choose to present in this form for ease of understanding. Otherwise, denoting $\tau_t = \mathrm{card}(\mathcal{U}_t) + \mathrm{card}(\{s \in [T] : t \in \mathcal{U}_s\}) + 1$ (in words, we count both the number of gradient unavailable at time t and the number of times g_t is unavailable for the computation of the played point after time t) and employing a suitable constant update learning rate $\eta_t \equiv \eta$ with $\gamma_t = \tau_t \eta$, we achieve a regret in $\mathcal{O}\left(\sqrt{\sum_{t=1}^T \tau_t \|g_t - \widetilde{g}_t\|^2}\right)$. Note that $\sum_{t=1}^T \tau_t = 2D + T$ and when $\widetilde{g}_t = 0$ this bound can be inferred from Theorem 3.1 by choosing $\pi = \mathrm{id}$.

4.2.2 *Necessity of Scale Separation for Robustness to Delays*

In this part, we discuss the *necessity* of having a relatively aggressive optimistic step compared to the update ($\gamma_t \geq \eta_t$) in order to be robust to delay.

For this, we consider linear losses $\ell_t = \langle g_t, \cdot \rangle$ and uniform delay τ (i.e., every feedback becomes available after a delay of τ time steps). We define the *τ-variation* of the loss sequence by $C_T^\tau = \sum_{t=1}^T \|g_t - g_{t-\tau}\|^2$ where we set $g_t = 0$ for $t \leq 0$. For ease of notation we further write $C_T^{\tau^+} = C_T^{\tau+1}$. The following corollary is immediate from Theorem 4.3.

Upper bound on regret of DOptDA against linear losses

Corollary 4.4. *In the context of linear losses $\ell_t = \langle g_t, \cdot \rangle$ and uniform delay τ ($\mathcal{S}_t = [t - \tau - 1]$ for all t), running* (DOptDA) *with $\widetilde{g}_t = g_{t-\tau-1}$ and constant learning rates $\eta = R/\sqrt{(2\tau+1)C_T^{\tau^+}}$ and $\gamma = (2\tau+1)\eta$ guarantees the following regret bound for any $z \in \mathcal{X}$ such that $R \geq \|z - X_1\|$*

$$\mathrm{Reg}_T(z) \leq R\sqrt{(2\tau+1)C_T^{\tau^+}}.$$

This results indicates that with an optimistic learning rate γ taken $2\tau + 1$ times bigger than the update learning rate η, one can guarantee a regret bound of the order of the square root of the $(\tau+1)$-variation. In contrast, we now demonstrate the impossibility to obtain a regret that is sub-linear in $C_T^{\tau^+}$ when $\gamma = \eta$ (or even when $\gamma \leq \tau\eta$).

Lower bound on regret of DOptDA against linear losses

Theorem 4.5. *Consider the setup of Corollary 4.4. Let $\eta = \eta(R, T, \tau, C_T^{\tau^+})$ be uniquely determined by $R \geq \|z - X_1\|$, the time horizon T, the uniform delay τ, and the $(\tau+1)$-variation $C_T^{\tau^+}$. If we run* (DOptDA) *with $\widetilde{g}_t = g_{t-\tau-1}$ and $\gamma \leq \tau\eta$, it is impossible to guarantee a regret in $o(\max(C_T^{\tau^+}, \sqrt{T}))$.*

Proof. Assume, for the sake of contradiction, that there exists $\eta = \eta(R, T, \tau, C_T^{\tau^+})$ and a corresponding γ with $\gamma \leq \tau\eta$ such that (DOptDA) with $\widetilde{g}_t = g_{t-\tau-1}$ guarantees a regret in $o(\max(C_T^{\tau^+}, \sqrt{T}))$. Formally, we define for this proof specifically a "run" of the algorithm as a composition a loss sequence, a delay mechanism, a initial point X_1, and a competing vector z, and denote by $\mathcal{C}(R, T, \tau, C_T^{\tau^+})$ the set of all the runs with time horizon T, $(\tau+1)$-variation $C_T^{\tau^+}$, uniform delay τ, and $\|z - X_1\| \leq R$. Then, fixing R and τ, for every $\varepsilon > 0$, we can find $N > 0$ such that if $\max(C_T^{\tau^+}, \sqrt{T}) \geq N$, the regret achieved by the algorithm for every instance in $\mathcal{C}(R, T, \tau, C_T^{\tau^+})$ is smaller than $\varepsilon \max(C_T^{\tau^+}, \sqrt{T})$. The proof then consists in finding two instances of $\mathcal{C}(R, T, \tau, C_T^{\tau^+})$ such that the regret

t	$\tau+2$	$\dots$	p	$p+1$	$\cdots$	$p+\tau+1$
x_t	$\eta+\gamma$	$\dots$	$(p-\tau-1)\eta+\gamma$	$(p-\tau)\eta+\gamma$	$\cdots$	$p\eta+\gamma$
g_t	-1			$+1$		

t	$p+\tau+2$	$\dots$	$2p$	$2p+1$	$\dots$	$2p+\tau+1$
x_t	$(p-1)\eta-\gamma$	$\dots$	$(\tau+1)\eta-\gamma$	$\tau\eta-\gamma$	$\dots$	$-\gamma$
g_t	$+1$			-1		

Figure 4.2: Illustration of the evolution of (DOptDA) for a period of feedback in the first example of the proof of Theorem 4.5. The time is taken modulo $2p$ starting from $t = \tau + 2$, that is, the first $\tau + 1$ rounds where the algorithm outputs 0 is not shown here.

achieved by the algorithm on these two instances can not be simultaneously smaller than $\varepsilon \max(C_T^{\tau^+}, \sqrt{T})$.

For this, we fix the delay τ, set $R = 1$ without loss of generality and explicit these two instances in the following ($\mathcal{X} = \mathbb{R}$):

1. Let $K, p > \tau$ be two positive integers. We first consider a loss sequence of length $2Kp + \tau + 1$ (i.e., $T = 2Kp + \tau + 1$) as illustrated below:

$$\underbrace{\underbrace{-1 \dots -1}_{p}\ \underbrace{+1 \dots +1}_{p}}_{2Kp \text{ losses}}\ \dots\ \underbrace{\underbrace{-1 \dots -1}_{p}\ \underbrace{+1 \dots +1}_{p}}_{2Kp \text{ losses}}\ \underbrace{-1 \dots -1}_{\tau+1}$$

A period is defined as a subsequence of $2p$ losses with p consecutive -1s followed by p consecutive $+1$s. The whole loss sequence is then composed of $2K$ periods followed by $\tau + 1$ consecutive -1s. We would like to compute the regret achieved by (DOptDA) with $\eta, \gamma, \widetilde{g}_t$ as specified in the statement and $X_1 = z = 0$.

For the first $\tau + 1$ steps, the algorithm stays at $X_1 = z$ so the accumulative regret is 0. For the remaining of the run, the algorithm goes through the same trajectory for each period of delayed feedback vectors it receives and this happens K times. To compute the regret, we just need to match the position of the iterate with the actual loss at each moment, which is done in Fig. 4.2 (as the loss vectors of a single period sum to 0, after receiving all the vectors from one period it is as if we started again from $X_1 = z = 0$). Notice that the algorithm uses the most recent vector it receives for the optimistic step.

The regret for each period of feedback is thus

$$\begin{aligned}
\mathrm{Reg}_{per} &= \frac{-(p-\tau-1)(p-\tau)\eta}{2} - (p-\tau-1)\gamma + \frac{(\tau+1)(2p-\tau)\eta}{2} + (\tau+1)\gamma \\
&\quad + \frac{(p-\tau-1)(p+\tau)\eta}{2} - (p-\tau-1)\gamma - \frac{(\tau+1)\tau\eta}{2} + (\tau+1)\gamma \\
&= (\tau+1)(p-\tau)\eta + (p-\tau-1)\tau\eta + 2(2\tau-p+2)\gamma \\
&= (\eta + 2\tau\eta - 2\gamma)p - 2\tau(\tau+1)\eta + (4\tau+4)\gamma.
\end{aligned}$$

Accordingly, the total regret is

$$\mathrm{Reg}_1 = K((\eta + 2\tau\eta - 2\gamma)p - 2\tau(\tau+1)\eta + (4\tau+4)\gamma) \geq K(p - 2\tau(\tau+1))\eta,$$

where for the inequality we use the fact that $\gamma \leq \tau\eta$.

Moreover, for every $m \in \mathbb{N}_0$, from time $2mp + \tau + 2$ to $2mp + 2p + \tau + 1$ the $(\tau+1)$-variation increases by $8(\tau+1)$: there are $\tau+1$ switches both from -1 to $+1$ and from $+1$ to -1 with each switch contributing 4 to the variation. Remember also that in the definition of the $C_T^{\tau^+}$ we compare the first $\tau+1$ losses with 0. For the whole sequence we therefore count $C_T^{\tau^+} = (8K+1)(\tau+1)$.

2. We now construct another example with the same $T, C_T^{\tau^+}$ as follows (with $p > 4\tau + 4$):

$$\underbrace{\underbrace{0\dots0}_{\tau+1}\ \underbrace{1\dots1}_{\tau+1}\ \dots\ \underbrace{0\dots0}_{\tau+1}\ \underbrace{1\dots1}_{\tau+1}}_{8K(\tau+1)\text{ losses}}\ \underbrace{0\dots0}_{2Kp-8K(\tau+1)}\ \underbrace{1\dots1}_{\tau+1}$$

In particular, $2Kp - 8K(\tau+1) > 2K > \tau + 1$. It follows immediately $C_T^{\tau^+} = (8K+1)(\tau+1)$ and of course $T = 2Kp + \tau + 1$.

Let $X_1 = 0$ and $z = -1$. In the sequence the loss 1 appears $(4K+1)(\tau+1)$ times while the remaining feedback are all 0s. Given that the last $\tau+1$ losses are never received by the algorithm, we have indeed always $x_t \geq -4K(\tau+1)\eta - \gamma$. The regret can therefore be lower bounded as:

$$\begin{aligned}
\mathrm{Reg}_2 &= \sum_{t=1}^{T} g_t(x_t + 1) \\
&= \sum_{t=1}^{T} g_t x_t + (4K+1)(\tau+1) \\
&\geq (4K+1)(\tau+1) - 4K(4K+1)(\tau+1)^2\eta - (4K+1)(\tau+1)\gamma \\
&\geq (4K+1)(\tau+1) - (4K+1)^2(\tau+1)^2\eta,
\end{aligned}$$

where in the last inequality we use again $\gamma \leq \tau\eta$.

Conclude. We choose K, p so that $p = (16K+9)(\tau+1)^2 + 2\tau(\tau+1) > 4\tau + 4$. Notice that T and $C_T^{\tau^+}$ can be made arbitrarily large. We run the algorithm in question on the two problem instances described above. We have on one side

$$\mathrm{Reg}_1 \geq K(p - 2\tau(\tau+1))\eta = (16K^2 + 9K)(\tau+1)^2\eta.$$

On the other side,

$$\begin{aligned}
\mathrm{Reg}_2 &\geq (4K+1)(\tau+1) - (4K+1)^2(\tau+1)^2\eta \\
&\geq (4K+1)(\tau+1) - (16K^2+9K)(\tau+1)^2\eta.
\end{aligned}$$

Recalling that $C_T^{\tau^+} = (8K+1)(\tau+1)$, the above shows

$$\mathrm{Reg}_1 + \mathrm{Reg}_2 \geq (4K+1)(\tau+1) \geq C_T^{\tau^+}/2.$$

Similarly, we have $T = 2Kp + \tau + 1 \leq (32K^2 + 22K)(\tau+1)^2$. As a consequence

$$\mathrm{Reg}_1 + \mathrm{Reg}_2 \geq (4K+1)(\tau+1) \geq \sqrt{T}/2.$$

To summarize, we have proven for some T and $C_T^{\tau^+}$ arbitrarily large, we can find two instances from $\mathcal{C}(R, T, \tau, C_T^{\tau^+})$ so that the regrets achieved by the algorithm on these two instances satisfy

$$\max(\mathrm{Reg}_1, \mathrm{Reg}_2) \geq \max(C_T^{\tau^+}, \sqrt{T})/2.$$

This is in contradiction with the initial hypothesis by choosing $\varepsilon = 1/2$. □

The above result is a generalization of the lower bound of Chiang et al. [45] that applies to the undelayed setup. However, in their proof, the learning rate was first fixed and then a loss sequence was constructed to yield large regret, which could possibly also prevent optimistic algorithms to achieve low regret. Our approach fixes this fallacy by informing the algorithm of the variation in advance so that optimistic algorithms provably obtain low regrets on these sequences (cf. Corollary 4.4). In the undelayed setting, we recover the result that the optimistic step is necessary to guarantee a regret in $O\left(\sqrt{\sum_{t=1}^{T}\|g_t - g_{t-1}\|^2}\right)$.

Finally, we also demonstrate that among all the online algorithms with the same prior information, the bound achieved in Corollary 4.4 is tight in its dependence on τ and $C_T^{\tau^+}$.

General lower bound for delayed online learning

Theorem 4.6. *Consider the setup of Corollary 4.4. No online learning algorithm that solely uses the information of T, τ, and $\overline{C^\tau} \geq C_T^{\tau^+}$ can guarantee regret $\mathrm{Reg}_T(z) = o(\sqrt{\tau\overline{C^\tau}})$ for any $z \in \mathcal{B}(x_1, 1)$ and any sequence of loss functions of length T and $(\tau+1)$-variation $C_T^{\tau^+}$.*

Proof. Let $N \in \mathbb{N}$ and $\mathfrak{A} = \mathfrak{A}(\tau, T, \overline{C^\tau})$ be an arbitrary online algorithm compatible with delayed feedback that at most uses the information about T, τ, and $\overline{C^\tau}$. We choose τ and $\overline{C^\tau}$ such that $\sqrt{\tau\overline{C^\tau}} \geq N$ and $p := \overline{C^\tau}/(4(\tau+1))$ is a sufficiently large positive integer. We set $T = (\tau+1)p$.

Following Shalev-Shwartz [245], we consider linear losses with either +1 or −1 loss vectors. For any online learning algorithm, we know by [245, Thm. 3] that there exists a such sequence and such that $\max_{z\in\mathcal{B}(x_1,1)} \mathrm{Reg}_p(z) = \Omega(\sqrt{p})$; that is, there exists an absolute constant c such that when p is large enough it holds $\max_{z\in\mathcal{B}(x_1,1)} \mathrm{Reg}_p(z) \geq c\sqrt{p}$.

We now apply the this lower bound to the following algorithm $\mathfrak{A}_{/\tau}$: We repeat each loss $\tau+1$ times, send each feedback after a delay of τ, run $\mathfrak{A}$ on this new loss sequence with delayed feedback and every $\tau+1$ iterations we return $x_k = \sum_{t=a_k}^{a_k+\tau} X_t/(\tau+1)$, where X_t is the iterate produced by $\mathfrak{A}$ on the constructed losses and $a_k = (k-1)(\tau+1)+1$. In more detail, for the loss sequence $g_1, \ldots, g_p$, at the end of iteration $k(\tau+1)$ to $k(\tau+1)+\tau$ we receive feedback g_k (except for the case $k = 0$ where we receive nothing) while we suffer loss $\langle g_k, x_t\rangle$ from iteration $a_k = (k-1)(\tau+1)+1$ to $a_k + \tau = k(\tau+1)$. This is a legitimate online algorithm because x_k can indeed be computed after receiving $g_1, \ldots, g_{k-1}$. Therefore, the $\Omega(\sqrt{p})$ lower bound applies to $\mathfrak{A}_{/\tau}$.

To conclude, we note that the regret achieved by $\mathfrak{A}$ on the constructed sequence is exactly $\tau+1$ times the regret achieved by $\mathfrak{A}_{/\tau}$ on the original sequence. Moreover, the $(\tau+1)$-variation $C_T^{\tau^+}$ of the constructed sequence is effectively bounded from above by $\overline{C^\tau}$ since $(\tau+1)+4(\tau+1)(p-1) < \overline{C^\tau}$ and the incurred regret is lower bounded by $c(\tau+1)\sqrt{p} \sim c\sqrt{\tau\overline{C^\tau}}/2$ (where $\sim$ stands for asymptotically equivalent). □

In summary, in this subsection we showed that using (DOptDA) *with a double learning rate strategy* enables to achieve an $O(\sqrt{\tau C_T^{\tau^+}})$ regret which is tight among online learning methods and out of reach of single learning rate (DOptDA). This justifies both the optimality of our approach and the necessity of the modification that we brought to the algorithm.

4.3 DELAYED ONLINE LEARNING WITH SLOW VARIATION

Now that we have established our foundational results concerning the optimistic variant of delayed dual averaging, we turn our attention to the specific case where the loss sequence varies slowly across iterations.

4.3.1 *Full Information Setup and Choices of Guess Vectors*

We start by investigating the choice of the guess vector $\widetilde{g}_t$. For this, we consider the case where the losses are differentiable and the full gradient $\nabla \ell_t$ is obtained as feedback (and not only $g_t = \nabla \ell_t(X_t)$). Using this kind of feedback, we can compute the gradient of the last received function at the current point immediately,[3] and use it as a guess for the current function's gradient. Formally, we make the following assumption.

Full-information feedback

Assumption 4.1. The loss functions are differentiable and the feedback associated to time step t is the whole vector field $V_t = \nabla \ell_t$, the evaluation of which at any point $x \in \mathbb{R}^d$ is immediate and does not induce any delay.

The requirement of having access to the entire $V_t = \nabla \ell_t$ is particularly satisfied in the "full-information" online learning model as we discussed in Remark 2.1. With this assumption, we can then set $\widetilde{g}_t = \widetilde{V}_t(X_t)$ where $\widetilde{V}_t$ is some *past* vector field (i.e., $\widetilde{V}_t = V_s$ for some $s \in \mathcal{S}_t$). However, even in this case, the point where the gradient is evaluated is still different from the point that is played (X_t versus $X_{t+\frac{1}{2}}$). We thus also need the gradient fields to be Lipschitz continuous to ensure that the difference $\|g_t - \widetilde{g}_t\|$ is small.

Lipschitz continuity of gradient fields

Assumption 4.2. There exists $L > 0$ such that for all $t \in \mathbb{N}$, V_t is L-Lipschitz continuous.

Given Assumptions 4.1 and 4.2, we are now in a position to formulate a regret bounds for a choice of $\widetilde{g}_t$ that can be relevant in many applications.

Regret bound for DOptDA with guess evaluated at X_t

Theorem 4.7. *Suppose that Assumptions 2.1, 4.1 and 4.2 hold and that the maximum delay is bounded by τ. Take $\widetilde{g}_t = \widetilde{V}_t(X_t)$, $\eta_{t+1} \leq \eta_t$, $(2\tau+1)\eta_t \leq \gamma_t$, and $2\gamma_t^2 L^2 \leq 1$. Then, the regret of* (DOptDA) *(evaluated at the points $X_{\frac{3}{2}}, \dots, X_{T+\frac{1}{2}}$) satisfies*

$$\mathrm{Reg}_T(z) \leq \frac{\|z - X_1\|^2}{2\eta_T} + \sum_{t=1}^{T} \gamma_t \|V_t(x_t) - \widetilde{V}_t(x_t)\|^2.$$

Proof. The proof is immediate from Theorem 4.3. Indeed,

$$\|V_t(X_{t+\frac{1}{2}}) - \widetilde{V}_t(X_t)\|^2 \leq 2\|V_t(X_{t+\frac{1}{2}}) - V_t(X_t)\|^2 + 2\|V_t(X_t) - \widetilde{V}_t(X_t)\|^2.$$

3 i.e., without any delay, the delays considered here are either due to communication between agents or inherent to the feedback mechanism.

Then, using the Lipschitz continuity of $\widetilde{V}_t$ and the condition $2\gamma_t^2 L^2 \leq 1$, we have:

$$2\|V_t(X_{t+\frac{1}{2}}) - V_t(X_t)\|^2 \leq 2L^2\|X_{t+\frac{1}{2}} - X_t\|^2 = 2\gamma_t^2 L^2 \|\widetilde{V}_t(X_t)\|^2 \leq \|\widetilde{V}_t(X_t)\|^2.$$

In other words, we have proven $\|g_t - \widetilde{g}_t\|^2 - \|\widetilde{g}_t\|^2 \leq 2\|V_t(X_t) - \widetilde{V}_t(X_t)\|^2$ and the bound follows. □

Theorem 4.7 reduces the problem of choosing an adequate vector $\widetilde{g}_t$ to that of choosing a vector field $\widetilde{V}_t$ which approximates well V_t. In our setup of full gradient feedback with a loss sequence evolving slowly over time, one natural option is reuse some recent function for the constitution of $\widetilde{V}_t$. Since we are in a distributed setting, the evolution of the loss functions may have both global and local components. We discuss these two typical cases below.[4]

Examples choice of the vector field $\widetilde{V}_t$

Example 4.1 (Global variation)**.** If the loss functions vary slowly following a global trend, we can timestamp every gradient field which makes it possible to choose $\widetilde{V}_t = V_{\tilde{t}}$ where $\tilde{t} = \max \mathcal{S}_t$, i.e., the active agent $i(t)$ uses the most recent data available at hand (independent of its source) when playing X_t. This would however require the agents to share the whole vector field V_t.

Example 4.2 (Local variation)**.** If the loss functions vary slowly for all the agents, the active agent $i(t)$ can choose the last feedback corresponding to a point it played, i.e., $\widetilde{V}_t = V_{\tilde{t}}$ where $\tilde{t} = \max\{s \in \mathcal{S}_t : i(s) = i(t)\}$. Compared to Example 4.1, we gain in terms of both data privacy and communication efficiency since only the gradients g_t need to be shared among the agents in this scenario.

Regret with constant learning rates

Denoting the total deviation of our approximation by $C_T = \sum_{t=1}^{T} \|V_t(x_t) - \widetilde{V}_t(x_t)\|^2$, Theorem 4.7 guarantees a regret of $O(R^2\tau L + R\sqrt{\tau C_T})$ for suitably chosen constant learning rate sequences $\eta_t \equiv \eta$ and $\gamma_t \equiv \gamma$. In both Examples 4.1 and 4.2, C_T encapsulates the temporal variation of the loss sequence. As such, we have effectively demonstrated how (DOptDA) can leverage this temporal variation to its advantage, providing significant benefits over traditional methods in scenarios where the loss sequence exhibits slow variation over time.

A remark on the name of the algorithm

Remark 4.3. When there is no delay, our algorithm simply uses $\widetilde{V}_t = V_{t-1}$ and hence $\widetilde{g}_t = V_{t-1}(X_t)$. The evaluation of the gradient at X_t makes it closer to the *dual extrapolation* algorithm of Nesterov [211], which is itself closely related to the *extra-gradient* method of Korpelevich [160]. On the other hand, the term "optimistic" is more frequently used to refer to the case where V_{t-1} is evaluated at $X_{t-\frac{1}{2}}$. We discuss this point in more detail in Section 5.3.

4.3.2 *Adaptive Learning Rate*

In general, the deviation C_T is not known in advance, and hence neither is the optimal choice of η and γ that allows us to obtain the aforementioned regret guarantee. To circumvent this issue, we can again design an adaptive learning rate schedule in the spirit of (AdaGrad-norm). For the sake of simplicity, we assume for the following result that the agents have access to τ a bound on the delay, G a bound on the magnitude of the loss gradient, and L a bound on the Lipschitz modulus of these gradients.

4 In the two cases, we may simply set $\widetilde{V}_t = 0$ when the corresponding set is empty.

Regret bound for DOptDA with adaptive learning rate

Theorem 4.8. *Suppose that Assumptions 2.1, 3.2, 4.1 and 4.2 hold and that the maximum delay is bounded by τ. Assume further that both V_t and $\widetilde{V}_t$ have their norm bounded by G. Then, running* (DOptDA) *with $\widetilde{g}_t = \widetilde{V}_t(X_t)$,*

$$\gamma_t = \min\left(\frac{R\sqrt{4\tau+1}}{2\sqrt{\left(\sum_{s\in\mathcal{S}_t}\|V_s(X_s)-\widetilde{V}_s(X_s)\|^2+4G^2(\tau+1)\right)}}, \frac{1}{\sqrt{2}L}\right),$$

and

$$\eta_t = \min\left(\frac{R}{2\sqrt{(4\tau+1)\left(\sum_{s\in\mathcal{S}_t}\|V_s(X_s)-\widetilde{V}_s(X_s)\|^2+4G^2(3\tau+1)\right)}}, \frac{1}{\sqrt{2}L(4\tau+1)}\right)$$

guarantees for all $z\in\mathcal{X}$ such that $\|z-X_1\|\le R$ and all $T\in\mathbb{N}$ that

$$\mathrm{Reg}_T(z) \le \max\left(\sqrt{2}R^2L(4\tau+1), 2R\sqrt{(4\tau+1)(C_T+4G^2(3\tau+1))}\right).$$

Remark 4.4. At the price of a worse dependence on the constants, we can use the difference $\|V_t(X_{t+\frac{1}{2}})-\widetilde{V}_t(X_t)\|$ instead of $\|V_t(X_t)-\widetilde{V}_t(X_t)\|$ in the computation of the learning rates, which prevents us from an extra evaluation of the vector field; see e.g., [140, Corollary 9].

Similar to before, we require Assumption 3.2 here to get a practical data-dependent guarantee in the multi-agent setup. Compared to the optimal regret that can be achieved with prior knowledge of C_T, the bound of Theorem 4.8 is only degraded by a constant factor. Implementing this learning rate schedule necessitates the computation of both γ_t and η_t, which in turn requires the agents to relay $\|V_t(X_t)-\widetilde{V}_t(X_t)\|$ in addition to $g_t = V_t(X_{t+\frac{1}{2}})$ after receiving V_t.

In order prove Theorem 4.8, we generalize both Theorem 4.3 and Theorem 4.7 to the case where the learning rate is non-increasing along a faithful permutation.

Proposition 4.9. *Suppose that Assumption 2.1 holds and that the maximum delay is bounded by τ. Consider a faithful permutation π and let* (DOptDA) *be run with learning rate sequences $(\eta_t)_{t\in[T]}$, $(\gamma_t)_{t\in[T]}$ satisfying $\eta_{\pi(t+1)}\le\eta_{\pi(t)}$ and $(4\tau+1)\max_{\{s:|s-t|\le\tau\}}\eta_s\le\gamma_t$ for all t. Then, the regret of the algorithm (evaluated at the points $X_{\frac{3}{2}},\dots,X_{T+\frac{1}{2}}$) satisfies*

$$\mathrm{Reg}_T(z) \le \frac{\|z-X_1\|^2}{2\eta_{\pi(T)}} + \sum_{t=1}^{T}\frac{\gamma_t}{2}\left(\|g_t-\widetilde{g}_t\|^2-\|\widetilde{g}_t\|^2\right).$$

Proof. We define the virtual iterates

$$\widetilde{X}_t = X_1 - \eta_{\pi(t)}\sum_{s=1}^{t-1} g_{\pi(s)}.$$

We then decompose

$$\langle g_t, X_{\pi(t)+\frac{1}{2}}-z\rangle = \langle g_t, X_{\pi(t)+\frac{1}{2}}-\widetilde{X}_{t+1}\rangle + \langle g_t, \widetilde{X}_{t+1}-z\rangle.$$

Following closely the proof of Theorem 4.3, we obtain

$$\mathrm{Reg}_T(z) \leq \frac{\|z - X_1\|^2}{2\eta_{\pi(T)}} + \sum_{t=1}^{T} \frac{\gamma_t}{2} \left(\|g_t - \widetilde{g}_t\|^2 - \|\widetilde{g}_t\|^2 \right) + \left(-\frac{\gamma_{\pi(t)}}{2} + \frac{(\mathrm{card}(\mathcal{U}_t^\pi) + 1)\eta_{\pi(t)}}{2} + \sum_{\pi(t) \in \mathcal{U}_l^\pi} \frac{\eta_{\pi(l)}}{2} \right) \|g_{\pi(t)}\|^2.$$

Invoking Proposition 3.8, we know that $[t]^\pi \setminus \mathcal{S}_{\pi(t)} \subseteq \{\pi(t) - \tau, ..., \pi(t) + \tau\}$. Given that $\pi(t) \notin [t-1]^\pi$, this implies $\mathcal{U}_t^\pi \subseteq \{\pi(t) - \tau, ..., \pi(t) - 1\} \cup \{\pi(t) + 1, ..., \pi(t) + \tau\}$. Therefore, $\mathrm{card}(\mathcal{U}_t^\pi) \leq 2\tau$ and if $\pi(t) \in \mathcal{U}_l^\pi$ then $|\pi(t) - \pi(l)| \leq \tau$ while $\pi(t) \neq \pi(l)$, which also shows $\mathrm{card}(\{l : \pi(t) \in \mathcal{U}_l^\pi\}) \leq 2\tau$. Accordingly,

$$\frac{(\mathrm{card}(\mathcal{U}_t^\pi) + 1)\eta_{\pi(t)}}{2} + \sum_{\pi(t) \in \mathcal{U}_l^\pi} \frac{\eta_{\pi(l)}}{2} \leq \frac{(4\tau + 1) \max_{\{s : |s - \pi(t)| \leq \tau\}} \eta_s}{2}.$$

With the assumption $\gamma_t \geq (4\tau + 1) \max_{\{s : |s-t| \leq \tau\}} \eta_s$, we effectively deduce

$$\mathrm{Reg}_T(z) \leq \frac{\|z - X_1\|^2}{2\eta_{\pi(T)}} + \sum_{t=1}^{T} \frac{\gamma_t}{2} \left(\|g_t - \widetilde{g}_t\|^2 - \|\widetilde{g}_t\|^2 \right).$$

This proves the theorem. □

Proposition 4.10. *Suppose that Assumptions 2.1, 4.1 and 4.2 hold and that the maximum delay is bounded by τ. Consider a faithful permutation π and take $\widetilde{g}_t = \widetilde{V}_t(X_t)$, $\eta_{\pi(t+1)} \leq \eta_{\pi(t)}$, $(4\tau + 1) \max_{\{s : |s-t| \leq \tau\}} \eta_s \leq \gamma_t$, and $2\gamma_t^2 L^2 \leq 1$. Then, the regret of (DOptDA) (evaluated at the points $X_{\frac{3}{2}}, \ldots, X_{T+\frac{1}{2}}$) satisfies*

$$\mathrm{Reg}_T(z) \leq \frac{\|z - X_1\|^2}{2\eta_{\pi(T)}} + \sum_{t=1}^{T} \gamma_t \|V_t(X_t) - \widetilde{V}_t(X_t)\|^2.$$

Proof. This is proved by applying Proposition 4.9 and bounding the term $\|g_t - \widetilde{g}_t\|^2 - \|\widetilde{g}_t\|^2$ as in the proof of Theorem 4.7. □

Thanks to Propositions 4.9 and 4.10, we can now provide regret guarantee for the case where the learning rate is not non-increasing. In particular, this enables us to prove Theorem 4.8.

Proof of Theorem 4.8. Let $\widetilde{C}_t = \sum_{s \in \mathcal{S}_t} \|V_s(X_s) - \widetilde{V}_s(X_s)\|^2$. We consider a permutation π such that (*i*) if $\widetilde{C}_s < \widetilde{C}_t$ then $\pi^{-1}(s) < \pi^{-1}(t)$; (*ii*) if $\widetilde{C}_s = \widetilde{C}_t$ and $s \in \mathcal{S}_t$ then $\pi^{-1}(s) < \pi^{-1}(t)$. The sequence $(\widetilde{C}_t)_t$ is non-decreasing along π (see e.g., proof of Proposition 3.5) and accordingly the learning rate sequence $(\eta_t)_{t \in [T]}$ is non-increasing along π. Moreover, if $s \in \mathcal{S}_t$, we have $\mathcal{S}_s \subseteq \mathcal{F}_s^{i(t)} \subseteq \mathcal{S}_t$ thanks to Assumption 3.2. This implies $\widetilde{C}_s \leq \widetilde{C}_t$; subsequently $\pi^{-1}(s) < \pi^{-1}(t)$. The above shows that π is a faithful permutation. The condition $2\gamma_t^2 L^2 \leq 1$ is automatically satisfied by the definition of γ_t. To apply Proposition 4.10, the last missing piece is to prove $(4\tau + 1) \max_{\{s : |s-t| \leq \tau\}} \eta_s \leq \gamma_t$. This boils down to showing that

$$\widetilde{C}_s + 4G^2(3\tau + 1) \geq \widetilde{C}_t + 4G^2(\tau + 1) \tag{4.7}$$

for all $s \in [T] \cap \{t-\tau, ..., t+\tau\}$. The maximum delay being bounded by τ, we have $|\operatorname{card}(\mathcal{S}_s) - \operatorname{card}(\mathcal{S}_t)| \leq |s-t| + \tau$. By bounding each $\|V_l(X_l) - \widetilde{V}_l(X_l)\|^2$ by $4G^2$, we indeed prove (4.7) for s such that $|s-t| \leq \tau$.

With all this at hand, applying Proposition 4.10 gives

$$\operatorname{Reg}_T(z) \leq \frac{\|z - X_1\|^2}{2\eta_{\pi(T)}} + \sum_{t=1}^{T} \gamma_t \|V_t(X_t) - \widetilde{V}_t(X_t)\|^2.$$

As the maximum delay is bounded by τ and the gradients are bounded by G, we have $\widetilde{C}_t + 4G^2(\tau+1) \geq C_t$. Invoking Lemma 2.6 then gives

$$\begin{aligned} \operatorname{Reg}_T(z) &\leq \frac{\|z - X_1\|^2}{2\eta_{\pi(T)}} + \frac{R\sqrt{4\tau+1}}{2} \sum_{t=1}^{T} \frac{1}{\sqrt{C_t}} \|V_t(X_t) - \widetilde{V}_t(X_t)\|^2 \\ &\leq \frac{R^2}{2\eta_{\pi(T)}} + R\sqrt{(4\tau+1)C_T}. \end{aligned} \tag{4.8}$$

We bound the second term by

$$R\sqrt{(4\tau+1)C_T} \leq R\sqrt{(4\tau+1)(\widetilde{C}_T + 4G^2(3\tau+1))} \leq \frac{R^2}{2\eta_T} \leq \frac{R^2}{2\eta_{\pi(T)}}. \tag{4.9}$$

Combining (4.8) and (4.9) we get $\operatorname{Reg}_T(z) \leq R^2/\eta_{\pi(T)}$. We can conclude by using the definition of $\eta_{\pi(T)}$ and $\widetilde{C}_{\pi(T)} \leq C_T$. □

Part II

ADAPTIVE LEARNING IN GAMES

5

FROM ONLINE LEARNING TO LEARNING IN GAMES

In this part of the thesis, we shift our focus to the specific setup of learning in games. Unlike Part I, where agents deal with arbitrary feedback and share a common objective, here the feedback arises from the intricate interaction among agents, each possessing their own distinct goals. This framework aptly models numerous real-world scenarios characterized by conflicting interests among agents, such as financial markets, traffic routing problems, and online auctions [90, 166, 217].

Inherent to this setup is a nontrivial interdependence between agents, where the optimal strategy of each player is intricately tied to those of others. This new context presents its own unique set of challenges that we strive to address in this part. Notably, we opt to set aside the complexity introduced by delays, which was the core focus of the first part of this manuscript. Although a few studies, such as those by Huang and Hu [132] and by Zhou et al. [295], have made commendable attempts to incorporate delays into learning in games, the addition of adaptive learning rates, a key pillar of this thesis as mentioned in Chapter 1, greatly amplifies this complexity.

In this chapter, we aim to provide an introductory overview of the learning-in-games framework. We elaborate on the objectives, clarify the underlying assumptions, and offer detailed discussions on the pivotal role of optimistic algorithms in this context. Further, we illuminate the connection between learning in games and variational inequalities. This chapter thus forms a comprehensive base, preparing us for the more intricate discussions and analyses that follow in the subsequent chapters.

OUTLINE OF THIS CHAPTER. This chapter is organized as follows. In Section 5.1, we introduce the framework of learning in games that forms the basis of our study. We also outline several important notions and assumptions that are essential to our analysis. Subsequently, we present variational inequalities and draw parallels between learning in games and variational inequalities in Section 5.2. Finally, we zoom in on optimistic algorithms and highlight their advantages in the realm of learning in games in Section 5.3, thereby setting the stage for our work in the forthcoming chapters.

5.1 LEARNING IN GAMES

Broadly speaking, a game is a mathematical model that captures the strategic interactions between multiple decision-makers, or *players*. Games can take various forms depending on the rules of play and information available to the players. For instance, games can be classified as extensive-form or normal-form games, symmetric or asymmetric games, and games with perfect or imperfect information. In this thesis, we focus exclusively on normal-form games with

At each round $t = 1, 2, \ldots$, each player $i \in \mathcal{N}$

- Plays an action $x_t^i \in \mathcal{X}^i$
- Suffers loss $\ell^i(x_t^1, \ldots, x_t^N)$ and receives corresponding feedback

Figure 5.1: The learning-in-games framework.

continuous action spaces and a finite set of players. We refer to these games as continuous games and provide the formal definition below.

Continuous games

Definition 5.1 (Continuous games)**.** A continuous game is a normal-form game played by a finite set of players $\mathcal{N} := 1, \ldots, N$. Each player $i \in \mathcal{N}$ is associated with a closed convex action set $\mathcal{X}^i \subseteq \mathbb{R}^{d_i}$ and a loss function $\ell^i \colon \mathfrak{X} \to \mathbb{R}$, where $\mathfrak{X} := \bigtimes_{i=1}^N \mathcal{X}^i$ denotes the game's joint action space.

Remark 5.1. In the literature of game theory, it is more common to use words such as reward, utility, or payoff instead of loss. In that case, the loss of a player should be the opposite of these quantities.

Continuous games are natural extensions of games with discrete action space. They arise in a variety of real-world problems, including resource allocation, network routing, and deep learning. More concretely, typical examples of continuous games include mixed extension of finite games, auctions, markets, saddle-point problems, and interactions between multiple neural networks such as generative adversarial networks (GANs) [101]. In this part of the thesis, we are going to focus on the "learning" aspect in continuous games. We will consider repeated interactions between players and investigate the outcome of these interactions in various situations.

NOTATIONS. Throughout the following, any quantity associated to player i (resp., all players but player i) is by default written with superscript i (resp. $-i$), and any ensemble of actions or functions whose definition involves multiple players is typeset in bold. In particular, we write $\mathbf{x} = (x^i, \mathbf{x}^{-i}) \in \mathfrak{X}$ for the action profile of all players, where x^i and $\mathbf{x}^{-i}$ respectively denote the action of player i and the joint action of all players other than i.

5.1.1 *Interaction Model, Regret, and Equlibrium*

Learning in games

In the multi-agent learning model that we consider, players interact with each other repeatedly via a continuous game. As shown in Fig. 5.1, during each round t of the process, each player i selects an action x_t^i from their action set $\mathcal{X}^i$ and suffers a loss $\ell^i(\mathbf{x}_t)$. At the end of the round, the players receive as feedback an estimate of their individual loss gradient [1]

$$g_t^i \approx V^i(\mathbf{x}_t) := \nabla_i \ell^i(x_t^i, \mathbf{x}_t^{-i}),$$

and the process repeats. Concretely, we have $g_t^i = V^i(\mathbf{x}_t)$ in the case of perfect feedback (Chapter 6), and we discuss the case of stochastic feedback in Chapters 7 and 8 (see Assumption 7.1).

1 Technically speaking, we may sometimes need to define V^i as a Lipschitz continuous selection of subgradient of $\ell^i(\cdot, \mathbf{x}_t^{-i})$. Our analysis still holds in this case.

Regret

From the viewpoint of a single player $i \in \mathcal{N}$, the loss of round t is exactly $\ell^i(\cdot, \mathbf{x}_t^{-i})$. Therefore, the notion of regret (Definition 2.1) can be readily translated into this game-theoretic setup as

$$\mathrm{Reg}_T^i(z^i) = \sum_{t=1}^{T} \ell^i(x_t^i, \mathbf{x}_t^{-i}) - \ell^i(z^i, \mathbf{x}_t^{-i}),$$

where $z^i \in \mathcal{X}^i$ is the fixed comparator. It is also convenient to define the regret with respect to a fixed comparator set $\mathcal{Z}^i \subseteq \mathcal{X}^i$ as

$$\mathrm{Reg}_T^i(\mathcal{Z}^i) \coloneqq \max_{z^i \in \mathcal{Z}^i} \mathrm{Reg}_T^i(z^i).$$

In fact, in the subsequent sections, we will mainly state our regret bounds using O notation, and we thus prefer to provide directly the bound with respect to an entire set.[2] We say that a sequence of play x_t^i of player i incurs *no regret* (at $\mathbf{x}_t^{-i}$) if $\mathrm{Reg}_T^i(\mathcal{Z}^i) = o(T)$ for every (compact) set of alternative strategies.

No-regret and coarse-correlated equilibrium

Besides the online learning interpretation of regret as a minimal performance guarantee, it is well known that the empirical frequency of no-regret play in *finite* games converges to the set of coarse correlated equilibriums (CCEs) —also known as the game's *Hannan set* [108, 110]. This provides yet another motivation to study algorithms that provably achieve low regret. Nonetheless, the aforementioned convergence result is to be taken with a grain of salt. First, the type of convergence involved does *not* concern the actual, day-to-day play but the empirical frequency of play. Moreover, the game's CCEs set may contain elements that fail even the most basic rationalizability axioms. In particular, Viossat and Zapechelnyuk [272] constructed a simple two-player game that admits CCEs supported *exclusively* on strictly dominated strategies.

Optimality at a single round

In this regard, it is more favorable to provide guarantee on the optimality of actual the iterate of play of each round. One such metric is the gap function.

Definition 5.2 (Gap function)**.** Let $i \in \mathcal{N}$ and $\mathcal{Z}^i \subseteq \mathcal{X}^i$. The gap function of player i with respect to the set $\mathcal{Z}^i$ is defined by

$$\mathrm{Gap}_{\mathcal{Z}^i}^i(\mathbf{x}) \coloneqq \ell^i(x^i, \mathbf{x}^{-i}) - \min_{z^i \in \mathcal{Z}^i} \ell^i(z^i, \mathbf{x}^{-i}).$$

In words, it is the best that the player could have gained by switching to any other strategy in $\mathcal{Z}^i$ when all players play $\mathbf{x}$.

The gap function evaluated at a played point x_t is sometimes referred to as instantaneous regret or simple regret in the literature. However, instead of providing a bound on these quantities, we instead focus on another closely related concept, that of Nash equilibrium.

Nash equilibrium

Definition 5.3 (Nash equilibrium)**.** A Nash equilibrium is strategy profile from which no player has incentive to deviate unilaterally. Formally, a point $\mathbf{x}_\star \in \mathcal{X}$ is a Nash equilibrium if for all $i \in \mathcal{N}$ and all $x^i \in \mathcal{X}^i$, $\ell^i(x_\star^i, \mathbf{x}_\star^{-i}) \leq \ell^i(x^i, \mathbf{x}_\star^{-i})$.

Nash equilibrium is without doubt the most widely spread solution concept in game theory. By definition we see that it corresponds exactly to the point $\mathbf{x}_\star$ at which $\mathrm{Gap}_{\mathcal{X}^i}^i(\mathbf{x}_\star) = 0$ for every $i \in \mathcal{N}$. In the following, we write $\mathfrak{X}_\star$ for the set of Nash equilibria of the game. The non-emptiness of $\mathfrak{X}_\star$ is guaranteed

2 In contrast, we made all constants in the regret bounds explicit in Part I and hence it is straightforward to derive a bound on $\mathrm{Reg}_T^i(\mathcal{Z}^i)$ from these bounds.

when $\mathfrak{X}$ is compact and the loss functions are continuous on $\mathfrak{X}$ and quasi-convex with respect to the players' own variables [62, 75, 96, 204].

5.1.2 *Adversarial Opponents and Self-Play*

In the model we just described, each player can exhibit diverse behaviors, which in turn results in interactions between players that can be arbitrarily complex. For the sake of analysis, we will mainly focus on the following two situations.

Adversarial opponents

- **Playing against adversarial opponents:** In this case, we aim to provide regret guarantee for a single player against arbitrary opponents. This brings us back to the online adversarial learning setup that we studied in Part I. The goal here is to show that the algorithm in question exhibits no regret as long as the feedback sequence is bounded.

Self-play

- **Self-play:** A more interesting scenario is when all the players are optimizing their own losses. In particular, we consider the situation when all the players adopt the same algorithm, or more generally, when all the players adopt an algorithm that satisfies a certain criteria.

In the latter case, the intricate interplay between no-regret and convergence to Nash equilibrium has attracted extensive attention in the literature; see e.g., [22, 82, 194, 273]. To shed light on this topic, we present below a proposition that summarizes some relations between regret, gap function, and Nash equilibrium.

Relation between regret, gap function, and Nash equilibrium

Proposition 5.1. *Assume that the players' loss functions are continuous. Then,*

(a) If the players stay at a Nash equilibrium, the players have no regret.

(b) If the the sequence of play $(\mathbf{x}_t)_{t\in\mathbb{N}}$ converges to a Nash equilibrium, then $\mathrm{Gap}^i_{\mathcal{Z}^i}(\mathbf{x}_t)$ converges to a non-positive number for any $i \in \mathcal{N}$ and compact set $\mathcal{Z}^i \subseteq \mathcal{X}^i$.

(c) If the sequence of play $(\mathbf{x}_t)_{t\in\mathbb{N}}$ converges and the players have no regret, then the point that $(\mathbf{x}_t)_{t\in\mathbb{N}}$ converges to must be a Nash equilibrium.

Proof. Point (a) is obvious. We prove the other two points below.

(b) It is known that $(\mathbf{x}_t)_{t\in\mathbb{N}}$ converges to a Nash equilibrium $\mathbf{x}_\star$. Moreover, the continuity of $\mathrm{Gap}^i_{\mathcal{Z}^i}$ is ensured by Berge's maximum theorem provided that $\mathcal{Z}^i$ is compact. It follows immediately that $\lim_{t\to+\infty} \mathrm{Gap}^i_{\mathcal{Z}^i}(\mathbf{x}_t) = \mathrm{Gap}^i_{\mathcal{Z}^i}(\mathbf{x}_\star) \leq 0$.

(c) Let us write $\mathbf{x}_\infty$ for the limit of $(\mathbf{x}_t)_{t\in\mathbb{N}}$. By the no-regret assumption, for all $i \in \mathcal{N}$ and $z^i \in \mathcal{X}^i$, we have $\sum_{t=1}^T \ell^i(\mathbf{x}_t) - \ell^i(z^i, \mathbf{x}_t^{-i}) = o(T)$, and accordingly $\liminf_{t\to+\infty} \ell^i(\mathbf{x}_t) - \ell^i(z^i, \mathbf{x}_t^{-i}) \leq 0$. On the other hand, by continuity of ℓ^i, we have

$$\lim_{t\to+\infty} \ell^i(\mathbf{x}_t) - \ell^i(z^i, \mathbf{x}_t^{-i}) = \ell^i(\mathbf{x}_\infty) - \ell^i(x^i_t, \mathbf{x}_\infty^{-i}).$$

In consequence, $\ell^i(\mathbf{x}_\infty) - \ell^i(z^i, \mathbf{x}_\infty^{-i}) \leq 0$ for all $z^i \in \mathcal{X}^i$ and this is true for all $i \in \mathcal{N}$. This shows that $\mathbf{x}_\infty$ is a Nash equilibrium by definition. □

No-regret does not imply convergence

Despite the positive results stated in Proposition 5.1, it is important to keep in mind that no-regret algorithms do no converge in general, and non-convergent examples are prevalent. This ranges from the appearance of cycles [56, 82, 195] to transition into chaos [44, 218]. More fundamentally, Hart and Mas-Colell [111] showed that no uncoupled dynamics is guaranteed to converge to a Nash equilibrium in all games. This aligns with the various negative complexity results for finding a Nash equilibrium [54, 58, 237].

In the light of the above, in this thesis we mostly restrict our attention to games for which more favorable guarantees can be derived.

5.1.3 *Assumptions on the Underlying Game*

Now that we have formulated the model in question, we move on to discussing the key assumptions that underpin our analysis.

Convexity

Assumption 5.1 (Convexity)**.** For all $i \in \mathcal{N}$, $\ell^i(\cdot, \mathbf{x}^{-i})$ is convex at all $\mathbf{x}^{-i}$.

As discussed in Part I, the convexity requirement in Assumption 5.1 is of paramount importance in the online learning literature, as it enables us to transform iterative gradient bounds to bona fide regret guarantees (Lemma 2.1). Moreover, convexity also plays an essential role in game theory provided that it contributes to the establishment of generalizations of the minimax theorem [62, 235] and simplifies the computation of equilibria.

Nonetheless, the convexity assumption alone is not sufficient for establishing the equilibrium convergence results. The algorithms that we study in this part make use of the slow-variation property of the feedback sequence that we explored in Chapter 4. We thus make the following regularity assumption.

Smoothness

Assumption 5.2 (Smoothness)**.** For all $i \in \mathcal{N}$, $\ell^i(\cdot, \mathbf{x}^{-i})$ is differentiable at all $\mathbf{x}^{-i}$ and the individual gradient of each player $V^i = \nabla_{x^i} \ell^i$ is L-Lipschitz continuous with respect to the Euclidean distance.

Assumption 5.2 is helpful for deriving improved regret bound, that is, the bound that grows slower than the standard $O(\sqrt{T})$ rate. As for the convergence result per se, we also rely on an additional stability assumption.

Variational stability

Definition 5.4 (Variational Stability)**.** A continuous game is *variationally stable* if the set $\mathfrak{X}_\star$ of Nash equilibria of the game is nonempty, the individual gradient field $V^i = \nabla_{x^i} \ell^i$ is well-defined for all $i \in \mathcal{N}$, and

$$\langle \mathbf{V}(\mathbf{x}), \mathbf{x} - \mathbf{x}_\star \rangle = \sum_{i=1}^{N} \langle V^i(\mathbf{x}), x^i - x^i_\star \rangle \geq 0 \quad \text{for all } \mathbf{x} \in \mathfrak{X}, \mathbf{x}_\star \in \mathfrak{X}_\star.$$

The game is *strictly variationally stable* if the above inequality holds as a strict inequality whenever $\mathbf{x} \notin \mathfrak{X}_\star$.

Assumption 5.3 (Variational Stability)**.** The underlying game is variationally stable.

Variational stability can be seen as a variant of the convexity assumption for multi-agent environments, where unilateral convexity assumptions do not suffice to give rise to a learnable game—for example, finite games are unilaterally linear, but finding a Nash equilibrium of a finite game is a PPAD-complete problem [54]. Our equilibrium convergence analyses would thus be focused on games that satisfy the variational stability condition. Some important families of games that are covered by this criterion are monotone games (i.e., $\mathbf{V}$ is monotone), which in their turn include convex-concave zero-sum games, zero-sum polymatrix games, and Cournot oligopolies. For the sake concreteness, let us provide two detailed examples below.

Zero-sum polymatrix games

Example 5.1 (Zero-sum polymatrix games [34, 124])**.** Polymatrix games are a class of finite games in which the loss of each player is determined by their pairwise interactions between other players. Each interaction is represented as an a bimatrix game, and the total loss for each player is the sum of the losses from these bimatrix games. Formally, for a mixed profile $\mathbf{x}$, we have

$$\ell^i(\mathbf{x}) = \sum_{j \neq i} (x^i)^\top A^{ij} x^j,$$

where A^{ij} is the matrix that dictates the loss of player i in the interaction with player j. A special case of polymatrix games is zero-sum polymatrix games. In these games, each interaction corresponds to a zero-sum bimatrix game, that is, $A^{ij} + A^{ji} = 0$. These games, when considered as continuous games over the mixed profiles, are monotone games.

Kelly auctions

Example 5.2 (Kelly auctions [151])**.** Consider an auction of K splittable resources among N bidders (players). For the k-th resource, let q_k and c_k denote respectively its available quantity and the entry barrier for bidding on it; for the i-th bidder, let b^i and r^i denote respectively the bidder's budget and marginal gain from obtaining a unit of resources. During play, each bidder submits a bid x_k^i for each resource k with the constraint $\sum_{k=1}^K x_k^i \leq b^i$. Resources are then allocated to bidders proportionally to their bids, so the i-th player gets $\rho_k^i = q_k x_k^i / (c_k + \sum_{i=1}^N x_k^i)$ units of the k-th resource. The utility of player $i \in \mathcal{N}$ is given by $u^i(\mathbf{x}) = \sum_{k=1}^K (r^i \rho_k^i - x_k^i)$, and the loss function is $\ell^i = -u^i$. This game is monotone as shown by Bravo et al. [27].

Beyond monotonicity, variational stability is still satisfied when $\mathbf{V}$ is only pseudo-monotone in the sense of Karamardian [147]—that is, for all $\mathbf{x}, \mathbf{x}' \in \mathfrak{X}$, we have

$$\langle \mathbf{V}(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle \geq 0 \quad \text{implies} \quad \langle \mathbf{V}(\mathbf{x}'), \mathbf{x}' - \mathbf{x} \rangle \geq 0.$$

As for an example that is not even pseudo-monotone but verifies the variational stability assumption, we refer the readers to [196].

5.2 VARIATIONAL INEQUALITIES

In this section, we delve into the concept of *variational inequality* (VI), which provides a unifying framework for understanding learning in convex games and extends its applicability to a variety of other domains. By drawing a connection between learning in games and VIs, we aim to demonstrate the broader relevance and versatility of the techniques and insights obtained from our study of online learning in convex games.

5.2.1 *Problem Formulation*

We start by introducing the VI problem. For that, we consider $\mathcal{X}$ a nonempty closed convex subset of $\mathbb{R}^d$, and $V : \mathcal{X} \to \mathbb{R}^d$ a single-valued operator on $\mathcal{X}$.

Stampacchia variational inequality

Definition 5.5 (Variational inequality)**.** The *variational inequality* (VI) problem associated to operator V and feasible set $\mathcal{X}$ is stated as:

$$\text{Find } x_\star \in \mathcal{X} \text{ such that } \langle V(x_\star), x - x_\star \rangle \geq 0 \text{ for all } x \in \mathcal{X}. \tag{VI}$$

A schematic representation of the condition in (VI) is presented in Fig. 5.2. In words, we aim to find $x_\star$ such that the opposite of $V(x_\star)$ lies in $\mathrm{NC}_{\mathcal{X}}(x_\star)$, the normal cone of $\mathcal{X}$ at $x_\star$. Mathematically, this can also be written as finding $x_\star \in \mathcal{X}$ such that

$$0 \in V(x_\star) + \mathrm{NC}_{\mathcal{X}}(x_\star) := \{V(x_\star) + v : v \in \mathrm{NC}_{\mathcal{X}}(x_\star)\}. \tag{5.1}$$

The above equation reformulates (VI) as an inclusion problem with set-valued mapping $V + \mathrm{NC}_{\mathcal{X}}$. In a similar vein, (VI) is also equivalent to the fixed-point problem of finding $x_\star$ such that $\Pi_{\mathcal{X}}(x_\star - V(x_\star)) = x_\star$. We summarize the above discussion in the proposition below.

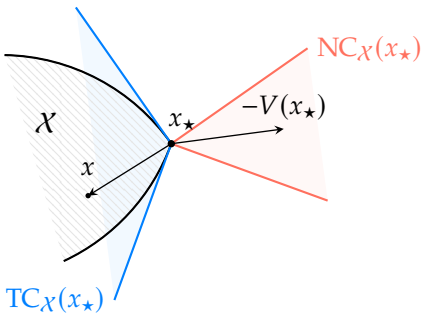


Figure 5.2: Schematic representation of the (VI) problem (adapted from [191]). $\mathrm{TC}_{\mathcal{X}}(x_\star)$ and $\mathrm{NC}_{\mathcal{X}}(x_\star)$ are respectively the tangent and the normal cones of $\mathcal{X}$ at $x_\star$.

Variational inequality, inclusion problem, fixed-point problem

Proposition 5.2 (Facchinei and Pang [73])**.** *The following three conditions on $x_\star \in \mathcal{X}$ equivalently define* (VI).

(i) $\langle V(x_\star), x - x_\star \rangle \geq 0$ *for all* $x \in \mathcal{X}$.

(ii) $0 \in V(x_\star) + \mathrm{NC}_{\mathcal{X}}(x_\star)$.

(iii) $\Pi_{\mathcal{X}}(x_\star - V(x_\star)) = x_\star$.

In the literature, the VI formulation that we just described is sometimes called *Stampacchia variational inequality* (SVI) [256], distinguishing it from what is referred to as *Minty variational inequality* (MVI) [94].

Minty variational inequality

Definition 5.6 (Minty variational inequality)**.** The *Minty variational inequality* (MVI) problem associated to operator V and feasible set $\mathcal{X}$ is stated as:

$$\text{Find } x_\star \in \mathcal{X} \text{ such that } \langle V(x), x - x_\star \rangle \geq 0 \text{ for all } x \in \mathcal{X}. \tag{MVI}$$

Compared to (VI), the operator V is now evaluated at any point $x \in \mathcal{X}$, rather than just at $x_\star$. Therefore, (VI) expresses a local condition on V whereas (MVI) deals with all function values on $\mathcal{X}$. The names *strong* and *weak* solutions are also occasionally used to indicate respectively the solutions of SVI and MVI. Importantly, they are related to each other by the well-known Minty Lemma.

Minty Lemma

Proposition 5.3 (Minty [197, Lem. 1])**.** *The following statements hold.*

(a) If V is pseudo-monotone, any solution of (VI) *is a solution of* (MVI).

(b) If V is continuous, any solution of (MVI) *is a solution of* (VI).

Beyond their use in optimization and in game theory as we will discussed in Section 5.2.3, VIs have also been fruitfully applied in various other domains, including obstacle problems [154] and contact mechanics [37]. For a comprehensive treatment of the subject, we refer the readers to the textbooks by Kinderlehrer and Stampacchia [153] and by Facchinei and Pang [73].

5.2.2 *Merit Functions*

Numerous merit functions have been proposed in the literature to measure the optimality of a point for a VI problem. Below we introduce some of the most relevant ones within our context. We continue to use the notations $\mathcal{X}$ for the feasible set and V for the operator in question.

Dual gap function

Definition 5.7 (Dual Gap)**.** Let $\mathcal{Z} \subseteq \mathcal{X}$ be a compact subset of $\mathcal{X}$. The (restricted) dual gap function is defined as

$$\mathrm{DGap}_{\mathcal{Z}}(x) \coloneqq \max_{z \in \mathcal{Z}} \langle V(z), x - z \rangle.$$

We have the following proposition by Nesterov [211].

Proposition 5.4 (Nesterov [211, Lem. 1])**.** *The function* DGap *is well-defined and convex. Assume additionally that V is continuous and monotone. Then*

(a) It holds $\mathrm{DGap}_{\mathcal{Z}}(x) \geq 0$ *for all* $x \in \mathcal{Z}$.

(b) If $x_\star \in \mathcal{Z}$ *is a solution of* (VI)*, then* $\mathrm{DGap}_{\mathcal{Z}}(x_\star) = 0$.

(c) If there exists $\hat{x}, x' \in \mathcal{X}$ *and* $\rho > 0$ *such that* $\mathrm{DGap}_{\mathcal{Z}}(\hat{x}) = 0$, $\mathcal{X} \cap \mathscr{B}(x', R) \subseteq \mathcal{Z}$, *and* $\|\hat{x} - x'\| < R$, *then* $\hat{x}$ *is a solution of* (VI).

In the light of Proposition 5.4, the dual gap function is frequently used to access the quality of a candidate solution when V is monotone [142, 211, 212]. In fact, the definition of DGap uses the (MVI) formulation and we effectively utilize the fact that (VI) and (MVI) are equivalent for continuous and monotone operator V in the proof of Proposition 5.4. Nonetheless, when V is non-monotone, it may be more appropriate to consider the primal gap function.

Primal gap function

Definition 5.8 (Primal Gap)**.** Let $\mathcal{Z} \subseteq \mathcal{X}$ be a compact subset of $\mathcal{X}$. The (restricted) primal gap function is defined as

$$\mathrm{PGap}_{\mathcal{Z}}(x) \coloneqq \max_{z \in \mathcal{Z}} \langle V(x), x - z \rangle.$$

Remark 5.2. The appearance of the unrestricted variants of the primal and the dual gap functions dates back to at least the work of Zukhovitskii, Polyak, and Primak [299, 300]. Their use have since been widely adopted in the VI literature [117, 167]. However, the name *gap function* may be used to refer to either of the two or other variants thereof. For clarity, we hence borrow the names *primal* and *dual* gaps from Facchinei and Pang [73], despite the fact that the dual gap function of [73] is actually the opposite of what we define in Definition 5.7.

Compared to the dual gap function, the primal gap function is no longer convex. It otherwise preserves other properties that we presented Proposition 5.4 without requiring the monotonicity of V. It can also be immediately seen that PGap is always larger than DGap when V is monotone.

On the downside, the definition of the (restricted) gap functions necessarily involves a compact set $\mathcal{Z}$. Selecting an appropriate compact set can be challenging when dealing with unbounded $\mathcal{X}$. To overcome this issue, there are several other alternatives that we can consider.

Natural residual

Definition 5.9 (Natural Residual)**.** Let $\eta > 0$ be a positive real number. The natural residual associated to η is defined as

$$\mathrm{Res}^{\mathrm{nat}}_{\eta}(x) \coloneqq \frac{1}{\eta} \|x - \Pi_{\mathcal{X}}(x - \eta V(x))\|_2.$$

Tangent residual

Definition 5.10 (Tangent Residual)**.** The tangent residual is defined as

$$\mathrm{Res}^{\tan}_{\eta}(x) \coloneqq \min\{\|V(x_\star) + v\|_2 : v \in \mathrm{NC}_{\mathcal{X}}(x_\star)\}.$$

The natural and the tangent residuals match respectively the fixed-point and the inclusion problem perspectives that we present in Proposition 5.2. Moreover, the two coincide and equal to the operator norm $\|V(x)\|_2$ when $\mathcal{X}$ is the entire space $\mathbb{R}^d$. For the general case, they are related by the following proposition.

Natural residual vs. tangent residual

Proposition 5.5 (Facchinei and Pang [73, Prop. 1.5.14])**.** *For any $\eta > 0$ and $x \in \mathcal{X}$, it holds*

$$\mathrm{Res}^{\mathrm{nat}}_{\eta}(x) \leq \mathrm{Res}^{\tan}(x).$$

In a similar spirit, the primal gap function is also related to the tangent residual.

Primal gap function vs. tangent residual

Proposition 5.6 (Cai et al. [35, Lem. 2])**.** *Let $x \in \mathcal{X}$, $\mathcal{Z} \subseteq \mathcal{X}$ be a compact set, and $R \coloneqq \max_{z \in \mathcal{Z}} \|x - z\|_2$. It holds*

$$\mathrm{PGap}_{\mathcal{Z}}(x) \leq R\, \mathrm{Res}^{\tan}(x).$$

From the above two propositions, it is clear that providing an upper bound on the tangent residual is the most favorable. Finally, the distance to the solution set $\mathrm{dist}(x, \mathcal{X}_\star)$, where $\mathcal{X}_\star$ is the solutions of (VI), is probably the most natural error metric that one can think of. When the problem is unconstrained (i.e., $\mathcal{X} = \mathbb{R}^d$) and V is $L-$Lipschitz continuous (i.e., $\|V(x) - V(x')\|_* \leq L\|x - x'\|$ for all x, x'), we have clearly $L\,\mathrm{dist}(x, \mathfrak{X}_\star) \geq \|V(x)\|_*$.

5.2.3 *Variational Inequalities and Learning in Games*

Throughout Part II of the thesis, we focus on the learning-in-games framework and present results within this context. However, our analysis can often be adapted to cope with VI problems and to provide bounds on the merit functions introduced in Section 5.2.2. Indeed, a significant portion of our analysis concentrates solely on the joint vector field, or the pseudo gradient, $\mathbf{V} = (V^i)_{i \in \mathcal{N}}$, of the players. This is possible thanks to the following first-order characterization of Nash equilibrium in convex games.

First-order characterization of Nash equilibrium

Proposition 5.7. *Assume Assumption 5.1 holds and each loss function ℓ^i is differentiable with respect to x^i_t. Then, a point $\mathbf{x}_\star$ is a Nash equilibrium if and only if it is a solution of the VI problem associated to operator $\mathbf{V}$ and feasible set $\mathfrak{X}$.*

Proposition 5.7 is an immediate consequence of the individual convexity of the loss functions. With the proposition in mind, to prove convergence to Nash equilibrium we often start by showing that any cluster point $\mathbf{x}_\infty$ of the played sequence $(\mathbf{x}_t)_{t \in \mathbb{N}}$ solves the VI problem $\forall \mathbf{x} \in \mathfrak{X}, \langle \mathbf{V}(\mathbf{x}_\infty), \mathbf{x} - \mathbf{x}_\infty \rangle \geq 0$.

Linearized regret

Regarding regret analysis, as in Part I we rely on the use of Lemma 2.1 and will systematically establish regret bounds for the linearized regret

$$\mathrm{LinReg}^i_T(\mathcal{Z}^i) = \max_{z^i \in \mathcal{Z}^i} \sum_{t=1}^{T} \langle V^i(\mathbf{x}_t), x^i_t - z^i \rangle.$$

From linearized regret to dual gap function

When $\mathbf{V}$ is monotone, this quantity is related to the dual gap function introduced earlier by the following inequality.

$$\frac{1}{T}\sum_{t=1}^{T} \langle \mathbf{V}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{z} \rangle \geq \frac{1}{T}\sum_{t=1}^{T} \langle \mathbf{V}(\mathbf{z}), \mathbf{x}_t - \mathbf{z} \rangle = \left\langle \mathbf{V}(\mathbf{z}), \frac{\sum_{t=1}^{T} \mathbf{x}_t}{T} - \mathbf{z} \right\rangle.$$

As a consequence, any upper bound on the sum of the players' linearized regrets can be directly translated into an upper bound on the dual gap function evaluated at the average iterate $\sum_{t=1}^{T} \mathbf{x}_t/T$. This provides yet another example on how our analysis can be relevant for VI problems beyond learning in games.

5.3 ALGORITHMS

In terms of algorithms, we examine the algorithms that are *optimistic*. These algorithms have several advantages over vanilla MD in learning in games, as we shall see below. For ease of notation, we omit the player index when describing the algorithms. That is, we consider a player with action set $\mathcal{X}$ who receives a sequence of feedback $(g_t)_{t\in\mathbb{N}}$.

5.3.1 *Optimistic Mirror Descent*

Optimistic mirror descent

To begin, we state the general *optimistic mirror descent* (OptMD) template that generalizes the OG method to account for the geometry induced by a regularizer h (the regularizer and Bregman divergence are defined as in Chapter 2). We recall that the player plays $x_t = X_{t+\frac{1}{2}}$ at round t. For a given learning rate sequence $(\eta_t)_{t\in\mathbb{N}}$, the algorithm writes

$$\begin{aligned} X_{t+\frac{1}{2}} &= \arg\min_{x\in\mathcal{X}} \langle \tilde{g}_t, x\rangle + \frac{D(x, X_t)}{\eta_t}, \\ X_{t+1} &= \arg\min_{x\in\mathcal{X}} \langle g_t, x\rangle + \frac{D(x, X_t)}{\eta_{t+1}}. \end{aligned} \tag{OptMD}$$

In Chapter 4, we have seen how the introduction of the optimistic step (the step that moves from X_t to $X_{t+\frac{1}{2}}$) helps achieve smaller regret when the optimistic guess $\tilde{g}_t$ is close to g_t. In fact, the regret bound of Proposition 4.1 also holds for (OptMD)—we just need to replace $\|z - X_1\|_2^2/2$ by $D(z, X_1)$ and use the corresponding dual norm to measure the variations.

Optimistic mirror descent with past feedback

As also suggested in Chapter 4, one natural choice is then to take $\tilde{g}_t$ to be the previous feedback received by the player, which is equivalent to $\tilde{g}_t = g_{t-1}$ under our current setup (we use the notation $g_0 = g_0^i = 0$ throughout the thesis). With this choice, low regret is guaranteed whenever the second-order path length $\sum_t \|g_t - g_{t-1}\|_\star^2$ is small. This may be notably the case when all the players take small move to optimize their losses. Formally, it has been shown in the literature that players enjoy poly-logarithmic or even constant regret when all of them employ (some instantiation of) (OptMD) with $\tilde{g}_t = g_{t-1}$ in certain classes of games [2, 3, 57, 77, 228].

Optimistic mirror descent for VI, mirror-prox, and extra-gradient

Equally of interest is the use of the (OptMD) algorithm for saddle-point problems, or for VIs in general. For example, we recover the *mirror-prox* (MP) algorithm of Nemirovski [207] if all the players adopt (OptMD) with $\tilde{g}_t = V^i(\mathbf{X}_t)$ (note that $\mathbf{X}_t$ is effectively defined in this case). Tracing back even further, the *extra-gradient* (EG) algorithm introduced in the pioneer work of Korpelevich [160] is essentially MP with quadratic regularizer. A large corpus of works has been dedicated to the study of EG and MP in the context of VI in the past few decades. This gives rise to a plethora of results that illuminate the convergence behavior of the algorithm under various assumptions [35, 73, 142, 196, 267].

Mirror-prox is not a valid algorithm for online learning

Nonetheless, it is worth noticing that the update of MP cannot be realized within our interaction model since it requires the player to have gradient feedback at an additional point $\mathbf{X}_t$. Although this can be partially solved by

reindexing the rounds and considering that a player plays both iterates X_t and $X_{t+\frac{1}{2}}$ and receives feedback $\tilde{g}_t$ and g_t, the resulting algorithm is no longer no-regret in the sense that we can find a game and a sequence of opponents' actions such that the regret evaluated at the played actions $(X_s)_{s\in\mathbb{N}/2}$ grows linearly while the feedback is bounded (see [99] for more details).

In contrast, the choice $\tilde{g}_t = g_{t-1}$ suggested earlier is indeed legitimate and leads to a no-regret algorithm. The use of this method can be traced to the work of Popov [225] where they demonstrated that the algorithm in question with quadratic regularizer converges to a solution in saddle-point problems. To summarize, we list below several important properties of the (OptMD) algorithm with $\tilde{g}_t = g_{t-1}$ that are particularly relevant in the scope of this thesis.

Advantages of optimistic mirror descent

Proposition 5.8. *Suppose Assumption 5.1 holds and that all players have access to perfect feedback $g_t^i = V^i(\mathbf{x}_t)$. The following statements hold true.*

(a) Let player i follow (OptMD) with $\tilde{g}_t^i = g_{t-1}^i$ and appropriate learning rate against a sequence of bounded feedback. Then, for any bounded set $\mathcal{Z}^i \subseteq \mathcal{X}^i$, we have

$$\mathrm{Reg}_T^i(\mathcal{Z}^i) = O(\sqrt{T}).$$

(b) Let Assumption 5.2 and Assumption 5.3 hold and all players follow (OptMD) with $\tilde{g}_t^i = g_{t-1}^i$ and appropriate learning rate. Then, for all players $i \in \mathcal{N}$ and bounded set $\mathcal{Z}^i \subseteq \mathcal{X}^i$, we have

$$\mathrm{Reg}_T^i(\mathcal{Z}^i) = O(1).$$

(c) Consider the situation described in (b). Under suitable "reciprocity" condition (see Section 6.4), the sequence of play converges to a Nash equilibrium.

Proof. We refer the readers to [125, 225, 228] for a series of related result. The proposition can otherwise be proved by modifying the analysis in Chapter 6 of adaptive methods. □

Given the desirable properties of (OptMD) outlined in Proposition 5.8, our primary objective in the subsequent sections is to extend these guarantees to a wider range of situations. In particular, we aim to address cases where the player has no information about the underlying game and the interaction paradigm with the help of adaptive learning rates (Chapter 6), as well as cases where feedback is corrupted by noise with the help of learning rate separation (Chapters 7 and 8). It is worth noting that neither *(b)* nor *(c)* can be achieved by vanilla MD. In fact, even in bilinear zero-sum games (which apparently satisfies Assumptions 5.2 and 5.3), the $\Omega(\sqrt{T})$ regret lower bound for MD has been demonstrated by Chen and Peng [43], and non-convergence of the algorithm is also well documented in the literature [56, 196].

5.3.2 *Optimistic Dual Averaging*

Optimistic dual averaging

As illustrated in Proposition 4.2, (OptMD) with dynamic learning rate can yield linear or even superlinear regret when the associated Bregman divergence is unbounded on the action set. To address this issue, we will also consider the *optimistic dual averaging* (OptDA) method.

$$\begin{aligned} X_{t+\frac{1}{2}} &= \arg\min_{x\in\mathcal{X}} \langle \tilde{g}_t, x\rangle + \frac{D(x, X_t)}{\eta_t}, \\ X_{t+1} &= \arg\min_{x\in\mathcal{X}} \sum_{s=1}^{t} \langle g_s, x\rangle + \frac{h(x)}{\eta_{t+1}}. \end{aligned} \tag{OptDA}$$

The algorithm outlined above notably encompasses dual extrapolation (DE) proposed by Nesterov [211] as a special case. This corresponds to the situation where all players adopt (OptDA) with $\tilde{g}_t = V^i(\mathbf{X}_t)$. However, akin to MP, DE is not a valid algorithm for online learning in games. It is also clear that (OptDA) is closely related to the (DOptDA) algorithm studied in Chapter 4. In fact, the latter operates with delayed feedback but is stated for the special case $\mathcal{X} = \mathbb{R}^d$ and $h(\cdot) = \|\cdot\|_2^2/2$. Finally, with the choice $\tilde{g}_t = g_{t-1}$, we obtain the algorithm that was independently introduced by Song et al. [255] under the name of optimistic dual extrapolation. This algorithm, for which a version of Proposition 5.8 holds, will play an important role in the remaining of this thesis.

A NOTE ON TERMINOLOGY. In existing literature, the terms *optimistic mirror descent* and *optimistic dual averaging* typically refer to the special cases where $\tilde{g}_t = g_{t-1}$. However, in the scope of thesis, we instead use them to represent the general templates. Generally speaking, we will rely on the associated equations to differentiate between algorithms, thus avoiding any potential confusion.

5.3.3 *Optimistic Gradient as Approximate Projection onto Separating Hyperplane*

Numerous attempts have been made in the literature to explain the success of optimistic gradient methods. These include for example approaches that conceptualize the optimization process as a discretization of its continuous-time counterpart [182], interpretations grounded in operator theory that regard these algorithms as approximations of proximal gradient [199], and the online learning with guess vector perspective that we adopted in Chapter 4.

In this subsection, we take yet another viewpoint, one that is well documented in the monograph [73]. This geometric interpretation treats each iteration of the algorithm as an approximate projection onto a separating hyperplane and is illustrated in Fig. 5.3. To explain it in detail, we consider the unconstrained Euclidean setup, in which case the update (OptMD) is written as

Update in the unconstrained Euclidean setup

$$X_{t+\frac{1}{2}} = X_t - \eta_t \tilde{g}_t, \quad X_{t+1} = X_t - \eta_{t+1} g_t.$$

We further look into the case where all the player adhere to the above algorithm, use the same learning rates, and receive perfect feedback $g_t = V^i(\mathbf{x}_t)$. This leads to the following global update rule (recall that $\mathbf{x}_t = \mathbf{X}_{t+\frac{1}{2}}$).

$$\mathbf{X}_{t+\frac{1}{2}} = \mathbf{X}_t - \eta_t \widetilde{\mathbf{g}}_t, \quad \mathbf{X}_{t+1} = \mathbf{X}_t - \eta_{t+1} \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}).$$

Separating hyperplane

Let us write $d = \sum_{i\in\mathcal{N}} d^i$ for the sum of the individual dimensions so that $\mathfrak{X} \subseteq \mathbb{R}^d$. We consider the hyperplane

$$\mathscr{H}_t = \{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}\rangle = 0\}.$$

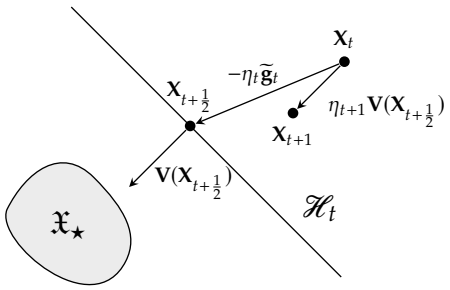


Figure 5.3: Illustration of the "separate and project" principle of the optimistic algorithms.

On one hand, when the game is variationally stable (Assumption 5.3), we have by definition that for all $\mathbf{x}_\star \in \mathfrak{X}_\star$,

$$\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star \rangle \geq 0.$$

On the other hand, if $\widetilde{\mathbf{g}}_t$ is effectively close to $\mathbf{g}_t$, we can expect

$$\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{X}_t \rangle = -\eta_{t+1}\langle \mathbf{g}_t, \widetilde{\mathbf{g}}_t \rangle \leq 0.$$

This is the easiest to illustrate through the choice $\widetilde{\mathbf{g}}_t = \mathbf{V}(\mathbf{X}_t)$ of EG. Then, if the pseudo-gradient is $\sqrt{N}L$-Lipschitz (Assumption 5.2) and the learning rate η_t is sufficiently small, it holds that

$$\begin{aligned}
\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{X}_t \rangle &= -\eta_t \langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{V}(\mathbf{X}_t) \rangle \\
&= \frac{\eta_t}{2}\left(\|\mathbf{V}(\mathbf{X}_t) - \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 - \|\mathbf{V}(\mathbf{X}_t)\|^2 - \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 \right) \\
&\leq \frac{\eta_t}{2}\left((\eta_t^2 NL^2 - 1)\|\mathbf{V}(\mathbf{X}_t)\|^2 - \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 \right) \\
&\leq -\frac{\eta_t}{2}\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 \\
&\leq 0. \qquad (5.2)
\end{aligned}$$

In summary, the hyperplane $\mathscr{H}_t$ effectively separates the solution set $\mathfrak{X}_\star$ from the current iterate $\mathbf{X}_t$. The update step $\mathbf{X}_{t+1} = \mathbf{X}_t - \eta_{t+1}\mathbf{V}(\mathbf{x}_t)$ moves $\mathbf{X}_t$ in the direction of the projection onto this hyperplane, and thereby moves it closer to the solutions.

Projection onto the the hyperplane

Going further, we can compute the projection of $\mathbf{X}_t$ onto the hyperplane.

$$\Pi_{\mathscr{H}_t}(\mathbf{X}_t) = \mathbf{X}_t - \eta_t \frac{\langle \mathbf{g}_t, \widetilde{\mathbf{g}}_t \rangle}{\|\mathbf{g}_t\|^2}\mathbf{g}_t$$

This shows that $\mathbf{X}_{t+\frac{1}{2}}$ is indeed "approximately" a projection onto the hyperplane when $\mathbf{g}_t \approx \widetilde{\mathbf{g}}_t$ and $\eta_t \approx \eta_{t+1}$. From the derivation of (5.2), we particularly know that for the case $\widetilde{\mathbf{g}}_t = \mathbf{V}(\mathbf{X}_t)$ we get $1/2 \leq \langle \mathbf{g}_t, \widetilde{\mathbf{g}}_t \rangle / \|\mathbf{g}_t\|^2$, so that if $\eta_{t+1} \leq \eta_t$ then the iterate is necessarily closer to $\mathscr{H}_t$ after the update.

As we conclude this discussion, it is worth noting that this perspective offers not just a conceptual understanding of the algorithm's mechanism, but also insightful cues for possible algorithmic modifications, as we will illustrate in Chapter 7. Moreover, this "separate and project" principle is not unique to optimistic methods. It has been effectively utilized for solving variational inequality and monotone inclusion problems, leading to a suite of algorithms

that directly employ this concept. For further examples and applications of this principle, we refer the reader to works such as [97, 116, 158, 159, 254].

6

LEARNING RATE ADAPTATION FOR GAMES WITH PERFECT FEEDBACK

This chapter incorporates material from Hsieh et al. [127]

Optimistic methods have proven to be valuable tools in online learning in games, as discussed in Section 5.3. However, their performance are significantly affected by the precise choice of learning rates (see Fig. 6.1). In light of this challenge, we turn our attention to adaptive learning rates in this chapter.

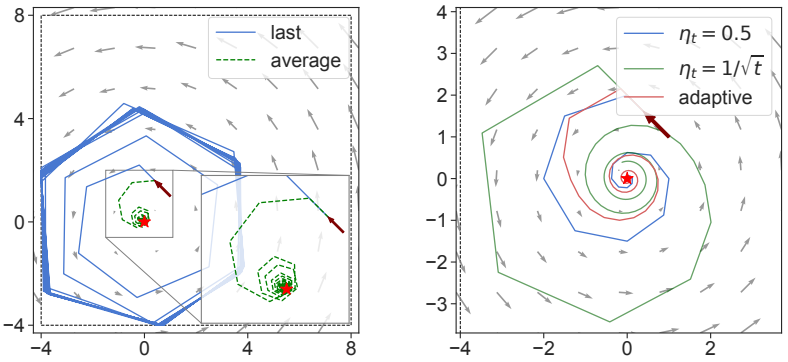
The study of adaptive learning rates in this context hold both practical and theoretical merit. On a practical level, it paves the way for the implementation of learning algorithms even when prior knowledge about the game's parameters is not available, fostering a broader application scope. Theoretically, it allows for a more nuanced understanding of the crucial elements required for an algorithm to achieve specific guarantees.

Indeed, focusing on scenarios where all players have access to perfect feedback $g_t^i = V^i(\mathbf{x}_t)$ throughout the chapter, we demonstrate that the local gradient feedback made available to the players is sufficient for achieving optimal regret guarantees and convergence to a relevant profile in various contexts. Importantly, by "sufficient" we mean that the players perform the entire learning algorithm, including the computation of their learning rates, at an individual level, without the need for prior knowledge about the game.

CONTRIBUTIONS AND OUTLINE. Expanding on the above discussion, the primary contribution of this chapter is the development of a range of adaptive optimistic policies with the following desirable properties (suppose that Assumptions 5.1 and 5.2 hold):

1. They do not require *any* prior tuning or knowledge of the game's parameters. Each player updates their individual step-size with purely local, individual gradient information. *No need of knowledge about game parameter*
2. They guarantee an order-optimal $O(\sqrt{T})$ regret bound against adversarial play, and they further enjoy *constant* social regret (i.e., the sum of the players' regrets) when all players employ one of these algorithms. *Optimal regret*
3. They guarantee convergence to to best response against convergent opponents if the action set of the focal player is compact. *Convergence to best response*
4. If all players follow one of these algorithms, the induced trajectory of play converges to a Nash equilibrium and the regret of each player is bounded as $O(1)$ in all variationally stable games. *Convergence to Nash equilibrium*

In terms of organization, we start by introducing our adaptive learning rates in Section 6.1. Following this, Section 6.2 details the optimistic algorithms that can successfully leverage these adaptive learning rates. Regret bounds and convergence analysis are respectively presented in Section 6.3 and Section 6.4. Finally, we close the chapter with some numerical illustrations in Section 6.5.



(a) (OptMD)/(OptDA) with constant learning rate $\eta_t \equiv 0.6$ causes played iterate to diverge and average iterate to converge to spurious point.

(b) Convergence of (OptMD)/(OptDA) with suitable learning rate. Adaptive method converges faster without requiring knowledge about the game.

Figure 6.1: The trajectories of play (and the time-average of one of these trajectories) obtained by running (OptMD)/(OptDA) with a quadratic regularizer on $\min_{\theta\in[-4,8]} \max_{\phi\in[-4,8]} \theta\phi$ using different learning rates (the trajectories of the two algorithms coincide on this example).

6.1 ADAPTIVE LEARNING RATE

The success of optimistic methods hinges critically on the careful tuning of learning rates. In particular, an optimistic algorithm could switch from convergent to non-convergent by a slight variation of its hyperparameters or a small perturbation of the game. We illustrate this via the following example.

Importance of learning rate tuning

Example 6.1 (Non-convergence of inappropriately tuned optimistic methods)**.** Consider the following bilinear zero-sum game with player variables θ and ϕ.

$$\mathcal{X}^1 = \mathcal{X}^2 = \mathbb{R}, \quad \ell^1(\theta,\phi) = \theta\phi = -\ell^2(\theta,\phi).$$

Suppose that both players run (OptMD) with quadratic regularizer $h^i(\cdot) = \|\cdot\|_2^2/2$ and constant learning rate $\eta_t^i \equiv \eta$ (note that (OptMD) and (OptDA) coincide in this case). Then, if $\eta < 1/\sqrt{3}$, the sequence of play converges to the game's unique Nash equilibrium at $(0,0)$. On the other hand, if the players misestimate the critical value $1/\sqrt{3}$ and choose $\eta \geq 1/\sqrt{3}$, the method no longer converges, in either the "ergodic" or "trajectory/last-iterate" sense (for a proof, see e.g., [289]). Moreover, as we show in Fig. 6.1a, this "off-equilibrium" behavior persists even if we restrict the players' actions to compact sets $\mathcal{X}^1 = \mathcal{X}^2 = [-4,8]$.

Adaptive learning rate

The above example elucidates the pitfalls of improperly chosen learning rates. Not only may the trajectory diverge, but its average could also converge to an irrelevant action profile, making such failure difficult to detect. In response to this predicament, we consider an adaptive policy in the spirit of Rakhlin and Sridharan [228], namely[1]

$$\eta_t^i = \frac{1}{\sqrt{1 + \sum_{s=1}^{t-1} \delta_s^i}} \quad \text{where} \quad \delta_t^i = \|g_t^i - g_{t-1}^i\|_{(i),*}^2. \tag{Adapt}$$

1 More generally speaking, all the results that we present in this chapter would still hold if we replace the 1 in the denominator of η_t^i by another positive constant τ^i and $\|\cdot\|_{(i),*}$ in the definition of δ_t^i by another norm in $\mathbb{R}^{d^i}$. Recall also that we have set $g_0^i = 0$.

In the above, $\|\cdot\|_{(i),*}$ is the dual norm of $\|\cdot\|_{(i)}$, which is itself the norm associated to player i's regularizer h^i. Intuitively, in the favorable case (e.g., when the environment is stationary), the increments δ_t^i eventually vanishes, so the policy (Adapt) will be a proxy for the "constant learning rate" case. By contrast, in a non-favorable / adversarial setting, we have $\delta_t^i = \Theta(1)$ and η_t^i decreases as $\Theta(1/\sqrt{t})$, which makes the algorithm robust.

We should also note here that (Adapt) involves *exclusively* player-specific quantities, and its computation only makes use of information that is available to each player *locally*. These methods thus sidestep the need for coordination between players or global knowledge about the game, which are often required by algorithms with aggressive predetermined learning rates, e.g., constant, or even by other adaptive methods [7, 177]. On the other hand, they achieve faster convergence compared to algorithms with more conservative learning rate schedule, e.g., $\eta_t = 1/\sqrt{t}$ (see Fig. 6.1b).

Importantly, although we also study adaptive learning rate in Part I, we have quite different objectives in the two setups. In Part I, our focus is to provide data-dependent regret bounds by dynamically adjusting the learning rate based on received feedback. Here, with a learning rate adjusted in a similar way, we can provide even stronger guarantees since the feedback comes from interaction with other players. Notably, we can show constant regret bound and convergence of learning trajectory in self-play, both of which are unattainable in the pure online learning setup. Furthermore, while we did look into adaptive learning rate of optimistic method in Section 4.3.2, the learning rate proposed there requires knowledge of various constants of the learning system. It is not the case anymore.

NOTATIONS. For the analysis in Sections 6.3 and 6.4, it will be convenient to write $\lambda_t^i := 1/\eta_t^i$ for the inverse of learning rate and define the norm on the joint action space as $\|(x^i)_{i\in\mathcal{N}}\| := \sqrt{\sum_{i=1}^N \|x^i\|_{(i)}^2}$.

6.2 A FAMILY OF OPTIMISTIC METHODS

A recurrent topic in this thesis is the compatibility of an online learning algorithm with dynamic learning rates. More precisely, in Chapters 2 and 4 we have explained that both MD and OptMD can incur linear or superlinear regret when run with dynamic learning rate sequence. To cope with this issue, we have introduced DA and OptDA. This section aims to broaden the scope of the considered algorithm and highlight the key property that an optimistic algorithm needs to possess to be used with our dynamic, and notably adaptive learning rate (Adapt).

For notational convenience, we omit the player index in this section's presentation, as our focus is on the algorithm taken by a single player.

6.2.1 *Compatibility with Dynamic Learning Rate*

The theoretical results of this chapter hold for a family of algorithms that we refer to as being *optimistic and compatible with dynamic learning rates*. We formally define this below.

Optimistic agorithms that are compatible with dynamic learning rate

Definition 6.1. A learning algorithm used by a player $i \in \mathcal{N}$ is said to be *optimistic and compatible with dynamic learning rate* if it operates with a regularizer

h and non-increasing positive learning rates $(\eta_t)_{t\in\mathbb{N}}$ to produce a sequence of iterates $(X_s)_{s\in\mathbb{N}/2}$ such that

1. At each round $t \in \mathbb{N}$, the player plays $x_t = X_{t+\frac{1}{2}}$ generated by

$$X_{t+\frac{1}{2}} = \operatorname*{arg\,min}_{x\in\mathcal{X}} \langle g_{t-1}, x\rangle + \frac{D(x, X_t)}{\eta_t}. \tag{6.1}$$

2. For some non-negative continuous functions $(\psi_t)_{t\in\mathbb{N}}$ and φ defined on $\mathcal{X}$, we have, for all $z \in \mathcal{X}$ and $t \in \mathbb{N}$,

$$\begin{aligned}\frac{\psi_{t+1}(z)}{\eta_{t+1}} \leq \frac{\psi_t(z)}{\eta_t} - \langle g_t, X_{t+\frac{1}{2}} - z\rangle + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\varphi(z)\\ + \langle g_t - g_{t-1}, X_{t+\frac{1}{2}} - X_{t+1}\rangle - \frac{D(X_{t+1}, X_{t+\frac{1}{2}})}{\eta_t} - \frac{D(X_{t+\frac{1}{2}}, X_t)}{\eta_t}.\end{aligned} \tag{6.2}$$

For brevity, in the following we will write $\mathfrak{A}_{ocd}$ for the family of such algorithms.

By replacing φ with $\max(\varphi, \psi_1)$ if needed, we may assume $\psi_1 \leq \varphi$ without loss of generality. The first point is common to the optimistic algorithms that we have considered so far and is important for our analysis on last-iterate convergence. In fact, with Eq. (6.1), $X_t = X_{t+\frac{1}{2}}$ implies that $\langle g_{t-1}, z - X_t\rangle \geq 0$ for all $z \in \mathcal{X}$. This thus shows that $\mathbf{X}_t$ is "almost" an equilibrium if $\mathbf{x}_{t-1} \approx \mathbf{X}_t$. We will exploit an argument of this kind in Section 6.4 (instead of $X_t = X_{t+\frac{1}{2}}$ we will only have $X_t \approx X_{t+\frac{1}{2}}$ for t large enough).

Regarding the second point, it relates the *energy* of round t to that of round $t+1$. It is thus used to show that this energy converges in convergence analysis of the learning dynamics. Moreover, summing up this inequality from $t = 1$ to T, we get directly an upper bound on the linearized regret as shown in the following lemma.

Preliminary bound on linearized regret

Lemma 6.1. *Let $(X_s)_{s\in\mathbb{N}/2}$ be the iterates produced by an algorithm of $\mathfrak{A}_{ocd}$ as described in Definition 6.1. Then*

$$\sum_{t=1}^{T}\langle g_t, X_{t+\frac{1}{2}} - z\rangle \leq \lambda_{T+1}\varphi(z) + \sum_{t=1}^{T}\frac{\|g_t - g_{t-1}\|_*^2}{\lambda_t} - \sum_{t=2}^{T}\frac{\lambda_{t-1}}{8}\|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2. \tag{6.3}$$

Proof. From (6.2), we get immediately

$$\begin{aligned}\sum_{t=1}^{T}\langle g_t, X_{t+\frac{1}{2}} - z\rangle &\leq \lambda_{T+1}\varphi(z) - \lambda_{T+1}\psi_{T+1}(z) + \sum_{t=1}^{T}\langle g_t - g_{t-1}, X_{t+\frac{1}{2}} - X_{t+1}\rangle\\ &\quad - \sum_{t=1}^{T}\lambda_t\left(D(X_{t+1}, X_{t+\frac{1}{2}}) + D(X_{t+\frac{1}{2}}, X_t)\right).\\ &= \lambda_{T+1}\varphi(z) - \lambda_{T+1}\psi_{T+1}(z)\\ &\quad - \lambda_1 D(X_{3/2}, X_1) - \frac{\lambda_T}{2}D(X_{T+1}, X_{T+\frac{1}{2}})\\ &\quad - \sum_{t=2}^{T}\left(\frac{\lambda_{t-1}}{2}D(X_t, X_{t-\frac{1}{2}}) + \lambda_t D(X_{t+\frac{1}{2}}, X_t)\right)\\ &\quad + \sum_{t=1}^{T}\left(\langle g_t - g_{t-1}, X_{t+\frac{1}{2}} - X_{t+1}\rangle - \frac{\lambda_t}{2}D(X_{t+1}, X_{t+\frac{1}{2}})\right)\end{aligned} \tag{6.4}$$

To deal with the summation in the second to last line, we use strong convexity of h to bound

$$\begin{aligned}\|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2 &\leq 2\|X_{t+\frac{1}{2}} - X_t\|^2 + 2\|X_t - X_{t-\frac{1}{2}}\|^2 \\ &\leq 4D(X_{t+\frac{1}{2}}, X_t) + 4D(X_t, X_{t-\frac{1}{2}}).\end{aligned}$$

As for the last line, we use Young's inequality to obtain

$$\begin{aligned}&\langle g_t - g_{t-1}, X_{t+\frac{1}{2}} - X_{t+1}\rangle - \frac{\lambda_t}{2} D(X_{t+1}, X_{t+\frac{1}{2}}) \\ &\leq \frac{\|g_t - g_{t-1}\|_*^2}{\lambda_t} + \frac{\lambda_t}{4}\|X_{t+\frac{1}{2}} - X_{t+1}\|^2 - \frac{\lambda_t}{4}\|X_{t+\frac{1}{2}} - X_{t+1}\|^2 \\ &= \frac{\|g_t - g_{t-1}\|_*^2}{\lambda_t}. \end{aligned} \tag{6.5}$$

Putting the above inequalities together we get (6.3). □

Inequality (6.3) in Lemma 6.1 is very similar to the *Regret bounded by Variations in Utilities* (RVU) property introduced by Syrgkanis et al. [259], but it now applies to an algorithm with dynamic learning rate (and, of course, to continuous action spaces). This upper bound, similar to Proposition 4.1, depends on the choice of the reference point z, and on the variation $\delta_t = \|g_t - g_{t-1}\|_*^2$. This also justifies our decision of taking the learning rate η_t to be almost the inverse of the square root of the second-order path length. Finally, when all the players play a such algorithm, the additional negative term cancels out the variation, leading to improved regret.

6.2.2 *Example Algorithms*

To begin, (OptDA) with $\tilde{g}_t = g_{t-1}$ fulfills the two conditions outlined in Definition 6.1.

Proposition 6.2. (OptDA) *with non-increasing learning rates and* $\tilde{g}_t = g_{t-1}$ *belongs to* $\mathfrak{A}_{ocd}$. *Moreover, inequality* (6.2) *is satisfied with* $\psi_t(z) = F(z, Y_t)$ *and* $\varphi(z) = h(z) - \min h$, *where* $Y_t = -\eta_t \sum_{s=1}^{t-1} g_s$.

(OptDA) *is in* $\mathfrak{A}_{ocd}$

Proof. With $\lambda_t = 1/\eta_t$, the update writes

$$\begin{aligned} X_t &= \arg\min_{x\in\mathcal{X}} \sum_{s=1}^{t-1} \langle g_s, x\rangle + \lambda_t h(x), \\ X_{t+\frac{1}{2}} &= \arg\min_{x\in\mathcal{X}} \langle g_{t-1}, x\rangle + \lambda_t D(x, X_t). \end{aligned} \tag{OptDA}$$

For the dual averaging step, as shown by (2.6) in the proof of Proposition 2.4, we have

$$\langle g_t, X_{t+1} - z\rangle \leq \lambda_t F(z, Y_t) - \lambda_{t+1} F(z, Y_{t+1}) - \lambda_t F(X_{t+1}, Y_t) + (\lambda_{t+1} - \lambda_t)\varphi(z). \tag{6.6}$$

As for the update of $X_{t+\frac{1}{2}}$, we note that $X_{t+\frac{1}{2}} = Q(\nabla h(X_t) - \eta_t g_{t-1})$. Therefore, invoking Lemma A.1 gives

$$\langle \nabla h(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - z\rangle \leq \langle \nabla h(X_t) - \eta_t g_{t-1}, X_{t+\frac{1}{2}} - z\rangle.$$

For the specific choice $z \leftarrow X_{t+1}$, using the three-point identity for Bregman divergence (A.1) we obtain

$$\begin{aligned}\langle g_{t-1}, X_{t+\frac{1}{2}} - X_{t+1}\rangle &\leq \lambda_t \langle \nabla h(X_t) - \nabla h(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - X_{t+1}\rangle \\ &= \lambda_t (D(X_{t+1}, X_t) - D(X_{t+1}, X_{t+\frac{1}{2}}) - D(X_{t+\frac{1}{2}}, X_t)). \end{aligned} \tag{6.7}$$

Since $F(X_{t+1}, Y_t) \geq D(X_{t+1}, X_t)$ by Lemma A.3, combining (6.6) and (6.7) leads to

$$\begin{aligned}\langle g_t, X_{t+\frac{1}{2}} - z\rangle &= \langle g_t - g_{t-1}, X_{t+\frac{1}{2}} - X_{t+1}\rangle + \langle g_{t-1}, X_{t+\frac{1}{2}} - X_{t+1}\rangle + \langle g_t, X_{t+1} - z\rangle \\ &\leq \lambda_t F(z, Y_t) - \lambda_{t+1} F(z, Y_{t+1}) + (\lambda_{t+1} - \lambda_t)\varphi(z) \\ &\quad + \langle g_t - g_{t-1}, X_{t+\frac{1}{2}} - X_{t+1}\rangle - \lambda_t D(X_{t+1}, X_{t+\frac{1}{2}}) - \lambda_t D(X_{t+\frac{1}{2}}, X_t).\end{aligned}$$

This proves the generated iterates of (OptDA) satisfy (6.2) with $\psi_t = F(\cdot, Y_t)$ and $\varphi = h - \min h$ and accordingly (OptDA) belongs to $\mathfrak{A}_{ocd}$. □

Remark 6.1. If we use two different regularizers h_1 and h_2 for the optimistic and the update steps of the algorithm, the above analysis still holds as long as $F_{h_2}(X_{t+1}, Y_t) \geq D_{h_1}(X_{t+1}, X_t)$. In particular, if there exists some constants c_1 and c_2 such that $h_1 = c_1 h$ and $h_2 = c_2 h$ for a certain regularizer h, then we can have $c_2 \geq c_1$. This corresponds to making the update learning rate smaller than the optimistic learning rate, as we have discussed in Chapter 4 and will also investigate in Chapters 7 and 8.

Dual stabilized optimistic mirror descent

Another algorithm that belongs to $\mathfrak{A}_{ocd}$ is *dual stabilized optimistic mirror descent* (DS-OptMD), which we state recursively as[2]

$$\begin{aligned}X_{t+\frac{1}{2}} &= \arg\min_{x\in\mathcal{X}} \langle g_{t-1}, x\rangle + \lambda_t D(x, X_t), \\ X_{t+1} &= \arg\min_{x\in\mathcal{X}} \langle g_t, x\rangle + \lambda_t D(x, X_t) + (\lambda_{t+1} - \lambda_t) D(x, X_1).\end{aligned} \tag{DS-OptMD}$$

The stabilization technique (i.e., the anchoring term that appears in the second line of the update) was first introduced in Fang et al. [76]. In the said paper, the authors show that unlike (MD), dual stabilized mirror can achieve no regret even when the Bregman diameter is unbounded. Moreover, by standard arguments [76, 143, 163], we can show that when the mirror map is interior-valued, i.e., $\operatorname{im} Q^i = \operatorname{ri} \mathcal{X}^i$ (here $\operatorname{ri} \mathcal{X}^i$ denotes the relative interior of $\mathcal{X}^i$), the update of (DS-OptMD) coincides with that of (OptDA).[3] One important example which falls into this situation is the (stabilized) optimistic multiplicative weights update (OMWU) algorithm [53], whose update can be written in a coordinate-wise way as follows

Optimistic multiplicative weights update

$$x^i_{t,k} = X^i_{t+\frac{1}{2},k} = \frac{\exp(-\eta^i_t(\sum_{s=1}^{t-1} g_{s,k} + g_{t-1,k}))}{\sum_{l=1}^{d_i} \exp(-\eta^i_t(\sum_{s=1}^{t-1} g_{s,l} + g_{t-1,l}))}. \tag{OMWU}$$

In the proposition below, we show that (DS-OptMD) is effectively optimistic and compatible with dynamic learning rates.

(DS-OptMD) *belongs to* $\mathfrak{A}_{ocd}$

Proposition 6.3. (DS-OptMD) *with non-increasing learning rates belongs to* $\mathfrak{A}_{ocd}$. *Moreover, inequality* (6.2) *is satisfied with* $\psi_t = D(\cdot, X_t)$ *and* $\varphi = D(\cdot, X_1)$.

2 Here we focus directly on the case $\tilde{g}_t = g_{t-1}$.

3 Precisely, this requires to take $\tilde{g}_t = g_{t-1}$ in (OptDA) and set $X_1 = \arg\min_{x\in\mathcal{X}^i} h^i(\mathcal{X})$ in (DS-OptMD).

Proof. By definition of the Bregman divergence and the mirror map, the second step is equivalent to

$$X_{t+1} = Q\left(\frac{\lambda_t}{\lambda_{t+1}}\nabla h(X_t) + (1 - \frac{\lambda_t}{\lambda_{t+1}})\nabla h(X_1) - \frac{g_t}{\lambda_{t+1}}\right).$$

This shows that the update of X_{t+1} consists in fact of a mixing step in the dual space with weight λ_t/λ_{t+1} followed by a standard mirror descent step. Applying Lemma A.1 gives

$$\langle \nabla h(X_{t+1}), X_{t+1} - z\rangle \leq \left\langle \frac{\lambda_t}{\lambda_{t+1}}\nabla h(X_t) + (1 - \frac{\lambda_t}{\lambda_{t+1}})\nabla h(X_1) - \frac{g_t}{\lambda_{t+1}}, X_{t+1} - z\right\rangle.$$

We rearrange the terms and use the three-point identity (A.1) to get

$$\begin{aligned}\langle g_t, X_{t+1} - z\rangle &\leq \lambda_t\langle \nabla h(X_t) - \nabla h(X_{t+1}), X_{t+1} - z\rangle \\ &\quad + (\lambda_{t+1} - \lambda_t)\langle \nabla h(X_1) - \nabla h(X_{t+1}), X_{t+1} - z\rangle \\ &\leq \lambda_t(D(z, X_t) - D(z, X_{t+1}) - D(X_{t+1}, X_t)) \\ &\quad + (\lambda_{t+1} - \lambda_t)(D(z, X_1) - D(z, X_{t+1}) - D(X_{t+1}, X_1)) \qquad (6.8)\end{aligned}$$

Since $X_{t+\frac{1}{2}}$ is computed exactly as in (OptDA), inequality (6.7) still holds. We conclude by putting together (6.8) and (6.7)

$$\begin{aligned}\langle g_t, X_{t+\frac{1}{2}} - z\rangle &= \langle g_t - g_{t-1}, X_{t+\frac{1}{2}} - X_{t+1}\rangle + \langle g_{t-1}, X_{t+\frac{1}{2}} - X_{t+1}\rangle + \langle g_t, X_{t+1} - z\rangle \\ &\leq \lambda_t D(z, X_t) - \lambda_{t+1}D(z, X_{t+1}) + (\lambda_{t+1} - \lambda_t)D(z, X_1) \\ &\quad + \langle g_t - g_{t-1}, X_{t+\frac{1}{2}} - X_{t+1}\rangle - \lambda_t D(X_{t+1}, X_{t+\frac{1}{2}}) - \lambda_t D(X_{t+\frac{1}{2}}, X_t).\end{aligned}$$

This prove the proposition. □

At this point, we have shown that both (OptDA) and (DS-OptMD) belong to $\mathfrak{R}_{ocd}$. Nonetheless, as we will discuss in Section 6.4, (DS-OptMD) might be more favorable if we want to guarantee the convergence of the trajectory of play in certain situations.

To complete the picture, it is worth noticing that although (OptMD) does so not satisfy (6.2) in general, it does when $\sup_{z,x\in\mathcal{X}^i} D(z, x) < +\infty$; in this case, $\psi_t = D(\cdot, X_t)$ and $\varphi \equiv \sup_{z,x\in\mathcal{X}^i} D(z, x)$. Finally, some other algorithm, such as optimistic follow the regularized leader [140, 198], satisfies (6.2) but not (6.1). As mentioned earlier, this means that the regret bound would still hold while our analysis for the trajectory convergence cannot apply.

6.3 OPTIMAL REGRET BOUNDS

In this section, we establish regret bounds for our algorithms under a variety of conditions, providing an indication of the efficacy of these algorithms.

6.3.1 *No-Regret Against Adversarial Opponents*

We start with the fallback regret guarantee against *any* sequence of play realized by the opponents. The following theorem is a direct consequence of Lemma 6.1 and the choice of our adaptive learning rate (Adapt). It shows that the algorithms that we consider in this chapter are indeed no-regret in the classical sense.

$O(\sqrt{T})$ against bounded adversarial feedback

Theorem 6.4. *Suppose that Assumption 5.1 holds, and a player $i \in \mathcal{N}$ adopts an algorithm of $\mathfrak{A}_{ocd}$ with adaptive learning rate (Adapt). Then, if $\mathcal{Z}^i \subseteq \mathcal{X}^i$ is bounded and $G = \sup_t \|g_t^i\|$ is finite, the regret incurred by the player is bounded as*

$$\mathrm{Reg}_T^i(\mathcal{Z}^i) = O(G\sqrt{T} + G^2).$$

Proof. Combining Lemma 6.1 and Lemma 2.6 we have

$$\begin{aligned}
\sum_{t=1}^{T} \langle g_t^i, X_{t+\frac{1}{2}}^i - z^i \rangle &\leq \lambda_{T+1}^i \varphi^i(z^i) + \sum_{t=1}^{T} \frac{\delta_t^i}{\lambda_t^i} \\
&= \lambda_{T+1}^i \varphi^i(z^i) + \sum_{t=1}^{T} \frac{\delta_t^i}{\lambda_{t+1}^i} + \sum_{t=1}^{T} \left(\frac{1}{\lambda_t^i} - \frac{1}{\lambda_{t+1}^i} \right) \delta_t^i \\
&\leq (2 + \varphi^i(z^i)) \sqrt{1 + \sum_{t=1}^{T} \delta_t^i} + 4 \sum_{t=1}^{T} \left(\frac{1}{\lambda_t^i} - \frac{1}{\lambda_{t+1}^i} \right) G^2 \\
&\leq (2 + \varphi^i(z^i)) \sqrt{1 + 4G^2 T} + 4G^2
\end{aligned}$$

We conclude by maximizing the above inequality over $z^i \in \mathcal{Z}^i$. □

6.3.2 *Bound on Social Regret in Self-Play*

We next provide a constant regret bound on the *social regret*, that is, the sum of all the players' regrets, when all the players adopt a suitable optimistic algorithm with our adaptive learning rate policy. For this result, we do *not* need to assume that the game is variationally stable.

Constant social regret

Theorem 6.5. *Suppose that Assumptions 5.1 and 5.2 hold and all players of the game adopt an algorithm of $\mathfrak{A}_{ocd}$ with adaptive learning rate (Adapt). Then, for every bounded comparator set $\mathcal{Z} := \bigtimes_{i \in \mathcal{N}} \mathcal{Z}^i \subseteq \mathfrak{X}$, the players' social regret is bounded as*

$$\sum_{i=1}^{N} \mathrm{Reg}_T^i(\mathcal{Z}^i) = O(1).$$

Proof. Let $\mathbf{z} = (z^i)_{i \in \mathcal{N}} \in \mathcal{Z}$. Since $\mathcal{Z}$ is bounded and φ^i is continuous, there exists $M^i > 0$ such that it always holds $\varphi^i(z^i) \leq M^i$. Invoking Lemma 6.1 we get

$$\begin{aligned}
\sum_{t=1}^{T} \langle g_t^i, X_{t+\frac{1}{2}}^i - z^i \rangle &\leq \lambda_{T+1}^i M^i + \|V^i(\mathbf{x}_1)\|_{(i),*}^2 \\
&\quad + \sum_{t=2}^{T} \left(\frac{\|V^i(\mathbf{x}_t) - V^i(\mathbf{x}_{t-1})\|_{(i),*}^2}{\lambda_t^i} - \frac{\lambda_{t-1}^i}{8} \|X_{t+\frac{1}{2}}^i - X_{t-\frac{1}{2}}^i\|_{(i)}^2 \right).
\end{aligned} \tag{6.9}$$

In the current setting, the realized joint action is $\mathbf{x}_t = \mathbf{X}_{t+\frac{1}{2}}$. With the norm on $\mathcal{X}$ defined in Section 6.1 , we have $\sum_{i=1}^{N} \|X_{t+\frac{1}{2}}^i - X_{t-\frac{1}{2}}^i\|_{(i)}^2 = \|\mathbf{X}_{t+\frac{1}{2}} - \mathbf{X}_{t-\frac{1}{2}}\|^2$. Note that $\lambda_t^i \geq 1$ for all t and i by definition. Summing (6.9) from $i = 1$ to N and maximizing over $z \in \mathcal{Z}$ then gives

$$\sum_{i=1}^{N} \mathrm{Reg}_T^i(\mathcal{Z}^i) \leq \sum_{i=1}^{N} \left(\lambda_{T+1}^i M^i + \|V^i(\mathbf{x}_1)\|_{(i),*}^2 \right)$$

$$+\sum_{t=2}^{T}\left(\sum_{i=1}^{N}\frac{\|V^i(\mathbf{X}_{t+\frac{1}{2}})-V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2_{(i),*}}{\lambda_t^i}-\frac{1}{8}\|\mathbf{X}_{t+\frac{1}{2}}-\mathbf{X}_{t-\frac{1}{2}}\|^2\right). \tag{6.10}$$

In the remainder of the proof, we show that the right-hand side of (6.10) is bounded from above by some constant. Since all the norms are equivalent in a finite dimensional space, from Assumption 5.2 we know that for every $i \in \mathcal{N}$, there exists $L^i > 0$ such that for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$,

$$\|V^i(\mathbf{x})-V^i(\mathbf{x}')\|_{(i),*} \leq L^i\|\mathbf{x}-\mathbf{x}'\|. \tag{6.11}$$

Subsequently,

$$\|\mathbf{X}_{t+\frac{1}{2}}-\mathbf{X}_{t-\frac{1}{2}}\|^2 \geq \sum_{i=1}^{N}\frac{1}{N(L^i)^2}\|V^i(\mathbf{X}_{t+\frac{1}{2}})-V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2_{(i),*}. \tag{6.12}$$

It is thus sufficient to show that for each $i \in \mathcal{N}$, there exists $C^i \in \mathbb{R}_+$ such that for all $T \in \mathbb{N}$,

$$\lambda^i_{T+1}M^i - \frac{1}{16N(L^i)^2}\sum_{t=2}^{T}\|V^i(\mathbf{X}_{t+\frac{1}{2}})-V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2_{(i),*} \leq C^i, \tag{6.13}$$

$$\sum_{t=2}^{T}\left(\frac{\|V^i(\mathbf{X}_{t+\frac{1}{2}})-V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2_{(i),*}}{\lambda_t^i}-\frac{1}{16N(L^i)^2}\|V^i(\mathbf{X}_{t+\frac{1}{2}})-V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2_{(i),*}\right) \leq C^i. \tag{6.14}$$

To simplify the notation, we will write $\alpha^i = 1/(16N(L^i)^2)$. We recall that $\lambda_t^i = \sqrt{1+\sum_{s=1}^{t-1}\delta_t^i}$ where $\delta_t^i = \|g_t^i - g_{t-1}^i\|^2_{(i),*}$. Using the inequality $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$, we can bound the left-hand side of (6.13) as following

$$\begin{aligned} M^i\sqrt{1+\sum_{s=1}^{T}\delta_t^i} - \alpha^i\sum_{t=2}^{T}\delta_t^i &\leq M^i\sqrt{1+\delta_1^i} + M^i\sqrt{\sum_{s=2}^{T}\delta_t^i} - \alpha^i\sum_{t=2}^{T}\delta_t^i \\ &= f^i\left(\sqrt{\sum_{t=2}^{T}\delta_t^i}\right). \end{aligned} \tag{6.15}$$

where $f^i\colon y \in \mathbb{R} \mapsto -\alpha^i y^2 + M^i y + M^i\sqrt{1+\delta_1^i}$ is a quadratic function with negative leading coefficient and is hence bounded from above. This proves (6.13) by setting $C^i \geq \max_{y\in\mathbb{R}_+} f^i(y)$.

Note that $(\lambda_t^i)_{t\in\mathbb{N}}$ is non-decreasing. Therefore, it either converges to some finite limit or tends to plus infinity. We write $\lim_{t\to+\infty}\lambda_t^i = \lambda^i \in \mathbb{R}_+ \cup \{+\infty\}$. To prove (6.14), we tackle the two cases separately:

Case 1, $\lambda^i \in \mathbb{R}_+$: In other words, $\sum_{t=2}^{+\infty}\delta_t^i$ is finite. Since $\lambda_t^i \geq 1$, by taking $C^i \geq \sum_{t=2}^{+\infty}\delta_t^i$ inequality (6.14) is verified.

Case 2, $\lambda^i = +\infty$: Then $\lim_{t\to+\infty} 1/\lambda_t^i = 0$. The quantity $t' = \min_t\{t : 1/\lambda_t^i \leq \alpha^i\}$ is well-defined and the inequality (6.14) is satisfied as long as $C^i \geq \sum_{t=2}^{t'-1}(1/\lambda_t^i - \alpha^i)\delta_t^i$.

To summarize, we have proved that (6.13) and (6.14) must hold for some $C^i \in \mathbb{R}_+$. Therefore, invoking (6.10) and (6.12) we have effectively proved that the social regret is bounded by a constant. □

Theorem 6.5 demonstrates the possibility of achieving improved regret with our algorithms. Compared to [259], which proved constant social regret bound for finite games, our theorem applies to any continuous game with smooth and convex losses. Furthermore, our approach incorporates adaptive learning rates and allows players to adopt different regularizers or even different template algorithms, enhancing the adoptability of the techniques in practice.

Remark 6.2. The relevance of social regret, as pointed out by Syrgkanis et al. [259], lies in its relationship with the convergence speed of average social welfare to the price of anarchy in a family of finite games known as *smooth games* [236]. Nonetheless, the importance of the above analysis goes beyond its immediate implications, as it also paves the way for the subsequent findings presented in this chapter (used especially for proving Proposition 6.7).

6.3.3 *Bound on Individual Regret in Self-Play*

Building upon the proof of Theorem 6.5, in the following theorem we strengthen the result for variationally stable games, providing constant regret bound on each player's *individual regret*.

Constant individual regret

Theorem 6.6. *Suppose that Assumptions 5.1–5.3 hold and all players of the game adopt an algorithm of $\mathfrak{A}_{ocd}$ with adaptive learning rate (Adapt). Then, for every bounded comparator set $\mathcal{Z}^i \subseteq \mathcal{X}^i$, the regret of player $i \in \mathcal{N}$ is bounded as*

$$\mathrm{Reg}_T^i(\mathcal{Z}^i) = O(1).$$

Theorem 6.6 provides both improved regret guarantee for individual players from a learning perspective and faster convergence to CCE from a computational perspective. It extends a range of results previously proved for *finite* two-player, zero-sum games under the use of different learning algorithms [55, 145, 228], while shaving off the logarithmic factors.

In order to prove the theorem, we start by showing that the inverse of the learning rate converges to a finite constant (which is equivalent to saying that the learning rate converges to a positive constant).

Convergence of learning rate to positive constant

Proposition 6.7. *Suppose that Assumptions 5.2 and 5.3 hold and that all players $i \in \mathcal{N}$ adopt an algorithm of $\mathfrak{A}_{ocd}$ with adaptive learning rate (Adapt). Then, for every $i \in \mathcal{N}$, the sequence $(\lambda_t^i)_{t\in\mathbb{N}}$ converges to a finite constant $\lambda^i \in \mathbb{R}_+$ (equivalently, $\sum_{t=1}^{+\infty} \delta_t^i < +\infty$).*

Proof. In this proof we borrow the notations from the proof of Theorem 6.5. First, summing the left-hand side of (6.9) from $i = 1$ to N leads to $\sum_{t=1}^{T} \langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{z}\rangle$. Since the game is variationally stable, we may take $\mathbf{z} \leftarrow \mathbf{x}_\star \in \mathfrak{X}_\star$ a Nash equilibrium of the game, which gurantees that $\langle \mathbf{V}(\mathbf{x}), \mathbf{x} - \mathbf{x}_\star\rangle \geq 0$ for all $\mathbf{x} \in \mathfrak{X}$. Summing (6.9) from $i = 1$ to N and using the Lipschitz continuity of the functions, similar to (6.10), we obtain

$$\begin{aligned} 0 \leq \sum_{i=1}^{N} \left(\lambda_{T+1}^i \varphi^i(x_\star^i) + \|V^i(\mathbf{x}_1)\|_{(i),*}^2\right) \\ + \sum_{t=2}^{T} \left(\sum_{i=1}^{N} \frac{\|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\mathbf{X}_{t-\frac{1}{2}})\|_{(i),*}^2}{\lambda_t^i} - \frac{1}{8}\|\mathbf{X}_{t+\frac{1}{2}} - \mathbf{X}_{t-\frac{1}{2}}\|^2\right). \end{aligned} \tag{6.16}$$

Combining (6.13) and (6.14) with the above inequality, we deduce that for any $i \in \mathcal{N}$, setting $\widetilde{C}^i = -2\sum_{j=1}^N C^j + C^i - \|\mathbf{V}(\mathbf{x}_1)\|_*^2$ ensures that for all $T \in \mathbb{N}$,

$$\varphi^i(x_\star^i)\sqrt{1 + \sum_{s=1}^{T} \delta_t^i} - \alpha^i \sum_{t=2}^{T} \delta_t^i \geq \widetilde{C}^i.$$

Invoking (6.15) then gives $f^i\left(\sqrt{\sum_{t=2}^T \delta_t^i}\right) \geq \widetilde{C}^i$. Since f^i is a quadratic function with negative leading coefficient, $\lim_{y\to+\infty} f^i(y) = -\infty$. Accordingly, $\sum_{t=2}^{+\infty} \delta_t^i$ is finite, which in turn implies $\lambda^i = \lim_{t\to+\infty} \lambda_t^i < +\infty$. □

Proposition 6.7 confirms our intuition that, under favorable circumstances, adaptive algorithms essentially mimic the performance of an algorithm with suitably chosen constant learning rate. With this result in hand, we are now poised to establish the constant regret bound for each player.

Proof of Theorem 6.6. From Lemma 6.1 we have

$$\sum_{t=1}^{T} \langle g_t^i, X_{t+\frac{1}{2}}^i - z^i \rangle \leq \lambda_{T+1}^i \varphi^i(z^i) + \sum_{t=1}^{T} \frac{\delta_t^i}{\lambda_t^i}. \tag{6.17}$$

As φ^i is continuous and $\mathcal{Z}^i$ is bounded, $M^i = \max_{z^i\in\mathcal{Z}^i} \varphi^i(z^i)$ is well-defined. Moreover, $1/\lambda_t^i \leq 1$ for all t. Maximizing (6.17) over $z^i \in \mathcal{Z}^i$ then gives

$$\mathrm{Reg}_T^i(\mathcal{Z}^i) \leq \lambda_{T+1}^i M^i + \sum_{t=1}^{T} \delta_t^i \leq \lambda^i M^i + \sum_{t=1}^{+\infty} \delta_t^i,$$

where $\lambda^i = \lim_{t\to+\infty} \lambda_t^i$ and $\sum_{t=1}^{+\infty} \delta_t^i$ are finite according to Proposition 6.7. We have thus proved $\mathrm{Reg}_T^i(\mathcal{Z}^i) = O(1)$. □

6.4 TRAJECTORY CONVERGENCE

The results Section 6.3 focused on "average" measures of performance, namely the players' individual and social regret. Even though the derived bounds are sharp, they cannot be used to draw meaningful conclusions for the players' *actual* sequence of play. Our analysis in this section shows that, in fact, the proposed learning methods do stabilize to a best response or a Nash equilibrium in a number of relevant cases.

6.4.1 *Reciprocity Conditions*

To show convergence of the trajectory, we require the sequence $(\psi_t)_{t\in\mathbb{N}}$ introduced in Definition 6.1 to satisfy the following assumption.

Assumption on the seuqence of energy functions

Assumption 6.1. For some norm $\|\cdot\|$ and its associated distance function dist, the sequence $(\psi_t)_{t\in\mathbb{N}}$ satisfies

(a) For any $t \in \mathbb{N}$ and $z \in \mathcal{X}^i$, $\psi_t(z) \geq (1/2)\|X_t - z\|^2$.

(b) For any compact set $\mathcal{K} \in \mathcal{X}^i$ and $\varepsilon > 0$, there exists $\rho > 0$ such that if $\mathrm{dist}(X_t, \mathcal{K}) \leq \rho$ then $\psi_t(\mathcal{K}) := \min_{z\in\mathcal{K}} \psi_t(z) \leq \varepsilon$.

Recall that for (OptDA) and (DS-OptMD), we have respectively $\psi_t = F(\cdot, Y_t)$, where $Y_t = -\eta_t \sum_{s=1}^{t-1} g_s$, and $\psi_t = D(\cdot, X_t)$ (see Propositions 6.2 and 6.3). Therefore, Assumption 6.1(a) is indeed verified (Lemma A.3). This ensures that the sequence $(X_t)_{t\in\mathbb{N}}$ converges to z whenever the metric ψ_t converges to 0. Assumption 6.1(b) ensures the converse and is implied by the reciprocity conditions that we define below.

Reciprocity conditions

Definition 6.2 (Bregman reciprocity [41, 156])**.** Bregman reciprocity is satisfied for regularizer h defined over $\mathcal{X}$ if for any $z \in \mathcal{X}$, and $(X_t)_{t\in\mathbb{N}}$ a sequence of primal points such that $X_t \to z$, it holds $D(z, X_t) \to 0$.

Definition 6.3 (Fenchel reciprocity [194])**.** Fenchel reciprocity is satisfied for regularizer h defined over $\mathcal{X}$ if for any $z \in \mathcal{X}$, and $(Y_t)_{t\in\mathbb{N}}$ a sequence of dual vectors such that $Q(Y_t) \to z$, we have $F(z, Y_t) \to 0$.

These two reciprocity conditions are commonly used in the literature for establishing last-iterate convergence results of MD- and DA-type methods [41, 194, 196]. It is straightforward to verify that Bregman reciprocity is implied by Fenchel reciprocity (Lemma A.5), but the opposite is generally not true. For example, when h is the quadratic regularizer, Bregman reciprocity always holds while Fenchel reciprocity is only guaranteed when $\mathcal{X}$ is a polytope. In this regard, (DS-OptMD) could be more favorable then (OptDA) in terms of trajectory convergence guarantee.

6.4.2 *Convergence to Best Response*

A fundamental consistency property for online learning in games is that any player should end up "best responding" to the action profile of all other players if their actions stabilize (or are stationary). Formally, a player $i \in \mathcal{N}$ is said to *converge to best response* if, whenever the action profile $\mathbf{x}_t^{-i}$ of all other players converges to some limit profile $\mathbf{x}_\infty^{-i} \in \prod_{j\neq i} \mathcal{X}^j$, the sequence of actions $x_t^i \in \mathcal{X}^i$ of the focal player $i \in \mathcal{N}$ converges itself to $\mathrm{BR}(\mathbf{x}_\infty^{-i}) := \arg\min_{x^i\in\mathcal{X}^i} \ell^i(x^i, \mathbf{x}_\infty^{-i})$. We establish this key requirement below.

Convergence to best response

Theorem 6.8. *Suppose that Assumptions 5.1 and 5.2 hold, and a player $i \in \mathcal{N}$ adopts an algorithm of $\mathfrak{A}_{ocd}$ with adaptive learning rate* (Adapt)*. Assume additionally that the algorithm verifies Assumption 6.1 and that $\mathcal{X}^i$ is compact. Then, if all other players' actions converge to a point $\mathbf{x}_\infty^{-i} \in \prod_{j\neq i} \mathcal{X}^j$, player i's realized actions converge to the best response to $\mathbf{x}_\infty^{-i}$.*

Proof. The proof of the theorem relies on a "trapping" argument. Specifically, we show that when the sequence X_t^i gets close to a best response, all subsequent iterates must remain in this neighborhood provided that t is sufficiently large. Subsequently, we also show that the sequence $(X_t^i)_{t\in\mathbb{N}}$ visits any neighborhood of $\mathrm{BR}(\mathbf{x}_\infty^{-i})$ infinitely many times. Therefore, for every neighborhood of $\mathrm{BR}(\mathbf{x}_\infty^{-i})$, the iterates eventually get trapped into that neighborhood, and we conclude by showing $\|X_{t+\frac{1}{2}}^i - X_t^i\|_{(i)}$ converges to zero.

We break down this argument into four steps below. To begin, let us define $x_\star^i \in \mathfrak{X}_\star^i := \mathrm{BR}(\mathbf{x}_\infty^{-i})$ and $M^i = \max_{x_\star^i\in\mathfrak{X}_\star^i} \varphi(x_\star^i)$.

(1) *Descent inequality.* By slightly modifying the proof of Lemma 6.1, we derive immediately that

$$\lambda_{t+1}^i \psi_{t+1}^i(x_\star^i) \leq \lambda_t^i \psi_t^i(x_\star^i) + (\lambda_{t+1}^i - \lambda_t^i) M^i + \frac{\delta_t^i}{\lambda_t^i} - \langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}}^i - x_\star^i \rangle \\ - \frac{\lambda_t^i}{4} \|X_{t+1}^i - X_{t+\frac{1}{2}}^i\|_{(i)}^2 - \frac{\lambda_t^i}{2} \|X_{t+\frac{1}{2}}^i - X_t^i\|_{(i)}^2, \tag{6.18}$$

The scalar product term is not necessarily non-negative, but with $\widetilde{\mathbf{X}}_{t+\frac{1}{2}} = (X_{t+\frac{1}{2}}^i, \mathbf{x}_\infty^{-i})$, $\mathbf{x}_\star = (x_\star^i, \mathbf{x}_\infty^{-i})$, and R the diameter of $\mathcal{X}^i$, we can decompose

$$\begin{aligned} \langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}}^i - x_\star^i \rangle &= \langle V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\widetilde{\mathbf{X}}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}}^i - x_\star^i \rangle \\ &\quad + \langle V^i(\widetilde{\mathbf{X}}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}}^i - x_\star^i \rangle \\ &\geq -R \|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\widetilde{\mathbf{X}}_{t+\frac{1}{2}})\|_{(i),*} + \ell^i(\widetilde{\mathbf{X}}_{t+\frac{1}{2}}) - \ell^i(\mathbf{x}_\star). \end{aligned} \tag{6.19}$$

In the inequality we have used the convexity of $\ell^i(\cdot, \mathbf{x}_\infty^{-i})$. Let us write $\ell_\star^i = \min_{x^i \in \mathcal{X}^i} \ell^i(x^i, \mathbf{x}_\infty^{-i})$ and define

$$f : \mathbf{x}^{-i} \mapsto \max_{z^i \in \mathcal{X}^i} \|V^i(z^i, \mathbf{x}^{-i}) - V^i(z^i, \mathbf{x}_\infty^{-i})\|_{(i),*}.$$

Then, combining (6.18), (6.19), using the definition of f, and minimizing with respect to $x_\star^i \in \mathfrak{X}_\star^i$ leads to

$$\lambda_{t+1}^i \psi_{t+1}^i(\mathfrak{X}_\star^i) \leq \lambda_t^i \psi_t^i(\mathfrak{X}_\star^i) + (\lambda_{t+1}^i - \lambda_t^i) M + \frac{\delta_t^i}{\lambda_t^i} + R f(\mathbf{X}_{t+\frac{1}{2}}^{-i}) \\ - (\ell^i(\widetilde{\mathbf{X}}_{t+\frac{1}{2}}) - \ell_\star^i) - \frac{\lambda_t^i}{4} \|X_{t+1}^i - X_{t+\frac{1}{2}}^i\|_{(i)}^2 - \frac{\lambda_t^i}{2} \|X_{t+\frac{1}{2}}^i - X_t^i\|_{(i)}^2. \tag{6.20}$$

(2) *Convergence of terms that precede with plus sign.* We define $\chi_t^i = (\lambda_{t+1}^i - \lambda_t^i) M + \delta_t^i / \lambda_t^i + R f(\mathbf{X}_{t+\frac{1}{2}}^{-i})$. As V^i is continuous, $\mathcal{X}^i$ is compact, and the iterates $(\mathbf{x}_t^{-i})_{t \in \mathbb{N}}$ converges and is hence bounded, the sequence of feedback received by player i is also bounded. Let us denote this bound by G. We first show that *a*) $(\lambda_{t+1}^i - \lambda_t^i)_{t \in \mathbb{N}}$ and *b*) $(\delta_t^i / \lambda_t^i)_{t \in \mathbb{N}}$ converge to 0.

This trivially holds if $\lim_{t \to +\infty} \lambda_t^i < +\infty$ (which is equivalent to $\sum_{t=1}^{+\infty} \delta_t^i < +\infty$). Otherwise, we have $\lambda_t^i \to +\infty$. Since $\delta_t^i \leq 4G^2$, we deduce the sequence *b*) $(\delta_t^i / \lambda_t^i)_{t \in \mathbb{N}}$ converges to 0. For the sequence *a*), we simply note that

$$\lambda_{t+1}^i - \lambda_t^i = \frac{(\lambda_{t+1}^i)^2 - (\lambda_t^i)^2}{\lambda_{t+1}^i + \lambda_t^i} = \frac{\delta_t^i}{\lambda_{t+1}^i + \lambda_t^i} \leq \frac{2G^2}{\lambda_t^i} \xrightarrow{\lambda_t^i \to +\infty} 0.$$

We now turn our attention to the third term that appears in the definition of χ_t. Since $\mathcal{X}^i$ is compact and V^i is continuous, the function f is continuous by Berge's maximum theorem. Accordingly, $f(\mathbf{X}_{t+\frac{1}{2}}^{-i})$ converges to 0 when t goes to infinity. Combining the above arguments we have shown that $\lim_{t \to +\infty} \chi_t^i = 0$.

(3) *Convergence of X_t^i to best response.* Below we show that for any $\varepsilon > 0$, we have $\psi_t^i(\mathfrak{X}_\star^i) \leq \varepsilon$ for all t large enough. This means $\lim_{t\to+\infty} \psi_t^i(\mathfrak{X}_\star^i) = 0$ and accordingly $\lim_{t\to+\infty} \operatorname{dist}(X_t^i, \mathfrak{X}_\star^i) = 0$ thanks to Assumption 6.1(a).

To begin, we establish some preliminary results for three different situations. Since $\mathfrak{X}_\star^i \subset \mathcal{X}^i$ is a compact set, Assumption 6.1(b) ensures the existence of $\rho > 0$ such that if $\operatorname{dist}(X_t^i, \mathfrak{X}_\star^i) \leq \rho$ then $\psi_t^i(\mathfrak{X}_\star^i) \leq \varepsilon$.

Case 1, $\operatorname{dist}(X_{t+\frac{1}{2}}^i, \mathfrak{X}_\star^i) \geq \rho/2$: By continuity of $\ell^i(\cdot, \mathbf{x}_\infty^{-i})$ and compacity of $\mathcal{X}^i$ this implies the existence of $c > 0$ such that $\ell^i(\widetilde{\mathbf{X}}_{t+\frac{1}{2}}) - \ell_\star^i \geq c$ whenever we are in this situation. As $\lim_{t\to+\infty} \chi_t^i = 0$, there exists $t_1 \in \mathbb{N}$ such that for all $t \geq t_1$, $\chi_t^i \leq c/2$. For any $t \geq t_1$, the inequality (6.20) then gives

$$\lambda_{t+1}^i \psi_{t+1}^i(\mathfrak{X}_\star^i) \leq \lambda_t^i \psi_t^i(\mathfrak{X}_\star^i) + \chi_t^i - c - \frac{\lambda_t^i}{4} \|X_{t+1}^i - X_{t+\frac{1}{2}}^i\|_{(i)}^2 \leq \lambda_t^i \psi_t^i(\mathfrak{X}_\star^i) - \frac{c}{2}.$$

Case 2, $\operatorname{dist}(X_{t+\frac{1}{2}}^i, \mathfrak{X}_\star^i) \leq \rho/2$ and $\|X_{t+1}^i - X_{t+\frac{1}{2}}^i\|_{(i)} \geq \rho/2$: We define $t_2 \in \mathbb{N}$ such that for all $t \geq t_2$, $\chi_t^i \leq \rho^2/32$. Then for $t \geq t_2$,

$$\lambda_{t+1}^i \psi_{t+1}^i(\mathfrak{X}_\star^i) \leq \lambda_t^i \psi_t^i(\mathfrak{X}_\star^i) + \chi_t^i - (\ell^i(\widetilde{\mathbf{X}}_{t+\frac{1}{2}}) - \ell_\star^i) - \frac{\rho^2}{16} \leq \lambda_t^i \psi_t^i(\mathfrak{X}_\star^i) - \frac{\rho^2}{32}.$$

Case 3, $\operatorname{dist}(X_{t+\frac{1}{2}}^i, \mathfrak{X}_\star^i) \leq \rho/2$ and $\|X_{t+1}^i - X_{t+\frac{1}{2}}^i\|_{(i)} \leq \rho/2$: By the triangular inequality this implies $\operatorname{dist}(X_{t+1}^i, \mathfrak{X}_\star^i) \leq \rho$ and accordingly $\psi_{t+1}^i(\mathfrak{X}_\star^i) \leq \varepsilon$ by the choice of ρ.

Putting all together: Let us consider the sequence $(U_t^i)_{t\in\mathbb{N}} \in (\mathbb{R}_+)^{\mathbb{N}}$ defined by $U_t^i = \lambda_t^i \psi_t^i(\mathfrak{X}_\star^i)$. For $t \geq \max(t_1, t_2)$, whenever we are in Case 1 or 2, we have $U_{t+1}^i \leq U_t^i - \min(c/2, \rho^2/32)$. Since $(U_t^i)_{t\in\mathbb{N}} \in (\mathbb{R}_+)^{\mathbb{N}}$ is non-negative, this can not happen for all $t \geq \max(t_1, t_2)$; this means Case 3 must happen for some $t' \geq \max(t_1, t_2)$. Note that for both Case 1 and 2 we get $\psi_{t+1}^i(\mathfrak{X}_\star^i) \leq \psi_t^i(\mathfrak{X}_\star^i)$. Therefore, with the three cases presented above we see that for all $t \geq t' + 1$ we have $\psi_t^i(\mathfrak{X}_\star^i) \leq \varepsilon$. We have thus proved that for any $\varepsilon > 0$, the energy $\psi_t^i(\mathfrak{X}_\star^i)$ becomes eventually smaller than ε.

(4) *Convergence of $X_{t+\frac{1}{2}}^i$ to best response.* Now that we know that X_t^i converges to best response, it is sufficient to show that $\|X_{t+\frac{1}{2}}^i - X_t^i\|_{(i)} \to 0$. Using (6.20) we can write

$$\|X_{t+\frac{1}{2}}^i - X_t^i\|_{(i)}^2 \leq 2\left(\psi_t^i(\mathfrak{X}_\star^i) - \psi_{t+1}^i(\mathfrak{X}_\star^i) + \frac{\chi_t^i}{\lambda_t^i}\right).$$

As the right-hand side of the above inequality tends to 0 when t goes to infinity, we conclude that $\|X_{t+\frac{1}{2}}^i - X_t^i\|_{(i)} \to 0$. □

As a direct consequence of the convergence to best response, we deduce that $\lim_{t\to+\infty} \operatorname{Gap}_{\mathcal{X}^i}^i(\mathbf{x}_t) = 0$ whenever the opponents' actions converge. Therefore, the action of the player becomes quasi-optimal as time goes by, in the sense that what they could earn by switching to any other strategy in a round tends to 0.

Remark 6.3. The compactness assumption can be removed if the opponents are stationary. In fact, in that case, we may reformulate the question as that of minimization of a convex function and the convergence to a minimum is then implied by the more general convergence to Nash equilibrium that we show in the next section.

6.4.3 *Convergence to Nash Equilibrium*

Moving forward, we proceed to establish results concerning the convergence of the players' trajectory of play to Nash equilibrium when all players employ an adaptive learning algorithm of $\mathfrak{A}_{ocd}$ in a variationally stable game.

Convergence to Nash equilibrium in (strictly) variationally stable games

Theorem 6.9. *Suppose that Assumptions 5.1 and 5.2 hold and all players of the game adopt an algorithm of $\mathfrak{A}_{ocd}$ with adaptive learning rate (Adapt). Assume additionally that all the players' algorithms verify Assumption 6.1. Then, the induced trajectory of play converges to a Nash equilibrium provided that either of the following is satisfied*

a) The game is strictly variationally stable.

b) The game is variationally stable and h^i is subdifferentiable on all of $\mathcal{X}^i$ for all i.

The convergence to a Nash equilibrium $\mathbf{x}_\star$ implies that for every $i \in \mathcal{N}$ and every compact set $\mathcal{Z}^i \in \mathcal{X}^i$, $\lim_{t\to+\infty} \operatorname{Gap}^i_{\mathcal{Z}^i}(\mathbf{x}_t) = \operatorname{Gap}^i_{\mathcal{Z}^i}(\mathbf{x}_\star) \leq 0$ (Proposition 5.1). Thus, in the long run, the players are individually satisfied with their own choices of each play compared to any other action they could have pick from a comparator set. Such convergence results for adaptive methods are scarce in the learning-in-games literature. Among these, the closest antecedent to ours is the work of Lin et al. [177] where the authors prove convergence to Nash equilibrium in unconstrained cocoercive games,[4] with an adaptive learning rate that is the same across player (and which therefore requires access to global information to be computed). In this regard, Theorem 6.9 extends a wide range of earlier equilibrium convergence results that were obtained with a constant or diminishing —but not *adaptive* —learning rate.

In order to prove Theorem 6.9, we start by showing that the distance between successive iterates (as indexed by $s \in \mathbb{N}/2$) converges to 0.

Lemma 6.10. *Suppose that Assumptions 5.2 and 5.3 hold and that all players of the game adopt an algorithm of $\mathfrak{A}_{ocd}$ with adaptive learning rate (Adapt). Then, $\|\mathbf{X}_{t+\frac{1}{2}} - \mathbf{X}_t\| \to 0$ and $\|\mathbf{X}_t - \mathbf{X}_{t-\frac{1}{2}}\| \to 0$ as $t \to +\infty$.*

Proof. Let $\mathbf{x}_\star$ be a Nash equilibrium. We apply (6.4) to $z^i \leftarrow x^i_\star$, and sum these bounds for $i = 1$ to N, with Young's inequality (6.5), we get

$$\frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\lambda_t^i\left(D^i(X^i_{t+1}, X^i_{t+\frac{1}{2}}) + D^i(X^i_{t+\frac{1}{2}}, X^i_t)\right) \leq \sum_{i=1}^{N}\left(\lambda^T_{t+1}\varphi^i(x^i_\star) + \sum_{t=1}^{+\infty}\frac{\delta^i_t}{\lambda^i_t}\right). \tag{6.21}$$

The right-hand side of (6.21) is finite by Proposition 6.7. With strong convexity of h^i, this implies

$$\sum_{t=1}^{+\infty}\left(\|\mathbf{X}_{t+1} - \mathbf{X}_{t+\frac{1}{2}}\|^2 + \|\mathbf{X}_{t+\frac{1}{2}} - \mathbf{X}_t\|^2\right) < +\infty.$$

Therefore, both $\|\mathbf{X}_{t+\frac{1}{2}} - \mathbf{X}_t\|$ and $\|\mathbf{X}_t - \mathbf{X}_{t-\frac{1}{2}}\|$ converge to 0 when $t \to +\infty$. □

Another important building block for the proof is the convergence of the energy function $\sum_{i=1}^N \lambda^i \psi^i_t(x^i_\star)$.

Lemma 6.11. *Suppose that Assumptions 5.2 and 5.3 hold and that all players $i \in \mathcal{N}$ adopt an algorithm of $\mathfrak{A}_{ocd}$ with adaptive learning rate (Adapt). Then, $\sum_{i=1}^N \lambda^i \psi^i_t(x^i_\star)$ converges for all Nash equilibrium $\mathbf{x}_\star \in \mathfrak{X}_\star$.*

4 The class of cocoercive games is defined by the property $\langle \mathbf{V}(\mathbf{x}) - \mathbf{V}(\mathbf{x}'), \mathbf{x} - \mathbf{x}'\rangle \geq (1/\beta)\|\mathbf{V}(\mathbf{x}) - \mathbf{V}(\mathbf{x}')\|_*^2$.

Proof. Let $\mathbf{x}_\star$ be a Nash equilibrium. From the descent inequality (6.2), it is straightforward to show that

$$\sum_{i=1}^N \lambda_{t+1}^i \psi_{t+1}^i(x_\star^i) \leq \sum_{i=1}^N \lambda_t^i \psi_t^i(x_\star^i) - \langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star \rangle + \sum_{i=1}^N \left((\lambda_{t+1}^i - \lambda_t^t)\varphi^i(x_\star^i) + \frac{\delta_t^i}{2\lambda_t^i} \right).$$

By the choice of $\mathbf{x}_\star$, $\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star \rangle \geq 0$. On the other hand, thanks to Proposition 6.7 we know that the term on the second line is summable. Therefore, by applying Lemma B.1, we deduce the convergence of $\sum_{i=1}^N \lambda_t^i \psi_t^i(x_\star^i)$. This in particular implies that $\psi_t^i(x_\star^i)$ is bounded above for all i and t; hence $\sum_{i=1}^N (\lambda^i - \lambda_t^i)\psi_t^i(x_\star^i)$ converges to 0, and the convergence of $\sum_{i=1}^N \lambda^i \psi_t^i(x_\star^i)$ follows immediately. □

With the above two lemmas, we are now ready prove Theorem 6.9.

Proof of Theorem 6.9. We first show that in both cases, a cluster point of $(\mathbf{X}_t)_{t\in\mathbb{N}}$ is necessarily a Nash equilibrium.

a) Let $\mathbf{x}_\infty$ be a cluster point of $(\mathbf{X}_t)_{t\in\mathbb{N}}$ and $\mathbf{x}_\star$ be a Nash equilibrium. The point $\mathbf{x}_\infty$ is also a cluster point of $(\mathbf{X}_{t+\frac{1}{2}})_{t\in\mathbb{N}}$ since $\lim_{t\to+\infty} \|\mathbf{X}_{t+\frac{1}{2}} - \mathbf{X}_t\| = 0$. From the proof of Theorem 6.5, we have $\sum_{t=1}^T \langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star \rangle = O(1)$. As $\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star \rangle \geq 0$ for all t, this implies $\lim_{t\to+\infty} \langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star \rangle = 0$. Subsequently, $\langle \mathbf{V}(\mathbf{x}_\infty), \mathbf{x}_\infty - \mathbf{x}_\star \rangle = 0$ by the continuity of $\mathbf{V}$, which shows that $\mathbf{x}_\infty$ must be a Nash equilibrium by the strict variational stability of the game.

b) Let $\mathbf{x}_\infty \in \mathcal{X}$ be a cluster point of $(\mathbf{X}_t)_{t\in\mathbb{N}}$. We recall that $X_{t+\frac{1}{2}}^i$ is obtained by

$$X_{t+\frac{1}{2}}^i = \arg\min_{x\in\mathcal{X}^i} \left\{ \langle V^i(\mathbf{X}_{t-\frac{1}{2}}), x \rangle + \lambda_t^i D^i(x, X_t^i) \right\}.$$

For any $z^i \in \mathcal{X}^i$, the optimality condition Lemma A.1 then gives

$$\langle V^i(\mathbf{X}_{t-\frac{1}{2}}) + \lambda_t^i \nabla h^i(X_{t+\frac{1}{2}}^i) - \lambda_t^i \nabla h^i(X_t^i), z^i - X_{t+\frac{1}{2}}^i \rangle \geq 0. \tag{6.22}$$

Let $(\mathbf{X}_{\omega(t)})_{t\in\mathbb{N}}$ be a subsequence that converges to $\mathbf{x}_\infty$. With $\lim_{t\to+\infty} \|\mathbf{X}_{t+\frac{1}{2}} - \mathbf{X}_t\| = 0$ and $\lim_{t\to+\infty} \|\mathbf{X}_t - \mathbf{X}_{t-\frac{1}{2}}\| = 0$ (Lemma 6.10), we deduce $\mathbf{X}_{\omega+\frac{1}{2}} \to \mathbf{x}_\infty$ and $\mathbf{X}_{\omega-\frac{1}{2}} \to \mathbf{x}_\infty$. Since both ∇h^i and V^i are continuous (∇h^i is a *continuous* selection of the subgradients of h^i) and $\mathcal{X}^i \subset \operatorname{dom} \partial h^i$, by substituting $t \leftarrow \omega(t)$ in (6.22) and letting t go to infinity, we get

$$\langle V^i(\mathbf{x}_\infty) + \lambda^i \nabla h^i(x_\infty^i) - \lambda^i \nabla h^i(x_\infty^i), z^i - x_\infty^i \rangle \geq 0.$$

In other words, for all $z^i \in \mathcal{X}^i$, it holds that

$$\langle \nabla_{x^i} \ell^i(\mathbf{x}_\infty), z^i - x_\infty^i \rangle \geq 0.$$

This is true for all $i \in \mathcal{N}$ and all $z^i \in \mathcal{X}^i$, which shows that $\mathbf{x}_\infty$ is indeed a Nash equilibrium thanks to the first-order characterization Proposition 5.7.

Conclude. Lemma 6.11 along with Assumption 6.1(a) implies the boundedness of $(\mathbf{X}_t)_{t\in\mathbb{N}}$. With the above we can readily show that $\operatorname{dist}(\mathbf{x}_t, \mathfrak{X}_\star) \to 0$ and

$\limsup_{t\to+\infty} \mathrm{Gap}^i_{\mathcal{Z}^i}(\mathbf{x}_t) \le 0$ for all i and every compact set $\mathcal{Z}^i \subset \mathcal{X}^i$ ($\mathbf{x}_t = \mathbf{X}_{t+\frac{1}{2}}$ is the realized action at time t).

Below, we further prove the convergence of the played iterates to a point using Assumption 6.1(b) and Lemma 6.11. The sequence $(\mathbf{X}_t)_{t\in\mathbb{N}}$, being bounded, necessarily possesses a cluster point which we denote by $\mathbf{x}_\infty$. We have proved that $\mathbf{x}_\infty$ must be a Nash equilibrium. Therefore, by Lemma 6.11 the sequence $\sum_{i=1}^N \lambda^i \psi^i_t(x^i_\infty)$ converges. In Assumption 6.1(b), we take $\mathcal{K} \leftarrow \{x^i_\infty\}$ and this means that when X^i_t is close enough to x^i_∞, $\psi^i_t(x^i_\infty)$ becomes arbitrarily small. Consequently, $\sum_{i=1}^N \lambda^i \psi^i_t(x^i_\infty)$ can only converge to 0. By invoking Assumption 6.1(a), we then get $\lim_{t\to+\infty} \mathbf{X}_t = \mathbf{x}_\infty$, and subsequently $\lim_{t\to+\infty} \mathbf{X}_{t+\frac{1}{2}} = \mathbf{x}_\infty$ thanks to Lemma 6.10. □

6.4.4 *Adaptive OMWU Converges in Finite Two-Player Zero-Sum Games*

Despite the generality of Theorem 6.9, it fails to cover the case where players use regularizer h^i whose subdifferential is not defined on the whole $\mathcal{X}^i$ (instead it is only defined on $\operatorname{ri}\mathcal{X}^i$) in a game that is variationally stable but not *strictly* so. One notable example is when the two players of a finite two-player zero-sum game use adaptive versions of OMWU. We address this case below.

Convergence of adaptive OMWU in finite two-player zero-sum games

Theorem 6.12. *Suppose that the players of a finite two-player zero-sum game follow* (OMWU) *with adaptive learning rate* (Adapt)*. Then the induced sequence of play converges to a Nash equilibrium.*

Prior to our work, last-iterate convergence of OMWU were shown in [53, 277]. Theorem 6.12 sharpens these results in two key aspects: (*i*) the players' learning rate is not contingent on the knowledge of game-specific constants; and (*ii*) we do not assume the existence of a *unique* Nash equilibrium.

Saddle-point formulation and essential strategy

To prove the theorem, we consider the saddle-point formulation of the problem. Let us denote respectively by $\theta \in \Delta_m$ and $\phi \in \Delta_n$ the mixed strategy of the first and the second player. A point $(\theta_\star, \phi_\star)$ is a Nash equilibrium if for all $\theta \in \Delta_m$ and $\phi \in \Delta_n$,

$$\theta_\star^\top A \phi_\star \le \theta^\top A \phi_\star, \quad \theta_\star^\top A \phi \le \theta_\star^\top A \phi_\star. \tag{6.23}$$

where A is the payoff matrix and without loss of generality we assume $\|A\|_\infty \le 1$. We define $v = \min_{\theta\in\Delta_m} \max_{\phi\in\Delta_n} \theta^\top A \phi$ as the value of the game and we write $x_{[k]}$ for the k-th coordinate of x. A pure strategy α^i of player i is called *essential* if there exists a Nash equilibrium in which player i plays α^i with positive probability. We have the following lemma from [195].

Lemma 6.13. *Let $A \in \mathbb{R}^{m\times n}$ be the game matrix for a finite two-player zero-sum game with value v. There is a Nash equilibrium $(\theta_\star, \phi_\star)$ such that each player plays each of their essential strategies with positive probability, and*

$$\forall k \notin \operatorname{supp}(\theta_\star),\ (A\phi_\star)_{[k]} > v, \quad \forall l \notin \operatorname{supp}(\phi_\star),\ (A^\top \theta_\star)_{[l]} < v.$$

Notations for the proof

With Lemma 6.13, we are now ready to define a series of notations that will be used in our proof of Theorem 6.12.

- We write $\mathbf{x}_\star = (\theta_\star, \phi_\star)$ for an equilibrium that meets the description of Lemma 6.13. As an immediate consequence, we have $(A\phi_\star)_{[k]} = v$ for all $k \in \operatorname{supp}(\theta_\star)$ and $(A^\top \theta_\star)_{[l]} = v$ for all $l \in \operatorname{supp}(\phi_\star)$.

- The minimum difference between the value and the payoff of a non-chosen strategy at $\mathbf{x}_\star$ is denoted by

$$\xi = \min\left\{\min_{k\notin\mathrm{supp}(\theta_\star)} (A\phi_\star)_{[k]} - v, v - \max_{l\notin\mathrm{supp}(\phi_\star)} (A^\top\theta_\star)_{[l]}\right\}.$$

- For any $\hat{\theta} \in \Delta_m$, we use the notation

$$\mathcal{V}_{\hat{\theta}} = \{\theta \in \Delta_m : \mathrm{supp}(\theta) \subseteq \mathrm{supp}(\hat{\theta})\}$$

  for the set of the points whose support is included in that of $\hat{\theta}$ and define $\mathcal{V}_{\hat{\phi}}$ in the same way for any $\hat{\phi} \in \Delta_n$.

- We use D_{KL} to represent the Bregman divergence induced by the negentropy regularizer, i.e., the KL divergence.

- We consider the following continuous gradient selections of the negentropy regularizers:

$$\nabla h^1 \colon (\theta_{[k]})_{k\in[m]} \to (\log\theta_{[k]})_{k\in[m]},\ \ \nabla h^2 \colon (\phi_{[l]})_{l\in[m]} \to (\log\phi_{[l]})_{l\in[n]}.$$

Having established the relevant notations, we proceed to prove the theorem. We start by presenting several lemmas that are crucial to the analysis. To begin, we show that ξ lies in the interval $(0, 1]$.

Preparatory lemmas

Lemma 6.14. *It holds that* $0 < \xi \leq 1$.

Proof. The fact that $\xi > 0$ is immediate from the definition of ξ (which uses Lemma 6.13). As for the upper bound, we note that

$$\begin{aligned}\xi &\leq \frac{\min_{k\notin\mathrm{supp}(\theta_\star)} (A\phi_\star)_{[k]} - v + v - \max_{l\notin\mathrm{supp}(\phi_\star)} (A^\top\theta_\star)_{[l]}}{2}\\ &\leq \frac{\|A\phi_\star\|_\infty + \|A^\top\theta_\star\|_\infty}{2}\\ &\leq 1.\end{aligned}$$

□

The next lemma allows us to construct a Nash equilibrium with the help of a point that fulfills a condition that is weaker than the one required for a Nash equilibrium. We draw inspiration from [277] in proving this lemma.

Lemma 6.15. *Let* $\hat{\mathbf{x}} = (\hat{\theta}, \hat{\phi}) \in \Delta_m \times \Delta_n$ *satisfy that for all* $(\theta, \phi) \in \mathcal{V}_{\theta_\star} \times \mathcal{V}_{\phi_\star}$,

$$(\theta - \hat{\theta})^\top A\hat{\phi} + \hat{\theta}^\top A(\hat{\phi} - \phi) \geq 0. \tag{6.24}$$

Then $\mathbf{x}' = (1 - \xi/2)\mathbf{x}_\star + (\xi/2)\hat{\mathbf{x}}$ *is also a Nash equilibrium.*

Proof. We rewrite the left-hand side of (6.24) as

$$\begin{aligned}(\theta - \hat{\theta})^\top A\hat{\phi} + \hat{\theta}^\top A(\hat{\phi} - \phi) &= \theta^\top A\hat{\phi} - v + v - \hat{\theta}^\top A\phi\\ &= \theta^\top A(\hat{\phi} - \phi_\star) + (\theta_\star - \hat{\theta})^\top A\phi. \end{aligned}\tag{6.25}$$

The second equality holds because $(\theta, \phi) \in \mathcal{V}_{\theta_\star} \times \mathcal{V}_{\phi_\star}$. With the choice $(\theta, \phi) \leftarrow (\theta_\star, \phi_\star)$ and (6.24) we then get

$$\theta_\star^\top A(\hat{\phi} - \phi_\star) + (\theta_\star - \hat{\theta})^\top A\phi_\star \geq 0.$$

This implies

$$\theta_\star^\top A(\hat{\phi} - \phi_\star) = (\theta_\star - \hat{\theta})^\top A\phi_\star = 0 \tag{6.26}$$

by the definition of Nash equilibrium (6.23).

We next prove that $(\theta_\star, \phi')$ is also a Nash equilibrium with $\phi' = (1 - \xi/2)\phi_\star + (\xi/2)\hat{\phi}$. From (6.26) we already have

$$\theta_\star^\top A\phi' = \theta_\star^\top A\phi_\star = v = \max_{\phi \in \Delta_n} \theta_\star^\top A\phi.$$

It remains to show that $\theta_\star^\top A\phi' = \min_{\theta \in \Delta_m} \theta^\top A\phi'$. By choosing $\phi = \phi_\star$ in (6.25), we know that for all $\theta \in \mathcal{V}_{\theta_\star}$, it holds $\theta^\top A(\hat{\phi} - \phi_\star) \geq 0$. In other words,

$$\forall k \in \mathrm{supp}(\theta_\star), \;\; (A(\hat{\phi} - \phi_\star))_{[k]} \geq 0 \tag{6.27}$$

Let $\theta \in \Delta_m$. We decompose

$$\theta^\top A\phi' = \sum_{k \in \mathrm{supp}(\theta_\star)} \theta_{[k]}(A\phi')_{[k]} + \sum_{k \notin \mathrm{supp}(\theta_\star)} \theta_{[k]}(A\phi')_{[k]}. \tag{6.28}$$

The first term can be bounded below using (6.27),

$$\begin{aligned}\sum_{k \in \mathrm{supp}(\theta_\star)} \theta_{[k]}(A\phi')_{[k]} &= \sum_{k \in \mathrm{supp}(\theta_\star)} \left(\frac{\xi}{2}\, \theta_{[k]}(A(\hat{\phi} - \phi_\star))_{[k]} + \theta_{[k]}(A\phi_\star)_{[k]} \right) \\ &\geq \sum_{k \in \mathrm{supp}(\theta_\star)} \theta_{[k]} v. \end{aligned} \tag{6.29}$$

We proceed to lower bound the second term

$$\begin{aligned}\sum_{k \notin \mathrm{supp}(\theta_\star)} \theta_{[k]}(A\phi')_{[k]} &\geq \sum_{k \notin \mathrm{supp}(\theta_\star)} \left(\theta_{[k]}(A\phi_\star)_{[k]} - \frac{\xi}{2}\, |\theta_{[k]}(A(\hat{\phi} - \phi_\star))_{[k]}| \right) \\ &\geq \sum_{k \notin \mathrm{supp}(\theta_\star)} \left(\theta_{[k]}(A\phi_\star)_{[k]} - \frac{\xi}{2}\, \theta_{[k]} \|A\|_\infty \|\hat{\phi} - \phi_\star\|_1 \right) \\ &\geq \sum_{k \notin \mathrm{supp}(\theta_\star)} \theta_{[k]}((A\phi_\star)_{[k]} - \xi) \\ &\geq \sum_{k \notin \mathrm{supp}(\theta_\star)} \theta_{[k]} v. \end{aligned} \tag{6.30}$$

In the last inequality we use the definition of ξ. Combining (6.28), (6.29), and (6.30) we have $\theta^\top A\phi' \geq v = \mathbf{x}_\star[\theta^\top]A\phi'$. We have therefore proved that $(\theta_\star, \phi')$ is a Nash equilibrium. In the same way we can show that with $\theta' = (1 - \xi/2)\theta_\star + (\xi/2)\hat{\theta}$, the point $(\theta', \phi_\star)$ is also a Nash equilibrium. We then conclude that $\mathbf{x}' = (\theta', \phi')$ is indeed a Nash equilibrium. □

Thanks to Lemma 6.15, we can now establish two important properties of the cluster points of the played sequence.

Lemma 6.16. *Suppose that the players of a two-player, finite zero-sum game follow* (OMWU) *with adaptive learning rate* (Adapt)*. Then, for any cluster point* $\mathbf{x}_\infty$ *of the sequence of play, we have*

(a) The point $\mathbf{x}'_\star = (1 - \xi/2)\mathbf{x}_\star + (\xi/2)\mathbf{x}_\infty$ *is a Nash equilibrium.*

(b) The two points $\mathbf{x}_\infty$ *and* $\mathbf{x}_\star$ *have the same support, i.e.,* $\mathrm{supp}(\mathbf{x}_\infty) = \mathrm{supp}(\mathbf{x}_\star)$.

Proof. Let $\mathbf{x}_\infty = (\theta_\infty, \phi_\infty)$ be a cluster point of $(\mathbf{X}_{t+\frac{1}{2}})_{t\in\mathbb{N}}$. By Lemma 6.10, the distance $\|\mathbf{X}_{t+\frac{1}{2}} - \mathbf{X}_t\|$ converges to 0 and thus $\mathbf{x}_\infty$ is also a cluster point of $(\mathbf{X}_t)_{t\in\mathbb{N}}$. Moreover, using Lemma 6.11, we know that $\lambda^1 D_{\mathrm{KL}}(\theta_\star, \theta_t) + \lambda^2 D_{\mathrm{KL}}(\phi_\star, \phi_t)$ are bounded from above. This implies that for all $k \in \mathrm{supp}(\theta_\star)$ and $l \in \mathrm{supp}(\phi_\star)$, the coordinates $\theta_{t,[k]}$ and $\phi_{t,[l]}$ are bounded from below. Accordingly, we deduce that $\mathrm{supp}(\theta_\star) \subseteq \mathrm{supp}(\theta_\infty)$ and $\mathrm{supp}(\phi_\star) \subseteq \mathrm{supp}(\phi_\infty)$.

We next show that the point $\mathbf{x}'_\star = (1-\xi/2)\mathbf{x}_\star + (\xi/2)\mathbf{x}_\infty$ is a Nash equilibrium. Let $\theta \in \mathcal{V}(\theta_\star)$. For any $t \in \mathbb{N}$, we define $\theta'_{t+\frac{1}{2}}$ such that $\theta'_{t+\frac{1}{2},[k]} = \theta_{[k]}$ for $k \in \mathrm{supp}(\theta_\infty)$ and $\theta'_{t+\frac{1}{2},[k]} = \theta_{t+\frac{1}{2},[k]}$ for $k \notin \mathrm{supp}(\theta_\infty)$. Applying the optimality condition (6.22) to $z^1 \leftarrow \theta'_{t+\frac{1}{2}}$ gives

$$\sum_{k\in\mathrm{supp}(\theta_\infty)} (V^1(\mathbf{X}_{t-\frac{1}{2}})_{[k]} + \log(\theta_{t+\frac{1}{2},[k]}) - \lambda^1_t(\log(\theta_{t,[k]})))(\theta_{[k]} - \theta_{t+\frac{1}{2},[k]}) \geq 0 \quad (6.31)$$

As in the proof of Theorem 6.9, we take a subsequence of $(\mathbf{X}_t)_{t\in\mathbb{N}}$ that converges to $\mathbf{x}_\infty$ and take the limit of (6.31) along this subsequence. The fact that $\theta_{\infty[k]} > 0$ for all $k \in \mathrm{supp}(\theta_\infty)$ ensures that the limits of the log terms are well defined. With the convergence of $\|\mathbf{X}_t - \mathbf{X}_{t-\frac{1}{2}}\|$ and $\|\mathbf{X}_{t+\frac{1}{2}} - \mathbf{X}_t\|$ to 0 (Lemma 6.10), we get

$$\sum_{k\in\mathrm{supp}(\theta_\infty)} V^1(\mathbf{x}_\infty)_{[k]}(\theta_{[k]} - \theta_{\infty[k]}) \geq 0.$$

Since $\mathrm{supp}(\theta_\star) \subseteq \mathrm{supp}(\theta_\infty)$, the above can be rewritten as $(\theta - \theta_\infty)^\top A\phi_\infty \geq 0$. In the same way, for all $\phi \in \mathcal{V}(\phi_\star)$, we have $\theta_\infty^\top A(\phi_\infty - \phi) \geq 0$. As a consequence, it follows from Lemma 6.15 that $\mathbf{x}'_\star$ is effectively a Nash equilibrium.

To conclude, we note that $\mathbf{x}'_\star$ being a Nash equilibrium, we have $\mathrm{supp}(\mathbf{x}'_\star) \subseteq \mathrm{supp}(\mathbf{x}_\star)$ by the choice of $\mathbf{x}_\star$. Along with $\mathrm{supp}(\mathbf{x}_\star) \subseteq \mathrm{supp}(\mathbf{x}_\infty)$ and $0 < \xi \leq 1$ Lemma 6.14, we deduce that $\mathrm{supp}(\mathbf{x}_\star) = \mathrm{supp}(\mathbf{x}_\infty)$. □

Finally, we show that the sequence of play has only one cluster point, and thus this cluster point must be a Nash equilibrium according to Proposition 5.1.

Proof of convergence via the uniqueness of the cluster point

Proof of Theorem 6.12. Let $\mathbf{x}_\infty = (\theta_\infty, \phi_\infty)$ and $\mathbf{x}'_\infty = (\theta'_\infty, \phi'_\infty)$ be two cluster points of $(\mathbf{X}_t)_{t\in\mathbb{N}}$ (equivalently, of $(\mathbf{X}_{t+\frac{1}{2}})_{t\in\mathbb{N}}$ because $\|\mathbf{X}_{t+\frac{1}{2}} - \mathbf{X}_t\|$ converges to 0). By Lemma 6.16, we know that $\mathbf{x}'_\star = (1-\xi/2)\mathbf{x}_\star + (\xi/2)\mathbf{x}_\infty$ is a Nash equilibrium and $\mathrm{supp}(\mathbf{x}_\infty) = \mathrm{supp}(\mathbf{x}_\star) = \mathrm{supp}(\mathbf{x}'_\infty)$. We write $\mathbf{x}'_\star = (\theta'_\star, \phi'_\star)$. Using Lemma 6.11, we can define

$$U_\infty = \lim_{t\to+\infty} \lambda^1 D_{\mathrm{KL}}(\theta_\star, \theta_t) + \lambda^2 D_{\mathrm{KL}}(\phi_\star, \phi_t)$$
$$U'_\infty = \lim_{t\to+\infty} \lambda^1 D_{\mathrm{KL}}(\theta'_\star, \theta_t) + \lambda^2 D_{\mathrm{KL}}(\phi'_\star, \phi_t).$$

Since $\mathrm{supp}(\theta_\infty) = \mathrm{supp}(\theta_\star) = \mathrm{supp}(\theta'_\infty)$ and $\mathrm{supp}(\phi_\infty) = \mathrm{supp}(\phi_\star) = \mathrm{supp}(\phi'_\infty)$, we can use the continuity of the KL divergence with respect to the second variable to deduce that

$$\lambda^1 D_{\mathrm{KL}}(\theta_\star, \theta_\infty) + \lambda^2 D_{\mathrm{KL}}(\phi_\star, \phi_\infty) = U_\infty = \lambda^1 D_{\mathrm{KL}}(\theta_\star, \theta'_\infty) + \lambda^2 D_{\mathrm{KL}}(\phi_\star, \phi'_\infty),$$
$$\lambda^1 D_{\mathrm{KL}}(\theta'_\star, \theta_\infty) + \lambda^2 D_{\mathrm{KL}}(\phi'_\star, \phi_\infty) = U'_\infty = \lambda^1 D_{\mathrm{KL}}(\theta'_\star, \theta'_\infty) + \lambda^2 D_{\mathrm{KL}}(\phi'_\star, \phi'_\infty).$$

In other words, we have

$$\begin{aligned}&\lambda^1 \sum_{k\in\operatorname{supp}(\theta_\star)} \theta_{\star[k]} \log \theta_{\infty[k]} + \lambda^2 \sum_{l\in\operatorname{supp}(\phi_\star)} \phi_{\star[l]} \log \phi_{\infty[l]} \\ &= \lambda^1 \sum_{k\in\operatorname{supp}(\theta_\star)} \theta_{\star[k]} \log \theta'_{\infty[k]} + \lambda^2 \sum_{l\in\operatorname{supp}(\phi_\star)} \phi_{\star[l]} \log \phi'_{\infty[l]},\end{aligned} \tag{6.32}$$

and

$$\begin{aligned}&\lambda^1 \sum_{k\in\operatorname{supp}(\theta_\star)} \theta'_{\star[k]} \log \theta_{\infty[k]} + \lambda^2 \sum_{l\in\operatorname{supp}(\phi_\star)} \phi'_{\star[l]} \log \phi_{\infty[l]} \\ &= \lambda^1 \sum_{k\in\operatorname{supp}(\theta_\star)} \theta'_{\star[k]} \log \theta'_{\infty[k]} + \lambda^2 \sum_{l\in\operatorname{supp}(\phi_\star)} \phi'_{\star[l]} \log \phi'_{\infty[l]},\end{aligned} \tag{6.33}$$

With $(\theta'_\star, \phi'_\star) = (1-\xi/2)\mathbf{x}_\star + (\xi/2)\mathbf{x}_\infty$ and $\xi > 0$, using (6.32) and (6.33) we get

$$\begin{aligned}&\lambda^1 \sum_{k\in\operatorname{supp}(\theta_\star)} \theta_{\infty[k]} \log \theta_{\infty[k]} + \lambda^2 \sum_{l\in\operatorname{supp}(\phi_\star)} \phi_{\infty[l]} \log \phi_{\infty[l]} \\ &= \lambda^1 \sum_{k\in\operatorname{supp}(\theta_\star)} \theta_{\infty[k]} \log \theta'_{\infty[k]} + \lambda^2 \sum_{l\in\operatorname{supp}(\phi_\star)} \phi_{\infty[l]} \log \phi'_{\infty[l]},\end{aligned}$$

Since $\operatorname{supp}(\theta_\star) = \operatorname{supp}(\theta'_\infty)$ and $\operatorname{supp}(\phi_\star) = \operatorname{supp}(\phi'_\infty)$, the above is thus equivalent to

$$\lambda^1 D_{\mathrm{KL}}(\theta_\infty, \theta'_\infty) + \lambda^2 D_{\mathrm{KL}}(\phi_\infty, \phi'_\infty) = 0$$

This shows $\mathbf{x}_\infty = \mathbf{x}'_\infty$, and therefore $(\mathbf{X}_{t+\frac{1}{2}})_{t\in\mathbb{N}}$ has only one cluster point (the existence of which is guaranteed by the compactness of the actions sets); in other words, the induced sequence of play converges. We know that the players have no regret thanks to Theorem 6.6. We can thus conclude with the help of Proposition 5.1. □

6.5 NUMERICAL ILLUSTRATIONS

In this section we experimentally illustrate our theoretical results through two variationally stable games and a general-sum finite game. Precisely, we investigate the following three different setups.

Finite two-player zero-sum game

- A finite two-player zero-sum game with 10×10 cost matrix whose elements are drawn uniformly at random from $[-1, +1]$: We let the two players play (DS-OptMD) respectively with negentropy and quadratic regularizers. Note that the convergence of this particular configuration can be proved following the proof of Theorem 6.12.

Kelly auction

- A resource allocation auction (Example 5.2) with 6 resources and 20 bidders: We fix $c_k = 1$, draw q_k and r^i uniformly at random from $[4, 6]$, and draw b^i uniformly at random from $[5, 10]$. Each player runs either (OptDA) or (DS-OptMD) with quadratic regularizer.

Three-player-matching-pennies game

- A three-player-matching-pennies game introduced by Jordan [138]: Each player has two pure strategies. Player 1 wants to match the pure strategy of player 2; player 2 wants to match the pure strategy of player 3; and player 3 wants to match the opposite of the pure strategy of player 1. Each player receives a loss of -1 if they match as desired, and 1 otherwise. It is straightforward to see that the unique equilibrium is achieved when all

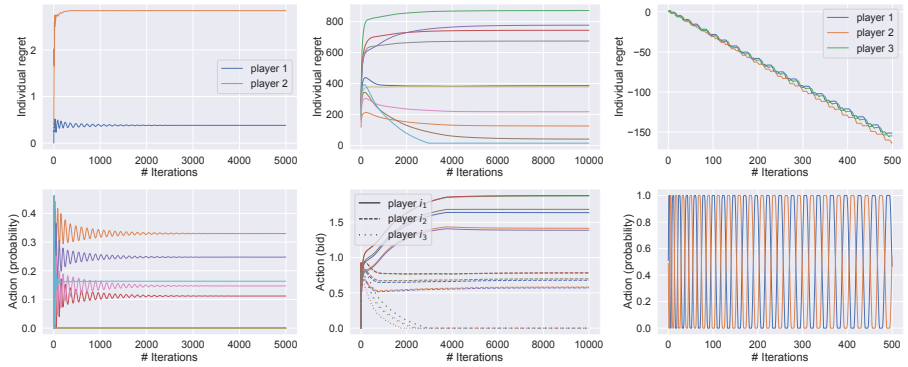


Figure 6.2 The (linearized) individual regret (top) and the realized actions (bottom, each line representing a coordinate of x_t^i) of a subset of players in a finite two-player zero-sum game (left), a resource allocation auction (middle), and a three-player matching-pennies game [138] (right). All the players use either adaptive (OptDA) or adaptive (DS-OptMD) as their learning strategies. We observe convergence of the realized actions and the regrets in the first two examples.

the players uniformly randomize. In this game, we let the three players run (DS-OptMD) with Euclidean regularizer.

Discussion on experimental results

As for the learning rates, we use the L2 norm in the definition of δ_t^i. The results are plotted in Fig. 6.2. Provided that the first two games are variationally stable, we observe the boundedness of individual regrets (measured with respect to the entire action set) and the convergence of iterates as predicted by our analysis. For the three-player-matching-pennies game, all the players oscillate between the two pure strategies, and have their individual regrets tend to minus infinity. We provided a theorem that partially characterized this behavior in [127] that this chapter draws from. Interested readers may also refer to the work of Anagnostides et al. [4] that further examined this phenomenon after the publication of our work.

7

DEALING WITH STOCHASTIC FEEDBACK I: TRAJECTORY CONVERGENCE

 # This chapter incorporates material from Hsieh et al. [126, 129]

In the previous chapter, we explored and tackled the challenge of adaptive learning rate tuning for optimistic online learning algorithms. This was a significant step toward the practical application of these methods. However, it relied on one crucial assumption: the availability of perfect gradient information. This assumption often does not hold in practice. For instance, in robotics, decisions are made based on noisy sensor data or incomplete observations of the environment [263]. Likewise, in machine learning, dealing with large datasets often necessitates the use of stochastic gradient methods [24]. Such stochasticity introduces an element of unpredictability and variance that can significantly affect the algorithm's performance and behavior.

Unconstrained setup

This chapter and the next are dedicated to addressing this issue. Our goal is to extend the theoretical guarantees of these algorithms (as presented in Proposition 5.8) to cases where the feedback is subject to stochastic perturbations. To do so, we will focus on the unconstrained setup $\mathcal{X}^i = \mathbb{R}^{d^i}$ and on the use of quadratic regularizer $h(\cdot)^i = \|\cdot\|^2/2$, where, for the sake of notational simplicity in the two chapters, we denote the L2 norm as $\|\cdot\| = \|\cdot\|_2$. Using $\tilde{g}_t^i = g_{t-1}^i$ then gives the OG algorithm.[1]

Optimistic gradient

$$X_{t+\frac{1}{2}}^i = X_t^i - \eta_t^i g_{t-1}^i, \quad X_{t+1}^i = X_t^i - \eta_{t+1}^i g_t^i. \tag{OG}$$

Extra-gradient

The focus of this chapter is on the (last-iterate) convergence of an algorithm to a Nash equilibrium. In this regard, although EG is not a valid online learning algorithm (see the discussion in Section 5.3), it can still be of interest when the goal is to compute an approximate equilibrium point, either when coordination between players are allowed, or more drastically, when there is a centralized entity that performs the computation. Since EG is only defined when used by all the players, we directly present the update for the joint iterate below.

$$\mathbf{X}_{t+\frac{1}{2}} = \mathbf{X}_t - \eta_t \hat{\mathbf{V}}_t, \quad \mathbf{X}_{t+1} = \mathbf{X}_t - \eta_{t+1} \hat{\mathbf{V}}_{t+\frac{1}{2}}. \tag{EG}$$

In the above, $\hat{\mathbf{V}}_t$ and $\hat{\mathbf{V}}_{t+\frac{1}{2}}$ are respectively the stochastic estimates of $\mathbf{V}(\mathbf{X}_t)$ and of $\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})$ (see Assumption 7.1). Moreover, we have assumed that all the players use the same learning rate η_t at round t.

Non-convergence and noise model

CONTRIBUTIONS AND OUTLINE. To motivate our analysis, we first present a counterexample in Section 7.1. We show that even in bilinear zero-sum games, where (EG) and (OG) with perfect feedback converge from any initialization,

1 In Chapter 4 we used the same name for the update with general guess vector $\tilde{g}_t$. In this chapter we use it to refer to the instantiation with $\tilde{g}_t = g_{t-1}$.

stochasticity in the feedback can cause these algorithms to fail. We also present our noise model in this section.

Learning rate separation and energy inequalities

In response to the failure we've observed, we propose EG+ and OG+ in Section 7.2. These algorithms utilize distinct learning rates for the optimistic and the update steps. Through energy inequalities, we illustrate the advantage of using a smaller update learning rate in the presence of stochastic feedback.

Global convergence

Building upon the aforementioned inequalities, we demonstrate a series of convergence results in Section 7.3. Precisely, we show that both EG+ and OG+ converge with probability 1 in all variationally stable games, and derive explicit convergence rates for the algorithms' last iterate under an additional error bound condition on the pseudo-gradient $\mathbf{V}$ of the game.

Local convergence

Finally, we demonstrate in Section 7.4 that it is possible to establish local versions of the aforementioned results, circumventing the sometimes impractical global assumptions. Concretely, we show that EG+ converges with (arbitrarily) high probability provided the game is locally variationally stable and the pseudo-gradient is locally Lipschitz around a first-order equilibrium point (i.e., a point where $\mathbf{V}(\mathbf{x}_\star) = 0$). We also provide expected convergence rate, conditioned on the convergence of the algorithm, when the Jacobian of $\mathbf{V}$ at the equilibrium point is invertible.

7.1 FEEDBACK MODEL AND FAILURE OF OPTIMISTIC METHODS

In this section, we present our model for the feedback oracle and illustrate through an example that noise in feedback may hinder the convergence of (EG) and (OG) in games where they would have otherwise converged.

7.1.1 *Noise Model*

Additive noise versus multiplicative noise

To account for stochasticity in the feedback, we consider two noise models, *additive noise* and *multiplicative noise*. To illustrate the difference between these two models, suppose we wish to estimate the value of some quantity $v \in \mathbb{R}$. Then, an estimate of v with additive noise is a random variable $\hat{v}_{\text{add}}$ of the form $\hat{v}_{\text{add}} = v + \xi_{\text{add}}$ for some zero-mean noise variable ξ_{add}; analogously, a multiplicative noise model for v is a random variable of the form $\hat{v}_{\text{mult}} = v(1 + \xi_{\text{mult}})$ for some zero-mean noise variable ξ_{mult}. The two models can be compared directly via the additive representation of the multiplicative noise model as $\hat{v}_{\text{mult}} = v + \xi_{\text{mult}} v$, which gives $\text{Var}[\xi_{\text{add}}] = v^2 \text{Var}[\xi_{\text{mult}}]$.

Feedback oracle

With all this in mind, we consider the following oracle feedback model: Let $\mathbf{X}_s = (X_s^i)_{i \in \mathcal{N}}$ collect the actions taken by all the players at time s for some $s \in \mathbb{N}/2$ (recall that we index the iterates of the optimistic methods by half-integer). The feedback received by player i at time s is $\hat{V}_s^i = V^i(\mathbf{X}_s) + \xi_s^i$, where ξ_s^i represents the aggregate measurement error relative to $V^i(\mathbf{X}_s)$. Moreover, with $(\mathcal{F}_s)_{s \in \mathbb{N}/2}$ the filtration such that $\boldsymbol{\xi}_s = (\xi_s^i)_{i \in \mathcal{N}}$ is $\mathcal{F}^i_{s+\frac{1}{2}}$-measurable but not $\mathcal{F}_s^i$-measurable, and $\mathbb{E}_s[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_s]$ the corresponding conditional expectation, we make the following assumption for the noise / measurement error vectors.

Zero-mean and variance control

Assumption 7.1. The noise vectors $(\boldsymbol{\xi}_s)_{s \in \mathbb{N}/2}$ satisfy the following requirements for some $\sigma_A, \sigma_M \geq 0$.

(a) *Zero-mean:* For all $i \in \mathcal{N}$ and $s \in \mathbb{N}/2$, $\mathbb{E}_s[\xi_s^i] = 0$.

(b) *Variance control:* For all $i \in \mathcal{N}$ and $s \in \mathbb{N}/2$, $\mathbb{E}_s[\|\xi_s^i\|^2] \leq \sigma_A^2 + \sigma_M^2 \|V^i(\mathbf{X}_s)\|^2$.

Special case: perfect feedback and absolute noise

Let us briefly discuss several special cases of Assumption 7.1. A first straightforward observation is that by setting $\sigma_A, \sigma_M = 0$ we recover the "perfect feedback" setup studied in Chapter 6, whereas for $\sigma_A = 0$, $\sigma_M > 0$ we obtain the standard "absolute noise" model that often serves as a context-agnostic model for stochastic first-order methods, cf. [142, 209] and references therein.

Special case: relative noise

A more intriguing case is when $\sigma_A = 0$ and $\sigma_M > 0$. This particular instance is sometimes referred to as "relative noise" [224], and it is widely used as a model for randomized coordinate descent methods [210], randomized player updates in game theory [8], and physical measurements in signal processing and control [243]. Although this is actually more general than the multiplicative noise model described above given that ξ_s^i and $V^i(\mathbf{X}_s)$ may not point to the same direction, we will, by a slight abuse of terminology, say that the noise is "multiplicative" whenever $\sigma_A = 0$. In our analysis, we show that the regret bounds and convergence rates have much better dependence on t in this case, essentially recovering the guarantees of the perfect information case. Otherwise, in the general case both σ_A and σ_M are positive, and we use the term "noise" to tacitly refer to the presence of both additive and multiplicative components.

Joint noise vector and joint feedback vector

NOISE OF ALL PLAYERS. Although Assumption 7.1 is stated for the individual noise component of each player, it is straightforward to translate it into results on the joint quantities as demonstrated below.

Proposition 7.1. *Suppose that Assumption 7.1 holds. Then, for all $s \in \mathbb{N}/2$, we have*

$$\mathbb{E}_s[\boldsymbol{\xi}_s] = 0, \quad \mathbb{E}_s[\|\boldsymbol{\xi}_s\|^2] \leq N\sigma_A^2 + \sigma_M^2\|\mathbf{V}(\mathbf{X}_s)\|^2. \tag{7.1}$$

Moreover, if $\mathbf{X}_s$ is $\mathcal{F}_s$-measurable for all $s \in \mathbb{N}/2$, we additionally have

$$\mathbb{E}_s[\hat{\mathbf{V}}_s] = \mathbf{V}(\mathbf{X}_s), \quad \mathbb{E}_s[\|\hat{\mathbf{V}}_s\|^2] \leq N\sigma_A^2 + (1+\sigma_M^2)\|\mathbf{V}(\mathbf{X}_s)\|^2 \tag{7.2}$$

Proof. This follows immediately from the assumptions. □

Natural filtration

Remark 7.1. For the algorithms that we consider in this chapter and the next, $(\mathcal{F}_s)_{s\in\mathbb{N}/2}$ is nothing but the natural filtration associated to $(\mathbf{X}_s)_{s\in\mathbb{N}/2}$ (this is for example *not* the case for an algorithm that always outputs a same vector). In particular, we will use the notation $\hat{\mathbf{V}}_{1/2} = \boldsymbol{\xi}_{1/2} = 0$ and we define $\mathcal{F}_1$ as the σ-algebra generated by $\mathbf{X}_1$, while $\mathcal{F}_{1/2}$ and $\mathcal{F}_0$ denote the trivial σ-algebra.

Notation for single-call methods

SINGLE-CALL VERSUS TWO-CALL METHODS. For the learning-in-games setup described in Section 5.1, we may write $g_t^i = V^i(\mathbf{x}_t) + \xi_t^i$ for the gradient feedback to player i at round t. When using optimistic methods, we have $\mathbf{x}_t = \mathbf{X}_{t+\frac{1}{2}}$ and thus $g_t^i = \hat{V}^i_{t+\frac{1}{2}}$ and $\xi_t^i = \xi^i_{t+\frac{1}{2}}$. While the equation $\xi_t^i = \xi^i_{t+\frac{1}{2}}$ cannot hold for "two-call" methods that also evaluate at $\mathbf{X}_t$ (e.g., EG), it is a valid notation for the "single-call" methods that only evaluate at $\mathbf{X}_{t+\frac{1}{2}}$ (e.g., OG). For this latter case, it will also be convenient to write $\mathcal{F}_t$ for $\mathcal{F}_{t+\frac{1}{2}}$ and $\mathbb{E}_t$ for $\mathbb{E}_{t+\frac{1}{2}}$.

7.1.2 *Non-convergence of EG and OG with Stochastic Feedback*

To ensure convergence of algorithms even in the face of stochasticity, it is often required to take a learning rate sequence that is *square summable but not summable*, i.e., $\sum_{t=1}^{+\infty}\eta_t = +\infty$ and $\sum_{t=1}^{+\infty}\eta_t^2 < +\infty$ [233, 276]. Nonetheless, in the following example, we show that even with such learning rate and even if the

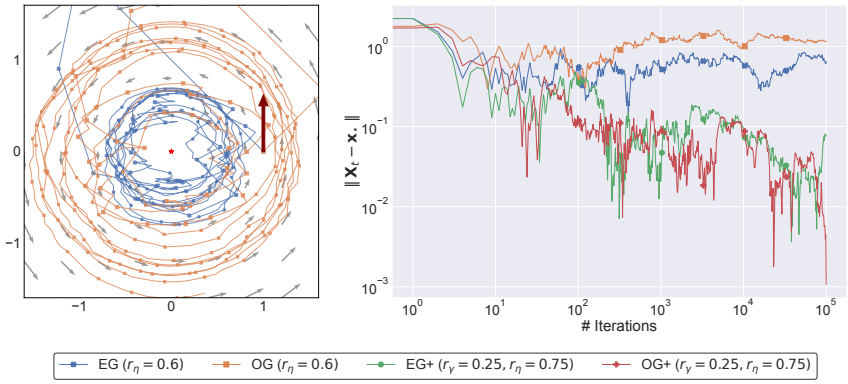


Figure 7.1: Trajectories of play (left) and distances to equilibrium (right) of (EG), (OG), (EG+), and (OG+) when they are run on the game $\min_{\theta\in\mathbb{R}} \max_{\phi\in\mathbb{R}} \theta\phi$ with stochastic feedback presented in Example 7.1. The learning rates are $\gamma_t = 1/(t+1)^{r_\gamma}$ and $\eta_t = 1/(t+1)^{r_\eta}$. For sake of readability, we only plot the trajectories of (EG) and (OG), and we only show the results for the iterates $(\mathbf{X}_t)_{t\in\mathbb{N}}$ but we observe the same qualitative convergence behaviors for $(\mathbf{X}_{t+\frac{1}{2}})_{t\in\mathbb{N}}$.

noise is almost surely bounded, neither (EG) nor (OG) with stochastic feedback converges in the bilinear zero-sum game we defined in Example 6.1.

Non-convergence under stochastic feedback

Example 7.1. Consider the following bilinear zero-sum game with player variables θ and ϕ.

$$\mathcal{X}^1 = \mathcal{X}^2 = \mathbb{R}, \quad \ell^1(\theta,\phi) = \theta\phi = -\ell^2(\theta,\phi). \tag{7.3}$$

If the feedback of the first player is perturbed by noise ξ_t^1 that takes value 1 and -1 with probability $1/2$ for each, then, as shown in Fig. 7.1, even with $\eta_t = 1/t^{0.6}$ which satisfies the square-summable-but-not-summable rule, the iterates of the neither (EG) nor (OG) (with the same learning rate for the two players) converges to the unique Nash equilibrium $(0,0)$.[2]

The above negative result suggests that, to cope with stochasticity of the feedback, further modification would be needed to stabilize the methods. With this in mind, we propose a simple fix in the next section to address this issue.

7.2 LEARNING RATE SEPARATION AND ENERGY INEQUALITIES

The goal of this section is to introduce EG+ and OG+, our backbone algorithms that enable last-iterate convergence when the feedback is corrupted by noise. Through a series of energy inequalities, we further elucidate how the incorporated learning rate separation mechanism assists in reducing the detrimental effects of noise.

7.2.1 *Learning Rate Separation as a Remedy: EG+ and OG+*

Viewed abstractly, the failure of (EG) and (OG) in the face of noisy feedback should be attributed to its inability of separating noise from the gradient variation $\|V^i(\mathbf{X}_t) - V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2$ (or $\|V^i(\mathbf{X}_{t-\frac{1}{2}}) - V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2$ for (OG)). In fact, in a noisy environment, the two consecutive pieces of feedback are only close

2 In [126], we formally show that (EG) with *any* learning rate sequence does not converge in Example 7.1. This indicates that the non-convergence observed here is inherent to the algorithm and not contingent on the choice of learning rate.

in expectation, so a player can only exploit this similarity when the noise is mitigated appropriately.

Intuition for learning rate separation

To overcome this difficulty, let us recall the interpretation of optimistic methods as approximate projection onto hyperplane that we present in Section 5.3.3. At each iteration, we move from $\mathbf{X}_t$ toward a separating hyperplane in the direction of $\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})$, and the distance between the hyperplane and the current iterate is in the order of $\eta_t \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|$. However, in the stochastic case we only have access to $\hat{\mathbf{V}}_{t+\frac{1}{2}}$, a stochastic estimate of $\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})$. Consequently, as suggested by the stochastic approximation literature, if we want the distance between the iterate and the hyperplane to go to 0, we cannot take the full step $\eta_t \hat{\mathbf{V}}_{t+\frac{1}{2}}$ but instead we need to take a more conservative step $\alpha_t \eta_t \hat{\mathbf{V}}_{t+\frac{1}{2}}$ for some α_t that goes to 0.

Learning rate separation

In other words, we should use different learning rates for the two steps of an iteration. This brings us to a strategy already explored in Chapter 4 in the context of online learning with delayed feedback: taking an optimistic step that is more aggressive than the update step. To distinguish from the standard optimistic methods with single learning rate sequence, we refer to the resulting methods as EG+ and OG+. They are defined with respect to two sequences of learning rates $(\gamma_t^i)_{t\in\mathbb{N}}$ and $(\eta_t^i)_{t\in\mathbb{N}}$ as follows (recall that $\hat{V}_{\frac{1}{2}}^i = 0$).

$$X_{t+\frac{1}{2}}^i = X_t^i - \gamma_t^i \hat{V}_t^i, \quad X_{t+1}^i = X_t^i - \eta_{t+1}^i \hat{V}_{t+\frac{1}{2}}^i. \tag{EG+}$$

$$X_{t+\frac{1}{2}}^i = X_t^i - \gamma_t^i \hat{V}_{t-\frac{1}{2}}^i, \quad X_{t+1}^i = X_t^i - \eta_{t+1}^i \hat{V}_{t+\frac{1}{2}}^i. \tag{OG+}$$

Alternative formulation of (OG+)

The above formulation effectively demonstrates the close relationship between (EG+) and (OG+). Moreover, similar to what we have shown in Section 4.1, with $x_t^i = X_{t+\frac{1}{2}}^i$ and $g_t^i = \hat{V}_{t+\frac{1}{2}}^i$, (OG+) can be written in the following form that directly relates the consecutive actions taken by a player.

$$x_{t+1}^i = x_t^i - (\gamma_{t+1}^i + \eta_{t+1}^i) g_t^i + \gamma_t^i g_{t-1}^i.$$

Throughout this chapter, we focus on the case where all the players take the same learning rates, i.e., $\gamma_t^i \equiv \gamma_t$ and $\eta_t^i \equiv \eta_t$ (the only exception being Lemma 7.5), an assumption that is later relaxed in Chapter 8 for the dual averaging variant. The key observation here is that by taking a larger optimistic step $\gamma_t^i > \eta_t^i$, the noise effectively becomes an order of magnitude smaller relative to the gradient variation. We demonstrate this via our energy inequalities in the remaining of this section. As a complement to this, we also provide empirical evidence in Fig. 7.1, showing that (EG+) and (OG+) indeed achieve convergence for the game and feedback model described in Example 7.1.

Double step-size extra-gradient: a note on naming

Remark 7.2. In our original paper [126], we used the name double step-size extra-gradient (DSEG) for the (EG+) algorithm. Later, the same algorithm was independently introduced by Diakonikolas et al. [65] under the name of EG+. The focus there was to solve minimax problems that admit a *weak Minty solution*. The name EG+ has since gain popularity in the literature. We follow this terminology, and in the similar spirit, use the names OG+ and OptDA+ (see Chapter 8) for the double-learning-rate variant of (OG) and of (OptDA).

7.2.2 *Generalized OG+*

Let us first analyze a general template that is run with two arbitrary sequences of vectors $(\tilde{g}_t)_{t\in\mathbb{N}}$ and $(g_t)_{t\in\mathbb{N}}$.

$$X_{t+\frac{1}{2}} = X_t - \gamma_t \tilde{g}_t, \quad X_{t+1} = X_t - \eta_{t+1} g_t. \tag{Generalized OG+}$$

We have the following preliminary result for this algorithm.

A basic decomposition for Generalized OG+

Proposition 7.2. *Let $(X_t)_{t\in\mathbb{N}}$ and $(X_{t+\frac{1}{2}})_{t\in\mathbb{N}}$ be generated by Generalized OG+. It holds for any $z \in \mathcal{X}$ and $t \in \mathbb{N}$ that*

$$\begin{aligned}\|X_{t+1} - z\|^2 = \|X_t - z\|^2 &- 2\eta_{t+1}\langle g_t, X_{t+\frac{1}{2}} - z\rangle \\ &- 2\gamma_t\eta_{t+1}\langle g_t, \tilde{g}_t\rangle + (\eta_{t+1})^2\|g_t\|^2.\end{aligned}$$

Proof. We develop directly

$$\begin{aligned}\|X_{t+1} - z\|^2 &= \|X_t - \eta_{t+1}g_t - z\|^2 \\ &= \|X_t - z\|^2 - 2\eta_{t+1}\langle g_t, X_t - z\rangle + (\eta_{t+1})^2\|g_t\|^2 \\ &= \|X_t - z\|^2 - 2\eta_{t+1}\langle g_t, X_{t+\frac{1}{2}} - z\rangle \\ &\quad - 2\gamma_t\eta_{t+1}\langle g_t, \tilde{g}_t\rangle + (\eta_{t+1})^2\|g_t\|^2,\end{aligned}$$

where in the last equality we use the fact that $X_t = X_{t+\frac{1}{2}} + \gamma_t\tilde{g}_t$. □

Proposition 7.2 is nothing but an elementary decomposition that relates two consecutive distance measures $\|X_{t+1} - z\|^2$ and $\|X_t - z\|^2$. In standard analysis one would proceed with

$$-2\langle g_t, \tilde{g}_t\rangle = \|g_t - \tilde{g}_t\|^2 - \|g_t\|^2 - \|\tilde{g}_t\|^2.$$

This gives rise to the approximation error term $\|g_t - \tilde{g}_t\|^2$ that shows up in all of our analysis of optimistic methods so far (see e.g., Proposition 4.1). Nonetheless, when the feedback is noisy, it encompasses both the gradient variation and the noise, making it impossible to cancel out the negative effect of noise. In view of this issue, we take a different approach, and show through careful analysis that the noise is actually an order smaller than the gradient variation in this scalar product term.

7.2.3 *Inequalities for EG+*

We start with the analysis of (EG+). We work directly with the global update and for ease of notation we write L_g for the global Lipschitz constant so that $\|\mathbf{V}(\mathbf{x}) - \mathbf{V}(\mathbf{x}')\| \leq L_g\|\mathbf{x} - \mathbf{x}'\|$.

Lemma 7.3. *Suppose that Assumptions 5.2 and 7.1 hold and all players run (EG+) with the same learning rates. Then, for all $i \in \mathcal{N}$ and $t \in \mathbb{N}$, it holds*

$$\begin{aligned}-2\,\mathbb{E}_t[\langle \hat{\mathbf{V}}_{t+\frac{1}{2}}, \hat{\mathbf{V}}_t\rangle] \leq \mathbb{E}_t\Big[&-\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 - \|\mathbf{V}(\mathbf{X}_t)\|^2 \\ &+ \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}) - \mathbf{V}(\mathbf{X}_t)\|^2 + 2\gamma_t L_g\|\xi_t\|^2\Big]\end{aligned} \tag{7.4}$$

Proof. The estimate $\hat{\mathbf{V}}_t$ being $\mathcal{F}_{t+\frac{1}{2}}$ measurable, applying the law of total expectation gives

$$\begin{aligned}\mathbb{E}_t[\langle \hat{\mathbf{V}}_{t+\frac{1}{2}}, \hat{\mathbf{V}}_t\rangle] &= \mathbb{E}_t[\langle \mathbb{E}_{t+\frac{1}{2}}[\hat{\mathbf{V}}_{t+\frac{1}{2}}], \hat{\mathbf{V}}_t^i\rangle] \\ &= \mathbb{E}_t[\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \hat{\mathbf{V}}_t\rangle] \\ &= \mathbb{E}_t[\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{V}(\mathbf{X}_t)\rangle + \langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \boldsymbol{\xi}_t\rangle]. \end{aligned} \tag{7.5}$$

We rewrite the first term as

$$2\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{V}(\mathbf{X}_t)\rangle = \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 + \|\mathbf{V}(\mathbf{X}_t)\|^2 - \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}) - \mathbf{V}(\mathbf{X}_t)\|^2. \tag{7.6}$$

Regarding the second term, we define

$$\tilde{\mathbf{X}}_{t+\frac{1}{2}} = \mathbf{X}_t - \gamma_t \mathbf{V}(\mathbf{X}_t) = \mathbf{X}_{t+\frac{1}{2}} + \gamma_t \boldsymbol{\xi}_t.$$

This is the leading state that we would obtain if the feedback were not noisy. Note that $\tilde{\mathbf{X}}_{t+\frac{1}{2}}$ is $\mathcal{F}_t$-measurable, and hence

$$\mathbb{E}_t[\langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}), \boldsymbol{\xi}_t\rangle] = \langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}), \mathbb{E}_t[\boldsymbol{\xi}_t]\rangle = 0.$$

It then follows from the Lipschitz continuity of $\mathbf{V}$ that

$$\begin{aligned}\mathbb{E}_t[-\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \boldsymbol{\xi}_t\rangle] &= \mathbb{E}_t[-\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}) - \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}), \boldsymbol{\xi}_t\rangle] \\ &\leq \mathbb{E}_t[L_g \|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_{t+\frac{1}{2}}\| \|\boldsymbol{\xi}_t\|] \\ &= \mathbb{E}_t[\gamma_t L_g \|\boldsymbol{\xi}_t\|^2] \end{aligned} \tag{7.7}$$

Putting (7.5), (7.6), and (7.7) together gives the desired inequality. □

In Lemma 7.3, we effectively separate the gradient variation $\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}) - \mathbf{V}(\mathbf{X}_t)\|^2$ from the noise term $2\gamma_t L_g \|\boldsymbol{\xi}_t\|^2$, which scales in γ_t and can thus be made arbitrarily small by decreasing γ_t. Along with Proposition 7.2, we then deduce immediately the benefit of using two different learning rate sequences, as suggested by the following lemma.

Energy inequality for EG+

Lemma 7.4. *Suppose that Assumptions 5.2, 5.3 and 7.1 hold and all players run (EG+) with the same learning rates. Then, for all $t \in \mathbb{N}$ and $\mathbf{x}_\star \in \mathfrak{X}_\star$, we have*

$$\begin{aligned}\mathbb{E}_t[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2] \leq \|\mathbf{X}_t - \mathbf{x}_\star\|^2 &- \gamma_t\eta_{t+1}(1 - a_t(1+\sigma_M^2) - b_t\sigma_M^2)\|\mathbf{V}(\mathbf{X}_t)\|^2 \\ &- \gamma_t\eta_{t+1}\left(1 - \frac{\eta_{t+1}(1+\sigma_M^2)}{\gamma_t}\right)\mathbb{E}_t[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2] \\ &+ \gamma_t\eta_{t+1}\left(\frac{\eta_{t+1}}{\gamma_t} + a_t + b_t\right)N\sigma_A^2, \end{aligned} \tag{7.8}$$

where $a_t = \gamma_t^2 L_g^2$ and $b_t = 2\gamma_t L_g$.

Proof. We apply Proposition 7.2 to the global update with $X_t \leftarrow \mathbf{X}_t$ and $z \leftarrow \mathbf{x}_\star$. Since the inequality holds for any realization, we can take expectation with respect to $\mathcal{F}_t$ to get

$$\begin{aligned}\mathbb{E}_t[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2] = \mathbb{E}_t[\|\mathbf{X}_t - \mathbf{x}_\star\|^2 &- 2\eta_{t+1}\langle \hat{\mathbf{V}}_{t+\frac{1}{2}}, \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star\rangle \\ &- 2\gamma_t\eta_{t+1}\langle \hat{\mathbf{V}}_{t+\frac{1}{2}}, \hat{\mathbf{V}}_t\rangle + (\eta_{t+1})^2\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|^2]. \end{aligned}$$

On one hand, it follows from Assumptions 5.3 and 7.1 that

$$\mathbb{E}_t[\eta_{t+1}\langle \hat{\mathbf{V}}_{t+\frac{1}{2}}, \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star\rangle] = \mathbb{E}_t[\eta_{t+1}\langle \mathbf{V}(X_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star\rangle] \geq 0.$$

On the other hand, the scalar product term $\mathbb{E}_t[-2\gamma_t\eta_{t+1}\langle \hat{\mathbf{V}}_{t+\frac{1}{2}}, \hat{\mathbf{V}}_t\rangle]$ can be bounded from above thanks to Lemma 7.3. Moreover, with Assumption 5.2 Lipschitz continuity of $\mathbf{V}$ and Assumption 7.1 on the noise (using inequality (7.2) precisely), we deduce that

$$\begin{aligned}\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}) - \mathbf{V}(\mathbf{X}_t)\|^2 &\leq L^2\|\mathbf{X}_{t+\frac{1}{2}} - \mathbf{X}_t\|^2 \\ &= \gamma_t^2 L_g^2 \|\hat{\mathbf{V}}_t\|^2 \\ &\leq \gamma_t^2 N L_g^2 \sigma_A^2 + \gamma_t^2 L_g^2 (1+\sigma_M^2)\|\mathbf{V}(\mathbf{X}_t)\|^2.\end{aligned}$$

We further bound all the noise terms with help of (7.1) and (7.2). This gives

$$\begin{aligned}\mathbb{E}_t[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2] \leq \|\mathbf{X}_t - \mathbf{x}_\star\|^2 &- \gamma_t\eta_{t+1}\left(1 - \frac{\eta_{t+1}(1+\sigma_M^2)}{\gamma_t}\right)\mathbb{E}_t[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2] \\ &- \gamma_t\eta_{t+1}(1 - 2\gamma_t L_g\sigma_M^2 - \gamma_t^2 L_g^2(1+\sigma_M^2))\|\mathbf{V}(\mathbf{X}_t)\|^2 \\ &+ (\gamma_t^3\eta_{t+1}L_g^2 + 2\gamma_t^2\eta_{t+1}L_g + \eta_{t+1}^2)N\sigma_A^2.\end{aligned}$$

This is exactly (7.8). □

Analyzing the bound of Lemma 7.4 term-by-term gives a clear picture of how aggressive exploration (i.e., taking a larger optimistic step) can be helpful:

- The term $-\gamma_t\eta_{t+1}(1 - a_t(1+\sigma_M^2) - b_t\sigma_M^2)\|\mathbf{V}(\mathbf{X}_t)\|^2$ provides a consistently negative contribution as long as γ_t is small enough.
- Similarly, the term $-\gamma_t\eta_{t+1}(1 - \eta_{t+1}(1+\sigma_M^2)/\gamma_t)\,\mathbb{E}_t[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2]$ is negative as long as $\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}) \neq 0$ and η_{t+1} is sufficiently small compared to γ_t.
- The term $\gamma_t\eta_{t+1}(\eta_{t+1}/\gamma_t + a_t + b_t)N\sigma_A^2$ is antagonistic and needs to be made as small as possible.

In summary, the usefulness of the scale separation between the exploration and the update mechanisms appears in two different places. First, it ensures the difference $1 - \eta_{t+1}(1+\sigma_M^2)/\gamma_t$ to be non-negative; this is relevant for *multiplicative* noise. Moreover, it also ensures the coefficient η_{t+1}^2 that appears in the last term is small compared to the coefficient $\gamma_t\eta_{t+1}$ of the negative terms; this is relevant for *additive* noise. Finally, the coefficients $\gamma_t^2\eta_{t+1}$ and $\gamma_t^3\eta_{t+1}$ also appear in the last term, and therefore γ_t would also need to be decreasing whenever $\sigma_A > 0$.

7.2.4 *Inequalities for OG+*

We proceed to establish the energy inequality of (OG+). For this, we first prove the counterpart of Lemma 7.3 for (OG+).

Lemma 7.5. *Suppose that Assumptions 5.2 and 7.1 hold and all players run* (OG+). *Then, for all $i \in \mathcal{N}$ and $t \geq 2$, it holds*

$$\begin{aligned}-2\,\mathbb{E}_{t-1}[\langle \hat{V}^i_{t+\frac{1}{2}}, \hat{V}^i_{t-\frac{1}{2}}\rangle] \leq \mathbb{E}_{t-1}\Big[&-\|V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2 - \|V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\ &+ \|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2\end{aligned}$$

$$+ L\left(2\gamma_t^i\sqrt{N}\|\xi^i_{t-\frac{1}{2}}\|^2 + \sum_{j=1}^{N}\frac{(\gamma_t^j+\eta_t^j)^2\|\xi^j_{t-\frac{1}{2}}\|^2}{2\sqrt{N}\gamma_t^i}\right)\Bigg]$$

Proof. The proof below follows closely that of Lemma 7.3. To begin, we apply the law of total expectation to get

$$\begin{aligned}\mathbb{E}_{t-1}[\langle\hat{V}^i_{t+\frac{1}{2}},\hat{V}^i_{t-\frac{1}{2}}\rangle] &= \mathbb{E}_{t-1}[\langle\mathbb{E}_t[\hat{V}^i_{t+\frac{1}{2}}],\hat{V}^i_{t-\frac{1}{2}}\rangle]\\ &= \mathbb{E}_{t-1}[\langle V^i(\mathbf{X}_{t+\frac{1}{2}}),\hat{V}^i_{t-\frac{1}{2}}\rangle]\\ &= \mathbb{E}_{t-1}[\langle V^i(\mathbf{X}_{t+\frac{1}{2}}),V^i(\mathbf{X}_{t-\frac{1}{2}})\rangle + \langle V^i(\mathbf{X}_{t+\frac{1}{2}}),\xi^i_{t-\frac{1}{2}}\rangle]. \qquad (7.9)\end{aligned}$$

We reformulate the first term as

$$2\langle V^i(\mathbf{X}_{t+\frac{1}{2}}),V^i(\mathbf{X}_{t-\frac{1}{2}})\rangle = \|V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2 + \|V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 - \|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2. \qquad (7.10)$$

Moving on to the second term, for all $j \in \mathcal{N}$, we define

$$\tilde{X}^j_{t+\frac{1}{2}} = X^j_{t+\frac{1}{2}} + (\gamma_t^j+\eta_t^j)\xi^j_{t-\frac{1}{2}}.$$

Similar to before, this serves as as a surrogate for $X^j_{t+\frac{1}{2}}$ and is obtained by removing the noise of round $t-1$. Equivalently, we can write

$$\tilde{X}^j_{t+\frac{1}{2}} = X^j_{t-1} - (\gamma_t^j+\eta_t^j)V^j(\mathbf{X}_{t-\frac{1}{2}}).$$

This shows that $\tilde{\mathbf{X}}_{t+\frac{1}{2}}$ is $\mathcal{F}_{t-1}$-measurable and hence

$$\mathbb{E}_{t-1}[\langle V^i(\tilde{\mathbf{X}}_{t+\frac{1}{2}}),\xi^i_{t-\frac{1}{2}}\rangle] = \langle V^i(\tilde{\mathbf{X}}_{t+\frac{1}{2}}),\mathbb{E}_{t-1}[\xi^i_{t-\frac{1}{2}}]\rangle = 0.$$

Moreover, by definition of $\tilde{\mathbf{X}}_{t+\frac{1}{2}}$, we have

$$\|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_{t+\frac{1}{2}}\|^2 = \sum_{j=1}^{N}\|X^j_{t+\frac{1}{2}} - \tilde{X}^j_{t+\frac{1}{2}}\|^2 = \sum_{j=1}^{N}(\gamma_t^j+\eta_t^j)^2\|\xi^j_{t-\frac{1}{2}}\|^2$$

Exploiting the Lipschitz continuity of V^i we derive that

$$\begin{aligned}\mathbb{E}_{t-1}[-\langle V^i(\mathbf{X}_{t+\frac{1}{2}}),\xi^i_{t-\frac{1}{2}}\rangle] &= \mathbb{E}_{t-1}[-\langle V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\tilde{\mathbf{X}}_{t+\frac{1}{2}}),\xi^i_{t-\frac{1}{2}}\rangle]\\ &\quad - \mathbb{E}_{t-1}[\langle V^i(\tilde{\mathbf{X}}_{t+\frac{1}{2}}),\xi^i_{t-\frac{1}{2}}\rangle]\\ &\le \mathbb{E}_{t-1}[L\|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_{t+\frac{1}{2}}\|\|\xi^i_{t-\frac{1}{2}}\|]\\ &\le \mathbb{E}_{t-1}\left[L\left(\frac{\|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_{t+\frac{1}{2}}\|^2}{4\gamma_t^i\sqrt{N}} + \gamma_t^i\sqrt{N}\|\xi^i_{t-\frac{1}{2}}\|^2\right)\right]\\ &= \mathbb{E}_{t-1}\left[L\left(\gamma_t^i\sqrt{N}\|\xi^i_{t-\frac{1}{2}}\|^2 + \sum_{j=1}^{N}\frac{(\gamma_t^j+\eta_t^j)^2\|\xi^j_{t-\frac{1}{2}}\|^2}{4\sqrt{N}\gamma_t^i}\right)\right]. \qquad (7.11)\end{aligned}$$

We conclude by combining (7.9), (7.10), and (7.11). □

Compared to Lemma 7.3, in Lemma 7.5 we state the inequality in terms of the feedback received by each individual player. We also take into account the possibility of using player-dependent learning rates, as we will show in Chapter 8 that the same inequality also applies to the dual averaging variant of the algorithm (OptDA+), for which we consider player-dependent learning rates in the analysis. Except for these differences, the result is much similar to that of Lemma 7.3. We manage to separate the gradient variation from the noise term which at most scales in $\max_{j\in\mathcal{N}}(\gamma_t^j)^2/\eta_t^i$ (assume that $\eta_t^j \leq \gamma_t^j$ for all j).

With Lemma 7.5 we are ready to prove our energy inequalities. We turn our attention back to player-independent learning rates and we first derive a *local* version of the inequality.

Individual energy inequality for OG+

Lemma 7.6. *Suppose that Assumptions 5.2 and 7.1 hold and all players run (OG+) with the same learning rate sequences. Then, for all $i \in \mathcal{N}$, $t \geq 2$, and $z^i \in \mathcal{X}^i$, it holds*

$$\begin{aligned}\mathbb{E}_{t-1}[\|X_{t+1}^i - z^i\|^2] \leq \mathbb{E}_{t-1}[&\|X_t^i - z^i\|^2 - 2\eta_{t+1}\langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}}^i - z^i\rangle \\ &- \gamma_t\eta_{t+1}(1 - 2\gamma_t\sqrt{N}L\sigma_M^2)\|V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\ &- \gamma_t\eta_{t+1}\left(1 - \frac{\eta_{t+1}(1+\sigma_M^2)}{\gamma_t}\right)\|V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2 \\ &+ \left(a_t L^2(1+\sigma_M^2) + \frac{b_t L\sigma_M^2}{\sqrt{N}}\right)\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\ &+ 3(\gamma_{t-1})^2\gamma_t\eta_{t+1}L^2\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2 \\ &+ (a_t NL^2 + (2\gamma_t^2\eta_{t+1} + b_t)\sqrt{N}L + (\eta_{t+1})^2)\sigma_A^2],\end{aligned}$$

where $a_t = 3\gamma_t^3\eta_{t+1} + 3\gamma_t\eta_t^2\eta_{t+1}$ and $b_t = (\gamma_t + \eta_t)^2\eta_{t+1}/2$.

Proof. Applying Proposition 7.2 to player i's update with $X_t \leftarrow X_t^i$ and $z \leftarrow z^i$ and taking expectation with respect to $\mathcal{F}_{t-1}$ gives

$$\begin{aligned}\mathbb{E}_{t-1}[\|X_{t+1}^i - z^i\|^2] = \mathbb{E}_{t-1}[&\|X_t^i - z^i\|^2 - 2\eta_{t+1}\langle \hat{V}_{t+\frac{1}{2}}^i, X_{t+\frac{1}{2}}^i - z^i\rangle \\ &- 2\gamma_t\eta_{t+1}\langle \hat{V}_{t+\frac{1}{2}}^i, \hat{V}_{t-\frac{1}{2}}^i\rangle + (\eta_{t+1})^2\|\hat{V}_{t+\frac{1}{2}}^i\|^2].\end{aligned}$$

The unbiasedness of noise Assumption 7.1(a) implies

$$\mathbb{E}_{t-1}[\eta_{t+1}\langle \hat{V}_{t+\frac{1}{2}}^i, X_{t+\frac{1}{2}}^i - z^i\rangle] = \eta_{t+1}\,\mathbb{E}_{t-1}[\langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}}^i - z^i\rangle]. \tag{7.12}$$

Invoking Lemma 7.5, we then obtain

$$\begin{aligned}\mathbb{E}_{t-1}[\|X_{t+1}^i - z^i\|^2] \leq \mathbb{E}_{t-1}\Big[&\|X_t^i - z^i\|^2 - 2\eta_{t+1}\langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}}^i - z^i\rangle \\ &- \gamma_t\eta_{t+1}(\|V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2 + \|V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2) \\ &+ \gamma_t\eta_{t+1}\|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\ &+ 2\gamma_t^2\eta_{t+1}\sqrt{N}L\|\xi_{t-\frac{1}{2}}^i\|^2 \\ &+ \frac{(\gamma_t+\eta_t)^2\eta_{t+1}L}{2\sqrt{N}}\|\xi_{t-\frac{1}{2}}\|^2 + (\eta_{t+1})^2\|\hat{V}_{t+\frac{1}{2}}^i\|^2\Big].\end{aligned} \tag{7.13}$$

We proceed to bound the variation $\|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2$ by

$$\begin{aligned}\|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 &\leq 3\|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\mathbf{X}_t)\|^2 + 3\|V^i(\mathbf{X}_t) - V^i(\mathbf{X}_{t-1})\|^2 \\ &\quad + 3\|V^i(\mathbf{X}_{t-1}) - V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\ &\leq 3\gamma_t^2 L^2 \|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2 + 3\eta_t^2 L^2 \|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2 + 3(\gamma_{t-1})^2 L^2 \|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2.\end{aligned}$$

To conclude, we plug this into (7.13) and bound the noise terms with Assumption 7.1. □

From Lemma 7.6 it is straightforward to obtain the global energy inequality.

Global energy inequality for OG+

Lemma 7.7. *Suppose that Assumptions 5.2, 5.3 and 7.1 hold and all players run (OG+). Then, for all $t \geq 2$ and $\mathbf{x}_\star \in \mathfrak{X}_\star$, it holds*

$$\begin{aligned}\mathbb{E}_{t-1}[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2] &\leq \mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2] + 3(\gamma_{t-1})^2 \gamma_t \eta_{t+1} N L^2 \|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2 \\ &\quad - \gamma_t \eta_{t+1}(1 - a_t(1+\sigma_M^2) - b_t \sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\ &\quad - \gamma_t \eta_{t+1}\left(1 - \frac{\eta_{t+1}(1+\sigma_M^2)}{\gamma_t}\right)\mathbb{E}_{t-1}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2] \\ &\quad + \gamma_t \eta_{t+1}\left(\frac{\eta_{t+1}}{\gamma_t} + a_t + b_t\right) N \sigma_A^2. \qquad (7.14)\end{aligned}$$

where $a_t = 3(\gamma_t^2 + \eta_t^2)NL^2$ and $b_t = (3\gamma_t + \eta_t^2/\gamma_t)\sqrt{N}L$.

Proof. We first apply the individual quasi-descent inequality Lemma 7.6 to $z^i \leftarrow x_\star^i$ and sum from $i = 1$ to N to obtain

$$\begin{aligned}\mathbb{E}_{t-1}[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2] \leq \mathbb{E}_{t-1}\Bigg[&\|\mathbf{X}_t - \mathbf{x}_\star\|^2 - 2\eta_{t+1}\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star\rangle \\ &- \gamma_t \eta_{t+1}(1 - 2\gamma_t \sqrt{N} L \sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\ &- \gamma_t \eta_{t+1}\left(1 - \frac{\eta_{t+1}(1+\sigma_M^2)}{\gamma_t}\right)\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 \\ &+ 3\gamma_t \eta_{t+1}(\gamma_t^2 + \eta_t^2) N L^2 (1+\sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\ &+ \frac{(\gamma_t + \eta_t)^2 \eta_{t+1} \sqrt{N} L \sigma_M^2}{2}\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\ &+ 3(\gamma_{t-1})^2 \gamma_t \eta_{t+1} N L^2 (1+\sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t-\frac{3}{2}})\|^2 \\ &+ (3\gamma_t \eta_{t+1}((\gamma_{t-1})^2 + \gamma_t^2 + \eta_t^2) N L^2 + (\eta_{t+1})^2 N \sigma_A^2 \\ &+ \left(2\gamma_t^2 \eta_{t+1} + \frac{(\gamma_t + \eta_t)^2 \eta_{t+1}}{2}\right)\sqrt{N} L N \sigma_A^2\Bigg].\end{aligned}$$

To get (7.14), we drop the scalar product $\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star\rangle$ which is non-negative by Assumption 5.3 and use the bound $(\gamma_t + \eta_t)^2 \leq 2\gamma_t^2 + 2\eta_t^2$. □

In Lemma 7.7 we recognize the terms that already appear in Lemma 7.4, albeit with a different definition for a_t and b_t, and with $\mathbf{X}_{t-\frac{1}{2}}$ replacing $\mathbf{X}_t$. In fact, a_t and b_t are still in the same order if $\eta_t \leq \gamma_t$. While the analysis becomes more complex due to our conditioning on the σ-algebra $\mathcal{F}_{t-1}$ of time $t-1$ and the presence of the additional term $3(\gamma_{t-1})^2 \gamma_t \eta_{t+1} N L^2 \|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2$, these factors do not really interfere the convergence of the algorithm.

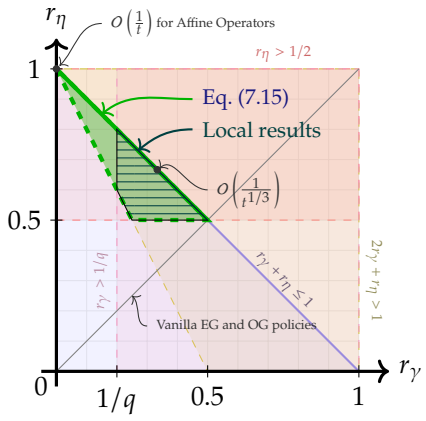


Figure 7.2: The stepsize exponents allowed by condition (7.15) for convergence (shaded green). Dashed lines are strict frontiers. Note that vanilla (EG) and (OG) (the separatrix $r_\eta = r_\gamma$) passes just outside of this region, explaining the methods' failures.

7.3 GLOBAL CONVERGENCE

In this section we present our main convergence results for (EG+) and for (OG+). Since we are in the unconstrained setup, with the convexity assumption (Assumption 5.1), $\mathbf{x}_\star$ is a Nash equilibrium if and only if $\mathbf{V}(\mathbf{x}_\star) = 0$. We thus focus on proving that our algorithm converges to a point $\mathbf{x}_\star$ such that $\mathbf{V}(\mathbf{x}_\star) = 0$.

7.3.1 *Asymptotic Convergence*

Robbins-Monro-like learning rate condition

As demonstrated in the energy inequalities of (EG+) and of (OG+) (Lemmas 7.4 and 7.7), $\gamma_t\eta_{t+1}$ should be as *large* as possible to ensure negative shift while $\gamma_t^2\eta_{t+1}$, $\gamma_t^3\eta_{t+1}$, and η_{t+1}^2 should be as *small* as possible to reduce the impact of the *additive* noise. This leads to the following requirement on the learning rates.

$$\sum_{t=1}^{+\infty} \gamma_t\eta_{t+1} = +\infty, \quad \sum_{t=1}^{+\infty} \gamma_t^2\eta_{t+1} < +\infty, \quad \sum_{t=1}^{+\infty} \eta_t^2 < +\infty. \tag{7.15}$$

Candidate learning rates

The objective of this subsection is to provide asymptotic convergence results for learning rate sequences that satisfy the above condition. It essentially posits that $\gamma_t \to 0$ and $\gamma_t/\eta_t \to 0$ as $t \to \infty$, and rules out the choice $\gamma_t = \eta_t$ which would yield the vanilla (EG) and (OG) algorithms. For further illustration, let us consider the following learning rate policy

$$\gamma_t = \frac{\gamma}{(t+\beta)^{r_\gamma}} \quad \text{and} \quad \eta_t = \frac{\eta}{(t+\beta)^{r_\eta}}$$

for some constants $\gamma, \eta, \beta > 0$ and exponents $r_\gamma, r_\eta \in [0, 1]$. Condition (7.15) then translates into $r_\gamma + r_\eta \leq 1$, $2r_\eta > 1$, and $2r_\gamma + r_\eta > 1$ as represented in Fig. 7.2. On the other hand, when the noise is *multiplicative*, i.e., $\sigma_A = 0$, (7.15) is no longer necessary and we only need $(1 + \sigma_M^2)\eta_{t+1} \leq \gamma_t$ instead. We will show almost sure convergence for this specific case in Chapter 8 (cf. Theorem 8.13).

Regarding the analyses, our proofs build upon the *stochastic quasi-Fejér monotonicity* [48, 49] of the iterates that can be derived from the energy inequalities and the learning rate conditions. With this, we then use the Robbins–Siegmund theorem [234] for almost-supermartingales to prove the almost sure convergence of the distance $\|\mathbf{X}_t - \mathbf{x}_\star\|$ to a finite random variable, that is, a random variable

that is finite almost surely. Given the importance of this argument, we provide the full statement of the Robbins–Siegmund theorem below.

Robbins–Siegmund theorem

Lemma 7.8 (Robbins and Sigmund [234]). *Consider a filtration $(\mathscr{G}_t)_{t\in\mathbb{N}}$ and four non-negative real-valued $(\mathscr{G}_t)_{t\in\mathbb{N}}$-adapted processes $(U_t)_{t\in\mathbb{N}}$, $(\alpha_t)_{t\in\mathbb{N}}$, $(\chi_t)_{t\in\mathbb{N}}$, $(\zeta_t)_{t\in\mathbb{N}}$ such that*

$$\mathbb{E}[U_1] < +\infty, \quad \left\| \prod_{t=1}^{+\infty} (1+\alpha_t) \right\|_\infty < +\infty, \quad \sum_{t=1}^{+\infty} \mathbb{E}[\chi_t] < +\infty,$$

and for all $t \in \mathbb{N}$,

$$\mathbb{E}[U_{t+1} \mid \mathscr{G}_t] \leq (1+\alpha_t)U_t + \chi_t - \zeta_t. \tag{7.16}$$

Then,

(a) $(U_t)_{t\in\mathbb{N}}$ converges almost surely to a finite random variable $U_\infty \in L^1$.

(b) $\sum_{t=1}^{+\infty} \mathbb{E}[\zeta_t] < +\infty$ and accordingly $\sum_{t=1}^{+\infty} \zeta_t < +\infty$ almost surely.

Remark 7.3. There exists another version of Robbins–Siegmund theorem with both weaker assumptions and weaker results. It focuses on almost sure convergence of the variables and does not deal with the integrability of the limits.

As we can see from the statement of the Robbins–Siegmund theorem, one of the required conditions is $\mathbb{E}[U_1] < +\infty$. This translates to the following assumption concerning the initialization of the algorithms.

Assumption on initialization: bounded second-order moment

Assumption 7.2. The algorithms of the players are initialized at points with finite second-moment, i.e., $\mathbb{E}[\|\mathbf{X}_1\|^2] < +\infty$.

As a direct consequence of Assumption 7.2, for any $\mathbf{x}_\star \in \mathfrak{X}_\star$ we have

$$\mathbb{E}[\|\mathbf{X}_1 - \mathbf{x}_\star\|^2] \leq 2\,\mathbb{E}[\|\mathbf{X}_1\|^2] + 2\,\mathbb{E}[\|\mathbf{x}_\star\|^2] < +\infty.$$

Besides the Robbins–Siegmund theorem, our proofs also utilize several other important lemmas for stochastic sequences. In order to maintain the flow of the main discussion, we defer the presentation of these lemmas to Appendix B.

CONVERGENCE OF EG+. We first establish the almost sure convergence of (EG+) to a Nash equilibrium in all variationally stable games under suitable learning rate condition.

Convergence of EG+

Theorem 7.9. *Suppose that Assumptions 5.1–5.3, 7.1 and 7.2 hold and all players run (EG+) with learning rate sequences $(\gamma_t)_{t\in\mathbb{N}}$ and $(\eta_t)_{t\in\mathbb{N}}$ satisfying (7.15) and*

$$\gamma_t \leq \min\left(\frac{1}{2L_g\sqrt{1+\sigma_M^2}}, \frac{1}{8L_g\sigma_M^2} \right), \quad \eta_{t+1} \leq \frac{\gamma_t}{1+\sigma_M^2} \quad \text{for all } t \in \mathbb{N}. \tag{7.17}$$

Then the iterate X_t converges almost surely to a Nash equilibrium.

Proof. The proof is divided into three steps.

(1) *With probability* 1, $\|\mathbf{X}_t - \mathbf{x}_\star\|$ *converges for all* $\mathbf{x}_\star \in \mathfrak{X}_\star$. Let $\mathbf{x}_\star \in \mathfrak{X}_\star$. Lemma 7.4 along with condition (7.17) gives

$$\begin{aligned}\mathbb{E}_t[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2] \leq \|\mathbf{X}_t - \mathbf{x}_\star\|^2 - \frac{\gamma_t\eta_{t+1}}{2}\|\mathbf{V}(\mathbf{X}_t)\|^2 \\ + \left((\eta_{t+1})^2 + \gamma_t^3\eta_{t+1}L_g^2 + 2\gamma_t^2\eta_{t+1}L_g\right) N\sigma_A^2.\end{aligned} \tag{7.18}$$

As for the rightmost term, it follows from the stepsize conditions $\sum_t \eta_t^2 < +\infty$, $\sum_t \gamma_t^2 \eta_{t+1} < \infty$, and $(\gamma_t)_{t\in\mathbb{N}}$ being upper-bounded that

$$\sum_{t=1}^{+\infty} (\eta_{t+1})^2 + \gamma_t^3 \eta_{t+1} L_g^2 + 2\gamma_t^2 \eta_{t+1} L_g < +\infty.$$

We can thus apply the Robbins–Siegmund theorem (Lemma 7.8) to inequality (7.18) with

$$\mathscr{G}_t \leftarrow \mathscr{F}_t, \quad U_t \leftarrow \|\mathbf{X}_t - \mathbf{x}_\star\|^2, \quad \alpha_t \leftarrow 0, \quad \zeta_t \leftarrow \frac{\gamma_t \eta_{t+1}}{2} \|\mathbf{V}(\mathbf{X}_t)\|^2,$$
$$\chi_t \leftarrow \left((\eta_{t+1})^2 + \gamma_t^3 \eta_{t+1} L_g^2 + 2\gamma_t^2 \eta_{t+1} L_g \right) N \sigma_A^2.$$

This gives that *(i)* $\sum_{t=1}^{+\infty} \gamma_t \eta_{t+1} \mathbb{E}[\|\mathbf{V}(\mathbf{X}_t)\|^2] < +\infty$ and *(ii)* $\|\mathbf{X}_t - \mathbf{x}_\star\|$ converges almost surely. As the seconds point holds true for all $\mathbf{x}_\star \in \mathfrak{X}_\star$, invoking Corollary B.7 we deduce that the event $\{\|\mathbf{X}_t - \mathbf{x}_\star\| \text{ converges for all } \mathbf{x}_\star \in \mathfrak{X}_\star\}$ happens with probability 1.

(2) *There exists an increasing function* $\omega \colon \mathbb{N} \to \mathbb{N}$ *such that* $\|\mathbf{V}(\mathbf{X}_{\omega(t)})\|^2$ *converges to* 0 *almost surely.* We have shown that $\sum_{t=1}^{+\infty} \gamma_t \eta_{t+1} \mathbb{E}[\|\mathbf{V}(\mathbf{X}_t)\|^2] < +\infty$. With $\sum_{t=1}^{+\infty} \gamma_t \eta_{t+1} = +\infty$ we then know that $\liminf \mathbb{E}[\|\mathbf{V}(X_t)\|^2] = 0$. Subsequently, we prove the claim with the help of Lemma B.5.

(3) *Conclude.* Let us define the event

$$\mathcal{E} = \{\|\mathbf{X}_t - \mathbf{x}_\star\| \textit{ converges for all } \mathbf{x}_\star \in \mathfrak{X}_\star;\ \|\mathbf{V}(\mathbf{X}_{\omega(t)})\|^2 \textit{ converges to } 0\}$$

We have $\mathbb{P}(\mathcal{E}) = 1$ from the two previous points. We next show that $(\mathbf{X}_t)_{t\in\mathbb{N}}$ converges to a point in $\mathfrak{X}_\star$ whenever $\mathcal{E}$ happens.

Let us take a realization of this event. Since $\mathfrak{X}_\star$ is non-empty, the convergence of $\|\mathbf{X}_t - \mathbf{x}_\star\|$ for a $\mathbf{x}_\star \in \mathfrak{X}_\star$ implies that $(\mathbf{X}_t)_{t\in\mathbb{N}}$ is bounded. $\mathbb{R}^d$ being finite-dimensional, we can then extract a subsequence of $(\mathbf{X}_{\omega(t)})_{t\in\mathbb{N}}$, which we denote by $(\mathbf{X}_{\omega(\psi(t))})_{t\in\mathbb{N}}$ that converges to a point $\mathbf{x}_\infty \in \mathbb{R}^d$. By continuity of $\mathbf{V}$, we have $\mathbf{V}(\mathbf{x}_\infty) = 0$, i.e., $\mathbf{x}_\infty \in \mathfrak{X}_\star$. Accordingly, with the definition of $\mathcal{E}$, we deduce the convergence of $\|\mathbf{X}_t - \mathbf{x}_\infty\|$, and

$$\lim_{t\to+\infty} \|\mathbf{X}_t - \mathbf{x}_\infty\| = \lim_{t\to+\infty} \|\mathbf{X}_{\omega(\psi(t))} - \mathbf{x}_\infty\| = \|\mathbf{x}_\infty - \mathbf{x}_\infty\| = 0.$$

This shows that $\mathbf{X}_t$ converges to a point of $\mathfrak{X}_\star$, namely $\mathbf{x}_\infty$, almost surely. □

Remark 7.4. We may try to further establish a stochastic version of Opial's Lemma [215, Lem. 2] for the conclusion part of the proof. However, we use here the weaker assumption that only the cluster points of a subsequence are guaranteed to be in the solution set. Therefore, it is not clear whether this can be generalized to infinite-dimensional Hilbert space or not.

Importantly, the result of Theorem 7.9 concerns the *last* iterate of the algorithm and does *not* require *strict* variational stability. In particular, this implies the almost sure convergence of the algorithm for bilinear problems like (7.3) where (EG) and (OG) do not converge. On the other hand, without learning rate separation, we can only show convergence either for the *average* iterate or for *strictly* variationally stable games [142, 146, 196].

CONVERGENCE OF OG+. In the same spirit, we can also show almost sure convergence of (OG+) when condition (7.15) is satisfied.

Theorem 7.10. *Suppose that Assumptions 5.1–5.3, 7.1 and 7.2 hold and all players run (OG+) with non-increasing learning rate sequences $(\gamma_t)_{t\in\mathbb{N}}$ and $(\eta_t)_{t\in\mathbb{N}}$ satisfying (7.15) and*

Convergence of OG+

$$\gamma_t \leq \min\left(\frac{1}{3L\sqrt{2N(1+\sigma_M^2)}}, \frac{1}{8\sqrt{N}L\sigma_M^2}\right), \quad \eta_t \leq \frac{\gamma_t}{2(1+\sigma_M^2)} \quad \text{for all } t \in \mathbb{N}. \tag{7.19}$$

Then, $\mathbf{X}_t$ converges almost surely to a Nash equilibrium.

The significance of Theorem 7.10 lies in the fact that (OG+) is a valid algorithm for online learning in games while (EG+) is not. Therefore, it suggests it is possible for the players to follow a specific learning algorithm and converge to a Nash equilibrium in all variationally stable games despite the stochasticity of the feedback.

The proof of Theorem 7.10 is however much more involved. First of all, with the additional term $3(\gamma_{t-1})^2\gamma_t\eta_{t+1}NL^2\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2$ in (7.14) we are not able to apply directly the Robbins–Siegmund Theorem. For this, we modify Lemma 7.7 to put the inequality in a more suitable form.

Lemma 7.11. *Suppose that Assumptions 5.2, 5.3 and 7.1 hold and all players run (OG+) with non-increasing learning rate sequences $(\gamma_t)_{t\in\mathbb{N}}$ and $(\eta_t)_{t\in\mathbb{N}}$ such that $\eta_t \leq \gamma_t$. Then, for all $t \geq 2$ and $\mathbf{x}_\star \in \mathfrak{X}_\star$, it holds*

Modified energy inequality for OG+

$$\begin{aligned}
&\mathbb{E}_{t-1}[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2 + c_{t+1}\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2] \\
&\quad\leq \mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2] + c_t\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2 \\
&\quad\quad - \gamma_t\eta_{t+1}(1 - a_t(1+\sigma_M^2) - b_t\sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\
&\quad\quad - \gamma_t\eta_{t+1}\left(1 - \frac{\eta_{t+1}(1+\sigma_M^2)}{\gamma_t}\right)\mathbb{E}_{t-1}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2] \\
&\quad\quad + (a_t\gamma_t\eta_{t+1} + b_t\gamma_t\eta_{t+1} + (\eta_{t+1})^2)N\sigma_A^2.
\end{aligned} \tag{7.20}$$

where $a_t = 9\gamma_t^2NL^2$, $b_t = 4\gamma_t\sqrt{N}L$, and $c_t = 3(\gamma_{t-1})^2\gamma_t\eta_{t+1}NL^2$.

Proof. This is shown by adding $3(\gamma_t)^2\gamma_{t+1}\eta_{t+2}NL^2\,\mathbb{E}_{t-1}[\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2]$ to the two sides of inequality (7.14). After that, we bound $\mathbb{E}_{t-1}[\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2]$ using (7.2) and bound various non-negative terms using $\eta_t \leq \gamma_t$, $\gamma_{t+1} \leq \gamma_t$, and $\eta_{t+2} \leq \eta_{t+1}$. □

The next challenge is to deal with the $\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2]$ term on the right-hand side (RHS) of (7.20). $\mathbf{X}_t$ is not $\mathcal{F}_{t-1}$-measurable because of the presence of noise $\xi_{t-\frac{1}{2}}$. To account for this, we introduce the surrogate sequence $(\tilde{\mathbf{X}}_t)_{t\in\mathbb{N}}$ defined by $\tilde{\mathbf{X}}_1 = \mathbf{X}_1$ and otherwise for $t \geq 2$

$$\tilde{\mathbf{X}}_t = \mathbf{X}_t + \eta_t\xi_{t-\frac{1}{2}} = \mathbf{X}_{t-1} - \eta_t\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}}). \tag{7.21}$$

This is similar to how we constructed $\tilde{\mathbf{X}}_{t+\frac{1}{2}}$ in the proofs of Lemmas 7.3 and 7.5. The vector $\tilde{\mathbf{X}}_t$ is $\mathcal{F}_{t-1}$-measurable and it holds that

$$\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2] = \mathbb{E}_{t-1}[\|\tilde{\mathbf{X}}_t - \eta_t\xi_{t-\frac{1}{2}} - \mathbf{x}_\star\|^2] = \|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|^2 + \eta_t^2\,\mathbb{E}_{t-1}[\|\xi_{t-\frac{1}{2}}\|^2]. \tag{7.22}$$

The use of $\tilde{\mathbf{X}}_t$ effectively allows us to prove the convergence of (OG+), as demonstrated below.

Proof. The proof is divided into four steps: In the first three steps we prove the almost sure convergence of the surrogate sequence $(\tilde{\mathbf{X}}_t)_{t\in\mathbb{N}}$ to a solution of $\mathfrak{X}_\star$ following the proof of Theorem 7.9; in the last step we show that this immediately implies the convergence of $(\mathbf{X}_t)_{t\in\mathbb{N}}$.

(1) *With probability* 1, $\|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|$ *converges for all* $\mathbf{x}_\star \in \mathfrak{X}_\star$. Let $x_\star \in \mathfrak{X}_\star$ and $c_t = 3(\gamma_{t-1})^2\gamma_t\eta_{t+1}NL^2$. Lemma 7.11 along with (7.19) implies that

$$\begin{aligned}\mathbb{E}_{t-1}[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2 + c_{t+1}\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2] \le \mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2] + c_t\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2 \\ - \frac{\gamma_t\eta_{t+1}}{2}\,\mathbb{E}_{t-1}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2] \\ + (9\gamma_t^3NL^2 + 4\gamma_t^2\sqrt{N}L + \eta_{t+1})\eta_{t+1}N\sigma_A^2.\end{aligned}$$

The summability of the rightmost term is guaranteed by (7.15) and the boundedness of $(\gamma_t)_{t\in\mathbb{N}}$. We can thus apply the Robbins–Siegmund theorem (Lemma 7.8) to the inequality with

$$\begin{aligned}&\mathcal{G}_t \leftarrow \mathcal{F}_{t-1}, \quad U_t \leftarrow \mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2] + c_t\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2, \quad \alpha_t \leftarrow 0,\\ &\zeta_t \leftarrow \frac{\gamma_t\eta_{t+1}}{2}\,\mathbb{E}_{t-1}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2], \quad \chi_t \leftarrow (9\gamma_t^3NL^2 + 4\gamma_t^2\sqrt{N}L + \eta_{t+1})\eta_{t+1}N\sigma_A^2.\end{aligned}$$

However, Lemma 7.11 only applies to $t \ge 2$. For $t = 1$, we have by convention $\hat{\mathbf{V}}_{1/2} = 0$, $\mathbf{X}_{3/2} = \mathbf{X}_1$ and $\mathbf{X}_2 = \mathbf{X}_1 - \eta_2\hat{\mathbf{V}}_{3/2}$, leading to

$$\begin{aligned}\mathbb{E}[\|\mathbf{X}_2 - \mathbf{x}_\star\|^2] &= \mathbb{E}[\|\mathbf{X}_1 - \mathbf{x}_\star\|^2 - 2\eta_2\langle\mathbf{V}(\mathbf{X}_{3/2}), \mathbf{X}_{3/2} - \mathbf{x}_\star\rangle + \eta_2^2\|\hat{\mathbf{V}}_{3/2}\|^2]\\ &\le \mathbb{E}[\|\mathbf{X}_1 - \mathbf{x}_\star\|^2 + \eta_2^2\|\hat{\mathbf{V}}_{3/2}\|^2],\end{aligned}$$

Subsequently,

$$\mathbb{E}[\|\mathbf{X}_2 - \mathbf{x}_\star\|^2 + c_2\|\hat{\mathbf{V}}_{1/2}\|^2] \le \mathbb{E}[\|\mathbf{X}_1 - \mathbf{x}_\star\|^2 + \eta_2^2(1+\sigma_M^2)\|\mathbf{V}(\mathbf{X}_1)\|^2] + \eta_2^2N\sigma_A^2 \tag{7.23}$$

We may thus choose $\zeta_1 = 0$ and $\chi_1 = \eta_2^2((1+\sigma_M^2)\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_1)\|^2] + N\sigma_A^2)$.

As a consequence of the Robbins–Siegmund theorem, we have both (*i*) the almost sure convergence of $\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2] + c_t\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2$ to a finite random variable U_∞ and (*ii*) $\sum_{t=2}^{+\infty}\gamma_t\eta_{t+1}\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2] < +\infty$.

To proceed, with equality (7.22) and bounds (7.1), (7.2) we get

$$\begin{aligned}&\sum_{t=2}^{+\infty}\mathbb{E}[\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2] + c_t\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2 - \|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|^2]\\ &\quad= \sum_{t=2}^{+\infty}\mathbb{E}[c_t\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2 + \eta_t^2\,\mathbb{E}_{t-1}[\|\xi_{t-\frac{1}{2}}\|^2]]\\ &\quad\le \sum_{t=3}^{+\infty}3(\gamma_{t-1})^2\gamma_t\eta_{t+1}NL^2((1+\sigma_M^2)\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t-\frac{3}{2}})\|^2] + N\sigma_A^2)\\ &\qquad+ \sum_{t=2}^{+\infty}\eta_t^2(\sigma_M^2\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2] + N\sigma_A^2)\\ &\quad\le \sum_{t=1}^{+\infty}\gamma_t\eta_{t+1}\left(\sigma_M^2 + \frac{1}{6}\right)\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2]\\ &\qquad+ \sum_{t=1}^{+\infty}((\eta_{t+1})^2 + 3\gamma_1\gamma_t^2\eta_{t+1}NL^2(1+\sigma_M^2))N\sigma_A^2 < +\infty.\end{aligned}$$

In the second to last inequality we use that $3\gamma_t^2 NL^2(1+\sigma_M^2) \leq 1/6$ which holds true thanks to condition (7.19). Invoking Lemma B.4, we then deduce

$$\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2] + c_t\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2 - \|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|^2 \xrightarrow{t\to+\infty} 0 \text{ almost surely.}$$

Thus, together with the almost sure convergence of $\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2] + c_t\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2$ to U_∞ we obtain the almost sure convergence of $\|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|^2$ to U_∞. To summarize, we have shown that for all $\mathbf{x}_\star \in \mathfrak{X}_\star$, the distance $\|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|$ almost surely converges. Applying Corollary B.7, we conclude that the event $\{\|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|$ converges for all $\mathbf{x}_\star \in \mathfrak{X}_\star\}$ happens with probability 1.

(2) *There exists an increasing function* $\omega\colon \mathbb{N} \to \mathbb{N}$ *such that* $\|\mathbf{V}(\mathbf{X}_{\omega(t)+\frac{1}{2}})\|^2 + \|\mathbf{X}_{\omega(t)+\frac{1}{2}} - \tilde{\mathbf{X}}_{\omega(t)}\|^2$ *converges to* 0 *almost surely.* From Lemma B.5, we know it is sufficient to show that

$$\liminf_{t\to+\infty} \mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 + \|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_t\|^2] = 0.$$

Since $\sum_{t=1}^{+\infty} \gamma_t\eta_{t+1} = +\infty$, the above is implied by

$$\sum_{t=2}^{+\infty} \gamma_t\eta_{t+1}\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 + \|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_t\|^2] < +\infty. \tag{7.24}$$

We already know that $\sum_{t=2}^{+\infty} \gamma_t\eta_{t+1}\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2] < +\infty$. It thus remains to deal with the term $\mathbb{E}[\|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_t\|^2]$. Using Assumption 7.1 and $\eta_t \leq \gamma_t$, we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_t\|^2] &= \mathbb{E}[\|\gamma_t\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}}) + (\gamma_t+\eta_t)\xi_{t-\frac{1}{2}}\|^2] \\
&= \gamma_t^2\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2] + (\gamma_t+\eta_t)^2\,\mathbb{E}[\|\xi_{t-\frac{1}{2}}\|^2] \\
&\leq \gamma_t^2(1+4\sigma_M^2)\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2] + 4\gamma_t^2 N\sigma_A^2.
\end{aligned}$$

With the summability of $(\gamma_t^2\eta_{t+1})_{t\in\mathbb{N}}$, $(\gamma_t\eta_{t+1}\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2])_{t\in\mathbb{N}}$, and the fact that the learning rates are non-increasing, we then obtain

$$\begin{aligned}
&\sum_{t=2}^{+\infty} \gamma_t\eta_{t+1}\,\mathbb{E}[\|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_t\|^2] \\
&\quad\leq \sum_{t=2}^{+\infty} \gamma_t\eta_{t+1}\,\mathbb{E}[\gamma_t^2(1+4\sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2 + 4\gamma_t^2 N\sigma_A^2] \\
&\quad\leq \sum_{t=1}^{+\infty} \gamma_1^2\gamma_t\eta_{t+1}(1+4\sigma_M^2)\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2] + \sum_{t=1}^{+\infty} 4\gamma_1\gamma_t^2\eta_{t+1}N\sigma_A^2 \\
&\quad< +\infty.
\end{aligned} \tag{7.25}$$

As a consequence, inequality (7.24) holds true and the claim can be deduced immediately.

(3) $(\tilde{\mathbf{X}}_t)_{t\in\mathbb{N}}$ *converges to a point in* $\mathfrak{X}_\star$ *almost surely.* Let us define the event

$$\begin{aligned}
\mathcal{E} = \{&\|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\| \text{ converges for all } \mathbf{x}_\star \in \mathfrak{X}_\star; \\
&\|\mathbf{V}(\mathbf{X}_{\omega(t)+\frac{1}{2}})\|^2 + \|\mathbf{X}_{\omega(t)+\frac{1}{2}} - \tilde{\mathbf{X}}_{\omega(t)}\|^2 \text{ converges to } 0\}
\end{aligned}$$

Combining the aforementioned two points we know that $\mathbb{P}(\mathcal{E}) = 1$. We now prove that $(\tilde{\mathbf{X}}_t)_{t\in\mathbb{N}}$ converges to a point in $\mathfrak{X}_\star$ for any realization of $\mathcal{E}$.

Let us consider a realization of $\mathcal{E}$. The set $\mathfrak{X}_\star$ being non-empty, the convergence of $\|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|$ for a $\mathbf{x}_\star \in \mathfrak{X}_\star$ implies the boundedness of $(\tilde{\mathbf{X}}_t)_{t\in\mathbb{N}}$. Therefore, we can extract $(\tilde{\mathbf{X}}_{\omega(\psi(t))})_t$, a subsequence of $(\tilde{\mathbf{X}}_{\omega(t)})_t$, that converges to a point $\mathbf{x}_\infty \in \mathbb{R}^d$. As $\lim_{t\to+\infty}\|\mathbf{X}_{\omega(\psi(t))+\frac{1}{2}} - \tilde{\mathbf{X}}_{\omega(\psi(t))}\|^2 = 0$, we deduce that $(\mathbf{X}_{\omega(\psi(t))+\frac{1}{2}})_t$ also converges to $\mathbf{x}_\infty \in \mathcal{X}$. Moreover, we also have $\lim_{t\to+\infty}\|\mathbf{V}(\mathbf{X}_{\omega(\psi(t))+\frac{1}{2}})\|^2 = 0$. By continuity of $\mathbf{V}$ we then know that $\mathbf{V}(\mathbf{x}_\infty) = 0$, i.e., $\mathbf{x}_\infty \in \mathfrak{X}_\star$. By definition of $\mathcal{E}$, this implies the convergence of $\|\tilde{\mathbf{X}}_t - \mathbf{x}_\infty\|$. The limit $\lim_{t\to+\infty}\|\tilde{\mathbf{X}}_t - \mathbf{x}_\infty\|$ is thus well defined and $\lim_{t\to+\infty}\|\tilde{\mathbf{X}}_t - \mathbf{x}_\infty\| = \lim_{t\to+\infty}\|\tilde{\mathbf{X}}_{\omega(\psi(t))} - \mathbf{x}_\infty\|$. However, $\lim_{t\to+\infty}\|\tilde{\mathbf{X}}_{\omega(\psi(t))} - \mathbf{x}_\infty\| = 0$ by the choice of $\mathbf{x}_\infty$. We have therefore $\lim_{t\to+\infty}\|\tilde{\mathbf{X}}_t - \mathbf{x}_\infty\| = 0$. Recalling that $\mathbf{x}_\infty \in \mathfrak{X}_\star$, we have indeed shown that $(\tilde{\mathbf{X}}_t)_{t\in\mathbb{N}}$ converges to a point in $\mathfrak{X}_\star$.

(4) *Conclude: $(\mathbf{X}_t)_{t\in\mathbb{N}}$ converges to a point in $\mathfrak{X}_\star$ almost surely .* We claim that $\|\mathbf{X}_t - \tilde{\mathbf{X}}_t\|$ converges to 0. In fact, similar to (7.25), it holds that

$$\sum_{t=1}^{+\infty} \mathbb{E}[\|\mathbf{X}_t - \tilde{\mathbf{X}}_t\|^2] = \sum_{t=2}^{+\infty} \eta_t^2\, \mathbb{E}[\|\xi_{t-\frac{1}{2}}\|^2] < +\infty.$$

Applying Lemma B.4 we then get almost sure convergence of $\|\mathbf{X}_t - \tilde{\mathbf{X}}_t\|$ to 0. Moreover, we have shown in the previous point that $(\tilde{\mathbf{X}}_t)_{t\in\mathbb{N}}$ converges to a point in $\mathfrak{X}_\star$ almost surely. Combining the above two arguments we obtain the almost sure convergence of $(\mathbf{X}_t)_{t\in\mathbb{N}}$ to a point in $\mathfrak{X}_\star$. □

So far, we have proved convergence of the iterates $(\mathbf{X}_t)_{t\in\mathbb{N}}$ to a Nash equilibrium. While it is sufficient for the computation of an equilibrium point, one should note that this is not the actual iterate of play in the learning-in-games setup. Instead, the players play $\mathbf{x}_t = \mathbf{X}_{t+\frac{1}{2}}$, and provided that the players use larger optimistic steps, the convergence of $\mathbf{X}_t$ does not necessarily imply the convergence of $\mathbf{X}_{t+\frac{1}{2}}$. In view of this, in the next theorem we derive sufficient condition for the latter to hold.

Convergence of the leading state of OG+

Theorem 7.12. *Suppose that Assumptions 5.1–5.3, 7.1 and 7.2 hold and all players run (OG+) with non-increasing learning rate sequences $(\gamma_t)_{t\in\mathbb{N}}$ and $(\eta_t)_{t\in\mathbb{N}}$ satisfying (7.15) and (7.19). Assume further that $\gamma_t^3 = O(\eta_t)$ and there exists $q \in (2, 4]$ and $\sigma > 0$ such that $\mathbb{E}[\|\xi_{t+\frac{1}{2}}\|^q] \le \sigma^q$ for all t and $\sum_{t=1}^{+\infty}\gamma_t^q < \infty$. Then, the actual point of play $\mathbf{X}_{t+\frac{1}{2}}$ converges almost surely to a Nash equilibrium.*

Proof. Since we already know that $(\mathbf{X}_t)_{t\in\mathbb{N}}$ converges to a point in $\mathfrak{X}_\star$ almost surely, it is sufficient to show that $\lim_{t\to+\infty}\|\mathbf{X}_t - \mathbf{X}_{t+\frac{1}{2}}\| = 0$ almost surely. By the update rule of OG+, we have, for $t \ge 2$, $\mathbf{X}_t - \mathbf{X}_{t+\frac{1}{2}} = \gamma_t\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}}) + \gamma_t\xi_{t-\frac{1}{2}}$. We will deal with the two terms separately. For the noise term, we notice that under the additional assumptions we have

$$\sum_{t=2}^{+\infty} \mathbb{E}[\|\gamma_t\xi_{t-\frac{1}{2}}\|^q] \le \sum_{t=2}^{+\infty} \gamma_t^q\sigma^q < +\infty.$$

Therefore, applying Lemma B.4 gives the almost sure convergence of $\|\gamma_t\xi_{t-\frac{1}{2}}\|$ to 0. As for the operator term, for $t \ge 3$ we bound

$$\|\gamma_t\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\| \le \gamma_t\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}}) - \mathbf{V}(\mathbf{X}_{t-1})\| + \gamma_t\|\mathbf{V}(\mathbf{X}_{t-1})\|.$$

On one hand, as $(\mathbf{X}_t)_{t\in\mathbb{N}}$ converges to a point in $\mathfrak{X}_\star$ almost surely, the term $\gamma_t\|\mathbf{V}(\mathbf{X}_{t-1})\|$ converges to 0 almost surely by continuity of $\mathbf{V}$. On the other hand, by Lipschitz continuity of $\mathbf{V}$ we have

$$\begin{aligned}
&\sum_{t=2}^{+\infty}\mathbb{E}[(\gamma_{t+1})^2\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})-\mathbf{V}(\mathbf{X}_t)\|^2]\\
&\quad\le\sum_{t=2}^{+\infty}(\gamma_{t+1})^2\gamma_t^2NL^2\,\mathbb{E}[\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2]\\
&\quad\le\sum_{t=2}^{+\infty}\gamma_t^4NL^2\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2]+\sum_{t=2}^{+\infty}\gamma_t^4NL^2\,\mathbb{E}[\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|^2]. \qquad (7.26)
\end{aligned}$$

Since $\gamma_t^3=O(\eta_t)$, there exists $C\in\mathbb{R}_+$ such that $\gamma_t^3\le C\eta_t$ for all $t\in\mathbb{N}$. Along with the summability of $(\gamma_t\eta_{t+1}\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2])_{t\in\mathbb{N}}$ we get

$$\sum_{t=2}^{+\infty}\gamma_t^4NL^2\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2]\le\sum_{t=2}^{+\infty}\gamma_{t-1}\eta_tCNL^2\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2]<+\infty. \qquad (7.27)$$

Regarding the noise term, we use (i) $\mathbb{E}[\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|^2]\le\sigma^2$ which holds true because $\mathbb{E}[\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|^q]\le\sigma^q$ with $q>2$, (ii) $\sum_{t=1}^{+\infty}\gamma_t^q<+\infty$, and (iii) $q\le 4$ to get

$$\sum_{t=2}^{+\infty}\gamma_t^4NL^2\,\mathbb{E}[\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|^2]\le\sum_{t=2}^{+\infty}\gamma_t^q\gamma_1^{4-q}NL^2\sigma^2<+\infty. \qquad (7.28)$$

Combining (7.26), (7.27), and (7.28) we obtain $\sum_{t=2}^{+\infty}\mathbb{E}[(\gamma_{t+1})^2\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})-\mathbf{V}(\mathbf{X}_t)\|^2]<+\infty$, which implies $\lim_{t\to+\infty}\gamma_{t+1}\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})-\mathbf{V}(\mathbf{X}_t)\|=0$ using Lemma B.4. In summary, we have shown the three sequences $(\gamma_t\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|)_{t\in\mathbb{N}}$, $(\gamma_t\|\mathbf{V}(\mathbf{X}_{t-1})\|)_{t\in\mathbb{N}}$, and $(\gamma_t\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})-\mathbf{V}(\mathbf{X}_{t-1})\|)_{t\in\mathbb{N}}$ converge almost surely to 0. As we have

$$\begin{aligned}
\|\mathbf{X}_t-\mathbf{X}_{t+\frac{1}{2}}\|&=\|\gamma_t\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})+\gamma_t\boldsymbol{\xi}_{t-\frac{1}{2}}\|\\
&\le\gamma_t\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})-\mathbf{V}(\mathbf{X}_{t-1})\|+\gamma_t\|\mathbf{V}(\mathbf{X}_{t-1})\|+\gamma_t\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|,
\end{aligned}$$

we can indeed conclude that $\lim_{t\to+\infty}\|\mathbf{X}_t-\mathbf{X}_{t+\frac{1}{2}}\|=0$ almost surely. □

7.3.2 *Convergence Rate*

In Section 7.3.1, we have shown that both (EG+) and (OG+) converge to a Nash equilibrium with suitably chosen learning rates. Nevertheless, it is unclear how fast this convergence is. In this subsection, we answer this question for games whose pseudo-gradient satisfies the following error bound condition.

Assumption 7.3 (Error bound)**.** For some $\nu>0$ and all $\mathbf{x}\in\mathbb{R}^d$, we have *Error bound condition*

$$\|\mathbf{V}(\mathbf{x})\|_*\ge\nu\,\mathrm{dist}(\mathbf{x},\mathfrak{X}_\star). \qquad (EB)$$

This kind of error bound is standard in the VI literature [73, 183]. It serves as a relatively mild condition under which geometric convergence of an algorithm's last iterate can be established [184, 253, 267]. In particular, it is satisfied by

- **Strongly monotone operators:** here, ν is the strong monotonicity modulus.

- **Affine operators:** for $\mathbf{V}(\mathbf{x}) = M\mathbf{x} + w$ where M is a $d \times d$ matrix and w is a d-dimensional vector, ν is the minimum non-zero singular value of M.

In this sense, Assumption 7.3 provides a unified umbrella for two types of problems that are typically considered poles apart. More generally speaking, (EB) is a special case of the metric subregularity concept in operator theory [67, 172, 175], and is also closely related to the Polyak–Łojasiewicz (PL) condition in optimization [148, 290]. Both of these are widely used in the literature to demonstrate the geometric convergence of algorithms.

CONVERGENCE RATE OF EG+. With our energy inequality Lemma 7.4 and the error bound condition, it is straightforward to derive convergence rates for the (EG+) algorithm.

Convergence rate of EG+

Theorem 7.13. *Suppose that Assumptions 5.1–5.3 and 7.1–7.3 hold and all players run (EG+) with learning rate sequences $(\gamma_t)_{t\in\mathbb{N}}$ and $(\eta_t)_{t\in\mathbb{N}}$ satisfying (7.15) and (7.17). Then:*

The case of constant learning rates

(a) If the learning rates are fixed at $\gamma_t \equiv \gamma$, $\eta_t \equiv \eta$, we have:

$$\mathbb{E}[\operatorname{dist}(\mathbf{X}_t, \mathfrak{X}_\star)^2] \le (1-\Delta)^{t-1}\,\mathbb{E}[\operatorname{dist}(\mathbf{X}_1, \mathfrak{X}_\star)^2] + \frac{C}{\Delta}$$

with constants $C = (\eta^2 + \gamma^3\eta L_g^2 + 2\gamma^2\eta L_g)N\sigma_A^2$ and $\Delta = \gamma\eta\nu^2/2$. In particular, the convergence is geometric if the noise is multiplicative (i.e., $\sigma_A = 0$).

The case of decreasing learning rates

(b) If the learning rates are set to $\gamma_t = \gamma/(t+\beta)^{\frac{1}{2}-q}$ and $\eta_{t+1} = \eta/(t+\beta)^{\frac{1}{2}+q}$ for some $\gamma, \eta > 0$, $q \in (0, 1/2)$, and $\beta \in \mathbb{N}$ such that $\gamma\eta\nu^2/2 > r := \min(1/2 - q, 2q)$, we have

$$\mathbb{E}[\operatorname{dist}(\mathbf{X}_t, \mathfrak{X}_\star)^2] \le \frac{C}{\Delta - r}\frac{1}{t^r} + o\left(\frac{1}{t^r}\right)$$

with constants C and Δ defined as in the previous point. In particular, the optimal rate is attained when $q = 1/6$, which gives $\mathbb{E}[\operatorname{dist}(\mathbf{X}_t, \mathfrak{X}_\star)^2] = O(1/t^{1/3})$.

Proof. From (7.18) and Assumption 7.3, we get immediately

$$\begin{aligned}\mathbb{E}_t[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2] \le \|\mathbf{X}_t - \mathbf{x}_\star\|^2 - \frac{\gamma_t\eta_{t+1}\nu^2}{2}\operatorname{dist}(\mathbf{X}_t, \mathfrak{X}_\star)^2 \\ + \left((\eta_{t+1})^2 + \gamma_t^3\eta_{t+1}L_g^2 + 2\gamma_t^2\eta_{t+1}L_g\right)N\sigma_A^2.\end{aligned}$$

By concavity of the minimum operator, we then obtain

$$\begin{aligned}\mathbb{E}_t\left[\min_{\mathbf{x}_\star\in\mathfrak{X}_\star}\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2\right] &\le \min_{\mathbf{x}_\star\in\mathfrak{X}_\star}\mathbb{E}_t[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2] \\ &\le \min_{\mathbf{x}_\star\in\mathfrak{X}_\star}\|\mathbf{X}_t - \mathbf{x}_\star\|^2 - \frac{\gamma_t\eta_{t+1}\nu^2}{2}\operatorname{dist}(\mathbf{X}_t, \mathfrak{X}_\star)^2 \\ &\quad + \left((\eta_{t+1})^2 + \gamma_t^3\eta_{t+1}L_g^2 + 2\gamma_t^2\eta_{t+1}L_g\right)N\sigma_A^2.\end{aligned}$$

In other words,

$$\begin{aligned}\mathbb{E}_t[\operatorname{dist}(\mathbf{X}_{t+1}, \mathfrak{X}_\star)^2] \le \left(1 - \frac{\gamma_t\eta_{t+1}\nu^2}{2}\right)\operatorname{dist}(\mathbf{X}_t, \mathfrak{X}_\star)^2 \\ + \left((\eta_{t+1})^2 + \gamma_t^3\eta_{t+1}L_g^2 + 2\gamma_t^2\eta_{t+1}L_g\right)N\sigma_A^2.\end{aligned}$$

We conclude by taking total expectation and apply respectively Lemma B.2 and Lemma B.3 to get points (a) and (b). □

The first part of Theorem 7.13 shows that, if (EG+) is run with constant learning rates, the initial condition is forgotten exponentially fast and the iterates converge to a neighborhood of $\mathfrak{X}_\star$. Furthermore, when dealing with multiplicative noise, the convergence isn't merely to a neighborhood of the solution, but is exact—the iterates converge precisely to the solution set itself, and do so at a geometric rate. More precisely, we can easily show that the number of iterations that is required to make the expected squared distance to the solution set to be smaller than ε in this case is

$$t_\varepsilon = O\left(L_g^2(1+\sigma_M^2)^2/\nu^2 \log(1/\varepsilon)\right).$$

On the other hand, we generally do not have exact convergence when $\sigma_A > 0$. To make the neighborhood in question small, we then need to decrease both γ and γ/η; this would be impossible for vanilla (EG) for which $\gamma/\eta = 1$. If we instead take decreasing sequences, an $O(1/t^{1/3})$ last-iterate convergence rate can be achieved, as shown in the second part of the theorem.

Remark 7.5. In [126], we further improve the $O(1/t^{1/3})$ rate to $O(1/t)$ for affine operators. To achieve this, we need to set the optimistic learning rate as constant and define the update learning rate in the form of $\eta_t = \eta/(t+\beta)$. Notably, when $\sigma_A > 0$, this results in the non-convergence of the intermediate states $(\mathbf{X}_{t+\frac{1}{2}})_{t\in\mathbb{N}}$, despite the convergence of the base states $(\mathbf{X}_t)_{t\in\mathbb{N}}$.

CONVERGENCE RATE OF OG+. We next establish the counterpart of Theorem 7.13 for (OG+).

Theorem 7.14. *Suppose that Assumptions 5.1–5.3 and 7.1–7.3 hold and all players run (OG+) with non-increasing learning rate sequences $(\gamma_t)_{t\in\mathbb{N}}$ and $(\eta_t)_{t\in\mathbb{N}}$ satisfying (7.15) and*

$$\gamma_t \leq \frac{1}{12\sqrt{N}L(1+\sigma_M^2)}, \quad \eta_t^2 \leq \frac{\gamma_t\eta_{t+1}}{2(1+\sigma_M^2)} \quad \text{for all } t \in \mathbb{N}. \tag{7.29}$$

Convergence rate of OG+

Then,

(a) *If the noise is multiplicative (i.e., $\sigma_A = 0$) and the learning rates are constant, we have $\mathbb{E}[\operatorname{dist}(\mathbf{X}_t, \mathfrak{X}_\star)^2] = O(\exp(-\rho t))$ for some $\rho > 0$.* *The case of constant learning rates*

(b) *If the learning rates are $\gamma_t = \gamma/(t+\beta)^{1/3}$ and $\eta_{t+1} = \eta/(t+\beta)^{2/3}$ for some $\gamma, \eta > 0$ and $\beta \in \mathbb{N}$ such that $\gamma\eta\nu^2 > 4/3$, we have $\mathbb{E}[\operatorname{dist}(\mathbf{X}_t, \mathfrak{X}_\star)^2] = O(1/t^{1/3})$.* *The case of decreasing learning rates*

Again, we have geometric convergence when the noise is multiplicative and $O(1/t^{1/3})$ convergence rate otherwise. The learning rate condition (7.29) is slightly different from than the one presented in Theorem 7.10, as the proof is based on another energy inequality that works directly with the surrogate iterate $\tilde{\mathbf{X}}_t$ introduced in (7.21).

Lemma 7.15. *Suppose that Assumptions 5.2, 5.3 and 7.1 hold and all players run (OG+) with the same predetermined learning rate sequences. Then, for all $t \geq 2$ and $\mathbf{x}_\star \in \mathfrak{X}_\star$, it holds*

$$\begin{aligned}\mathbb{E}_{t-1}[\|\tilde{\mathbf{X}}_{t+1} - \mathbf{x}_\star\|^2] \leq \|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|^2 + 3(\gamma_{t-1})^2\gamma_t\eta_{t+1}NL^2\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2 \\ - (\gamma_t\eta_{t+1}(1 - a_t(1+\sigma_M^2) - b_t\sigma_M^2) - \eta_t^2\sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2\end{aligned}$$

$$- \gamma_t \eta_{t+1} \left(1 - \frac{\eta_{t+1}}{\gamma_t}\right) \mathbb{E}_{t-1}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2] + (a_t \gamma_t \eta_{t+1} + b_t \gamma_t \eta_{t+1} + \eta_t^2) N \sigma_A^2, \tag{7.30}$$

where $a_t = 3(\gamma_t^2 + \eta_t^2)NL^2$, *and* $b_t = (3\gamma_t + \eta_t^2/\gamma_t)\sqrt{N}L$.

Proof. The inequality can be shown by slightly modifying the proofs of Lemmas 7.6 and 7.7. First, we use Proposition 7.2 to get

$$\mathbb{E}_{t-1}[\|\tilde{\mathbf{X}}_{t+1} - \mathbf{x}_\star\|^2] = \mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2 - 2\eta_{t+1}\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star\rangle - 2\gamma_t \eta_{t+1}\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \hat{\mathbf{V}}_{t-\frac{1}{2}}\rangle + (\eta_{t+1})^2 \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2].$$

The second and the third terms on the RHS of the equality can be bounded as before.[3] We conclude with the help of (7.22) and (7.1). □

Lemma 7.15 is crucial for our proof because it allows us to derive recursive inequality for the squared distance to the solution set $\text{dist}(\tilde{\mathbf{X}}_t, \mathbf{x}_\star)^2$ as done in proving Theorem 7.13. This cannot be achieved with the presence of expectation in $\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2]$ on the RHS of (7.14). Equipped with Lemma 7.15, we now present the proof of the theorem.

Proof Theorem 7.14. We first notice that condition (7.29) guarantees that

$$8\gamma_t^2 NL^2(1 + \sigma_M^2)^2 + 5\gamma_t \sqrt{N}L(1 + \sigma_M^2) \le \frac{1}{2}. \tag{7.31}$$

With $\eta_{t+1} \le \gamma_t$ and $\nu \le L$ we then have

$$3\gamma_t \sqrt{N}L + \frac{\gamma_t \eta_{t+1} \nu^2}{4} \le 3\gamma_t \sqrt{N}L(1 + \sigma_M^2) + \gamma_t^2 NL^2(1 + \sigma_M^2)^2 \le 1.$$

Therefore,

$$3(\gamma_{t-1})^2 \gamma_t \eta_{t+1} NL^2 \|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2 \le \left(1 - \frac{\gamma_t \eta_{t+1} \nu^2}{4}\right)(\gamma_{t-1})^2 \eta_t \sqrt{N}L \|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2.$$

We now turn back to the quasi-descent inequality (7.30). We add the term $\mathbb{E}_{t-1}[\gamma_t^2 \eta_{t+1} \sqrt{N}L \|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2]$ to both sides of the inequality, bound it from above using (7.2), and simplify (with notably $\eta_t^2 \sigma_M^2 \le \gamma_t \eta_{t+1}/2$), leading to

$$\begin{aligned}
&\mathbb{E}_{t-1}[\|\tilde{\mathbf{X}}_{t+1} - \mathbf{x}_\star\|^2 + \gamma_t^2 \eta_{t+1} \sqrt{N}L \|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2] \\
&\le \|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|^2 + \left(1 - \frac{\gamma_t \eta_{t+1} \nu^2}{4}\right)(\gamma_{t-1})^2 \eta_t \sqrt{N}L \|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2 \\
&\quad - \gamma_t \eta_{t+1} \left(\frac{1}{2} - (6\gamma_t^2 NL^2 + 5\gamma_t \sqrt{N}L)(1 + \sigma_M^2)\right) \|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\
&\quad - \frac{\gamma_t \eta_{t+1}}{2} \mathbb{E}_{t-1}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2] \\
&\quad + (6\gamma_t^3 \eta_{t+1} NL^2 + 5\gamma_t^2 \eta_{t+1} \sqrt{N}L + \eta_t^2) N \sigma_A^2.
\end{aligned}$$

3 Note that the first thing that we do for the equivalence of these terms in previous proofs is exactly to drop the noise vector $\xi_{t+\frac{1}{2}}$; see (7.9) and (7.12).

It follows from Young's inequality, Lipschitz continuity of $\mathbf{V}$, and the error bound condition that

$$\begin{aligned}\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 &\geq \frac{1}{2}\|\mathbf{V}(\tilde{\mathbf{X}}_t)\|^2 - \|\mathbf{V}(\tilde{\mathbf{X}}_t) - \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 \\ &\geq \frac{\nu^2}{2}\operatorname{dist}(\tilde{\mathbf{X}}_t, \mathfrak{X}_\star)^2 - NL^2\|\gamma_t \mathbf{V}(\mathbf{X}_{t-\frac{1}{2}}) + (\gamma_t + \eta_t)\xi_{t-\frac{1}{2}}\|^2.\end{aligned}$$

With Assumption 7.1 we can bound

$$\mathbb{E}_{t-1}[\|\gamma_t \mathbf{V}(\mathbf{X}_{t-\frac{1}{2}}) + (\gamma_t + \eta_t)\xi_{t-\frac{1}{2}}\|^2] \leq \gamma_t^2((1 + 4\sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2 + 4N\sigma_A^2).$$

Combining the above three inequalities, we get

$$\begin{aligned}&\mathbb{E}_{t-1}[\|\tilde{\mathbf{X}}_{t+1} - \mathbf{x}_\star\|^2 + \gamma_t^2\eta_{t+1}\sqrt{NL}\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2] \\ &\leq \|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|^2 - \frac{\gamma_t\eta_{t+1}\nu^2}{4}\operatorname{dist}(\tilde{\mathbf{X}}_t, \mathfrak{X}_\star)^2 \\ &\quad + \left(1 - \frac{\gamma_t\eta_{t+1}\nu^2}{4}\right)(\gamma_{t-1})^2\eta_t\sqrt{NL}\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2 \\ &\quad - \gamma_t\eta_{t+1}\left(\frac{1}{2} - (8\gamma_t^2NL^2 + +5\gamma_t\sqrt{NL})(1 + \sigma_M^2)\right)\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\ &\quad + (8\gamma_t^3\eta_{t+1}NL^2 + 5\gamma_t^2\eta_{t+1}\sqrt{NL} + \eta_t^2)N\sigma_A^2. \end{aligned} \tag{7.32}$$

Moreover, we have

$$(8\gamma_t^2NL^2 + 5\gamma_t\sqrt{NL})(1 + \sigma_M^2) \leq 8\gamma_t^2NL^2(1 + \sigma_M^2)^2 + 5\gamma_t\sqrt{NL}(1 + \sigma_M^2).$$

We can thus drop the forth term on the RHS of (7.32) thanks to (7.31). Using the concavity of the minimum operator as in the proof of Theorem 7.13, we get

$$\begin{aligned}&\mathbb{E}_{t-1}[\operatorname{dist}(\tilde{\mathbf{X}}_{t+1}, \mathbf{x}_\star)^2 + \gamma_t^2\eta_{t+1}\sqrt{NL}\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2] \\ &\leq \left(1 - \frac{\gamma_t\eta_{t+1}\nu^2}{4}\right)\left(\operatorname{dist}(\tilde{\mathbf{X}}_t, \mathfrak{X}_\star)^2 + (\gamma_{t-1})^2\eta_t\sqrt{NL}\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2\right) \\ &\quad + (8\gamma_t^3\eta_{t+1}NL^2 + 5\gamma_t^2\eta_{t+1}\sqrt{NL} + \eta_t^2)N\sigma_A^2. \end{aligned} \tag{7.33}$$

Taking total expectation and applying either Lemma B.2 or Lemma B.3 to the above inequality with $a_t \leftarrow \mathbb{E}[\operatorname{dist}(\tilde{\mathbf{X}}_t, \mathfrak{X}_\star)^2 + (\gamma_{t-1})^2\eta_t\sqrt{NL}\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2]$, we can derive convergence rate for a_t (either $O(\exp(-\rho t))$ or $O(1/t^{1/3})$) for the two learning rate schemes described in the statement.[4] To conclude, we note that

$$\begin{aligned}\mathbb{E}[\operatorname{dist}(\mathbf{X}_t, \mathfrak{X}_\star)^2] &= \mathbb{E}\left[\mathbb{E}_{t-1}\left[\min_{\mathbf{x}_\star \in \mathfrak{X}_\star}\|\mathbf{X}_t - \mathbf{x}_\star\|^2\right]\right] \\ &\leq \mathbb{E}\left[\min_{\mathbf{x}_\star \in \mathfrak{X}_\star}\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2]\right] \\ &= \mathbb{E}\left[\min_{\mathbf{x}_\star \in \mathfrak{X}_\star}\|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|^2 + \eta_t^2\,\mathbb{E}_{t-1}[\|\xi_{t-\frac{1}{2}}\|^2]\right] \\ &\leq \mathbb{E}\left[\operatorname{dist}(\tilde{\mathbf{X}}_t, \mathfrak{X}_\star)^2 + \eta_t^2\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2\right].\end{aligned}$$

4 Although (7.33) only holds for $t \geq 2$, one can easily show that a_2 is finite.

As $\eta_t^2 = \Theta(\gamma_t^2 \eta_{t+1})$ in the two cases, we have actually $\mathbb{E}[\mathrm{dist}(\mathbf{X}_t, \mathfrak{X}_\star)^2] = O(a_t + a_{t+1})$. This ends the proof. □

7.4 LOCAL CONVERGENCE

In this section, we shift our focus to localized analysis, examining the stochastic stability of the (EG+) dynamics around equilibrium points (we do not consider (OG+) here for the sake of simplicity). A notable advantage of this approach is the considerably relaxed set of assumptions compared to the global ones utilized in the previous section. Concretely, let us consider first-order equilibrium points defined as $\{\mathbf{x}_\star \in \mathbb{R}^d : \mathbf{V}(\mathbf{x}_\star) = 0\}$. Then, the local assumptions are stated with respect to a first-order equilibrium $\mathbf{x}_\star$ and a neighborhood $\mathcal{U}$ of it as below.

Local assumptions on the operator

Assumption 7.4. The operator $\mathbf{V}$ is L-Lipschitz continuous on $\mathcal{U}$, i.e., for all $\mathbf{x}, \mathbf{x}' \in \mathcal{U}$,

$$\|\mathbf{V}(\mathbf{x}') - \mathbf{V}(\mathbf{x})\|_* \leq L\|\mathbf{x}' - \mathbf{x}\|.$$

Assumption 7.5. The operator $\mathbf{V}$ satisfies $\langle \mathbf{V}(\mathbf{x}), \mathbf{x} - \mathbf{x}_\star \rangle \geq 0$ for all $\mathbf{x} \in \mathcal{U}$.

Assumption 7.4 and Assumption 7.5 mirror respectively Assumption 5.2 and Assumption 5.3, but they impose restrictions solely on the behavior of $\mathbf{V}$ within a neighborhood $\mathcal{U}$. These two assumptions are essential for the convergence of (EG+). On the other hand, while a localized version of Assumption 5.1 ensures that $\mathbf{x}_\star$ is a local Nash equilibrium [229], it is not necessary for the convergence analysis itself.

Moving forward, we also consider a localized version of Assumption 7.1 for noise or measurement error.

Local assumption on noise

Assumption 7.6. The noise vectors $(\xi_t)_{s \in \mathbb{N}/2}$ satisfy satisfies the following requirements for some for some $q > 2$ and $\sigma \geq 0$.

(a) *Zero-mean:* $\mathbb{E}_s[\xi_s]\, \mathbb{1}_{\{\mathbf{X}_s \in \mathcal{U}\}} = 0$.

(b) *Moment control:* $\mathbb{E}_s[\|\xi_s\|^q]\, \mathbb{1}_{\{\mathbf{X}_s \in \mathcal{U}\}} \leq \sigma^q$.

Similar to before, Assumption 7.6 asserts that the noise only needs to be "well-behaved" when the played point lies in the neighborhood $\mathcal{U}$. For technical reason, with $q > 2$ we require here the boundedness of a higher order moment. While this is a stronger assumption in terms of moment control requirement, it can still incorporate a multiplicative noise component. In fact, we may choose $\mathcal{U}$ to be a bounded set without loss of generality. Then the existence of σ in Assumption 7.6 is implied by the existence of $\sigma_A, \sigma_M \geq 0$ such that for all s

$$\mathbb{E}_s[\|\xi_s\|^q]\, \mathbb{1}_{\{\mathbf{X}_s \in \mathcal{U}\}} \leq \sigma_A^q + \sigma_M^q \|\mathbf{V}(\mathbf{X}_s)\|^q.$$

Precisely, we just take $\sigma = (\sigma_A^q + \sigma_M^q \max_{\mathbf{x} \in \mathcal{U}} \|\mathbf{V}(\mathbf{x})\|^q)^{1/q}$. Although separating the two components may allow us to show improved convergence rate when $\sigma_A = 0$ as in Theorems 7.13 and 7.14, we do not pursue in this direction.

7.4.1 *Stability of Equilibrium*

As a principle pillar of our analysis, we first show that any first-order equilibrium of the game is *Lyapunov stable in probability* [152] relative to (EG+). To state this result, we define $\tilde{\mathbf{X}}_{t+\frac{1}{2}} = \mathbf{X}_t - \gamma_t \mathbf{V}(\mathbf{X}_t)$ as in the proof of Lemma 7.3.

Stochastic stability of the equilibrium under EG+

Theorem 7.16. *Let $\mathbf{x}_\star$ be a first-order equilibrium such that Assumptions 7.4–7.6 are satisfied on $\mathcal{U} = \mathcal{B}_\rho(\mathbf{x}_\star)$ for some $\rho > 0$. We fix a tolerance level $\delta \in (0,1)$. For every $\alpha \in (0,1)$, there is a neighborhood $\mathcal{U}_\alpha$ of $\mathbf{x}_\star$ and a constant $\Gamma > 0$ such that if (EG+) is initialized at $\mathbf{X}_1 \in \mathcal{U}_\alpha$ and is run with stepsizes satisfying $\sum_{t=1}^{+\infty} \gamma_t \eta_{t+1} = \infty$, $\sum_{t=1}^{+\infty} \eta_t^2 < \Gamma$, $\sum_{t=1}^{+\infty} \gamma_t^2 \eta_{t+1} < \Gamma$ and $\sum_{t=1}^{+\infty} \gamma_t^q < \Gamma$, then*

$$\mathcal{E}_\infty^\alpha = \{\mathbf{X}_{t+\frac{1}{2}} \in \mathcal{B}_\rho(\mathbf{x}_\star), \mathbf{X}_t, \tilde{\mathbf{X}}_{t+\frac{1}{2}} \in \mathcal{B}_{\alpha\rho}(\mathbf{x}_\star) \text{ for all } t \in \mathbb{N}\}$$

occurs with probability at least $1-\delta$, i.e., $\mathbb{P}(\mathcal{E}_\infty^\alpha \mid \mathbf{X}_1 \in \mathcal{U}_\alpha) \geq 1 - \delta$.

To put it in simple terms, Theorem 7.16 states that as long as (EG+) is initialized sufficiently close to the equilibrium point and run with sufficiently small learning rates (while ensuring $\sum_{t=1}^{+\infty} \gamma_t \eta_{t+1} = +\infty$), the iterates produced by the algorithm remain close to $\mathbf{x}_\star$ with a probability arbitrarily close to 1. This stability result would allow us to prove our main convergence theorems in Theorems 7.19 and 7.21.

Nonetheless, proving Theorem 7.16 is itself challenging. Unlike in the case of perfect feedback where the stability of an equilibrium is a direct consequence of the (quasi-)Fejér monotonicity of the iterates, proving the stochastic stability requires a careful analysis of the involved stochastic sequence. In particular, we need to control the total noise accumulating from each noisy step, a task which is made difficult by the fact that the norm of the feedback can only be upper bounded recursively and thus depends on previous iterates. To tackle this challenge, we introduce the following lemma for bounding a recursive stochastic process.

A lemma on recursive stochastic process

Lemma 7.17. *Consider a filtration $(\mathcal{G}_t)_{t\in\mathbb{N}}$ and four $(\mathcal{G}_t)_{t\in\mathbb{N}}$-adapted processes $(U_t)_{t\in\mathbb{N}}$, $(\zeta_t)_{t\in\mathbb{N}}$, $(\chi_t)_{t\in\mathbb{N}}$, $(\vartheta_t)_{t\in\mathbb{N}}$ such that $(\chi_t)_{t\in\mathbb{N}}$ is non-negative and the following recursive inequality is satisfied for all $t \geq 1$*

$$U_{t+1} \leq U_t - \zeta_t + \chi_{t+1} + \vartheta_{t+1}.$$

Fixing a constant $C > 0$, we define the events $(\mathcal{A}_t)_{t\in\mathbb{N}}$ by $\mathcal{A}_1 \coloneqq \{U_1 \leq C/2\}$ and $\mathcal{A}_t \coloneqq \{U_t \leq C\} \cap \{\chi_t \leq C/4\}$ for $t \geq 2$. We consider also the decreasing sequence of events $(\mathcal{E}_t)_{t\in\mathbb{N}}$ defined by $\mathcal{E}_t \coloneqq \bigcap_{1\leq s\leq t} \mathcal{A}_s$. If the following four assumptions hold true

(i) $\forall t, \zeta_t \mathbb{1}_{\mathcal{E}_t} \geq 0$,

(ii) $\forall t, \mathbb{E}[\vartheta_{t+1} \mid \mathcal{G}_t] \mathbb{1}_{\mathcal{E}_t} = 0$,

(iii) $\mathbb{P}(\mathcal{A}_1) > 0$,

(iv) $\sum_{t=1}^{\infty} \mathbb{E}[(\vartheta_{t+1}^2 + \chi_{t+1}) \mathbb{1}_{\mathcal{E}_t}] \leq \delta\varepsilon\, \mathbb{P}(\mathcal{A}_1)$,

where $\varepsilon = \min(C^2/16, C/4)$ and $\delta \in (0,1)$, then $\mathbb{P}\left(\bigcap_{t\geq 1} \mathcal{A}_t \mid \mathcal{A}_1\right) \geq 1 - \delta$.

Proof. Let us start by introducing the following two $(\mathcal{G}_t)_{t\in\mathbb{N}}$-adapted submartingale sequences

$$S_t \coloneqq \sum_{s=2}^{t} \vartheta_s \quad \text{and} \quad Q_t \coloneqq S_t^2 + \sum_{s=2}^{t} \chi_s.$$

Subsequently, we define an auxiliary sequence of events

$$\mathcal{H}_t \coloneqq \mathcal{A}_1 \cap \{\max_{2\leq s\leq t} Q_s \leq \varepsilon\}$$

which is also decreasing. With this at hand, we are ready to start our proof.

(1) *Inclusion* $\mathcal{H}_t \subset \mathcal{E}_t$. We prove the inclusion by induction. The statement is true when $t = 1$ as $\mathcal{H}_1 = \mathcal{E}_1 = \mathcal{A}_1$. For $t \geq 2$, we write

$$U_t \leq U_1 - \sum_{s=1}^{t-1} \zeta_s + \sum_{s=1}^{t-1} \chi_{s+1} + \sum_{s=1}^{t-1} \vartheta_{s+1}. \tag{7.34}$$

By induction hypothesis, $\mathcal{H}_{t-1} \subset \mathcal{E}_{t-1}$, and thus for all $s \leq t-1$, we have $\mathcal{H}_t \subset \mathcal{E}_{t-1} \subset \mathcal{E}_s$. Combining with *(i)* we deduce that for any realization of $\mathcal{H}_t$, $\sum_{s=1}^{t-1} \zeta_s \geq 0$. On the other hand, by definition of $\mathcal{H}_t$, it holds $Q_t \mathbb{1}_{\mathcal{H}_t} \leq \varepsilon$. This implies

$$\left(\sum_{s=1}^{t-1} \vartheta_{s+1}\right) \mathbb{1}_{\mathcal{H}_t} = S_t \mathbb{1}_{\mathcal{H}_t} \leq \sqrt{\varepsilon} \leq C/4, \tag{7.35}$$

$$\left(\sum_{s=1}^{t-1} \chi_{s+1}\right) \mathbb{1}_{\mathcal{H}_t} \leq \varepsilon \leq C/4. \tag{7.36}$$

Finally as $\mathcal{H}_t \subset \mathcal{A}_1$ we have $U_1 \mathbb{1}_{\mathcal{H}_t} \leq C/2$. Therefore, for any realization of $\mathcal{H}_t$, using (7.34) gives

$$U_t \leq C/2 - 0 + C/4 + C/4 = C.$$

In the meantime (7.36) ensures as well $\chi_t \mathbb{1}_{\mathcal{H}_t} \leq C/4$ and we have thus proven $\mathcal{H}_t \subset \mathcal{A}_t$. Using $\mathcal{H}_t \subset \mathcal{H}_{t-1} \subset \mathcal{E}_{t-1}$, we conclude $\mathcal{H}_t \subset \mathcal{E}_t$.

(2) *Recursive bound on* $\mathbb{E}[Q_t \mathbb{1}_{\mathcal{H}_{t-1}}]$. Since $\mathcal{H}_{t-1} \subseteq \mathcal{H}_{t-2}$, it holds $\mathcal{H}_{t-1} = \mathcal{H}_{t-2} \setminus (\mathcal{H}_{t-2} \setminus \mathcal{H}_{t-1})$. We can therefore decompose

$$\begin{aligned}
\mathbb{E}[Q_t \mathbb{1}_{\mathcal{H}_{t-1}}] &= \mathbb{E}[(Q_t - Q_{t-1}) \mathbb{1}_{\mathcal{H}_{t-1}}] + \mathbb{E}[Q_{t-1} \mathbb{1}_{\mathcal{H}_{t-1}}] \\
&= \mathbb{E}[(\vartheta_t^2 + 2\vartheta_t S_{t-1} + \chi_t) \mathbb{1}_{\mathcal{H}_{t-1}}] \\
&\quad + \mathbb{E}[Q_{t-1} \mathbb{1}_{\mathcal{H}_{t-2}}] - \mathbb{E}[Q_{t-1} \mathbb{1}_{\mathcal{H}_{t-2} \setminus \mathcal{H}_{t-1}}].
\end{aligned}$$

From the law of total expectation, $\mathcal{H}_{t-1} \subset \mathcal{E}_{t-1}$ and *(ii)* we have

$$\mathbb{E}[\vartheta_t S_{t-1} \mathbb{1}_{\mathcal{H}_{t-1}}] = \mathbb{E}[\mathbb{E}[\vartheta_t \mid \mathcal{G}_{t-1}] S_{t-1} \mathbb{1}_{\mathcal{H}_{t-1}}] = 0.$$

As $\vartheta_t^2 + \chi_t$ is non-negative, using again $\mathcal{H}_{t-1} \subset \mathcal{E}_{t-1}$, we get

$$\mathbb{E}[(\vartheta_t^2 + \chi_t) \mathbb{1}_{\mathcal{H}_{t-1}}] \leq \mathbb{E}[(\vartheta_t^2 + \chi_t) \mathbb{1}_{\mathcal{E}_{t-1}}].$$

By definition for any realization in $\mathcal{H}_{t-2} \setminus \mathcal{H}_{t-1}$, it holds $Q_{t-1} > \varepsilon$ and thus

$$\mathbb{E}[Q_{t-1} \mathbb{1}_{\mathcal{H}_{t-2} \setminus \mathcal{H}_{t-1}}] \geq \varepsilon \, \mathbb{E}[\mathbb{1}_{\mathcal{H}_{t-2} \setminus \mathcal{H}_{t-1}}] = \varepsilon \, \mathbb{P}(\mathcal{H}_{t-2} \setminus \mathcal{H}_{t-1}).$$

Combining the above we deduce the following recursive bound

$$\mathbb{E}[Q_t \mathbb{1}_{\mathcal{H}_{t-1}}] \leq \mathbb{E}[Q_{t-1} \mathbb{1}_{\mathcal{H}_{t-2}}] + \mathbb{E}[(\vartheta_t^2 + \chi_t) \mathbb{1}_{\mathcal{E}_{t-1}}] - \varepsilon \, \mathbb{P}(\mathcal{H}_{t-2} \setminus \mathcal{H}_{t-1}). \tag{7.37}$$

(3) *Conclude.* Summing (7.37) from $t = 3$ to T we obtain

$$\begin{aligned}
\mathbb{E}[Q_T \mathbb{1}_{\mathcal{H}_{T-1}}] &\leq \mathbb{E}[Q_2 \mathbb{1}_{\mathcal{H}_1}] + \sum_{t=3}^{T} \mathbb{E}[(\vartheta_t^2 + \chi_t) \mathbb{1}_{\mathcal{E}_{t-1}}] - \varepsilon \sum_{t=3}^{T} \mathbb{P}(\mathcal{H}_{t-2} \setminus \mathcal{H}_{t-1}) \\
&= \sum_{t=2}^{T} \mathbb{E}[(\vartheta_t^2 + \chi_t) \mathbb{1}_{\mathcal{E}_{t-1}}] - \varepsilon \, \mathbb{P}(\mathcal{A}_1 \setminus \mathcal{H}_{T-1}),
\end{aligned} \tag{7.38}$$

where in the second line we use $Q_2 = \vartheta_2^2 + \chi_2$, $\mathcal{H}_1 = \mathcal{E}_1 = \mathcal{A}_1$ and $\mathcal{H}_1 \setminus \mathcal{H}_{T-1} = \dot{\bigcup}_{3\le t\le T}(\mathcal{H}_{t-2} \setminus \mathcal{H}_{t-1})$ with $\dot{\bigcup}$ denoting the disjoint union (true since $(\mathcal{H}_t)_{t\ge1}$ is a decreasing sequence of events). By repeating the same arguments that are used before and using the fact that Q_T is non-negative, we have

$$\begin{aligned}\mathbb{P}(\mathcal{A}_1 \setminus \mathcal{H}_T) &= \mathbb{P}(\mathcal{H}_{T-1} \setminus \mathcal{H}_T) + \mathbb{P}(\mathcal{A}_1 \setminus \mathcal{H}_{T-1}) \\ &\le \frac{1}{\varepsilon}\,\mathbb{E}[Q_T\,\mathbb{1}_{\mathcal{H}_{T-1}\setminus\mathcal{H}_T}] + \mathbb{P}(\mathcal{A}_1 \setminus \mathcal{H}_{T-1}) \\ &\le \frac{1}{\varepsilon}\,\mathbb{E}[Q_T\,\mathbb{1}_{\mathcal{H}_{T-1}}] + \mathbb{P}(\mathcal{A}_1 \setminus \mathcal{H}_{T-1}). \end{aligned} \tag{7.39}$$

(7.39), (7.38) along with *(iii)* lead to

$$\mathbb{P}(\mathcal{A}_1 \setminus \mathcal{H}_T) \le \frac{1}{\varepsilon}\sum_{t=2}^{T}\mathbb{E}[(\vartheta_t^2 + \chi_t)\,\mathbb{1}_{\mathcal{E}_{t-1}}] \le \delta\,\mathbb{P}(\mathcal{A}_1).$$

Subsequently,

$$\mathbb{P}(\mathcal{H}_T \mid \mathcal{A}_1) = 1 - \frac{\mathbb{P}(\mathcal{A}_1 \setminus \mathcal{H}_T)}{\mathbb{P}(\mathcal{A}_1)} \ge 1 - \delta.$$

With $\mathcal{H}_T \subset \mathcal{E}_T$ this also gives $\mathbb{P}(\mathcal{E}_T \mid \mathcal{A}_1) \ge 1 - \delta$. We notice that $\bigcap_{t\ge1}\mathcal{E}_t = \bigcap_{t\ge1}\mathcal{A}_t$. As $(\mathcal{E}_t)_{t\ge1}$ is decreasing, by continuity from above we conclude

$$\mathbb{P}\left(\bigcap_{t\ge1}\mathcal{A}_t \mid \mathcal{A}_1\right) = \lim_{t\to\infty}\mathbb{P}(\mathcal{E}_t \mid \mathcal{A}_1) \ge 1 - \delta. \qquad \square$$

Remark 7.6. Lemma 7.17 requires $\mathbb{P}(\mathcal{A}_1) > 0$. For our local convergence and stability results, this means that the probability of $\mathbf{X}_1$ falling into the appropriate neighborhood should be strictly larger than 0.

Note that the inequality of Lemma 7.17 concerns the random variables themselves and not their expectations. We thus need an energy inequality that holds without taking expectation. We provide such inequality below.

An energy inequality for EG+ without expectation

Lemma 7.18. *For all $\mathbf{x}_\star \in \mathfrak{X}_\star$, $t \in \mathbb{N}$, the iterates generated by (EG+) satisfy the following inequality*

$$\begin{aligned}\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2 \le \|\mathbf{X}_t - \mathbf{x}_\star\|^2 &- 2\eta_{t+1}\langle \mathbf{V}(\mathbf{X}_{t+\frac12}), \mathbf{X}_{t+\frac12} - \mathbf{x}_\star\rangle \\ &- 2\gamma_t\eta_{t+1}\|\mathbf{V}(\mathbf{X}_t)\|(\|\mathbf{V}(\mathbf{X}_t)\| - \|\mathbf{V}(\tilde{\mathbf{X}}_{t+\frac12}) - \mathbf{V}(\mathbf{X}_t)\|) \\ &- 2\eta_{t+1}\langle \xi_{t+\frac12}, \mathbf{X}_t - \mathbf{x}_\star\rangle - 2\gamma_t\eta_{t+1}\langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac12}), \xi_t\rangle \\ &+ 2\gamma_t\eta_{t+1}\|\hat{\mathbf{V}}_t\|\|\mathbf{V}(\mathbf{X}_{t+\frac12}) - \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac12})\| + \eta_{t+1}^2\|\hat{\mathbf{V}}_{t+\frac12}\|^2 \end{aligned} \tag{7.40}$$

Moreover, if we assume Assumption 7.4 for some open set $\mathcal{U}$ and that $\mathbf{X}_t$, $\tilde{\mathbf{X}}_{t+\frac12}$, $\mathbf{X}_{t+\frac12}$ all lie in $\mathcal{U}$, then

$$\begin{aligned}\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2 \le \|\mathbf{X}_t - \mathbf{x}_\star\|^2 &- 2\eta_{t+1}\langle \mathbf{V}(\mathbf{X}_{t+\frac12}), \mathbf{X}_{t+\frac12} - \mathbf{x}_\star\rangle - 2\eta_{t+1}\langle \xi_{t+\frac12}, \mathbf{X}_t - \mathbf{x}_\star\rangle \\ &- 2\gamma_t\eta_{t+1}(1 - \gamma_t L)\|\mathbf{V}(\mathbf{X}_t)\|^2 - 2\gamma_t\eta_{t+1}\langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac12}), \xi_t\rangle \\ &+ 2\gamma_t^2\eta_{t+1}L\|\xi_t\|\|\hat{\mathbf{V}}_t\| + \eta_{t+1}^2\|\hat{\mathbf{V}}_{t+\frac12}\|^2. \end{aligned} \tag{7.41}$$

Proof. We develop

$$\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2 = \|\mathbf{X}_t - \mathbf{x}_\star\|^2 - 2\eta_{t+1}\langle \mathbf{V}(\mathbf{X}_{t+\frac12}), \mathbf{X}_t - \mathbf{x}_\star\rangle$$

$$- 2\eta_{t+1}\langle \xi_{t+\frac{1}{2}}, \mathbf{X}_t - \mathbf{x}_\star\rangle + \eta_{t+1}^2 \|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|^2.$$

We further develop the second term on the RHS of the equality

$$\begin{aligned}\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_t - \mathbf{x}_\star\rangle &= \langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star\rangle + \gamma_t\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \hat{\mathbf{V}}_t\rangle \\ &= \langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star\rangle + \gamma_t\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}) - \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}), \hat{\mathbf{V}}_t\rangle \\ &\quad + \gamma_t\langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}), \hat{\mathbf{V}}_t\rangle.\end{aligned}$$

To deal with the last term

$$\begin{aligned}\langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}), \hat{\mathbf{V}}_t\rangle &= \langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}), \mathbf{V}(\mathbf{X}_t)\rangle + \langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}), \xi_t\rangle \\ &= \langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}) - \mathbf{V}(\mathbf{X}_t), \mathbf{V}(\mathbf{X}_t)\rangle + \|\mathbf{V}(\mathbf{X}_t)\|^2 + \langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}), \xi_t\rangle.\end{aligned}$$

By combing all the above, we readily get (7.40) with Cauchy's inequality. If Assumption 7.4 holds on a set that $\mathbf{X}_t$, $\tilde{\mathbf{X}}_{t+\frac{1}{2}}$, $X_{t+\frac{1}{2}}$ belong to, we can further bound

$$\begin{aligned}2\gamma_t\eta_{t+1}\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}) - \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}})\|\|\hat{\mathbf{V}}_t\| &\le 2\gamma_t^2\eta_{t+1}L\|\xi_t\|\|\hat{\mathbf{V}}_t\|, \\ 2\gamma_t\eta_{t+1}\|\mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}) - \mathbf{V}(\mathbf{X}_t)\|\|\mathbf{V}(\mathbf{X}_t)\| &\le 2\gamma_t^2\eta_{t+1}L\|\mathbf{V}(\mathbf{X}_t)\|^2,\end{aligned}$$

which gives (7.41). □

With the above two lemmas, we are now ready to prove the stability of (EG+).

Proof of Theorem 7.16. We would like to apply Lemma 7.17, but instead of indexing by $t \in \mathbb{N}$, we index by $s \in \mathbb{N}/2$. We invoke (7.40) from Lemma 7.18 and set the random variables accordingly

$$\begin{aligned}\underbrace{\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2}_{U_{t+1}} \le \underbrace{\|\mathbf{X}_t - \mathbf{x}_\star\|^2}_{U_t} &\underbrace{- 2\eta_{t+1}\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star\rangle}_{\zeta_{t+\frac{1}{2}}} \\ &\underbrace{- 2\gamma_t\eta_{t+1}\|\mathbf{V}(\mathbf{X}_t)\|(\|\mathbf{V}(\mathbf{X}_t)\| - \|\mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}) - \mathbf{V}(\mathbf{X}_t)\|)}_{\zeta_t} \\ &+ \underbrace{(-2\eta_{t+1}\langle \xi_{t+\frac{1}{2}}, \mathbf{X}_t - \mathbf{x}_\star\rangle)}_{\vartheta_{t+1}} + \underbrace{(-2\gamma_t\eta_{t+1}\langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}), \xi_t\rangle)}_{\vartheta_{t+\frac{1}{2}}} \\ &\underbrace{+ 2\gamma_t\eta_{t+1}\|\hat{\mathbf{V}}_t\|\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}) - \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}})\| + \eta_{t+1}^2\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|^2}_{\chi_{t+1}} \qquad (7.42)\end{aligned}$$

We additionally define $\chi_{t+\frac{1}{2}} := \gamma_t^q\|\xi_t\|^q$ and $U_{t+\frac{1}{2}} := U_t - \zeta_t + \chi_{t+\frac{1}{2}} + \vartheta_{t+\frac{1}{2}}$ so that (7.42) implies $U_{t+1} \le U_{t+\frac{1}{2}} - \zeta_{t+\frac{1}{2}} + \chi_{t+1} + \vartheta_{t+1}$. With the definition of $U_{t+\frac{1}{2}}$ the inequality (7.34) is indeed verified with all half integers. We shall next verify that the assumptions *(i)*, *(ii)* and *(iii)* in Lemma 7.17 are satisfied for a C that is properly chosen. Let us write $\rho' := \alpha\rho$, define $\mathcal{U}' = \mathscr{B}_{\rho'}(\mathbf{x}_\star)$, and denote by M the supremum of $\|\mathbf{V}(\mathbf{x})\|$ for $\mathbf{x} \in \mathcal{U}'$. We choose $C := \min(\rho'^2/9, 4(\rho'/3)^q)$ and set Γ small enough to guarantee $\gamma_t \le \min(\rho'/(3M), 1/L)$.

(1) *Inclusion* $\mathcal{E}_t \subset \{\mathbf{X}_t, \tilde{\mathbf{X}}_{t+\frac{1}{2}} \in \mathcal{U}'\}$ *and* $\mathcal{E}_{t+\frac{1}{2}} \subset \{\mathbf{X}_t, \tilde{\mathbf{X}}_{t+\frac{1}{2}}, \mathbf{X}_{t+\frac{1}{2}} \in \mathcal{U}'\}$. Since $\mathcal{E}_t \subset \mathcal{A}_t$, for any realization of $\mathcal{E}_t$, we have $\|\mathbf{X}_t - \mathbf{x}_\star\|^2 \le C \le \rho'^2/9$. It follows

$$\|\tilde{\mathbf{X}}_{t+\frac{1}{2}} - \mathbf{x}_\star\|^2 \le 2\|\mathbf{X}_t - \mathbf{x}_\star\|^2 + 2\gamma_t^2\|\mathbf{V}(\mathbf{X}_t)\|^2 \le \frac{2\rho'^2}{9} + 2\gamma_t^2 M^2 \le \frac{4\rho'^2}{9}.$$

We have shown $\mathcal{E}_t \subset \{\mathbf{X}_t, \tilde{\mathbf{X}}_{t+\frac{1}{2}} \in \mathcal{U}'\}$. On the other hand, $\mathcal{E}_{t+\frac{1}{2}} \subset \mathcal{A}_t \cap \mathcal{A}_{t+\frac{1}{2}} \subset \{U_t \le C\} \cap \{\chi_{t+\frac{1}{2}} \le C/4\}$. Therefore for any realization of $\mathcal{E}_{t+\frac{1}{2}}$,

$$\gamma_t^q \|\xi_t\|^q = \chi_{t+\frac{1}{2}} \le \frac{C}{4} \le (\rho'/3)^q.$$

Subsequently,

$$\begin{aligned}\|\mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star\|^2 &\le 3\|\mathbf{X}_t - \mathbf{x}_\star\|^2 + 3\gamma_t^2\|\mathbf{V}(\mathbf{X}_t)\|^2 + 3\gamma_t^2\|\xi_t\|^2 \\ &\le \frac{\rho'^2}{3} + \frac{\rho'^2}{3} + 3\left(\frac{\rho'}{3}\right)^2 \le \rho'^2.\end{aligned}$$

This proves $\mathcal{E}_{t+\frac{1}{2}} \subset \{\mathbf{X}_t, \tilde{\mathbf{X}}_{t+\frac{1}{2}}, \mathbf{X}_{t+\frac{1}{2}} \in \mathcal{U}'\}$.

(2) *Assumption (i).* We start by $\zeta_{t+\frac{1}{2}} \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}} \ge 0$. This is true because $\mathcal{E}_{t+\frac{1}{2}} \subset \{\mathbf{X}_{t+\frac{1}{2}} \in \mathcal{U}'\}$ and $\mathcal{U}' \subset \mathscr{B}_\rho(\mathbf{x}_\star)$, which allows us to apply Assumption 7.5 to obtain $\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star\rangle \ge 0$ whenever $\mathcal{E}_{t+\frac{1}{2}}$ occurs. Similarly, by $\mathcal{E}_t \subset \{\mathbf{X}_t, \tilde{\mathbf{X}}_{t+\frac{1}{2}} \in \mathcal{U}'\}$ and Assumption 7.4 we then have

$$\zeta_t \mathbb{1}_{\mathcal{E}_t} \ge 2\gamma_t\eta_{t+1}(1 - \gamma_t L)\|\mathbf{V}(\mathbf{X}_t)\|^2 \ge 0.$$

(3) *Assumption (ii).* This is immediate from Assumption 7.6(a), $\mathcal{E}_t \subset \{\mathbf{X}_t \in \mathscr{B}_\rho(\mathbf{x}_\star)\}$, $\mathcal{E}_{t+\frac{1}{2}} \subset \{\mathbf{X}_{t+\frac{1}{2}} \in \mathscr{B}_\rho(\mathbf{x}_\star)\}$, and the law of the total expectation.

(4) *Assumption (iii).* By using that $\mathcal{E}_t \subset \{\tilde{\mathbf{X}}_{t+\frac{1}{2}} \in \mathcal{U}'\}$ and $\mathcal{E}_t \subset \{\mathbf{X}_t \in \mathscr{B}_\rho(\mathbf{x}_\star)\}$, we get

$$\begin{aligned}\mathbb{E}[\vartheta_{t+\frac{1}{2}}^2 \mathbb{1}_{\mathcal{E}_t}] &\le 4\gamma_t^2\eta_{t+1}^2 \mathbb{E}[\|\mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}})\|^2 \mathbb{1}_{\mathcal{E}_t}\|\xi_t\|^2 \mathbb{1}_{\mathcal{E}_t}] \\ &\le 4\gamma_t^2\eta_{t+1}^2 M^2 \mathbb{E}[\|\xi_t\|^2 \mathbb{1}_{\{\mathbf{X}_t \in \mathscr{B}_\rho(\mathbf{x}_\star)\}}] \le 4\gamma_t^2\eta_{t+1}^2 M^2\sigma^2.\end{aligned}$$

For the last inequality we use Assumption 7.6(b) and Jensen's inequality to bound $\mathbb{E}[\|\xi_t\|^2 \mathbb{1}_{\{\mathbf{X}_t \in \mathscr{B}_\rho(\mathbf{x}_\star)\}}]$. Similarly,

$$\begin{aligned}\mathbb{E}[\|\xi_t\| \mathbb{1}_{\{\mathbf{X}_t \in \mathscr{B}_\rho(\mathbf{x}_\star)\}}] &\le \sigma, \\ \mathbb{E}[\|\xi_{t+\frac{1}{2}}\|^2 \mathbb{1}_{\{\mathbf{X}_{t+\frac{1}{2}} \in \mathscr{B}_\rho(\mathbf{x}_\star)\}}] &\le \sigma^2.\end{aligned}$$

Using $\mathcal{E}_{t+\frac{1}{2}} \subset \{\mathbf{X}_t, \tilde{\mathbf{X}}_{t+\frac{1}{2}}, \mathbf{X}_{t+\frac{1}{2}} \in \mathcal{U}'\}$ and Assumption 7.4 then gives

$$\begin{aligned}\mathbb{E}[\chi_{t+1} \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] &\le 2\gamma_t^2\eta_{t+1}L \mathbb{E}[\|\xi_t\|(\|\mathbf{V}(\mathbf{X}_t)\| + \|\xi_t\|) \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] \\ &\quad + \eta_{t+1}^2 \mathbb{E}[(\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 + \|\xi_{t+\frac{1}{2}}\|^2) \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] \\ &\le 2\gamma_t^2\eta_{t+1}L \mathbb{E}[\|\xi_t\|^2 \mathbb{1}_{\{\mathbf{X}_t \in \mathscr{B}_\rho(\mathbf{x}_\star)\}}] \\ &\quad + 2\gamma_t^2\eta_{t+1}L \mathbb{E}[\|\xi_t\| \mathbb{1}_{\{\mathbf{X}_t \in \mathscr{B}_\rho(\mathbf{x}_\star)\}}\|\mathbf{V}(\mathbf{X}_t)\| \mathbb{1}_{\{\mathbf{X}_t \in \mathcal{U}'\}}] \\ &\quad + \eta_{t+1}^2(\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 \mathbb{1}_{\{\mathbf{X}_{t+\frac{1}{2}} \in \mathcal{U}'\}}] + \mathbb{E}[\|\xi_{t+\frac{1}{2}}\|^2 \mathbb{1}_{\{\mathbf{X}_{t+\frac{1}{2}} \in \mathscr{B}_\rho(\mathbf{x}_\star)\}}])\end{aligned}$$

$$\leq 2\gamma_t^2\eta_{t+1}L(M\sigma+\sigma^2)+\eta_{t+1}^2(M^2+\sigma^2). \tag{7.43}$$

By similar arguments and in particular by invoking $\mathcal{E}_{t+\frac{1}{2}} \subset \{U_t \leq C\}$ and the definition of C, it follows

$$\mathbb{E}[\vartheta_{t+1}^2\,\mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] \leq \frac{4}{9}\eta_{t+1}^2\rho'^2\sigma^2,$$

Combining the above with $\mathbb{E}[\chi_{t+\frac{1}{2}}\,\mathbb{1}_{\mathcal{E}_t}] \leq \gamma_t^q\sigma^q$, we have

$$\begin{aligned}
&\sum_{s\in 1,3/2,\dots}(\vartheta_{s+\frac{1}{2}}^2+\chi_{s+\frac{1}{2}})\,\mathbb{1}_{\mathcal{E}_s}\\
&\qquad\leq\sum_{t=1}^{\infty}\Big(\gamma_t^q\sigma^q+2\gamma_t^2\eta_{t+1}L(M\sigma+\sigma^2)+4\gamma_t^2\eta_{t+1}^2M^2\sigma^2\\
&\qquad\qquad+\eta_{t+1}^2(M^2+\sigma^2+\frac{4}{9}\rho'^2\sigma^2)\Big)\\
&\qquad\leq\left(\sigma^q+2L(M\sigma+\sigma^2)+\frac{4}{L}M^2\sigma^2+M^2+\sigma^2+\frac{4}{9}\rho'^2\sigma^2\right)\Gamma.
\end{aligned}$$

We can thus pick Γ small enough to make *(iii)* verified.

(5) *Conclude.* We set $\mathcal{U}_\alpha = \mathcal{B}_{\sqrt{C/2}}(\mathbf{x}_\star)$ so that $\mathcal{A}_1 = \{\mathbf{X}_1 \in \mathcal{U}_\alpha\}$. By invoking Lemma 7.17 we get $\mathbb{P}\left(\bigcap_{t\geq 1}\mathcal{A}_t \mid \mathcal{A}_1\right) \geq 1-\delta$. Additionally, (1) along with $\mathcal{U}' \subset \mathcal{B}_\rho(\mathbf{x}_\star)$ imply $\bigcap_{t\geq 1}\mathcal{A}_t \subset \mathcal{E}_\infty^\alpha$, concluding the proof. □

7.4.2 *Asymptotic Convergence*

In the next theorem we strengthen Theorem 7.16 to a high-probability convergence result. It additionally requires the equilibrium $\mathbf{x}_\star$ to be isolated (i.e., there is a neighborhood $\mathcal{U}$ of $\mathbf{x}_\star$ such that $\mathbf{x}_\star$ is the only equilibrium in $\mathcal{U}$), but can otherwise be regarded as the localized version of Theorem 7.9.

High-probability convergence of EG+

Theorem 7.19. *Let $\mathbf{x}_\star$ be an isolated first-order equilibrium such that Assumptions 7.4–7.6 are satisfied on $\mathcal{U} = \mathcal{B}_\rho(\mathbf{x}_\star)$ for some $\rho > 0$. For every $\delta > 0$, if (EG+) is run with small enough γ_t, η_t satisfying $\sum_{t=1}^{+\infty}\gamma_t\eta_{t+1} = \infty$. and initialized sufficiently close to $\mathbf{x}_\star$, its iterates converge to $\mathbf{x}_\star$ with probability at least $1-\delta$.*

Proof. Let $\alpha \in (0,1)$. By Theorem 7.16, we know that if (EG+) is run as stated in Theorem 7.19, the event $\mathcal{E}_\infty^\alpha$ occurs with probability $1-\delta$. With this at hand we are ready to prove the large probability convergence result. For $t \in \mathbb{N}$, let us define the following events

$$\begin{aligned}
\mathcal{E}_t &\coloneqq \{\mathbf{X}_s, \tilde{\mathbf{X}}_{s+\frac{1}{2}} \in \mathcal{B}_{\alpha\rho}(\mathbf{x}_\star) \text{ for all } s \in \{1,\dots,t\}\},\\
\mathcal{E}_{t+\frac{1}{2}} &\coloneqq \mathcal{E}_t \cap \{\mathbf{X}_{s+\frac{1}{2}} \in \mathcal{B}_\rho(\mathbf{x}_\star) \text{ for all } s \in \{1,\dots,t\}\}.
\end{aligned}$$

We notice that $\mathcal{E}_\infty^\alpha = \bigcap_{t\geq 1}\mathcal{E}_{t+\frac{1}{2}}$. We would like to establish a recursive inequality in the form of (7.16) by taking $U_t = \|\mathbf{X}_t - \mathbf{x}_\star\|\,\mathbb{1}_{\mathcal{E}_{t-\frac{1}{2}}}$. The main difficulty consists in controlling the term $\mathbb{E}_t[\langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}), \xi_t\rangle\,\mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}]$, which is generally non-zero as $\mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}$ is not $\mathcal{F}_t$-measurable. To achieve this, we rely on the following key observation.

$$\mathbb{E}_t[\xi_t\,\mathbb{1}_{\mathcal{E}_t}] = \mathbb{E}_t[\xi_t\,\mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] + \mathbb{E}_t[\xi_t\,\mathbb{1}_{\mathcal{E}_t\setminus\mathcal{E}_{t+\frac{1}{2}}}].$$

As $\mathbb{1}_{\mathcal{E}_t}$ is $\mathscr{F}_t$-measurable and $\mathcal{E}_t \subset \{\mathbf{X}_t \in \mathscr{B}_\rho(\mathbf{x}_\star)\}$, $\mathbb{E}_t[\xi_t \mathbb{1}_{\mathcal{E}_t}]$ is indeed zero and this implies

$$\|\mathbb{E}_t[\xi_t \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}]\| = \|\mathbb{E}_t[\xi_t \mathbb{1}_{\mathcal{E}_t \setminus \mathcal{E}_{t+\frac{1}{2}}}]\|. \tag{7.44}$$

The problem then reduces to finding an upper bound of $\|\mathbb{E}_t[\xi_t \mathbb{1}_{\mathcal{E}_t \setminus \mathcal{E}_{t+\frac{1}{2}}}]\|$. By definition, for any realization of $\mathcal{E}_t \setminus \mathcal{E}_{t+\frac{1}{2}}$, $\tilde{\mathbf{X}}_{t+\frac{1}{2}} \in \mathscr{B}_{\alpha\rho}(\mathbf{x}_\star)$ and $X_{t+\frac{1}{2}} \notin \mathscr{B}_\rho(\mathbf{x}_\star)$. Since $\mathbf{X}_{t+\frac{1}{2}} = \tilde{\mathbf{X}}_{t+\frac{1}{2}} - \gamma_t \xi_t$, we deduce

$$\mathcal{E}_t \setminus \mathcal{E}_{t+\frac{1}{2}} \subset \{\|\gamma_t \xi_t\| \geq (1-\alpha)\rho\}.$$

Therefore, using $\mathcal{E}_t \subset \{\mathbf{X}_t \in \mathscr{B}_\rho(\mathbf{x}_\star)\}$ along with the Chebyshev's inequality yields

$$\mathbb{P}(\mathcal{E}_t \setminus \mathcal{E}_{t+\frac{1}{2}} \mid \mathscr{F}_t) \leq \mathbb{P}\left(\|\xi_t\| \mathbb{1}_{\{\mathbf{X}_t \in \mathscr{B}_\rho(\mathbf{x}_\star)\}} \geq \frac{(1-\alpha)\rho}{\gamma_t} \,\middle|\, \mathscr{F}_t\right) \leq \frac{\sigma^2 \gamma_t^2}{(1-\alpha)^2 \rho^2}.$$

Applying the Cauchy–Schwarz inequality leads to

$$\|\mathbb{E}_t[\xi_t \mathbb{1}_{\mathcal{E}_t \setminus \mathcal{E}_{t+\frac{1}{2}}}]\| \leq \sqrt{\mathbb{E}_t[\|\xi_t \mathbb{1}_{\mathcal{E}_t}\|^2]} \sqrt{\mathbb{E}_t[\mathbb{1}^2_{\mathcal{E}_t \setminus \mathcal{E}_{t+\frac{1}{2}}}]} \leq \frac{\sigma^2 \gamma_t}{(1-\alpha)\rho}. \tag{7.45}$$

Then, by using (7.44), (7.45) and $\mathcal{E}_{t+\frac{1}{2}} \subset \mathcal{E}_t$, we get

$$\begin{aligned}
\mathbb{E}_t[\langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}), \xi_t \rangle \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] &= \mathbb{E}_t[\langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}) \mathbb{1}_{\mathcal{E}_t}, \xi_t \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}} \rangle] \\
&= \langle \mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}) \mathbb{1}_{\mathcal{E}_t}, \mathbb{E}_t[\xi_t \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] \rangle \\
&\leq \|\mathbf{V}(\tilde{\mathbf{X}}_{t+\frac{1}{2}}) \mathbb{1}_{\mathcal{E}_t}\| \|\mathbb{E}_t[\xi_t \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}]\| \\
&\leq \frac{M\sigma^2 \gamma_t}{(1-\alpha)\rho},
\end{aligned} \tag{7.46}$$

where $M \coloneqq \sup_{x \in \mathscr{B}_\rho(\mathbf{x}_\star)} \|\mathbf{V}(x)\|$. We can now derive a recursive bound on $\mathbb{E}[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\| \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}]$ by invoking Lemma 7.18. The inequality (7.41) multiplied by $\mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}$ holds true by definition of $\mathcal{E}_{t+\frac{1}{2}}$ and Assumption 7.4. The desired inequality can then be obtained by taking expectation conditioned on $\mathscr{F}_t$. On the one hand, we use

$$\begin{aligned}
\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star \rangle \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}} &\geq 0 \\
\mathbb{E}_t[\langle \xi_{t+\frac{1}{2}}, \mathbf{X}_t - \mathbf{x}_\star \rangle \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] = \mathbb{E}_t[\langle \mathbb{E}_{t+\frac{1}{2}}[\xi_{t+\frac{1}{2}}] \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}, \mathbf{X}_t - \mathbf{x}_\star \rangle] &= 0.
\end{aligned}$$

On the other hand, the last two terms of (7.41) can be bounded similarly as in (7.43) and the antepenultimate term can now be bounded thanks to (7.46). We then obtain

$$\begin{aligned}
\mathbb{E}_t[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2 \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] \leq \mathbb{E}_t[\|\mathbf{X}_t - \mathbf{x}_\star\|^2 \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] \\
- 2\gamma_t \eta_{t+1}(1 - \gamma_t L) \mathbb{E}_t[\|\mathbf{V}(\mathbf{X}_t)\|^2 \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] + \eta_{t+1}^2 (M^2 + \sigma^2) \\
+ 2\gamma_t^2 \eta_{t+1} \frac{M\sigma^2}{(1-\alpha)\rho} + 2\gamma_t^2 \eta_{t+1} L(M\sigma + \sigma^2).
\end{aligned} \tag{7.47}$$

We additionally suppose that γ_t is small enough such that $\gamma_t L \leq 1/2$, and set

$$\begin{aligned}\zeta_t &= \min\left(\|\mathbf{X}_t - \mathbf{x}_\star\|^2, \gamma_t \eta_{t+1} \|\mathbf{V}(\mathbf{X}_t)\|^2\right),\\ M_1 &= \frac{2M\sigma^2}{(1-\alpha)\rho} + 2L(M\sigma + \sigma^2),\\ M_2 &= M^2 + \sigma^2.\end{aligned} \tag{7.48}$$

It follows from (7.47) that

$$\mathbb{E}_t[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2 \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] \leq \mathbb{E}_t[(\|\mathbf{X}_t - \mathbf{x}_\star\|^2 - \zeta_t)\mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] + \gamma_t^2 \eta_{t+1} M_1 + \eta_{t+1}^2 M_2.$$

As $\|\mathbf{X}_t - \mathbf{x}_\star\|^2 - \zeta_t \geq 0$ and $\mathcal{E}_{t+\frac{1}{2}} \subset \mathcal{E}_{t-\frac{1}{2}}$, this implies

$$\mathbb{E}_t[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2 \mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] \leq \|\mathbf{X}_t - \mathbf{x}_\star\|^2 \mathbb{1}_{\mathcal{E}_{t-\frac{1}{2}}} - \zeta_t \mathbb{1}_{\mathcal{E}_{t-\frac{1}{2}}} + \gamma_t^2 \eta_{t+1} M_1 + \eta_{t+1}^2 M_2.$$

Invoking the the Robbins–Siegmund theorem (Lemma 7.8) gives the almost sure convergence of $\sum_{t=1}^{+\infty} \zeta_t \mathbb{1}_{\mathcal{E}_{t-\frac{1}{2}}}$ and $\|\mathbf{X}_t - \mathbf{x}_\star\|^2 \mathbb{1}_{\mathcal{E}_{t-\frac{1}{2}}}$. We use $\mathbb{P}(\mathcal{E}_\infty^\alpha) > 1 - \delta$ and deduce that

$$\mathbb{P}\left(\underbrace{\mathcal{E}_\infty^\alpha \cap \left\{\sum_{t=1}^{\infty} \zeta_t \mathbb{1}_{\mathcal{E}_{t-\frac{1}{2}}} < \infty\right\} \cap \left\{\|\mathbf{X}_t - \mathbf{x}_\star\|^2 \mathbb{1}_{\mathcal{E}_{t-\frac{1}{2}}} \ \textit{converges}\right\}}_{\mathcal{E}}\right) \geq 1 - \delta.$$

Since $\mathcal{E}_\infty^\alpha = \bigcap_{t\geq 1} \mathcal{E}_{t+\frac{1}{2}}$, for any realization of the above event it holds $\sum_{t=1}^{+\infty} \zeta_t < \infty$ and $\|\mathbf{X}_t - \mathbf{x}_\star\|^2$ converges. We assume by contradiction that $\|\mathbf{X}_t - \mathbf{x}_\star\|^2$ converges to some constant $U_\infty > 0$. From the summability of $(\zeta_t)_{t\in\mathbb{N}}$ we know that $\zeta_t \to 0$ and therefore for all t large enough we have in fact $\zeta_t = \gamma_t \eta_{t+1} \|\mathbf{V}(\mathbf{X}_t)\|^2$. It follows that $\sum_{t=1}^{+\infty} \gamma_t \eta_{t+1} \|\mathbf{V}(\mathbf{X}_t)\|^2 < \infty$. Repeating the arguments of the proof of Theorem 7.9 we then show that $\|\mathbf{X}_t - \mathbf{x}_\star\| \to 0$, which is a contradiction (we take ρ small enough so that $\mathbf{x}_\star$ is the only equilibrium point in $\mathscr{B}_\rho(\mathbf{x}_\star)$). We have therefore proved that $\|\mathbf{X}_t - \mathbf{x}_\star\| \to 0$ for any realization of $\mathcal{E}$. In conclusion, $\mathbf{X}_t$ converges to $\mathbf{x}_\star$ with probability at least $1 - \delta$. □

7.4.3 *Convergence Rate*

We proceed to present local convergence rate for (EG+). We focus specifically on equilibrium points that satisfy the following Jacobian regularity condition.

Invertibility of Jacobian

Assumption 7.7. $\mathbf{V}$ is differentiable at $\mathbf{x}_\star$ and its Jacobian matrix $\mathrm{Jac}_{\mathbf{V}}(x_\star)$ is invertible.

The link between Assumptions 7.3 and 7.7 is provided below.

From invertibility of Jacobian to local error bound

Proposition 7.20. *If a solution $\mathbf{x}_\star$ satisfies Assumption 7.7, then for every $\varepsilon > 0$, there is a neighborhood $\mathcal{U}$ of $\mathbf{x}_\star$ such that the error bound condition (EB) is satisfied on $\mathcal{U}$ with constant $\nu = \varsigma_{\min} - \varepsilon$ where $\varsigma_{\min}$ denotes the smallest singular value of $\mathrm{Jac}_{\mathbf{V}}(\mathbf{x}_\star)$.*

Proof. By definition of Jacobian we have

$$\mathbf{V}(\mathbf{x}) = \mathbf{V}(\mathbf{x}_\star) + \mathrm{Jac}_V(\mathbf{x}_\star)(\mathbf{x} - \mathbf{x}_\star) + o(\|\mathbf{x} - \mathbf{x}_\star\|). \tag{7.49}$$

By the min-max principle of singular value it holds

$$\|\mathrm{Jac}_{\mathbf{V}}(\mathbf{x}_\star)(\mathbf{x}-\mathbf{x}_\star)\| \geq \varsigma_{\min}\|\mathbf{x}-\mathbf{x}_\star\|. \tag{7.50}$$

Since $\mathbf{V}(\mathbf{x}_\star)=0$, combining (7.49) and (7.50) gives

$$\|\mathbf{V}(\mathbf{x})\| \geq \varsigma_{\min}\|\mathbf{x}-\mathbf{x}_\star\| - o(\|\mathbf{x}-\mathbf{x}_\star\|).$$

We conclude by noticing $\mathrm{dist}(\mathbf{x},\mathfrak{X}_\star)=\|\mathbf{x}-\mathbf{x}_\star\|$ when $\mathcal{U}$ is small enough. □

Proposition 7.20 elucidates that if a solution $\mathbf{x}_\star$ satisfies the Jacobian regularity condition, a local error bound can be established. This extends the applicability of our quantitative convergence results to much broader scenarios—on condition that we manage to establish a localized version of it. This brings us to the following theorem.

Local convergence rate of EG+

Theorem 7.21. *Let $\mathbf{x}_\star$ be a first-order equilibrium such that Assumptions 7.4–7.7 are satisfied on $\mathcal{U}=\mathscr{B}_\rho(\mathbf{x}_\star)$ for some $\rho>0$. Assume further that $q>3$ and (EG+) is run with stepsize parameters of the form $\gamma_t=\gamma/(t+\beta)^{1/3}$ and $\eta_{t+1}=\eta/(t+\beta)^{2/3}$. Then, for every $\delta>0$, when $\beta,\eta>0$ and $\gamma\geq\eta$ are taken large enough, there exist neighborhoods $\mathcal{U}_1,\mathcal{U}_2$ of $\mathbf{x}_\star$ and an event $\mathcal{E}_{\mathcal{U}_1}$ such that:*

(a) $\mathbb{P}(\mathcal{E}_{\mathcal{U}_1} \mid \mathbf{X}_1\in\mathcal{U}_1)\geq 1-\delta$.

(b) $\mathbb{P}(\mathbf{X}_t\in\mathcal{U}_2 \text{ for all } t \mid \mathcal{E}_{\mathcal{U}_1})=1$.

(c) $\mathbb{E}[\|\mathbf{X}_t-\mathbf{x}_\star\|^2 \mid \mathcal{E}_{\mathcal{U}_1}] = O\Big(1/t^{1/3}\Big)$

In words, if (EG+) is not initialized too far from $\mathbf{x}_\star$, the iterates $\mathbf{X}_t$ remain close to $\mathbf{x}_\star$ with probability at least $1-\delta$ and, conditioned on this event, $\mathbf{X}_t$ converges to $\mathbf{x}_\star$ at a rate $O(1/t^{1/3})$ in mean square error.

Proof. Both (a) and (b) are direct consequences of Theorem 7.16. In effect, since $q>3$, the sum of the series $\sum_{t=1}^{+\infty}\eta_{t+1}^2$, $\sum_{t=1}^{+\infty}\gamma_t^2\eta_{t+1}$ and $\sum_{t=1}^{+\infty}\gamma_t^q$ can be made arbitrarily small by taking sufficiently large β. Moreover, $\mathbf{x}_\star$ is an isolated solution because $\mathrm{Jac}_{\mathbf{V}}(\mathbf{x}_\star)$ is non-singular. Therefore, taking $\mathcal{E}_{\mathcal{U}_1}:=\mathcal{E}_\infty^\alpha$, $\mathcal{U}_1:=\mathcal{U}^\alpha$ and $\mathcal{U}_2:=\mathscr{B}_{\alpha\rho}(\mathbf{x}_\star)$ readily gives (a) and (b).

Finally, to guarantee (c), we need to have α small enough and enforce $\gamma\eta\varsigma_{\min}^2(1-\gamma_1 L)>1/6$. In fact, from $\gamma\eta\varsigma_{\min}^2(1-\gamma_1 L)>1/6$ we deduce the existence of $\varepsilon\in(0,\varsigma_{min})$ such that $\gamma\eta(\varsigma_{\min}-\varepsilon)^2(1-\gamma_1 L)>1/6$. Since $\mathrm{Jac}_{\mathbf{V}}(\mathbf{x}_\star)$ is non-singular, by Proposition 7.20 we can choose $\alpha>0$ so that the error bound condition (EB) is satisfied on $\mathscr{B}_{\alpha\rho}(\mathbf{x}_\star)$ with $\nu=\varsigma_{\min}-\varepsilon$. Let M_1, M_2 be defined as in (7.48). We then obtained from (7.47)

$$\begin{aligned}\mathbb{E}[\|\mathbf{X}_{t+1}-\mathbf{x}_\star\|^2\mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] \leq (1-2\gamma_t\eta_{t+1}\nu^2(1-\gamma_t L))\,\mathbb{E}[\|\mathbf{X}_t-\mathbf{x}_\star\|^2\mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}]\\ +\gamma_t^2\eta_{t+1}M_1+\eta_{t+1}^2M_2.\end{aligned}$$

Suppose additionally that β is large enough such that $2\gamma_t\eta_{t+1}\nu^2(1-\gamma_t L)\leq 1$. Then, by using $\mathcal{E}_{t+\frac{1}{2}}\subset\mathcal{E}_{t-\frac{1}{2}}$, we get

$$\begin{aligned}\mathbb{E}[\|\mathbf{X}_{t+1}-\mathbf{x}_\star\|^2\mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] \leq (1-2\gamma_t\eta_{t+1}\nu^2(1-\gamma_t L))\,\mathbb{E}[\|\mathbf{X}_t-\mathbf{x}_\star\|^2\mathbb{1}_{\mathcal{E}_{t-\frac{1}{2}}}]\\ +\gamma_t^2\eta_{t+1}M_1+\eta_{t+1}^2M_2\end{aligned}$$

Therefore, with the specified stepsize policy and the condition $\gamma\eta\nu^2(1-\gamma_1 L) > 1/6$, applying Lemma B.3 yields $\mathbb{E}[\|\mathbf{X}_{t+1}-\mathbf{x}_\star\|^2\,\mathbb{1}_{\mathcal{E}_{t+\frac{1}{2}}}] = O(1/t^{1/3})$. Finally

$$\mathbb{E}[\|\mathbf{X}_t-\mathbf{x}_\star\|^2 \mid \mathcal{E}_\infty^\alpha] = \frac{\mathbb{E}[\|\mathbf{X}_t-\mathbf{x}_\star\|^2\,\mathbb{1}_{\mathcal{E}_\infty^\alpha}]}{\mathbb{P}(\mathcal{E}_\infty^\alpha)} \le \frac{\mathbb{E}[\|\mathbf{X}_t-\mathbf{x}_\star\|^2\mathbb{1}_{\mathcal{E}_{t-\frac{1}{2}}}]}{1-\delta},$$

which proves $\mathbb{E}[\|\mathbf{X}_t-\mathbf{x}_\star\|^2 \mid \mathcal{E}_\infty^\alpha] = O(1/t^{1/3})$. □

This theorem offers insight into the rate of convergence under the much lighter local error bound condition. Taken together, Theorems 7.9 and 7.21 show that for all variationally stable games with a non-degenerate equilibrium point, employing the suggested learning rate policy yields an asymptotic $O(1/t^{1/3})$ rate. In more detail, the last point of Theorem 7.21 shows that, with the same kind of learning rate as in the second part of Theorem 7.13, we can retrieve a $O(1/t^{1/3})$ convergence rate provided that the iterates stay close to the solution.

8

DEALING WITH STOCHASTIC FEEDBACK II: NO-REGRET AND ADAPTIVE LEARNING

 # This chapter incorporates material from Hsieh et al. [129]

The preceding chapter examined how optimistic methods with learning rate separation can achieve trajectory convergence when the feedback is corrupted by noise. While our analysis showcased promising results, it suffered from two limitations. First, these methods rely on coordination among players to agree on a common learning rate. Second, the learning rates have to be carefully adjusted for each situation to make sure the guarantees hold up.

To tackle these limitations, we draw inspiration from Chapter 6 and devise an *adaptive* method in this chapter. This algorithm can be independently run by each player without knowledge of either the game or of the noise profile. Moreover, as in Chapter 6, it offers (near-)optimal guarantees in various scenarios with a unified learning rate policy, thus eliminating the need for both player coordination and the selection of situation-specific learning rates.

This chapter also pays particular attention to the multiplicative noise setup. As we have seen previously, this type of noise often leads to faster convergence. In this chapter, we further elucidate this point by deriving bounds on the norm of the pseudo-gradient and regret. Precisely, we focus on the notion of *pseudo-regret*.

Pseduo-regret

Definition 8.1 (Pseudo-regret)**.** For any player $i \in \mathcal{N}$, comparator set $\mathcal{Z}^i \subseteq \mathcal{X}^i$, and time horizon $T \in \mathbb{N}$, the *pseudo-regret* of this player relative to $\mathcal{Z}^i$ after T rounds is defined as

$$\overline{\mathrm{Reg}}_T^i(\mathcal{Z}^i) = \max_{z^i \in \mathcal{Z}^i} \mathbb{E}\left[\sum_{t=1}^{T} [\ell^i(x_t^i, \mathbf{x}_t^{-i}) - \ell^i(z^i, \mathbf{x}_t^{-i})]\right].$$

The expectation is taken over both the randomness of the feedback and of the algorithm.

Remark 8.1. Another closely related concept is *expected regret*, defined by

$$\mathbb{E}\left[\mathrm{Reg}_T^i(z)\right] = \mathbb{E}\left[\max_{z^i \in z^i} \sum_{t=1}^{T} [\ell^i(x_t^i, \mathbf{x}_t^{-i}) - \ell^i(z^i, \mathbf{x}_t^{-i})]\right].$$

Expected regret versus pseudo-regret

Compared to pseudo-regret, it targets the action that is optimal against the sequence of *realized* losses rather than optimal against the expected losses. Clearly, we always have $\overline{\mathrm{Reg}}_T^i(\mathcal{Z}) \leq \mathbb{E}[\mathrm{Reg}_T^i(\mathcal{Z})]$. While we focus on deriving bounds for pseudo-regret throughout this chapter, most of these bounds can be obtained for the expected regret as well. This follows from a standard technique that we detail in Appendix C.

OptDA+: OptDA with learning rate separation

For the sake of brevity, we will simply refer to pseudo-regret as "regret" hereinafter. Then, to ensure no-regret in the adversarial scenario, our algorithm of choice is OptDA+, the double-learning-rate variant of (OptDA). Its update is recursively stated as

$$X^i_{t+\frac{1}{2}} = X^i_t - \gamma^i_t g^i_{t-1}, \quad X^i_{t+1} = X^i_1 - \eta^i_{t+1} \sum_{s=1}^{t} g^i_s. \tag{OptDA+}$$

As before, γ^i_t and η^i_t are respectively the optimistic and the update learning rates of the algorithm. Note that (OG+) and (OptDA+) coincide when the update learning rate η^i_t is taken constant.

CONTRIBUTIONS AND OUTLINE. The focus of this chapter is on the following three types of results: bounds on sum of squared pseudo-gradient norms, bounds on regret, and last-iterate convergence under multiplicative noise. To achieve these, we lay out our preliminary inequalities for (OptDA+) in Section 8.1, which serve as the bedrock of our subsequent analysis. We then apply these inequalities to provide specific results for predetermined and adaptive learning rates in Section 8.2 and Section 8.3, respectively. In addition to the typical $O(\sqrt{T})$ regret guarantee for the adversarial scenario and the general noise model, we specifically show that both the regret and the sum of squared pseudo-gradient norms of (OptDA+) can be bounded by a *constant* when the noise is multiplicative. Finally, we corroborate our theoretical results with numerical illustrations in Section 8.4, demonstrating the efficacy of our methods in a bilinear game and a toy GAN model.

NOTATIONS RELATED TO THE LEARNING RATES. For any $\mathbf{x} = (x^i)_{i\in\mathcal{N}} \in \mathfrak{X} = \mathbb{R}^d$ and $\boldsymbol{\alpha} = (\alpha^i)_{i\in\mathcal{N}} \in \mathbb{R}^N_+$, we write the weighted norm as $\|\mathbf{x}\|_{\boldsymbol{\alpha}} = \sqrt{\sum_{i=1}^N \alpha^i \|x^i\|^2}$. The weights $\boldsymbol{\alpha}$ will be taken as a function of the learning rates. It will thus be convenient to write $\boldsymbol{\eta}_t = (\eta^i_t)_{i\in\mathcal{N}}$ and $\boldsymbol{\gamma}_t = (\gamma^i_t)_{i\in\mathcal{N}}$ for the joint learning rates. The arithmetic manipulation and the comparisons of these vectors should be taken elementwisely. For example, the element-wise division is $1/\boldsymbol{\eta}_t = (1/\eta^i_t)_{i\in\mathcal{N}}$. Provided that $\boldsymbol{\eta}_1$ is not needed for the update of (OptDA+), we will simply use the notation $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$ for our analysis.

8.1 PRELIMINARY INEQUALITIES

The goal of this section is provide a set of template inequalities that hold under minimal constraints on the learning rates. Importantly, although we have implicitly assumed that our learning rates are deterministic, i.e., $\mathcal{F}_0$-measurable up to now, this will not be the case for learning rates that are adjusted adaptively based on stochastic feedback. To account for this, we make the following assumption concerning the measurability of the learning rates in this section.

Measurability of learning rates

Assumption 8.1. For all $t \in \mathbb{N}$, the learning rates γ^i_{t+1} and η^i_{t+1} are $\mathcal{F}^i_t$-measurable.

Assumption 8.1 essentially suggests that γ^i_{t+1} and η^i_{t+1}, which are respectively used in the computation of $X^i_{t+\frac{3}{2}}$ and X^i_{t+1}, cannot be defined by incorporating any information that is only available from time t. This is slightly stricter than the natural assumption which posits that a player can only determine their learning rates based on received information, as it excludes the possibility

of using g_t^i in defining γ_{t+1}^i and η_{t+1}^i (recall that g_t^i is not $\mathcal{F}_t$-measurable but $\mathcal{F}_{t+1}$-measurable). Nonetheless, it guarantees that $\mathbb{E}_t[\gamma_{t+1}^i \eta_{t+1}^i \xi_{t+\frac{1}{2}}^i] = 0$, which is crucial for our analysis.

8.1.1 *Generalized OptDA+*

Similar to Section 7.2.2, we first analyze a generalized version of (OptDA+) that operates with two arbitrary sequences of vectors $(\tilde{g}_t)_{t\in\mathbb{N}}$ and $(g_t)_{t\in\mathbb{N}}$.

$$X_{t+\frac{1}{2}} = X_t - \gamma_t \tilde{g}_t, \quad X_{t+1} = X_1 - \eta_{t+1} \sum_{s=1}^{t} g_s. \tag{Generalized OptDA+}$$

Our goal here is to establish an (in)equality that would serve the same role as what Proposition 7.2 does in the analysis of (OG+). This necessitates incorporating the weight $1/\eta_t$ into the definition of the Lyapunov function. With the notation $\eta_1 = \eta_2$, some basic calculations give us the following proposition.

A basic decomposition for Generalized OptDA+

Proposition 8.1. *Let $(X_t)_{t\in\mathbb{N}}$ and $(X_{t+\frac{1}{2}})_{t\in\mathbb{N}}$ be generated by Generalized OptDA+. It holds for any $z \in \mathcal{X}$ and $t \in \mathbb{N}$ that*

$$\begin{aligned}\frac{\|X_{t+1} - z\|^2}{\eta_{t+1}} &= \frac{\|X_t - z\|^2}{\eta_t} - \frac{\|X_t - X_{t+1}\|^2}{\eta_t} \\ &\quad + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|X_1 - z\|^2 - \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|X_1 - X_{t+1}\|^2 \\ &\quad - 2\langle g_t, X_{t+\frac{1}{2}} - z\rangle - 2\gamma_t\langle g, \tilde{g}_t\rangle + 2\langle g_t, X_t - X_{t+1}\rangle.\end{aligned}$$

Proof. Using $g_t = (X_t - X_1)/\eta_t - (X_{t+1} - X_1)/\eta_{t+1}$, we can write

$$\begin{aligned}\langle g_t, X_{t+1} - z\rangle &= \left\langle \frac{X_t - X_1}{\eta_t} - \frac{X_{t+1} - X_1}{\eta_{t+1}}, X_{t+1} - z \right\rangle \\ &= \frac{1}{\eta_t}\langle X_t - X_{t+1}, X_{t+1} - z\rangle + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\langle X_1 - X_{t+1}, X_{t+1} - z\rangle \\ &= \frac{1}{2\eta_t}(\|X_t - z\|^2 - \|X_{t+1} - z\|^2 - \|X_t - X_{t+1}\|^2) \\ &\quad + \left(\frac{1}{2\eta_{t+1}} - \frac{1}{2\eta_t}\right)(\|X_1 - z\|^2 - \|X_{t+1} - z\|^2 - \|X_1 - X_{t+1}\|^2).\end{aligned}$$

Multiplying the equality by 2 and rearranging, we get

$$\begin{aligned}\frac{\|X_{t+1} - z\|^2}{\eta_{t+1}} &= \frac{\|X_t - z\|^2}{\eta_t} - \frac{\|X_t - X_{t+1}\|^2}{\eta_t} + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|X_1 - z\|^2 \\ &\quad - \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|X_1 - X_{t+1}\|^2 - 2\langle g_t, X_{t+1} - z\rangle.\end{aligned}$$

We conclude with the equality

$$\begin{aligned}\langle g_t, X_{t+1} - z\rangle &= \langle g_t, X_{t+1} - X_t\rangle + \langle g_t, X_t - X_{t+\frac{1}{2}}\rangle + \langle g_t, X_{t+\frac{1}{2}} - z\rangle \\ &= \langle g_t, X_{t+1} - X_t\rangle + \gamma_t\langle g_t, \tilde{g}_t\rangle + \langle g_t, X_{t+\frac{1}{2}} - z\rangle,\end{aligned}$$

where we have used $X_t = X_{t+\frac{1}{2}} + \gamma_t \tilde{g}_t$. □

With the standard assumption on the non-increasingness of the learning rates, we can then further refine the above the result into the following corollary.

Corollary 8.2. *Let $(X_t)_{t\in\mathbb{N}}$ and $(X_{t+\frac{1}{2}})_{t\in\mathbb{N}}$ be generated by Generalized OptDA+. For any $z \in \mathcal{X}$ and $t \in \mathbb{N}$, if $\eta_{t+1} \le \eta_t$, it holds that*

$$\frac{\|X_{t+1}-z\|^2}{\eta_{t+1}} \le \frac{\|X_t-z\|^2}{\eta_t} + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|X_1-z\|^2 - 2\langle g_t, X_{t+\frac{1}{2}}-z\rangle \\ - 2\gamma_t\langle g, \tilde{g}_t\rangle + \eta_t^2\|g_t\|^2 + \min\left(\eta_t\|g_t\|^2 - \frac{\|X_t-X_{t+1}\|^2}{2\eta_t}, 0\right).$$

Proof. This is immediate from Proposition 8.1 by applying Young's inequality. More precisely, we use $(1/\eta_{t+1} - 1/\eta_t)\|X_1 - X_{t+1}\|^2 \ge 0$ and

$$2\langle g_t, X_t - X_{t+1}\rangle \\ \le \min\left(\eta_t\|g_t\|^2 + \frac{\|X_t-X_{t+1}\|^2}{\eta_t}, 2\eta_t\|g_t\|^2 + \frac{\|X_t-X_{t+1}\|^2}{2\eta_t}\right). \qquad \square$$

In, Corollary 8.2 we recognize the two scalar product terms $\langle g_t, X_{t+\frac{1}{2}} - z\rangle$, $\langle g_t, \tilde{g}_t\rangle$, and the squared norm term $\|g_t\|^2$ that also appear in Proposition 7.2. The absence of player-dependent coefficient in front of $\langle g_t, X_{t+\frac{1}{2}} - z\rangle$ enables the use of player-dependent learning rates in our analysis, as we may just sum this up from $i = 1$ to N and take expectation to get $\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{z}\rangle$. This term is non-negative under Assumption 5.3 by choosing $\mathbf{z} \leftarrow \mathbf{x}_\star$ to be a Nash equilibrium. While it is also possible to put Proposition 7.2 in this form, this gives rise to an additional term $(1/\eta_{t+1} - 1/\eta_t)\|X_t - z\|^2$ and it is unclear how we can control it.

8.1.2 *Inequalities for OptDA+*

The success of our analysis for double-learning-rate optimistic gradient methods hinges on our careful treatment of the scalar product $\langle g_t, \tilde{g}_t\rangle$. Below we show that the bound proved in Lemma 7.5 for this quantity is still valid for (OptDA+).

Lemma 8.3. *Suppose that Assumptions 5.2 and 7.1 hold and all players run* (OptDA+) *with learning rates satisfying Assumption 8.1. Then, for all $i \in \mathcal{N}$ and $t \ge 2$, it holds*

$$-2\,\mathbb{E}_{t-1}[\langle \hat{V}^i_{t+\frac{1}{2}}, \hat{V}^i_{t-\frac{1}{2}}\rangle] \le \mathbb{E}_{t-1}\Bigg[- \|V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2 - \|V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\ + \|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\ + L\left(2\gamma_t^i\sqrt{N}\|\xi^i_{t-\frac{1}{2}}\|^2 + \sum_{j=1}^N \frac{(\gamma_t^j+\eta_t^j)^2\|\xi^j_{t-\frac{1}{2}}\|^2}{2\sqrt{N}\gamma_t^i}\right)\Bigg]$$

Proof. This lemma is proved in the same way as Lemma 7.5. The only difference is in that with $\tilde{X}^j_{t+\frac{1}{2}} = X^j_{t+\frac{1}{2}} + (\eta_t^j + \gamma_t^j)\xi^j_{t-\frac{1}{2}}$, we now have

$$\tilde{X}^j_{t+\frac{1}{2}} = X_1^i - \eta_t^j \sum_{s=1}^{t-2} \hat{V}^j_{s+\frac{1}{2}} - (\eta_t^j + \gamma_t^j)V^j(\mathbf{X}_{t-\frac{1}{2}}).$$

Thanks to Assumption 8.1, the surrogate $\tilde{\mathbf{X}}_{t+\frac{1}{2}}$ is always $\mathcal{F}_{t-1}$-measurable and the proof thus remains valid. □

Equipped with Corollary 8.2 and Lemma 8.3, we next establish a series of energy inequalities and regret and pseudo-gradient bounds for (OptDA+). We start with the energy inequality at the level of each individual.

Individual energy inequality for OptDA+

Lemma 8.4. *Suppose that Assumptions 5.2 and 7.1 hold and all players run (OptDA+) with non-increasing learning rates satisfying Assumption 8.1. Then, for all $i \in \mathcal{N}$, $t \geq 2$, and $z^i \in \mathcal{X}^i$, it holds*

$$\mathbb{E}_{t-1}\left[\frac{\|X^i_{t+1} - z^i\|^2}{\eta^i_{t+1}}\right] \leq \mathbb{E}_{t-1}\left[\frac{\|X^i_t - z^i\|^2}{\eta^i_t} + \left(\frac{1}{\eta^i_{t+1}} - \frac{1}{\eta^i_t}\right)\|X^i_1 - z^i\|^2\right. \tag{8.1a}$$

$$- 2\langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X^i_{t+\frac{1}{2}} - z^i\rangle \tag{8.1b}$$

$$- \gamma^i_t(\|V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2 + \|V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2) \tag{8.1c}$$

$$+ \gamma^i_t\|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 + \eta^i_t\|\hat{V}^i_{t+\frac{1}{2}}\|^2 \tag{8.1d}$$

$$+ \min\left(-\frac{\|X^i_t - X^i_{t+1}\|^2}{2\eta^i_t} + \eta^i_t\|\hat{V}^i_{t+\frac{1}{2}}\|^2, 0\right) \tag{8.1e}$$

$$\left. + 2(\gamma^i_t)^2\sqrt{N}L\|\xi^i_{t-\frac{1}{2}}\|^2 + \frac{L}{2\sqrt{N}}\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|^2_{(\boldsymbol{\eta}_t+\boldsymbol{\gamma}_t)^2}\right]. \tag{8.1f}$$

Proof. This is an immediate by combining Corollary 8.2 and Lemma 8.3. We just notice that as γ^i_t is $\mathcal{F}_{t-1}$-measurable, we have $\mathbb{E}_{t-1}[\gamma^i_t\langle\hat{V}^i_{t+\frac{1}{2}}, \hat{V}^i_{t-\frac{1}{2}}\rangle] = \gamma^i_t\,\mathbb{E}_{t-1}[\langle\hat{V}^i_{t+\frac{1}{2}}, \hat{V}^i_{t-\frac{1}{2}}\rangle]$. □

To gain insight on how Lemma 8.4 is used in our analysis, it is beneficial to dissect this bound and examine each term in detail.

- The weighted squared distance to z^i, i.e., $\|X^i_t - z^i\|^2/\eta^i_t$, telescopes in the analysis on aggregate performance measures (e.g., regret) and otherwise plays the role of energy in equilibrium convergence analysis.
- $(1/\eta^i_{t+1} - 1/\eta^i_t)\|X^i_1 - z^i\|^2$ in (8.1a) also telescopes in the analysis on aggregate performance measures. This leads to a term in the order of $1/\eta^i_T$ when we sum to $t = T$. However, it plays a more delicate role in convergence analysis, and this is what prevents us from showing last-iterate convergence of the algorithm when $\sigma_A > 0$.
- The linearized regret of each player is obtained by summing the pairing terms in (8.1b). On the other hand, taking $\mathbf{x}_\star \in \mathfrak{X}_\star$, $z^i \leftarrow x^i_\star$, and summing from $i = 1$ to N, we obtain $-2\langle\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star\rangle$, which is non-positive by Assumption 5.3, and can thus be dropped from the inequality.
- The negative term in (8.1c) provides a consistent negative drift that partially cancels out the noise.
- Thanks to the smoothness assumption, the gradient variation in (8.1d) can be partially cancelled out by the negative path variation in (8.1e) when we sum up over t and i. This leaves out terms that are in the order of $\gamma^i_t(\gamma^j_t)^2$.

- The remaining terms in lines (8.1d), (8.1e), and (8.1f) are of the order $(\gamma_t^j)^2 + \eta_t^i$. To ensure that they are sufficiently small with respect to the decrease of (8.1c), we again need both $(\gamma_t^j)_{j\in\mathcal{N}}$ and η_t^i/γ_t^i to be small.

With the above in mind, we next provide a template regret bound that holds for a large range of learning rate sequences, laying out the foundation of our subsequent regret analysis.

Bound on linearized regret

Lemma 8.5. *Suppose that Assumptions 5.2 and 7.1 hold and all players run* (OptDA+) *with non-increasing learning rates satisfying Assumption 8.1 and $\boldsymbol{\eta}_t \leq \boldsymbol{\gamma}_t$ for all $t \in \mathbb{N}$. Then, for all $i \in \mathcal{N}$, $T \in \mathbb{N}$, and $z^i \in \mathcal{X}^i$, we have*

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X^i_{t+\frac{1}{2}} - z^i\rangle\right] \leq \mathbb{E}\left[\frac{\|X_1^i - z^i\|^2}{2\eta^i_{T+1}} + \frac{1}{2}\sum_{t=1}^{T}\eta_t^i\|\hat{V}^i_{t+\frac{1}{2}}\|^2 + \sum_{t=2}^{T}\gamma_t^i L^2\left(3\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2_{\gamma_t^2} + \frac{3}{2}\|\mathbf{X}_t - \mathbf{X}_{t-1}\|^2\right) + \sum_{t=2}^{T}((\gamma_t^i)^2\sqrt{N}L\|\xi^i_{t-\frac{1}{2}}\|^2 + \frac{L}{\sqrt{N}}\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|^2_{\gamma_t^2})\right].$$

Proof. Applying Lemma 8.4, dropping the non-positive terms in (8.1c), (8.1e), and taking total expectation gives

$$\mathbb{E}\left[\frac{\|X^i_{t+1} - z^i\|^2}{\eta^i_{t+1}}\right] \leq \mathbb{E}\left[\frac{\|X_t^i - z^i\|^2}{\eta_t^i} + \left(\frac{1}{\eta^i_{t+1}} - \frac{1}{\eta_t^i}\right)\|X_1^i - z^i\|^2 + \eta_t^i\|\hat{V}^i_{t+\frac{1}{2}}\|^2 - 2\langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X^i_{t+\frac{1}{2}} - z^i\rangle + \gamma_t^i\|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 + 2(\gamma_t^i)^2\sqrt{N}L\|\xi^i_{t-\frac{1}{2}}\|^2 + \frac{L}{2\sqrt{N}}\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|^2_{(\boldsymbol{\eta}_t+\boldsymbol{\gamma}_t)^2}\right]. \tag{8.2}$$

The above inequality holds for $t \geq 2$. As for $t = 1$, we notice that with $X_2^i = X_1^i - \eta_2^i\hat{V}^i_{3/2}$, we have in fact

$$\|X_2^i - z^i\|^2 = \|X_1^i - z^i\|^2 - 2\eta_2^i\langle\hat{V}^i_{3/2}, X_1^i - z^i\rangle + (\eta_2^i)^2\|\hat{V}^i_{3/2}\|^2.$$

As $X^i_{3/2} = X_1^i = 0$ and $\eta_1^i = \eta_2^i$, the above implies

$$\mathbb{E}\left[\langle V^i(\mathbf{X}_{3/2}), X^i_{3/2} - z^i\rangle\right] = \mathbb{E}\left[\frac{\|X_1^i - z^i\|^2}{2\eta_2^i} - \frac{\|X_2^i - z^i\|^2}{2\eta_2^i} + \frac{\eta_1^i\|\hat{V}^i_{3/2}\|^2}{2}\right]. \tag{8.3}$$

Summing (8.2) from $t = 2$ to T, dividing by 2, adding (8.3), and using $\boldsymbol{\eta}_t \leq \boldsymbol{\gamma}_t$ leads to

$$\sum_{t=1}^{T}\mathbb{E}[\langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X^i_{t+\frac{1}{2}} - z^i\rangle] \leq \frac{1}{2}\mathbb{E}\left[\frac{\|X_1^i - z^i\|^2}{\eta^i_{T+1}} + \sum_{t=1}^{T}\eta_t^i\|\hat{V}^i_{t+\frac{1}{2}}\|^2 + \sum_{t=2}^{T}\gamma_t^i\|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 + 2L\sum_{t=2}^{T}((\gamma_t^i)^2\sqrt{N}\|\xi^i_{t-\frac{1}{2}}\|^2 + \frac{1}{\sqrt{N}}\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|^2_{\gamma_t^2})\right].$$

Furthermore, thanks to Lipschitz continuity of V^i, we can bound the difference term by

$$\begin{aligned}\|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 &\leq 3\|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\mathbf{X}_t)\|^2 + 3\|V^i(\mathbf{X}_t) - V^i(\mathbf{X}_{t-1})\|^2 \\ &\quad + 3\|V^i(\mathbf{X}_{t-1}) - V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 \\ &\leq 3L^2\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2_{\gamma_t^2} + 3L^2\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2_{(\gamma_{t-1})^2} + 3L^2\|\mathbf{X}_t - \mathbf{X}_{t-1}\|^2. \end{aligned} \tag{8.4}$$

Combining the above two inequalities and using $\hat{\mathbf{V}}_{1/2} = 0$ gives the desired inequality. □

In Sections 8.2 and 8.3, we will use Lemma 8.5 to bound each player's regret in the case where they all play (OptDA+). Nonetheless, as we can see from the statement, this in turn requires bounds on the second-order path length and on the sum of the squared pseudo-gradient norms. As a stepping stone toward such bound, we derive a global energy inequality for (OptDA+), which additionally, as in Chapter 7, will also server as an important building block for the derivation of trajectory convergence results.

Global energy inequality for OptDA+

Lemma 8.6. *Suppose that Assumptions 5.2, 5.3 and 7.1 hold and all players run (OptDA+) with non-increasing learning rates satisfying Assumption 8.1. Then, for all $t \geq 2$ and $\mathbf{x}_\star \in \mathfrak{X}_\star$, if $\boldsymbol{\eta}_t \leq \boldsymbol{\gamma}_t$, we have*

$$\begin{aligned}\mathbb{E}_{t-1}[\|\mathbf{X}_{t+1} - \mathbf{x}_\star\|^2_{1/\boldsymbol{\eta}_{t+1}}] \leq \mathbb{E}_{t-1}[&\|\mathbf{X}_t - \mathbf{x}_\star\|^2_{1/\boldsymbol{\eta}_t} + \|\mathbf{X}_1 - \mathbf{x}_\star\|^2_{1/\boldsymbol{\eta}_{t+1} - 1/\boldsymbol{\eta}_t} \\ &- \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2_{\boldsymbol{\gamma}_t} - \|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2_{\boldsymbol{\gamma}_t} \\ &- \|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2_{1/(2\boldsymbol{\eta}_t)} + 3\|\mathbf{V}(\mathbf{X}_t) - \mathbf{V}(\mathbf{X}_{t-1})\|^2_{\boldsymbol{\gamma}_t} \\ &+ 3L^2(\|\boldsymbol{\gamma}_t\|_1\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2_{\boldsymbol{\gamma}_t^2} + \|\boldsymbol{\gamma}_{t-1}\|_1\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2_{(\boldsymbol{\gamma}_{t-1})^2}) \\ &+ 4\sqrt{N}L\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|^2_{\boldsymbol{\gamma}_t^2} + 2\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|^2_{\boldsymbol{\eta}_t}]. \end{aligned} \tag{8.5}$$

Proof. On one hand, we have

$$\min\left(-\frac{\|X_t^i - X_{t+1}^i\|^2}{2\eta_t^i} + \eta_t^i\|\hat{V}^i_{t+\frac{1}{2}}\|^2, 0\right) \leq -\frac{\|X_t^i - X_{t+1}^i\|^2}{2\eta_t^i} + \eta_t^i\|\hat{V}^i_{t+\frac{1}{2}}\|^2.$$

On the other hand, we use (8.4) but keep the $3\|V^i(\mathbf{X}_t) - V^i(\mathbf{X}_{t-1})\|^2$ term instead of bounding it from above. That is,

$$\begin{aligned}\|V^i(\mathbf{X}_{t+\frac{1}{2}}) - V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 &\leq 3L^2\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2_{\gamma_t^2} + 3L^2\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2_{(\gamma_{t-1})^2} \\ &\quad + 3\|V^i(\mathbf{X}_t) - V^i(\mathbf{X}_{t-1})\|^2. \end{aligned}$$

Plugging the previous two inequalities into Lemma 8.4 with $z^i \leftarrow x^i_\star$, summing from $i = 1$ to N, and using $\langle \mathbf{V}(\mathbf{X}_{t+\frac{1}{2}}), \mathbf{X}_{t+\frac{1}{2}} - \mathbf{x}_\star\rangle \geq 0$, $\boldsymbol{\eta}_t \leq \boldsymbol{\gamma}_t$, and $\|\boldsymbol{\gamma}_t\|_1 \leq \|\boldsymbol{\gamma}_{t-1}\|_1$ gives (8.5). □

Lemma 8.6 shares a lot of similarity with Lemma 7.7. There are however two important differences. First, as mentioned previously, the squared distance is now weighted by $1/\boldsymbol{\eta}_t$. This is common to energy inequalities for DA-type methods (see the proof of Proposition 2.4). However, it is also this very aspect that hinders our ability to prove last-iterate convergence of these methods

when additive noise is present, i.e., when $\sigma_A > 0$.[1] Second, the presence of $-\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2_{1/(2\eta_t)} + 3\|\mathbf{V}(\mathbf{X}_t) - \mathbf{V}(\mathbf{X}_{t-1})\|^2_{\gamma_t}$ is also be specific to (OptDA+). It is written in this form because unlike (OG+), we do not have simple ways to develop these variation terms here. Otherwise, they can still be controlled using the smoothness of the losses.

We close this section with a bound on the sum of pseudo-gradient norms. It can also be understood as a bound on the second-order path length as we have $-\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2_{1/(2\eta_t)}$ on the RHS of the inequality.

Bound on sum of squared pseudo-gradient norms

Lemma 8.7. *Suppose that Assumptions 5.2, 5.3 and 7.1 hold and all players run (OptDA+) with non-increasing learning rates satisfying Assumption 8.1 and $\boldsymbol{\eta}_t \leq \boldsymbol{\gamma}_t$ for all $t \in \mathbb{N}$. Then, for all $T \in \mathbb{N}$ and $\mathbf{x}_\star \in \mathfrak{X}_\star$, we have*

$$\begin{aligned}
&\sum_{t=2}^{T} \mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2_{\gamma_t} + \|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2_{\gamma_t}] \\
&\quad\leq \mathbb{E}\left[\|\mathbf{X}_1 - \mathbf{x}_\star\|^2_{1/\eta_{T+1}} + \sum_{t=1}^{T}\left(3\|\mathbf{V}(\mathbf{X}_t) - \mathbf{V}(\mathbf{X}_{t+1})\|^2_{\gamma_t} - \|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2_{1/(2\eta_t)}\right)\right. \\
&\qquad\left. + \sum_{t=2}^{T} 6\|\boldsymbol{\gamma}_t\|_1 L^2 \|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2_{\gamma_t^2} + \sum_{t=2}^{T} 4\sqrt{N}L\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|^2_{\gamma_t^2} + \sum_{t=1}^{T} 2\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|^2_{\eta_t}\right]. \quad (8.6)
\end{aligned}$$

Proof. This is a direct consequence of Lemma 8.6. In fact, taking total expectation of (8.5) and summing from $t = 2$ to T gives already

$$\begin{aligned}
&\sum_{t=2}^{T} \mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2_{\gamma_t} + \|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2_{\gamma_t}] \\
&\quad\leq \mathbb{E}\left[\|\mathbf{X}_2 - \mathbf{x}_\star\|^2_{1/\eta_2} + \|\mathbf{X}_1 - \mathbf{x}_\star\|^2_{1/\eta_{T+1} - 1/\eta_2}\right. \\
&\qquad + \sum_{t=2}^{T}(3\|\mathbf{V}(\mathbf{X}_t) - \mathbf{V}(\mathbf{X}_{t-1})\|^2_{\gamma_t} - \|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2_{1/(2\eta_t)}) \\
&\qquad\left. + \sum_{t=2}^{T} 6\|\boldsymbol{\gamma}_t\|_1 L^2 \|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2_{\gamma_t^2} + \sum_{t=2}^{T} 4\sqrt{N}L\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|^2_{\gamma_t^2} + \sum_{t=2}^{T} 2\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|^2_{\eta_t}\right]. \quad (8.7)
\end{aligned}$$

We have in particular used $\hat{\mathbf{V}}_{1/2} = 0$ to bound

$$\begin{aligned}
&\sum_{t=2}^{T} 3L^2(\|\boldsymbol{\gamma}_t\|_1 \|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2_{\gamma_t^2} + \|\boldsymbol{\gamma}_{t-1}\|_1 \|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|^2_{(\gamma_{t-1})^2}) \\
&\quad= \sum_{t=2}^{T} 3\|\boldsymbol{\gamma}_t\|_1 L^2 \|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2_{\gamma_t^2} + \sum_{t=3}^{T} \|\boldsymbol{\gamma}_t\|_1 3NL^2 \|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2_{\gamma_t^2} \\
&\quad\leq \sum_{t=2}^{T} 6\|\boldsymbol{\gamma}_t\|_1 L^2 \|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2_{\gamma_t^2}.
\end{aligned}$$

1 While we can try to put the inequality in the form of Lemma 7.7, the appearance of other additional terms would still undermine the convergence proof.

To obtain (8.6), we further bound

$$\sum_{t=2}^{T} 3\|\mathbf{V}(\mathbf{X}_t) - \mathbf{V}(\mathbf{X}_{t-1})\|^2_{\gamma_t} = \sum_{t=1}^{T-1} 3\|\mathbf{V}(\mathbf{X}_t) - \mathbf{V}(\mathbf{X}_{t+1})\|^2_{\gamma_{t+1}}$$
$$\leq \sum_{t=1}^{T} 3\|\mathbf{V}(\mathbf{X}_t) - \mathbf{V}(\mathbf{X}_{t+1})\|^2_{\gamma_t} \tag{8.8}$$

For $t = 1$, we use (8.3) with $z^i \leftarrow x^i_\star$; that is

$$\mathbb{E}\left[\frac{\|X^i_2 - x^i_\star\|^2}{\eta^i_2}\right] = \mathbb{E}\left[\frac{\|X^i_1 - x^i_\star\|^2}{\eta^i_2} - 2\langle V^i(\mathbf{X}_{3/2}), X^i_1 - x^i_\star\rangle + \eta^i_1\|\hat{V}^i_{3/2}\|^2\right].$$

Since $\mathbf{X}_{3/2} = \mathbf{X}_1$, summing the above inequality from $i = 1$ to N leads to

$$\mathbb{E}[\|\mathbf{X}_2 - \mathbf{x}_\star\|^2_{1/\eta_2}] = \mathbb{E}[\|\mathbf{X}_1 - \mathbf{x}_\star\|^2_{1/\eta_2} - 2\langle \mathbf{V}(\mathbf{X}_{3/2}), \mathbf{X}_{3/2} - \mathbf{x}_\star\rangle + \|\hat{\mathbf{V}}_{3/2}\|^2_{\eta_1}].$$

Assumption 5.3 ensures $\langle \mathbf{V}(\mathbf{X}_{3/2}), \mathbf{X}_{3/2} - \mathbf{x}_\star\rangle \geq 0$ and therefore

$$\mathbb{E}[\|\mathbf{X}_2 - \mathbf{x}_\star\|^2_{1/\eta_2}] \leq \mathbb{E}[\|\mathbf{X}_1 - \mathbf{x}_\star\|^2_{1/\eta_2} + \|\hat{\mathbf{V}}_{3/2}\|^2_{\eta_2}]$$
$$\leq \mathbb{E}[\|\mathbf{X}_1 - \mathbf{x}_\star\|^2_{1/\eta_2} + 2\|\hat{\mathbf{V}}_{3/2}\|^2_{\eta_1} - \|\mathbf{X}_1 - \mathbf{X}_2\|^2_{(1/2\eta_1)}]. \tag{8.9}$$

Combining (8.7), (8.8), and (8.9) gives exactly (8.6). □

8.2 OPTDA+ WITH PREDETERMINED LEARNING RATES

As a warm up, we consider in this section learning rate sequences that are fixed from the beginning of play, that is, the learning rates must be $\mathscr{F}_1$-measurable.[2] Similar to what we have seen in last chapter (Eq. (7.17) and Eq. (7.19)), the optimistic learning rate γ^i_t and the ratio η^i_t/γ^i_t should both be small enough. Precisely, in the case where all player play (OptDA+), we require the following inequalities to be satisfied for all $t \in \mathbb{N}$ and $i \in \mathcal{N}$.

Upper bound on learning rates and their ratio

$$\gamma^i_t \leq \frac{1}{2L}\min\left(\frac{1}{\sqrt{3N(1+\sigma^2_M)}}, \frac{1}{4\sqrt{N}\sigma^2_M}\right) \quad \text{and} \quad \eta^i_t \leq \frac{\gamma^i_t}{4(1+\sigma^2_M)}. \tag{8.10}$$

Compared to the results that we obtain in the next section for adaptive learning rates, we rely on weaker assumptions here and the constants involved in the bounds are systematically smaller.

8.2.1 *No-Regret Against Adversarial Opponents*

To begin, we first state a worst-case regret bound when played against arbitrary opponents.

Theorem 8.8. *Suppose that Assumptions 5.1 and 7.1 hold and player i run* (OptDA+) *with non-increasing learning rates $\gamma^i_t = \Theta(1/t^{\frac{1}{2}-r})$ and $\eta^i_t = \Theta(1/\sqrt{t})$ for some*

Regret bound against bounded adversarial feedback

2 To avoid redundancy, we will not restate this when presenting the results of this section.

$r \in [0, 1/4]$. Then, if there exists $G \in \mathbb{R}_+$ such that $\sup_{\mathbf{x}\in\mathbb{R}^d} \|V^i(\mathbf{x})\| \le G$, it holds for any bounded set $\mathcal{Z}^i$ with $R^2 \ge \sup_{z^i\in\mathcal{Z}^i} \mathbb{E}[\|X_1^i - z^i\|^2]$ that

$$\overline{\mathrm{Reg}}_T^i(\mathcal{Z}^i) = O\left(R^2\sqrt{T} + ((1+\sigma_M^2)G^2 + \sigma_A^2)T^{\frac{1}{2}+r}\right).$$

Proof. Let $z^i \in \mathcal{Z}^i$. From Corollary 8.2 and Young's inequality we get

$$\begin{aligned}
\langle \hat{V}^i_{t+\frac{1}{2}}, X^i_{t+\frac{1}{2}} - z^i\rangle &\le \frac{\|X_t^i - z^i\|^2}{2\eta_t^i} - \frac{\|X_{t+1}^i - z^i\|^2}{2\eta_{t+1}^i} - \frac{\|X_t^i - X_{t+1}^i\|^2}{2\eta_t^i} \\
&\quad + \left(\frac{1}{2\eta_{t+1}^i} - \frac{1}{2\eta_t^i}\right)\|X_1^i - z^i\|^2 - \gamma_t^i\langle \hat{V}^i_{t+\frac{1}{2}}, \hat{V}^i_{t-\frac{1}{2}}\rangle + \eta_t^i\|\hat{V}^i_{t+\frac{1}{2}}\|^2 \\
&\le \frac{R^2}{2\eta_t^i} - \frac{\|X_{t+1}^i - z^i\|^2}{2\eta_{t+1}^i} - \frac{\|X_t^i - X_{t+1}^i\|^2}{2\eta_t^i} + \eta_t^i\|\hat{V}^i_{t+\frac{1}{2}}\|^2 \\
&\quad + \left(\frac{1}{2\eta_{t+1}^i} - \frac{1}{2\eta_t^i}\right)\|X_1^i - z^i\|^2 + \frac{\gamma_t^i}{2}(\|\hat{V}^i_{t+\frac{1}{2}}\|^2 + \|\hat{V}^i_{t-\frac{1}{2}}\|^2)
\end{aligned}$$

As $\mathbf{V}^i_{1/2} = 0$ and $\eta_1^i = \eta_2^i$, summing the above from $t = 1$ to T gives

$$\sum_{t=1}^T \langle \hat{V}^i_{t+\frac{1}{2}}, X^i_{t+\frac{1}{2}} - z^i\rangle \le \frac{\|X_1^i - z^i\|^2}{2\eta_{T+1}^i} - \sum_{t=1}^T \frac{\|X_t^i - X_{t+1}^i\|^2}{2\eta_t^i} + \sum_{t=1}^T(\gamma_t^i + \eta_t^i)\|\hat{V}^i_{t+\frac{1}{2}}\|^2. \tag{8.11}$$

Dropping the non-positive term and taking expectation leads to

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X^i_{t+\frac{1}{2}} - z^i\rangle\right] \\
&\le \mathbb{E}\left[\frac{\|X_1^i - z^i\|^2}{2\eta_{T+1}^i} + \sum_{t=1}^T(\gamma_t^i + \eta_t^i)((1+\sigma_M^2)\|V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2 + \sigma_A^2)\right] \\
&\le \frac{R^2}{2\eta_{T+1}^i} + \sum_{t=1}^T(\gamma_t^i + \eta_t^i)((1+\sigma_M^2)G^2 + \sigma_A^2)
\end{aligned}$$

The claim then follows immediately from the choice of the learning rates. □

Remark 8.2. Instead of assuming the operator to be bounded on the entire space, we may simply assume the feedback to be bounded as done in Theorem 6.4.

The no-regret guarantee provided in Theorem 8.8 should be no surprise to the readers as it follows almost immediately from the standard analysis of optimistic gradient methods for online learning. The only subtlety is that we additionally introduce an exponent r. From the proposition we see clearly that taking smaller r (i.e., smaller optimistic step) is more favorable in the adversarial regime. This is because arbitrarily different successive feedback may make the optimistic step harmful rather than helpful. Nonetheless, as we shall see in the next section, taking larger r (i.e., larger optimistic steps) may be more beneficial when all the players use (OptDA+).

8.2.2 *Fast Convergence of Pseudo-Gradient in Self-Play*

We next shift our attention to the case where all the players take their actions according to (OptDA+). To begin, we present our results that quantify the performance of the algorithm with bounds on pseudo-gradient norms.

Bound on pseudo-gradient norm

Theorem 8.9. *Suppose that Assumptions 5.2, 5.3, 7.1 and 7.2 hold and all players run (OptDA+) with non-increasing learning rate sequences $(\gamma_t^i)_{t\in\mathbb{N}}$ and $(\eta_t^i)_{t\in\mathbb{N}}$ satisfying (8.10). We have*

(a) If there exists $r \in [0, 1/4]$ such that $\gamma_t^j = O(1/t^{1/4})$, $\gamma_t^j = \Omega(1/t^{\frac{1}{2}-r})$, and $\eta_t^j = \Theta(1/\sqrt{t})$ for all $j \in \mathcal{N}$, then

$$\sum_{t=1}^{T} \mathbb{E}[\|\mathbf{V}(X_{t+\frac{1}{2}})\|^2] = O\left(T^{1-r}\right)$$

(b) If the noise is multiplicative (i.e., $\sigma_A = 0$) and the learning rates are constant $\gamma_t \equiv \gamma$, $\eta_t \equiv \eta$, then

$$\sum_{t=1}^{T} \mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2] \leq \frac{2}{\min_{i\in\mathcal{N}} \gamma^i} \mathbb{E}\left[\mathrm{dist}_{1/\eta}(\mathbf{X}_1, \mathfrak{X}_\star)^2 + \|\mathbf{V}(\mathbf{X}_1)\|_{\gamma_t}^2\right]$$

In particular, if the equalities hold in (8.10), then the above is in $O(N^2L^2(1+\sigma_M^2)^3)$.

Theorem 8.19 provides an indicator on the "convergence speed" of the algorithm. However, it is not really a guarantee on the last iterate $\mathbf{X}_t$; instead, it measures the average quality of the iterates and suggests that the trajectory of play would get arbitrarily close to the set of equilibria over time (while it can still oscillate between being close and being far from the equilibria).

The case of general noise

Precisely, when an additive component is present in the noise, the best we can achieve is with the choice $r = 1/4$. This gives a rate of $O(t^{-1/4})$ for the average squared pseudo-gradient norm, and this rate worsens as we decrease r, that is, when we take smaller optimistic step. In the extreme case, with $r = 0$, we end up with a trivial $O(T)$ bound on the sum and thus no guarantee on the convergence rate. In the meantime, it is this very component that allows us to get the optimal $O(\sqrt{T})$ regret in Theorem 8.8 when faced with adversarial opponents. This reveals a tension between the adversarial and the self-play setups when it comes to choosing r, as already discussed in Section 8.2.1.

The case of multiplicative noise

When the noise is multiplicative, we can further obtain a constant bound on the sum and thus an $O(1/t)$ rate for the average squared pseudo-gradient norm. This represents a dramatic improvement in performance, and matches the rate of the algorithm run with perfect feedback [35, 103]. Of course, there is no hope to guarantee no regret when using constant learning rates. This will be addressed by our adaptive learning rate policy of Section 8.3.

second-order path length

Another metric on the trajectory that we may consider here is the second-order path length $\sum_{t=1}^{+\infty} \|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2$. When this quantity is smaller, the iterates move less from one one round to another and are thus more "stable". In the following lemma, we first provide an bound on both the sum of squared pseudo-gradient norms and the second-order path length without specifying the form of the learning rates.

Lemma 8.10. *Suppose that Assumptions 5.2, 5.3 and 7.1 hold and all players run* (OptDA+) *with non-increasing learning rates satisfying* (8.10). *Then, for all* $T \in \mathbb{N}$ *and* $\mathbf{x}_\star \in \mathfrak{X}_\star$, *we have*

$$\begin{aligned}\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_{\gamma_t}^2] + \sum_{t=1}^{T} 21\|\gamma_1\|_\infty NL^2\,\mathbb{E}[\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2]\\ \leq \mathbb{E}[\|\mathbf{X}_1 - \mathbf{x}_\star\|_{1/\eta_{T+1}}^2 + \|\mathbf{V}(\mathbf{X}_1)\|_{\gamma_1}^2]\\ + \sum_{t=1}^{T}\left(6\|\gamma_t\|_\infty^3 NL^2 + 4\|\gamma_t\|_\infty^2\sqrt{N}L + 2\|\eta_t\|_\infty\right)N\sigma_A^2\end{aligned}$$

Proof. We first apply Lemma 8.7 to obtain (8.6). We bound the expectations of the following three terms separately.

$$\begin{aligned}A_t &= 3\|\mathbf{V}(\mathbf{X}_t) - \mathbf{V}(\mathbf{X}_{t+1})\|_{\gamma_t}^2 - \|\mathbf{X}_t - \mathbf{X}_{t+1}\|_{1/(2\eta_t)}^2,\\ B_t &= 6\|\gamma_t\|_1 L^2\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|_{\gamma_t^2}^2 + 4\sqrt{N}L\|\xi_{t-\frac{1}{2}}\|_{\gamma_t^2}^2, \quad C_t = 2\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|_{\eta_t}^2.\end{aligned}$$

To bound A_t, we first use $\eta_t \leq \gamma_t/(4(1+\sigma_M^2)) \leq \|\gamma_1\|_\infty/(4(1+\sigma_M^2))$ to get

$$\|\mathbf{X}_t - \mathbf{X}_{t+1}\|_{1/(2\eta_t)}^2 \geq \frac{2(1+\sigma_M^2)}{\|\gamma_1\|_\infty}\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2.$$

Moreover, with $\|\gamma_1\|_\infty^2 \leq 1/(12NL^2(1+\sigma_M^2))$ we have

$$\frac{2(1+\sigma_M^2)}{\|\gamma_1\|_\infty} \geq 24NL^2(1+\sigma_M^2)^2\|\gamma_1\|_\infty \geq 24NL^2\|\gamma_1\|_\infty.$$

On the other hand, with the Lipschitz continuity of $(V^i)_{i\in\mathcal{N}}$ it holds

$$3\|\mathbf{V}(\mathbf{X}_t) - \mathbf{V}(\mathbf{X}_{t+1})\|_{\gamma_t}^2 \leq \sum_{i=1}^{N} 3\gamma_t^i L^2\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2 \leq 3\|\gamma_1\|_\infty NL^2\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2.$$

Combining the above inequalities we deduce that $A_t \leq -21\|\gamma_1\|_\infty NL^2\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2$ and accordingly

$$\mathbb{E}[A_t] \leq \mathbb{E}[-21\|\gamma_1\|_\infty NL^2\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2]. \tag{8.12}$$

We proceed to bound $\mathbb{E}[B_t]$. Using Assumption 7.1 and the law of total expectation, we get

$$\begin{aligned}\mathbb{E}[B_t] &= \mathbb{E}\left[\,\mathbb{E}_{t-1}[6\|\gamma_t\|_1 L^2\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|_{\gamma_t^2} + 4\sqrt{N}L\|\xi_{t-\frac{1}{2}}\|_{\gamma_t}^2]\right]\\ &= \mathbb{E}\left[\sum_{i=1}^{N}\left(6\|\gamma_t\|_1(\gamma_t^i)^2L^2\,\mathbb{E}_{t-1}[\|\hat{V}_{t-\frac{1}{2}}^i\|^2] + 4(\gamma_t^i)^2\sqrt{N}L\,\mathbb{E}_{t-1}[\|\xi_{t-\frac{1}{2}}^i\|^2]\right)\right]\\ &\leq \mathbb{E}\left[6\|\gamma_t\|_\infty^2 NL^2(1+\sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|_{\gamma_t}^2 + 4\|\gamma_t\|_\infty\sqrt{N}L\sigma_M^2\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|_{\gamma_t}^2\right.\\ &\qquad\left. + (6\|\gamma_t\|_\infty^3 NL^2 + 4\|\gamma_t\|_\infty^2\sqrt{N}L)N\sigma_A^2\right].\end{aligned} \tag{8.13}$$

Similarly, $\boldsymbol{\eta}_{t+1}$ being $\mathscr{F}_1$-measurable and in particular $\mathscr{F}_t$-measurable, we have

$$\mathbb{E}[C_t] = \mathbb{E}[\mathbb{E}_t[2\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|^2_{\boldsymbol{\eta}_t}]] \leq \mathbb{E}\left[2(1+\sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2_{\boldsymbol{\eta}_t} + 2\|\boldsymbol{\eta}_t\|_\infty N\sigma_A^2\right]. \quad (8.14)$$

Putting together (8.6), (8.12), (8.13), and (8.14), we get

$$\begin{aligned}
&\sum_{t=2}^{T} \mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2_{\boldsymbol{\gamma}_t - 2(1+\sigma_M^2)\boldsymbol{\eta}_t} + (1 - a_t(1+\sigma_M^2) - b_t\sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2_{\boldsymbol{\gamma}_t}] \\
&\quad \leq \mathbb{E}\left[\|\mathbf{X}_1 - \mathbf{x}_\star\|^2_{1/\boldsymbol{\eta}_{T+1}} + 2(1+\sigma_M^2)\|\mathbf{V}(\mathbf{X}_{3/2})\|^2_{\boldsymbol{\eta}_1} - \sum_{t=1}^{T} 21\|\boldsymbol{\gamma}_1\|_\infty NL^2\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2\right. \\
&\qquad \left. + \sum_{t=1}^{T}(a_t + b_t + 2\|\boldsymbol{\eta}_t\|_\infty)N\sigma_A^2\right],
\end{aligned}$$

where $a_t = 6\|\boldsymbol{\gamma}_t\|^3_\infty NL^2$ and $b_t = \|\boldsymbol{\gamma}_t\|^2_\infty 4\sqrt{N}L$. We conclude by using $\mathbf{X}_{3/2} = \mathbf{X}_1$ and noticing that under our learning rate requirement it is always true that $1 - 6\|\boldsymbol{\gamma}_t\|^2_\infty NL^2(1+\sigma_M^2) - 4\|\boldsymbol{\gamma}_t\|_\infty\sqrt{N}L\sigma_M^2 \geq 0$ and $\boldsymbol{\gamma}_t - 2(1+\sigma_M^2)\boldsymbol{\eta}_t \geq \boldsymbol{\gamma}_t/2$. □

To prove Theorem 8.9, we then instantiate Lemma 8.10 with specific learning rates for both $\sigma_A \neq 0$ and for $\sigma_A = 0$.

Proof of Theorem 8.9. Let us define $a_t = 6\|\boldsymbol{\gamma}_t\|^3_\infty NL^2 + 4\|\boldsymbol{\gamma}_t\|^2_\infty\sqrt{N}L + 2\|\boldsymbol{\eta}_t\|_\infty$. From Lemma 8.10 we know that for all $\mathbf{x}_\star \in \mathfrak{X}_\star$, it holds

$$\sum_{s=1}^{T} \mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2_{\boldsymbol{\gamma}_t/2}]] \leq \|\mathbf{X}_1 - \mathbf{x}_\star\|^2_{1/\boldsymbol{\eta}_{T+1}} + \|\mathbf{V}(\mathbf{X}_1)\|^2_{\boldsymbol{\gamma}_1} + \sum_{t=1}^{T} a_t N\sigma_A^2,$$

Since the learning rates are decreasing, we can lower bound $\boldsymbol{\gamma}_t$ by $\boldsymbol{\gamma}_t \geq \boldsymbol{\gamma}_T \geq \min_{i\in\mathcal{N}} \gamma_T^i$. Accordingly,

$$\sum_{s=1}^{T} \mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2]] \leq \frac{2}{\min_{i\in\mathcal{N}} \gamma_T^i}\left(\|\mathbf{X}_1 - \mathbf{x}_\star\|^2_{1/\boldsymbol{\eta}_{T+1}} + \|\mathbf{V}(\mathbf{X}_1)\|^2_{\boldsymbol{\gamma}_1} + \sum_{t=1}^{T} a_t N\sigma_A^2\right),$$

The result then follows immediately from our learning rate choices. For (a), we observe with $\|\boldsymbol{\gamma}_t\|_\infty = O(1/t^{\frac{1}{4}})$ and $\|\boldsymbol{\eta}_t\|_\infty = O(1/\sqrt{t})$, we have $\sum_{t=1}^{T} a_t = O(\sqrt{T})$, while $\gamma_t^j = \Omega(1/t^{\frac{1}{2}-r})$, and $\eta_t^j = \Omega(1/\sqrt{t})$ guarantee $1/\min_{i\in\mathcal{N}} \gamma_t^T = O(T^{\frac{1}{2}-r})$ and $1/\min_{i\in\mathcal{N}} \eta_{t+1}^T = O(\sqrt{T})$. For (b), we take $\mathbf{x}_\star = \arg\min_{\mathbf{x}\in\mathfrak{X}_\star} \|\mathbf{X}_1 - \mathbf{x}\|_{1/\boldsymbol{\eta}}$. □

8.2.3 *Improved Regret in Self-Play*

Moving on, we dig into the regret guarantee for each individual player.

Regret bound in self-play

Theorem 8.11. *Suppose that Assumptions 5.1–5.3, 7.1 and 7.2 hold and all players run* (OptDA+) *with non-increasing learning rate sequences* $(\gamma_t^i)_{t\in\mathbb{N}}$ *and* $(\eta_t^i)_{t\in\mathbb{N}}$ *satisfying* (8.10). *For any* $i \in \mathcal{N}$ *and bounded set* $\mathcal{Z}^i \subset \mathcal{X}^i$ *with* $R^2 \geq \sup_{z^i\in\mathcal{Z}^i} \mathbb{E}[\|X_1^i - z^i\|^2]$ *, we have*

(a) If $\gamma_t^j = O(1/t^{\frac{1}{4}})$ *and* $\eta_t^j = \Theta(1/\sqrt{t})$ *for all* $j \in \mathcal{N}$*, then*

$$\overline{\mathrm{Reg}}_T^i(\mathcal{Z}^i) = O\left(\sqrt{T}\right).$$

(b) *If the noise is multiplicative (i.e., $\sigma_A = 0$) and the learning rates are constant $\gamma_t \equiv \gamma$, $\eta_t \equiv \eta$, then*

$$\overline{\mathrm{Reg}}_T^i(\mathcal{Z}^i) \leq \frac{R^2}{2\eta^i} + \frac{5}{4}\,\mathbb{E}\left[\mathrm{dist}_{1/\eta}(\mathbf{X}_1, \mathfrak{X}_\star)^2 + \|\mathbf{V}(\mathbf{X}_1)\|_\gamma^2\right].$$

In particular, if the equalities hold in (8.10), the above is in $O(N^2L(1+\sigma_M^2)^2)$.

The case of general noise

The first part of Theorem 8.11 guarantees the standard $O(\sqrt{T})$ regret in the presence of additive noise, in accordance with existing results in the literature. Moreover, unlike Theorem 8.9, this result holds for any $\gamma_t^j = O(1/t^{\frac{1}{4}})$. This is not surprising providing that we can already guarantee the $O(\sqrt{T})$ regret for (DA). Note also that Theorem 8.11(a) is not a consequence of Theorem 8.8 as there is no a priori reason for the feedback of the players to be bounded.

The case of multiplicative noise

What is more surprising is the second part of Theorem 8.11 which shows that when the noise is multiplicative (i.e., when $\sigma_A = 0$), it is still possible to achieve constant regret. This is of course closely related to the fast convergence results that we have shown in Theorem 8.9. More precisely, we make use of Lemma 8.10. For that, we first refine Lemma 8.5 as follows.

Lemma 8.12. *Suppose that Assumptions 5.2 and 7.1 hold and all players run (OptDA+) with non-increasing learning rates satisfying (8.10). Then, for all $i \in \mathcal{N}$, $T \in \mathbb{N}$, and $z^i \in \mathcal{X}^i$, we have*

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T \langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X^i_{t+\frac{1}{2}} - z^i\rangle\right] \leq \mathbb{E}\Bigg[&\frac{\|X_1^i - z^i\|^2}{2\eta^i_{T+1}} + \sum_{t=1}^T \frac{5}{8}\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2_{\gamma_t}\\ &+ \sum_{t=1}^{T-1} \frac{3\|\gamma_1\|_\infty L^2}{2}\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2\\ &+ \sum_{t=1}^T \left(3\|\gamma_t\|_\infty^3 NL^2 + 2\|\gamma_t\|_\infty^2\sqrt{N}L + \eta_t^i\right)\sigma_A^2\Bigg].\end{aligned}$$

Proof. Thanks to Lemma 8.5 and Assumption 7.1, we can bound

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^T \langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X^i_{t+\frac{1}{2}} - z^i\rangle\right]\\ &\leq \mathbb{E}\Bigg[\frac{\|X_1^i - z^i\|^2}{2\eta^i_{T+1}} + \frac{1}{2}\sum_{t=1}^T \eta_t^i\left((1+\sigma_M^2)\|V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2 + \sigma_A^2\right)\\ &\quad + \sum_{t=2}^T \gamma_t^i L^2\left(3(1+\sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2_{\gamma_t^2} + 3\|\gamma_t^2\|_1\sigma_A^2 + \frac{3}{2}\|\mathbf{X}_t - \mathbf{X}_{t-1}\|^2\right)\\ &\quad + \sum_{t=2}^T (\gamma_t^i)^2\sqrt{N}L(\sigma_M^2\|V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 + \sigma_A^2)\\ &\quad + \sum_{t=2}^T \frac{L}{\sqrt{N}}(\sigma_M^2\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|^2_{\gamma_t^2} + \|\gamma_t^2\|_1\sigma_A^2)\Bigg].\end{aligned}$$

In the following, we further bound the above using (*i*) $\eta_t^i \le \gamma_t^i/(4(1+\sigma_M^2))$, (*ii*) $\boldsymbol{\gamma}_{t+1} \le \boldsymbol{\gamma}_t$, (*iii*) $\alpha_t^i \|V^i(\mathbf{x})\|^2 \le \|\mathbf{V}(\mathbf{x})\|_{\boldsymbol{\alpha}}^2$ for any $\boldsymbol{\alpha} \in \mathbb{R}_+^N$ and $\mathbf{x} \in \mathbb{R}^d$, and (*iv*) $\|\boldsymbol{\alpha}\|_\infty = \max_{i\in\mathcal{N}} \alpha^i$ and in particular $\|\boldsymbol{\alpha}^2\|_1 \le N\|\boldsymbol{\alpha}\|_\infty^2$ for $\boldsymbol{\alpha} \in \mathbb{R}_+^N$.

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X^i_{t+\frac{1}{2}} - z^i\rangle\right] \\
&\le \mathbb{E}\left[\frac{\|X_1^i - z^i\|^2}{2\eta^i_{T+1}} + \sum_{t=2}^{T} 3\|\boldsymbol{\gamma}_t\|_\infty^2 L^2\left((1+\sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|_{\boldsymbol{\gamma}_t}^2 + \|\boldsymbol{\gamma}_t\|_\infty N\sigma_A^2\right)\right. \\
&\quad + \sum_{t=2}^{T}\|\boldsymbol{\gamma}_t\|_\infty L\sigma_M^2\left(\gamma_t^i\sqrt{N}\|V^i(\mathbf{X}_{t-\frac{1}{2}})\|^2 + \frac{1}{\sqrt{N}}\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|_{\boldsymbol{\gamma}_t}^2\right) \\
&\quad + \sum_{t=2}^{T} 2\|\boldsymbol{\gamma}_t\|_\infty^2\sqrt{N}L\sigma_A^2 \\
&\quad \left. + \sum_{t=2}^{T}\frac{3\|\boldsymbol{\gamma}_t\|_\infty L^2}{2}\|\mathbf{X}_t - \mathbf{X}_{t-1}\|^2 + \frac{1}{2}\sum_{t=1}^{T}\left(\frac{\gamma_t^i}{4}\|V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2 + \eta_t^i\sigma_A^2\right)\right] \\
&\le \mathbb{E}\left[\frac{\|X_1^i - z^i\|^2}{2\eta^i_{T+1}} + \sum_{t=1}^{T-1}\frac{3\|\boldsymbol{\gamma}_1\|_\infty L^2}{2}\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2\right. \\
&\quad + \sum_{t=1}^{T}\left(3\|\boldsymbol{\gamma}_t\|_\infty^2 L^2(1+\sigma_M^2) + 2\|\boldsymbol{\gamma}_t\|_\infty\sqrt{N}L\sigma_M^2 + \frac{1}{8}\right)\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_{\boldsymbol{\gamma}_t}^2 \\
&\quad \left. + \frac{1}{2}\sum_{t=1}^{T}\left(6\|\boldsymbol{\gamma}_t\|_\infty^3 NL^2\sigma_A^2 + 4\|\boldsymbol{\gamma}_t\|_\infty^2\sqrt{N}L\sigma_A^2 + \eta_t^i\sigma_A^2\right)\right].
\end{aligned}
$$

To conclude, we notice that under that our learning rate requirements (8.10) it holds that $3\|\boldsymbol{\gamma}_t\|_\infty^2 L^2(1+\sigma_M^2) + 2\|\boldsymbol{\gamma}_t\|_\infty\sqrt{N}L\sigma_M^2 \le 1/2$. □

Our main regret guarantees of (OptDA+) with predetermined learning rates then follows from the combination of Lemmas 8.10 and 8.12.

Proof. Let $z^i \in \mathcal{Z}^i$ and $\mathbf{x}_\star = \arg\min_{\mathbf{x}\in\mathfrak{X}_\star}\|\mathbf{X}_1 - \mathbf{x}\|_{1/\boldsymbol{\eta}}$. Similar to the proof of Theorem 8.9, we define $a_t = 3\|\boldsymbol{\gamma}_t\|_\infty^3 NL^2 + 2\|\boldsymbol{\gamma}_t\|_\infty^2\sqrt{N}L + \|\boldsymbol{\eta}_t\|_\infty$. Combining Lemma 8.10 and Lemma 8.12, we know that

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X^i_{t+\frac{1}{2}} - z^i\rangle\right] \\
&\le \mathbb{E}\left[\frac{R^2}{2\eta^i_{T+1}} + \sum_{t=1}^{T} a_t\sigma_A^2 + \frac{5}{4}\left(\|\mathbf{X}_1 - \mathbf{x}_\star\|_{1/\boldsymbol{\eta}_{T+1}}^2 + \|\mathbf{V}(\mathbf{X}_1)\|_{\boldsymbol{\gamma}_1}^2 + 2\sum_{t=1}^{T} a_t N\sigma_A^2\right)\right].
\end{aligned}
$$

The claims of the theorem follow immediately. □

8.2.4 *Convergence to Equilibrium under Multiplicative Noise*

Finally, as in Chapters 6 and 7, we also demonstrate trajectory convergence of the algorithm. We focus here on the multiplicative noise scenario and show convergence for players that use constant learning rates. Nonetheless, we fail to prove a convergence result for the general case where $\sigma_A > 0$, due to the challenges that we highlighted in Section 8.1.

Convergence to Nash equilibrium under multiplicative noise

Theorem 8.13. *Suppose that Assumptions 5.1–5.3, 7.1 and 7.2 hold with $\sigma_A = 0$ and all players run (OptDA+) with constant learning rates satisfying (8.10). Then, both $\mathbf{X}_t$ and $\mathbf{X}_{t+\frac{1}{2}}$ converge almost surely to a Nash equilibrium.*

Proof. Following the proof of Theorem 7.10, we define $\tilde{\mathbf{X}}_1 = \mathbf{X}_1$ and for all $i \in \mathcal{N}$, $t \geq 2$,

$$\tilde{X}_t^i = X_t^i + \eta^i \xi_{t-\frac{1}{2}}^i = -\eta^i \sum_{s=1}^{t-2} \hat{V}_{t+\frac{1}{2}}^i - \eta^i V^i(\mathbf{X}_{t-\frac{1}{2}}).$$

$\tilde{\mathbf{X}}_t$ serves a surrogate for $\mathbf{X}_t$ and is $\mathcal{F}_{t-1}$-measurable. Our first step is to show that

$$\textit{With probability } 1, \ \|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|_{1/\eta} \textit{ converges for all } \mathbf{x}_\star \in \mathfrak{X}_\star.$$

For this, we fix $\mathbf{x}_\star \in \mathfrak{X}_\star$ and apply the Robbins–Siegmund theorem (Lemma 7.8) to inequality (8.5) of Lemma 8.6 with

$$\begin{aligned}
&\mathcal{G}_t \leftarrow \mathcal{F}_{t-1}, \quad U_t \leftarrow \mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|_{1/\eta}^2], \quad \alpha_t \leftarrow 0,\\
&\zeta_t \leftarrow \mathbb{E}_{t-1}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_\gamma^2] + \|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|_\gamma^2,\\
&\chi_t \leftarrow \mathbb{E}_{t-1}[3\|\mathbf{V}(\mathbf{X}_t) - \mathbf{V}(\mathbf{X}_{t-1})\|_\gamma^2 + 4\sqrt{N}L\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|_{\gamma^2}^2\\
&\qquad\qquad + 3L^2(\|\boldsymbol{\gamma}\|_1\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|_{\gamma^2}^2 + \|\boldsymbol{\gamma}\|_1\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|_{\gamma^2}^2) + 2\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|_\eta^2].
\end{aligned}$$

For $t = 1$ we use (8.9); thus $\zeta_1 = 0$ and $\chi_1 = \|\hat{\mathbf{V}}_{3/2}\|_\eta^2$. To see that the Robbins–Siegmund theorem is effectively applicable, we use Assumptions 5.2 and 7.1 with $\sigma_A = 0$ to establish[3]

$$\begin{aligned}
\mathbb{E}[\chi_t] \leq \mathbb{E}[&3\|\boldsymbol{\gamma}\|_\infty L^2\|\mathbf{X}_t - \mathbf{X}_{t-1}\|^2\\
&+ (4\|\boldsymbol{\gamma}\|_\infty\sqrt{N}L\sigma_M^2 + 3\|\boldsymbol{\gamma}\|_\infty^2 NL^2(1+\sigma_M^2))\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|_\gamma^2]\\
&+ 3\|\boldsymbol{\gamma}\|_\infty^2 NL^2(1+\sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t-\frac{3}{2}})\|_\gamma^2 + 2(1+\sigma_M^2)\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_\eta^2].
\end{aligned}$$

With $2(1+\sigma_M^2)\boldsymbol{\eta} \leq \boldsymbol{\gamma}$, it follows immediately from Lemma 8.10 that $\sum_{t=1}^{+\infty}\mathbb{E}[\chi_t] < +\infty$. the Robbins–Siegmund theorem thus ensures the almost sure convergence of $\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2]$ to a finite random variable. By definition of $\tilde{X}_t^i$, we have

$$\begin{aligned}
\mathbb{E}_{t-1}[\|X_t^i - x_\star^i\|^2] &= \mathbb{E}_{t-1}[\|\tilde{X}_t^i - \eta^i\xi_{t-\frac{1}{2}}^i - x_\star^i\|^2]\\
&= \|\tilde{X}_t^i - x_\star^i\|^2 + (\eta^i)^2\,\mathbb{E}_{t-1}[\|\xi_{t-\frac{1}{2}}^i\|^2].
\end{aligned}$$

Subsequently,

$$\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|_{1/\eta}^2] = \|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|_{1/\eta}^2 + \mathbb{E}_{t-1}[\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|_\eta^2].$$

Therefore, by Assumption 7.1 with $\sigma_A = 0$ and Lemma 8.10 we get

$$\begin{aligned}
\sum_{t=2}^{+\infty}\mathbb{E}[\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|_{1/\eta}^2] - \|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|_{1/\eta}^2] &= \sum_{t=2}^{+\infty}\mathbb{E}[\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|_\eta^2]\\
&\leq \sum_{t=2}^{+\infty}\sigma_M^2\,\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|_\eta^2] < +\infty.
\end{aligned}$$

3 For $t = 1$ and $t = 2$, we remove the terms that involve either $\mathbf{X}_{1/2}$, $\mathbf{X}_0$, or $\mathbf{X}_{-1/2}$.

Following the proof of Theorem 7.10, we deduce with the help of Lemma B.4 and Corollary B.7 that the claimed argument is effectively true, i.e., with probability 1, $\|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|_{1/\eta}$ converges for all $\mathbf{x}_\star \in \mathfrak{X}_\star$.

Since $\|\mathbf{X}_t - \tilde{\mathbf{X}}_t\|^2 = \|\xi_{t-\frac{1}{2}}\|_{\eta^2}^2$ and $\|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_t\|^2 = \sum_{i=1}^N \|\gamma^i \hat{V}^i_{t-\frac{1}{2}} + \eta^i \xi^i_{t-\frac{1}{2}}\|^2$, applying the multiplicative noise assumption, Lemma 8.10, and Lemma B.4 we deduce that both $\|\mathbf{X}_t - \tilde{\mathbf{X}}_t\|$ and $\|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_t\|$ converge to 0 almost surely. Moreover, Lemma 8.10 along with Lemma B.4 also imply the almost sure convergence of $\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|$ to 0. In summary, we have shown that the event

$$\mathcal{E} := \left\{ \begin{array}{c} \|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|_{1/\eta} \text{ converges for all } \mathbf{x}_\star \in \mathfrak{X}_\star, \\ \lim_{t\to+\infty} \|\mathbf{X}_t - \tilde{\mathbf{X}}_t\| = 0, \quad \lim_{t\to+\infty} \|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_t\| = 0, \quad \lim_{t\to+\infty} \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\| = 0 \end{array} \right\}$$

happens almost surely. To conclude, we just need to show that $\mathbf{X}_t$ and $\mathbf{X}_{t+\frac{1}{2}}$ converge to a point in $\mathfrak{X}_\star$ whenever $\mathcal{E}$ happens. The convergence of $\|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|_{1/\eta}$ for a point $\mathbf{x}_\star$ in particular implies the boundedness of $(\tilde{\mathbf{X}}_t)_{t\in\mathbb{N}}$. Therefore, $(\tilde{\mathbf{X}}_t)_{t\in\mathbb{N}}$ has at least a cluster point, which we denote by $\mathbf{x}_\infty$. Provided that $\lim_{t\to+\infty}\|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_t\| = 0$, the point $\mathbf{x}_\infty$ is clearly also a cluster point of $(\mathbf{X}_{t+\frac{1}{2}})_{t\in\mathbb{N}}$. By $\lim_{t\to+\infty}\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\| = 0$ and the continuity of $\mathbf{V}$ we then have $\mathbf{V}(\mathbf{x}_\infty) = 0$, i.e., $\mathbf{x}_\infty \in \mathfrak{X}_\star$. This in turn implies that $\|\tilde{\mathbf{X}}_t - \mathbf{x}_\infty\|_{1/\eta}$ converges, so this limit can only be 0. In other words, $(\tilde{\mathbf{X}}_t)_{t\in\mathbb{N}}$ converges to $\mathbf{x}_\infty$; we conclude with $\lim_{t\to+\infty}\|\mathbf{X}_t - \tilde{\mathbf{X}}_t\| = 0$ and $\lim_{t\to+\infty}\|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_t\| = 0$. □

8.3 OPTDA+ WITH ADAPTIVE LEARNING RATES

Let's now turn to the ultimate goal of this chapter: the development of an algorithm that is both adaptive and resilient to noise.

8.3.1 *Adaptivity in the Face of Noise*

The learning rates that we have examined so far in Chapters 7 and 8 require tuning based on various model parameters. Nonetheless, even though a player might be aware of their own loss function, there is little hope that the noise-related parameters are also known by the player. On the other hand, the adaptive method proposed in Chapter 6 can achieve neither last-iterate convergence nor constant regret bound when the feedback is corrupted by noise. Our goal in this section, therefore, is to combine the strengths of these methods and to circumvent their respective limitations. We strive to deliver the best of both worlds: an algorithm that is adaptive and simultaneously capable of handling noisy feedback.

Desiderata for our algorithm

In pursuit of this objective, we work toward the design of adaptive methods that boast the following desirable properties:

- The method should be implementable by every individual player using only local information and without any prior knowledge of the setting's parameters (for the noise profile and the game alike).
- The method should guarantee sublinear individual regret against any bounded feedback sequence.
- When employed by all players, the method should guarantee $O(\sqrt{T})$ regret under additive noise and $O(1)$ regret under multiplicative noise.

- When employed by all players, the method should converge to a Nash equilibrium under multiplicative noise.

Adaptive learning rates

In order to achieve the above, inspired by the learning rate requirements of Theorems 8.8, 8.9 and 8.11, we consider the following adaptive learning rate schedule that is defined with respect to some $r \in (0, 1/4]$.

$$\gamma_t^i = \frac{1}{\left(1 + \sum_{s=1}^{t-2} \|g_s^i\|^2\right)^{\frac{1}{2}-r}}, \quad \eta_t^i = \frac{1}{\sqrt{1 + \sum_{s=1}^{t-2} \left(\|g_s^i\|^2 + \|X_s^i - X_{s+1}^i\|^2\right)}}. \tag{Adapt+}$$

As in (AdaGrad-norm), the sum of the squared norm of the feedback appears in the denominator. This helps controlling the various positive terms appearing in Lemma 8.4, such as $\|\xi_{t-\frac{1}{2}}\|^2_{(\eta_t+\gamma_t)^2}$ and $\eta_t^i \|\hat{V}^i_{t+\frac{1}{2}}\|^2$. Nonetheless, this sum is not taken to the same exponent in the definition of the two learning rates. This scale separation ensures that the contribution of the term $-\gamma_t^i \|V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2$ appearing in (8.1c) remains negative, and it is the key for deriving constant regret under multiplicative noise. As a technical detail, the path variation $\|X_s^i - X_{s+1}^i\|^2$ is involved in the definition of η_t^i for controlling the difference between the gradient variation and the path variation. Finally, we do not include the previous received feedback g_{t-1}^i in the definition of γ_t^i and η_t^i. This makes these learning rates $\mathcal{F}_{t-1}$-measurable so that Assumption 8.1 is verified.

Compared to (Adapt), we use feedback norms and not feedback variations to define the learning rates here. This is feasible because we now work with *unconstrained* action sets, and the negative terms in Lemma 8.4 are effectively the squared norms of the pseudo-gradients. While we believe a learning rate policy that is much closer to the one described in (Adapt) can also fulfill the aforementioned criteria, our analysis deals specifically with the (Adapt+) learning rate rule.

Ultimately, our goal here is to recover automatically the learning rate schedules of Theorems 8.9 and 8.11. This implies that γ_t^i and η_t^i should at least be in the order of $\Omega(1/t^{\frac{1}{2}-r})$ and $\Omega(1/\sqrt{t})$, suggesting the following boundedness assumptions on the feedback.

Boundedness of operator and noise

Assumption 8.2. There exist $G, \bar{\sigma} \in \mathbb{R}_+$ such that (*i*) $\|V^i(\mathbf{x})\| \leq G$ for all $i \in \mathcal{N}$, $\mathbf{x} \in \mathbb{R}^d$; and (*ii*) $\|\xi_t^i\| \leq \bar{\sigma}$ for all $i \in \mathcal{N}$, $t \in \mathbb{N}$ with probability 1.

These assumptions are commonplace in the adaptive method literature, as evidenced by a series of works in the context of learning in games and VIs [6, 15, 72, 149]. It also aligns with Assumption 2.2 that we have assumed in Part I for adversarial online learning. We note, however, that the boundedness of the gradient was not assumed in Chapter 6, thanks to a specific treatment on the gradient variation in the proof of Theorem 6.5. Unfortunately, that same technique is not applicable here due to the presence of noise.

In a similar spirit, we also need to strengthen our assumption on the initialization of the algorithm.

Assumption on initialization: almost sure boundedness

Assumption 8.3. There exist $\rho \in \mathbb{R}_+$ and $\mathbf{x}_\star \in \mathfrak{X}_\star$ such that $\|\mathbf{X}_1 - \mathbf{x}_\star\|_\infty \leq \rho$ with probability 1.

Provided that both Assumptions 8.2 and 8.3 assume the inequalities to hold *almost surely*, all the inequalities that we are going to see in this section only hold almost surely. To avoid repetition, we will not mention this explicitly in the following.

8.3.2 *Preliminary Lemmas*

In this subsection, we present several basic lemmas concerning our adaptive learning rates. For ease of notation, we introduce the following quantities

$$\Lambda_t^i = \sum_{s=1}^{t} \|\hat{V}_{s+\frac{1}{2}}^i\|^2, \quad \Gamma_t^i = \sum_{s=1}^{t} \|X_s^i - X_{s+1}^i\|^2.$$

The learning rate rule (Adapt+) can then be stated as

$$\gamma_t^i = \frac{1}{(1+\Lambda_{t-2}^i)^{\frac{1}{2}-r}}, \quad \eta_t^i = \frac{1}{\sqrt{1+\Lambda_{t-2}^i+\Gamma_{t-2}^i}}$$

To begin, we state the apparent fact that Λ_t^i grows at most linearly under Assumption 8.2.

Lemma 8.14. *Suppose that Assumption 8.2 holds. Then, for all $i \in \mathcal{N}$ and $T \in \mathbb{N}$, we have*

$$\Lambda_T^i \leq 2(G^2+\bar{\sigma}^2)T.$$

Proof. Using Assumption 8.2, we deduce that

$$\|\hat{V}_{t+\frac{1}{2}}^i\|^2 \leq 2\|V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2 + 2\|\xi_{t+\frac{1}{2}}^i\|^2 \leq 2G^2 + 2\bar{\sigma}^2,$$

The claimed inequality is then immediate from the definition of Λ_T^i. □

Along with a generalization of Lemma 2.6 that we present in Appendix B (Lemma B.8), we can then derive the following bound on the weighted sum of the squared norms of the feedback.

Lemma 8.15. *Suppose that Assumption 8.2 holds. Then, for all $s \in \mathbb{N}_0$, $q \in [0,1)$, $i \in \mathcal{N}$, and $T \in \mathbb{N}$, we have*

$$\sum_{t=1}^{T} \frac{\|\hat{V}_{t+\frac{1}{2}}^i\|^2}{(1+\Lambda_{t-s}^i)^q} \leq \frac{(\Lambda_T^i)^{1-q}}{1-q} + 2s(G^2+\bar{\sigma}^2).$$

Proof. Since $1/(1+\Lambda_t^i)^q \leq 1/(1+\Lambda_{t-s}^i)^q$ and $\|\hat{V}_{t+\frac{1}{2}}^i\|^2 \leq 2G^2+2\bar{\sigma}^2$, we have

$$\left(\frac{1}{(1+\Lambda_{t-s}^i)^q} - \frac{1}{(1+\Lambda_t^i)^q}\right)\|\hat{V}_{t+\frac{1}{2}}^i\|^2 \leq \left(\frac{1}{(1+\Lambda_{t-s}^i)^q} - \frac{1}{(1+\Lambda_t^i)^q}\right)2(G^2+\bar{\sigma}^2).$$

Subsequently, it follows from Lemma B.8 that

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\|\hat{V}_{t+\frac{1}{2}}^i\|^2}{(1+\Lambda_{t-s}^i)^q} &= \sum_{t=1}^{T}\left(\frac{\|\hat{V}_{t+\frac{1}{2}}^i\|^2}{(1+\Lambda_t^i)^q} + \left(\frac{1}{(1+\Lambda_{t-s}^i)^q} - \frac{1}{(1+\Lambda_t^i)^q}\right)\|\hat{V}_{t+\frac{1}{2}}^i\|^2\right) \\
&\leq \sum_{t=1}^{T} \frac{\|\hat{V}_{t+\frac{1}{2}}^i\|^2}{(1+\Lambda_t^i)^q} + \sum_{t=1}^{T}\left(\frac{1}{(1+\Lambda_{t-s}^i)^q} - \frac{1}{(\Lambda_t^i)^q}\right)2(G^2+\bar{\sigma}^2) \\
&\leq \frac{(\Lambda_T^i)^{1-q}}{1-q} + \sum_{t=-s+1}^{0} \frac{2(G^2+\bar{\sigma}^2)}{(1+\Lambda_t^i)^q}
\end{aligned}$$

$$= \frac{(\Lambda_T^i)^{1-q}}{1-q} + 2s(G^2 + \bar{\sigma}^2). \qquad \square$$

We also state a variant of the above result that applies to the joint feedback vectors.

Lemma 8.16. *Suppose that Assumption 8.2 holds and $(\boldsymbol{\alpha}_t)_{t\in\mathbb{N}}$ is a sequence of non-negative N-dimensional vectors such that $\alpha_t^i \leq 1/(1+\Lambda_{t-s}^i)^q$. Then, for all Then, for all $s \in \mathbb{N}_0$, $q \in [0,1)$, and $T \in \mathbb{N}$, we have*

$$\sum_{t=1}^{T} \|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|_{\boldsymbol{\alpha}_t}^2 \leq \frac{(\Lambda_T^i)^{1-q}}{1-q} + 2Ns(G^2 + \bar{\sigma}^2).$$

Proof. This is immediate from Lemma 8.15. $\square$

Both Lemma 8.15 and Lemma 8.16 are essential for our analysis as they allow us to control the sum of the positive terms that show up in (8.1d), (8.1e), and (8.1f). The new upper bound then just contains powers of Λ_t^i which we can further control in several ways.

Finally, we present a lemma for bounding the inverse of η_t^i that separates the squared norms of the feedback from the path variations.

Lemma 8.17. *Consider the learning rates defined as in (Adapt+). For any $i \in \mathcal{N}$, $T \in \mathbb{N}$, and $a, b \in \mathbb{R}_+$, we have*

$$\frac{a}{\eta_{T+1}^i} - b\sum_{t=1}^{T} \frac{\|X_t^i - X_{t+1}^i\|^2}{\eta_t^i} \leq a\sqrt{1+\Lambda_{T-1}^i} + \frac{a^2}{4b}.$$

Proof. On one hand, we have

$$\frac{a}{\eta_{T+1}^i} = a\sqrt{1+\Lambda_{T-1}^i+\Gamma_{T-1}^i} \leq a\sqrt{1+\Lambda_{T-1}^i} + a\sqrt{\Gamma_{T-1}^i}.$$

On the other hand, with $\eta_t^i \leq 1$, it holds

$$b\sum_{t=1}^{T} \frac{\|X_t^i - X_{t+1}^i\|^2}{\eta_t^i} \geq b\sum_{t=1}^{T} \|X_t^i - X_{t+1}^i\|^2 \geq b\Gamma_{T-1}^i.$$

Let us define the function $f\colon y \in \mathbb{R} \mapsto -by^2 + ay$. Then

$$a\sqrt{\Gamma_{T-1}^i} - b\Gamma_{T-1}^i \leq \max_{y\in\mathbb{R}} f(y) = \frac{a^2}{4b}.$$

Combining the above inequalities gives the desired result. $\square$

8.3.3 *No-Regret Against Adversarial Opponents*

As usual, we derive regret bounds for the algorithm in question when it is used against adversarial opponents.

Regret bound against bounded adversarial feedback

Theorem 8.18. *Suppose that Assumptions 5.1, 7.1 and 8.2 hold and player i runs*

(OptDA+) *with learning rates* (Adapt+). *Then, for any bounded set* $\mathcal{Z}^i$ *with* $R \geq \sup_{z^i \in \mathcal{Z}^i} \|X_1^i - z^i\|$, *it holds*

$$\overline{\mathrm{Reg}}_T^i(\mathcal{Z}^i) = \mathcal{O}\left(((G^2 + \bar{\sigma}^2)T)^{\frac{1}{2}+r} + R^2(G + \bar{\sigma})\sqrt{T} + R^4 + G^2 + \bar{\sigma}^2\right).$$

Proof. To begin, we notice that inequality (8.11) that we established in the proof of Theorem 8.8 still holds here for any $z^i \in \mathcal{Z}^i$. Furthermore, applying Lemma 8.17 with $a \leftarrow R^2/2$, $b \leftarrow 1/2$ leads to

$$\frac{R^2}{2\eta_{T+1}^i} - \sum_{t=1}^{T} \frac{\|X_t^i - X_{t+1}^i\|^2}{2\eta_t^i} \leq \frac{R^2\sqrt{1 + \Lambda_{T-1}^i}}{2} + \frac{R^4}{8}.$$

On the other hand, invoking Lemma 8.15 with $q \leftarrow 1/2 - r$ and $q \leftarrow 1/2$ guarantees that

$$\sum_{t=1}^{T} (\gamma_t^i + \eta_t^i)\|\hat{V}_{t+\frac{1}{2}}^i\|^2 \leq \frac{2(\Lambda_T^i)^{1/2+r}}{1+2r} + 2\sqrt{\Lambda_T^i} + 8(G^2 + \bar{\sigma}^2).$$

Putting the above inequalities together, we obtain

$$\max_{z^i \in \mathcal{Z}^i} \mathbb{E}\left[\sum_{t=1}^{T} \langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}}^i - z^i\rangle\right] \leq \mathbb{E}\left[\frac{R^2\sqrt{1 + \Lambda_{T-1}^i}}{2} + \frac{2(\Lambda_T^i)^{1/2+r}}{1+2r} + 2\sqrt{\Lambda_T^i}\right] + \frac{R^4}{8} + 8(G^2 + \bar{\sigma}^2).$$

We conclude with the help of Lemma 8.14 and the convexity of ℓ^i. □

Theorem 8.18 provides exactly the same rate as Theorem 8.8, illustrating in this way the benefit of taking a smaller r for achieving smaller regret against adversarial opponents. Nonetheless, similar to before, we will also see below that taking smaller r is less favorable in the self-play scenario. In particular, we require $r > 0$ in order to obtain convergence and constant regret under multiplicative noise, and this prevents us from obtaining the optimal $\mathcal{O}(\sqrt{T})$ regret in fully adversarial environments.

8.3.4 *Fast Convergence of Pseudo-Gradient in Self-Play*

We next derive bounds on the norm of the pseudo-gradients to characterize the convergence speed of the method.

Bound on norm of pseudo-gradient

Theorem 8.19. *Suppose that Assumptions 5.2, 5.3, 7.1, 8.2 and 8.3 hold and all players run* (OptDA+) *with adaptive learning rates* (Adapt+). *Then,*

(a) It holds that $\sum_{t=1}^{T} \mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2] = \mathcal{O}\left(T^{1-r}\right)$.

(b) If the noise is multiplicative, it holds almost surely that $\sum_{t=1}^{+\infty} \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 < +\infty$.

Again, our bounds match those obtained for the non-adaptive version of the algorithm (cf. Theorem 8.9). Of course, this relies on stronger assumptions, and, more importantly, we can no longer show that these pseudo-gradient norms are summable in L2 when the noise is multiplicative. Instead, we can only show their squares are summable almost surely. A consequence of this difference is

that we are not able to provide a constant bound on the *expected regret* here (see Appendix B). However, we are still able to prove constant *pseudo-regret* bound as we will show in Section 8.3.5

Let us now dive into the proof of Theorem 8.19. For this, we present a series of important lemmas that will be used in next two subsections as well. We start with a lemma that controls the difference between the gradient variation and the path variation. As argued earlier, this is the reason that we include $\|X_s^i - X_{s+1}^i\|^2$ in the definition of η_t^i.

Bound on gradient variation

Lemma 8.20. *Suppose that Assumptions 5.2 and 8.2 hold and the learning rates be defined as in* (Adapt+)*, then for all $T \in \mathbb{N}$, we have*

$$\sum_{t=1}^{T}\left(3\|\mathbf{V}(\mathbf{X}_t)-\mathbf{V}(\mathbf{X}_{t+1})\|_{\gamma_t}^2 - \|\mathbf{X}_t-\mathbf{X}_{t+1}\|_{1/(4\eta_t)}^2\right) \le 432N^3L^6 + 24N^2G^2.$$

Proof. For all $i \in \mathcal{N}$, let us define

$$\bar{t}^i := \max\left\{t \in \{0,...,T\} : \eta_t^i \ge \frac{1}{12NL^2}\right\},$$

where we set $\eta_0^i = 1/(12NL^2)$ to ensure that $\bar{t}^i$ is always well-defined. By the definition of η_t^i, the inequality $\eta_{\bar{t}^i}^i \ge 1/(12NL^2)$ implies $\Gamma_{\bar{t}^i-2}^i \le 144N^2L^2$. We next define the sets

$$\mathcal{T} := \bigcup_{i\in\mathcal{N}}\{\bar{t}^i-1, \bar{t}^i\} \cap \{1,...,T\}$$

Clearly, $\mathrm{card}(\mathcal{T}) \le 2N$. With $\gamma_t \le 1$, Assumptions 5.2 and 8.2, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T} 3\|\mathbf{V}(\mathbf{X}_t)-\mathbf{V}(\mathbf{X}_{t+1})\|_{\gamma_t}^2 \\
&\quad\le \sum_{t=1}^{T} 3\|\mathbf{V}(\mathbf{X}_t)-\mathbf{V}(\mathbf{X}_{t+1})\|^2 \\
&\quad= \sum_{t\in[T]\setminus\mathcal{T}} 3\|\mathbf{V}(\mathbf{X}_t)-\mathbf{V}(\mathbf{X}_{t+1})\|^2 + \sum_{t\in\mathcal{T}} 3\|\mathbf{V}(\mathbf{X}_t)-\mathbf{V}(\mathbf{X}_{t+1})\|^2 \\
&\quad\le \sum_{t\in[T]\setminus\mathcal{T}} 3NL^2\|\mathbf{X}_t-\mathbf{X}_{t+1}\|^2 + \sum_{t\in\mathcal{T}} 6\left(\|\mathbf{V}(\mathbf{X}_t)\|^2 + \|\mathbf{V}(\mathbf{X}_{t+1})\|^2\right) \\
&\quad\le \sum_{i=1}^{N}\sum_{t\in[T]\setminus\mathcal{T}} 3NL^2\|X_t^i - X_{t+1}^i\|^2 + \sum_{t\in\mathcal{T}} 12NG^2 \\
&\quad\le \sum_{i=1}^{N} 3NL^2\Bigg(\underbrace{\sum_{t=1}^{\bar{t}^i-2}\|X_t^i-X_{t+1}^i\|^2}_{\Gamma_{\bar{t}^i-2}^i \le 144N^2L^2} + \sum_{t=\bar{t}^i+1}^{T}\|X_t^i-X_{t+1}^i\|^2\Bigg) + 24N^2G^2
\end{aligned} \tag{8.15}$$

On the other hand, by the choice of $\bar{t}^i$ we know that $1/\eta_t^i \ge 12NL^2$ for all $t \ge \bar{t}^i + 1$; hence

$$\sum_{t=1}^{T}\|\mathbf{X}_t-\mathbf{X}_{t+1}\|_{1/(4\eta_t)}^2 = \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\|X_t^i-X_{t+1}^i\|^2}{4\eta_t^i}$$

$$\geq \sum_{i=1}^{N} \sum_{t=\bar{t}^i+1}^{T} \frac{\|X_t^i - X_{t+1}^i\|^2}{4\eta_t^i}$$
$$\geq \sum_{i=1}^{N} \sum_{t=\bar{t}^i+1}^{T} 3NL^2 \|X_t^i - X_{t+1}^i\|^2. \tag{8.16}$$

Combining (8.15) and (8.16) gives the desired result. □

Equipped with Lemma 8.20 and the lemmas introduced in Section 8.3.2, we are now in position to establish an upper bound on the summation of two quantities: a weighted sum of squared pseudo-gradient norms, and the algorithm's second-order path length. This upper bound is crucial for our analysis.

Auxiliary bound on norm of pseudo-gradient and second-order path length

Lemma 8.21. *Suppose that Assumptions 5.2, 5.3, 7.1, 8.2 and 8.3 hold and all players run (OptDA+) with adaptive learning rates (Adapt+). Then, for all $T \in \mathbb{N}$ we have*

$$\sum_{t=1}^{T} \mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_{\gamma_t}^2] + \frac{1}{8} \sum_{t=1}^{T} \mathbb{E}[\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2] \leq c_1 \sum_{i=1}^{N} \mathbb{E}\left[\sqrt{\Lambda_T^i}\right] + c_2,$$

where

$$c_1 = 12NL^2 + 8\sqrt{N}L + \rho^2 + 4,$$
$$c_2 = 432N^3L^6 + 24N^2G^2 + (12NL^2 + 8\sqrt{N}L + 8)(NG^2 + N\bar{\sigma}^2) + N\rho^2 + 2N\rho^4.$$

Proof. As in the proof of Lemma 8.10, we proceed to bound in expectation the sum of the following quantities

$$A_t = 3\|\mathbf{V}(\mathbf{X}_t) - \mathbf{V}(\mathbf{X}_{t+1})\|_{\gamma_t}^2 - \|\mathbf{X}_t - \mathbf{X}_{t+1}\|_{1/(2\eta_t)}^2,$$
$$B_t = 6\|\boldsymbol{\gamma}_t\|_1 L^2 \|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|_{\gamma_t^2}^2 + 4\sqrt{N}L\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|_{\gamma_t^2}^2, \quad C_t = 2\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|_{\eta_t}^2.$$

Thanks to Lemma 8.20, we know that the sum of A_t can be bounded directly without taking expectation by

$$\sum_{t=1}^{T} A_t = \sum_{t=1}^{T} \left(3\|\mathbf{V}(\mathbf{X}_t) - \mathbf{V}(\mathbf{X}_{t+1})\|_{\gamma_t}^2 - \|\mathbf{X}_t - \mathbf{X}_{t+1}\|_{1/(4\eta_t)}^2 - \|\mathbf{X}_t - \mathbf{X}_{t+1}\|_{1/(4\eta_t)}^2\right)$$
$$\leq 432N^3L^6 + 24N^2G^2 - \sum_{t=1}^{T} \|\mathbf{X}_t - \mathbf{X}_{t+1}\|_{1/(4\eta_t)}^2. \tag{8.17}$$

To obtain the above inequality we have also used $\boldsymbol{\eta}_t \leq 1$. To bound $\mathbb{E}[B_t]$, we use $\mathbb{E}[\|\boldsymbol{\xi}_{t-\frac{1}{2}}\|_{\gamma_t^2}^2] \leq \mathbb{E}[\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|_{\gamma_t^2}^2]$, $\|\boldsymbol{\gamma}_t\|_1 \leq N$, and Lemma 8.16 (as $(\gamma_{t+1}^i)^2 \leq 1/\sqrt{1+\Lambda_{t-1}^i}$) to obtain

$$\sum_{t=2}^{T} \mathbb{E}[B_t] \leq \mathbb{E}\left[\sum_{t=2}^{T} (6NL^2 + 4\sqrt{N}L)\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|_{\gamma_t^2}^2\right]$$
$$= \mathbb{E}\left[\sum_{t=1}^{T-1} (6NL^2 + 4\sqrt{N}L)\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|_{(\gamma_{t+1})^2}^2\right]$$

$$\leq (6NL^2 + 4\sqrt{N}L)\left(2N(G^2+\bar{\sigma}^2) + \sum_{i=1}^{N} 2\,\mathbb{E}\left[\sqrt{\Lambda_{T-1}^i}\right]\right). \tag{8.18}$$

Similarly, the sum of C_t can be bounded in expectation by

$$\sum_{t=1}^{T}\mathbb{E}[C_t] \leq 8N(G^2+\bar{\sigma}^2) + \sum_{i=1}^{N} 4\,\mathbb{E}\left[\sqrt{\Lambda_T^i}\right]. \tag{8.19}$$

Let us choose $\mathbf{x}_\star$ as the one given by Assumption 8.3 so that $\|\mathbf{X}_1 - \mathbf{x}_\star\|_\infty \leq \rho$. Plugging (8.17), (8.18), and (8.19) into (8.6) of Lemma 8.7, we get readily

$$\begin{aligned}
&\sum_{t=2}^{T}\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_{\gamma_t}^2 + \|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|_{\gamma_t}^2] + \sum_{t=1}^{T}\mathbb{E}[\|\mathbf{X}_t - \mathbf{X}_{t+1}\|_{1/(8\eta_t)}^2] \\
&\leq \mathbb{E}[\|\mathbf{X}_1 - \mathbf{x}_\star\|_{1/\eta_{T+1}}^2] - \sum_{t=1}^{T}\|\mathbf{X}_t - \mathbf{X}_{t+1}\|_{1/(8\eta_t)}^2 \\
&\quad + (12NL^2 + 8\sqrt{N}L + 4)\sum_{i=1}^{N}\mathbb{E}\left[\sqrt{\Lambda_T^i}\right] \\
&\quad + 432N^3L^6 + 24N^2G^2 + (12NL^2 + 8\sqrt{N}L + 8)(NG^2 + N\bar{\sigma}^2)
\end{aligned} \tag{8.20}$$

Using Lemma 8.17, we can then further bound the RHS of (8.20) with

$$\begin{aligned}
&\|\mathbf{X}_1 - \mathbf{x}_\star\|_{1/\eta_{T+1}}^2 - \sum_{t=1}^{T}\|\mathbf{X}_t - \mathbf{X}_{t+1}\|_{1/(8\eta_t)}^2 \\
&= \sum_{i=1}^{N}\left(\frac{\|X_1^i - x_\star^i\|^2}{\eta_{T+1}^i} - \sum_{t=1}^{T}\frac{\|X_t^i - X_{t+1}^i\|^2}{8\eta_t^i}\right) \\
&\leq \sum_{i=1}^{N}\left(\|X_1^i - x_\star^i\|^2\sqrt{1+\Lambda_{T-1}^i} + 2\|X_1^i - x_\star^i\|^4\right) \\
&\leq N\rho^2 + 2N\rho^4 + \sum_{i=1}^{N}\rho^2\sqrt{\Lambda_{T-1}^i}.
\end{aligned} \tag{8.21}$$

Finally, using $\eta_t \leq 1$ and $\gamma_2 = \gamma_1$, the left-hand side (LHS) of (8.20) can be bounded from below by

$$\begin{aligned}
&\sum_{t=2}^{T}\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_{\gamma_t}^2 + \|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|_{\gamma_t}^2] + \sum_{t=1}^{T}\mathbb{E}[\|\mathbf{X}_t - \mathbf{X}_{t+1}\|_{1/(8\eta_t)}^2] \\
&\geq \sum_{t=1}^{T}\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_{\gamma_t}^2] + \frac{1}{8}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2].
\end{aligned} \tag{8.22}$$

Combining (8.20), (8.21), and (8.22) gives the desired result. □

Lemma 8.21 will be used multiple times in the remaining of the section. For ease of notation, we will continue to denote the two constants by c_1 and c_2 without redefining them whenever this is the case.

Bound for the General Noise Model

From Lemma 8.21 and Lemma 8.25 we can readily derive our main results for the general noise model.

Proof of Theorem 8.19(a). With Lemma 8.14, for $t \in \{1, ..., T\}$, we can lower bound the learning rate γ_t^i by

$$\gamma_t^i = \frac{1}{(1+\Lambda_{t-2}^i)^{\frac{1}{2}-r}} \geq \frac{1}{(1+2\max(t-2,0)(G^2+\bar{\sigma}^2))^{\frac{1}{2}-r}} \geq \frac{1}{(1+2T(G^2+\bar{\sigma}^2))^{\frac{1}{2}-r}}.$$

Lemma 8.21 thus guarantees

$$\frac{\sum_{t=1}^T \mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2]}{(1+2T(G^2+\bar{\sigma}^2))^{\frac{1}{2}-r}} \leq c_1 \sum_{i=1}^N \mathbb{E}\left[\sqrt{\Lambda_t^T}\right] + c_2.$$

Using again Lemma 8.14 we know that $\mathbb{E}\left[\sqrt{\Lambda_t^T}\right] = O(\sqrt{T})$ and thus we have effectively $\sum_{t=1}^T \mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2] = O\left(T^{1-r}\right)$. □

Bound for the Multiplicative Noise Model

The case of multiplicative noise is more delicate. From Lemma 8.21 it is clear that we need to control Λ_t^i. We achieve this via the following lemma.

Bound on Λ_t^i

Lemma 8.22. *Suppose that Assumptions 5.2, 5.3, 7.1, 8.2 and 8.3 hold and all players run (OptDA+) with adaptive learning rates (Adapt+). Assume additionally Assumption 7.1 with $\sigma_A = 0$. Then, with constants c_1 and c_2 defined in Lemma 8.21, we have for any $T \in \mathbb{N}$ that*

$$\sum_{i=1}^N \mathbb{E}\left[(1+\Lambda_T^i)^{\frac{1}{2}+r}\right] \leq N\left((1+\sigma_M^2)c_1 + 1 + \frac{(1+\sigma_M^2)c_2}{N}\right)^{1+\frac{1}{2r}}, \tag{8.23}$$

$$\sum_{i=1}^N \mathbb{E}\left[\sqrt{1+\Lambda_T^i}\right] \leq N\left((1+\sigma_M^2)c_1 + 1 + \frac{(1+\sigma_M^2)c_2}{N}\right)^{\frac{1}{2r}}, \tag{8.24}$$

$$\sum_{i=1}^N \mathbb{E}[\Gamma_T^i] \leq 8Nc_1\left((1+\sigma_M^2)c_1 + 1 + \frac{(1+\sigma_M^2)c_2}{N}\right)^{\frac{1}{2r}} + 8c_2. \tag{8.25}$$

Proof. From Lemma 8.21 we know that

$$\sum_{t=1}^T \mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_{\gamma_t}^2] \leq c_1 \sum_{i=1}^N \mathbb{E}\left[\sqrt{\Lambda_T^i}\right] + c_2,$$

Since $\boldsymbol{\gamma}_t$ is $\mathcal{F}_t$-measurable, using the $\sigma_A = 0$ and the law of total expectation we get

$$\begin{aligned}\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_{\gamma_t}^2] &= \sum_{i=1}^N \mathbb{E}[\gamma_t^i \mathbb{E}_t[\|V^i(\mathbf{X}_{t+\frac{1}{2}})]\|^2] \\ &\geq \sum_{i=1}^N \mathbb{E}\left[\gamma_t^i \mathbb{E}_t\left[\frac{\|\hat{V}_{t+\frac{1}{2}}^i\|^2}{1+\sigma_M^2}\right]\right] = \frac{\mathbb{E}[\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|_{\gamma_t}^2]}{1+\sigma_M^2}.\end{aligned}$$

The learning rates γ_t being non-increasing, we can then bound from below the sum of $\mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2_{\gamma_t}]$ by

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2_{\gamma_t}] &\geq \frac{1}{1+\sigma_M^2} \sum_{t=1}^{T} \mathbb{E}[\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|^2_{\gamma_t}] \\
&\geq \frac{1}{1+\sigma_M^2} \sum_{t=1}^{T} \mathbb{E}[\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|^2_{\gamma_{T+2}}] \\
&= \frac{1}{1+\sigma_M^2} \sum_{i=1}^{N} \mathbb{E}\left[\frac{\sum_{t=1}^{T} \|\hat{V}^i_{t+\frac{1}{2}}\|^2}{(1+\Lambda_T^i)^{\frac{1}{2}-r}}\right] \\
&= \frac{1}{1+\sigma_M^2} \sum_{i=1}^{N} \mathbb{E}\left[\frac{\Lambda_T^i + 1 - 1}{(1+\Lambda_T^i)^{\frac{1}{2}-r}}\right] \\
&\geq -\frac{N}{1+\sigma_M^2} + \frac{1}{1+\sigma_M^2} \sum_{i=1}^{N} \mathbb{E}\left[(1+\Lambda_T^i)^{\frac{1}{2}+r}\right].
\end{aligned}$$

As a consequence, we have shown that

$$\sum_{i=1}^{N} \mathbb{E}\left[(1+\Lambda_T^i)^{\frac{1}{2}+r}\right] \leq (1+\sigma_M^2)c_1 \sum_{i=1}^{N} \mathbb{E}\left[\sqrt{\Lambda_T^i}\right] + (1+\sigma_M^2)c_2 + N,$$

Subsequently,

$$\sum_{i=1}^{N} \mathbb{E}\left[(1+\Lambda_T^i)^{\frac{1}{2}+r}\right] \leq \left((1+\sigma_M^2)c_1 + 1 + \frac{(1+\sigma_M^2)c_2}{N}\right) \sum_{i=1}^{N} \mathbb{E}\left[\sqrt{1+\Lambda_T^i}\right].$$

We deduce (8.23) and (8.24) with the help of Lemma B.9 taking $p \leftarrow 1/2 + r$, $r \leftarrow 1/2$, $c \leftarrow (1+\sigma_M^2)c_1 + 1 + (1+\sigma_M^2)c_2/N$, and $a^i \leftarrow 1 + \Lambda_T^i$. Plugging (8.24) into Lemma 8.21 gives (8.25). □

As a direct consequence of Lemma 8.22, we show that the learning rates almost surely converge to positive constants. Therefore, akin to what we have seen in Chapter 6, the algorithm is basically capable of figure out itself the right constant learning rates to use and stick to it.

Convergence of adaptive learning rates to positive constants

Proposition 8.23. *Suppose that Assumptions 5.2, 5.3, 7.1 and 8.2 hold with $\sigma_A = 0$ and all players run (OptDA+) with adaptive learning rates (Adapt+). Then,*

(a) With probability 1, for all $i \in \mathcal{N}$, $(\Lambda_t^i)_{t\in\mathbb{N}}$ and $(\Gamma_t^i)_{t\in\mathbb{N}}$ converge to finite constant.

(b) With probability 1, for all $i \in \mathcal{N}$, the learning rates $(\gamma_t^i)_{t\in\mathbb{N}}$ and $(\eta_t^i)_{t\in\mathbb{N}}$ converge to positive constants.

Proof. We notice that (b) is a direct consequence of (a) so we will only show (a) below. For this, we make use of Lemma 8.22 and Lemma B.4. In fact, $(\sqrt{\Lambda_t^i})_{t\in\mathbb{N}}$ is clearly non-decreasing and by Lemma 8.22, $\sup_{t\in\mathbb{N}} \mathbb{E}[\sqrt{\Lambda_t^i}] < +\infty$. Therefore, Lemma B.4 ensures the almost sure convergence of $(\sqrt{\Lambda_t^i})_{t\in\mathbb{N}}$ to a finite random variable, which in turn implies that $(\Lambda_t^i)_{t\in\mathbb{N}}$ converges to a finite constant almost surely. Similarly, $(\Gamma_t^i)_{t\in\mathbb{N}}$ is non-decreasing and $\sup_{t\in\mathbb{N}} \mathbb{E}[\Gamma_t^i] < +\infty$ by Lemma 8.22. We thus deduce by Lemma B.4 that $(\Gamma_t^i)_{t\in\mathbb{N}}$ converges to finite constant almost surely. □

Intuitively, Proposition 8.23 also suggests that the improvements that we demonstrated in Section 8.2 for the multiplicative setup, which were achieved with constant learning rates, can to a large extent be extended to our current analysis. This is effectively the case, and in particular, we can now prove the second part of Theorem 8.19.

Proof of Theorem 8.19(b). We define $\boldsymbol{\gamma}_\infty = \lim_{t\to+\infty} \boldsymbol{\gamma}_t$ as the limit of the optimistic learning rate vector. This quantity is indeed well-defined as $(\gamma_t^i)_{t\in\mathbb{N}}$ is non-negative and non-increasing for every $i \in \mathcal{N}$, which also implies that

$$\sum_{t=1}^{+\infty} \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_{\boldsymbol{\gamma}_t}^2 \geq \sum_{t=1}^{+\infty} \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_{\boldsymbol{\gamma}_\infty}^2 \geq \min_{i\in\mathcal{N}} \gamma_\infty^i \sum_{t=1}^{+\infty} \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2.$$

Consequently, whenever (i) $C \coloneqq \sum_{t=1}^{+\infty} \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_{\boldsymbol{\gamma}_t}^2$ is finite; and (ii) $\min_{i\in\mathcal{N}} \gamma_\infty^i > 0$, it holds that

$$\sum_{t=1}^{+\infty} \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2 \leq \frac{C}{\min_{i\in\mathcal{N}} \gamma_\infty^i} < +\infty.$$

In the remaining of the proof, we show that both *(i)* and *(ii)* hold almost surely, from which we then deduce that the sum $\sum_{t=1}^{+\infty} \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2$ is almost surely finite.

In fact, from Proposition 8.23 we know already that *(ii)* holds almost surely. As for *(i)*, we combine Lemma 8.21 and Lemma 8.22 to get $\sum_{t=1}^{+\infty} \mathbb{E}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_{\boldsymbol{\gamma}_t}^2] < +\infty$. After that, we use Lemma B.4 to deduce that $\sum_{t=1}^{+\infty} \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_{\boldsymbol{\gamma}_t}^2 < +\infty$ almost surely. This concludes the proof. □

8.3.5 *Improved Regret in Self-Play*

Importantly, the (Adapt+) learning rate schedule also ensures optimal dependence on the time horizon for regret in self-play.

Regret bound in self-play

Theorem 8.24. *Suppose that Assumptions 5.1–5.3, 7.1, 8.2 and 8.3 hold and all players run* (OptDA+) *with learning rates given by* (Adapt+)*. Then, for any $i \in \mathcal{N}$ and bounded set $\mathcal{Z}^i \subset \mathcal{X}^i$, we have*

(a) The regret is bounded as

$$\overline{\mathrm{Reg}}_T^i(\mathcal{Z}^i) = O\left(\sqrt{T}\right).$$

(b) If the noise is multiplicative (i.e., $\sigma_A = 0$), we further get

$$\overline{\mathrm{Reg}}_T^i(\mathcal{Z}^i) = O\left(\exp\left(\frac{1}{2r}\right)\right).$$

We recover here the bounds presented in Theorem 8.11, and similarly, the $O(\sqrt{T})$ regret does not depend on the choice of r (it can even be shown for $r \leq 0$). On the contrary, the bound tailored to the multiplicative noise setup, despite being constant with respect to T, has an exponential dependence on $1/r$. This along with Theorems 8.18 and 8.19 underscore the inherent trade-off in the choice of r: larger values of r favor the situation where all players adopt adaptive (OptDA+), while smaller values of r provide better fallback guarantees in adversarial environments.

To prove Theorem 8.24, we refine Lemma 8.5 for adaptive learning rates.

Lemma 8.25. *Suppose that Assumptions 5.1–5.3, 7.1, 8.2 and 8.3 hold and all players run (OptDA+) with adaptive learning rates (Adapt+). Then, for all $i \in \mathcal{N}$, $T \in \mathbb{N}$, and bounded set $\mathcal{Z}^i \subset \mathcal{X}^i$ with $R \geq \sup_{z^i \in \mathcal{Z}^i} \|X_1^i - z^i\|$, it holds that*

$$\begin{aligned}\overline{\mathrm{Reg}}_T^i(\mathcal{Z}^i) \leq \mathbb{E}\Bigg[&\left(\frac{R^2}{2} + 2\sqrt{N}L + 1\right)\sqrt{\Lambda_T^i} + (6L^2 + 2L)\sum_{j=1}^N \sqrt{\Lambda_{T-1}^j} \\ &+ \frac{R^2\sqrt{\Gamma_{T-1}^i}}{2} + \frac{3L^2}{2}\sum_{t=1}^{T-1}\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2 \\ &+ \frac{R^2}{2} + (6NL^2 + 2NL + 2\sqrt{N}L + 2)(G^2 + \bar{\sigma}^2)\Bigg].\end{aligned}$$

Proof. The learning rate γ_t being $\mathcal{F}_{t-1}$-measurable, from Assumption 7.1(a) we deduce $\mathbb{E}[(\gamma_t^i)^2\|\xi_{t-\frac{1}{2}}^i\|^2] \leq \mathbb{E}[(\gamma_t^i)^2\|\hat{V}_{t-\frac{1}{2}}^i\|]$ and subsequently $\mathbb{E}[\|\xi_{t-\frac{1}{2}}\|_{\gamma_t^2}^2] \leq \mathbb{E}[\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|_{\gamma_t^2}^2]$. Plugging these two inequalities into the inequality of Lemma 8.5 and using $\gamma_t^i \leq 1$ gives

$$\begin{aligned}&\max_{z^i \in \mathcal{Z}^i} \mathbb{E}\left[\sum_{t=1}^T \langle V^i(\mathbf{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}}^i - z^i\rangle\right] \\ &\leq \mathbb{E}\Bigg[\frac{R^2}{2\eta_{T+1}^i} + \sum_{t=2}^T \gamma_t^i L^2\left(3\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|_{\gamma_t^2}^2 + \frac{3}{2}\|\mathbf{X}_t - \mathbf{X}_{t-1}\|^2\right) \\ &\quad + \sum_{t=2}^T((\gamma_t^i)^2\sqrt{N}L\|\hat{V}_{t-\frac{1}{2}}^i\|^2 + \frac{L}{\sqrt{N}}\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|_{\gamma_t^2}^2) + \frac{1}{2}\sum_{t=1}^T \eta_t^i\|\hat{V}_{t+\frac{1}{2}}^i\|^2\Bigg] \\ &\leq \mathbb{E}\Bigg[\frac{R^2\sqrt{1 + \Lambda_{T-1}^i + \Gamma_{T-1}^i}}{2} + \sum_{t=1}^{T-1}(3L^2 + L)\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|_{(\gamma_{t+1})^2}^2 + \frac{3L^2}{2}\sum_{t=1}^{T-1}\|\mathbf{X}_t - \mathbf{X}_{t+1}\|^2 \\ &\quad + \sum_{t=1}^{T-1}(\gamma_{t+1}^i)^2\sqrt{N}L\|\hat{V}_{t+\frac{1}{2}}^i\|^2 + \frac{1}{2}\sum_{t=1}^T \eta_t^i\|\hat{V}_{t+\frac{1}{2}}^i\|^2\Bigg].\end{aligned}$$

Since we have both $(\gamma_{t+1}^i)^2 \leq 1/\sqrt{1 + \Lambda_{t-1}^i}$ and $\eta_t^i \leq 1/\sqrt{1 + \Lambda_{t-2}^i}$, applying Lemma 8.15 leads to

$$\begin{aligned}&\sum_{t=1}^{T-1}(\gamma_{t+1}^i)^2\sqrt{N}L\|\hat{V}_{t+\frac{1}{2}}^i\|^2 + \frac{1}{2}\sum_{t=1}^T \eta_t^i\|\hat{V}_{t+\frac{1}{2}}^i\|^2 \\ &\quad \leq 2\sqrt{N}L\left(\sqrt{\Lambda_{T-1}^i} + G^2 + \bar{\sigma}^2\right) + \sqrt{\Lambda_T^i} + 2(G^2 + \bar{\sigma}^2).\end{aligned}$$

Similarly, using Lemma 8.16 we deduce

$$\sum_{t=1}^{T-1}(3L^2 + L)\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|_{(\gamma_{t+1})^2}^2 \leq (3L^2 + L)\left(2N(G^2 + \bar{\sigma}^2) + \sum_{j=1}^N 2\sqrt{\Lambda_{T-1}^j}\right)$$

Putting the above inequalities together and using $\sqrt{1 + \Lambda_{T-1}^i + \Gamma_{T-1}^i} \leq 1 + \sqrt{\Lambda_{T-1}^i} + \sqrt{\Gamma_{T-1}^i}$ gives the desired result. □

With all the lemmas that we have established so far, it is how straightforward to prove Theorem 8.24.

Proof of Theorem 8.24. To prove (a), we note that with Lemma 8.14, we have clearly

$$\mathbb{E}\left[\left(\frac{R^2}{2}+2\sqrt{N}L+\frac{L}{2}\right)\sqrt{\Lambda_t^T}+(6L^2+2L)\sum_{j=1}^{N}\sqrt{\Lambda_{T-1}^j}\right]=O\left(\sqrt{T}\right).$$

Next, thanks to Lemma 8.21 we can bound

$$\begin{aligned}
&\mathbb{E}\left[\frac{R^2\sqrt{\Gamma_{T-1}^i}}{2}+\frac{3L^2}{2}\sum_{t=1}^{T-1}\|\mathbf{X}_t-\mathbf{X}_{t+1}\|^2\right]\\
&\quad\le\frac{R^2}{2}+\mathbb{E}\left[\left(\frac{R^2}{2}+\frac{3L^2}{2}\right)\sum_{t=1}^{T-1}\|\mathbf{X}_t-\mathbf{X}_{t+1}\|^2\right]\\
&\quad\le\frac{R^2}{2}+(4R^2+12L^2)\left(c_1\sum_{i=1}^{N}\mathbb{E}\left[\sqrt{\Lambda_t^{T-1}}\right]+c_2\right).
\end{aligned}$$

This is again in $O(\sqrt{T})$. Plugging the above into Lemma 8.25 shows the regret is indeed in $O(\sqrt{T})$. As for (b) it is immediate by combining Lemma 8.22 and Lemma 8.25. □

8.3.6 *Convergence to Equilibrium under Multiplicative Noise*

In closing, we prove the almost sure last-iterate convergence of adaptive (OptDA+) under multiplicative noise.

Convergence to Nash equilirium under multiplicative noise

Theorem 8.26. *Suppose that Assumptions 5.1–5.3, 7.1, 8.2 and 8.3 hold with $\sigma_A = 0$ and all players run (OptDA+) with adaptive learning rates (Adapt+). Then, both $(\mathbf{X}_t)_{t\in\mathbb{N}}$ and $(\mathbf{X}_{t+\frac{1}{2}})_{t\in\mathbb{N}}$ converge to a Nash equilibrium almost surely.*

Proof. We follow closely the proof of Theorem 8.13. To begin, we fix $\mathbf{x}_\star\in\mathfrak{X}_\star$ and show that we can always apply the Robbins–Siegmund theorem (Lemma 7.8) to inequality (8.5) of Lemma 8.6 (or inequality (8.9) for $t=1$). This gives, for $t\ge 2$,

$$\begin{aligned}
&\mathscr{G}_t=\mathscr{F}_{t-1},\quad U_t=\mathbb{E}_{t-1}[\|\mathbf{X}_t-\mathbf{x}_\star\|_{1/\eta_t}^2],\quad \alpha_t=0,\\
&\zeta_t=\mathbb{E}_{t-1}[\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|_{\gamma_t}^2]+\|\mathbf{V}(\mathbf{X}_{t-\frac{1}{2}})\|_{\gamma_t}^2,\\
&\chi_t=\mathbb{E}_{t-1}[3\|\mathbf{V}(\mathbf{X}_t)-\mathbf{V}(\mathbf{X}_{t-1})\|_{\gamma_t}^2+\|\mathbf{X}_1-\mathbf{x}_\star\|_{1/\eta_{t+1}-1/\eta_t}^2+4\sqrt{N}L\|\xi_{t-\frac{1}{2}}\|_{\gamma_t^2}^2\\
&\qquad+3L^2(\|\boldsymbol{\gamma}_t\|_1\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|_{\gamma_t^2}^2+\|\boldsymbol{\gamma}_{t-1}\|_1\|\hat{\mathbf{V}}_{t-\frac{3}{2}}\|_{(\gamma_{t-1})^2}^2)+2\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\|_{\eta_t}^2].
\end{aligned}$$

As for $t=1$, we replace the above with $\zeta_t=0$ and $\chi_t=\|\hat{\mathbf{V}}_{3/2}\|_{\eta_1}^2$. Using Assumption 5.2, (8.18), and (8.19), we can bound the sum of the expectation of χ_t by

$$\sum_{t=1}^{T}\mathbb{E}[\chi_t]\le\sum_{t=1}^{T-1}3L^2\,\mathbb{E}[\|\mathbf{X}_t-\mathbf{X}_{t+1}\|^2]+\sum_{i=1}^{N}\left(\|X_1^i-x_\star^i\|^2\,\mathbb{E}\left[\sqrt{1+\Lambda_{T-1}^i+\Gamma_{T-1}^i}\right]\right)$$

$$+ (6NL^2 + 4\sqrt{N}L)\left(2N(G^2+\bar{\sigma}^2) + \sum_{i=1}^{N} 2\,\mathbb{E}\left[\sqrt{\Lambda_{T-1}^i}\right]\right) +$$
$$+ 8N(G^2+\bar{\sigma}^2) + \sum_{i=1}^{N} 4\,\mathbb{E}\left[\sqrt{\Lambda_T^i}\right]$$

It then follows immediately from Lemma 8.22 that $\sum_{t=1}^{+\infty}\mathbb{E}[\chi_t] < +\infty$. With the Robbins–Siegmund theorem Lemma 7.8 we deduce that $\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2_{1/\eta_t}]$ converges almost surely to a finite random variable.

As in the proof of Theorem 8.13, we next define $\tilde{\mathbf{X}}_1 = \mathbf{X}_1$ and for all $i \in \mathcal{N}$, $t \geq 2$,

$$\tilde{X}_t^i = X_t^i + \eta_t^i \xi_{t-\frac{1}{2}}^i = -\eta_t^i \sum_{s=1}^{t-2} \hat{V}_{t+\frac{1}{2}}^i - \eta_t^i V^i(\mathbf{X}_{t-\frac{1}{2}}).$$

Then,

$$\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2_{1/\eta_t}] = \|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|^2_{1/\eta_t} + \mathbb{E}_{t-1}[\|\xi_{t-\frac{1}{2}}\|^2_{\eta_t}].$$

Using $\mathbb{E}_{t-1}[\|\xi_{t-\frac{1}{2}}^i\|^2] \leq \mathbb{E}_{t-1}[\|\hat{V}_{t-\frac{1}{2}}^i\|^2]$, the law of total expectation, the fact that $\boldsymbol{\eta}_t$ is $\mathscr{F}_{t-1}$-measurable, Lemma 8.16, and Lemma 8.22, we then get

$$\begin{aligned}
\sum_{t=2}^{+\infty} \mathbb{E}[\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2_{1/\eta_t}] - \|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|^2_{1/\eta_t}] &= \sum_{t=2}^{+\infty} \mathbb{E}[\|\xi_{t-\frac{1}{2}}\|^2_{\eta_t}] \\
&\leq \sum_{t=2}^{+\infty} \mathbb{E}[\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2_{\eta_t}] \\
&\leq 2N(G^2+\bar{\sigma}^2) + \sup_{t\in\mathbb{N}} \sum_{i=1}^{N} 2\,\mathbb{E}\left[\sqrt{\Lambda_t^i}\right] \\
&< +\infty. \qquad (8.26)
\end{aligned}$$

Invoking Lemma B.4 we deduce that $\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2_{1/\eta_t}] - \|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|^2_{1/\eta_t}$ almost surely converges to 0. Since we have shown $\mathbb{E}_{t-1}[\|\mathbf{X}_t - \mathbf{x}_\star\|^2_{1/\eta_t}]$ almost surely converges to a finite random variable, we now know that $\|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|^2_{1/\eta_t}$ almost surely converges to this finite random variable as well. To summarize, we have shown that for any $\mathbf{x}_\star \in \mathfrak{X}_\star$, $\|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|_{1/\eta_t}$ converges almost surely.

To proceed, let us define $\boldsymbol{\eta}_\infty = \lim_{t\to+\infty} \boldsymbol{\eta}_t$. This limit always exists because $(\eta_t^i)_{t\in\mathbb{N}}$ is a non-negative non-increasing sequence for every $i \in \mathcal{N}$. Moreover, by Proposition 8.23 we know that $\boldsymbol{\eta}_\infty$ is positive almost surely, and thus $(1/\boldsymbol{\eta}_\infty)$, the limit of $(1/\boldsymbol{\eta}_t)_{t\in\mathbb{N}}$ is finite almost surely. Applying Lemma B.6 with $\mathcal{Z} \leftarrow \mathfrak{X}_\star$, $\mathbf{u}_t \leftarrow \tilde{\mathbf{X}}_t$, and $\boldsymbol{\alpha}_t \leftarrow 1/\boldsymbol{\eta}_t$, we then deduce that with probability 1, $\|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|_{1/\eta_\infty}$ converges for all $\mathbf{x}_\star \in \mathfrak{X}_\star$.

Next, with $\|\mathbf{X}_t - \tilde{\mathbf{X}}_t\|^2 = \|\xi_{t-\frac{1}{2}}\|^2_{\eta_t^2}$ and $\|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_t\|^2 = \sum_{i=1}^N \|\gamma_t^i \hat{V}_{t-\frac{1}{2}}^i + \eta_t^i \xi_{t-\frac{1}{2}}^i\|^2$, following the reasoning of (8.26), we get both $\sum_{t=1}^{+\infty}\mathbb{E}[\|\mathbf{X}_t - \tilde{\mathbf{X}}_t\|^2] < +\infty$ and $\sum_{t=1}^{+\infty}\mathbb{E}[\|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_t\|^2] < +\infty$. By Lemma B.4 we then know that $\|\mathbf{X}_t - \tilde{\mathbf{X}}_t\|$ and $\|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_t\|$ converge to 0 almost surely. Finally, from Theorem 8.19(b) we know that $\|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|$ converges to 0 almost surely. To conclude, let us define the event

$$\mathcal{E} := \left\{ \begin{array}{c} 1/\boldsymbol{\eta}_\infty \text{ is finite and } \|\tilde{\mathbf{X}}_t - \mathbf{x}_\star\|_{1/\eta_\infty} \text{ converges for all } \mathbf{x}_\star \in \mathfrak{X}_\star, \\ \lim\limits_{t\to+\infty} \|\mathbf{X}_t - \tilde{\mathbf{X}}_t\| = 0, \quad \lim\limits_{t\to+\infty} \|\mathbf{X}_{t+\frac{1}{2}} - \tilde{\mathbf{X}}_t\| = 0, \quad \lim\limits_{t\to+\infty} \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\| = 0 \end{array} \right\}$$

We have shown that $\mathbb{P}(\mathcal{E}) = 1$. Moreover, with the arguments of Theorem 8.13 we deduce that whenever $\mathcal{E}$ happens both $(\mathbf{X}_t)_{t\in\mathbb{N}}$ and $(\mathbf{X}_{t+\frac{1}{2}})_{t\in\mathbb{N}}$ converge to a point in $\mathfrak{X}_\star$, and this ends the proof. □

8.4 NUMERICAL ILLUSTRATIONS

In this section, we numerically illustrate the performance of the algorithms that we studied in the previous two chapters via experiments on a bilinear game and on a non-monotone game. All the players use the same algorithm in the these experiments.

8.4.1 *A Bilinear Zero-sum Game*

For this part we consider the simple problem of finding the Nash equilibrium of the game

$$\mathcal{X}^1 = \mathcal{X}^2 = \mathbb{R}, \quad \ell^1(\theta, \phi) = \theta\phi = -\ell^2(\theta, \phi).$$

As already mentioned in Examples 6.1 and 7.1, the unique Nash equilibrium of this game is located at $(0, 0)$.

Algorithms of interest

We focus here on four different algorithms: gradient descent/ascent (GDA) also known as the vanilla gradient method, (OG), (OG+), and AdaOptDA+, i.e., (OptDA+) with learning rate (Adapt+). For both (GDA) and (OG) we use decreasing learning rate of the form $\eta_t = \eta/(t+1)^{r_\eta}$. As for (OG+) we use either constant learning rates or $\gamma_t = \gamma/(t+1)^{r_\gamma}$ and $\eta_t = \eta/(t+1)^{r_\eta}$. We take $r = 1/4$ for AdaOptDA+. In terms of feedback oracle, we examine two situations that correspond respectively to the case of additive noise ($\sigma_A > 0$, $\sigma_M = 0$) and multiplicative noise ($\sigma_A = 0$, $\sigma_M > 0$). We use respectively $\gamma = \eta = 1$ and $\gamma = \eta = 0.1$ in the definition of the decreasing learning rates in these two cases.

The case of additive noise

ALMOST SURELY BOUNDED ADDITIVE NOISE. We first explore the stochastic feedback as described in Example 7.1. In this scenario, the second player receives perfect feedback while the feedback of the first player is deteriorated by an additive noise which assumes either 1 or -1 with equal probability. The metrics used to evaluate performance, as illustrated in Fig. 8.1, are: (*i*) the first player's regret with respect to the comparator point 0 (*ii*) the distance between the players' base state $\mathbf{X}_t$ and the Nash equilibrium, and (*iii*) the distance between the players' played/leading state $\mathbf{X}_{t+\frac{1}{2}}$ and the Nash equilibrium.

The results for this case is shown in the top row of Fig. 8.1. The leftmost figure indicates that the adoption of optimistic strategies indeed yields significantly lower regret, though these regrets still grow slowly over time. The only exception is represented by the purple curve, leading to high regret due to the use of a constant optimistic step. In this particular case, we have the convergence of the base state $\mathbf{X}_t$ but the divergence of the played/leading state $\mathbf{X}_{t+\frac{1}{2}}$. This intriguing phenomenon is in accordance with our analysis in [126] regarding the use of (EG+) for affine operators (note however that for (EG+) the leading sate $\mathbf{X}_{t+\frac{1}{2}}$ roughly stays at a constant distance from the equilibrium instead of diverging). For (OG+) with both learning rates decreasing and AdaOptDA+, we observe convergence of both the base and the leading states, and these convergence happen at a comparable speed.

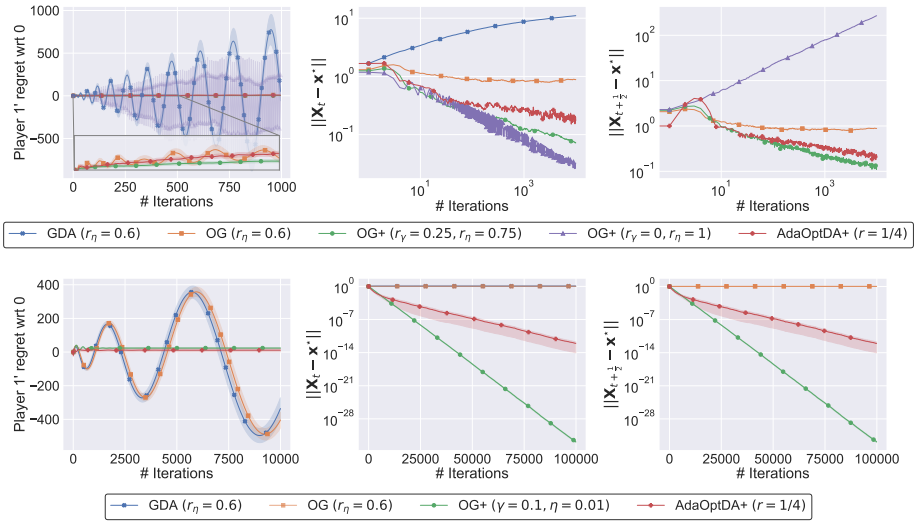


Figure 8.1: Regret of the first player with respect to 0 and the distance between the iterates and the Nash equilibrium when the players run one of the shown algorithms in the bilinear game $\min_{\theta\in\mathbb{R}} \max_{\phi\in\mathbb{R}} \theta\phi$. In the top row, the feedback is affected by an additive noise, while in the bottom row, the feedback is affected by a multiplicative noise. The results are averaged over 50 runs and shaded areas indicate standard errors.

The case of multiplicative noise

PROBLEM WITH A FINITE SUM STRUCTURE. We next consider noise that arises from the sampling of a problem that admits a finite-sum structure. For this, let us define $\mathcal{L}_1(\theta,\phi) = 3\theta\phi$ and $\mathcal{L}_2(\theta,\phi) = -\theta\phi$ so that $\ell^1 = -\ell^2 = (\mathcal{L}_1 + \mathcal{L}_2)/2$. At each round, we randomly draw $\mathcal{L}_1$ or $\mathcal{L}_2$ with probability $1/2$ and return the gradient of the sampled function as feedback. Assumption 7.1 is clearly satisfied here with $\sigma_A = 0$ and $\sigma_M = 2$, so the noise is multiplicative.

The results shown in the bottom row of Fig. 8.1 reveals that (GDA) and (OG) exhibit similar behavior in this setup. The iterates cycle around the equilibrium at a fixed distance,[4] leading to a regret that oscillates but whose magnitude grows over time. As for (OG+) (with constant learning rates since the noise is multiplicative) and AdaOptDA+, we indeed have constant regret and geometric convergence. This aligns with our theory though we are not able to prove the geometric convergence of AdaOptDA+.

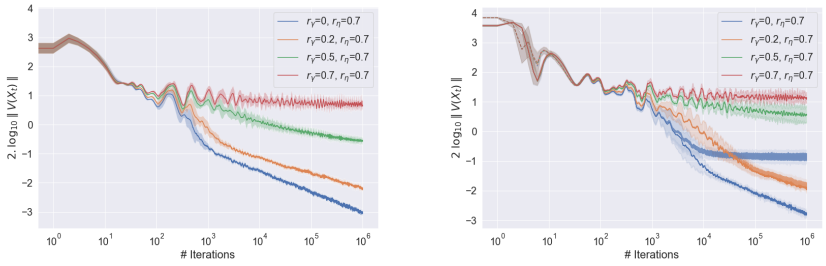
8.4.2 *Linear Quadratic Gaussian GAN*

Moving on, we examine below the behaviors of (EG+) and (OG+) in a non convex-concave linear quadratic Gaussian GAN model as defined by Daskalakis et al. [56] and Nagarajan and Kolter [201]. This is a saddle-point problem with the following objective.

$$\mathcal{L}(J,W) = \mathbb{E}_{a\sim\mathcal{N}(0,\Sigma)}[a^\top W a] - \mathbb{E}_{\xi\sim\mathcal{N}(0,I)}[\xi^\top J^\top W J \xi].$$

Here, a and ξ are vectors in $\mathbb{R}^d$ for some $d > 0$ while J, W, and Σ are matrices of size $d \times d$. I denotes the $d \times d$ identity matrix. The first player plays J and wants to minimize $\mathcal{L}$, while the second player plays W and aims to maximize $\mathcal{L}$. This corresponds to the WGAN formulation [9] without clipping when data are sampled from a normal distribution with covariance matrix Σ, i.e., $a \sim \mathcal{N}(0,\Sigma)$,

4 while these iterate may eventually spiral out, this happens at a very low speed due to the use of decreasing learning rates



(a) EG+ (b) OG+

Figure 8.2: Convergence of (EG+) and (OG+) in the linear quadratic Gaussian GAN model. For (OG+), the dashed lines and the solid lines depict respectively the results for the leading states and the base states. All curves are averaged over 10 runs with the shaded area indicating the standard deviation.

and the generator and the discriminator are respectively defined by $G(\xi) = J\xi$, $D(a) = a^\top W a$. The stochasticity is induced by the sampling of a and ξ. For the experiments we take a mini-batch of size 128 and a and ξ of dimension 10. Σ is a randomly generated positive definite matrix whose eigenvalues are uniformly drawn from the interval $[1, 2]$.

For the algorithms we use the learning rates

$$\gamma_t = \frac{\gamma}{(t+19)^{r_\gamma}} \quad \text{and} \quad \eta_{t+1} = \frac{\eta}{(t+19)^{r_\eta}}.$$

The choice of 19 is arbitrary and the main goal is to avoid too fast decrease of the learning rates in the early iterations, as otherwise it is hard to see the progress. However, in this case for fixed γ and η the initial learning rates γ_1 and η_2 vary according to the exponents r_γ and r_η. To account for this, we rather fix the initial learning rates and set $\gamma = \gamma_1 \cdot 20^{r_\eta}$ and $\eta = \eta_2 \cdot 20^{r_\eta}$. Provided that the game is non-monotone and may possess multiple equilibria, the squared norm of V is traced as the convergence measure.

As we can see from Fig. 8.2, the convergence speeds of (EG+) and (OG+) are positively related to the difference $r_\eta - r_\gamma$, echoing the results that we have shown in Theorem 7.13. Moreover, again, choosing $r_\gamma < r_\eta$ is necessary for the convergence of the algorithm.

DISCUSSION AND CONCLUDING REMARKS

9

CONCLUSION AND PERSPECTIVES

Conclusion

This thesis delves into the intricacies of decision-making in multi-agent systems, focusing primarily on two aspects: communication delays and conflicts of interest that arise from agents' selfish behaviors. These elements introduce unique challenges, absent in single-agent settings, and understanding their impact on system performance is of paramount importance.

In Part I, we worked toward this goal by examining the impact of delays in cooperative online learning. We analyzed the regret of a version of dual averaging that operates with delayed feedback. Precisely, the action taken by a learner is computed with the sum of the feedback gradients that they have received so far. Building on the template regret bound demonstrated in Theorem 3.1, we developed a series of adaptive learning rates that lead to regret bounds that self-adapt to both the the magnitudes of the feedback and the delays associated with each piece of feedback. In Chapter 4, we further showed that by incorporating an optimistic step with an optimistic learning rate that is in the order of the maximum delay, we can mitigate the impact of delays when the loss functions vary slowly over time.

Subsequently, in Part II, we dedicated ourselves to the problem of learning in continuous games. We respectively looked into adaptive methods, noisy feedback, and their combination, in Chapter 6, 7, and 8. Our algorithms are based on optimistic mirror descent and optimistic dual averaging. The adaptive learning rate scheme resembles that of AdaGrad and allows the algorithms to be tuned individually by each player even in the absence of prior knowledge about the game or the noise profile. To handle noisy feedback, we proposed to employ a double-learning-rate schedule for the algorithms, akin to what we do to handle delays with optimism in Chapter 4. These algorithms enjoy (near-)optimal regret guarantees in various situations, while achieving convergence to a Nash equilibrium in variationally stable games (Definition 5.4) when followed by all the players, all of which indicate the promise of our approaches.

Despite our efforts, we acknowledge that the real-world scenario is invariably more complex. Not only we have more complicated models for both cooperative and competitive learning, but also the nature of multi-agent interaction is rarely entirely cooperative or competitive. Still, it is our hope that the insights offered in this thesis will help deepen our comprehension and enhance our ability to design solutions for this intricate field of learning in multi-agent systems.

Perspectives

Looking forward, our research paves the way for numerous unexplored questions, each presenting exciting opportunities for future inquiry. In the following, I enumerate some of these questions that have captured my attention during my Ph.D. This also affords me an opportunity to touch upon related work from a wider perspective. While the upcoming list is far from exhaustive, it should provide a useful guide pointing to some of the active research areas in this field.

To begin, we first discuss several points related to the algorithms that we study and the specific results that we obtain in this manuscript.

DOUBLE-LEARNING-RATE METHODS WITH CONSTRAINED ACTION SET(S). In Chapters 4, 7 and 8, we restrict ourselves to the unconstrained setup. Naturally, one might wonder if these results extend to the constrained setup. Unfortunately, such extensions are not straightforward. Concerning the use of optimistic steps in addressing delays as studied in Chapter 4, Flaspohler et al. [80] indeed investigated the more general MD and FTRL algorithms. Nevertheless, their results significantly differ from ours and their analysis does not allow to recover the $\sqrt{\tau}$ dependence that we can obtain from Theorem 4.3.

As it relates to the setup for learning in games with noisy feedback considered in Chapters 7 and 8, it turns out that a simple double-learning-rate strategy falls short in the general constrained case. This indicates a need for further modifications to the algorithm. For example, it appears that performing an optimistic step with some average of the past gradients can be helpful.

Equally, it is essential to remember that there are existing techniques that have proven effective in achieving convergence under similar conditions, such as mini-batching [23, 134] and Tikhonov regularization / Halpern iteration [32, 161]. While these methods offer different strategic approaches, they also come with their own set of challenges. For example, mini-batching appears to be less applicable in the online setup, whereas the use of Tikhonov regularization can slow down the algorithm in situations where geometric convergence could otherwise be achieved. Another interesting attempt was recently made by Pethick et al. [222], where a variant of (EG+) with an additional bias correction term for the optimistic step is considered. An important caveat is however that their method requires to sample the two stochastic gradients of an iteration with the same random seed, presenting substantial challenges for its adaptation to online setups.

THE ROLE OF LEARNING RATE SEPARATION IN OPTIMISTIC METHODS. Independent of our work, the benefit of having a larger optimistic step has also been demonstrated in other situations. Zhang and Yu [289] and Fasoulakis et al. [79] respectively showed faster convergence of OG+ and OMWU+ (a version of OMWU with separate learning rate tuning) when the optimistic learning rate is taken larger than the update one. In another series of work, it is established that learning rate separation helps achieve convergence in a larger family of games—those for which a *weak Minty solution* exists [65, 171, 221, 222]. This leads us to ponder: Is the effectiveness of learning rate separation merely coincidental? Or could there potentially be a deeper rational that explains its utility in these diverse scenarios?

LAST-ITERATE CONVERGENCE OF STOCHASTIC DUAL AVERAGING. In Chapter 8, we were unable to show last-iterate convergence of (OptDA+) under the general noise model. The challenge is tied closely to the use of the dual averaging template, and more specifically, to the application of a uniform and vanishing learning rate to all the feedback. Considering the simplicity of the algorithm and the perceived importance of this result, it is puzzling that no conclusive results exist regarding the convergence or non-convergence of the last iterate of dual averaging when run with stochastic feedback.

LAST-ITERATE CONVERGENCE RATE IN MONOTONE GAMES WITH STOCHASTIC FEEDBACK. Only recently have a series of papers been successful in demys-

tifying the last-iterate convergence rate of (EG) and (OG) in monotone games [35, 99, 100, 102]. As measured by the tangent residual (as defined in Definition 5.10), this rate stands at $O(1/\sqrt{t})$. Notably, the rate can be further improved to $O(1/t)$ by integrating an anchoring step [33, 36, 171]. However, such results for the stochastic setup are still scarce at present.

Our result in Theorems 7.13 and 7.14 translates into a rate in $O(1/t^{1/6})$ on the norm of pseudo-gradient, which appears suboptimal. One might then conjecture that by wisely adjusting the batch size in mini-batching or the regularization parameter in Tikhonov regularization we can get a rate in $O(1/t^{1/4})$ under stochastic feedback. However, even this may be suboptimal. In fact, with more involved methods, Cai et al. [32] obtained an $O(1/t^{1/3})$ rate for stochastic monotone inclusion problems, with Chen and Luo [42] further improving this to a near-optimal $O(1/\sqrt{t})$ rate for stochastic convex-concave saddle-point problems. Unfortunately, these algorithms are not suited to the learning-in-games setup that we consider. Therefore, it remains to be seen if such guarantees can be achieved for an algorithm that conforms to our protocol while fulfilling the basic no-regret requirement.

We next move on to discuss the criteria that we have used to evaluate our algorithms.

BEYOND EXTERNAL REGRET. In this thesis, we focus exclusively on bounding the agents' *external/static* regret, as defined in Definition 2.1. This is arguably one of the simplest forms of regret that one can imagine. Nonetheless, depending on the context, this may not always be the most appropriate measure to consider. For example, taking into account the non-stationary nature of the problem, one may want to instead control the dynamic regret [105, 298], which is evaluated with respect to a sequence of comparator points that evolve over time.

Turning our attention to the cooperative setup studied in Chapters 3 and 4, we could also think of a form of regret that uses a different comparator point for each agent. Even though it might appear that agents just need to minimize their own losses and ignore completely the others in this situation, it turns out that communication between agents can still be advantageous if the tasks that they tackle exhibit certain similarities. Such a concept has been formalized by Cavallanti et al. [38] and Cesa-Bianchi et al. [40], among others. Regarding the setup of learning in games that we examine in Part II, we may want to account for the fact that opponents could have modified their actions if we had chosen a different sequence of actions. This consideration leads to the notion of policy regret [10, 11].

MORE COMPLEX INTERACTION PARADIGMS. In Part II of this manuscript, we primarily assess the performance of our algorithms in two specific situations: the adversarial scenario and the case where all players utilize the same type of algorithm. However, real-world applications often present complexities that go beyond this dichotomy. For instance, it is possible that only a subset of the players deviate from their prescribed policy. Such occurrences urge us to investigate the robustness of the results obtained in the self-play scenario against these partial deviations.

Another compelling direction would be to consider *opponent shaping* [85, 181], where players do not just respond to the behavior of others but actively seek to influence it. Such a scenario reflects a deeper level of strategic thinking and opens up interesting avenues for exploration. Are there specific tactics that

prove particularly effective for shaping the behavior of opponents? Can the idea of no-regret learning be extended to cover these more sophisticated strategies?

Moving forward, another promising direction is to extend our results to more complex setups, either by considering all the factors that we have tackled simultaneously, or by considering a more fundamental change of paradigm.

TIME-VARYING GAMES. In Part I, we address a situation where the agents' feedback is non-stationary, arbitrary, and even adversarial. In Part II, we then consider a scenario where this non-stationarity originates from the interaction with other agents through an underlying game, which we assume to be fixed across time. Yet, in many real-world applications, the game itself evolves over time. This paradigm of time-varying games can be regarded as the de facto mix of the non-stationary online learning framework and the learning-in-games setup, and requires the use of different criteria to evaluate the algorithms [5, 71, 293]. Naturally, dynamic regret that we just mentioned turns out to be particularly relevant for this situation. Moreover, insights gleaned from time-varying optimization [51] might provide valuable guidance here.

LEARNING IN GAMES WITH NON-INSTANTANEOUS FEEDBACK. Another notable distinction between Parts I and II of this manuscript is that we sidestep issues related to communication latency in Part II. Indeed, in Part II, we work under the assumption that players receive immediate feedback upon executing each action, a condition that that may not generally hold true. Recognizing these limitations, several works have addressed the communication aspect of learning in games. This body of work encompasses models that adopt the abstraction of delayed feedback [118, 295], similar to our approach in Chapter 3, as well as those incorporating a more explicit model with a communication graph [89, 238, 257, 262, 286]. These studies are complementary to our research, and combining these approaches with our methodologies constitutes a promising avenue for future exploration.

BEYOND VARIATIONALLY STABLE GAMES. While this thesis primarily focuses on continuous normal-form games, and especially variationally stable games, considerable efforts have been made within the community to extend such low-regret and last-iterate guarantees to a wider scope of games. These efforts encompass, one one hand, poly-logarithmic regret results obtained in general-sum finite games [2, 3, 57], extensive-form games [78], and general convex games [77], and on the other hand, last-iterate convergence results for zero-sum extensive-form games [169, 223], alongside several dichotomy results for two-player general-sum finite games [4, 61], and a number of attempts in achieving convergence in various types of non-convex games [59, 65, 83, 176] (a game is non-convex if Assumption 5.1 is not satisfied).

The study of these setups, combined with the different challenges that we aim to address in this thesis, all contribute to a more comprehensive and nuanced understanding of the learning-in-games framework. The study of non-convex games is of particular importance given the current surge of interest in deep learning techniques, which almost inevitably involve non-convex optimization landscapes. Nonetheless, such games present considerable challenges, and even the existence of a Nash equilibrium is not guaranteed [58]. For a thorough discussion on the difficulties and potential directions for learning in non-convex games, we invite the reader to consult the monograph [52] (note that what

we call "non-convex game" here is referred to as "non-concave game" in this monograph as it considers players that aim to maximize their payoffs).

PAYOFF-BASED LEARNING. The algorithms examined in this manuscript assume access to the gradient feedback or an unbiased estimate thereof. However, in many practical scenarios, even the latter assumption can be too demanding. Instead, a learner may only have access to the payoff information at the taken action, whether it be the conversion rate of a recommendation or the outcome of an auction. This situation is often referred to as the bandit setup [30, 168].

A large corpus of work has been dedicated to addressing bandit learning in multi-agent systems. In cooperative bandits, the agents collaborate to minimize their regrets, similar to the model considered in Part I. For this scenario, algorithms have been developed to deal with either stochastic or adversarial feedback, while taking into account the delays induced by communication [17, 39, 164, 187]. Information sharing among agents can turn out to be particularly helpful here as it allows the agents to perform more aggressive exploration.

Concurrently, payoff-based learning has also been extensively studied in learning in games. In fact, some of the most well-known algorithms for learning in games such as fictitious play [28] and regret matching [110] are widely recognized as payoff-based methods. Particularly relevant to the results presented in Part II is the convergence of bandit learning in strictly monotone games, as shown by Bravo et al. [27] and Tatarenko and Kamgarpour [260]. These approaches employ a single-point estimator to deduce a (biased) estimate of the gradient from bandit feedback, in the spirit of [81, 213]. More recently, Gao and Pavel [91] and Tatarenko and Kamgarpour [261] extended these results to (non-strictly) monotone games through the incorporation of a Tikhonov regularization scheme. A few works, namely [118, 133], have also tried to address the challenges posed by delays, learning in games, and bandit feedback all at the same time. All these works can provide valuable insights into how we can extend our results to cope with bandit feedback.

Finally, we would also like to highlight two important subjects that are less related to the works realized in this thesis, but also fall under the umbrella topic of "decision-making in multi-agent systems".

META-LEARNING IN MULTI-AGENT SYSTEMS. Despite the non-stationary nature and heterogeneous composition of many real-world multi-agent environments, recurring patterns and shared structures are often present. This is observed in diverse scenarios, such as traffic networks [122], which despite changing constantly, tend to exhibit regular trends related to the time of day or week. In federated learning [144], users across different devices strive to improve their personal models based on locally collected data, while also leveraging shared knowledge across the network to enhance their learning efficiency. Similarly, for an autonomous robotic teams, successful strategies in one context often prove useful in another, even when facing a diverse range of scenarios [131, 266]. This shared structure across varying tasks or environments is the foundation of *meta-learning*, or *learning to learn* [264], a powerful tool for extracting and applying an *inductive bias* across tasks, enabling effective adaptation to the ever-evolving contexts of multi-agent systems.

Some recent works in this direction include for example [98, 109, 150, 173]. Concretely, Kayaalp et al. [150] and Li et al. [173] focused on meta-learning for decentralized optimization, respectively addressing the problem of meta-learning the initialization parameter and the mixing weights of the models.

The problem of meta-learning in games is explored by Harris et al. [109], who proposed an online learning-based approach that provably achieves lower regret when the games presented to the players obey a certain notion of similarity. On the other hand, Goktas et al. [98] investigated learning to compute a (generalized) Nash equilibrium across a class of (pseudo-)games, and they introduced a generative adversarial learning mechanism for this purpose.

While these works have made important contributions toward understanding and implementing meta-learning in multi-agent contexts, there remains much to explore. Continued advancements in this field could revolutionize how we approach decision-making and learning in multi-agent systems, leading to a new generation of adaptable and robust solutions for such complex, dynamic environments.

MULTI-AGENT REINFORCEMENT LEARNING. Following the breakthroughs in deep learning methods, recent years have witnessed substantial advancements in the domain of *multi-agent reinforcement learning* (MARL). These developments have enabled the solution of a spectrum of complex games such as Go [252], StarCraft [271], and Diplomacy [74], as well as practical applications in traffic control [46], stock trading [170], and robotics [240], just to name a few. For an overview on this topic, readers may refer to the surveys [31, 104, 292].

Despite the tremendous success of these methods, our theoretical understanding about them remain relatively limited. From a mathematical perspective, MARL is mostly modeled as a *stochastic game*. Introduced in the seminal work of Shapley [249], this model is also known under the name of Markov game [179], highlighting the Markovian aspect of the framework. Numerous works have explored the efficiency and convergence of algorithms within this setting —for an appetizer, see e.g., [16, 95, 242, 277, 283, 291] and references therein. However, these investigations are generally constrained in the types of games considered (mostly restricted to two-player, zero-sum, or potential games) and the variety of function approximation utilized, if any.

Overall, while MARL has seen impressive progress and is increasingly being applied to solve complex problems, many challenges remain. These include developing a deeper theoretical understanding, handling sophisticated aspects like communication learning [84], partial observability [66, 220], and memory mechanisms [93], and expanding the scope of current algorithmic approaches. Addressing these issues is not only essential for further advancements in the field, but also critical in building more believable agents in multi-agent systems [219]. As we continue to refine and enhance MARL, the potential for these systems to mirror, predict, and interact with real-world complex behaviors grows dramatically.

BIBLIOGRAPHY

[1] Jacob D. Abernethy, Peter L. Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Conference on Learning Theory*, 2008.

[2] Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 736–749, 2022.

[3] Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Tuomas Sandholm. Uncoupled learning dynamics with $o(\log t)$ swap regret in multiplayer games. In *Advances in Neural Information Processing Systems*, volume 35, pages 3292–3304, 2022.

[4] Ioannis Anagnostides, Gabriele Farina, Ioannis Panageas, and Tuomas Sandholm. Optimistic mirror descent either converges to nash or to strong coarse correlated equilibria in bimatrix games. In *Advances in Neural Information Processing Systems*, 2022.

[5] Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On the convergence of no-regret learning dynamics in time-varying games. *arXiv preprint arXiv:2301.11241*, 2023.

[6] Kimon Antonakopoulos, Veronica Belmega, and Panayotis Mertikopoulos. An adaptive mirror-prox method for variational inequalities with singular operators. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[7] Kimon Antonakopoulos, Veronica Belmega, and Panayotis Mertikopoulos. Adaptive extra-gradient methods for min-max optimization and games. In *International Conference on Learning Representations*, 2021.

[8] Kimon Antonakopoulos, Thomas Pethick, Ali Kavis, Panayotis Mertikopoulos, and Volkan Cevher. Sifting through the noise: Universal first-order methods for stochastic variational inequalities. In *Advances in Neural Information Processing Systems*, volume 34, pages 13099–13111, 2021.

[9] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[10] Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In *International Coference on Machine Learning*, pages 1747–1754, 2012.

[11] Raman Arora, Michael Dinitz, Teodor Vanislavov Marinov, and Mehryar Mohri. Policy regret in repeated games. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[12] Mahmoud Assran, Arda Aytekin, Hamid Reza Feyzmahdavian, Mikael Johansson, and Michael G Rabbat. Advances in asynchronous parallel and distributed optimization. *Proceedings of the IEEE*, 108(11):2013–2031, 2020.

[13] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pages 322–331. IEEE, 1995.

[14] Peter Auer, Nicolo Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

[15] Francis Bach and Kfir Y Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Conference on Learning Theory*, pages 164–194. PMLR, 2019.

[16] Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pages 551–560. PMLR, 2020.

[17] Yogev Bar-On and Yishay Mansour. Individual regret in cooperative nonstochastic multi-armed bandits. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[18] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[19] Dimitri Bertsekas and John Tsitsiklis. *Parallel and distributed computation: numerical methods*. Athena Scientific, 2015.

[20] David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.

[21] Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(6), 2007.

[22] Avrim Blum, Eyal Even-Dar, and Katrina Ligett. Routing without regret: on convergence to Nash equilibria of regret-minimizing in routing games. In *PODC '06: Proceedings of the 25th annual ACM SIGACT-SIGOPS symposium on Principles of Distributed Computing*, pages 45–52, 2006.

[23] Radu Ioan Boţ, Panayotis Mertikopoulos, Mathias Staudigl, and Phan Tu Vuong. Minibatch forward-backward-forward methods for solving stochastic variational inequalities. *Stochastic Systems*, 11(2):112–139, 2021.

[24] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.

[25] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

[26] Mario Bravo and Panayotis Mertikopoulos. On the robustness of learning in games with stochastically perturbed payoff observations. *Games and Economic Behavior*, 103, John Nash Memorial issue:41–66, May 2017.

[27] Mario Bravo, David Leslie, and Panayotis Mertikopoulos. Bandit learning in concave n-person games. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[28] George W. Brown. Iterative solution of games by fictitious play. In *Activity Analysis of Production and Allocation*. Wiley, New York, 1951.

[29] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[30] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[31] Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

[32] Xufeng Cai, Chaobing Song, Cristóbal Guzmán, and Jelena Diakonikolas. Stochastic halpern iteration with variance reduction for stochastic monotone inclusions. In *Advances in Neural Information Processing Systems*, volume 35, pages 24766–24779, 2022.

[33] Yang Cai and Weiqiang Zheng. Doubly optimal no-regret learning in monotone games. In *International Conference on Machine Learning*, 2023.

[34] Yang Cai, Ozan Candogan, Constantinos Daskalakis, and Christos Papadimitriou. Zero-sum polymatrix games: A generalization of minmax. *Mathematics of Operations Research*, 41(2):648–655, 2016.

[35] Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *Advances in Neural Information Processing Systems*, volume 35, pages 33904–33919, 2022.

[36] Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Accelerated algorithms for monotone inclusions and constrained nonconvex-nonconcave min-max optimization. *arXiv preprint arXiv:2206.05248*, 2022.

[37] Anca Capatina. *Variational inequalities and frictional contact problems*, volume 31. Springer, 2014.

[38] Giovanni Cavallanti, Nicolo Cesa-Bianchi, and Claudio Gentile. Linear algorithms for online multitask classification. *The Journal of Machine Learning Research*, 11: 2901–2934, 2010.

[39] Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Delay and cooperation in nonstochastic bandits. *Journal of Machine Learning Research*, 20(17):1–38, 2019.

[40] Nicolò Cesa-Bianchi, Pierre Laforgue, Andrea Paudice, and Massimiliano Pontil. Multitask online mirror descent. *Transactions of Machine Learning Research*, 2022.

[41] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3): 538–543, 1993.

[42] Lesi Chen and Luo Luo. Near-optimal algorithms for making the gradient small in stochastic minimax optimization. *arXiv preprint arXiv:2208.05925*, 2022.

[43] Xi Chen and Binghui Peng. Hedging in games: Faster convergence of external and swap regrets. In *Advances in Neural Information Processing Systems*, 2020.

[44] Yun Kuen Cheung and Georgios Piliouras. Vortices instead of equilibria in minmax optimization: Chaos and butterfly effects of online learning in zero-sum games. In *Conference on Learning Theory*, 2019.

[45] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Conference on Learning Theory*, 2012.

[46] Tianshu Chu, Jie Wang, Lara Codecà, and Zhaojian Li. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1086–1095, 2019.

[47] Kuo-Liang Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3):463–483, 1954.

[48] Patrick L. Combettes. Quasi-Fejérian analysis of some optimization algorithms. In *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*, pages 115–152. Elsevier, New York, NY, USA, 2001.

[49] Patrick L. Combettes and Jean-Christophe Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248, 2015.

[50] Lorenzo Croissant, Marc Abeille, and Clément Calauzènes. Real-time optimisation for online learning in auctions. In *International Conference on Machine Learning*, pages 2217–2226. PMLR, 2020.

[51] Emiliano Dall'Anese, Andrea Simonetto, Stephen Becker, and Liam Madden. Optimization and learning with information streams: Time-varying algorithms and applications. *IEEE Signal Processing Magazine*, 37(3):71–83, 2020.

[52] Constantinos Daskalakis. Non-concave games: A challenge for game theory's next 100 years. 2022.

[53] Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *ITCS '19: Proceedings of the 10th Conference on Innovations in Theoretical Computer Science*, 2019.

[54] Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, 39(1): 195–259, 2009.

[55] Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 235–254. SIAM, 2011.

[56] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 2018.

[57] Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[58] Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1466–1478, 2021.

[59] Constantinos Daskalakis, Noah Golowich, Stratis Skoulakis, and Manolis Zampetakis. Stay-on-the-ridge: Guaranteed convergence to local minimax equilibrium in nonconvex-nonconcave games. *arXiv preprint arXiv:2210.09769*, 2022.

[60] Nair Maria Maia De Abreu. Old and new results on algebraic connectivity of graphs. *Linear algebra and its applications*, 423(1):53–73, 2007.

[61] Étienne de Montbrun and Jérôme Renault. Convergence of optimistic gradient descent ascent in bilinear games. *arXiv preprint arXiv:2208.03085*, 2022.

[62] Gerard Debreu. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences*, 38(10):886–893, 1952.

[63] Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by d-adaptation. In *International Conference on Machine Learning*, 2023.

[64] Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345):118–121, 1974.

[65] Jelena Diakonikolas, Constantinos Daskalakis, and Michael I Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR, 2021.

[66] Jilles Dibangoye and Olivier Buffet. Learning to act in decentralized partially observable mdps. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2018.

[67] Asen L Dontchev and R Tyrrell Rockafellar. *Implicit functions and solution mappings: A view from variational analysis*, volume 11. Springer, 2009.

[68] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

[69] John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011.

[70] Muhammad Nouman Durrani and Jawwad A Shamsi. Volunteer computing: requirements, challenges, and solutions. *Journal of Network and Computer Applications*, 39:369–380, 2014.

[71] Benoit Duvocelle, Panayotis Mertikopoulos, Mathias Staudigl, and Dries Vermeulen. Multiagent online learning in time-varying games. *Mathematics of Operations Research*, 48(2):914–941, 2023.

[72] Alina Ene and Huy Lê Nguyen. Adaptive and universal algorithms for variational inequalities with optimal convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6559–6567, 2022.

[73] Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research. Springer, 2003.

[74] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624): 1067–1074, 2022.

[75] Ky Fan. Fixed-point and minimax theorems in locally convex topological linear spaces. *Proceedings of the National Academy of Sciences*, 38(2):121–126, 1952.

[76] Huang Fang, Nick Harvey, Victor Portella, and Michael Friedlander. Online mirror descent and dual averaging: keeping pace in the dynamic case. In *International Conference on Machine Learning*, pages 3008–3017, 2020.

[77] Gabriele Farina, Ioannis Anagnostides, Haipeng Luo, Chung-Wei Lee, Christian Kroer, and Tuomas Sandholm. Near-optimal no-regret learning dynamics for general convex games. In *Advances in Neural Information Processing Systems*, volume 35, pages 39076–39089, 2022.

[78] Gabriele Farina, Chung-Wei Lee, Haipeng Luo, and Christian Kroer. Kernelized multiplicative weights for 0/1-polyhedral games: Bridging the gap between learning in extensive-form and normal-form games. In *International Conference on Machine Learning*, 2022.

[79] Michail Fasoulakis, Evangelos Markakis, Yannis Pantazis, and Constantinos Varsos. Forward looking best-response multiplicative weights update methods for bilinear zero-sum games. In *International Conference on Artificial Intelligence and Statistics*, pages 11096–11117. PMLR, 2022.

[80] Genevieve E Flaspohler, Francesco Orabona, Judah Cohen, Soukayna Mouatadid, Miruna Oprescu, Paulo Orenstein, and Lester Mackey. Online learning with optimism and delay. In *International Conference on Machine Learning*, pages 3363–3373. PMLR, 2021.

[81] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

[82] Lampros Flokas, Emmanouil Vasileios Vlatakis-Gkaragkounis, and Georgios Piliouras. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. In *Advances in Neural Information Processing Systems*, 2019.

[83] Lampros Flokas, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Georgios Piliouras. Solving min-max optimization with hidden structure via gradient descent ascent. In *Advances in Neural Information Processing Systems*, volume 34, pages 2373–2386, 2021.

[84] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in neural information processing systems*, volume 29, 2016.

[85] Jakob N. Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 122–130, 2018.

[86] Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2): 379–390, 1998.

[87] Mauro Franceschelli and Paolo Frasca. Proportional dynamic consensus in open multi-agent systems. In *2018 IEEE 57th Annual Conference on Decision and Control (CDC)*, pages 900–905, 2018.

[88] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55 (1):119–139, 1997.

[89] Paul Frihauf, Miroslav Krstic, and Tamer Basar. Nash equilibrium seeking in noncooperative games. *IEEE Transactions on Automatic Control*, 57(5):1192–1207, 2011.

[90] Drew Fudenberg and Jean Tirole. *Game theory*. MIT press, 1991.

[91] Bolin Gao and Lacra Pavel. Bandit learning with regularized second-order mirror descent. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 5731–5738. IEEE, 2022.

[92] Mario Gerla, Eun-Kyu Lee, Giovanni Pau, and Uichin Lee. Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds. In *2014 IEEE world forum on internet of things (WF-IoT)*, pages 241–246. IEEE, 2014.

[93] Samuel J Gershman and Nathaniel D Daw. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual review of psychology*, 68:101–128, 2017.

[94] Franco Giannessi. On minty variational principle. *New Trends in Mathematical Programming: Homage to Steven Vajda*, pages 93–99, 1998.

[95] Angeliki Giannou, Kyriakos Lotidis, Panayotis Mertikopoulos, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. On the convergence of policy gradient methods to nash equilibria in general stochastic games. In *Advances in Neural Information Processing Systems*, volume 35, pages 7128–7141, 2022.

[96] IL Gilicksberg. A further generalization of the kakutani fixed point theorem with application to nash equilibrium points. *Proc Natl Acad Sci*, 38:170–174, 1952.

[97] Pontus Giselsson. Nonlinear forward-backward splitting with projection correction. *SIAM Journal on Optimization*, 31(3):2199–2226, 2021.

[98] Denizalp Goktas, David C Parkes, Ian Gemp, Luke Marris, Georgios Piliouras, Romuald Elie, Guy Lever, and Andrea Tacchetti. Generative adversarial equilibrium solvers. *arXiv preprint arXiv:2302.06607*, 2023.

[99] Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *Advances in Neural Information Processing Systems*, volume 33, pages 20766–20778, 2020.

[100] Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory*, pages 1758–1784. PMLR, 2020.

[101] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[102] Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient method: O (1/k) last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pages 366–402. PMLR, 2022.

[103] Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. In *Advances in Neural Information Processing Systems*, 2022.

[104] Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, pages 1–49, 2022.

[105] Eric C Hall and Rebecca M Willett. Online convex optimization in dynamic environments. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):647–662, 2015.

[106] P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Probability and Mathematical Statistics. Academic Press, New York, 1980.

[107] Nadav Hallak, Panayotis Mertikopoulos, and Volkan Cevher. Regret minimization in stochastic non-convex learning via a proximal-gradient approach. In *International Conference on Machine Learning*, pages 4008–4017. PMLR, 2021.

[108] James Hannan. Approximation to Bayes risk in repeated play. In *Contributions to the Theory of Games, Volume III*, volume 39 of *Annals of Mathematics Studies*, pages 97–139. Princeton University Press, Princeton, NJ, 1957.

[109] Keegan Harris, Ioannis Anagnostides, Gabriele Farina, Mikhail Khodak, Steven Wu, and Tuomas Sandholm. Meta-learning in games. In *The Eleventh International Conference on Learning Representations*, 2023.

[110] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

[111] Sergiu Hart and Andreu Mas-Colell. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review*, 93(5):1830–1836, 2003.

[112] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

[113] Elad Hazan and Comandur Seshadhri. Efficient learning algorithms for changing environments. In *International conference on machine learning*, pages 393–400, 2009.

[114] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[115] Elad Hazan, Karan Singh, and Cyril Zhang. Efficient regret minimization in non-convex games. In *International Conference on Machine Learning*, pages 1433–1441. PMLR, 2017.

[116] Yiran He. A new double projection algorithm for variational inequalities. *Journal of Computational and Applied Mathematics*, 185(1):166–173, 2006.

[117] Donald W Hearn. The gap function of a convex program. *Operations Research Letters*, 1(2):67–71, 1982.

[118] Amélie Héliou, Panayotis Mertikopoulos, and Zhengyuan Zhou. Gradient-free online learning in continuous games with delayed rewards. In *International conference on machine learning*, pages 4172–4181. PMLR, 2020.

[119] Julien M Hendrickx and Samuel Martin. Open multi-agent systems: Gossiping with random arrivals and departures. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 763–768. IEEE, 2017.

[120] Julien M Hendrickx and Michael G Rabbat. Stability of decentralized gradient descent in open multi-agent systems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 4885–4890. IEEE, 2020.

[121] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer, 2001.

[122] Martin Hoefer, Vahab S Mirrokni, Heiko Röglin, and Shang-Hua Teng. Competitive routing over time. *Theoretical Computer Science*, 412(39):5420–5432, 2011.

[123] Saghar Hosseini, Airlie Chapman, and Mehran Mesbahi. Online distributed optimization via dual averaging. In *2013 IEEE 52nd Annual Conference on Decision and Control (CDC)*, pages 1484–1489. IEEE, 2013.

[124] Joseph T Howson. Equilibria of polymatrix games. *Management Science*, 18 (5-part-1):312–318, 1972.

[125] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[126] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *Advances in Neural Information Processing Systems*, volume 33, pages 16223–16234, 2020.

[127] Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to nash equilibrium. In *Conference on Learning Theory*, 2021.

[128] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Optimization in open networks via dual averaging. In *2021 IEEE 60th Annual Conference on Decision and Control (CDC)*, 2021.

[129] Yu-Guan Hsieh, Kimon Antonakopoulos, Volkan Cevher, and Panayotis Mertikopoulos. No-regret learning in games with noisy feedback: Faster rates and adaptivity via learning rate separation. In *Advances in Neural Information Processing Systems*, 2022.

[130] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Multi-agent online optimization with delays: Asynchronicity, adaptivity, and optimism. *Journal of Machine Learning Research*, 2022.

[131] Ye Hu, Mingzhe Chen, Walid Saad, H Vincent Poor, and Shuguang Cui. Distributed multi-agent meta learning for trajectory design in wireless drone networks. *IEEE Journal on Selected Areas in Communications*, 39(10):3177–3192, 2021.

[132] Yuanhanqing Huang and Jianghai Hu. On the convergence rates of a nash equilibrium seeking algorithm in potential games with information delays. In *2023 American Control Conference (ACC)*, pages 1080–1085. IEEE, 2023.

[133] Yuanhanqing Huang and Jianghai Hu. A bandit learning method for continuous games under feedback delays with residual pseudo-gradient estimate. *arXiv preprint arXiv:2303.16433*, 2023.

[134] Alfredo N. Iusem, Alejandro Jofré, Roberto I. Oliveira, and Philip Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

[135] Franck Iutzeler, Pascal Bianchi, Philippe Ciblat, and Walid Hachem. Asynchronous distributed optimization using a randomized alternating direction method of multipliers. In *52nd IEEE conference on decision and control (CDC)*, pages 3671–3676. IEEE, 2013.

[136] Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag, 2nd edition, 2003.

[137] Jiyan Jiang, Wenpeng Zhang, Jinjie GU, and Wenwu Zhu. Asynchronous decentralized online learning. In *Advances in Neural Information Processing Systems*, 2021.

[138] James S Jordan. Three problems in learning mixed-strategy nash equilibria. *Games and Economic Behavior*, 5(3):368–386, 1993.

[139] Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *30th AAAI Conference on Artificial Intelligence*, 2016.

[140] Pooria Joulani, András György, and Csaba Szepesvári. A modular analysis of adaptive (non-) convex optimization: Optimism, composite objectives, and variational bounds. In *International Conference on Algorithmic Learning Theory*, pages 681–720, 2017.

[141] Pooria Joulani, András György, and Csaba Szepesvári. Think out of the "box": Generically-constrained asynchronous composite optimization and hedging. In *Advances in Neural Information Processing Systems*, 2019.

[142] Anatoli Juditsky, Arkadi Semen Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[143] Anatoli Juditsky, Joon Kwon, and Éric Moulines. Unifying mirror descent and dual averaging. *Mathematical Programming*, 199(1-2):793–830, 2023.

[144] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[145] Ehsan Asadi Kangarshahi, Ya-Ping Hsieh, Mehmet Fatih Sahin, and Volkan Cevher. Let's be honest: An optimal no-regret framework for zero-sum games. In *International Conference on Machine Learning*, pages 2488–2496, 2018.

[146] Aswin Kannan and Uday V Shanbhag. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Computational Optimization and Applications*, 74(3):779–820, 2019.

[147] S Karamardian. Complementarity problems over cones with monotone and pseudomonotone maps. *Journal of Optimization Theory and Applications*, 18(4): 445–454, 1976.

[148] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer, 2016.

[149] Ali Kavis, Kfir Y Levy, Francis Bach, and Volkan Cevher. Unixgrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[150] Mert Kayaalp, Stefan Vlaski, and Ali H Sayed. Dif-maml: Decentralized multi-agent meta-learning. *IEEE Open Journal of Signal Processing*, 3:71–93, 2022.

[151] Frank P Kelly, Aman K Maulloo, and David Kim Hong Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49(3):237–252, 1998.

[152] Rafail Z. Khasminskii. *Stochastic Stability of Differential Equations*. Number 66 in Stochastic Modelling and Applied Probability. Springer-Verlag, Berlin, 2 edition, 2012.

[153] D. Kinderlehrer and G. Stampacchia. *An Introduction to Variational Inequalities and Their Applications*. Academic Press, New York, 1980.

[154] David Kinderlehrer. Variational inequalities and free boundary problems. *Bulletin of the American Mathematical Society*, 84(1):7–26, 1978.

[155] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[156] Krzysztof C Kiwiel. Proximal minimization methods with generalized bregman functions. *SIAM journal on control and optimization*, 35(4):1142–1168, 1997.

[157] Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

[158] Igor Konnov. *Combined relaxation methods for variational inequalities*, volume 495. Springer Science & Business Media, 2001.

[159] IV Konnov, S Schaible, and JC Yao. Combined relaxation method for mixed equilibrium problems. *Journal of Optimization Theory and Applications*, 126(2), 2005.

[160] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Èkonom. i Mat. Metody*, 12:747–756, 1976.

[161] Jayash Koshal, Angelia Nedic, and Uday V Shanbhag. Regularized iterative stochastic approximation methods for stochastic variational inequality problems. *IEEE Transactions on Automatic Control*, 58(3):594–609, 2012.

[162] Walid Krichene, Maximilian Balandat, Claire Tomlin, and Alexandre Bayen. The hedge algorithm on a continuum. In *International Conference on Machine Learning*, pages 824–832. PMLR, 2015.

[163] Joon Kwon and Panayotis Mertikopoulos. A continuous-time approach to online optimization. *Journal of Dynamics and Games*, 4(2):125–148, 2017.

[164] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference (ECC)*, pages 243–248. IEEE, 2016.

[165] John Langford, Alexander J Smola, and Martin Zinkevich. Slow learners are fast. In *Advances in Neural Information Processing Systems*, pages 2331–2339, 2009.

[166] Rida Laraki, Jérôme Renault, and Sylvain Sorin. *Mathematical foundations of game theory*. Springer, 2019.

[167] Torbjörn Larsson and Michael Patriksson. A class of gap functions for variational inequalities. *Mathematical Programming*, 64:53–79, 1994.

[168] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[169] Chung-Wei Lee, Christian Kroer, and Haipeng Luo. Last-iterate convergence in extensive-form games. In *Advances in Neural Information Processing Systems*, volume 34, pages 14293–14305, 2021.

[170] Jae Won Lee, Jonghun Park, O Jangmin, Jongwoo Lee, and Euyseok Hong. A multiagent approach to q-learning for daily stock trading. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(6):864–877, 2007.

[171] Sucheol Lee and Donghwan Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[172] D Leventhal. Metric subregularity and the proximal point method. *Journal of Mathematical Analysis and Applications*, 360(2):681–688, 2009.

[173] Shuangtong Li, Tianyi Zhou, Xinmei Tian, and Dacheng Tao. Learning to collaborate in decentralized learning of personalized models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9766–9775, 2022.

[174] Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *International Conference on Artificial Intelligence and Statistics*, pages 983–992. PMLR, 2019.

[175] Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Convergence rates with inexact non-expansive operators. *Mathematical Programming*, 159:403–434, 2016.

[176] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

[177] Tianyi Lin, Zhengyuan Zhou, Panayotis Mertikopoulos, and Michael I Jordan. Finite-time last-iterate convergence for multi-agent learning in games. In *International Conference on Machine Learning*, pages 6161–6171, 2020.

[178] Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

[179] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[180] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[181] Christopher Lu, Timon Willi, Christian A Schroeder De Witt, and Jakob Foerster. Model-free opponent shaping. In *International Conference on Machine Learning*, pages 14398–14411. PMLR, 2022.

[182] Haihao Lu. An o (sr)-resolution ode framework for understanding discrete-time algorithms and applications to the linear convergence of minimax problems. *Mathematical Programming*, 194(1-2):1061–1112, 2022.

[183] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[184] Yura Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, pages 1–28, 2019.

[185] Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229, 2017.

[186] Yuyi Mao, Changsheng You, Jun Zhang, Kaibin Huang, and Khaled B Letaief. A survey on mobile edge computing: The communication perspective. *IEEE communications surveys & tutorials*, 19(4):2322–2358, 2017.

[187] David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[188] Brendan McMahan and Matthew Streeter. Delay-tolerant algorithms for asynchronous distributed online learning. In *Advances in Neural Information Processing Systems*, pages 2915–2923, 2014.

[189] H Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18(1):3117–3166, 2017.

[190] Jean-François Mertens, Sylvain Sorin, and Shmuel Zamir. *Repeated games*, volume 55. Cambridge University Press, 2015.

[191] Panayotis Mertikopoulos. *Online Optimization and Learning in Games: Theory and Applications*. Habilitation à Diriger des Recherches (HDR), Université Grenoble-Alpes, December 2019.

[192] Panayotis Mertikopoulos and William H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, November 2016.

[193] Panayotis Mertikopoulos and Mathias Staudigl. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization*, 28(1):163–197, January 2018.

[194] Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173 (1-2):465–507, January 2019.

[195] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717. SIAM, 2018.

[196] Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference on Learning Representations*, 2019.

[197] George J Minty. On the generalization of a direct method of the calculus of variations. *Bulletin of the American Mathematical Society*, 73(3):314–321, 1967.

[198] Mehryar Mohri and Scott Yang. Accelerating online convex optimization via adaptive prediction. In *International Conference on Artificial Intelligence and Statistics*, pages 848–856, 2016.

[199] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

[200] Daniel K Molzahn, Florian Dörfler, Henrik Sandberg, Steven H Low, Sambuddha Chakrabarti, Ross Baldick, and Javad Lavaei. A survey of distributed optimization and control algorithms for electric power systems. *IEEE Transactions on Smart Grid*, 8(6):2941–2962, 2017.

[201] Vaishnavh Nagarajan and J Zico Kolter. Gradient descent gan optimization is locally stable. In *Advances in Neural Information Processing Systems*, pages 5585–5595, 2017.

[202] Eduardo F Nakamura, Antonio AF Loureiro, and Alejandro C Frery. Information fusion for wireless sensor networks: Methods, models, and classifications. *ACM Computing Surveys (CSUR)*, 39(3):9–es, 2007.

[203] Shravan Narayanamurthy, Markus Weimer, Dhruv Mahajan, Tyson Condie, and Sundararajan Sellamanickam. Towards resource-elastic machine learning. In *NIPS 2013 BigLearn Workshop*, 2013.

[204] John Nash. Non-cooperative games. *Annals of mathematics*, pages 286–295, 1951.

[205] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

[206] Angelia Nedić, Alex Olshevsky, and Michael G Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.

[207] Arkadi Semen 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.

[208] Arkadi Semen Nemirovski and David Berkovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, NY, 1983.

[209] Arkadi Semen Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[210] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Number 87 in Applied Optimization. Kluwer Academic Publishers, 2004.

[211] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

[212] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[213] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

[214] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. *Algorithmic game theory*. Cambridge university press, 2007.

[215] Zdzisław Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bulletin of the American Mathematical Society*, 73(4): 591–597, 1967.

[216] Francesco Orabona and Dávid Pál. Scale-free online learning. *Theoretical Computer Science*, 716:50–69, 2018.

[217] Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994.

[218] Gerasimos Palaiopanos, Ioannis Panageas, and Georgios Piliouras. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. In *Advances in Neural Information Processing Systems*, 2017.

[219] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

[220] Ronald Parr and Stuart Russell. Approximating optimal policies for partially observable stochastic domains. In *IJCAI*, volume 95, pages 1088–1094, 1995.

[221] Thomas Pethick, Puya Latafat, Panagiotis Patrinos, Olivier Fercoq, and Volkan Cevhera. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *International Conference on Learning Representations*, 2022.

[222] Thomas Pethick, Olivier Fercoq, Puya Latafat, Panagiotis Patrinos, and Volkan Cevher. Solving stochastic weak minty variational inequalities without increasing batch size. In *International Conference on Learning Representations*, 2023.

[223] Georgios Piliouras, Lillian Ratliff, Ryann Sim, and Stratis Skoulakis. Fast convergence of optimistic gradient ascent in network zero-sum extensive form games. In *Algorithmic Game Theory: 15th International Symposium, SAGT 2022, Colchester, UK, September 12–15, 2022, Proceedings*, pages 383–399. Springer, 2022.

[224] Boris Teodorovich Polyak. *Introduction to Optimization*. Optimization Software, New York, NY, USA, 1987.

[225] Leonid Denisovich 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.

[226] Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. In *Advances in Neural Information Processing Systems*, pages 1270–1278, 2015.

[227] Michael Rabbat and Robert Nowak. Distributed optimization in sensor networks. In *International Symposium on Information Processing in Sensor Networks*, pages 20–27, 2004.

[228] Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, 2013.

[229] Lillian J Ratliff, Samuel A Burden, and S Shankar Sastry. Characterization and computation of local nash equilibria in continuous games. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 917–924. IEEE, 2013.

[230] Usman Raza, Alessandro Camerra, Amy L Murphy, Themis Palpanas, and Gian Pietro Picco. Practical data prediction for real-world wireless sensor networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(8):2231–2244, 2015.

[231] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in neural information processing systems*, volume 24, 2011.

[232] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[233] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[234] Herbert Robbins and David Sigmund. A convergence theorem for nonnegative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Academic Press, New York, NY, 1971.

[235] J. B. Rosen. Existence and uniqueness of equilibrium points for concave N-person games. *Econometrica*, 33(3):520–534, 1965.

[236] Tim Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM (JACM)*, 62(5):1–42, 2015.

[237] Aviad Rubinstein. Inapproximability of nash equilibrium. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 409–418, 2015.

[238] Farzad Salehisadaghiani and Lacra Pavel. Distributed nash equilibrium seeking: A gossip-based algorithm. *Automatica*, 72:209–216, 2016.

[239] William H Sandholm. *Population games and evolutionary dynamics*. MIT press, 2010.

[240] Guillaume Sartoretti, Yue Wu, William Paivine, TK Satish Kumar, Sven Koenig, and Howie Choset. Distributed reinforcement learning for multi-robot decentralized collective construction. In *Distributed Autonomous Robotic Systems: The 14th International Symposium*, pages 35–49. Springer, 2019.

[241] Jean-Paul Sartre. *Huis clos*. Gallimard, 1945.

[242] Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Basar, and Asuman Ozdaglar. Decentralized q-learning in zero-sum markov games. In *Advances in Neural Information Processing Systems*, volume 34, pages 18320–18334, 2021.

[243] Gesualdo Scutari, Daniel P Palomar, Francisco Facchinei, and Jong-Shi Pang. Convex optimization, game theory, and variational inequality theory. *IEEE Signal Processing Magazine*, 27(3):35–49, 2010.

[244] Shahin Shahrampour and Ali Jadbabaie. Distributed online optimization in dynamic environments using mirror descent. *IEEE Transactions on Automatic Control*, 63(3):714–725, 2017.

[245] Shai Shalev-Shwartz. *Online learning: Theory, algorithms, and applications*. PhD thesis, Hebrew University of Jerusalem, 2007.

[246] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[247] Shai Shalev-Shwartz and Yoram Singer. Convex repeated games and Fenchel duality. In *Advances in Neural Information Processing Systems*, pages 1265–1272. MIT Press, 2006.

[248] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

[249] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39 (10):1095–1100, 1953.

[250] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018.

[251] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge computing: Vision and challenges. *IEEE internet of things journal*, 3(5):637–646, 2016.

[252] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[253] Mikhail V Solodov. Convergence rate analysis of iteractive algorithms for solving variational inequality problems. *Mathematical Programming*, 96(3):513–528, 2003.

[254] Mikhail V Solodov and Benar Fux Svaiter. A new projection method for variational inequality problems. *SIAM Journal on Control and Optimization*, 37(3):765–776, 1999.

[255] Chaobing Song, Zhengyuan Zhou, Yichao Zhou, Yong Jiang, and Yi Ma. Optimistic dual extrapolation for coherent non-monotone variational inequalities. In *Advances in Neural Information Processing Systems*, volume 33, pages 14303–14314, 2020.

[256] Guido Stampacchia. Formes bilineaires coercitives sur les ensembles convexes. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 258:4413–4416, 1964.

[257] Miloš S Stanković, Karl Henrik Johansson, and Dušan M Stipanović. Distributed seeking of nash equilibria in mobile sensor networks. In *49th IEEE Conference on Decision and Control (CDC)*, pages 5598–5603. IEEE, 2010.

[258] Arun Sai Suggala and Praneeth Netrapalli. Online non-convex learning: Following the perturbed leader is optimal. In *International Conference on Algorithmic Learning Theory*, pages 845–861. PMLR, 2020.

[259] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, volume 28, pages 2989–2997, 2015.

[260] Tatiana Tatarenko and Maryam Kamgarpour. Learning generalized nash equilibria in a class of convex games. *IEEE Transactions on Automatic Control*, 64(4):1426–1439, 2018.

[261] Tatiana Tatarenko and Maryam Kamgarpour. Bandit learning in convex non-strictly monotone games. *arXiv preprint arXiv:2009.04258*, 2020.

[262] Tatiana Tatarenko, Wei Shi, and Angelia Nedić. Accelerated gradient play algorithm for distributed nash equilibrium seeking. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 3561–3566. IEEE, 2018.

[263] Sebastian Thrun. Probabilistic robotics. *Communications of the ACM*, 45(3):52–57, 2002.

[264] Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

[265] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

[266] Predrag T Tošić and Ricardo Vilalta. A unified framework for reinforcement learning, co-learning and meta-learning how to coordinate in collaborative multi-agent systems. *Procedia Computer Science*, 1(1):2217–2226, 2010.

[267] Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2): 237–252, June 1995.

[268] John Tsitsiklis, Dimitri Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE transactions on automatic control*, 31(9):803–812, 1986.

[269] John N Tsitsiklis. *Problems in decentralized decision making and computation*. PhD thesis, Massachusetts Institute of Technology, 1984.

[270] Vineeth S Varma, Irinel-Constantin Morărescu, and Dragan Nešić. Open multi-agent systems with discrete states and stochastic interactions. *IEEE Control Systems Letters*, 2(3):375–380, 2018.

[271] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[272] Yannick Viossat and Andriy Zapechelnyuk. No-regret dynamics and fictitious play. *Journal of Economic Theory*, 148(2):825–842, March 2013.

[273] Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Thanasis Lianeas, Panayotis Mertikopoulos, and Georgios Piliouras. No-regret learning and mixed nash equilibria: They do not mix. In *Advances in Neural Information Processing Systems*, volume 33, pages 1380–1391, 2020.

[274] Xiaofei Wang, Yiwen Han, Victor CM Leung, Dusit Niyato, Xueqiang Yan, and Xu Chen. Convergence of edge computing and deep learning: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(2):869–904, 2020.

[275] Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *The Journal of Machine Learning Research*, 21(1):9047–9076, 2020.

[276] Madanlal Tilakchand Wasan. *Stochastic approximation*. Cambridge University Press, 2004.

[277] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Conference on learning theory*, pages 4259–4299. PMLR, 2021.

[278] Jörgen W Weibull. *Evolutionary game theory*. MIT press, 1997.

[279] Marcelo J Weinberger and Erik Ordentlich. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, 2002.

[280] Qingyun Wu, Hongning Wang, Liangjie Hong, and Yue Shi. Returning is believing: Optimizing long-term user engagement in recommender systems. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1927–1936, 2017.

[281] Tianyu Wu, Kun Yuan, Qing Ling, Wotao Yin, and Ali H Sayed. Decentralized consensus optimization with asynchrony and delays. *IEEE Transactions on Signal and Information Processing over Networks*, 4(2):293–307, 2017.

[282] Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, October 2010.

[283] Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pages 3674–3682. PMLR, 2020.

[284] Feng Yan, Shreyas Sundaram, SVN Vishwanathan, and Yuan Qi. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2483–2493, 2012.

[285] Tao Yang, Xinlei Yi, Junfeng Wu, Ye Yuan, Di Wu, Ziyang Meng, Yiguang Hong, Hong Wang, Zongli Lin, and Karl H Johansson. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019.

[286] Maojiao Ye and Guoqiang Hu. Distributed nash equilibrium seeking by a consensus based approach. *IEEE Transactions on Automatic Control*, 62(9):4811–4818, 2017.

[287] Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[288] Constantin Zalinescu. *Convex analysis in general vector spaces*. World scientific, 2002.

[289] Guojun Zhang and Yaoliang Yu. Convergence of gradient methods on bilinear zero-sum games. In *International Conference on Learning Representations*, 2020.

[290] Hui Zhang. New analysis of linear convergence of gradient-type methods via unifying error bound conditions. *Mathematical Programming*, 180(1-2):371–416, 2020.

[291] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, pages 5872–5881. PMLR, 2018.

[292] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.

[293] Mengxiao Zhang, Peng Zhao, Haipeng Luo, and Zhi-Hua Zhou. No-regret learning in time-varying zero-sum games. In *International Conference on Machine Learning*, pages 26772–26808. PMLR, 2022.

[294] Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. In *OPT2020: 12th Annual Workshop on Optimization for Machine Learning*, 2020.

[295] Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Peter W Glynn, and Claire Tomlin. Countering feedback delays in multi-agent learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[296] Cai-Nicolas Ziegler. *Towards decentralized recommender systems*. PhD thesis, Verlag nicht ermittelbar, 2005.

[297] Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[298] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, pages 928–936, 2003.

[299] SI Zuhovickii, RA Poljak, and M Eo Primak. Numerical methods of finding equilibrium points of n-person games. *Proceedings of the First Winter School of Mathematical Programming in Drogobych*, pages 93–130, 1969.

[300] Semen Izrailevich Zukhovitskii, Roman A Polyak, and ME Primak. Two methods of finding equilibrium points of concave n-person games. In *Doklady Akademii Nauk*, volume 185, pages 24–27. Russian Academy of Sciences, 1969.

APPENDIX

A

BREGMAN DIVERGENCES, MIRROR MAPS, AND FENCHEL COUPLINGS

In this appendix, we present several basic properties of the Bregman divergence, the mirror map, and the Fenchel coupling. We follow the notations of Chapter 2. The learner's action set and the associated regularizer are thus respectively $\mathcal{X}$ and h. Moreover, the regularizer h is assumed to be 1-strongly convex relative to an ambient norm $\|\cdot\|$. Let us also recall the definition of the Bregman divergence, the Fenchel coupling, and the mirror map

$$\begin{aligned} D(z,x) &= h(z) - h(x) - \langle \nabla h(x), z - x \rangle, \\ F(z,y) &= h(z) + h^*(y) - \langle y, z \rangle, \\ Q(y) &= \arg\min_{x \in \mathcal{X}} \langle -y, x \rangle + h(x). \end{aligned}$$

The auxiliary results that we are going to present below concerning these three quantities are not new (see e.g., [142, 194, 209] and references therein); however, the set of hypotheses used to obtain them varies widely in the literature, so we still provide the proofs for the sake of completeness.

To begin, our first lemma concerns the optimality condition of the mirror map. This result is widely used for the analysis of MD-type methods (see e.g., Propositions 2.2 and 6.3).

Lemma A.1. *Let h be a regularizer on $\mathcal{X}$. Then, for all $x \in \operatorname{dom} \partial h$ and all $y \in \mathbb{R}^d$, we have*

$$x = Q(y) \iff y \in \partial h(x).$$

Moreover, if $x = Q(y)$, it holds for all $z \in \mathcal{X}$ that

$$\langle \nabla h(x), x - z \rangle \leq \langle y, x - z \rangle.$$

Proof. For the first claim, we have by the definition of the mirror map $x = Q(y)$ if and only if $0 \in \partial h(x) - y$, i.e., $y \in \partial h(x)$. For the second claim, it suffices to show it holds for all $p \in \operatorname{ri} \mathcal{X}$ (by continuity). To do so, we can define

$$\phi(t) = h(x + t(p - x)) - [h(x) + \langle y, x + t(p - x) \rangle].$$

Since h is strongly convex and $y \in \partial h(x)$ by the previous claim, it follows that $\phi(t) \geq 0$ with equality if and only if $t = 0$. Moreover, as $\operatorname{ri} \mathcal{X} \subset \operatorname{dom} \partial h$, $\nabla h(x + t(p - x))$ is well-defined and $\psi(t) = \langle \nabla h(x + t(p - x)) - y, p - x \rangle$ is a continuous selection of subgradients of ϕ. Given that ϕ and ψ are both continuous on $[0,1]$, it follows that ϕ is continuously differentiable and $\phi' = \psi$ on $[0,1]$. Thus, with $\phi(t) \geq 0 = \phi(0)$ for all $t \in [0,1]$, we conclude that $\phi'(0) = \langle \nabla h(x) - y, p - x \rangle \geq 0$, from which our claim follows. □

We continue with the "three-point identities" which are used in Chapters 2 and 6 to derive the recurrent relationship between the divergence measures of different steps.

Lemma A.2. *Let h be a regularizer on $\mathcal{X}$. Then, for all $z \in \mathcal{X}$ and all $x, x' \in \operatorname{dom} \partial h$, we have*

$$\langle \nabla h(x') - \nabla h(x), x - z \rangle = D(z, x') - D(z, x) - D(x, x'). \tag{A.1}$$

Similarly, writing $x = Q(y)$, for all $z \in \mathcal{X}$ and all $y, y' \in \mathbb{R}^d$, we have

$$\langle y' - y, x - z \rangle = F(z, y') - F(z, y) - F(x, y'). \tag{A.2}$$

Proof. We start with the Bregman version. By definition,

$$\begin{aligned} D(p, x') &= h(p) - h(x') - \langle \nabla h(x'), p - x' \rangle \\ D(p, x) &= h(p) - h(x) - \langle \nabla h(x), p - x \rangle \\ D(x, x') &= h(x) - h(x') - \langle \nabla h(x'), x - x' \rangle. \end{aligned}$$

The result then follows by adding the two last lines and subtracting the first. On the other hand, in order to show the Fenchel coupling version we write

$$\begin{aligned} F(p, y') &= h(p) + h^*(y') - \langle y', p \rangle \\ F(p, y) &= h(p) + h^*(y) - \langle y, p \rangle. \end{aligned}$$

Then, by subtracting the above we obtain

$$\begin{aligned} F(p, y') - F(p, y) &= h(p) + h^*(y') - \langle y', p \rangle - h(p) - h^*(y) + \langle y, p \rangle \\ &= h^*(y') - h^*(y) - \langle y' - y, p \rangle \\ &= h^*(y') - \langle y, Q(y) \rangle + h(Q(y)) - \langle y' - y, p \rangle \\ &= h^*(y') - \langle y, x \rangle + h(x) - \langle y' - y, p \rangle \\ &= h^*(y') + \langle y' - y, x \rangle - \langle y', x \rangle + h(x) - \langle y' - y, p \rangle \\ &= F(x, y') + \langle y' - y, x - p \rangle \end{aligned}$$

and our proof is complete. □

Since $x = Q(\nabla h(x))$ and $F(z, \nabla h(x)) = D(z, x)$, the identity (A.1) is indeed a special case of (A.2). In the general case, the Fenchel coupling and the Bregman divergence can be related by the following lemma.

Lemma A.3. *Let h be a regularizer on $\mathcal{Z}$. Then, for all $z \in \mathcal{Z}$ and $y \in \mathbb{R}^d$, it holds*

$$F(z, y) \geq D(z, Q(y)) \geq \frac{\|z - Q(y)\|^2}{2}.$$

Proof. For the first inequality we have,

$$\begin{aligned} F(z, y) &= h(z) + h^*(y) - \langle y, z \rangle \\ &= h(z) - h(Q(y)) + \langle y, Q(y) \rangle + \langle y, -z \rangle \\ &= h(z) - h(Q(y)) - \langle y, z - Q(y) \rangle \end{aligned}$$

Since $y \in \partial h(Q(y))$, by Lemma A.1 we get

$$\langle \nabla h(Q(y)), Q(y) - z \rangle \leq \langle y, Q(y) - z \rangle$$

With all the above we then have

$$\begin{aligned} F(z,y) &= h(z) - h(Q(y)) - \langle y, z - Q(y)\rangle \\ &\geq h(z) - h(Q(y)) - \langle \nabla h(Q(y)), z - Q(y)\rangle \\ &= D(z, Q(y)) \end{aligned}$$

and the result follows. The second inequality follows directly from the fact that the regularizer h is 1-strongly convex relative to $\|\cdot\|$. □

We next prove the non-expansiveness of the mirror map which are used multiple times in the analysis of Chapter 3. (for a reference, see e.g., [121, Chapter E, Thm. 4.2.1], or [288, Cor. 3.5.11]).

Lemma A.4. *The mirror map is non-expansive, i.e.,* $\|P(y) - P(y')\| \leq \|y - y'\|_*$ *for all* $y, y' \in \mathbb{R}^d$.

Proof. Let $x = P(y)$ and $x' = P(y')$. By definition of the mirror map,

$$x = \arg\min_{\hat{x}\in\mathcal{X}} \langle -y, \hat{x}\rangle + h(\hat{x}), \quad x' = \arg\min_{\hat{x}\in\mathcal{X}} \langle -y', \hat{x}\rangle + h(\hat{x}).$$

The optimality condition implies that $y \in \partial h(x)$ and $y' \in \partial h(x')$. Hence, with the Cauchy–Schwarz inequality and the 1-strong convexity of h with respect to $\|\cdot\|$, we have

$$\|y - y'\|_*\|x' - x\| \geq \langle y' - y, x' - x\rangle \geq \|x - x'\|^2.$$

It follows immediately $\|y - y'\|_* \geq \|x - x'\|$. □

Remark A.1. Precisely, P is non-expansive because we are assuming that the strong convexity constant of h is 1. Otherwise it would just be Lipschitz continuous, and clearly this would only influence our results by a constant factor (that depends on the strong convexity constant of h).

We end up with a simple result showing that Bregman reciprocity (Definition 6.2) is implied by Fenchel reciprocity (Definition 6.3). This is used in Chapter 6 to advocate the use of (DS-OptMD) in the place of (OptDA).

Lemma A.5. *Let* $\mathcal{X}$ *be the action set and* h *be a regularizer over* $\mathcal{X}$. *If Fenchel reciprocity is satisfied, then Bregman reciprocity is also satisfied.*

Proof. We have $x = Q(\nabla h(x))$. Hence

$$F(z, \nabla h(x)) = h(z) - h(Q(\nabla h(x))) - \langle \nabla h(x), z - Q(\nabla h(x))\rangle = D(z, x).$$

Consider $(X_t)_{t\in\mathbb{N}}$ a sequence of points of $\mathcal{X}$ such that $X_t \to z$. This means $Q(Y_t) \to z$ for $Y_t = \nabla h(X_t)$ and accordingly $F(z, Y_t) = D(z, X_t)$ tends to 0. □

B

TECHNICAL LEMMAS ON NUMERICAL AND STOCHASTIC SEQUENCES

This appendix collects a series of fundamental lemmas regarding numerical and stochastic sequences. These lemmas, while crucial for the proofs developed throughout this thesis, are presented separately here due to their general applicability. The aim is to prevent the interruption of the main flow of the thesis with these technical details. Some of these lemmas might be straightforward, while others might require a deeper understanding of the concepts involved.

CONVERGENCE OF QUASI-DECREASING SEQUENCE. The following lemma, used in the proof of Lemma 6.11, states that a "quasi-decreasing" sequence converges. This kind of result is useful when dealing with quasi-Fejér monotone sequences.

Lemma B.1. *Let $(U_t)_{t\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}}$ be a non-negative sequence and $(\chi_t)_{t\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}}$ be summable such that, for all $t \in \mathbb{N}$,*

$$U_{t+1} \leq U_t + \chi_t. \tag{B.1}$$

Then, $(U_t)_{t\in\mathbb{N}}$ converges.

Proof. Since $(\chi_t)_{t\in\mathbb{N}}$ is summable, we can define $U'_t = U_t + \sum_{s=t}^{+\infty} \chi_s \in \mathbb{R}_+$. Inequality (B.1) then implies $U'_{t+1} \leq U'_t$. Therefore, $(U'_t)_{t\in\mathbb{N}}$ converges, and accordingly $(U_t)_{t\in\mathbb{N}}$ converges. □

Remark B.1. We can also derive this lemma from the Robbins–Siegmund theorem by taking only the trivial σ-algebra but this is clearly using a sledgehammer to crack a nut.

CONVERGENCE RATE DERIVATION. In the optimization literature, one can find a myriad of lemmas designed for the purpose of establishing convergence rates associated with specific (predetermined) learning rates. The crafting of these learning rates, with the aim of achieving certain convergence rates, is indeed a discipline in itself. Below we present two such lemmas that we exploit in Chapter 7 to prove convergence rates under error bound conditions. The reader is referred to the work of Polyak [224] for a collection of results of this type.

Lemma B.2. *Let $(a_t)_{t\in\mathbb{N}}$ be a sequence of real numbers such that for all t,*

$$a_{t+1} \leq (1-c)a_t + c',$$

where $1 > c > 0$ and $c' > 0$. Then,

$$a_t \leq (1-c)^{t-1}a_1 + \frac{c'}{c}.$$

The above lemma comes into play when an algorithm is run with constant learning rate sequences, whereas we resort to the following two lemmas in case of decreasing learning rate sequences.

Lemma B.3 (Chung [47, Lemma 1]). *Let $(a_t)_{t\in\mathbb{N}}$ be a sequence of real numbers and $\beta \in \mathbb{N}$ such that for all t,*

$$a_{t+1} \leq \left(1 - \frac{c}{t+\beta}\right) a_t + \frac{c'}{(t+\beta)^{r+1}},$$

where $c > r > 0$ and $c' > 0$. Then,

$$a_t \leq \frac{c'}{c-r}\frac{1}{t^r} + o\left(\frac{1}{t^r}\right).$$

Proof. See Chung [47, Lemma 1]. □

LEMMAS ON STOCHASTIC SEQUENCES. We now shift our focus to several lemmas on stochastic sequences that are extensively used in Chapters 7 and 8. The first one translates a bound of expectation into almost sure boundedness. It is a special case of Doob's martingale convergence theorem [106], but we also provide another elementary proof below. Note that we use the term finite random variable to refer to those random variables which are finite almost surely.

Lemma B.4. *Let $(U_t)_{t\in\mathbb{N}}$ be a sequence of non-decreasing and non-negative real-valued random variables. If there exists constant $C \in \mathbb{R}$ such that*

$$\forall\, t \in \mathbb{N}, \quad \mathbb{E}[U_t] \leq C.$$

Then $(U_t)_{t\in\mathbb{N}}$ converges almost surely to a finite random variable. In particular, for any sequence of non-negative real-valued random variables $(\chi_t)_{t\in\mathbb{N}}$, the fact that $\sum_{t=1}^{+\infty} \mathbb{E}[\chi_t] < +\infty$ implies $\sum_{t=1}^{+\infty} \chi_t < +\infty$ almost surely, and accordingly $\lim_{t\to+\infty} \chi_t = 0$ almost surely.

Proof. Let U_∞ be the pointwise limit of $(U_t)_{t\in\mathbb{N}}$. Applying Beppo Levi's lemma we deduce that U_∞ is also measurable and $\lim_{t\to+\infty} \mathbb{E}[U_t] = \mathbb{E}[U_\infty]$. Accordingly, $\mathbb{E}[U_\infty] \leq C$. The random variable U_∞ being non-negative, $\mathbb{E}[U_\infty] \leq C < +\infty$ implies that U_∞ is finite almost surely, which concludes the first statement of the lemma. The second statement is derived from the first statement by setting $U_t = \sum_{s=1}^{t} \chi_s$. □

The next lemma is essential for building almost sure last-iterate convergence, as it allows to extract a convergent subsequence.

Lemma B.5. *Let $(U_t)_{t\in\mathbb{N}}$ be a sequence of non-negative real-valued random variables such that*

$$\liminf_{t\to+\infty} \mathbb{E}[U_t] = 0.$$

Then, (i) there exists a subsequence $(U_{\omega(t)})_{t\in\mathbb{N}}$ of $(U_t)_{t\in\mathbb{N}}$ that converges to 0 almost surely;[1] and accordingly (ii) it holds almost surely that $\liminf_{t\to+\infty} U_t = 0$.

Proof. Since $\liminf_{t\to+\infty} \mathbb{E}[U_t] = 0$, we can extract a subsequence $(U_{\omega(t)})_{t\in\mathbb{N}}$ such that for all $t \in \mathbb{N}$, $\mathbb{E}[U_{\omega(t)}] \leq 2^{-t}$. This gives $\sum_{t=1}^{+\infty} \mathbb{E}[U_{\omega(t)}] < +\infty$

1 We remark that the choice of the subsequence does not depend on the realization but only the distribution of the random variables.

and invoking Lemma B.4 we then know that $\sum_{t=1}^{+\infty} U_{\omega(t)} < +\infty$ almost surely, which in turn implies that $U_{\omega(t)}$ converges to 0 almost surely. To prove (*ii*), we just notice that for any realization such that $\lim_{t\to+\infty} U_{\omega(t)} = 0$, we have $0 = \lim_{t\to+\infty} U_{\omega(t)} \geq \liminf_{t\to+\infty} U_t \geq 0$ and thus the equalities must hold, i.e., $\liminf_{t\to+\infty} U_t = 0$. □

Finally, since the solution may not be unique, we need a to translate a result with respect to a single point to one that applies to the entire set. This is achieved through the following lemma (we reuse the notation from Chapter 8 for vector in dimension d split in N components and its weighted norm with respect to a vector of dimension N).

Lemma B.6. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a closed set, $(\mathbf{u}_t)_{t\in\mathbb{N}}$ be a sequence of $\mathbb{R}^d$-valued random variable, and $(\boldsymbol{\alpha}_t)_{t\in\mathbb{N}}$ be a sequence of $\mathbb{R}^N$-valued random variable such that*

(a) For all $i \in \mathcal{N}$, $\alpha_1^i \geq 1$, $(\alpha_t^i)_{t\in\mathbb{N}}$ is non-decreasing and converges to a finite constant almost surely.

(b) For all $\mathbf{x} \in \mathcal{K}$, $\|\mathbf{u}_t - \mathbf{x}\|_{\boldsymbol{\alpha}_t}$ converges almost surely.

Then, with probability 1, the vector $\boldsymbol{\alpha}_\infty = \lim_{t\to+\infty} \boldsymbol{\alpha}_t$ is well-defined, finite, and $\|\mathbf{u}_t - \mathbf{x}\|_{\boldsymbol{\alpha}_\infty}$ converges for all $\mathbf{x} \in \mathcal{K}$.

Proof. As $\mathbb{R}^d$ is a separable metric space, $\mathcal{K}$ is also separable and we can find a countable set $\mathcal{Z}$ such that $\mathcal{K} = \mathrm{cl}(\mathcal{Z})$. Let us define the event

$$\mathcal{E} \coloneqq \{\boldsymbol{\alpha}_\infty = \lim_{t\to+\infty} \boldsymbol{\alpha}_t \text{ is well-defined and finite;} \\ \|\mathbf{u}_t - \mathbf{z}\|_{\boldsymbol{\alpha}_t} \text{ converges for all } \mathbf{z} \in \mathcal{Z}.\}$$

The set $\mathcal{Z}$ being countable, from (a) and (b) we then know that $\mathbb{P}(\mathcal{E}) = 1$. In the following, we show that $\|\mathbf{u}_t - \mathbf{x}\|_{\boldsymbol{\alpha}_\infty}$ converges for all $\mathbf{x} \in \mathcal{K}$ whenever $\mathcal{E}$ happens, which concludes our proof.

Let us now consider a realization of $\mathcal{E}$. We first establish the convergence of $\|\mathbf{u}_t - \mathbf{z}\|_{\boldsymbol{\alpha}_\infty}$ for any $\mathbf{z} \in \mathcal{Z}$. To begin, the convergence of $\|\mathbf{u}_t - \mathbf{z}\|_{\boldsymbol{\alpha}_t}$ implies the boundedness of this sequence, from which we deduce immediately the boundedness of $\|\mathbf{u}_t - \mathbf{z}\|$ as $\|\mathbf{u}_t - \mathbf{z}\| \leq \|\mathbf{u}_t - \mathbf{z}\|_{\boldsymbol{\alpha}_t}$ by $\boldsymbol{\alpha}_t \geq \boldsymbol{\alpha}_1 \geq 1$. In other words, $C = \sup_{t\in\mathbb{N}} \|\mathbf{u}_t - \mathbf{z}\|$ is finite. Furthermore, we have

$$0 \leq \|\mathbf{u}_t - \mathbf{z}\|^2_{\boldsymbol{\alpha}_\infty} - \|\mathbf{u}_t - \mathbf{z}\|^2_{\boldsymbol{\alpha}_t} = \sum_{i=1}^{N} (\alpha_\infty^i - \alpha_t^i)\|u_t^i - z^i\|^2 \leq \sum_{i=1}^{N} (\alpha_\infty^i - \alpha_t^i)C^2. \quad \text{(B.2)}$$

Since $\alpha_\infty^i - \alpha_t^i$ converges to 0 when t goes to infinity, from (B.2) we get immediately $\lim_{t\to+\infty}(\|\mathbf{u}_t - \mathbf{z}\|^2_{\boldsymbol{\alpha}_\infty} - \|\mathbf{u}_t - \mathbf{z}\|^2_{\boldsymbol{\alpha}_t}) = 0$. This shows that $\|\mathbf{u}_t - \mathbf{z}\|^2_{\boldsymbol{\alpha}_\infty}$ converges to $\lim_{t\to+\infty} \|\mathbf{u}_t - \mathbf{z}\|^2_{\boldsymbol{\alpha}_t}$, which exists by definition of $\mathcal{E}$. We have thus shown the convergence of $\|\mathbf{u}_t - \mathbf{z}\|_{\boldsymbol{\alpha}_\infty}$.

To conclude, we need to show that $\|\mathbf{u}_t - \mathbf{x}\|_{\boldsymbol{\alpha}_\infty}$ in fact converges for all $\mathbf{x} \in \mathcal{K}$. Let $\mathbf{x} \in \mathcal{K}$. As $\mathcal{Z}$ is dense in $\mathcal{K}$, there exists a sequence of points $(\mathbf{z}_k)_{k\in\mathbb{N}}$ with $\mathbf{z}_k \in \mathcal{Z}$ for all $k \in \mathbb{N}$ such that $\lim_{k\to+\infty} \mathbf{z}_k = \mathbf{x}$. For any $t, k \in \mathbb{N}$, the triangular inequality implies

$$-\|\mathbf{z}_k - \mathbf{x}\|_{\boldsymbol{\alpha}_\infty} \leq \|\mathbf{u}_t - \mathbf{x}\|_{\boldsymbol{\alpha}_\infty} - \|\mathbf{u}_t - \mathbf{z}_k\|_{\boldsymbol{\alpha}_\infty} \leq \|\mathbf{z}_k - \mathbf{x}\|_{\boldsymbol{\alpha}_\infty}.$$

Since $\mathbf{z}_k \in \mathcal{Z}$, we have shown that $\lim_{t\to+\infty} \|\mathbf{u}_t - \mathbf{z}_k\|$ exists. Subsequently, we get

$$-\|\mathbf{z}_k - \mathbf{x}\|_{\boldsymbol{\alpha}_\infty} \leq \liminf_{t\to+\infty} \|\mathbf{u}_t - \mathbf{x}\|_{\boldsymbol{\alpha}_\infty} - \lim_{t\to+\infty} \|\mathbf{u}_t - \mathbf{z}_k\|_{\boldsymbol{\alpha}_\infty}$$

$$\leq \limsup_{t\to+\infty} \|\mathbf{u}_t - \mathbf{x}\|_{\boldsymbol{\alpha}_\infty} - \lim_{t\to+\infty} \|\mathbf{u}_t - \mathbf{z}_k\|_{\boldsymbol{\alpha}_\infty}$$
$$\leq \|\mathbf{z}_k - \mathbf{x}\|_{\boldsymbol{\alpha}_\infty}.$$

Taking the limit as $k \to +\infty$, we deduce that $\lim_{k\to+\infty} \lim_{t\to+\infty} \|\mathbf{u}_t - \mathbf{z}_k\|_{\boldsymbol{\alpha}_\infty}$ exists and

$$\liminf_{t\to+\infty} \|\mathbf{u}_t - \mathbf{x}\|_{\boldsymbol{\alpha}_\infty} = \lim_{k\to+\infty} \lim_{t\to+\infty} \|\mathbf{u}_t - \mathbf{z}_k\|_{\boldsymbol{\alpha}_\infty} = \limsup_{t\to+\infty} \|\mathbf{u}_t - \mathbf{x}\|_{\boldsymbol{\alpha}_\infty}.$$

This shows the convergence of $\|\mathbf{u}_t - \mathbf{x}\|_{\boldsymbol{\alpha}_\infty}$. □

In Lemma B.6, we consider a sequence of random weights $(\boldsymbol{\alpha}_t)_{t\in\mathbb{N}}$ so that it also applies when the learning rates are not constant but converge. We only use this lemma in its most general form in the proof of Theorem 8.26. Otherwise, the following corollary is sufficient.

Corollary B.7. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a closed set, $(\mathbf{u}_t)_{t\in\mathbb{N}}$ be a sequence of $\mathbb{R}^d$-valued random variable, and $\boldsymbol{\alpha} \in \mathbb{R}^N$ such that $\alpha^i \geq 1$ for all $i \in \mathcal{N}$, and for all $\mathbf{x} \in \mathcal{K}$, $\|\mathbf{u}_t - \mathbf{x}\|_{\boldsymbol{\alpha}}$ converges almost surely. Then, with probability 1, $\|\mathbf{u}_t - \mathbf{x}\|_{\boldsymbol{\alpha}}$ converges for all $\mathbf{x} \in \mathcal{K}$.*

Remark B.2. The subtlety in Lemma B.6 and Corollary B.7 is in the position of the quantifier "for all". Initially, we have a probability event for each point and each event holds with probability 1. With the theorem, we show that all these events can be grouped together to form a single event that holds with probability 1.

ADAPTIVE LEARNING RATES. We close this appendix with two lemmas that are used in Section 8.3 for the analysis of adaptive (OptDA+). The first one extends Lemma 2.6 to deal with any exponent $r \in [0, 1)$.

Lemma B.8. *Let $T \in \mathbb{N}$, $\varepsilon > 0$, and $r \in [0, 1)$. For any sequence of non-negative real numbers $a_1, \dots, a_T$, it holds*

$$\sum_{t=1}^{T} \frac{a_t}{\left(\varepsilon + \sum_{s=1}^{t} a_s\right)^r} \leq \frac{1}{1-r} \left(\sum_{t=1}^{T} a_t\right)^{1-r}. \tag{B.3}$$

Proof. The function $y \in \mathbb{R}_+ \mapsto y^{1-r}$ is concave and has derivative $y \mapsto (1-r)/y^r$. Therefore, it holds for every $y, z > 0$ that

$$z^{1-r} \leq y^{1-r} + \frac{1-r}{y^r}(z - y).$$

For $\varepsilon' \in (0, \varepsilon)$, we apply the above inequality to $y = \varepsilon' + \sum_{s=1}^{t} a_s$ and $z = \varepsilon' + \sum_{s=1}^{t-1} a_s$. This gives

$$\begin{aligned}\frac{1}{1-r}\left(\varepsilon' + \sum_{s=1}^{t-1} a_s\right)^{1-r} &\leq \frac{1}{1-r}\left(\varepsilon' + \sum_{s=1}^{t} a_s\right)^{1-r} - \frac{a_t}{\left(\varepsilon' + \sum_{s=1}^{t} a_s\right)^r} \\ &\leq \frac{1}{1-r}\left(\varepsilon' + \sum_{s=1}^{t} a_s\right)^{1-r} - \frac{a_t}{\left(\varepsilon + \sum_{s=1}^{t} a_s\right)^r}. \end{aligned} \tag{B.4}$$

Moreover, at $t = 1$ we have

$$\frac{a_1}{(\varepsilon + a_1)^r} \leq (\varepsilon' + a_1)^{1-r} \leq \frac{1}{1-r}(\varepsilon' + a_1)^{1-r}. \tag{B.5}$$

Summing (B.4) from $t = 2$ to T, adding (B.5), and rearranging leads to

$$\sum_{t=1}^{T} \frac{a_t}{\left(\varepsilon + \sum_{s=1}^{t} a_s\right)^r} \leq \frac{1}{1-r}\left(\varepsilon' + \sum_{t=1}^{T} a_t\right)^{1-r}.$$

Provided that the above inequality holds for any $\varepsilon' \in (0, \varepsilon)$, we obtain (B.3) by taking $\varepsilon' \to 0$. □

Remark B.3. Similar to Lemma 2.6, we can also take $\varepsilon = 0$ in Lemma B.8 if we adopt the notation $0/0 = 0$. To see this, assume that $a_{t'}$ is the first non-zero number in the sequence (set $t' = T + 1$ if such number does not exist). We can ignore the sum up to $t' - 1$ and for the remaining terms we use (B.3) starting from t' and take $\varepsilon \to 0$.

As for the second lemma, it allows us to bound the moments of a collection of random variables through an inequality that relates their moments of different orders. This lemma is used in the proof of Lemma 8.22.

Lemma B.9. *Let $p, r, c \in \mathbb{R}_+$ such that $p > r$, and $(a^1, \ldots, a^N)$ be a collection of N non-negative real-valued random variables. If*

$$\sum_{i=1}^{N} \mathbb{E}[(a^i)^p] \leq c \sum_{i=1}^{N} \mathbb{E}[(a^i)^r], \tag{B.6}$$

Then $\sum_{i=1}^{N} \mathbb{E}[(a^i)^p] \leq N c^{\frac{p}{p-r}}$ and $\sum_{i=1}^{N} \mathbb{E}[(a^i)^r] \leq N c^{\frac{r}{p-r}}$.

Proof. Since $p > r$, the function $y \in \mathbb{R}_+ \mapsto y^{\frac{r}{p}}$ is concave. Applying Jensen's inequality for the expectation gives $\mathbb{E}[(a^i)^r] \leq \mathbb{E}[(a^i)^p]^{\frac{r}{p}}$. Next, we apply Jensen's inequality for the average to obtain

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[(a^i)^p]^{\frac{r}{p}} \leq \left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[(a^i)^p]\right)^{\frac{r}{p}}. \tag{B.7}$$

Along with inequality (B.6) we then get

$$\sum_{i=1}^{N} \mathbb{E}[(a^i)^p] \leq c \sum_{i=1}^{N} \mathbb{E}[(a^i)^r] \leq c N^{1-\frac{r}{p}} \left(\sum_{i=1}^{N} \mathbb{E}[(a^i)^p]\right)^{\frac{r}{p}}. \tag{B.8}$$

In other words

$$\left(\sum_{i=1}^{N} \mathbb{E}[(a^i)^p]\right)^{1-\frac{r}{p}} \leq c N^{1-\frac{r}{p}}.$$

Taking both sides of the inequality to the power of $p/(p-r)$, we obtain effectively

$$\sum_{i=1}^{N} \mathbb{E}[(a^i)^p] \leq N c^{\frac{p}{p-r}}$$

The second inequality combines the above with second part of (B.8). □

C

FROM PSEUDO-REGRET TO EXPECTED REGRET

In this appendix, we explain how can provide bound on expected regret for the algorithms and setup described in Chapter 8.

To begin, we first observe that Lemma 8.5 can in fact be restated as a bound on $\mathbb{E}\left[\max_{z^i \in \mathcal{Z}^i} \sum_{t=1}^{T} \langle \hat{V}^i_{t+\frac{1}{2}}, x^i_t - z^i \rangle\right]$ for any comparator set $z^i \subseteq \mathcal{X}^i$ as below.

Lemma C.1. *Suppose that Assumptions 5.2 and 7.1 hold and all players run* (OptDA+) *with non-increasing learning rates satisfying Assumption 8.1 and $\boldsymbol{\eta}_t \leq \gamma_t$ for all $t \in \mathbb{N}$. Then, for all $i \in \mathcal{N}$, $T \in \mathbb{N}$, and bounded set $\mathcal{Z}^i \subset \mathcal{X}^i$ with $R \geq \sup z^i \in \mathcal{Z}^i \|X^i_1 - z^i\|$, we have*

$$\mathbb{E}\left[\max_{z^i \in \mathcal{Z}^i} \sum_{t=1}^{T} \langle \hat{V}^i_{t+\frac{1}{2}}, x^i_t - z^i \rangle\right] \leq \mathbb{E}\left[\frac{R^2}{2\eta^i_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \eta^i_t \|\hat{V}^i_{t+\frac{1}{2}}\|^2 + \sum_{t=2}^{T} \gamma^i_t L^2 \left(3\|\hat{\mathbf{V}}_{t-\frac{1}{2}}\|^2_{\gamma^i_t} + \frac{3}{2}\|\mathbf{X}_t - \mathbf{X}_{t-1}\|^2\right) + \sum_{t=2}^{T} ((\gamma^i_t)^2 \sqrt{N} L \|\xi^i_{t-\frac{1}{2}}\|^2 + \frac{L}{\sqrt{N}} \|\boldsymbol{\xi}_{t-\frac{1}{2}}\|^2_{\gamma^i_t})\right].$$

Proof. Let us define $z^i_\star = \arg\max_{z^i \in \mathcal{Z}} \langle \hat{V}^i_{t+\frac{1}{2}}, x^i_t - z^i \rangle$. This is a random variable whose value depends on the actual feedback. We note however that $\langle \hat{V}^i_{t+\frac{1}{2}}, x^i_t - z^i \rangle$ is in fact exactly the scalar term that appears in Corollary 8.2. Therefore, we just need to conduct the proof without further modifying this term (in particular, we *do not* have anymore $\mathbb{E}[\langle \hat{V}^i_{t+\frac{1}{2}}, x^i_t - z^i_\star \rangle] = \mathbb{E}[\langle V^i(\mathbf{X}_{t+\frac{1}{2}}), x^i_t - z^i_\star \rangle]$) and we get the desired result. □

Lemma C.1 provides a bound for the linearized regret evaluated with respect to the noisy feedback. Nonetheless, what we need is a bound for the linearized regret evaluated with respect to the noiseless feedback, from which we can then deduce a bound on the expected regret $\mathbb{E}[\mathrm{Reg}^i_T(\mathcal{Z}^i)]$ using the convexity assumption (Assumption 5.1). To achieve this, we rely on the following lemma.

Lemma C.2. *Let $i \in \mathcal{N}$ and $\mathcal{Z}^i \subset \mathcal{X}^i$ be a compact. Let $R = \max_{z^i} \|x^i_1 - z^i\|$. Then, under Assumption 7.1, for any $\mathcal{F}_t$-adapted sequence of played points $(\mathbf{x}_t)_{t \in \mathbb{N}}$, it holds*

$$\mathbb{E}\left[\max_{z^i \in \mathcal{Z}^i} \sum_{t=1}^{T} \langle V^i(\mathbf{x}_t), x^i_t - z^i \rangle\right] \leq \mathbb{E}\left[\max_{z^i \in \mathcal{Z}^i} \sum_{t=1}^{T} \langle \hat{V}^i_t, x^i_t - z^i \rangle\right] + R\sqrt{\sum_{t=1}^{T} \mathbb{E}[\|\xi^i_t\|^2]}$$

Proof. Let $z^i_\star = \arg\max_{z^i \in \mathcal{Z}} \langle V^i(x^i_t), x^i_t - z^i \rangle$. We remark that $z^i_\star$ is a random variable that depends on the realization of the noises up to time t. Since $\hat{V}^i_t = V^i(\mathbf{x}_t) + \xi^i_t$, we have

$$\langle V^i(\mathbf{x}_t), x^i_t - z^i_\star \rangle = \langle \hat{V}^i_t, x^i_t - z^i_\star \rangle - \langle \xi^i_t, x^i_t - x^i_1 \rangle - \langle \xi^i_t, x^i_1 - z^i_\star \rangle \tag{C.1}$$

Provided that both x^i_t and x^i_1 are $\mathcal{F}_t$-measurable, with the law of total expectation we deduce that the expectation of the second term is 0,

$$\mathbb{E}[\langle \xi^i_t, x^i_t - x^i_1 \rangle] = \mathbb{E}[\mathbb{E}_t[\langle \xi^i_t, x^i_t - x^i_1 \rangle]] = \mathbb{E}[\langle \mathbb{E}_t[\xi^i_t], x^i_t - x^i_1 \rangle] = 0.$$

On the other hand, the sum of the third term can be bounded using the definition of R

$$\mathbb{E}\left[\sum_{t=1}^T \langle \xi^i_t, x^i_1 - z^i_\star \rangle\right] \leq \mathbb{E}\left[\left\|\sum_{t=1}^T \xi^i_t\right\| \|x^i_1 - z^i_\star\|\right] \leq \mathbb{E}\left[R\left\|\sum_{t=1}^T \xi^i_t\right\|\right].$$

Applying Jensen's inequality we get

$$\mathbb{E}\left[\left\|\sum_{t=1}^T \xi^i_t\right\|\right] \leq \sqrt{\mathbb{E}\left[\left\|\sum_{t=1}^T \xi^i_t\right\|^2\right]} = \sqrt{\sum_{t=1}^T \mathbb{E}[\|\xi^i_t\|^2]}.$$

Combining the above we obtain

$$\mathbb{E}\left[\sum_{t=1}^T \langle V^i(\mathbf{x}_t), x^i_t - z^i_\star \rangle\right] \leq \mathbb{E}\left[\sum_{t=1}^T \langle \hat{V}^i_t, x^i_t - z^i_\star \rangle\right] + R\sqrt{\sum_{t=1}^T \mathbb{E}[\|\xi^i_t\|^2]}$$

To conclude, we upper bound the second term by $\mathbb{E}\left[\max_{z^i \in \mathcal{Z}^i} \sum_{t=1}^T \langle \hat{V}^i_t, x^i_t - z^i \rangle\right]$. □

With Lemmas C.1 and C.2, we see immediately that compared to the pseudo-regret bound that we provided in Chapter 8, the only additional term is $R\sqrt{\sum_{t=1}^T \mathbb{E}[\|\xi^i_t\|^2]}$. Thanks to Assumption 7.1, we have

$$\begin{aligned} R\sqrt{\sum_{t=1}^T \mathbb{E}[\|\xi^i_{t+\frac{1}{2}}\|^2]} &\leq R\sqrt{\sum_{t=1}^T (\sigma_M^2\, \mathbb{E}[\|V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2] + \sigma_A^2)} \\ &\leq R\sqrt{\sigma_M^2 \sum_{t=1}^T \mathbb{E}[\|V^i(\mathbf{X}_{t+\frac{1}{2}})\|^2]} + R\sigma_A\sqrt{T}. \end{aligned}$$

Therefore, using Theorem 8.9 we see immediately that the $O(\sqrt{T})$ and $O(1)$ regret bounds of Theorem 8.11 are still valid in the case of predetermined learning rates that we studied in Section 8.2. As for the case of adaptive learning rates that we examined in Section 8.3, we have directly $R\sqrt{\sum_{t=1}^T \mathbb{E}[\|\xi^i_{t+\frac{1}{2}}\|^2]} \leq R\bar{\sigma}\sqrt{T}$ under Assumption 8.2. As such, the only regret bound that we are not able to recover is the $O(1)$ bound of Theorem 8.24. This is because we are not able to show that the sum $\sum_{t=1}^{+\infty} \|\mathbf{V}(\mathbf{X}_{t+\frac{1}{2}})\|^2$ is finite in expectation when we use adaptive learning rates in the multiplicative noise regime (instead, we show it is finite almost surely in Theorem 8.19).

D

LIST OF PUBLICATIONS

This appendix provides an up-to-date list of the author's scientific publications.

CONFERENCE PAPERS

[1] Yu-Guan Hsieh, Shiva Kasiviswanathan, Branislav Kveton, and Patrick Bloebaum. Thompson Sampling with Diffusion Generative Prior. In *International Conference on Machine Learning*, 2023.

[2] Yu-Guan Hsieh, Kimon Antonakopoulos, Volkan Cevher, and Panayotis Mertikopoulos. No-Regret Learning in Games with Noisy Feedback: Faster Rates and Adaptivity via Learning Rate Separation. In *Advances in Neural Information Processing Systems*, 2022.

[3] Yu-Guan Hsieh, Shiva Kasiviswanathan, and Branislav Kveton. Upliting Bandits. In *Advances in Neural Information Processing Systems*, 2022.

[4] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Optimization in Open Networks via Dual Averaging. In *IEEE Conference on Decision and Control*, 2021.

[5] Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive Learning in Continuous Games: Optimal Regret Bounds and Convergence to Nash Equilibrium. In *Conference on Learning Theory*, 2021.

[6] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore Aggressively, Update Conservatively: Stochastic Extragradient Methods with Variable Stepsize Scaling. In *Advances in Neural Information Processing Systems*, 2020.

[7] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the Convergence of Single-Call Stochastic Extra-Gradient Methods. In *Advances in Neural Information Processing Systems*, 2019.

[8] Yu-Guan Hsieh, Gang Niu, and Masashi Sugiyama. Classification from Positive, Unlabeled and Biased Negative Data. In *International Conference on Machine Learning*, 2019.

JOURNAL PAPERS

[1] Yu-Guan Hsieh, Yassine Laguel, Franck Iutzeler, and Jérôme Malick. Push–Pull with Device Sampling. *IEEE Transactions on Automatic Control*, 2023.

[2] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Multi-agent Online Optimization with Delays: Asynchronicity, Adaptivity, and Optimism. *Journal of Machine Learning Research*, 2022.

LIST OF FIGURES

Figure 1.1 Illustrations of the learning setups considered in this thesis. 5
Figure 1.2 Schematic representation of (DOptDA). 6
Figure 1.3 The trajectories of play when running (OptDA) with quadratic regularizer on $\min_{\theta\in[-4,8]} \max_{\phi\in[-4,8]} \theta\phi$ using different learning rates. 8
Figure 1.4 The behaviors of (OG), (OG+), and adaptive (OptDA+) on the game $\min_{\theta\in\mathbb{R}} \max_{\phi\in\mathbb{R}} \theta\phi$ when the feedback is corrupted by (multiplicative) noise. 9
Figure 1.5 Illustration of the uplifting bandit model. This example has 3 arms, 5 variables, and each arm affects 2 variables. Dash lines indicate the variables' payoffs follow the baseline distribution by default. 10
Figure 1.6 Overview of the meta-learning for bandits with diffusion prior framework. 11
Figure 2.1 The online learning framework. 16
Figure 2.2 Schematic representations of mirror descent. 20
Figure 2.3 Schematic representations of dual averaging. 24
Figure 3.1 Illustration of the considered multi-agent online-learning setup: the case of coordinator-worker architecture. 33
Figure 3.2 Illustration of the considered multi-agent online-learning setup: the case of decentralized open network. 33
Figure 3.3 Illustration of the type of feedback sequences that may occur in a multi-agent setting. 34
Figure 3.4 Dependency graphs for the two examples of Fig. 3.3. 37
Figure 3.5 Schematic representation of the communication graphs used in our experiments. These diagrams are simplified representations of the actual graphs, which contain 64 nodes, and mainly depict the general structure and connectivity. 58
Figure 3.6 Comparison of (D-DDA) and (DGD). For a static network we plot in (a) the averaged suboptimality. For an open network we plot in (b) the averaged instantaneous suboptimality (3.22) and the averaged running loss (3.23). 59
Figure 4.1 Schematic representation of (DOptDA). 68
Figure 4.2 Illustration of the evolution of (DOptDA) for a period of feedback in the first example of the proof of Theorem 4.5. 70
Figure 5.1 The learning-in-games framework. 82
Figure 5.2 Schematic representation of the (VI) problem (adapted from [191]). $\mathrm{TC}_{\mathcal{X}}(x_\star)$ and $\mathrm{NC}_{\mathcal{X}}(x_\star)$ are respectively the tangent and the normal cones of $\mathcal{X}$ at $x_\star$. 87
Figure 5.3 Illustration of the "separate and project" principle of the optimistic algorithms. 93
Figure 6.1 The trajectories of play when running (OptMD)/(OptDA) with quadratic regularizer on $\min_{\theta\in[-4,8]} \max_{\phi\in[-4,8]} \theta\phi$ using different learning rates 96
Figure 6.2 The (linearized) individual regret and the realized actions of a subset of players in a finite two-player zero-sum game, a resource allocation auction, and a three-player matching-pennies game. All the players use either adaptive OptDA or adaptive DS-OptMD as their learning strategies. 116
Figure 7.1 Trajectories of play and distances to equilibrium of (EG), (OG), (EG+), and (OG+) when run on the game $\min_{\theta\in\mathbb{R}} \max_{\phi\in\mathbb{R}} \theta\phi$ with stochastic feedback presented in Example 7.1. 120

Figure 7.2 The stepsize exponents allowed by condition (7.15) for convergence (shaded green). Dashed lines are strict frontiers. Note that vanilla (EG) and (OG) (the separatrix $r_\eta = r_\gamma$) passes just outside of this region, explaining the methods' failures. 128
Figure 8.1 Regret of the first player with respect to 0 and the distance between the iterates and the Nash equilibrium when the players run (GDA), (OG), (OG+), or adaptive (OptDA+) in the bilinear game $\min_{\theta\in\mathbb{R}} \max_{\phi\in\mathbb{R}} \theta\phi$ under stochastic feedback. 182
Figure 8.2 Convergence of (EG+) and (OG+) in the linear quadratic Gaussian GAN model. 183

ACRONYMS

CCE	coarse correlated equilibrium
DA	dual averaging
DDA	delayed dual averaging
D-DDA	decentralized delayed dual averaging
DE	dual extrapolation
DGD	decentralized gradient descent
DOptDA	delayed optimistic dual averaging
DS-OptMD	dual stabilized optimistic mirror descent
EG	extra-gradient
FTRL	follow the regularized leader
GAN	generative adversarial network
GDA	gradient descent/ascent
LAD	least absolute deviation
MARL	multi-agent reinforcement learning
MD	mirror descent
MP	mirror-prox
MVI	Minty variational inequality
MWU	multiplicative weights update
OptDA	optimistic dual averaging
OG	optimistic gradient
OMWU	optimistic multiplicative weights update
OptMD	optimistic mirror descent
SVI	Stampacchia variational inequality
VI	variational inequality

COLOPHON

This manuscript was typeset with LaTeX 2ε using Hermann Zapf's Palatino type face (the actual Type 1 PostScript fonts used were URW Palladio L and FPL). The monospaced text (hyperlinks, etc.) was typeset in *Bera Mono*, originally developed by Bitstream, Inc. as "Bitstream Vera" (with Type 1 PostScript fonts by Malte Rosenau and Ulrich Dirr).

The typographic style of this dissertation was inspired by the authoritative genius of Bringhurst's *Elements of Typographic Style*, ported to LaTeX by André Miede, the original designer of the `classicthesis` template. Any unsightly deviations from these works should be attributed solely to the author's (not always successful) efforts to conform to the awkward A4 paper size.

Decision-Making in Multi-Agent Systems: Delays, Adaptivity, and Learning in Games

Grenoble, November 7, 2023

Yu-Guan Hsieh