Continuous and Discrete Dynamics
For Online Learning and Convex Optimization

by

Walid Krichene

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

and the Designated Emphasis

in

Communication, Computation and Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Alex M. Bayen, Chair
Professor Peter L. Bartlett
Professor Nikhil Srivastava

Fall 2016

Continuous and Discrete Dynamics
For Online Learning and Convex Optimization

Walid Krichene

Abstract

Continuous and Discrete Dynamics
For Online Learning and Convex Optimization

by

Walid Krichene

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences
and the Designated Emphasis in
Communication, Computation and Statistics

University of California, Berkeley

Professor Alex M. Bayen, Chair

Online learning and convex optimization algorithms have become essential tools for solving problems in modern machine learning, statistics and engineering. Many algorithms for online learning and convex optimization can be interpreted as a discretization of a continuous time process, and studying the continuous time dynamics offers many advantages: the analysis is often simpler and more elegant in continuous time, it provides insights and leads to new interpretations of the discrete process, and streamlines the design of new algorithms, obtained by deriving the dynamics in continuous time, then discretizing. In this thesis, we apply this paradigm to two problems: the study of decision dynamics for online learning in games, and the design and analysis of accelerated methods for convex optimization.

In the first part of the thesis, we study online learning dynamics for a class of games called non-atomic convex potential games, which are used for example to model congestion in transportation and communication networks. We make a connection between the discrete Hedge algorithm for online learning, and an ODE on the simplex, known as the replicator dynamics. We study the asymptotic properties of the ODE, then by discretizing the ODE and using results from stochastic approximation theory, we derive a new class of online learning algorithms with asymptotic convergence guarantees. We further give a more refined analysis of these dynamics and their convergence rates. Then, using the Hedge algorithm as a model of decision dynamics, we pose and study two related problems: the problem of estimating the learning rates of the Hedge algorithm given observations on its sequence of decisions, and the problem of optimal control under Hedge dynamics.

In the second part, we study first-order accelerated dynamics for constrained convex optimization. We develop a method to design an ODE for the problem using an inverse Lyapunov argument: we start from an energy function that encodes the constraints of the problem and the desired convergence rate, then design an ODE tailored to that energy function. Then, by carefully discretizing the ODE, we obtain a family of accelerated algorithms with opti-

mal rate of convergence. This results in a unified framework to derive and analyze most known first-order methods, from gradient descent and mirror descent to their accelerated versions. We give different interpretations of the ODE, inspired from physics and statistics. In particular, we give an averaging interpretation of accelerated dynamics, and derive simple sufficient conditions on the averaging scheme to guarantee a given rate of convergence. We also develop an adaptive averaging heuristic that empirically speeds up the convergence, and in many cases performs significantly better than popular heuristics such as restarting.

To my parents, Sami and Ibtissème.
To my sister, Syrine.

Contents

List of Figures

List of Algorithms

Acknowledgments

The five years of my graduate studies at Berkeley have been some of the happiest and most intellectually gratifying years of my life, and this is due in large part to the professors and friends I collaborated with during these years. There are too many people who had a positive impact on my academic and personal life to list here, and I apologize in advance to anyone whom I neglected to mention.

I must begin by thanking my Ph.D. advisors, Alex and Peter, for their guidance and their support throughout the years. Alex has been an outstanding mentor and friend, and he gave me a great deal of freedom in defining my research agenda and finding my own interests. Without his encouragements, I would not have been able to work on such a wide range of topics, from control theory and convex optimization to machine learning and statistics. His vision kept me grounded and focused, and his patience and advice helped me hone the different skills needed to navigate graduate school, from writing research papers and giving talks, to teaching classes and organizing reading groups. I started working with Peter during the third year of my Ph.D., after taking his phenomenal class on Learning in Sequential Decision Problems. I have a great deal of admiration for Peter and his scientific maturity, and the extent of his knowledge and technical ability is simply incredible. He has been an unfailing source of inspiration, and I enjoyed every one of our discussions, which never failed to give me new ideas to try. I felt immediately welcome in his research group, and his reading group has given rise to some of the most fascinating discussions I have had in the last few years.

Berkeley offered me a great environment to learn from the best, and do research alongside the brightest professors and students in the field. And even though it might seem intimidating to interact with the best and brightest, I always felt that my ideas were appreciated, even as a starting Ph.D. student. There are many other faculty whom I interacted with, and who had a great impact on my approach to research and teaching; it is their classes and their teaching that maintained my sense of wonder and my desire to learn: Claire Tomlin, Shankar Sastry and Murat Arcak, who taught the best control theory classes I ever took and who made me feel appreciated in the control community; Laurent El Ghaoui, who was extremely kind and helpful to me, and whose expertise in convex optimization is unmatched; Satish Rao who taught one of the most fun classes I took during my graduate studies, and who provided some very helpful pointers that were the starting point of much of my work on online learning. I would also like to thank my mathematics professors, Michael Christ who made me fall in love with topology and measure theory again, David Aldous for his amazing probability theory class, and Nikhil Srivastava for some illuminating discussions on convex analysis, and for being very kind to be on my dissertation committees, both for my M.A. and Ph.D. theses.

I have also collaborated with some outstanding graduate students during my time at Berkeley. I took all of my math classes with Roy, Jupiter and Max, who became some of my dearest friends. I would not have enjoyed these classes nearly as much without them. I will fondly remember the many weekends spent together going through notes and working

on homework problems. I will also miss our Rockafellar reading group with Roy and Dan, who shared my excitement and passion for convex analysis. I also enjoyed working with Max on learning on infinite action sets, and thank him for his dedication and his ability to work through some intricate and subtle proofs. I have also supervised many undergraduate researchers in the last few years, and I enjoyed collaborating with every one of them. I have to thank Benjamin in particular, whose scientific maturity and mathematical insight were quite impressive. His contributions appear in much of the first part of my thesis. I would also like to thank Syrine for her outstanding work on stochastic optimization, and both Kiet and Chedly for their meticulous work on the routing game web application.

I would like to thank my dear friends for sharing some great memories over the years: Roy, Katie, Dan and Jerôme for many fun board game nights, Sandra and Aaron for memorable camping trips, Samy for fun tennis games and cooking experiments, Jupiter for sharing his talent and passion for math, Marouen, Omar and Alan for being wonderful travel companions.

Finally, I cannot thank my family enough for being there for me every step of the way, and for believing in me. My parents, Sami and Ibtissème, gave me their love and caring and everything a child could hope for, and helped me develop and maintain my curiosity and love for mathematics throughout the years, by helping me in school when I was young, encouraging me to go to the math olympiads, and later to prépa school, and to pursue the career that I truly wanted. It is their love and their encouragements that kept me going during the difficult times. I also thank my sister, Syrine, for bringing me joy. Her optimism, her curiosity and her kindness make her the best sister one could hope for. I am very proud of her, and I love her dearly.

Chapter 1

Introduction

> The most practical thing in the world is a good theory.
>
> *H. von Helmholtz*

1.1 From continuous time ODEs to discrete time algorithms

Many discrete algorithms for online learning and convex optimization can be interpreted as a discretization of a continuous-time dynamics. Perhaps the simplest and oldest example is the gradient descent algorithm. If we seek to minimize a differentiable convex function f on $\mathbb{R}^n$, gradient descent can be written as a sequence of iterates $(x^{(t)})$ satisfying $x^{(t+1)} = x^{(t)} - \eta_t \nabla f(x^{(t)})$, where η_t is a positive step size. This difference equation can be interpreted as a discrete-time approximation of the ODE $\dot{X}(t) = -\nabla f(X(t))$, with discretization step η_t. While most algorithms are inherently discrete, studying the continuous-time process can be useful for many reasons. The analysis is often simpler in continuous-time, and can benefit from the well-established theory of differential equations and dynamical systems. It can also provide intuition, and new insights into the discrete process, and can help guide the design and analysis of new algorithms. For example, an important question in the analysis of many discrete algorithms is the asymptotic behavior of the trajectories, and whether they converge to a given set (this could be e.g. the set of minimizers of a convex function, or the set of equilibria of a game). Convergence of solution trajectories is often simpler to prove in continuous-time, and can be done for instance by exhibiting a Lyapunov function for the invariant set, that is, a function that is non-increasing along solution trajectories, and that is minimal on the invariant set. Once the convergence is established in continuous time, one can then discretize the ODE, and attempt to prove convergence using a discrete-time counterpart of the Lyapunov function. In this thesis, we explore some of these techniques in the context of two classes of problems: online learning dynamics in games, studied in the

first part of the thesis, and accelerated dynamics for convex optimization, studied in the second part.

1.2 Online learning and games

Online learning theory studies sequential decision problems, in which a decision maker iteratively chooses an action and observes outcomes. This model of sequential decision is relevant to many systems, from physical systems such as transportation networks and power networks (the network users make decisions as new information becomes available), to online systems, such as online advertising and auctions.

Many of these systems can be modeled as games, and one can study their Nash equilibria [95], which describe strategies for players such that no player has an incentive to unilaterally deviate. However, the Nash equilibrium concept may not always offer a good descriptive model of actual behavior of players. Besides the assumption of rationality, which can be questioned [129], the Nash equilibrium usually assumes that players have a good description of the game, of the other players, and of their utilities, which is not realistic for many large-scale distributed systems.

One alternative model of player behavior is repeated play [92, 51, 90], sometimes called learning models [40] or adjustment models [54]. In such models, one assumes that each player makes decisions *iteratively* (instead of playing a one-shot game), and uses the outcome of each iteration to adjust their next decision. Formally, at every iteration t, player k makes a decision $x_k^{(t)}$, then observes an outcome $\ell_k^{(t)}$ (e.g. a vector of losses of all the possible actions), so from the perspective of each player, this is a sequential decision problem. These problems are of course coupled through the outcomes, since $\ell_k^{(t)}$ depends on $x_k^{(t)}$ but also on $x_{k'}^{(t)}$ for $k' \neq k$. This is illustrated in Figure 1.1.

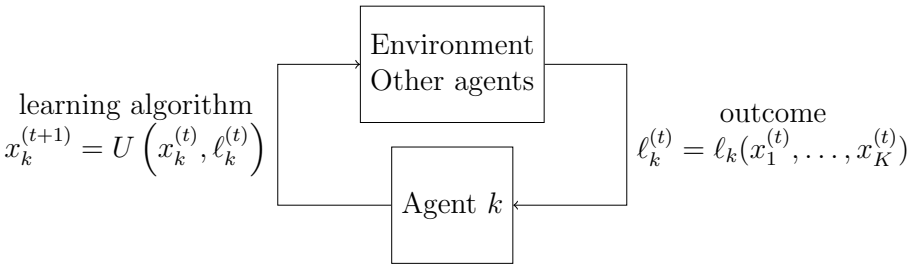


Figure 1.1: Coupled sequential decision problems. Each player faces an online learning problem, and the decisions of the different players are coupled through their loss functions.

In such models, a natural question is whether the joint decision dynamics of the players converge to some equilibrium set, for example to the Nash equilibrium of the game if it were to be played as a one-shot game. This question has a long history in game theory and mathematics, and dates back to the work of Hannan [57], who defined the regret, and Blackwell [25] who defined approachability, and both concepts have become key in the

design and analysis of online learning algorithms. For example, regret-based dynamics in games have been studied in [5, 89], and by Hart and Mas-Colell, both in continuous [61] and discrete time [60, 59, 58, 62]. See also [40] and references therein. Regret is also central in other classes of online learning problems, such as bandit problems [35, 34], and online convex optimization [63, 126, 11]. Blackwell approachability has been used to study learning in games, for example in [39], and many connections are made between regret and approachability [1, 107].

Continuous-time dynamics have also been studied for several classes of games, see for example [67, 135, 66, 123, 19], in which different families of ODEs are used to describe the time evolution of the decision dynamics of player populations. In [124], Sandholm studies convergence for the class of potential games. He shows that dynamics which satisfy a positive correlation condition with respect to the potential function of the game converge to the set of stationary points of the vector field (usually, a superset of Nash equilibria). In [68], Hofbauer and Sandholm study the convergence of a class of dynamics called excess payoff target (EPT), for the class of stable games. In [51], Fox and Shamma extend these convergence results to passive evolutionary dynamics, and give a dynamical systems interpretation. Some approaches even generalize the class of dynamics and consider differential inclusions instead of differential equations, see [21, 22].

Our contributions

In the first part of the thesis, we study learning dynamics in the class of nonatomic population games that admit a convex potential (which will be formally defined in Chapter 2). This class of games can be used to model the interaction of large populations of players, and have a special structure due to the existence of the convex potential. This will allow us to apply three different techniques to study learning dynamics:

1. First, we will analyze regret-based dynamics in Chapter 2.
2. Second, we study a continuous-time learning dynamics in Chapter 3, known as the replicator dynamics, and study the asymptotic behavior of its solutions. Then building on results from stochastic approximation theory [18], we show in Chapter 4 how the replicator ODE can be discretized while preserving convergence. We call the resulting algorithms approximate replicator (AREP).
3. Third, we use techniques from stochastic convex optimization, to analyze, in Chapter 5, the convergence properties of a class of dynamics based on the mirror descent method.

These convergence results are presented from the weakest to the strongest: Regret-based dynamics have the weakest convergence guarantee, we show that the sequence of decisions converges in the sense of Cesàro, i.e. that the weighted averages converge. For the approximate replicator dynamics, we show almost sure convergence. For mirror descent dynamics, we derive explicit convergence rates.

Hedge algorithm The Hedge algorithm will be central in our discussion. It is perhaps one of the most well studied online learning algorithms, also known as the multiplicative weights update [4] in the computer science literature, the exponentiated gradient algorithm [72] or the entropic descent algorithm [15] in the optimization literature, as well as log-linear learning [28, 91] in the economics and game theory literature. It is also known to be an instance of the mirror descent family of methods due to Nemirovski and Yudin [98], which we discuss in detail in Appendix B.

Using the Hedge algorithm, we will see in particular that the connection between discrete time and continuous time dynamics can be useful in both directions: In Chapter 3, we show that the continuous-time replicator equation can be motivated as the continuous-time limit of the Hedge algorithm. In Chapter 4, we show that by carefully discretizing the replicator ODE, we can obtain a larger family of algorithms (which contains the Hedge algorithm), while preserving convergence. This is achieved by ensuring that the discrete trajectory is close, in a sense to be made precise in Chapter 4, to the continuous solution trajectories of the ODE. And since the latter are guaranteed to converge to the equilibrium set, we can provide guarantees on the discrete process.

Routing games The routing game is a special case of a nonatomic population game, which can be used to model congestion in many cyber physical systems in which non-cooperative players compete for shared resources, such as transportation networks [16, 119] (the resources being roads) and communication networks [106] (the resources being communication links). Our study of nonatomic population games is motivated in particular by routing games, which we will use in many of the numerical examples provided throughout the thesis.

Modeling decision dynamics Beyond the design and analysis of learning algorithms and their convergence properties, we study the problem of modeling the decision dynamics of players. As argued by Marden and Shamma in [92], online learning can be used not only as a prescriptive tool, used to solve sequential decision problems, but also as a descriptive tool, used to model the behavior of players. We explore the second point of view in Chapter 6 and 7. First, we consider the problem of estimating the learning rates of a decision maker that follows the Hedge algorithm. More precisely, we suppose that we can observe the sequence of decisions that obey the Hedge dynamics, with unknown learning rates, and show how the learning rates can be estimated. We consider the Hedge model in particular since it is both an instance of the AREP class studied in Chapter 4, and the mirror descent class studied in Chapter 5.

To apply this method on field data, we implement a web application that simulates a routing game. Players can use the application to participate in a simultaneous, online version of the game, and make sequential decisions on how to allocate their traffic on a shared network (without directly interacting or observing the decisions of other players). We use this experiment to study some qualitative aspects of decision dynamics, and test our learning rate estimation approach. The results indicate that the Hedge algorithm can be descriptive

of actual decision dynamics. In Chapter 7, we study the related problem of optimal control of a population of online learners who follow the Hedge dynamics. Assuming we have estimated the learning rates, we pose the problem of optimally controlling the game in order to minimize a given objective. Due to the presence of the non-linear Hedge constraints, this problem is non-convex, but we propose a method for finding a local minimizer, using the adjoint method from optimal control theory [48, 110]. We derive the adjoint equations associated to the Hedge dynamics and apply the approach to routing game examples: both a toy network to illustrate the qualitative behavior of the method, and a model of a real highway network to show the potential impact of this approach.

1.3 Accelerated dynamics for convex optimization

Convex optimization is an essential tool in many engineering, statistics, machine learning and economics problems, see for example [30] for a brief overview of some of these applications. First-order methods have seen a resurgence of interest due to the significant increase in both size and dimensionality of the data sets typically encountered in machine learning and other applications, which makes higher-order methods computationally intractable in most cases [103, 69, 33]. Many of these algorithms can be interpreted as a discretization of a continuous time ODE. For example, the mirror descent family for constrained convex optimization was originally derived by Nemirovski and Yudin [98] as a discretization of an ODE that was tailored to a specific Lyapunov function. Continuous-time dynamics for optimization have been studied for a long time, e.g. [32, 64, 26], and proving convergence results in continuous time often uses simple and elegant Lyapunov arguments. By discretizing the continuous dynamics, one can then design discrete algorithms for convex optimization, and to prove convergence in discrete time, one can attempt to use a discrete counterpart of the original Lyapunov function. Although it is hard to guarantee that the discretization will preserve the Lyapunov function, many such approaches have been successful. In particular, Su et al. show in [130] that Nesterov's accelerated method [102] can be obtained as a discretization of a a second-order ODE, for which they exhibit a Lyapunov function, and Attouch et al. [6] further study the properties of the its solutions trajectories and its convergence rates. This continuous-time interpretation also allowed the design of restarting heuristics, which empirically improve the speed of convergence, such as [105].

Our contributions

In the second part of this thesis, we study dynamics for constrained convex optimization, in continuous and discrete time. We start by reviewing the continuous-time interpretations of two important optimization methods: Nesterov's accelerated method, proposed by Nesterov in [102], and the mirror descent method, proposed by Nemirovski and Yudin [98]. We show in Chapter 8 that these two ideas can be combined to derive a general family of accelerated mirror descent dynamics for constrained optimization, using a simple Lyapunov argument.

This family generalizes the ODE studied by [6, 130], which only applies to unconstrained convex problems.

We show that the solution trajectories of the ODE converge to the set of minimizers of the objective function at a quadratic rate. We also show that the dynamics can be naturally described as a coupling of a dual variable that accumulates gradients with weights $\eta(t)$, and a primal variable obtained as the weighted average of the mirrored trajectory, using weights $w(t)$. This interpretation motivates the study of generalized averaging schemes in Chapter 9, in which we give sufficient conditions on the weight functions η and w to achieve a desired rate in continuous time. We also propose an adaptive averaging heuristic which adaptively computes the weights (instead of using a predefined weight function of time), essentially by reducing weights on portions of the trajectory that make the least progress, and show that this heuristic preserves the Lyapunov function of the accelerated dynamics, making it the first such heuristic with convergence guarantees.

In Chapter 10, we propose a discretization of the accelerated mirror descent ODE which has a quadratic convergence rate, and prove that a discrete version of the adaptive averaging heuristic also preserves the quadratic rate. We show several numerical examples on simplex-constrained problems to illustrate the qualitative behavior of these methods. In particular, we compare adaptive averaging to the restarting heuristics developed in [105, 130], and show that it compares favorably to restarting, with significant improvements in many cases.

1.4 Bibliographic notes

Most of the work reported in this thesis is adapted from previously published research. Chapters 2, 3 and 4 on the replicator dynamics and approximate replicator algorithms are based on [79, 80, 44]. Chapter 5 on stochastic mirror descent dynamics is based on [81, 76]. Chapter 6 on learning rate estimation is based on [84], Chapter 8 on accelerated mirror descent is based on [77], and portions of Chapter 9 and 10 on generalized and adaptive averaging are based on [78]. Finally, part of Appendix B is base on [77] and Appendix C is based on [82].

Part I

Online Learning Dynamics and Nonatomic Potential Games

Chapter 2

Online Learning in Convex Potential Games

2.1 Introduction

Nonatomic potential games are games that model the interaction of populations of players, and such that the set of players in each population is endowed with a measurable set structure with a nonatomic measure [124, 123]. One of the most well-studied families of nonatomic potential games are congestion games [73, 104], which motivate our results. These are non-cooperative games that model the interaction of players who share resources. Each player makes a decision on which resources to utilize. The individual decisions of players result in a resource allocation at the population scale. Resources which are highly utilized become congested, and the corresponding players incur higher losses. For example, the resources can be edges in a transportation or a communication network, and each player has a source vertex and a destination vertex on the graph, and needs to send traffic between the two. Each player chooses a path, and the joint decision of all players determines the congestion on each edge. The more a given edge is utilized, the more congested it is, creating delays for those players using that edge.

Congestion games and their equilibria have been studied in the transportation literature since the seminal work of Wardrop [134] and Beckman [16], and more recently in computer science, see [119] for a comprehensive introduction and related work. The set of Nash equilibria of the congestion game is known to coincide with the set of minimizers of a convex potential function. This was proved by Rosenthal for the atomic congestion game in [118], and later generalized. Thus computing the set of Nash equilibria can be done efficiently if one is given the exact formulation of the game, including the congestion functions of every resource, and the description of all populations. A natural generalization of the congestion game is given by convex potential games, in which the Rosenthal potential is generalized to any convex potential function.

Characterizing the Nash equilibria of potential games, and congestion games in particular,

gives useful insights, such as the loss of efficiency due to selfishness of players. One popular measure of inefficiency is the price of anarchy, introduced by Koutsoupias and Papadimitriou in [75], and studied in the case of congestion games by Roughgarden et al. in [122, 121]. Many approaches have been proposed since to alleviate the inefficiency of equilibria, either through incentivization [106] or by controlling a subset of the population [120].

Online learning dynamics While characterizing Nash equilibria of the game gives many insights, it does not model how players *arrive to the equilibrium*. Studying the game in a repeated setting can help answer this question. Additionally, most realistic scenarios do not correspond to a one-shot game, but rather a repeated setting in which each player faces a sequential decision problem, observes outcomes, and may update their strategies given the previous outcomes. This motivates the study of the game and the population dynamics in an online learning framework.

Arguably, a good model for learning should be distributed (no centralization between players), and should have realistic information requirements. For example in congestion games, one should not expect the players to have an accurate model of congestion of each resource. Players should be able to learn simply by observing the outcomes of their previous actions, and potentially those of other players. No-regret learning is of particular interest here, as many regret-minimizing algorithms are easy to implement by individual players, and only require the player losses to be revealed, see for example [40] and the references therein. The Hedge algorithm (also known as the multiplicative weights algorithm [4], or the exponentiated gradient method [72]) is a famous example of regret-minimizing algorithms. It was applied to learning in games by Freund and Schapire in [53]. The Hedge algorithm will be central in our discussion, as it will motivate the study of the continuous-time replicator equation in the next chapter.

Organization of Part I In this chapter, we start by formally defining nonatomic potential games and congestion games in Section 2.2. We give some preliminary results on the characterization of Nash equilibria as the set of solutions to a convex problem. We then define the online learning model in Section 2.4. We give a first convergence result in Section 2.5: we show in Theorem 2 that if the regret is sublinear for all populations, then the sequence of mass distributions converges, in the sense of Cesàro, to the set of Nash equilibria. We also show that as a consequence, a dense subsequence converges to the set of Nash equilibria. In Section 2.6, we review the Hedge algorithm for online learning, and some of its properties.

While our learning model is inherently discrete, it can be helpful to study continuous-time dynamics for learning, and to view discrete learning algorithms as a discretization of the continuous-time dynamics. In Chapter 3, we show that by taking the continuous-time limit of the Hedge algorithm, we obtain an ODE known as the replicator equation. We study properties of its stationary points, and show that all Nash equilibria of the game are stationary points (but the converse is not true in general), and show in Theorem 3 that solution trajectories converge to the set of Nash equilibria and derive an explicit rate of

convergence. We further study stability of stationary points by linearizing the dynamics: we show in Theorem 4 that all stable stationary points are Nash equilibria, and in Theorem 5, that under a non-degeneracy assumption, all Nash equilibria are exponentially stable.

In Chapter 4, we go back to discrete algorithms for online learning, and study a family of algorithms that can be obtained as a discretization of the replicator ODE. We first propose a deterministic discretization and prove that it guarantees sublinear regret in Theorem 6. Then using results from stochastic approximation theory, we show in Theorem 8 that a class of approximate replicator algorithms converges almost surely to the set of Nash equilibria.

While this guarantees convergence of a large family of algorithms, the stochastic approximation analysis does not provide convergence rates. In Chapter 5, we consider a different family of learning dynamics, obtained by applying the stochastic mirror descent method to the problem of minimizing the potential function of the game. In particular, we propose a heterogeneous formulation of the dynamics, in which different populations can use different algorithms and learning rates, and show that under mild assumptions on their learning rates, the sequence of their decisions is guaranteed to converge.

This defines a model of distributed learning, which enjoys several convergence guarantees. In Chapters 6 and 7, we propose and explore the approach of using mirror descent as a model of decision dynamics, in problems in which a coordinator interacts with a population of online learners. In Chapter 6, we propose a simple method to estimate the unknown learning rates of a decision maker who follows the Hedge dynamics, assuming that we can observe the sequence of decisions generated by the algorithm. We test this method using a web application, in which we simulate the routing game, and study the qualitative behavior of decision makers. We conclude the first part in Chapter 7, where we study a control problem, in which a coordinator can choose the decisions of a subset of the population, and the rest of the population is assumed to follow Hedge dynamics. This defines an optimal control problem under non-linear dynamics, which we propose to solve using different methods. In particular, we derive the adjoint equations of the Hedge dynamics, and show how the method can be applied to optimal routing on a transportation network.

2.2 Non atomic potential games and Nash equilibria

A nonatomic population game is given by a set $\mathcal{S}$ of players, endowed with a structure of measure space, $(\mathcal{S}, \Sigma, m)$, where Σ is a σ-algebra of measurable subsets, and m is a finite Lebesgue measure. The measure is non-atomic, in the sense that single-player sets are null-sets for m. The player set is partitioned into K populations, $\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_K$, such that the total mass $m(\mathcal{S}_k)$ is non-zero for all k. Each population is characterized by an action set, $\mathcal{A}_k$.

The joint actions of players in population k can be represented by an action profile $A_k : \mathcal{S}_k \to \mathcal{A}_k$, that specifies the action of each player. The function $s \mapsto A_k(s)$ is assumed to be $\mathcal{S}$-measurable ($\mathcal{A}_k$ is equipped with the counting measure). Given a joint action profile $A = (A_1, \ldots, A_K)$, a more concise, macroscopic description of the joint action of players is

given by the mass distribution, i.e. the proportion of players choosing action $a \in \mathcal{A}_k$, which we denote by

$$x_{k,a} = \frac{1}{m(\mathcal{S}_k)} \int_{s \in \mathcal{S}_k} 1_{A_k(s)=a} dm(s). \tag{2.1}$$

so that $x_k \in \Delta^{\mathcal{A}_k}$, the probability simplex over the action set

$$\Delta^{\mathcal{A}_k} = \{x \in \mathbb{R}_+^{\mathcal{A}_k} : \sum_{a \in \mathcal{A}_k} x_a = 1\}.$$

Note that x_k depends on A_k, but we keep this dependence implicit to simplify the notation. We denote by $x = (x_1, \ldots, x_K) \in \Delta^{\mathcal{S}_1} \times \cdots \times \Delta^{\mathcal{S}_K}$ the product mass distribution of all populations (which we also refer to as the joint mass distribution), and we will denote $\Delta = \Delta^{\mathcal{A}_1} \times \cdots \times \Delta^{\mathcal{A}_K}$ the product of simplices.

The joint mass distribution x determines the losses of all players as follows: for all k, we are given a vector valued function

$$\ell_k : \Delta \to \mathbb{R}^{\mathcal{A}_k},$$

such that $\ell_{k,a}(x)$ is the loss of action $a \in \mathcal{A}_k$, incurred by any player in population k who chooses action a. Finally, we denote by $\ell(x)$ the tuple $\ell(x) = (\ell_1(x), \ldots, \ell_K(x))$.

Definition 1. *A nonatomic game with action sets $\mathcal{A}_k$ and losses ℓ_k is a convex potential game if there exists a convex function f, differentiable on Δ with Lipschitz gradient, and positive reals $\kappa_1, \ldots, \kappa_K$, such that for all $x \in \Delta$ and all k,*

$$\nabla_{x_k} f(x) = \kappa_k \ell_k(x), \tag{2.2}$$

where $\nabla_{x_k} f(x)$ denotes the gradient of f with respect to x_k.

In other words, the loss functions of the game coincide (up to scaling by κ) with the gradient field of a convex function. In the remainder of the chapter, we will study such games. First, we define and characterize the Nash equilibria of nonatomic convex potential games.

Nash equilibria

Definition 2 (Nash equilibrium of a nonatomic convex potential game)**.**
A product distribution $x \in \Delta$ is a Nash equilibrium of the game if for all k, and all $a \in \mathcal{A}_k$ such that $x_{k,a} > 0$, $\ell_{k,a'}(x) \geq \ell_{k,a}(x)$ for all $a' \in \mathcal{A}_k$. The set of Nash equilibria will be denoted by $\mathcal{N}$.

Definition 2 implies that, for a population $\mathcal{S}_k$, all actions with non-zero mass have equal losses, and actions with zero mass have greater losses. Therefore almost all players incur the same loss.

In finite player games, a Nash equilibrium is defined to be an action profile $A : \mathcal{S} \to \mathcal{A}$ such that no player has an incentive to unilaterally deviate [95], that is, no player can strictly decrease her loss by unilaterally changing her action. We show that this condition (referred to as the Nash condition) holds for *almost all players* if and only if the mass distribution x induced by A is a Nash equilibrium in the sense of Definition 2.

Proposition 1. *A distribution x is a Nash equilibrium if and only if for any joint action profile A which induces the distribution x, almost all players have no incentive to unilaterally deviate from A.*

Proof. First, we observe that, given an action profile $A = (A_1, \dots, A_K)$, when a single player s changes her strategy, this does not affect the distribution x. This follows from the definition of the distribution, $x_{k,a} = \frac{1}{m(\mathcal{S}_k)} \int_{\mathcal{S}_k} 1_{A(s)=a} dm(s)$. Changing the action profile A on a null-set $\{s\}$ does not affect the integral.

Now, assume that almost all players have no incentive to unilaterally deviate. That is, for all k, for almost all $s \in \mathcal{S}_k$,

$$\forall a' \in \mathcal{A}_k,\ \ell_{k,a'}(x') \geq \ell_{A(s)}(x), \tag{2.3}$$

where x' is the distribution obtained when s unilaterally changes her action from $A(s)$ to a'. By the previous observation, $x' = x$. As a consequence, condition (2.3) becomes: for almost all s, and for all a', $\ell_{k,a'}(x) \geq \ell_{k,A(s)}(x)$. Therefore, integrating over the set $\{s \in \mathcal{S}_k : A(s) = a\}$, we have for all k,

$$\ell_{k,a'}(x) x_{k,a} \geq \ell_{k,a}(x) x_{k,a},\ \forall a'$$

which implies that x is a Nash equilibrium in the sense of Definition 2. Conversely, if A is an action profile, inducing distribution x, such that the Nash condition does not hold for a set of players with positive measure, then there exists k_0 and a subset $S \subset \mathcal{S}_{k_0}$ with $m(S) > 0$, such that every player in S can strictly decrease her loss by changing her action. Let $S_a = \{s \in S : A(s) = a\}$, then S is the disjoint union $S = \cup_{a \in \mathcal{A}_{k_0}} S_a$, and there exists a_0 such that $m(S_{a_0}) > 0$. Therefore

$$x_{k_0,a_0} = \frac{m\left(\{s \in \mathcal{S}_{k_0} : A(s) = a_0\}\right)}{m(\mathcal{S}_{k_0})} \geq \frac{m(S_{a_0})}{m(\mathcal{S}_{k_0})} > 0.$$

Let $s \in S_{a_0}$. Since s can strictly decrease her loss by unilaterally changing her action, there exists a_1 such that $\ell_{k_0,a_1}(x) < \ell_{k_0,A(s)}(x) = \ell_{k_0,a_0}(x)$. But since $x_{k_0,a_0} > 0$, x is not a Nash equilibrium. □

Next, we give a characterization of Nash equilibria in terms of the minimizers of the potential f.

Theorem 1. *$\mathcal{N}$ is the set of minimizers of f on the product of simplices Δ. It is a non-empty convex compact set. We denote by $f^\star$ the value of f on $\mathcal{N}$.*

Proof. First, observe that Definition 2 is equivalent to the following condition:

$$\begin{aligned} x \in \mathcal{N} &\Leftrightarrow \forall x' \in \Delta, \langle \ell_k(x), x'_k - x_k \rangle \geq 0, \ \forall k \\ &\Leftrightarrow \forall x' \in \Delta, \frac{1}{\kappa_k} \langle \nabla_{x_k} f(x), x'_k - x_k \rangle \geq 0, \ \forall k \\ &\Leftrightarrow \forall x' \in \Delta, \langle \nabla f(x), x' - x \rangle \geq 0, \end{aligned}$$

which corresponds to the first-order optimality conditions for minimizing the function f over Δ, see for example Section 3.1.3 in [30]. □

This characterization of Nash equilibria is useful since it allows one to compute an equilibrium by solving a convex optimization problem. It will also be useful in studying online learning dynamics both in continuous and discrete time.

Mixed strategies

The Nash equilibria we have described so far are *pure strategy* equilibria, since each player s deterministically plays a single action $A(s)$. We now extend the model to allow mixed strategies. That is, the action of a player s is a random variable $A(s)$ with distribution $\pi(s)$.

We show that when players use mixed strategies, provided they randomize independently, the resulting Nash equilibria are, in fact, the same as those given in Definition 2. The key observation is that under independent randomization, the resulting mass distributions x_k are random variables with zero variance, thus they are essentially deterministic.

To formalize the probabilistic setting, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A mixed strategy profile is given by the functions $A_k : \mathcal{S}_k \to (\Omega \to \mathcal{A}_k)$, assumed $\Sigma \times \mathcal{F}$-measurable. For all $s \in \mathcal{S}_k$ and $a \in \mathcal{A}_k$, let $\pi_{k,a}(s) = \mathbb{P}[A(s) = a]$. Similarly to the deterministic case, the mixed strategy profile A determines the distributions x_k, which are, in this case, random variables, given by $x_{k,a} = \frac{1}{m(\mathcal{S}_k)} \int_{\mathcal{S}_k} 1_{A(s)=a} dm(s)$.

Nevertheless, assuming players randomize independently, the mass distribution is almost surely equal to its expectation, as stated in the following proposition. The assumption of independent randomization is a reasonable one, since players are non-cooperative.

Proposition 2. *Under independent randomization,*

$$\forall k, \textit{ almost surely, } x_k = \mathbb{E}[x_k] = \frac{1}{m(\mathcal{S}_k)} \int_{\mathcal{S}_k} \pi_k(s) dm(s). \tag{2.4}$$

Proof. Fix k and let $a \in \mathcal{A}_k$. Since $(s, \omega) \mapsto 1_{A(s)=a}(\omega)$ is a non-negative bounded $\Sigma \times \mathcal{F}$-

measurable function, we can apply Tonelli's theorem and write:

$$\begin{aligned}\mathbb{E}\left[x_{k,a}\right] &= \mathbb{E}\left[\frac{1}{m(\mathcal{S}_k)}\int_{\mathcal{S}_k} 1_{A(s)=a} dm(s)\right] \\ &= \frac{1}{m(\mathcal{S}_k)}\int_{\mathcal{S}_k} \mathbb{E}\left[1_{A(s)=p}\right] dm(s) \\ &= \frac{1}{m(\mathcal{S}_k)}\int_{\mathcal{S}_k} \pi_{k,a}(s) dm(s).\end{aligned}$$

Similarly,

$$\begin{aligned}m(\mathcal{S}_k)^2 \operatorname{var}\left[x_{k,a}\right] &= \mathbb{E}\left(\int_{\mathcal{S}_k} 1_{A(s)=a} dm(s)\right)^2 - \left(\int_{\mathcal{S}_k} \pi_{k,a}(s) dm(s)\right)^2 \\ &= \int_{\mathcal{S}_k}\int_{\mathcal{S}_k} \mathbb{E}\, 1_{A(s)=a;A(s')=a} dm(s) dm(s') - \int_{\mathcal{S}_k}\int_{\mathcal{S}_k} \pi_{k,a}(s)\pi_{k,a}(s') dm(s) dm(s') \\ &= \int_{\mathcal{S}_k\times\mathcal{S}_k} \left(\mathbb{P}[A(s)=a; A(s')=a] - \pi_{k,a}(s)\pi_{k,a}(s')\right) d(m\times m)(s,s').\end{aligned}$$

Then observing that the diagonal $D = \{(s,s)\colon s \in \mathcal{S}_k\}$ is an $(m \times m)$-nullset (this follows for example from Proposition 251T in [52]), we can restrict the integral to the set $\mathcal{S}_k \times \mathcal{S}_k \setminus D$, on which $\mathbb{P}[A(s) = a; A(s') = s] = \pi_{k,a}(s)\pi_{k,a}(s')$, by the independent randomization assumption. This proves that $\operatorname{var}\left[x_{k,a}\right] = 0$. Therefore $x_{k,a} = \mathbb{E}\left[x_{k,a}\right]$ almost surely. □

2.3 Congestion games

In this section, we give an example of a nonatomic population game with a convex potential. To fully specify the game, we simply need to define the action set and the loss function of each population.

In the congestion game, a finite set $\mathcal{R}$ of resources is shared by the players. For each population k, the action set $\mathcal{A}_k$ is given by a collection of non-empty subsets of $\mathcal{R}$. Given a mass distribution $x \in \Delta$, we define, for all $r \in \mathcal{R}$, the resource load to be the total mass of players utilizing r:

$$\phi_r(x) = \sum_{k=1}^{K} m(\mathcal{S}_k) \sum_{a\in\mathcal{A}_k : r\in a} x_{k,a}. \tag{2.5}$$

Note that the vector of resource loads ϕ is a linear function of the distribution x, and can be written as

$$\phi(x) = \bar{M}x$$

where $\bar{M} = \left(\, m(\mathcal{S}_1)M_1 \,\middle|\, \ldots \,\middle|\, m(\mathcal{S}_K)M_K \,\right)$, and for each k, M_k is an incidence matrix given by

$$M_{k,(r,a)} = \begin{cases} 1 & \text{if } r \in a, \\ 0 & \text{otherwise.}\end{cases}$$

The resource loads determine the losses of all players as follows: the loss associated to a resource r is given by $c_r(\phi_r(x))$, where c_r are given congestion functions, assumed to satisfy the following:

Assumption 1. *The congestion functions c_r are non-negative, non-decreasing, Lipschitz-continuous functions.*

Then the loss of an action $a \in \mathcal{A}_k$ is the sum of the losses of resources in a, i.e.

$$\ell_{k,a}(x) = \sum_{r \in a} c_r(\phi_r(x)) = \sum_{r \in a} c_r((\bar{M}x)_r) = (M^\top c(\bar{M}x))_{k,a}, \tag{2.6}$$

where M is the incidence matrix $M = \begin{pmatrix} M_1 | & \dots & | M_K \end{pmatrix}$, and $c(\phi)$ is the vector $(c_r(\phi_r))_{r \in \mathcal{R}}$.

A motivating example: the routing game

A routing game is a special case of a congestion game, studied for example in [119]. The game has an underlying graph structure, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with vertex set $\mathcal{V}$ and edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. In this case, the resource set is equal to the edge set, $\mathcal{R} = \mathcal{E}$, and the actions are paths on the graph. Routing games are used to model congestion on transportation or communication networks. Each population $\mathcal{S}_k$ is characterized by a common origin vertex $o_k \in \mathcal{V}$ and a common destination vertex $d_k \in \mathcal{V}$. In a transportation setting, players represent drivers traveling from o_k to d_k; in a communication setting, players send packets from o_k to d_k. The action set $\mathcal{A}_k$ is a set of paths connecting o_k to d_k. In other words, each player chooses a path connecting his or her source and destination vertices. The mass of players $x_{k,a}$ can then be thought of as the total flow on path a, and the resource load $\phi_r(x)$ is the edge flow. Finally, the congestion functions $\phi_r \mapsto c_r(\phi_r)$ determine the delay (or latency) incurred by each player. The assumption that the delay function is increasing simply describes the intuitive fact that the more an edge is utilized, the more congested it becomes, and the more latency the players who use that edge incur. Finally, by Definition 2, a Nash equilibrium corresponds to a distribution x such that for each population S_k, all paths with non-zero mass have equal losses, and paths with zero mass have higher losses.

The Rosenthal potential function

We now exhibit a convex potential function for the congestion game. Consider the function

$$f^{\text{Rosenthal}}(x) = \sum_{r \in \mathcal{R}} \int_0^{(\bar{M}x)_r} c_r(u)du, \tag{2.7}$$

defined on the product of simplices $\Delta = \Delta^{\mathcal{A}_1} \times \cdots \times \Delta^{\mathcal{A}_K}$. $f^{\text{Rosenthal}}$ is called the Rosenthal potential function, and was introduced in [118] for the congestion game with finitely many players, and later generalized to the nonatomic case. It can be viewed as the composition of

the function $g : \phi \in \mathbb{R}_+^{\mathcal{R}} \mapsto \sum_{r\in\mathcal{R}} \int_0^{\phi_r} c_r(u)du$ and the linear function $x \mapsto \bar{M}x$. Since for all r, c_r is, by assumption, non-negative, g is differentiable, non-negative and $\nabla g(\phi) = (c_r(\phi_r))_{r\in\mathcal{R}}$. And since c_r are non-decreasing, g is convex. Therefore $f^{\text{Rosenthal}}$ is convex as the composition of a convex and a linear function.

A simple application of the chain rule gives $\nabla f^{\text{Rosenthal}}(x) = \bar{M}^\top c(\bar{M}x)$. Thus,

$$\forall k,\ \nabla_{x_k} f^{\text{Rosenthal}}(x) = m(\mathcal{S}_k) {M_k}^\top c(\bar{M}x) = m(\mathcal{S}_k)\ell_k(x),$$

where the last equality follows from Equation (2.6). Therefore $f^{\text{Rosenthal}}$ is a potential function for the congestion game, in the sense of Definition 1, with $\kappa_k = m(\mathcal{S}_k)$. By Theorem 1, the set of Nash equilibria of the congestion game (also called Wardrop equilibria in the transportation literature, in reference to [134]), coincides with the set of minimizers of $f^{\text{Rosenthal}}$ over Δ.

We also observe that when the congestion functions c_r are strictly increasing, the function g is strictly convex, and the set of minimizers has the following simple structure: $\mathcal{N} = \{x \in \Delta : \bar{M}x = \phi^\star\}$, where $\phi^\star$ is the unique solution to the problem

$$\begin{array}{ll} \text{minimize} & g(\phi) \\ \text{subject to} & \phi = \bar{M}x \\ & x \in \Delta, \end{array}$$

where uniqueness follows by strict convexity of g. Beyond computing Nash equilibria, we seek to study learning dynamics, which model how players arrive at the set $\mathcal{N}$. This is discussed in the next section.

2.4 The online learning model

We propose a model of repeated play, in which each player $s \in \mathcal{S}_k$ faces an online learning problem with full feedback, and applies an online learning algorithm, as defined below.

Online learning problem with full feedback

Given an action set $\mathcal{A}$, the online learning problem with loss sequence $(\ell^{(\tau)})$ consists in choosing, at each iteration τ, a probability distribution $\pi^{(\tau)} \in \Delta^{\mathcal{A}}$, sampling an action $A^{(\tau)} \sim \pi^{(\tau)}$, then observing the loss vector $\ell^{(\tau)}$. The loss incurred at iteration τ is then $\ell^{(\tau)}_{A^{(\tau)}}$, and the expected loss is $\left\langle \ell^{(\tau)}, \pi^{(\tau)} \right\rangle$.

Definition 3 (Online learning algorithm)**.** *Given an online learning problem with full feedback, an online learning algorithm is a sequence of functions indexed by τ, that we refer to as the update rules, that map the current distribution and the current loss vector to the next distribution*

$$U^{(\tau)} :\ \Delta^{\mathcal{A}} \times \mathbb{R}^{\mathcal{A}} \to \Delta^{\mathcal{A}}.$$

Note that this definition can be generalized, by making the update rule depend on the entire history of losses and previous distributions, but we refrain from making this generalization to simplify the discussion and the notation. The online learning framework is summarized in Algorithm 1 below.

Algorithm 1 Online learning problem with full feedback, on an action set $\mathcal{A}$ and with sequence of losses $(\ell^{(\tau)})$.

1: Input: Initial distribution $\pi^{(0)} \in \Delta^{\mathcal{A}}$ and learning algorithm $(U^{(\tau)})$.
2: **for** each iteration $\tau \in \mathbb{N}$ **do**
3: The player draws an action from $A^{(\tau)} \sim \pi^{(\tau)}$.
4: A vector of losses $\ell^{(\tau)}$ is revealed to the player, who incurs loss $\ell^{(\tau)}_{A^{(\tau)}}$.
5: The player updates
$$\pi^{(\tau+1)} = U^{(\tau)}(\pi^{(\tau)}, \ell^{(\tau)}).$$
6: **end for**

A natural measure of performance of online learning algorithms is given by the regret, which we define next. Since the game is played for infinitely many iterations, we may assume that the losses are discounted over time. This is a common technique in infinite-horizon optimal control for example, and can be motivated from an economic perspective by considering that losses are devalued over time.

Let $(\gamma_\tau)_{\tau\in\mathbb{N}}$ denote a sequence of discount factors (which can be constant, in which case the losses are not discounted), and which satisfies the following assumption.

Assumption 2. *The sequence of discount factors $(\gamma_\tau)_{\tau\in\mathbb{N}}$ is assumed to be positive non-increasing, and non-summable.*

A note on monotonicity of the discount factors A similar definition of discounted regret is used for example by Cesa-Bianchi and Lugosi in Section 3.2 of [40]. However, in their definition, the sequence of discount factors is *increasing*. This can be motivated by the following argument: present observations may provide better information than past, stale observations. While this argument is accurate in many applications, it does not serve our purpose of convergence of population strategies. In our discussion, the standing assumption is that discount factors are *non-increasing*.

On iteration τ, the player draws $A^{(\tau)} \sim \pi^{(\tau)}$ and incurs loss $\gamma_\tau \ell^{(\tau)}_{A^{(\tau)}}$. The cumulative discounted loss up to iteration T, is then defined to be

$$L^{(T)} = \sum_{\tau=0}^{T} \gamma_\tau \ell^{(\tau)}_{A^{(\tau)}}, \tag{2.8}$$

which is a random variable, since the action $A^{(\tau)}$ is random. Its expectation is

$$\mathbb{E}[L^{(T)}] = \sum_{\tau=0}^{T} \gamma_\tau \, \mathbb{E}\left[\ell^{(\tau)}_{A^{(\tau)}}\right] = \sum_{\tau=0}^{T} \gamma_\tau \left\langle \pi^{(\tau)}, \ell^{(\tau)} \right\rangle .$$

Similarly, we define the cumulative discounted loss for a fixed action $a \in \mathcal{A}_k$

$$\mathscr{L}_a^{(T)} = \sum_{\tau=0}^{T} \gamma_\tau \ell_a^{(\tau)}. \tag{2.9}$$

We can now define the discounted regret.

Definition 4 (Regret)**.** *Consider an online learning algorithm on an action set $\mathcal{A}$, with sequence of losses $(\ell^{(\tau)})$, and let $(\pi^{(\tau)})$ be the sequence of decisions generated by the algorithm. Then the discounted regret up to iteration T, is the random variable*

$$\begin{aligned} R^{(T)} &= \max_{a\in\mathcal{A}} \sum_{\tau=0}^{T} \gamma_\tau (\ell^{(\tau)}_{A^{(\tau)}} - \ell^{(\tau)}_a), \\ &= L^{(T)} - \min_{a\in\mathcal{A}} \mathscr{L}_a^{(T)}. \end{aligned} \tag{2.10}$$

Its expectation is given by

$$\mathbb{E}[R^{(T)}] = \max_{a\in\mathcal{A}} \sum_{\tau=0}^{T} \gamma_\tau (\left\langle \pi^{(\tau)}, \ell^{(\tau)} \right\rangle - \ell^{(\tau)}_a).$$

The algorithm $U^{(\tau)}$ is said to have sublinear discounted regret if, for any sequence of losses $(\ell^{(\tau)})$, and any initial strategy $\pi^{(0)}$,

$$\lim_{T\to\infty} \frac{\left[R^{(T)}\right]_+}{\sum_{\tau=0}^{T} \gamma_\tau} = 0 \text{ almost surely.} \tag{2.11}$$

where x_+ denotes the positive part of x. If the condition holds for $\mathbb{E}\left[R^{(T)}\right]$, we say that the algorithm has sublinear discounted regret in expectation*.*

We observe that, in the definition of the regret, one can replace the minimum over the set $\mathcal{A}$ by a minimum over the simplex $\Delta^{\mathcal{A}}$, $\min_{a\in\mathcal{A}} \mathscr{L}_a^{(T)} = \min_{\pi\in\Delta^{\mathcal{A}}} \left\langle \pi, \mathscr{L}^{(T)} \right\rangle$, since the minimizers of a bounded linear function on a polytope lie on the set of its extremal points. Therefore, the discounted regret compares the performance of the online learning algorithm to the *best constant strategy in hindsight*. If the algorithm has sublinear regret, its average performance is, asymptotically, as good as the performance of any constant strategy, regardless of the sequence of losses $(\ell^{(\tau)})$.

Online learning in the nonatomic game

We assume that each player $s \in \mathcal{S}_k$ faces the online learning problem on $\mathcal{A}_k$, with losses given by $\ell_k(x^{(\tau)})$, and follows an online learning algorithm $(U_s^{(\tau)})$. In other words, each player solves a sequential decision problem, and the problems are coupled through the mass distribution $x^{(\tau)}$, which is determined by the joint decision of all players.

The decision of all players can be represented, as defined above, by functions $\pi_k^{(\tau)} : \mathcal{S} \to \Delta^{\mathcal{A}_k}$, such that for each player $s \in \mathcal{S}_k$, $\pi_k^{(\tau)}(s)$ is a probability distribution over $\mathcal{A}_k$, and players randomize independently by drawing an action $A_k^{(\tau)}(s)$ from $\pi_k^{(\tau)}(s)$. As discussed in Section 2.2, this induces, at the level of each population $\mathcal{S}_k$, a mass distribution $x_k^{(\tau)}$, a random variable with zero variance and expectation given by the integral (2.4),

$$x_k^{(\tau)} = \frac{1}{m(\mathcal{S}_k)} \int_{\mathcal{S}_k} \pi_k(s) dm(s), \text{ a.s.}$$

These, in turn, determine losses $\ell_k(x^{(\tau)})$, which are revealed to all players in population $\mathcal{S}_k$, and this marks the end of iteration τ. Players can then use this information to update their strategies using the update rule of their learning algorithm. The online learning framework is summarized in Algorithm 2.

Algorithm 2 Online learning in the nonatomic, convex potential game

1: Input: For every player $s \in \mathcal{A}_k$, an initial mixed strategy $\pi_k^{(0)}(s) \in \Delta^{\mathcal{A}_k}$ and an online learning algorithm $(U_s^{(\tau)})$.
2: **for** each iteration $\tau \in \mathbb{N}$ **do**
3: For all k, each player $s \in \mathcal{A}_k$ independently draws an action from $\pi_k^{(\tau)}(s)$. This determines the mass distribution $x^{(\tau)}$.
4: The vector of losses $\ell_k(x^{(\tau)})$ is revealed to players in $\mathcal{S}_k$.
5: Players update their mixed strategies:

$$\pi_k^{(\tau+1)}(s) = U_s^{(\tau)}(\pi_k^{(\tau)}(s), \ell_k(x^{(\tau)})).$$

6: **end for**

Population-wide regret

Let $L^{(T)}(s)$ and $R^{(T)}(s)$ denote the discounted cumulative loss and regret of player s, respectively. In order to analyze the population dynamics, we define a population-wide cumulative discounted loss $L_k^{(T)}$, and discounted regret $R_k^{(T)}$ as follows:

$$L_k^{(T)} = \frac{1}{m(\mathcal{S}_k)} \int_{\mathcal{S}_k} L^{(T)}(s) dm(s), \tag{2.12}$$

$$R_k^{(T)} = \frac{1}{m(\mathcal{S}_k)} \int_{\mathcal{S}_k} R^{(T)}(s) dm(s) = L_k^{(T)} - \min_{a \in \mathcal{A}_k} \mathscr{L}_{k,a}^{(T)}. \tag{2.13}$$

Since $L^{(T)}(s)$ is random, $L_k^{(T)}$ is also a random variable. However, it is, in fact, almost surely equal to its expectation. Indeed, recalling that $x_{k,a}^{(\tau)}$ is the proportion of players who chose action a at iteration τ, we can write

$$\begin{aligned} L_k^{(T)} &= \sum_{\tau=0}^{T} \gamma_\tau \frac{1}{m(\mathcal{S}_k)} \sum_{a \in \mathcal{A}_k} \int_{\{s \in \mathcal{S}_k : A^{(\tau)}(s) = a\}} \ell_{k,a}(x^{(\tau)}) dm(s) \\ &= \sum_{\tau=0}^{T} \gamma_\tau \sum_{a \in \mathcal{A}_k} x_{k,a}^{(\tau)} \ell_{k,a}(x^{(\tau)}) \\ &= \sum_{\tau=0}^{T} \gamma_\tau \left\langle x_k^{(\tau)}, \ell_k(x^{(\tau)}) \right\rangle. \end{aligned}$$

Thus, assuming players randomize independently, $x^{(\tau)}$ is almost surely deterministic by Proposition 2, and so is $L_k^{(T)}$. The same holds for $R_k^{(T)}$.

Next, we show that if the individual regrets are sublinear in expectation, then the population regrets are sublinear. This relies on the following observation: By Definition 1 of the potential game, the losses coincide with a gradient which is assumed to be Lipschitz. Thus the losses are continuous functions on the compact set Δ, thus bounded. Let $\rho > 0$ such that for all k, all $a \in \mathcal{A}_k$ and all $x \in \Delta$, $\ell_{k,a}(x) \in [0, \rho]$. Then it is straightforward to show the following.

Proposition 3. *For all k and all $s \in \mathcal{S}_k$, $\frac{L^{(T)}(s)}{\sum_{\tau=0}^{T} \gamma_\tau} \in [0, \rho]$ and $\frac{\left[R^{(T)}(s)\right]_+}{\sum_{\tau=0}^{T} \gamma_\tau} \in [0, \rho]$.*

Proposition 4. *If almost every player $s \in \mathcal{S}_k$ applies an online learning algorithm with sublinear regret in expectation, then the population-wide regret $R_k^{(T)}$ is also sublinear.*

Proof. By the previous observation, we have, almost surely,

$$R_k^{(T)} = \mathbb{E}\left[R_k^{(T)}\right] = \frac{1}{m(\mathcal{S}_k)} \int_{\mathcal{S}_k} \mathbb{E}\left[R^{(T)}(s)\right] dm(s),$$

where the second equality follows from Tonelli's theorem. Taking the positive part and using Jensen's inequality, we have

$$\frac{1}{\sum_{\tau=0}^{T} \gamma_\tau} \left[R_k^{(T)}\right]_+ \leq \frac{1}{m(\mathcal{S}_k)} \int_{\mathcal{S}_k} \frac{1}{\sum_{\tau=0}^{T} \gamma_\tau} \left[\mathbb{E}\left[R^{(T)}(s)\right]\right]_+ dm(s).$$

By assumption, $\frac{\left[\mathbb{E}\left[R^{(T)}(s)\right]\right]_+}{\sum_{\tau=0}^{T} \gamma_\tau}$ converges to 0 for all s, and by Proposition 3, it is bounded uniformly in s. Thus the result follows by applying the dominated convergence theorem. □

In the next section, we provide a first convergence guarantee of the sequence of mass distributions, when the population regret is sublinear.

2.5 Convergence of sublinear regret dynamics in the sense of Cesàro

As discussed in Proposition 4, if almost every player applies an algorithm with sublinear discounted regret in expectation, then the population-wide discounted regret is sublinear (almost surely). We now show that under these conditions, the sequence of distributions $(x^{(\tau)})$ converges in the sense of Cesàro. That is, $\sum_{\tau\leq T}\gamma_\tau x^{(\tau)}/\sum_{\tau\leq T}\gamma_\tau$ converges to the set of Nash equilibria. We also show that we have convergence of a dense subsequence, under certain conditions on the discount sequence (γ_τ). First, we give some definitions.

Definition 5 (Convergence in the sense of Cesàro)**.** *Fix a sequence of positive weights $(\gamma_\tau)_{\tau\in\mathbb{N}}$. A sequence $(u^{(\tau)})$ of elements of a normed vector space $(F,\|\cdot\|)$ converges to $u\in F$ in the sense of Cesàro with respect to (γ_τ) if*

$$\lim_{T\to\infty}\frac{\sum_{\tau\in\mathbb{N}:\tau\leq T}\gamma_\tau u^{(\tau)}}{\sum_{\tau\in\mathbb{N}:\tau\leq T}\gamma_\tau}=u.$$

We write $u^{(\tau)}\xrightarrow{(\gamma_\tau)}u$.

The Stolz-Cesàro theorem states that if $(u^{(\tau)})_\tau$ converges to u, then it converges in the sense of Cesàro with respect to any non-summable sequence $(\gamma_\tau)_\tau$, see for example [94]. The converse is not true in general. However, if a sequence converges *absolutely* in the sense of Cesàro, i.e. $\|u^{(\tau)}-u\|\xrightarrow{(\gamma_\tau)}0$, then we can show that a dense subsequence of $(u^{(\tau)})_\tau$ converges to u. To prove this, we first show that absolute Cesàro convergence implies statistical convergence, in the sense defined below.

Definition 6 (Statistical convergence)**.** *Fix a sequence of positive weights $(\gamma_\tau)_\tau$. A sequence $(u^{(\tau)})_{\tau\in\mathbb{N}}$ of elements of a normed vector space $(F,\|\cdot\|)$ converges to $u\in F$ statistically with respect to (γ_τ) if for all $\epsilon>0$, the set of indices $\mathcal{I}_\epsilon=\{\tau\in\mathbb{N}\colon\|u^{(\tau)}-u\|\geq\epsilon\}$ has zero density with respect to (γ_τ). The density of a subset of integers $\mathcal{I}\subset\mathbb{N}$, with respect to (γ_τ), is defined to be the limit, if it exists*

$$\lim_{T\to\infty}\frac{\sum_{\tau\in\mathcal{I}:\tau\leq T}\gamma_\tau}{\sum_{\tau\in\mathbb{N}:\tau\leq T}\gamma_\tau}.$$

Lemma 1. *If $(u^{(\tau)})_\tau$ converges to u absolutely in the sense of Cesàro with respect to (γ_τ), then it converges to u statistically with respect to (γ_τ).*

Proof. Let $\epsilon>0$. We have for all $T\in\mathbb{N}$,

$$0\leq\frac{\sum_{\tau\in\mathcal{I}_\epsilon\colon\tau\leq T}\gamma_\tau\epsilon}{\sum_{\tau\in\mathbb{N}\colon\tau\leq T}\gamma_\tau}\leq\frac{\sum_{\tau\in\mathbb{N}:\tau\leq T}\gamma_\tau\|u^{(\tau)}-u\|}{\sum_{\tau\in\mathbb{N}:\tau\leq T}\gamma_\tau},$$

which converges to 0 since $(u^{(\tau)})_\tau$ converges to u absolutely in the sense of Cesàro. Therefore $\mathcal{I}_\epsilon$ has zero density for all ϵ. □

We can now show convergence of a dense subsequence.

Proposition 5. *If $(u^{(\tau)})_{\tau\in\mathbb{N}}$ converges to u absolutely in the sense of Cesàro with respect to (γ_τ), then there exists a subset of indices $\mathcal{T}\subset\mathbb{N}$ of density one, such that the subsequence $(u^{(\tau)})_{\tau\in\mathcal{T}}$ converges to u.*

Proof. By Lemma 1, for all $\epsilon>0$, the set $\mathcal{I}_\epsilon=\{\tau\in\mathbb{N}\colon \|u^{(\tau)}-u\|\geq\epsilon\}$ has zero density. We will construct a set $\mathcal{I}\subset\mathbb{N}$ of zero density, such that the subsequence $(u_\tau)_{\tau\in\mathbb{N}\setminus\mathcal{I}}$ converges. For all $k\in\mathbb{N}^*$, let $p_k(T)=\sum_{\tau\in\mathcal{I}_{\frac{1}{k}}\colon \tau\leq T}\gamma_\tau$. Since $\frac{p_k(T)}{\sum_{\tau\in\mathbb{N}\colon \tau\leq T}\gamma_\tau}$ converges to 0 as $T\to\infty$, there exists $T_k>0$ such that for all $T\geq T_k$, $\frac{p_k(T)}{\sum_{\tau\in\mathbb{N}\colon \tau\leq T}\gamma_\tau}\leq\frac{1}{k}$. Without loss of generality, we can assume that $(T_k)_{k\in\mathbb{N}^*}$ is increasing. Now, let $\mathcal{I}=\bigcup_{k\in\mathbb{N}^*}(\mathcal{I}_{\frac{1}{k}}\cap\{T_k,\dots,T_{k+1}-1\})$. Then we have for all $k\in\mathbb{N}^*$, $\mathcal{I}\cap\{0,\dots,T_{k+1}-1\}=\left(\cup_{j=1}^k\mathcal{I}_{\frac{1}{j}}\right)\cap\{0,\dots,T_{k+1}-1\}$. But since $\mathcal{I}_1\subset\mathcal{I}_{\frac{1}{2}}\subset\cdots\subset\mathcal{I}_{\frac{1}{k}}$, we have $\mathcal{I}\cap\{0,\dots,T_{k+1}-1\}\subset\mathcal{I}_{\frac{1}{k}}\cap\{0,\dots,T_{k+1}-1\}$, thus for all T such that $T_k\leq T<T_{k+1}$, we have

$$\frac{\sum_{\tau\in\mathcal{I}\colon \tau\leq T}\gamma_\tau}{\sum_{\tau\in\mathbb{N}\colon \tau\leq T}\gamma_\tau}\leq\frac{\sum_{\tau\in\mathcal{I}_{\frac{1}{k}}\colon \tau\leq T}\gamma_\tau}{\sum_{\tau\in\mathbb{N}\colon \tau\leq T}\gamma_\tau}=\frac{p_k(T)}{\sum_{\tau\in\mathbb{N}\colon \tau\leq T}\gamma_\tau}\leq\frac{1}{k},$$

which proves that $\mathcal{I}$ has zero density.

Let $\mathcal{T}=\mathbb{N}\setminus\mathcal{I}$. We have that $\mathcal{T}$ has density one, and it remains to prove that the subsequence $(u^{(\tau)})_{\tau\in\mathcal{T}}$ converges to u. Since $\mathcal{T}$ has density one, it has infinitely many elements, and for all k, there exists $S_k\in\mathcal{T}$ such that $S_k\geq T_k$. For all $\tau\in\mathcal{T}$ with $\tau\geq S_k$, there exists $k'\geq k$ such that $T_{k'}\leq\tau<T_{k'+1}$. Since $\tau\notin\mathcal{I}$ and $T_{k'}\leq\tau<T_{k'+1}$, we must have $\tau\notin\mathcal{I}_{\frac{1}{k'}}$, therefore $\|u^{(\tau)}-u\|<\frac{1}{k'}\leq\frac{1}{k}$. This proves that $(u^{(\tau)})_{\tau\in\mathcal{T}}$ converges to u. □

We now present the main result of this section, which concerns the convergence of a subsequence of population distributions $(x^{(\tau)})$ to the set $\mathcal{N}$ of Nash equilibria. We say that $(x^{(\tau)})$ converges to $\mathcal{N}$ if $d(x^{(\tau)},\mathcal{N})\to0$, where $d(x,\mathcal{N})=\inf_{\nu\in\mathcal{N}}\|x-\nu\|$.

Theorem 2. *Consider a congestion game with discount factors $(\gamma_\tau)_\tau$ satisfying Assumption 2. Assume that for all $k\in\{1,\dots,K\}$, population k has sublinear discounted regret. Then the sequence of distributions $(x^{(\tau)})_\tau$ converges to the set of Nash equilibria in the sense of Cesàro with respect to (γ_τ). Furthermore, there exists a dense subsequence $(x_\tau)_{\tau\in\mathcal{T}}$ which converges to $\mathcal{N}$.*

To prove the theorem, we will use the following fact:

Lemma 2. *A sequence $(\nu^{(\tau)})$ in Δ converges to $\mathcal{N}$ only if $(f(\nu^{(\tau)}))$ converges to $f^\star$, the value of f on $\mathcal{N}$.*

Proof. Indeed, suppose by contradiction that $f(\nu^{(\tau)})\to f^\star$ but $\nu^{(\tau)}\not\to\mathcal{N}$. Then there would exist $\epsilon>0$ and a subsequence $(\nu^{(\tau)})_{\tau\in\mathcal{T}}$, $\mathcal{T}\subset\mathbb{N}$ such that $d(\nu^{(\tau)},\mathcal{N})\geq\epsilon$ for all $\tau\in\mathcal{T}$. Since Δ is compact, we can extract a further subsequence $(\nu^{(\tau)})_{\tau\in\mathcal{T}'}$ which converges to some $\nu\notin\mathcal{N}$. But by continuity of f, $(f(\nu^{(\tau)}))_{\tau\in\mathcal{T}'}$ converges to $f(\nu)>f^\star$, a contradiction. □

Proof of Theorem 2. Consider the potential function f. By convexity of f and the expression (2.2) of its gradient, we have for all τ and for all $x \in \Delta$:

$$f(x^{(\tau)}) - f(x) \leq \left\langle \nabla f(x^{(\tau)}), x^{(\tau)} - x \right\rangle = \sum_{k=1}^{K} \kappa_k \left\langle \ell_k(x^{(\tau)}), x_{k,a}^{(\tau)} - x_{k,a} \right\rangle,$$

then taking the weighted sum up to iteration T,

$$\begin{aligned}
\sum_{\tau=0}^{T} \gamma_\tau (f(x^{(\tau)}) - f(x)) &\leq \sum_{k=1}^{K} \kappa_k \left[\sum_{\tau=0}^{T} \gamma_\tau \left\langle x_k^{(\tau)}, \ell_k(x^{(\tau)}) \right\rangle - \left\langle x_k, \sum_{\tau=0}^{T} \gamma_\tau \ell_k(x^{(\tau)}) \right\rangle \right] \\
&= \sum_{k=1}^{K} \kappa_k \left[L_k^{(T)} - \left\langle x_k, \mathscr{L}_k^{(T)} \right\rangle \right] \\
&\leq \sum_{k=1}^{K} \kappa_k R_k^{(T)},
\end{aligned}$$

where for the last inequality, we use the fact that $\left\langle x_k, \mathscr{L}_k^{(T)} \right\rangle \geq \min_{a \in \mathcal{A}_k} \mathscr{L}_{k,a}^{(T)}$. In particular, when x is a Nash equilibrium, by Theorem 1, $f(x) = \min_{x \in \Delta^{\mathcal{A}_1} \times \cdots \times \Delta^{\mathcal{A}_K}} f(x) = f^\star$, thus

$$\frac{\sum_{\tau=0}^{T} \gamma_\tau |f(x^{(\tau)}) - f^\star|}{\sum_{\tau=0}^{T} \gamma_\tau} \leq \sum_{k=1}^{K} \kappa_k \frac{R_k^{(T)}}{\sum_{\tau=0}^{T} \gamma_\tau}.$$

Since the population-wide regret $R_k^{(T)}$ is assumed to be sublinear for all k, we have $|f(x^{(\tau)}) - f^\star| \xrightarrow{(\gamma_\tau)} 0$. By Proposition 5, there exists $\mathcal{T} \subset \mathbb{N}$ of density one, such that $(f(x^{(\tau)}))_{\tau \in \mathcal{T}}$ converges to $f^\star$. And it follows that $(x^{(\tau)})_{\tau \in \mathcal{T}}$ converges to $\mathcal{N}$. This proves the second part of the theorem. To prove the first part, we observe that, by convexity of f,

$$f^\star \leq f\left(\frac{\sum_{\tau=0}^{T} \gamma_\tau x^{(\tau)}}{\sum_{\tau=0}^{T} \gamma_\tau} \right) \leq \frac{\sum_{\tau=0}^{T} \gamma_\tau f(x^{(\tau)})}{\sum_{\tau=0}^{T} \gamma_\tau} = f^\star + \frac{\sum_{\tau=0}^{T} \gamma_\tau (f(x^{(\tau)}) - f^\star)}{\sum_{\tau=0}^{T} \gamma_\tau},$$

and the upper bound converges to $f^\star$. Therefore $\left(\frac{\sum_{\tau \leq T} \gamma_\tau x^{(\tau)}}{\sum_{\tau \leq T} \gamma_\tau} \right)_T$ converges to $\mathcal{N}$. □

2.6 The Hedge algorithm

We now present the Hedge algorithm, one example of online learning algorithms with sublinear regret. It is also known as the multiplicative weights algorithm [4], and as the exponentiated gradient descent [72] or the entropic mirror descent algorithm [15] in the convex optimization literature. It is also studied in the economics and game theory literature and is usually referred to as log-linear learning [28, 91]. We will use the Hedge algorithm to motivate the study of the continuous time replicator equation in the next chapter.

Definition 7 (Hedge algorithm)**.** *Consider an online learning problem on an action set $\mathcal{A}$ with loss functions $(\ell^{(\tau)})$. The Hedge algorithm with initial distribution $\pi^{(0)} \in \Delta^{\mathcal{A}}$ and learning rates $(\eta_\tau)_{\tau\in\mathbb{N}}$ is an online learning algorithm $(U^{(\tau)})$ such that the τ-th update function is given by*

$$\pi^{(\tau+1)} = U^{(\tau)}(\pi^{(\tau)}, \ell^{(\tau)}) \propto \left(\pi_a^{(\tau)} \exp\left(-\eta_\tau \ell_a^{(\tau)}\right)\right)_{a\in\mathcal{A}} \tag{2.14}$$

Algorithm 3 Hedge algorithm with learning rates (η_t).

1: Input: Initial distribution $\pi^{(0)} \in \Delta^{\mathcal{A}}$.
2: **for** each iteration $\tau \in \mathbb{N}$ **do**
3: Draw an action $A^{(\tau)} \sim \pi^{(\tau)}$.
4: Observe a vector of losses $\ell^{(\tau)}$, incur loss $\ell^{(\tau)}_{A^{(\tau)}}$.
5: Update

$$\pi^{(\tau+1)} \propto \left(\pi_a^{(\tau)} \exp\left(-\eta_\tau \ell_a^{(\tau)}\right)\right)_{a\in\mathcal{A}}$$

6: **end for**

Intuitively, the Hedge algorithm updates the distribution by computing, at each iteration, a set of action weights, then normalizing the vector of weights. The weight of an action a is obtained by multiplying its probability at the previous iteration $\pi_a^{(\tau)}$, by a term which is exponentially decreasing in $\ell_a^{(\tau)}$, the loss of action a. Thus, the higher the loss of a at iteration τ, the lower the probability of selecting a at the next iteration. The parameter η_τ can be interpreted as a learning rate. As $\eta_\tau \to 0$, $\pi^{(\tau+1)}$ tends to $\pi^{(\tau)}$, and as $\eta_\tau \to \infty$, $\pi^{(\tau+1)}$ puts all probability mass on $\arg\min_{a\in\mathcal{A}} \ell_a^{(\tau)}$. The Hedge algorithm is discussed in more detail in Appendix B: it is shown in Section B.6 to be an instance of the mirror descent method.

Remark 1. *The sequence of distributions given by the Hedge algorithm also satisfies*

$$\pi^{(\tau+1)} \propto \left(\pi_a^{(0)} \exp - \sum_{t=0}^{\tau} \eta_t \ell_a^{(t)}\right)_{a\in\mathcal{A}} \tag{2.15}$$

This follows from the update equation (2.14) and a simple induction on τ. In particular, when $\eta_\tau = \gamma_\tau$, the term $\sum_{t=0}^{\tau} \eta_t \ell_a^{(t)}$ coincides with the cumulative discounted loss $\mathscr{L}_a^{(\tau)}$ defined in (2.9). This motivates using the discount factors γ_τ as learning rates. We discuss this in the next proposition.

Proposition 6. *Consider an online learning problem with a sequence of discount factors $(\gamma_\tau)_{\tau\in\mathbb{N}}$ satisfying Assumption 2, and suppose that the losses $(\ell^{(\tau)})$ are in $[0, \rho]$, uniformly in τ. Then the Hedge algorithm with learning rates $\eta_\tau = \frac{\gamma_\tau}{\rho}$ satisfies the following regret bound: for any sequence of losses $(\ell^{(\tau)})$ and any initial strategy $\pi^{(0)}$,*

$$\mathbb{E}[R^{(T)}] \leq -\rho \log \pi^{(0)}_{\min} + \frac{\rho}{8} \sum_{\tau=0}^{T} \gamma_\tau^2,$$

where $\pi^{(0)}_{\min} = \min_{a\in\mathcal{A}} \pi^{(0)}_a$.

Proof. Given an initial strategy $\pi^{(0)}$, define $\xi : \mathbb{R}^{\mathcal{A}} \to \mathbb{R}, u \mapsto \log\left(\sum_{a\in\mathcal{A}} \pi^{(0)}_a \exp(-u_a/\rho)\right)$. Recalling the expression of the cumulative action loss $\mathscr{L}^{(\tau)}_a = \sum_{t=0}^{\tau} \gamma_t \ell^{(t)}_a$, we have for all $\tau \geq 0$:

$$\begin{aligned}
\xi(\mathscr{L}^{(\tau+1)}) - \xi(\mathscr{L}^{(\tau)}) &= \log\left(\sum_{a\in\mathcal{A}} \frac{\pi^{(0)}_a \exp\left(-\mathscr{L}^{(\tau)}_a/\rho\right)}{\sum_{a'\in\mathcal{A}} \exp\left(-\mathscr{L}^{(\tau)}_{a'}/\rho\right)} \exp\left(-\gamma_{\tau+1}\ell^{(\tau+1)}_a/\rho\right)\right) \\
&= \log\left(\sum_{a\in\mathcal{A}} \pi^{(\tau+1)}_a \exp\left(-\gamma_{\tau+1}\ell_a(x^{(\tau+1)})/\rho\right)\right) \\
&\leq -\gamma_{\tau+1} \sum_{a\in\mathcal{A}} \pi^{(\tau+1)}_a \frac{\ell_a(x^{(\tau+1)})}{\rho} + \frac{\gamma^2_{\tau+1}}{8}
\end{aligned}$$

The last inequality follows from Hoeffding's lemma, since $0 \leq \ell^{(\tau)}_a/\rho \leq 1$. Summing over $\tau \in \{-1, \dots, T-1\}$, we have for all a:

$$\xi(\mathscr{L}^{(T)}) - \xi(\mathscr{L}^{(-1)}) \leq -\sum_{\tau=0}^{T} \gamma_\tau \sum_{a\in\mathcal{A}} \pi^{(\tau)}_a \frac{\ell^{(\tau)}_a}{\rho} + \frac{1}{8}\sum_{\tau=0}^{T} \gamma^2_\tau$$

where $\xi(\mathscr{L}^{(-1)}) = \xi(0) = 0$. By monotonicity of the log function, we have for all $a_0 \in \mathcal{A}$, $\log(\pi^{(0)}_{a_0} \exp(-\mathscr{L}^{(T)}_{a_0}/\rho)) \leq \xi(\mathscr{L}^{(T)})$, thus

$$-\frac{\mathscr{L}^{(\tau)}_{a_0}}{\rho} + \log \pi^{(0)}_{a_0} \leq \xi(\mathscr{L}^{(T)}) \leq -\sum_{\tau=0}^{T} \gamma_\tau \sum_{a\in\mathcal{A}} \pi^{(\tau)}_a \frac{\ell^{(\tau)}_a}{\rho} + \frac{1}{8}\sum_{\tau=0}^{T} \gamma^2_\tau.$$

Rearranging, we have for all $a \in \mathcal{A}$

$$\sum_{\tau=0}^{T} \gamma_\tau \sum_{a\in\mathcal{A}} \pi^{(\tau)}_a \ell^{(\tau)}_a - \mathscr{L}^{(T)}_{a_0} \leq -\frac{\rho}{8} \log \pi^{(0)}_{a_0} + \rho \sum_{\tau=0}^{T} \gamma^2_\tau,$$

and we obtain the desired inequality by maximizing both sides over $a_0 \in \mathcal{A}$. □

The previous proposition provides an upper-bound on the expected regret of the Hedge algorithm, of the form

$$\frac{\mathbb{E}\left[R^{(T)}\right]}{\sum_{\tau\leq T} \gamma_\tau} \leq -\rho \log \pi^{(0)}_{\min} \frac{1}{\sum_{\tau\leq T} \gamma_\tau} + \frac{\rho}{8} \frac{\sum_{\tau\leq T} \gamma^2_\tau}{\sum_{\tau\leq T} \gamma_\tau}.$$

In particular, if the discount factors (γ_τ) satisfy $\lim_{T\to\infty} \frac{\sum_{\tau\leq T} \gamma^2_\tau}{\sum_{\tau\leq T} \gamma_\tau} = 0$, this proves that the discounted regret is sub-linear. This also provides a bound on the convergence rate. For

example, if $\gamma_\tau = \frac{1}{\tau}$, then the upper-bound is $\mathcal{O}\left(\frac{1}{\log T}\right)$, which converges to zero as $T \to \infty$, albeit slowly. A better bound can be obtained for sequences of discount factors which are not square-summable, for example, taking $\gamma_\tau \sim \sqrt{\frac{\log \tau}{\tau}}$, the upper-bound is $\mathcal{O}\left(\sqrt{\frac{\log T}{T}}\right)$.

We now have one example of an online learning algorithm with sublinear discounted regret. In the next chapter, we will study a continuous-time limit of the Hedge algorithm, and show that this results in an ODE, and study its stationary points, their stability, and their relationship to the set of Nash equilibria of the game.

Chapter 3

Replicator dynamics in convex potential games

The replicator dynamics is a continuous-time ODE that describes the evolution of a probability distribution, and that has been used to model the dynamics of populations of players in evolutionary game theory [67, 135, 66]. It has a long list of applications in biological and ecological systems, see [128] for a survey, and has also been studied in the context of viability theory [8], since it provides an elementary example of viability on the probability simplex.

We first show that the replicator ODE can be obtained as a continuous-time limit of the Hedge learning algorithm applied to the potential game defined in the previous chapter. Then we study the properties of its solution trajectories and its stationary points, and exhibit, in Section 3.3, several Lyapunov functions for different invariant sets. In particular, we prove in Theorem 3 that the solutions converge to the set of Nash equilibria of the game, and give a rate of convergence. Then, by studying the spectrum of the linearized system around stationary points, we show that any stationary point that is not a Nash equilibrium is unstable (Theorem 4), and under a strict monotonicity assumption, that Nash equilibria are exponentially stable (Theorem 5). We illustrate these results on a congestion game example.

In the next chapter, we will go back to studying discrete-time learning algorithms, by discretizing the solution trajectories of the replicator ODE. In particular, we will use the properties of continuous-time solutions derived in this chapter, together with results from stochastic approximation theory, to prove convergence of the discretized dynamics.

3.1 The replicator ODE as a continuous-time limit of the Hedge algorithm

To motivate the study of the replicator dynamics from an online learning point of view, we first derive the continuous-time replicator dynamics as a limit of the discrete Hedge dynamics, as discussed below. Consider a nonatomic game with a convex potential f as in Definition 1, and suppose that in each population $\mathcal{S}_k$, all players start from the same initial

distribution $\pi_k^{(0)} \in \Delta^{\mathcal{A}_k}$, and apply the Hedge algorithm with the same learning rates (η_τ). As a result, the sequence of distributions $(x_k^{(\tau)})$ satisfies the Hedge update rule (2.14). That is,

$$x_{k,a}^{(\tau+1)} \propto x_{k,a}^{(\tau)} e^{-\eta_\tau \ell_{k,a}(x^{(\tau)})}.$$

Now suppose the existence of an underlying C^1 function $X_k(t)$ defined on $\mathbb{R}_+$, and suppose that $x_k^{(\tau)}$ corresponds to a discretization of this continuous trajectory, at times T_τ, $\tau \in \mathbb{N}$, such that the time steps are given by a decreasing, vanishing, non-summable sequence (ϵ_τ), i.e. $T_{\tau+1} - T_\tau = \epsilon_\tau$, and $x_k^{(\tau)} = X_k(T_\tau)$. Then we have for all k and all $a \in \mathcal{A}_k$, using Landau notation:

$$\begin{aligned}
X_{k,a}(T_{\tau+1}) &= x_{k,a}^{(\tau+1)} \\
&= x_{k,a}^{(\tau)} \frac{e^{-\eta_\tau \ell_{k,a}(x^{(\tau)})}}{\sum_{a' \in \mathcal{A}_k} x_{k,a'}^{(\tau)} e^{-\eta_\tau \ell_{k,a'}(x^{(\tau)})}} \\
&= x_{k,a}^{(\tau)} \frac{1 - \eta_\tau \ell_{k,a}(x^{(\tau)}) + o(\eta_\tau)}{1 - \eta_\tau \sum_{a' \in \mathcal{A}_k} x_{k,a'}^{(\tau)} \ell_{k,a'}(x^{(\tau)}) + o(\eta_\tau)} \\
&= X_{k,a}(T_\tau) \left[1 + \eta_\tau \left\langle \ell_k(X(T_\tau)), X_k(T_\tau) \right\rangle - \eta_\tau \ell_{k,a}(X(T_\tau))\right] + o(\eta_\tau).
\end{aligned}$$

Thus, rearranging,

$$\frac{X_{k,a}(T_{\tau+1}) - X_{k,a}(T_\tau)}{T_{\tau+1} - T_\tau} \frac{\epsilon_\tau}{\eta_\tau} = X_{k,a}(T_\tau)(\langle \ell_k(X(T_\tau)), X_k(T_\tau)\rangle - \ell_{k,a}(X(T_\tau))) + o(1).$$

In particular, if we take the discretization time steps ϵ_τ to be equal to the sequence of learning rates η_τ, the expression simplifies, and taking the limit as $\eta_\tau \to 0$, we obtain the following ODE system:

$$\text{Replicator} \begin{cases} \forall k,\ \forall a \in \mathcal{A}_k, \dot{X}_{k,a}(t) = X_{k,a}(t) \left(\langle \ell_k(X(t)), X_k(t)\rangle - \ell_{k,a}(X(t))\right) \\ X(0) \in \mathring{\Delta} \end{cases} \tag{3.1}$$

Here $\ell_k : \Delta \to \mathbb{R}_+^{\mathcal{A}_k}$ coincides with the gradient of the potential, by Equation (2.2).

This ODE is known as the replicator ODE, and is a common tool in evolutionary game theory [135]. It has also been studied to model dynamics of populations in routing games, see for example [49] and Chapter III.29 in [104].

We will assume that the initial condition $X(0)$ is taken in the relative interior of the product of simplices,

$$\mathring{\Delta} = \{x \in \Delta^{\mathcal{A}_1} \times \cdots \times \Delta^{\mathcal{A}_K} : \forall k,\ \forall a \in \mathcal{A}_k, x_{k,a} > 0\}.$$

We require the initial distribution to have positive probability mass on all actions for the following reason: whenever $X_{k,a}(0) = 0$, any solution trajectory will have $X_{k,a}(t) \equiv 0$ identically. It is impossible for such trajectories to converge to the set of Nash equilibria $\mathcal{N}$ if the

support of equilibria in $\mathcal{N}$ contains the action a. In other words, the replicator dynamics cannot expand the support of the initial distribution, therefore we require that the initial distribution be supported everywhere.

Equation (3.1) can be written concisely as $\dot{X}(t) = F(X(t))$, where F is a vector field given by

$$\begin{aligned} F :&\Delta \to \mathcal{H} \\ &x \mapsto F_{k,a}(x) = x_{k,a}(\langle \ell_k(x), x_k \rangle - \ell_{k,a}(x)). \end{aligned}$$

Here, Δ is the product of simplices, and $\mathcal{H}$ is the product $\mathcal{H} = \mathcal{H}^{\mathcal{A}_1} \times \cdots \times \mathcal{H}^{\mathcal{A}_K}$, where

$$\mathcal{H}^{\mathcal{A}_k} = \left\{ v_k \in \mathbb{R}^{\mathcal{A}_k} : \sum_{a \in \mathcal{A}_k} v_{k,a} = 0 \right\}$$

is the hyperplane parallel to the simplex $\Delta^{\mathcal{A}_k}$. Indeed, we have for all k and all $x_k \in \Delta^{\mathcal{A}_k}$,

$$\begin{aligned} \sum_{a \in \mathcal{A}_k} F_{k,a}(x) &= \sum_{a \in \mathcal{A}_k} x_{k,a}(\langle \ell_k(x), x_k \rangle - \ell_{k,a}) \\ &= \langle \ell_k(x), x_k \rangle \sum_{a \in \mathcal{A}_k} x_{k,a} - \sum_{a \in \mathcal{A}_k} \ell_{k,a}(x) x_{k,a} \\ &= 0. \end{aligned}$$

This ensures that the derivatives remain in the direction of the simplex, and will be used to prove that the solution trajectory remains in the simplex.

Existence, uniqueness, and viability of the solution

Proposition 7. *The ODE* (3.1) *has a unique maximal (i.e. defined on a maximal interval) C^1 solution $X(t)$ which remains in $\mathring{\Delta}$ and is defined on all of $\mathbb{R}_+$.*

Proof. First, since ∇f is assumed to be Lipschitz continuous, so is the vector field F. Thus we have existence and uniqueness of a maximal C^1 solution by the Cauchy-Lipschitz theorem (e.g. Theorem 2.5 in [132]).

To show that the solution remains in the relative interior of Δ, we observe that for all t and for all k,

$$\sum_{a \in \mathcal{A}_k} \dot{X}_{k,a}(t) = \sum_{a \in \mathcal{A}_k} F_{k,a}(X(t)) = 0,$$

since we observed that F_k maps to $\mathcal{H}_k$. Therefore, $\sum_{a \in \mathcal{A}} X_{k,a}(t)$ is constant and equal to 1 (since the initial point is in the simplex). To show that $X_{k,a}(t) > 0$ for all t in the solution domain, assume by contradiction that there exists $t_0 > 0$ and $a_0 \in \mathcal{A}_k$ such that $X_{a_0}(t_0) = 0$. Since the solution trajectories are continuous, we can assume, without loss of generality, that

t_0 is the infimum of all such times (thus for all $t < t_0$, $X_{a_0}(t) > 0$). Now consider the new system given by

$$\begin{aligned} \dot{\tilde{X}}_{k,a} &= \tilde{X}_{k,a}\left(\left\langle \ell_k(\tilde{X}(t)), \tilde{X}_k \right\rangle - \ell_{k,a}(\tilde{X})\right) && \forall a \neq a_0 \\ \tilde{X}_{k,a}(t_0) &= X_a(t_0) && \forall a \neq a_0 \end{aligned}$$

and $\tilde{X}_{a_0}(t)$ is constant equal to 0. Any solution of the new system, defined on $(t_0 - \delta, t_0]$, is also a solution of Equation (3.1). Since $X(t_0) = \tilde{X}(t_0)$, we have $X \equiv \tilde{X}$ by uniqueness of the solution. This leads to a contradiction since by assumption, for all $t < t_0$, $X_{a_0}(t) > 0$ but $\tilde{X}_{a_0}(t) = 0$.

This proves that X remains in $\mathring{\Delta}$. Furthermore, since Δ is compact, we have by Theorem 2.4 in [71] that the solution is defined on $\mathbb{R}_+$ (otherwise it would eventually leave any compact set). □

3.2 Stationary points

We now identify stationary points of the dynamics, i.e. points $x \in \Delta$ such that $F(x) = 0$. The set of all such stationary points is denoted $\mathcal{RN}$.

Proposition 8. *A product distribution $x \in \Delta$ is a stationary point of the ODE (3.1) if and only if for all k, the losses $(\ell_{k,a}(x), a \in \text{support}(x_k))$ are equal. Furthermore, $\mathcal{RN}$ is compact, and the potential function f takes finitely many values on $\mathcal{RN}$.*

Proof. From Equation (3.1), we have

$$\begin{aligned} F_{k,a}(x) = 0 &\Leftrightarrow x_{k,a} = 0 \text{ or } \ell_{k,a}(x) = \langle \ell_k(x), x_k \rangle \\ &\Leftrightarrow \forall a \in \text{support}(x_k),\ \ell_{k,a}(x) = \langle \ell_k(x), x_k \rangle\,, \end{aligned} \tag{3.2}$$

which proves the first part of the claim. To prove the second part, we observe that for any stationary point $x^\dagger \in \mathcal{RN}$ if we let $\mathcal{A}_k^\dagger$ be the support of $x_k^\dagger$, then the condition (3.2) is equivalent to the Nash equilibrium Definition 2 for the modified game played on the action sets $\mathcal{A}_k^\dagger$. Since Nash equilibria are the minimizers of the potential f by Theorem 1, we have

$$x^\dagger \in \mathcal{N}^\dagger = \underset{x \in \Delta_1^\dagger \times \cdots \times \Delta_K^\dagger}{\arg\min}\ f(x),$$

where $\Delta_k^\dagger$ is the set of distributions with support contained in $\mathcal{A}_k^\dagger$. The set of minimizers $\mathcal{N}^\dagger$ is compact. Since there are finitely many possible supports, we have that $\mathcal{RN}$ is compact as the finite union of the compact sets $\mathcal{N}^\dagger$, and f takes finitely many values on $\mathcal{RN}$. □

In particular, any Nash equilibrium satisfies this assumption, and is a stationary point for the replicator dynamics. However, a stationary point is not necessarily a Nash equilibrium, since one may have a stationary point with $x_{k,a_0} = 0$ and $\ell_{k,a_0}(x)$ strictly lower than $\langle \ell_k(x), x_k \rangle$.

A stationary point $x^\dagger \in \mathcal{RN}$ given by Proposition 8 is also called *restricted Nash equilibrium* (hence the notation $\mathcal{RN}$), see e.g. [49], since it is a Nash equilibrium for the game if the action set is restricted to the support of $x^\dagger$.

3.3 Lyapunov functions and convergence to Nash equilibria

In this section, we exhibit several Lyapunov functions for the invariant sets $\mathcal{RN}$ and $\mathcal{N}$, and show that the solution trajectories converge to $\mathcal{N}$, with an explicit convergence rate.

Lyapunov function

Given an invariant set Γ for the ODE, we say that a differentiable function $V : \Delta \to \mathbb{R}$ is a Lyapunov function for Γ, in reference to Aleksandr Mikhailovich Lyapunov [88], if along solution trajectories $X(t)$ of the ODE, $V(X(t))$ is constant on Γ, and decreasing outside of Γ. The time derivative of $V(X(t))$ is given by

$$\frac{d}{dt}V(X(t)) = \left\langle \nabla V(X(t)), \dot{X}(t) \right\rangle = \langle \nabla V(X(t)), F(X(t)) \rangle ,$$

thus V is a Lyapunov function for Γ if and only if

$$\begin{cases} \langle \nabla V(x), F(x) \rangle = 0 & \forall x \in \Gamma \\ \langle \nabla V(x), F(x) \rangle < 0 & \forall x \notin \Gamma. \end{cases}$$

We start by showing that the potential function f is a Lyapunov function for the set of stationary points $\mathcal{RN}$.

Proposition 9. *The convex potential function f is a Lyapunov function for the set of stationary points $\mathcal{RN}$ under the replicator ODE.*

Proof. Recall that by Equation (2.2), we have that $\nabla_{x_k} f(x) = \kappa_k \ell_k(x)$. Taking the time derivative of the Lyapunov function along the solution trajectory, we have

$$\begin{aligned} \frac{d}{dt}f(X(t)) &= \left\langle \nabla f(X(t)), \dot{X}(t) \right\rangle \\ &= \sum_{k=1}^{K} \sum_{a \in \mathcal{A}_k} \nabla_{k,a} f(X(t)) X_{k,a}(t) (\langle \ell_k(X(t)), X_k(t) \rangle - \ell_{k,a}(X(t))) \\ &= \sum_{k=1}^{K} \kappa_k \left[\langle \ell_k(X(t)), X_k(t) \rangle^2 - \left\langle (\ell_k^2(X(t))), X_k(t) \right\rangle \right] \leq 0, \end{aligned} \tag{3.3}$$

where each term of the sum is non-positive by Jensen's inequality, and equals zero if and only if for all k, $\ell_{k,a}(X(t))$ is constant on the support of $X_k(t)$. This condition is exactly the condition of Proposition 8 characterizing stationary points. □

This proves that the solution trajectories of the replicator ODE converge to the set of its stationary points. In fact, we can prove convergence to the set of Nash equilibria (a subset of stationary points), using a second Lyapunov function defined as follows. Let $x^\star \in \mathcal{N}$ be a Nash equilibrium, and consider the function

$$V_{\mathrm{KL}}(x) = \sum_{k=1}^{K} \kappa_k D_{\mathrm{KL}}(x_k^\star, x_k), \tag{3.4}$$

where $D_{\mathrm{KL}}(x_k^\star, x_k)$ is the Kullback-Leibler divergence, defined as follows

$$D_{\mathrm{KL}}(x^\star, x) = \sum_{a \in \mathcal{A}} x_a^\star \ln \frac{x_a^\star}{x_a}.$$

Note that along solution trajectories of the system, the function $V_{\mathrm{KL}}(X(t))$ is finite for all t since by Proposition 7, the solution remains in the relative interior of the simplex.

Proposition 10. *V_{KL} is a Lyapunov function for the set of Nash equilibria $\mathcal{N}$ under the replicator ODE.*

Proof. Taking the time derivative of D_{KL} along the solution trajectory, we have

$$\begin{aligned}
\frac{d}{dt} D_{\mathrm{KL}}(X_k(t)) &= -\sum_{a \in \mathcal{A}_k} x_{k,a}^\star \frac{\dot{X}_{k,a}(t)}{X_{k,a}(t)} \\
&= -\sum_{a \in \mathcal{A}_k} x_{k,a}^\star (\langle \ell_k(X_k(t)), X_k(t) \rangle - \ell_{k,a}(X(t))) \\
&= -\langle \ell_k(X_k(t)), X_k(t) \rangle + \langle \ell_k(X(t)), x_k^\star \rangle .
\end{aligned}$$

Then, by Equation (2.2), we have that $\nabla_{x_k} f(x) = \kappa_k \ell_k(x)$, thus

$$\begin{aligned}
\frac{d}{dt} V_{\mathrm{KL}}(X(t)) &= \frac{d}{dt} \sum_{k=1}^{K} \kappa_k D_{\mathrm{KL}}(x_k^\star, X_k(t)) \\
&= \langle \nabla f(X(t)), x^\star - X(t) \rangle \\
&\leq f^\star - f(X(t)),
\end{aligned} \tag{3.5}$$

where the last inequality is by convexity of f. To conclude, we simply recall that the set of Nash equilibria coincides with the set of minimizers of the convex potential f by Theorem 1, thus for all $x \in \mathcal{N}$, $\langle \nabla f(x), x^\star - x \rangle = 0$ by first-order optimality, and for all $x \notin \mathcal{N}$, $\langle \nabla f(x), x^\star - x \rangle \leq f^\star - f(x) < 0$, which concludes the proof. □

Next, by carefully combining the two Lyapunov functions f and V_{KL}, we exhibit a time-varying Lyapunov function, which will allow us to prove an explicit convergence rate. Consider the function

$$V(x, t) := t(f(x) - f^\star) + V_{\mathrm{KL}}(x), \tag{3.6}$$

and let

$$\mathcal{V}(t) := V(X(t), t).$$

Since the set of Nash equilibria coincides with the set of minimizers of f on Δ, it suffices by Lemma 2, in order to prove convergence of the solution to $\mathcal{N}$, to show that $f(X(t))$ converges to $f^\star$, the minimum of f on Δ.

Theorem 3. *The unique solution $X(t)$ of the replicator ODE (3.1) satisfies, for all $t > 0$,*

$$f(X(t)) - f^\star \leq \frac{\mathcal{V}(0)}{t}.$$

In particular, $X(t)$ converges to the set of Nash equilibria $\mathcal{N}$.

Proof. Using the bounds (3.3) and (3.5) on the time-derivative of V_{KL} and f, we have

$$\frac{d}{d}\mathcal{V}(t) = f(X(t)) - f^\star + t\frac{d}{dt}f(X(t)) + \frac{d}{dt}V_{\mathrm{KL}}(X(t)) \leq 0$$

Therefore, $\mathcal{V}(t)$ is a non-increasing function of t. Finally,

$$f(X(t)) - f^\star \leq \frac{\mathcal{V}(t)}{t} \leq \frac{\mathcal{V}(0)}{t},$$

since the KL divergence is non-negative, and $\mathcal{V}$ is non-increasing. □

The Lyapunov function used in this proof is a special case of the Lyapunov function (8.3) studied in Chapter 8, which we use to prove convergence of the mirror descent ODE. We will further study the replicator ODE in Chapter 9 in the second part of the thesis, and show, in particular, that it is an instance of the mirror descent ODE, and that it can be accelerated by averaging (so that solution trajectories converge faster to the set of minimizers of the potential).

In the remainder of the chapter, we will further study the stability of the stationary points for a single population. In particular, we show that all stable stationary points are Nash equilibria, then under an additional monotonicity assumption, we show that Nash equilibria are, in fact, exponentially stable (a stronger result than Theorem 3).

3.4 Linearizing the dynamics around stationary points

We now assume that $K = 1$, and we omit the subscripts k to simplify notation. We also assume that the potential function is twice differentiable, i.e. that the loss function $\ell = \kappa\nabla f$ is differentiable. To study stability of the stationary points, we derive the eigenvalues $\mathfrak{S}$ of the Jacobian of the vector field at stationary points. Note that the vector field is continuously

differentiable by assumption on f, so the Jacobian exists and is continuous. Define, for $x \in \Delta$,

$$\bar{\ell}(x) := \langle \ell(x), x \rangle .$$

Then, we can rewrite the vector field F as follows

$$\begin{aligned} F : \Delta &\to \mathcal{H}_{\mathcal{A}} \\ x &\mapsto -\operatorname{diag}(\ell(x))x + x\bar{\ell}(x) \end{aligned} \tag{3.7}$$

where $\operatorname{diag} : \mathbb{R}^{\mathcal{A}} \to \mathbb{R}^{\mathcal{A}\times\mathcal{A}}$ is the operator which maps a vector to the diagonal matrix whose diagonal elements are given by the the vector entries. Observing that $\bar{\ell}(x) = \mathbf{1}_{\mathcal{A}}^{\top} \operatorname{diag}(\ell(x))x$, where $\mathbf{1}_{\mathcal{A}}$ is a vector in $\mathbb{R}^{\mathcal{A}}$ whose entries are all equal to one, we can write

$$\begin{aligned} F(x) &= -\operatorname{diag}(\ell(x))x + x\mathbf{1}_{\mathcal{A}}^{\top} \operatorname{diag}(\ell(x))x \\ &= -\left[I_{\mathcal{A}} - x\mathbf{1}_{\mathcal{A}}^{\top}\right] \operatorname{diag}(\ell(x))x \\ &= -\Psi(x)L(x)x \end{aligned} \tag{3.8}$$

where

$$\begin{aligned} \Psi(x) &= I_{\mathcal{A}} - x\mathbf{1}_{\mathcal{A}}^{\top} \\ L(x) &= \operatorname{diag}(\ell(x)) \end{aligned}$$

are $\mathcal{A} \times \mathcal{A}$ matrices. This matrix form of F will be useful to derive the linearized system.

As observed in the previous section, F is defined on $\Delta^{\mathcal{A}}$ and has values in $\mathcal{H}_{\mathcal{A}}$, the hyperplane orthogonal to the unit vector $\mathbf{1}_{\mathcal{A}}$. But F can also be viewed as a function from $\mathbb{R}^{\mathcal{A}}$ to itself. We first derive the Jacobian of F viewed as function from $\mathbb{R}^{\mathcal{A}}$ to $\mathbb{R}^{\mathcal{A}}$, denoted $\nabla F(x)$, and then consider its restriction to $\mathcal{H}$ to obtain $\mathfrak{S}$ as the eigenvalues of the restriction $\nabla F(x)_{|\mathcal{H}}$.

Lemma 3. *The Jacobian of F is given by*

$$\nabla F(x) = \bar{\ell}(x) I_{\mathcal{A}} - \Psi(x) \operatorname{diag}(x) \nabla \ell(x) - \Psi(x) L(x). \tag{3.9}$$

Proof. Let $DF(x)$ be the differential of F at x, and let e_a be a vector of the canonical basis. Then $DF(x)(e_a)$ is the directional derivative of F in the direction of e_a. From the matrix form of F given in Equation (3.8), and using the product rule of differentials, we have

$$\begin{aligned} DF(x)(e_a) &= -D\Psi(x)(e_a)L(x)x - \Psi(x)DL(x)(e_a)x - \Psi(x)L(x)e_a \\ &= e_a\mathbf{1}_{\mathcal{A}}^{\top}L(x)x - \Psi(x)\operatorname{diag}(\nabla\ell(x)e_a)x - \Psi(x)L(x)e_a \\ &= e_a\ell(x)^{\top}x - \Psi(x)\operatorname{diag}(x)\nabla\ell(x)e_a - \Psi(x)L(x)e_a \\ &= \left(\bar{\ell}(x)I_{\mathcal{A}} - \Psi(x)\operatorname{diag}(x)\nabla\ell(x) - \Psi(x)L(x)\right)e_a \end{aligned} \tag{3.10}$$

where we used the product rule in the first equality, we used the expression of the following differentials in the second equality

$$D\Psi(x)(e_a) = -e_a \mathbf{1}_{\mathcal{A}}^\top$$
$$DL(x)(e_a) = \operatorname{diag}(\nabla \ell(x) e_a),$$

and the fact that $\operatorname{diag}(u)v = \operatorname{diag}(v)u$ in the third equality. This proves the claim. □

Now that we have the expression of the Jacobian, we are ready to prove the first stability result.

3.5 Instability of non-Nash stationary points

Theorem 4. *If x is a stationary point of system* (3.1) *but not a Nash equilibrium, then x is unstable.*

Proof. Let x be a stationary point of ODE (3.1). Let $\mathcal{A}^*$ be the support of x and $\mathcal{A}^\diamond = \mathcal{A} \setminus \mathcal{A}^*$. Without loss of generality, we assume that in the vector representation of x, the support corresponds to the first elements. Finally, for a vector $v \in \mathbb{R}^{\mathcal{A}}$, we write v^* as a shorthand for $(v_a)_{a \in \mathcal{A}^*}$ and $v^\diamond$ as a shorthand for $(v_a)_{a \in \mathcal{A}^\diamond}$. Finally, we write ∇_* and $\nabla_\diamond$ the gradients taken along x^* and $x^\diamond$, respectively. Then we can calculate the different terms in the expression (3.9) of the Jacobian. First, by simple algebraic manipulation,

$$\begin{aligned}\Psi(x) &= \begin{pmatrix} I_{\mathcal{A}^*} & 0 \\ 0 & I_{\mathcal{A}_*} \end{pmatrix} - \begin{pmatrix} x^* \\ 0 \end{pmatrix} \begin{pmatrix} \mathbf{1}_{\mathcal{A}^*}^\top & \mathbf{1}_{\mathcal{A}^\diamond}^\top \end{pmatrix} \\ &= \begin{pmatrix} \Psi^*(x^*) & -x^* \mathbf{1}_{\mathcal{A}^\diamond}^\top \\ 0 & I_{\mathcal{A}^\diamond} \end{pmatrix},\end{aligned}$$

and

$$\begin{aligned}\operatorname{diag}(x)\nabla \ell(x) &= \begin{pmatrix} \operatorname{diag}(x^*) & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \nabla_* \ell^*(x) & \nabla_\diamond \ell^*(x) \\ \nabla_* \ell^\diamond(x) & \nabla_\diamond \ell^\diamond(x) \end{pmatrix} \\ &= \begin{pmatrix} \operatorname{diag}(x^*)\nabla_* \ell^*(x) & \operatorname{diag}(x^*)\nabla_\diamond \ell^*(x) \\ 0 & 0 \end{pmatrix}.\end{aligned}$$

Thus, taking the product, we have

$$\Psi(x)\operatorname{diag}(x)\nabla \ell(x) = \begin{pmatrix} \Psi^*(x^*)\operatorname{diag}(x^*)\nabla_* \ell^*(x) & \Psi^*(x^*)\operatorname{diag}(x^*)\nabla_\diamond \ell^*(x) \\ 0 & 0 \end{pmatrix}.$$

Finally,

$$\begin{aligned}\Psi(x)L(x) &= \begin{pmatrix} \Psi^*(x^*) & -x^* \mathbf{1}_{\mathcal{A}^\diamond}^\top \\ 0 & I_{\mathcal{A}^\diamond} \end{pmatrix} \begin{pmatrix} \operatorname{diag}(\ell^*(x)) & 0 \\ 0 & \operatorname{diag}(\ell^\diamond(x)) \end{pmatrix} \\ &= \begin{pmatrix} \bar{\ell}(x) I_{\mathcal{A}^*} - \bar{\ell}(x) x^* \mathbf{1}_{\mathcal{A}^*}^\top & -x^* \ell^\diamond(x)^\top \\ 0 & \operatorname{diag}(\ell^\diamond(x)) \end{pmatrix},\end{aligned}$$

where we used the fact that for a stationary point x, the losses are equal on the support of x (Proposition 8), so that $\operatorname{diag}(\ell^*(x)) = \bar{\ell}(x) I_{\mathcal{A}^*}$.

Combining these terms in the expression of the Jacobian given in Lemma 3, we obtain

$$\begin{aligned}\nabla F(x) &= \bar{\ell}(x) I_{\mathcal{A}} - \Psi(x) \operatorname{diag}(x) \nabla \ell(x) - \Psi(x) L(x) \\ &= \begin{pmatrix} -\Psi^*(x^*) \operatorname{diag}(x^*) \nabla_* \ell^*(x) + \bar{\ell}(x) x^* \mathbf{1}_{\mathcal{A}^*}^\top & -\Psi^*(x^*) \operatorname{diag}(x^*) \nabla_\diamond \ell^*(x) + x^* \ell^\diamond(x)^\top \\ 0 & \bar{\ell}(x) I_{\mathcal{A}^\diamond} - \operatorname{diag}(\ell^\diamond(x)). \end{pmatrix}\end{aligned} \tag{3.11}$$

Thus $\nabla F(x)$ is an upper block-triangular matrix of the form

$$\nabla F(x) = \begin{pmatrix} A & C \\ 0 & B \end{pmatrix}$$

with

$$\begin{aligned} A &= \bar{\ell}(x) x^* \mathbf{1}_{\mathcal{A}^*}^\top - \Psi^*(x^*) \operatorname{diag}(x^*) \nabla_* \ell^*(x) \\ B &= \bar{\ell}(x) I_{\mathcal{A}^\diamond} - \operatorname{diag}(\ell^\diamond(x)) \\ C &= (x^* \ell^\diamond(x)^\top - \Psi^*(x^*) \operatorname{diag}(x^*) \nabla_\diamond \ell^*(x)). \end{aligned}$$

The stability of x is determined by the restriction of $\nabla F(x)$ to $\mathcal{H}_{\mathcal{A}}$. Let $\alpha \in \mathcal{H}_{\mathcal{A}}$ and write $\alpha = \begin{pmatrix} \alpha^* \\ \alpha^\diamond \end{pmatrix}$. Then $\mathbf{1}_{\mathcal{A}^*}^\top \alpha^* + \mathbf{1}_{\mathcal{A}^\diamond}^\top \alpha^\diamond = 0$. Then we have

$$\begin{aligned} \nabla F(x) \begin{pmatrix} \alpha^* \\ \alpha^\diamond \end{pmatrix} &= \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} \begin{pmatrix} \alpha^* \\ \alpha^\diamond \end{pmatrix} \\ &= \begin{pmatrix} -\Psi^*(x^*) \operatorname{diag}(x^*) \nabla_* \ell^*(x) \alpha^* + \bar{\ell}(x) x^* \mathbf{1}_{\mathcal{A}^*}^\top \alpha^* + C \alpha^\diamond \\ B \alpha^\diamond \end{pmatrix} \\ &= \begin{pmatrix} -\Psi^*(x^*) \operatorname{diag}(x^*) \nabla_* \ell^*(x) \alpha^* - \bar{\ell}(x) x^* \mathbf{1}_{\mathcal{A}^\diamond}^\top \alpha^\diamond + C \alpha^\diamond \\ B \alpha^\diamond \end{pmatrix} \\ &= \begin{pmatrix} -\Psi^*(x^*) \operatorname{diag}(x^*) \nabla_* \ell^*(x) & -\bar{\ell}(x) x^* \mathbf{1}_{\mathcal{A}^\diamond}^\top + C \\ 0 & B \end{pmatrix} \begin{pmatrix} \alpha^* \\ \alpha^\diamond \end{pmatrix} \\ &= \begin{pmatrix} \tilde{A} & \tilde{C} \\ 0 & B \end{pmatrix} \begin{pmatrix} \alpha^* \\ \alpha^\diamond \end{pmatrix} \end{aligned}$$

where we used the fact that $\mathbf{1}_{\mathcal{A}^*}^\top \alpha^* = -\mathbf{1}_{\mathcal{A}^\diamond}^\top \alpha^\diamond$ in the third equality, and defined

$$\begin{aligned} \tilde{A} &= -\Psi^*(x^*) \operatorname{diag}(x^*) \nabla_* \ell^*(x) \\ \tilde{C} &= -\bar{\ell}(x) x^* \mathbf{1}^\top + C. \end{aligned}$$

Then the restriction of $\nabla F(x)$ to $\mathcal{H}_{\mathcal{A}}$ coincides with the restriction of $\tilde{\nabla} F(x)$ to $\mathcal{H}$, where

$$\tilde{\nabla} F(x) = \begin{pmatrix} \tilde{A} & \tilde{C} \\ 0 & B \end{pmatrix}$$

The benefit of the latter formulation is that the range of $\tilde{\nabla}F(x)$ is a subset of $\mathcal{H}_{\mathcal{A}}$, since

$$\begin{aligned}
&\mathbf{1}_{\mathcal{A}}^{\top}\tilde{\nabla}F(x)\\
&= \begin{pmatrix}\mathbf{1}_{\mathcal{A}^*}^{\top} & \mathbf{1}_{\mathcal{A}^\diamond}^{\top}\end{pmatrix}\begin{pmatrix}-\Psi^*(x^*)\operatorname{diag}(x^*)\nabla_*\ell^*(x) & -\bar{\ell}(x)x^*\mathbf{1}_{\mathcal{A}^\diamond}^{\top} + x^*\ell^\diamond(x)^{\top} - \Psi^*(x^*)\operatorname{diag}(x^*)\nabla_\diamond\ell^*(x)\\ 0 & \bar{\ell}(x)I_{\mathcal{A}^\diamond} - \operatorname{diag}(\ell^\diamond(x))\end{pmatrix}\\
&= \begin{pmatrix}0 & \mathbf{1}_{\mathcal{A}^*}^{\top}[-\bar{\ell}(x)x^*\mathbf{1}_{\mathcal{A}^\diamond}^{\top} + x^*\ell^\diamond(x)^{\top}] + \mathbf{1}_{\mathcal{A}^\diamond}^{\top}[\bar{\ell}(x)I_{\mathcal{A}^\diamond} - \operatorname{diag}(\ell^\diamond(x))]\end{pmatrix}\\
&= \begin{pmatrix}0 & -\bar{\ell}(x)\mathbf{1}_{\mathcal{A}^\diamond}^{\top} + \ell^\diamond(x)^{\top} + \bar{\ell}(x)\mathbf{1}_{\mathcal{A}^\diamond}^{\top} - \ell^\diamond(x)^{\top}\end{pmatrix}\\
&= \begin{pmatrix}0 & 0\end{pmatrix},
\end{aligned}$$

where we used the fact that $\mathbf{1}_{\mathcal{A}^*}^{\top}\Psi^*(x) = 0$, and that $\mathbf{1}_{\mathcal{A}^*}^{\top}x^* = 1$. Therefore, $\mathcal{H}_{\mathcal{A}}$ is an invariant subspace of $\tilde{\nabla}F(x)$, and the spectrum $\mathfrak{S}$ is given by the spectrum of $\tilde{\nabla}F(x)$, from which we remove one zero (corresponding to the eigenvector $\mathbf{1}_{\mathcal{A}}$). Next, since $\tilde{\nabla}F(x)$ is block upper-triangular, with diagonal blocks $\tilde{A}$ and B, we have, using the expression of B,

$$Sp(\tilde{\nabla}F(x)) = Sp(\tilde{A}) \cup \{\bar{\ell}(x) - \ell_a^\diamond(x)\}_{a\in\mathcal{A}^\diamond}$$

Therefore

$$\mathfrak{S} = Sp(\hat{A}) \cup \{\bar{\ell}(x) - \ell_a^\diamond(x)\}_{a\in\mathcal{A}^\diamond}$$

where $\hat{A}$ is the restriction of $\tilde{A}$ to $\mathcal{H}_{\mathcal{A}^*} = \mathbf{1}_{\mathcal{A}^*}^{\perp}$.

To conclude the proof, suppose that x is a stationary point but not a Nash equilibrium, i.e. $x \in \mathcal{RA} \setminus \mathcal{N}$. Then there exists $a \in \mathcal{A}^\diamond$ such that $\bar{\ell}(x) - \ell_a^\diamond(x) > 0$, and it follows that $\mathfrak{S}$ contains at least one strictly positive eigenvalue, therefore x is unstable (by Theorem 3.7 in [71] for example). □

3.6 Exponential stability of Nash equilibria

In order to prove the converse of Theorem 4, we need to study the eigenvalues of $\hat{A}$, the restriction to $\mathcal{H}_{\mathcal{A}^*}$ of $\tilde{A} = -\Psi^*(x^*)\operatorname{diag}(x^*)\nabla_*\ell^*(x)$.

Lemma 4. *The matrix $\Psi^*(x^*)\operatorname{diag}(x^*)$ is symmetric positive semidefinite and its restriction to $\mathcal{H}_{\mathcal{A}^*}$ is positive definite.*

Proof. We have $\Psi^*(x^*)\operatorname{diag}(x^*) = \operatorname{diag}(x^*) - x^*(x^*)^{\top}$ is symmetric. We also have

$$\mathbf{1}_{\mathcal{A}^*}^{\top}\Psi^*(x^*)\operatorname{diag}(x^*)\mathbf{1}_{\mathcal{A}^*} = 0,$$

and for all $y \in \mathcal{H}_{\mathcal{A}^*}$,

$$y^{\top}\Psi^*(x^*)\operatorname{diag}(x^*)y = \sum_{a\in\mathcal{A}^*} x_a^* y_a^2 - \left(\sum_{a\in\mathcal{A}^*} x_a^* y_a\right)^2$$

which, by Jensen's inequality, is strictly positive except at $y = 0$. □

Lemma 5. *Let R and S be two symmetric matrices such that R is positive-definite and S is positive-semidefinite. Then the product RS is diagonalizable, has non-negative eigenvalues and has the same number of zero eigenvalues as S (with the same eigenvectors).*

Proof. Since R is positive definite, there exists a positive definite matrix $\bar{R}$ such that $R = \bar{R}^2$. Then we have

$$\bar{R}^{-1}RS\bar{R} = \bar{R}S\bar{R}$$

thus RS is similar to the symmetric matrix $\bar{R}S\bar{R}$, and is diagonalizable.

Consider the function $h : x \mapsto RSx$ and the inner product $\langle x; y\rangle = x^\top R^{-1}y$. We have

$$\langle h(x); y\rangle = x^\top SRR^{-1}y = x^\top Sy.$$

Thus if λ is an eigenvalue of h with eigenvector x, then

$$\langle h(x); x\rangle = \lambda\langle x; x\rangle$$

i.e. $\lambda = \frac{x^\top Sx}{x^\top R^{-1}x}$ which is non-negative since $S \succeq 0$ and $R \succ 0$. Furthermore, $\lambda = 0$ if and only if $Sx = 0$, which proves the claim. □

We can now show a partial converse of Theorem 4.

Theorem 5. *Assume that $\nabla\ell$ is symmetric, positive definite. Then x is a Nash equilibrium only if x is a locally exponentially stable stationary point of the replicator dynamics* (3.1).

Proof. Suppose x is a Nash equilibrium. Then it is a stationary point of the system (3.1). To show that it is exponentially stable, recall that the eigenvalues of the Jacobian are given by:

$$\mathfrak{S} = Sp(\hat{A}) \cup \{\bar{\ell}(x) - \ell_a^\diamond(x)\}_{a\in\mathcal{A}^\diamond}$$

where $\hat{A}$ is the restriction of $-\Psi^*(x^*)\operatorname{diag}(x^*)\nabla_*\ell^*(x)$ to $\mathcal{H}_{\mathcal{A}^*}$. By Lemma 4, $\Psi^*(x^*)\operatorname{diag}(x^*)$ has a positive definite restriction to $\mathcal{H}_{\mathcal{A}^*}$, and $\nabla\ell^*(x)$ is positive definite as a diagonal block of $\nabla f(x)$, which is positive definite by assumption. Therefore, applying Lemma 5, we have that all eigenvalues of $\hat{A}$ are negative. Therefore x is exponentially asymptotically stable (for example by Theorem 4.7 of [71]). □

Application to congestion games

We now consider the special case of congestion games to illustrate the assumption that $\nabla\ell(x)$ is positive definite. Consider the congestion game defined in Section 2.3, in which the loss function is given by

$$\ell(x) = M^\top c(\bar{M}x)$$

where M is an incidence matrix, and c_r are congestion functions assumed to be non-negative non-decreasing by Assumption 1. We further assume that c_r are differentiable. Then we have the following result.

Lemma 6. *For the congestion game with differentiable congestion functions, the gradient $\nabla_* \ell^*(x)$ is a symmetric positive-semidefinite matrix. Furthermore, if the congestion functions are strictly increasing and the incidence matrix M is injective, then $\nabla_* \ell^*(x)$ is positive-definite.*

Proof. We have $\ell^*(x) = M_*^\top c(\bar{M}x)$, where M_* is the submatrix with columns $a \in \mathcal{A}^*$, the support of x, and $\bar{M} = m(\mathcal{S})M$ is the incidence matrix scaled by the total mass of the population. Thus

$$\nabla_* \ell^* = m(\mathcal{S}) M_*^\top \nabla c(\bar{M}x) M_*$$

where, by definition of c,

$$\nabla c(\phi) = \operatorname{diag}(\{c_r'(\phi_r)\}_{r \in \mathcal{R}}).$$

Thus $\nabla_* \ell^*(x)$ is symmetric, and since c_r is non-increasing for all r, it is positive semidefinite. If all congestion functions are strictly increasing and M is injective, then $\nabla_* \ell^*(x)$ is positive definite. □

Note that the incidence matrix may not be injective in general, since $M \in \{0,1\}^{\mathcal{R} \times \mathcal{A}}$ and $|\mathcal{A}| = 2^{\mathcal{R}}$ in the worst case. The expression of the spectrum suggests that if we could find a more concise representation of the game, by reducing the number of actions, the dynamics may converge faster in the reduced game. This is discussed in the next section.

Reducing the size of the congestion game

We observe that if an action a_0 is a conic combination of other actions, then the congestion game without a_0 is equivalent, in a sense defined below, to the original game, allowing us to reduce the size of the action set $\mathcal{A}$.

More precisely, consider a given action $a_0 \in \mathcal{A}$, and let $\bar{\mathcal{A}} = \mathcal{A} \setminus \{a_0\}$. Assume a_0 is a conic combination of actions in $\bar{\mathcal{A}}$, that is, $M_{a_0} = \sum_{a \in \bar{\mathcal{A}}} \lambda_a M_a$ for some non-negative coefficients λ_a. First, we must have $\sum_{a \in \bar{\mathcal{A}}} \lambda_a \geq 1$: since a_0 is non-empty (by assumption, no action is empty), there exists r such that $M_{r,a_0} = 1$. But

$$M_{r,a_0} = \sum_{a \in \bar{\mathcal{A}}} \lambda_a M_{r,a} \leq \sum_{a \in \bar{\mathcal{A}}} \lambda_a$$

since $M_{r,a} \in \{0,1\}$, which proves the claim.

Proposition 11. *Let $a_0 \in \mathcal{A}$, denote $\bar{\mathcal{A}} = \mathcal{A} \setminus \{a_0\}$, and assume a_0 is a conic combination of actions in $\bar{\mathcal{A}}$.*

If $y \in \Delta^{\bar{\mathcal{A}}}$ is a Nash equilibrium for the game without a_0, then the vector x (obtained by augmenting y with a 0 on a_0) is a Nash equilibrium for the original game.

Proof. We have for all $a \in \operatorname{support}(y)$, the loss $\ell_a(y)$ is equal to $\bar{\ell}(y)$ the minimum loss across all actions in $\bar{\mathcal{A}}$. Then if c_y is the vector of resource congestions, the loss of action a_0 under

distribution y is

$$\begin{aligned}\ell_{a_0}(y) = M_{a_0}^\top c_y &= \sum_{a\in\bar{\mathcal{A}}} \lambda_a M_a^\top c_y = \sum_{a\in\bar{\mathcal{A}}} \lambda_a \ell_a(y) \\ &\geq \sum_{a\in\bar{\mathcal{A}}} \lambda_a \bar{\ell}(y) \geq \bar{\ell}(y).\end{aligned}$$

Thus y augmented by 0 on a_0 is an equilibrium of the original game. □

Proposition 12. *Let $a_0 \in \mathcal{A}$, denote $\bar{\mathcal{A}} = \mathcal{A} \setminus \{a_0\}$, and assume a_0 is a conic combination of actions in $\bar{\mathcal{A}}$.*

If $x \in \Delta^{\mathcal{A}}$ is a Nash equilibrium for the original game, then $y \in \Delta^{\bar{\mathcal{A}}}$ defined by

$$y_a = x_a + \frac{\lambda_a}{\sum_{a'\in\bar{\mathcal{A}}} \lambda_{a'}} x_{a_0}$$

is a Nash equilibrium for the game without a_0.

Proof. First, y is, by definition, a distribution over $\bar{\mathcal{A}}$. To show that it is a Nash equilibrium of the reduced game, we argue that y and x induce the same resource loads, that is, $Mx = My$. To show this, we observe that if $a_0 \in \text{support}(x)$, we must have $\sum_{a\in\bar{\mathcal{A}}} \lambda_a = 1$. Indeed, if $x_{a_0} > 0$, then by definition of a Nash equilibrium, $\ell_{a_0}(x) \leq \ell_a(x)$ for all a. But

$$\ell_{a_0}(x) = \sum_{a\in\bar{\mathcal{A}}} \lambda_a \ell_a(x) \geq \sum_{a\in\bar{\mathcal{A}}} \lambda_a \ell_{a_0}(x)$$

therefore $\sum_{a\in\bar{\mathcal{A}}} \lambda_a \leq 1$, which, combined with the previous observation that $\sum_{a\in\bar{\mathcal{A}}} \lambda_a \geq 1$, proves the claim.

Now we consider two cases: if $x_{a_0} = 0$, then we have immediately $My = Mx$. If $x_{a_0} > 0$, then we have $\sum_{a\in\bar{\mathcal{A}}} \lambda_a = 1$ and

$$\begin{aligned}My &= \sum_{a\in\bar{\mathcal{A}}} y_a M_a \\ &= \sum_{a\in\bar{\mathcal{A}}} x_a M_a + x_{a_0} \sum_{a\in\bar{\mathcal{A}}} \lambda_a M_a \\ &= \sum_{a\in\bar{\mathcal{A}}} x_a M_a + x_{a_0} M_{a_0} \\ &= Mx.\end{aligned}$$

Therefore the distribution y induces the same resource loads as x, hence the same losses, and y is a Nash equilibrium of the reduced game. □

With the previous propositions, one can reduce the size of the game by removing a_0 from the set of actions, and obtain an equivalent game. Applying this argument repeatedly, we can reduce $\mathcal{A}$ to a minimal set $\hat{\mathcal{A}}$. One way to compute such a minimal set is to find a Hilbert basis of the family $\{M_a\}_{a\in\mathcal{A}}$, $H = \{M_a\}_{a\in\hat{\mathcal{A}}}$, and use $\hat{\mathcal{A}}$ as the reduced action set.

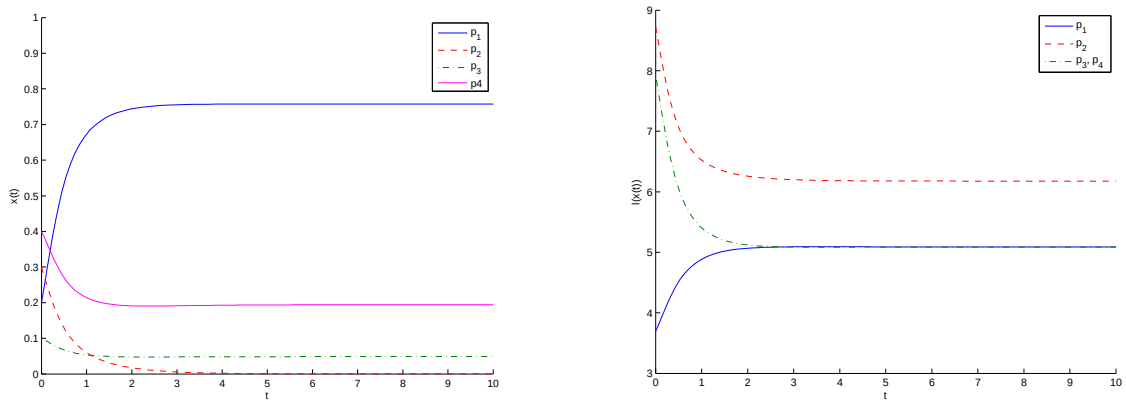


Figure 3.1: Trajectory of $x(t)$ (left) and evolution of loss functions (right). We have convergence to the set of Nash equilibria: on the support of the limit distribution, all action losses are equal.

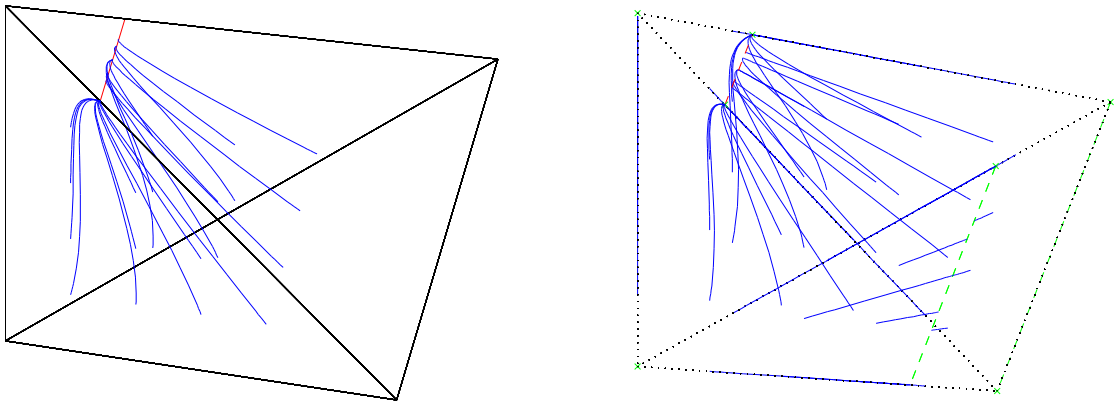


Figure 3.2: Solution trajectory $X(t)$ in the simplex $\Delta^{\mathcal{A}}$ (represented as a Tetrahedron). Starting from different initial conditions in the interior of the simplex, we have convergence to different points in $\mathcal{N}$ (represented by a red dashed line). Other stationary points are in dashed green lines. In particular, the entire face $\{x : x_1 = x_2 = 0\}$ is stationary.

3.7 Numerical example

We illustrate the replicator dynamics on a congestion game example. Suppose we have three resources $\mathcal{R} = \{r_1, r_2, r_3\}$, with quadratic congestion functions

$$c_1(\phi_1) = \frac{1}{2}(1 + \phi_1)^2$$
$$c_2(\phi_2) = (1 + \phi_2)^2$$
$$c_3(\phi_1) = 2(1 + \phi_3)^2$$

and consider the following actions

$$p_1 = \{r_1, r_2\}, p_2 = \{r_2, r_3\}, p_3 = \{r_3, r_1\}, p_4 = \{r_3, r_1\}.$$

In particular, we have $p_4 = p_3$. In particular, we do not have uniqueness of the Nash equilibrium in this case. The set of Nash equilibria is given by

$$\mathcal{N} = \{x : x_1 = .757, x_2 = 0, x_3 + x_4 = .2426\}.$$

If we apply the replicator dynamics from the initial distribution $x_0 = \begin{pmatrix} .2 & .3 & .1 & .4 \end{pmatrix}^\top$, we obtain the trajectories shown in Figure 3.1.

Starting from different initial conditions in the interior of the simplex, $\mathring{\Delta}$, we have convergence to different points in the set of Nash equilibria $\mathcal{N}$. This is illustrated in Figure 3.2. If we start on the boundary of the simplex, we may have convergence to stationary points which are not Nash equilibria. Any face of the simplex is invariant for the dynamics (and so is any intersection of faces). Therefore a stationary point which is unstable in the entire simplex may be stable if we restrict the dynamics to an invariant face.

In the next chapter, we will go back to studying discrete-time dynamics, and show that a large family of online learning methods can be obtained by discretizing the replicator ODE, and prove convergence of these methods.

Chapter 4

Discretizing the Replicator Dynamics

In this chapter, we study online learning algorithms that can be obtained as a discretization of the replicator ODE (3.1). We start with a simple forward Euler discretization in Section 4.1 with decreasing step sizes, and show that the resulting algorithm has sublinear regret guarantees similar to the Hedge algorithm reviewed in Section 2.6. We then consider a general family of discrete-time, stochastic algorithms, obtained as a stochastic discretization of the ODE. More precisely, if the ODE is given by $\dot{X}(t) = F(X(t))$, then a stochastic discretization with step sizes (η_τ) is given by $x^{(\tau+1)} = x^{(\tau)} + \eta_\tau(F(x^{(\tau)}) + U^{(\tau)})$, where $U^{(\tau)}$ is a sequence of stochastic perturbations that satisfy conditions which are detailed in Proposition 14. We call this family the AREP algorithms, for approximate replicator. The theory of stochastic approximation exhibits links between the properties of solution trajectories of the ODE (such as convergence to stable stationary points) and properties of its discretization. In Section 4.2, we review and illustrate some of these results, which are mostly lifted from [18]. Then we show, in Theorem 8 that under AREP algorithms, the sequence of mass distributions is guaranteed to converge to the set of Nash equilibria, almost surely. This is a strong convergence result for a large class of algorithms, in particular, it is stronger than the convergence in the sense of Cesàro for sublinear regret algorithms, which we proved in Theorem 2. However, the stochastic approximation analysis does not allow us to characterize convergence rates of the algorithms. In the next chapter, we consider a particular class of learning dynamics, given by the stochastic mirror descent method, which we can analyze to provide convergence rates.

4.1 Euler discretization of the replicator ODE: the REP algorithm

Inspired by the continuous-time replicator dynamics, we propose a discrete-time update rule for online learning, by discretizing the ODE (3.1). The resulting algorithm guarantees sublinear regret and is simple to implement. We call it the REP algorithm in reference to the replicator ODE.

Since we seek to provide guarantees on the regret of the algorithm for any sequence of bounded loss vectors (not necessarily loss vectors of a potential game, $\ell(x) = \nabla f(x)$), we first decouple, in the ODE (3.1), the dependence on the mass distribution x and the loss ℓ. The vector field F can be written in the following form: for all k, $F_k(x) = G_k(x, \ell(x))$ where for all a,

$$G_{k,a}(x, \ell) = x_{k,a} \left(\langle x_k, \ell_k \rangle - \ell_{k,a} \right).$$

Then suppose that we are given a C^1 function of time, $\ell_k(t)$, and a distribution $\pi(t)$ that obeys the dynamics

$$\dot{\pi}_{k,a}(t) = G_{k,a}(\pi(t), \ell(t)) = \pi_{k,a}(t) \left(\langle \pi_k(t), \ell_k(t) \rangle - \ell_{k,a}(t) \right). \tag{4.1}$$

Using a forward Euler discretization (see e.g. Chapter 2 in [36]) with decreasing, non-summable step sizes (η_τ), we can define discrete times T_τ such that $T_0 = 0$ and $T_{\tau+1} = T_\tau + \eta_\tau$, and let $\pi^{(\tau)} = \pi(T_\tau)$, and $\ell_k^{(\tau)} = \ell_k(T_\tau)$. Then approximating the time derivative

$$\dot{\pi}(T_\tau) \approx \frac{\pi(T_\tau + \eta_\tau) - \pi(T_\tau)}{\eta_\tau} = \frac{\pi^{(\tau+1)} - \pi^{(\tau)}}{\eta_\tau},$$

we propose the following update rule:

Definition 8 (Discrete Replicator algorithm)**.** *Consider an online learning problem with a sequence of losses $(\ell^{(\tau)})$ in $[0, \rho]$ uniformly in τ. The REP algorithm with initial distribution $\pi^{(0)} \in \Delta^{\mathcal{A}}$ and learning rates $(\eta_\tau)_{\tau \in \mathbb{N}}$ with $\eta_\tau \leq \frac{1}{\rho}$, is an online learning algorithm $(U^{(\tau)})$ such that the τ-th update function is given by*

$$\pi_a^{(\tau+1)} = U^{(\tau)}(\pi^{(\tau)}, \ell^{(\tau)})_a = \pi_a^{(\tau)} + \eta_\tau \pi_a^{(\tau)} \left(\left\langle \pi^{(\tau)}, \ell^{(\tau)} \right\rangle - \ell_a^{(\tau)} \right). \tag{4.2}$$

Algorithm 4 REP algorithm with learning rates (η_t).

1: Input: Initial distribution $\pi^{(0)} \in \Delta^{\mathcal{A}}$.
2: **for** each iteration $\tau \in \mathbb{N}$ **do**
3: Draw an action $A^{(\tau)} \sim \pi^{(\tau)}$.
4: Observe a vector of losses $\ell^{(\tau)}$, incur loss $\ell_{A^{(\tau)}}^{(\tau)}$.
5: Update

$$\pi_a^{(\tau+1)} = \pi_a^{(\tau)} + \eta_\tau \pi_a^{(\tau)} \left(\left\langle \pi^{(\tau)}, \ell^{(\tau)} \right\rangle - \ell_a^{(\tau)} \right).$$

6: **end for**

Under the REP update, the sequence of distributions $\pi^{(\tau)}$ remains in simplex $\Delta^{\mathcal{A}}$, provided $\eta_\tau \leq \frac{1}{\rho}$: Indeed, for all $\tau \in \mathbb{N}$, we have

$$\sum_{a \in \mathcal{A}} \pi_a^{(\tau+1)} = \sum_{a \in \mathcal{A}} \pi_a^{(\tau)} + \eta_\tau \pi_a^{(\tau)} \left(\left\langle \ell^{(\tau)}, \pi^{(\tau)} \right\rangle - \ell_a^{(\tau)} \right) = \sum_{a \in \mathcal{A}} \pi_a^{(\tau)},$$

and if $\eta_\tau \leq 1$, then

$$1 + \eta_\tau \left(\left\langle \ell^{(\tau)}, \pi^{(\tau)} \right\rangle - \ell_a^{(\tau)} \right) \geq 1 - \rho\eta_\tau \geq 0,$$

which guarantees that $\pi^{(\tau)}$ remains in $\Delta^{\mathcal{A}}$.

We now show that when the losses are discounted by (γ_τ), the REP update rule with learning rates proportional to (γ_τ) has sublinear regret. First, we prove the following lemma, for general online learning problems with signed losses.

Lemma 7. *Consider an online learning problem, with a finite action set $\mathcal{A}$, and sequence of losses, $m^{(\tau)}$, and suppose that the losses are bounded uniformly in $[-1, 1]$. Suppose that the losses are discounted by a sequence of discount factors (γ_τ), with $\gamma_\tau \leq \frac{1}{2}$ for all τ. Then the learning algorithm defined by the update rule*

$$\pi^{(\tau+1)} \propto \left(\pi_a^{(\tau)} (1 - \gamma_\tau m_a^{(\tau)}) \right)_{a \in \mathcal{A}} \tag{4.3}$$

has the following regret bound: for all T and all $a \in \mathcal{A}$,

$$\sum_{0 \leq \tau \leq T} \gamma_\tau \left\langle m^{(\tau)}, \pi^{(\tau)} \right\rangle \leq -\log \pi_{\min}^{(0)} + \sum_{0 \leq \tau \leq T} \gamma_\tau m_a^{(\tau)} + \sum_{0 \leq \tau \leq T} \gamma_\tau^2 |m_a^{(\tau)}|,$$

where $\pi_{\min}^{(0)} = \min_{a \in \mathcal{A}} \pi_a^{(0)}$.

Proof. We extend the proof of Theorem 2.1 in [4] to the discounted case. By a simple induction, we have for all t, $\pi^{(t)}$ is proportional to the vector $w^{(t)}$ defined by

$$w_a^{(t)} = \pi_a^{(0)} \prod_{0 \leq \tau < t} (1 - \gamma_\tau m_a^{(\tau)}).$$

Define the function $\xi^{(t)} = \sum_a w_a^{(t)}$. Then $\pi_a^{(t)} = \frac{w_a^{(t)}}{\xi^{(t)}}$, and we have for all t:

$$\begin{aligned}
\xi^{(t+1)} &= \sum_a w_a^{(t+1)} \\
&= \sum_a w_a^{(t)} (1 - \gamma_t m_a^{(t)}) \\
&= \xi^{(t)} - \gamma_t \sum_a m_a^{(t)} \pi_a^{(t)} \xi^{(t)} \\
&= \xi^{(t)} \left(1 - \gamma_t \left\langle m^{(t)}, \pi^{(t)} \right\rangle \right) \\
&\leq \xi^{(t)} e^{-\gamma_t \left\langle m^{(t)}, \pi^{(t)} \right\rangle}.
\end{aligned}$$

Thus, by induction on t,

$$\xi^{(t+1)} \leq \exp\left(- \sum_{0 \leq \tau \leq t} \gamma_\tau \left\langle m^{(\tau)}, \pi^{(\tau)} \right\rangle \right).$$

We also have for all a,

$$\xi^{(t+1)} \geq w_a^{(t+1)} \geq \pi_{\min}^{(0)} \prod_{0 \leq \tau \leq t} (1 - \gamma_t m_a^{(\tau)}).$$

Combining the previous bounds on $\xi^{(T)}$ and taking logarithms, we have

$$\sum_{0 \leq \tau \leq T} \gamma_\tau \left\langle m^{(\tau)}, \pi^{(\tau)} \right\rangle \leq -\log \pi_{\min}^{(0)} - \sum_{0 \leq \tau \leq T} \log(1 - \gamma_\tau m_a^{(\tau)}).$$

To obtain the desired bound, it suffices to observe that for all $m \in [-1, 1]$ and $\gamma \in [0, \frac{1}{2}]$, $-\log(1 - \gamma m) \leq \gamma m + \gamma^2 |m|$. □

Theorem 6. *Consider an online learning problem, with action set $\mathcal{A}$, and sequence of uniformly bounded losses, $\ell^{(\tau)} \in [0, \rho]$. Suppose that the losses are discounted by a sequence (γ_τ) that satisfies Assumption 2 and is bounded by $\frac{1}{2}$. Then the regret of the REP algorithm with learning rates $\eta_\tau = \frac{\gamma_\tau}{\rho}$ satisfies the following bound*

$$R^{(T)}(s) \leq -\rho \log \pi_{\min}^{(0)} + \rho \sum_{0 \leq \tau \leq T} \gamma_\tau^2.$$

In particular, when $\frac{\sum_{0 \leq \tau \leq T} \gamma_\tau^2}{\sum_{0 \leq \tau \leq T} \gamma_\tau} \to 0$, we have $\lim_{T \to \infty} \frac{\left[R^{(T)}\right]_+}{\sum_{\tau \leq T} \gamma_\tau} \leq 0$.

Proof. Let

$$r_a^{(\tau)} = \left\langle \pi^{(\tau)}, \ell^k(x^{(\tau)}) \right\rangle - \ell_a(x^{(\tau)}) \in [-\rho, \rho]$$

be the *instantaneous regret* of the player. Then the REP update can be viewed as a multiplicative-weights algorithm with update rule (4.3), in which the vector of signed losses is given by $m_a^{(\tau)} = -\frac{r_a{}^{(\tau)}}{\rho} \in [-1, 1]$, and discount factors (γ_τ). Observing that $\left\langle r^{(\tau)}, \pi^{(\tau)} \right\rangle = 0$, we have by Lemma 7, for all $a \in \mathcal{A}$:

$$\frac{1}{\rho} \sum_{0 \leq \tau \leq T} \gamma_\tau r_a^{(\tau)} \leq -\log \pi_{\min}^{(0)} + \sum_{0 \leq \tau \leq T} \gamma_\tau^2.$$

Rearranging and taking the maximum over $a \in \mathcal{A}$, we have

$$R^{(T)}(s) \leq -\rho \log \pi_{\min}^{(0)} + \rho \sum_{0 \leq \tau \leq T} \gamma_\tau^2,$$

which proves the claim. □

Interestingly, we can show the the REP update rule corresponds to the solution to a regularized version of the greedy update $\min_{\pi \in \Delta^{\mathcal{A}}} \left\langle \pi, \ell^{(\tau)} \right\rangle$.

Proposition 13. *The REP update rule* (4.2) *with* $\eta_\tau < \frac{1}{\rho}$ *and* $\pi^{(\tau)} > 0$ *is solution to the following problem:*

$$\{\pi^{(\tau+1)}\} = \arg\min_{\pi\in\Delta} \eta_\tau \left\langle \pi, \ell^{(\tau)} \right\rangle + D(\pi\|\pi^{(\tau)}),$$

where $D(\pi\|\nu) = \frac{1}{2}\sum_{a\in\mathcal{A}} \nu_a \left(\frac{\pi_a}{\nu_a} - 1\right)^2$.

The regularization function $D(\pi\|\nu)$ is known as the $\mathcal{X}^2$ divergence, and can be interpreted as the f-divergence (or Csiszàr divergence, in reference to Csiszár [42]), associated to the convex function $f(x) = (x-1)^2$. Note that the Hedge algorithm has a similar interpretation as the minimizer of a regularized linear function, in which the $\mathcal{X}^2$ divergence is replace by the KL divergence, see for example [15]. This will also be discussed in Chapter 5.

Proof. Define the Lagrangian of the problem: for $\lambda \in \mathbb{R}_+^{\mathcal{A}}$ and $\nu \in \mathbb{R}$,

$$L(\pi, \lambda, \nu) = \eta_\tau \left\langle \pi_a, \ell^{(\tau)} \right\rangle + \frac{1}{2}\sum_{a\in\mathcal{A}} \pi_a^{(\tau)} \left(\frac{\pi_a}{\pi_a^{(\tau)}} - 1\right)^2 - \nu\left(\sum_{a\in\mathcal{A}} \pi_a - 1\right) - \langle \lambda, \pi\rangle ,$$

where $\nu \in \mathbb{R}$ is the dual variable for the constraint $\sum_{a\in\mathcal{A}} \pi_a = 1$ and λ is the dual variable for the constraint $\pi \geq 0$. The gradient of $\mathcal{L}$ with respect to π is given by

$$\forall a \in \mathcal{A},\ \frac{\partial}{\partial \pi_a} L(\pi, \lambda, \nu) = \eta_\tau \ell_a^{(\tau)} + \left(\frac{\pi_a}{\pi_a^{(\tau)}} - 1\right) - \lambda_a - \nu.$$

Then $(\pi^\star, \lambda^\star, \nu^\star)$ are primal-dual optimal if and only if they satisfy the following KKT conditions:

$$\begin{cases} \frac{\pi_a^\star}{\pi_a^{(\tau)}} = 1 + \lambda_a^\star + \nu^\star - \eta_\tau \ell_a^{(\tau)} \\ \sum_{a\in\mathcal{A}} \pi_a^\star = 1 \\ \pi^\star \geq 0. \end{cases}$$

Multiplying by $\pi_a^{(\tau)}$ and taking the sum over $a \in \mathcal{A}$, we have

$$1 = 1 - \eta_\tau \left\langle \pi^{(\tau)}, \ell^{(\tau)} \right\rangle + \left\langle \pi^{(\tau)}, \lambda^\star \right\rangle + \nu^\star,$$

i.e. $\nu^\star = \eta_\tau \left\langle \pi^{(\tau)}, \ell^{(\tau)} \right\rangle - \left\langle \lambda^\star, \ell^{(\tau)} \right\rangle$. Plugging in the expression of $\pi^\star$,

$$\pi_a^\star = \pi_a^{(\tau)} + \eta_\tau \pi_a^{(\tau)} (\left\langle \pi^{(\tau)}, \ell^{(\tau)} \right\rangle - \ell_a^{(\tau)}) + \pi_a^{(\tau)} (\lambda_a^\star - \left\langle \pi^{(\tau)}, \lambda^\star \right\rangle). \tag{4.4}$$

To conclude, it suffices to show that $\lambda^\star = 0$. Note that since the losses are in $[0, \rho]$, we have for all a, $\left\langle \pi^{(\tau)}, \ell^{(\tau)} \right\rangle - \ell_a^{(\tau)} \geq -\rho$, and since $\eta_\tau \leq \frac{1}{\rho}$, $1 + \eta_\tau (\left\langle \pi^{(\tau)}, \ell^{(\tau)} \right\rangle - \ell_a^{(\tau)}) > 0$.

Now, we have for all $a \in \operatorname{support} \lambda^\star$, by the complementary slackness condition, $\pi_a^\star = 0$, thus

$$\lambda_a^\star - \left\langle \pi^{(\tau)}, \lambda^\star \right\rangle = -\eta_\tau (\left\langle \pi^{(\tau)}, \ell^{(\tau)} \right\rangle - \ell_a^{(\tau)}) < 0.$$

Suppose by contradiction that $\lambda^\star$ has a nonempty support, then multiplying the last inequality by $\pi_a^{(\tau)}$ and summing over $a \in \operatorname{support} \lambda^\star$, we would have $\left\langle \pi^{(\tau)}, \lambda^\star \right\rangle < \left\langle \pi^{(\tau)}, \lambda^\star \right\rangle$, a contradiction. Therefore $\lambda^\star = 0$ and the expression (4.4) of $\pi^\star$ reduces to the REP update rule (4.2). □

4.2 Results from the theory of stochastic approximation

In the remainder of the chapter, our goal will be to define a family of discrete-time learning algorithms which guarantee the convergence of the mass distributions $(x^{(\tau)})$. The idea is to show that the discrete process $(x^{(\tau)})_{\tau\in\mathbb{N}}$, approaches, in a certain sense that we will make precise in this section, the solution trajectories of the continuous-time replicator ODE. Then one can show, using a Lyapunov function, that any limit point of the discrete process must lie in the set of stationary points $\mathcal{RN}$. With an additional argument, we show that, in fact, limit points lie in the set $\mathcal{N}$ of Nash equilibria.

We start by reviewing results from the theory of stochastic approximation. The results of this section are adapted from [18], due to Benaïm. Let $\mathcal{D}$ be a bounded subset of $\mathbb{R}^n$, and consider a dynamical system given by the ODE

$$\dot{X}(t) = F(X(t)) \tag{4.5}$$

where $F : \mathcal{D} \to \mathbb{R}^n$ is a continuous globally integrable vector field, with unique integral curves which remain in $\mathcal{D}$. Let Φ be the associated flow function such that $t \mapsto \Phi_t(x^{(0)})$ is the solution trajectory of (4.5) with initial condition $X(0) = x^{(0)}$.

Discrete-time approximation and asymptotic pseudo trajectory

Let $(x^{(\tau)})_\tau$ be a discrete-time (stochastic) process with values in $\mathcal{D}$, and $(\tau_\tau)_\tau$ a sequence of positive real numbers (step sizes) such that $\sum_{\tau\in\mathbb{N}} \eta_\tau = \infty$ and $\lim_{\tau\to\infty} \eta_\tau = 0$. We say that $(x^{(\tau)})_\tau$ is a discrete-time approximation of the dynamical system (4.5) with step sizes (η_τ) and perturbations $(U^{(\tau)})$ if it satisfies, $\forall\tau$,

$$x^{(\tau+1)} - x^{(\tau)} = \eta_\tau \left(F(x^{(\tau)}) + U^{(\tau+1)}\right). \tag{4.6}$$

Note that it is always possible to define a sequence of perturbations $U^{(\tau)}$ such that Equation (4.6), simply by defining $U^{(\tau+1)} = \frac{x^{(\tau+1)}-x^{(\tau)}}{\eta_\tau} - F(x^{(\tau)})$. However, in order to relate the asymptotic behavior of the discrete process $x^{(\tau)}$ to the solution trajectories of the ODE (4.5), one needs to impose additional assumptions on the perturbations $U^{(\tau)}$, as we discuss next.

Given such a discrete-time approximation, we can define the affine interpolated process of $(x^{(\tau)})$: let $T_\tau = \sum_{t=0}^{\tau} \eta_t$ as in Section 4.1.

Definition 9 (Affine interpolated process)**.** *The continuous time affine interpolated process of the discrete process $(x^{(\tau)})_{\tau\in\mathbb{N}}$ is the function $M : \mathbb{R}_+ \to \mathbb{R}^n$ defined as*

$$M(T_\tau + s) = x^{(\tau)} + s\frac{x^{(\tau+1)} - x^{(\tau)}}{\eta_\tau}, \quad \forall\tau \in \mathbb{N} \text{ and } \forall s \in [0, \eta_\tau).$$

We now define what it means for a discrete process to approach the trajectories of the system (4.5). A continuous function X is said to be an asymptotic pseudo trajectory for the flow Φ if for all T,

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} \|X(t+h) - \Phi_h(X(t))\| = 0.$$

Asymptotic pseudo trajectories (APT) have been introduced by Benaïm and Hirsch in [20], and further studied in [17]. They are useful in relating the asymptotic behavior of discrete processes to the solutions of the ODE. The next proposition gives sufficient conditions for an affine interpolated process to be an APT.

Proposition 14 (Proposition 4.1 in [18])**.** *Let M be the affine interpolated process of the discrete-time approximation $(x^{(\tau)})$, and assume that for all $T > 0$*

$$\lim_{\tau_1\to\infty} \max_{\tau_2:\ \sum_{\tau=\tau_1}^{\tau_2} \eta_\tau < T} \left\| \sum_{\tau=\tau_1}^{\tau_2} \eta_\tau U^{(\tau+1)} \right\| = 0. \tag{4.7}$$

Then M is an APT of the flow Φ induced by the vector field F.

Furthermore, we have the following sufficient condition for property (4.7) to hold:

Proposition 15. *Let $(x^{(\tau)})_{\tau\in\mathbb{N}}$ be a discrete time approximation of the system (4.5). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_\tau)_{\tau\in\mathbb{N}}$ a filtration of $\mathcal{F}$. Suppose that the perturbations satisfy the Robbins-Monro conditions: for all $\tau \in \mathbb{N}$,*

i) $U^{(\tau)}$ is measurable with respect to $\mathcal{F}_\tau$

ii) $\mathbb{E}[U^{(\tau+1)}|\mathcal{F}_\tau] = 0$

Furthermore, suppose that there exists $q \geq 2$ such that

$$\begin{cases} \sup_{\tau\in\mathbb{N}} \mathbb{E}[\|U^{(\tau)}\|^q] < \infty \\ \sum_{\tau\in\mathbb{N}} \eta_\tau^{1+q/2} < \infty. \end{cases}$$

Then, condition (4.7) of Proposition 14 holds with probability one.

Chain transitivity

Next, Theorem 7 gives an important property of limit points of bounded asymptotic pseudo-trajectories.

Definition 10 (Pseudo-orbit and chain transitivity)**.** *A (δ, T)-pesudo-orbit from $a \in \mathcal{D}$ to $b \in \mathcal{D}$ is a finite sequence of partial trajectories. It is given by a sequence of points $(T_\tau, x^{(\tau)}), \tau \in \{0, \ldots, m-1\}$ (with $T_\tau \geq T$ for all τ) and the corresponding sequence of partial trajectories*

$$\{\Phi_t(x^{(\tau)}) \colon 0 \leq t \leq T_\tau\};\ \tau = 0, \ldots, m-1,$$

such that $d(x^{(0)}, a) < \delta$, $d(\Phi_{T_\tau}(x^{(\tau)}), x^{(\tau+1)}) < \delta$ *for all* τ, *and* $x^{(m)} = b$.

The conditions are illustrated in Figure 4.1. We write $a \hookrightarrow_{\delta,T} b$ *if there exists a* (δ, T)*-pesudo-orbit from* a *to* b. *We write* $a \hookrightarrow b$ *if* $a \hookrightarrow_{\delta,T} b$ *for all* $\delta, T > 0$. *The flow* Φ *is said to be chain transitive if* $a \hookrightarrow b$ *for all* $a, b \in \mathcal{D}$.

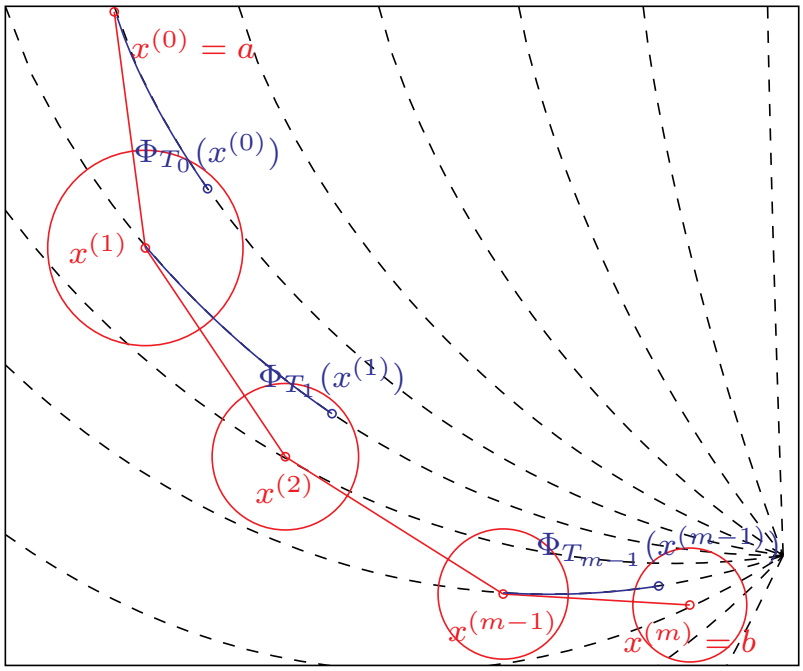


Figure 4.1: A (δ, T)-pseudo-orbit for the flow Φ, from a to b.

In the remainder of this section, let $\Gamma \subset \mathcal{D}$ be a compact invariant set for Φ, that is, $\Phi_t(\Gamma) \subseteq \Gamma$ for all $t \in \mathbb{R}_+$.

Definition 11 (Internally chain transitive set)**.** *The compact invariant set* Γ *is internally chain transitive if the restriction of* Φ *to* Γ *is chain transitive.*

Theorem 7 (Theorem 5.7 in [18])**.** *Let* M *be a bounded APT of* (4.5). *Then the limit set*

$$L(M) = \bigcap_{t \geq 0} \mathrm{cl}\,\{M(s) \colon s \geq t\}$$

is internally chain transitive. Here cl *denotes the closure of a set.*

Finally, we give the following property of internally chain transitive sets:

Proposition 16 (Proposition 6.4 in [18])**.** *Let* $\Gamma \subset \mathcal{D}$ *be a compact invariant set and suppose that there exists a Lyapunov function* $V : \mathcal{D} \to \mathbb{R}$ *for* Γ *(that is,* V *is continuous and the time derivative* $\langle \nabla V(x), F(x) \rangle < 0$ *for all* $x \notin \Gamma$*) such that* $V(\Gamma)$ *has empty interior. Then every internally chain transitive set* L *is contained in* Γ *and* V *is constant on* L.

Proposition 16, combined with Theorem 7, provides a powerful tool for proving convergence of a discrete time approximation $(x^{(\tau)})$ of the ODE: if Γ is an invariant set for the ODE, and we can exhibit a Lyapunov function for Γ, then we know by Theorem 7 and Proposition 16 that the limit set $L(M)$ of any APT M is contained in Γ. This is true in particular for the affine interpolated process of $(x^{(\tau)})$ if the discretization satisfies the conditions of Proposition 14. But any limit point of the sequence $(x^{(\tau)})$ is also a limit point of its affine interpolated process, which proves that all limit points of $(x^{(\tau)})$ lie inside the compact set Γ, i.e. that $(x^{(\tau)})$ converges to Γ.

4.3 The approximate replicator class (AREP)

Now, we are ready to define a class of online learning algorithms which we call AREP for approximate replicator. An AREP online algorithm is simply a discrete time approximation, in the sense of (4.6), of the replicator ODE (4.1), with perturbations that satisfy the condition of Proposition 14.

Definition 12 (AREP algorithm)**.** *Consider an online learning problem with action set $\mathcal{A}$ and loss sequence $(\ell^{(\tau)})$, such that the losses are uniformly bounded in $[0, \rho]$. An online learning algorithm with output sequence $(\pi^{(\tau)})_{\tau\in\mathbb{N}}$, is said to be an approximate replicator (AREP) algorithm if its update equation can be written as*

$$\pi_a^{(\tau+1)} - \pi_a^{(\tau)} = \eta_\tau \left(\pi_a^{(\tau)}(\left\langle \pi^{(\tau)}, \ell^{(\tau)} \right\rangle - \ell_a^{(\tau)}) + U_a^{(\tau)}\right) \tag{4.8}$$

where $(U^{(\tau)})_{\tau\in\mathbb{N}}$ is a sequence of stochastic perturbations with values in $\mathbb{R}^{\mathcal{A}}$, and which satisfies condition (4.7) *a.s.*

In particular, the REP algorithm given in Definition 8 is an AREP algorithm in which the perturbations are identically zero. It turns out that the Hedge algorithm also belongs to the AREP class, as shown in the following proposition.

Proposition 17. *The Hedge algorithm with non-increasing, non-summable learning rates (η_τ) is an AREP algorithm.*

Proof. Let $(\pi^{(\tau)})_{\tau\in\mathbb{N}}$ be the sequence of strategies, and let $(\ell^{(\tau)})$ be any sequence of losses. By definition of the Hedge algorithm, we have

$$\pi_a^{(\tau+1)} = \pi_a^{(\tau)} \exp\left(-\eta_\tau \ell_a^{(\tau)}\right) / \sum_{a'\in\mathcal{A}} \pi_{a'}^{(\tau)} \exp\left(-\eta_\tau \ell_{a'}^{(\tau)}\right),$$

which we can write in the form of Equation (4.8), with perturbation terms

$$U_a^{(\tau+1)} = \frac{\pi_a^{(\tau)}}{\eta_\tau}\left[\exp\left(-\eta_\tau(\ell_a^{(\tau)} - \tilde{\ell}^{(\tau)})\right) + \eta_\tau(\ell_a^{(\tau)} - \tilde{\ell}^{(\tau)}) - 1\right] + \pi_a^{(\tau)}(\tilde{\ell}^{(\tau)} - \bar{\ell}^{(\tau)})$$

where

$$\tilde{\ell}^{(\tau)} = -\frac{1}{\eta_\tau} \log \sum_{a' \in \mathcal{A}} \pi_{a'}^{(\tau)} \exp\left(-\eta_\tau \ell_{a'}(x^{(\tau)})\right),$$
$$\bar{\ell}^{(\tau)} = \left\langle \pi^{(\tau)}, \ell^{(\tau)} \right\rangle.$$

Letting $\theta(x) = e^x - x - 1$, we have for all $a \in \mathcal{A}$:

$$U_a^{(\tau+1)} = \frac{\pi_a^{(\tau)}}{\eta_\tau} \theta\left(-\eta_\tau(\ell_a^{(\tau)} - \tilde{\ell}^{(\tau)}\right) + \pi_a^{(\tau)}(\tilde{\ell}^{(\tau)} - \bar{\ell}^{(\tau)}).$$

The first term is a $\mathcal{O}(\eta_\tau)$ as $\theta(x)$ is equivalent to $x^2/2$ as x tends to 0. To bound the second term, we have by concavity of the logarithm

$$\tilde{\ell}^{(\tau)} = -\frac{1}{\eta_\tau} \log \sum_{a' \in \mathcal{A}} \pi_{a'}^{(\tau)} \exp\left(-\eta_\tau \ell_{a'}^{(\tau)}\right) \leq \sum_{a' \in \mathcal{A}} \pi_{a'}^{(\tau)} \ell_{a'}^{(\tau)} = \bar{\ell}^{(\tau)}.$$

And by Hoeffding's lemma,

$$\log \sum_{a' \in \mathcal{A}} \pi_{a'} \exp\left(-\eta_\tau \ell_{a'}^{(\tau)}\right) \leq -\eta_\tau \sum_{a' \in \mathcal{A}} \pi_{a'}^{(\tau)} \ell_{a'}^{(\tau)} + \frac{\eta_\tau^2}{8}.$$

Rearranging, we have $0 \leq \bar{\ell}^{(\tau)} - \tilde{\ell}^{(\tau)} \leq \frac{\eta_\tau}{8}$, therefore $U_a^{(\tau+1)} = \mathcal{O}(\eta_\tau)$, and

$$\left\| \sum_{\tau=\tau_1}^{\tau_2} \eta_\tau U^{(\tau+1)} \right\| = \mathcal{O}\left(\sum_{\tau=\tau_1}^{\tau_2} \eta_t^2 \right).$$

Finally, since $\eta_\tau \downarrow 0$, for any fixed T, $\max_{\tau_2 : \sum_{\tau=\tau_1}^{\tau_2} \eta_\tau \leq T} \sum_{\tau_1}^{\tau_2} \eta_\tau^2$ converges to 0 as $\tau_1 \to \infty$, therefore condition (4.7) is verified. □

4.4 Convergence of AREP

We now give the main convergence result of this chapter.

Theorem 8. *Consider the online learning model in nonatomic, convex potential games, defined in Section 2.4, and suppose that the mass distributions $(x^{(\tau)})_\tau$ have sublinear regret, and obey an AREP update rule, where, for each population k the loss function $\ell_k^{(\tau)}$ is given by the congestion game loss $\ell_k(x^{(\tau)})$. That is, for all k and all $a \in \mathcal{A}_k$,*

$$x_{k,a}^{(\tau+1)} - x_{k,a}^{(\tau)} = \eta_\tau \left(x_{k,a}^{(\tau)} \left(\left\langle x_k^{(\tau)}, \ell_k(x^{(\tau)}) \right\rangle - \ell_{k,a}(x^{(\tau)}) \right) + U_{k,a}^{(\tau)} \right)$$

where $U_k^{(\tau)}$ are sequences of stochastic perturbations that satisfy condition (4.7). Then $(x^{(\tau)})$ converges to the set of Nash equilibria $\mathcal{N}$, almost surely.

Proof. By Proposition 14, the affine interpolated process M of $(x^{(\tau)})_\tau$ is an APT of the continuous-time replicator ODE, $\dot{X} = F(X)$. Thus by Theorem 7, the limit set $L(M)$ is internally chain transitive.

Consider the set of stationary points $\mathcal{RN}$, which is invariant by definition, and by Proposition 8, it is compact and the potential function f takes finitely many values on $\mathcal{RN}$. Additionally, f is a Lyapunov function for $\mathcal{RN}$ by Proposition 9, therefore we can apply Proposition 16 to conclude that the set of limit points $L(M)$ is contained in $\mathcal{RN}$ and f is constant over $L(M)$. Let $v^\star$ be this constant value.

Next, we show that $f(x^{(\tau)})$ converges to $v^\star$. Let $\hat{v}$ be a limit point of $f(x^{(\tau)})$. Then by Lemma 2, $\hat{v} = f(\hat{x})$ where $\hat{x}$ is a limit point of $(x^{(\tau)})$. But $\hat{x} \in L(M)$, thus $\hat{v} = f(\hat{x}) = v^\star$. This shows that the bounded sequence $(f(x^{(\tau)}))$ has a unique limit point $v^\star$, therefore it converges to $v^\star$. To conclude, it suffices to show that $v^\star = f^\star$, the minimum of the potential function f. To show that $v^\star = f^\star$, observe that since $f(x^{(\tau)}) \to v^\star$, we also have $f(x^{(\tau)}) \xrightarrow{(\gamma_\tau)} v^\star$. But since the populations have sublinear discounted regret, by Theorem 2, $f(x^{(\tau)}) \xrightarrow{(\gamma_\tau)} f^\star$. By uniqueness of the limit, we must have $v^\star = f^\star$. □

Note that Theorem 8 assumes that the AREP update rule is applied to the population dynamics $(x^{(\tau)})$, not to individual strategies $\pi^{(\tau)}(s)$. One sufficient condition for $x^{(\tau)}$ to satisfy an AREP update is that for each k, all players in $\mathcal{S}_k$ start from a common initial distribution $\pi_k^{(0)} = x_k^{(0)}$, and apply the same update rule. This guarantees that for all τ and for all $s \in \mathcal{S}_k$, $x_k^{(\tau)} = \pi_k^{(\tau)}(s)$.

Convergence of the REP and Hedge algorithms

We apply Theorem 8 to show convergence of the REP and Hedge algorithms.

Corollary 1. *In nonatomic, convex potential games, if $(x^{(\tau)})$ obeys the REP update rule with non-increasing, non-summable learning rates (η_τ) and such that $\eta_\tau \leq \frac{1}{2\rho}$, then $x^{(\tau)} \to \mathcal{N}$.*

Proof. The REP update rule is a discounted no-regret algorithm by Theorem 6, and it is an AREP algorithm with zero perturbations, so we can apply Theorem 8. □

Corollary 2. *In nonatomic, convex potential games, if $(x^{(\tau)})$ obeys the Hedge update rule with non-increasing, non-summable learning rates η_τ, then $x^{(\tau)} \to \mathcal{N}$.*

Proof. By Proposition 6, Hedge has sublinear regret, and by Proposition 17, Hedge is an AREP algorithm, and we can apply Theorem 8. □

A numerical example

We illustrate this convergence result with a routing game example, as defined in Section 2.2. Consider the network given in Figure 4.2. In this example, two populations of players share the network, the first population sends packets from v_0 to v_1, and the second

population from v_2 to v_3. The paths (actions) available to each population are given by $\mathcal{A}_1 = \{(v_0, v_1), (v_0, v_4, v_5, v_1), (v_0, v_5, v_1)\}$, $\mathcal{A}_2 = \{(v_2, v_3), (v_2, v_4, v_5, v_3), (v_2, v_4, v_3)\}$.

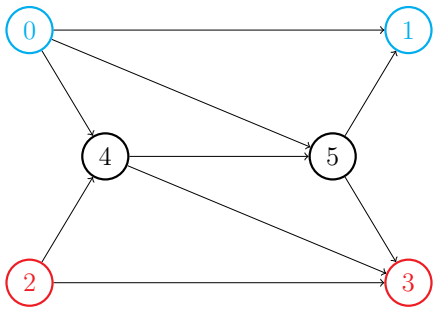


Figure 4.2: Routing game with two populations of players.

We simulate the population dynamics under the discounted Hedge algorithm with a harmonic sequence of learning rates, $\eta_\tau = \frac{1}{\tau}$. The results are shown in Figure 4.3.

Conclusion

Starting from the replicator ODE and the properties of its solution trajectories, we showed that a family of discrete time algorithms can be obtained by taking a discretization of the ODE, with perturbations. We showed that under suitable conditions on the perturbations (Proposition 14), the sequence of distributions is guaranteed to converge to the set of Nash equilibria, almost surely. However, the stochastic approximation analysis done in this chapter does not allow us to characterize convergence rates of the discrete dynamics. In the next chapter, we will study a second class of dynamics for which we can derive explicit convergence rates.

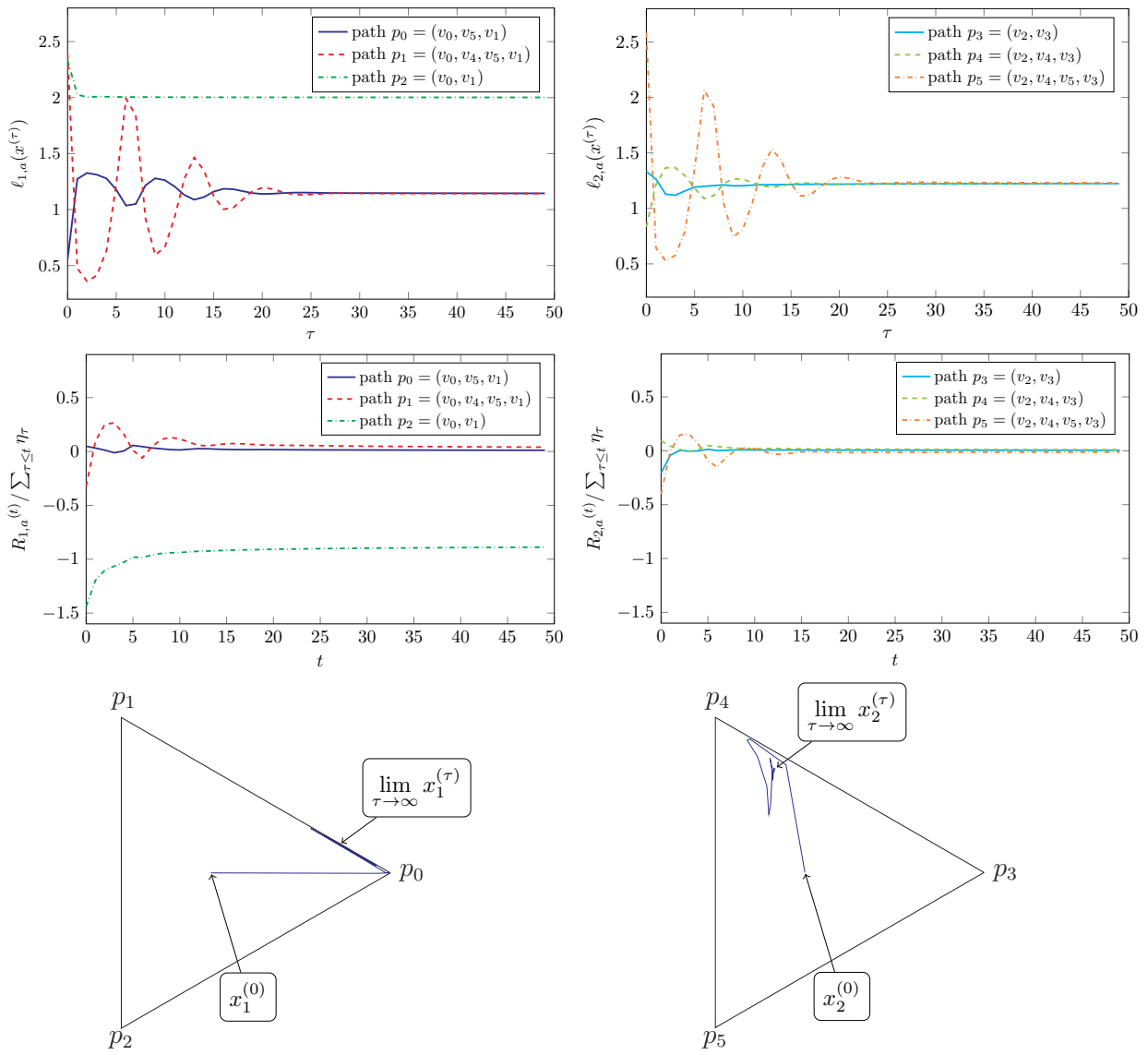


Figure 4.3: Simulation of the population dynamics under the discounted Hedge algorithm, initialized at the uniform distribution. The trajectories of the population strategies $x_k^{(\tau)}$ are shown in the simplex for each population (bottom). The path losses $\ell_{k,a}(x^{(\tau)})$ (top) converge to a common value on the support of the Nash equilibrium. The sequences of discounted regrets (middle) confirm that the population regret is sub-linear, i.e. $\limsup_{t \to \infty} \frac{[R_k^{(t)}]_+}{\sum_{\tau \le t} \eta_\tau} = 0$.

Chapter 5

Stochastic Mirror Descent Dynamics

In this chapter, we seek to design learning algorithms for the convex potential games, that are robust to measurement noise and other stochastic perturbations, and for which we can provide convergence rates. More precisely, we extend the previous learning model defined in Chapter 2, by assuming that, at each iteration, instead of observing the exact loss vector, the player rather observes a stochastic vector, the conditional expectation of which is (almost surely) equal to the true loss. This is a natural extension for two reasons: the losses can be inherently stochastic, or the observation or measurement of the loss can be noisy. For example, in the routing game, which models congestion in transportation and communication networks, the loss corresponds to delays on the network, which may be hard to measure exactly, and which may depend on external, stochastic variables such as weather.

Stochastic learning models have been studied in online learning theory, e.g. Bubeck and Cesa-Bianchi [34], adaptive control theory, e.g. Kumar and Varaiya [83], as well as convex optimization, e.g. Nemirovski et al. [96], [70]. Adapting ideas from these works, we propose a family of stochastic distributed learning algorithms and study their convergence.

Since convergence of the mass distributions $(x^{(\tau)})$ is equivalent to convergence of the potential values $f(x^{(\tau)})$ to the minimum $f^\star$ over Δ, we can use tools and methods from stochastic convex optimization to study learning dynamics. In particular, we will use mirror descent, a general method for first-order optimization, which we will review in this chapter, and further study in the second part of the thesis. In our model, we assume that every population k follows a stochastic mirror descent algorithm, with different learning rates $(\eta_k^{(t)})$ for different populations.

Contributions

There are two aspects of the learning model that we will particularly care about. First, we consider models in which different populations follow different dynamics, and in particular, different learning rates. We refer to such models as heterogeneous, and their analysis can be more challenging. Second, we seek to prove strong convergence of the sequence of mass distributions $(x^{(\tau)})$, rather than a weaker notion of convergence such as convergence in the

sense of Cesàro, which is also more challenging for stochastic methods, as most convergence proofs consider a sequence of averages, see for example [111, 96, 113]. In [127], Shamir and Zhang prove that for stochastic gradient descent with step size $\eta^{(t)} = 1/\sqrt{t}$, the sequence of iterates $(x^{(t)})$ converges at a rate $\mathcal{O}(\ln t/\sqrt{t})$. Using a similar technique, we extend their result to stochastic mirror descent, and show that for heterogeneous learning dynamics, with learning rates of the form $\eta_k^{(t)} = \theta_k t^{-\alpha_k}$, the sequence of iterates $x^{(\tau)}$ converges in expectation to the set of equilibria, at a $\mathcal{O}(\ln t/t^{\min(\alpha_{\min}, 1-\alpha_{\max})})$ rate (the fastest corresponding rate is $\mathcal{O}(\ln t/\sqrt{t})$).

In the homogeneous case (all populations use the same learning rates, but not necessarily the same Bregman divergences), we show that the mass distributions converge almost surely to the set of equilibria, without additional assumptions on regularity or strong convexity. In particular, convergence holds even when the equilibrium is not unique, an assumption which is usually made when proving almost sure convergence of stochastic methods, e.g. [29], and which we manage to relax. Finally, for strongly convex potential functions, we show that the distance to equilibrium converges to 0 in expectation.

5.1 Distributed Stochastic Mirror Descent (DSMD)

We start by giving a brief review of the mirror descent method, and define the stochastic mirror descent dynamics that we will study in the chapter.

Mirror descent

Mirror descent is a general method for constrained convex optimization, proposed by Nemirovsky and Yudin [98]. Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{X} \end{array}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function defined on a convex, compact set $\mathcal{X} \subset \mathbb{R}^n$, and call $f^\star$ the minimum value of f on $\mathcal{X}$.

There are many interpretations of mirror descent, discussed for example in [98, 15, 10], and many others. We discuss these interpretations in more detail in Appendix B. In this chapter, we take the following point of view: mirror descent can be interpreted, as observed by Beck and Teboulle [15], as minimizing, at each iteration t, a local approximation of the objective function around the current iterate, as follows:

$$\begin{aligned} x^{(t+1)} &= \arg\min_{x \in \mathcal{X}} \left\langle \nabla f(x^{(t)}), x \right\rangle + \frac{1}{\eta^{(t)}} D_\psi(x, x^{(t)}) \\ &= \arg\min_{x \in \mathcal{X}} f(x^{(t)}) + \left\langle \nabla f(x^{(t)}), x - x^{(t)} \right\rangle + \frac{1}{\eta^{(t)}} D_\psi(x, x^{(t)}), \end{aligned} \tag{5.1}$$

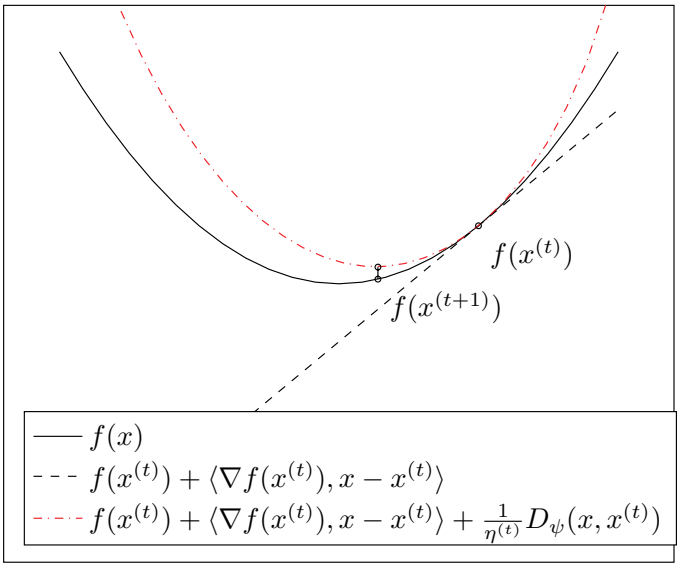


Figure 5.1: Mirror Descent iteration

where $D_\psi(x, x^{(t)})$, is the Bregman divergence associated to a strongly convex function $\psi : \mathcal{X} \to \mathbb{R}$, and defined as follows:

$$D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle .$$

Here, we assume that ψ is differentiable on $\mathcal{X}$ to simplify the discussion, but this can be relaxed, see Appendix B. By convexity of ψ, the Bregman divergence is non-negative and convex in its first argument. The function ψ is said to be μ-strongly convex w.r.t. a reference norm $\|\cdot\|$ (not necessarily the Euclidean norm) if for all $x, y \in \mathcal{X}$, $D_\psi(x, y) \geq \frac{\mu}{2}\|x - y\|^2$.

In the minimization problem (5.1), the first term, $f(x^{(t)}) + \left\langle \nabla f(x^{(t)}), x - x^{(t)} \right\rangle$ is the first-order Taylor approximation of the function around the current iterate, and the Bregman divergence term $D_\psi(x, x^{(t)})$ penalizes deviations from $x^{(t)}$. The parameter $\eta^{(t)}$ is a step size or learning rate, which determines the tradeoff between the two terms. Thus, mirror descent minimizes, at each iteration, a local approximation of the function, penalized by a Bregman divergence term. This is illustrated in Figure 5.1.

One special case of mirror descent is projected gradient descent, which can be obtained by taking $\psi(x) = \frac{1}{2}\|x\|_2^2$, in which case the Bregman divergence is the squared Euclidean distance, $D_\psi(x, y) = \frac{1}{2}\|x - y\|_2^2$, see Section B.3 in the appendix. For more examples, and a detailed discussion on the properties of Bregman divergences, see Appendix B.

Stochastic optimization

To allow for stochastic perturbation in the learning model, we consider the stochastic optimization setting, given as follows: suppose that at iteration t, we have access to a stochastic

vector $\hat{\ell}^{(t)}$, such that the conditional expectation $\ell^{(t)} = \mathbb{E}[\hat{\ell}^{(t)}|\mathcal{F}_{t-1}]$ is equal to $\nabla f(x^{(t)})$ almost surely, where $(\mathcal{F}_t)$ is the natural filtration of the stochastic process $(\hat{\ell}^{(t)})$.

This stochastic optimization framework is motivated by our study of learning dynamics, but it is also useful in solving problems in which computing the exact gradient can be prohibitively expensive, such as large-scale convex optimization, where the objective function is a sum of individual convex loss terms over a large set of samples, that is, $f(x) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \ell(x, z_i)$. A cheap estimate of the gradient of f can then be obtained by randomly drawing a small subset of samples $\mathcal{I}^{(t)} \subset \mathcal{I}$, and defining $\hat{\ell}^{(t)}$ to be $\hat{\ell}^{(t)} = \frac{1}{|\mathcal{I}^{(t)}|}\sum_{i\in\mathcal{I}^{(t)}} \nabla_x \ell(x^{(t)}, z_i)$.

The stochastic version of mirror descent is simply obtained by replacing, in the update equation (5.1), the gradient term $\nabla f(x)$ with its stochastic counterpart $\hat{\ell}^{(t)}$. Thus the algorithm generates a random sequence of iterates $x^{(t)}$, such that $x^{(t)}$ is $\mathcal{F}_{t-1}$-measurable (since at each iteration, $x^{(t)}$ is determined by $x^{(t-1)}$ and $\hat{\ell}^{(t-1)}$). We will assume that the first iterate $x^{(1)}$ is deterministic, i.e. $\mathcal{F}_0$ is trivial.

Distributed optimization on a cartesian product

We will assume that the feasible set $\mathcal{X}$ can be written as the cartesian product $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_K$. This is motivated by our problem of learning in the nonatomic convex potential game, since the feasible set is the product of simplices $\Delta = \Delta^{\mathcal{A}_1} \times \cdots \times \Delta^{\mathcal{A}_K}$. When the feasible set is a cartesian product, we can take the distance generating function ψ to be the sum

$$\psi(x) = \sum_{k=1}^{K} \psi_k(x_k), \tag{5.2}$$

in which case the Bregman divergence $D_\psi(x, y)$ is the sum of divergences $D_{\psi_k}(x_k, y_k)$, and the mirror descent update problem (5.1) decomposes

$$\begin{aligned} x^{(t+1)} &= \arg\min_{x\in\mathcal{X}} \left\langle \hat{\ell}^{(t)}, x \right\rangle + \frac{1}{\eta^{(t)}} D_\psi(x, x^{(t)}) \\ &= \arg\min_{x\in\mathcal{X}} \sum_{k=1}^{K} \left\langle \hat{\ell}_k^{(t)}, x_k \right\rangle + \frac{1}{\eta^{(t)}} D_{\psi_k}(x_k, x_k^{(t)}), \end{aligned}$$

thus $x^{(t+1)}$ can be obtained by solving K mirror updates, one on each feasible set $\mathcal{X}_k$. In order to allow more flexibility in our model, we will further assume that we can use different learning rates $\eta_k^{(t)}$ for different updates. Finally, the distributed stochastic mirror descent model is summarized in Algorithm 5. We call the algorithm homogeneous if the K updates use the same sequence of learning rates (but not necessarily the same Bregman divergence), and heterogeneous otherwise.

Assumption 3. *Throughout the chapter, we make the following assumptions:*

Algorithm 5 Distributed Stochastic Mirror Descent (DSMD) with Bregman divergences D_{ψ_k} and learning rates $(\eta_k^{(t)})$.

for $t \in \mathbb{N}$ **do**
  for each $k \in \{1, \dots, K\}$ **do**
    Observe $\hat{\ell}_k^{(t)}$ with $\ell^{(t)} := \mathbb{E}[\hat{\ell}^{(t)} | \mathcal{F}_{t-1}] \overset{a.s.}{=} \nabla f(x^{(t)})$.
    Update

$$x_k^{(t+1)} = \arg\min_{x_k \in \mathcal{X}_k} \left\langle \hat{\ell}_k^{(t)}, x_k \right\rangle + \frac{1}{\eta_k^{(t)}} D_{\psi_k}(x_k, x_k^{(t)}) \tag{5.3}$$

  end for
end for

(i) For each k, the distance generating function ψ_k is strongly convex w.r.t. a reference norm $\|\cdot\|$, and the corresponding Bregman divergence is bounded on $\mathcal{X}_k$, that is, there exists $\mu_k > 0$ and $D_k > 0$ such that for all $x, y \in \mathcal{X}_k$, $\frac{\mu_k}{2}\|x - y\|^2 \leq D_{\psi_k}(x, y) \leq D_k$,

(ii) The noisy gradient vectors are uniformly square integrable in the dual norm, that is, there exists $G > 0$ such that for all t, $\mathbb{E}\left[\|\hat{\ell}^{(t)}\|_*^2\right] \leq G^2$.

5.2 A stochastic model of learning in nonatomic potential games

In this section, we show how the distributed optimization model of Algorithm 5 applies to our problem of learning in nonatomic convex potential games.

By Definition 1, there exists a convex potential function f and scalars κ_k such that for all k and all $x \in \Delta^{\mathcal{A}_k}$, $\kappa_k \ell_k(x) = \nabla_{x_k} f(x)$. We extend the learning model of Chapter 2 to allow stochastic perturbations of the loss vectors. That is, we now suppose that at iteration t, population k observes a stochastic vector $\hat{\ell}_k^{(t)}$, which is unbiased in the sense that

$$\mathbb{E}\left[\hat{\ell}_k^{(t)} | \mathcal{F}_{t-1}\right] = \kappa_k \ell_k(x^{(t)}) = \nabla_{x_k} f(x^{(t)}).$$

Then we assume that each population updates its mass distribution $x_k^{(t)}$ by applying a mirror descent algorithm on its feasible set $\mathcal{X}_k = \Delta^{\mathcal{A}_k}$ and with learning rates $(\eta_k^{(\tau)})$. Note that we scaled the loss vector by κ_k, so that the conditional expectation of $\hat{\ell}^{(t)}$ is the gradient of the potential. The learning model is then a special case of Algorithm 5, with feasible sets $\mathcal{X}_k = \Delta^{\mathcal{A}_k}$.

Although the dynamics is motivated by the learning problem in potential games, we will analyze it in the general case, since the convergence results are of interest in the broader context of first-order, stochastic optimization.

5.3 Convergence in the sense of Cesàro

A fundamental lemma

The following lemma is an essential step in proving the convergence results of this chapter. It is a straightforward generalization of Lemma 2.1 in Nemirovski et al. [96].

Proposition 18. *Consider the DSMD algorithm with Bregman divergences D_{ψ_k} and decreasing learning rates $(\eta_k^{(t)})$ and let $(x^{(t)})$ be the resulting stochastic process. Then for all k, all for all τ, and all $\mathcal{F}_{\tau-1}$ measurable x_k,*

$$D_{\psi_k}(x_k, x_k^{(\tau+1)}) \leq D_{\psi_k}(x_k, x_k^{(\tau)}) - \eta^{(\tau)} \left\langle \hat{\ell}_k^{(\tau)}, x_k^{(\tau)} - x_k \right\rangle + \frac{(\eta^{(\tau)})^2}{2\mu_k} \|\hat{\ell}_k^{(\tau)}\|_*^2, \tag{5.4}$$

additionally, for all $t_2 > t_1 \geq 1$, and all $\mathcal{F}_{t_1-1}$-measurable x_k,

$$\sum_{\tau=t_1}^{t_2} \mathbb{E}\left[\left\langle \ell_k^{(\tau)}, x_k^{(\tau)} - x_k \right\rangle\right] \leq \frac{\mathbb{E}\left[D_{\psi_k}(x_k, x_k^{(t_1)})\right]}{\eta_k^{(t_1)}} + D_k \left(\frac{1}{\eta_k^{(t_2)}} - \frac{1}{\eta_k^{(t_1)}}\right) + \frac{G^2}{2\mu_k} \sum_{\tau=t_1}^{t_2} \eta_k^{(\tau)}. \tag{5.5}$$

This bound can be interpreted as a regret bound when the feasible set is a simplex: Taking the supremum over $x_k \in \Delta^{\mathcal{A}_k}$, $\sup_{x_k \in \Delta^{\mathcal{A}_k}} \sum_{\tau=t_1}^{t_2} \left\langle \ell_k^{(\tau)}, x_k^{(\tau)} - x_k \right\rangle$ is the cumulative regret of population k, as defined in Definition 4. Even when $\mathcal{X}_k$ is a general convex set, this quantity is also defined to be the regret in the context of online convex optimization, see for example [137, 63].

We begin by proving convergence in the sense of Cesàro, and show that if the algorithm has a sublinear regret in expectation, then $f\left(\mathbb{E}\left[\bar{x}^{(t)}\right]\right)$ converges to $f^\star$, where $\bar{x}^{(t)} = \frac{\sum_{\tau=1}^t x^{(\tau)}}{t}$. This can be guaranteed when $(\eta_k^{(t)})$ have appropriate decay rates, as in the following Corollary.

Theorem 9. *Consider the DSMD method with $\eta_k^{(t)} = \theta_k t^{-\alpha_k}$, with $\theta_k > 0$ and $\alpha_k \in (0, 1)$. Then*

$$f\left(\mathbb{E}\left[\bar{x}^{(t)}\right]\right) - f^\star \leq \sum_{k=1}^K \left(\frac{D_k}{\theta_k t^{1-\alpha_k}} + \frac{\theta_k}{1-\alpha_k} \frac{G^2}{2\mu_k} \frac{1}{t^{\alpha_k}}\right).$$

The bound is $\mathcal{O}\left(t^{-\min(\alpha_{\min}, 1-\alpha_{\max})}\right)$, where $\alpha_{\min}$ and $\alpha_{\max}$ are, respectively, the smallest and largest rate α_k.

Proof. Let $x^\star$ be a minimizer of f over $\mathcal{X}$. We have by convexity of f and the fact that $\ell^{(\tau)} \overset{a.s.}{=} \nabla f(x^{(\tau)})$,

$$\begin{aligned} f\left(\mathbb{E}\left[\bar{x}^{(t)}\right]\right) - f^\star &\leq \frac{\sum_{\tau=1}^t \mathbb{E}\left[f(x^{(\tau)}) - f^\star\right]}{t} \\ &\leq \sum_{k=1}^K \frac{\sum_{\tau=1}^t \mathbb{E}\left[\left\langle \ell_k^{(\tau)}, x_k^{(\tau)} - x_k^\star \right\rangle\right]}{t}. \end{aligned}$$

Then by Proposition 18, and since $x^\star$ is $\mathcal{F}_0$-measurable (deterministic),

$$\begin{aligned} f\left(\mathbb{E}\left[\bar{x}^{(t)}\right]\right) - f^\star &\leq \frac{\sum_{\tau=1}^{t} \mathbb{E}\left[\left\langle \ell^{(\tau)}, x^{(\tau)} - x \right\rangle\right]}{t} \\ &\leq \sum_{k=1}^{K} \frac{\mathbb{E}\left[D_{\psi_k}(x_k^\star, x_k^{(1)})\right]}{\eta_k^{(1)} t} + \frac{D_k}{t}\left(\frac{1}{\eta_k^{(t)}} - \frac{1}{\eta_k^{(1)}}\right) + \frac{G^2}{2\mu_k}\frac{\sum_{\tau=1}^{t}\eta_k^{(\tau)}}{t} \\ &\leq \sum_{k=1}^{K} \frac{D_k}{t\eta_k^{(t)}} + \frac{G^2}{2\mu_k}\frac{\sum_{\tau=1}^{t}\eta_k^{(\tau)}}{t}. \end{aligned}$$

Finally, since $u \mapsto u^{-\alpha}$ is a decreasing function over $\mathbb{R}^+$, $\sum_{\tau=1}^{t}\eta_k^{(\tau)} \leq \theta_k \int_0^t u^{-\alpha_k} du = \frac{\theta_k}{1-\alpha_k} t^{1-\alpha_k}$, which concludes the proof. □

5.4 Convergence of heterogeneous DSMD

We now analyze the convergence of $\mathbb{E}\left[f(x^{(t)})\right]$ under the heterogeneous DSMD model with learning rates $\eta_k^{(t)} = \theta_k t^{-\alpha_k}$, $\alpha_k \in (0,1)$. Shamir and Zhang [127] prove the convergence of the last iterate in the case of stochastic gradient descent (a special case of SMD) for $\alpha = \frac{1}{2}$. Our analysis uses their technique and extends it to the mirror descent method, heterogeneous learning rates, and general $\alpha_k \in (0,1)$.

Theorem 10. *Consider the DSMD method with learning rates $\eta_k^{(t)} = \theta_k t^{-\alpha_k}$. Then for all $t \geq 1$,*

$$\mathbb{E}\left[f(x^{(t)})\right] - f^\star \leq \sum_{k=1}^{K}\left(\frac{D_k}{\theta_k}\frac{1}{t^{1-\alpha_k}} + \frac{\theta_k G^2}{2\mu_k(1-\alpha_k)}\frac{1}{t^{\alpha_k}}\right)(2 + \ln t)\,. \tag{5.6}$$

This bound is a $\mathcal{O}(t^{-\min(\alpha_{\min}, 1-\alpha_{\max})}\ln t)$.

Proof. Let t be fixed. Adapting the proof of Shamir and Zhang [127], we define S_m to be

$$S_m = \frac{1}{m+1}\sum_{\tau=t-m}^{t}\mathbb{E}\left[f(x^{(\tau)})\right].$$

We have by convexity of f,

$$\begin{aligned} \sum_{\tau=t-m}^{t}\mathbb{E}\left[f(x^{(\tau)}) - f(x^{(t-m)})\right] &\leq \sum_{\tau=t-m}^{t}\mathbb{E}\left[\left\langle \ell^{(\tau)}, x^{(\tau)} - x^{(t-m)}\right\rangle\right] \\ &= \sum_{k=1}^{K}\sum_{\tau=t-m}^{t}\mathbb{E}\left[\left\langle \ell_k^{(\tau)}, x_k^{(\tau)} - x_k^{(t-m)}\right\rangle\right]. \end{aligned}$$

and applying Proposition 18 with $t_1 = t - m$, $t_2 = t$, and $x_k = x_k^{(t-m)}$, which is $\mathcal{F}_{t-m-1}$-measurable, we have

$$\sum_{\tau=t-m}^{t} \mathbb{E}\left\langle \ell_k^{(\tau)}, x_k^{(\tau)} - x_k^{(t-m)} \right\rangle \leq D_k \left(\frac{1}{\eta_k^{(t)}} - \frac{1}{\eta_k^{(t-m)}} \right) + \frac{G^2 \theta_k}{2\mu_k} \sum_{\tau=t-m}^{t} \tau^{-\alpha_k}$$
$$\leq \frac{D_k \left(t^{\alpha_k} - (t-m)^{\alpha_k}\right)}{\theta_k} + \frac{\theta_k G^2 (t^{1-\alpha_k} - (t-m-1)^{1-\alpha_k})}{2\mu_k(1-\alpha_k)},$$

where we used the integral bound $\sum_{\tau=t-m}^{t} \tau^{-\alpha_k} \leq \int_{t-m-1}^{t} u^{-\alpha_k} du$. To simplify this bound, we can use the fact that $-(t-m-1)^{-\alpha_k} \leq -t^{-\alpha_k}$ and write

$$t^{1-\alpha_k} - (t-m-1)^{1-\alpha_k} \leq \frac{t - (t-m-1)}{t^{\alpha_k}} = \frac{m+1}{t^{\alpha_k}}.$$

Similarly, $t^{\alpha_k} - (t-m)^{\alpha_k} \leq \frac{m}{t^{1-\alpha_k}}$. Therefore

$$\sum_{\tau=t-m}^{t} \mathbb{E}\left[f(x^{(\tau)}) - f(x^{(t-m)})\right] \leq \sum_{k=1}^{K} \left(\frac{D_k}{\theta_k} \frac{m+1}{t^{1-\alpha_k}} + \frac{\theta_k G^2}{2\mu_k(1-\alpha_k)} \frac{m+1}{t^{\alpha_k}} \right). \tag{5.7}$$

Dividing by $m+1$, we have

$$-\mathbb{E}\left[f(x^{(t-m)})\right] \leq -S_m + \sum_{k=1}^{K} \left(\frac{D_k}{\theta_k} \frac{1}{t^{1-\alpha_k}} + \frac{\theta_k G^2}{2\mu_k(1-\alpha_k)} \frac{1}{t^{\alpha_k}} \right).$$

Therefore

$$S_{m-1} = \frac{1}{m}\left((m+1)S_m - \mathbb{E}\left[f(x^{t-m})\right]\right)$$
$$\leq S_m + \sum_{k=1}^{K} \left(\frac{D_k}{\theta_k} \frac{1}{t^{1-\alpha_k}} + \frac{\theta G^2}{2\mu_k(1-\alpha_k)} \frac{1}{t^{\alpha_k}} \right) \frac{1}{m}. \tag{5.8}$$

We seek to derive a bound on $\mathbb{E}\left[f(x^{(t)})\right] - f^\star = S_0 - f^\star$, thus, we can sum inequality (5.8) for $m \in \{1, \ldots, t\}$, and obtain

$$S_0 - f^\star \leq S_{t-1} - f^\star + \sum_{k=1}^{K} \left(\frac{D_k}{\theta_k} \frac{1}{t^{1-\alpha_k}} + \frac{\theta_k G^2}{2\mu_k(1-\alpha_k)} \frac{1}{t^{\alpha_k}} \right) \sum_{m=1}^{t-1} \frac{1}{m}, \tag{5.9}$$

and from Theorem 9, we have

$$S_{t-1} - f^\star \leq \sum_{k=1}^{K} \left(\frac{D_k}{\theta_k t^{1-\alpha_k}} + \frac{\theta_k G^2}{2\mu_k(1-\alpha_k)} \frac{1}{t^{\alpha_k}} \right). \tag{5.10}$$

Finally, combining the inequalities (5.9) and (5.10) and using the fact that $\sum_{m=1}^{t-1} \frac{1}{m} \leq 1 + \ln t$, gives the desired bound. □

In particular, for $\psi(x) = \frac{1}{2}\|x\|_2^2$, which is strongly convex with respect to the Euclidean norm with constant $\mu = 1$, the algorithm reduces to stochastic gradient descent. Then, taking $\alpha = \frac{1}{2}$ yields the bound of Theorem 2 in [127].

5.5 Convergence of homogeneous DSMD

In this section, we study convergence properties of the DSMD model when all the updates use the same sequence of learning rates, $\eta_k^{(t)} = \eta^{(t)}$ for all k. The Bregman divergence, can however, be different for different k. Observe that by scaling the Bregman divergence D_{ψ_k} and the learning rate $\eta_k^{(t)}$ by the same constant, the mirror update (5.3) is unchanged, thus the learning rate sequences only have to be equal up to constant scaling.

Almost sure convergence

First, we show almost sure convergence to the set of minimizers of f. Let us denote the set of minimizers by $\mathcal{X}^\star := \arg\min_{x\in\mathcal{X}} f(x)$. We say that a sequence $x^{(t)}$ converges to $\mathcal{X}^\star$, and write $x^{(t)} \to \mathcal{X}^\star$, if $d(x^{(t)}, \mathcal{X}^\star) \to 0$ as $t \to \infty$ where d is the distance to the set $d(x, \mathcal{X}^\star) = \inf_{y\in\mathcal{X}} \|x - y\|$.

Theorem 11. *Consider the homogeneous DSMD method, and suppose that* $\sum_{t=1}^{\infty} \eta^{(t)} = \infty$ *and* $\sum_{t=1}^{\infty} (\eta^{(t)})^2 < \infty$. *Then*

$$x^{(t)} \xrightarrow{a.s.} \mathcal{X}^\star.$$

Note that a similar almost sure convergence result is known in the stochastic optimization literature, see for example [29]. However, such results assume uniqueness of the minimizer. We relax this uniqueness assumption by analyzing the Bregman divergence to $\mathcal{X}^\star$. In the proof, we will use the following result of convergence of almost super martingales, due to Robbins and Siegmund [116].

Theorem 12 ([116])**.** *A stochastic process* $d^{(\tau)}$ *adapted to the filtration* $\mathcal{F}_\tau$ *is an almost super martingale if there exist non-negative adapted processes* $\xi^{(\tau)}, \zeta^{(\tau)}$ *such that*

$$\mathbb{E}[d^{(\tau+1)}] \le d^{(\tau)} + \xi^{(\tau)} - \zeta^{(\tau)}.$$

If $\sum_\tau \xi^{(\tau)} < \infty$ *a.s., then almost surely,* $\lim_\tau d^{(\tau)}$ *exists, is finite, and* $\sum_\tau \zeta^{(\tau)} < \infty$.

Proof. Recall that we can define a distance generating function on the product $\mathcal{X}$, by taking $\psi(x) = \sum_{k=1}^K \psi_k(x_k)$ as in (5.2), and the corresponding Bregman divergence is the sum of Bregman divergences. Now let $D_\psi(\mathcal{X}^\star, x) = \inf_{x^\star\in\mathcal{X}^\star} D_\psi(x^\star, x)$. Since D_ψ is continuous and $\mathcal{X}^\star$ is compact (it is a closed subset of the compact set $\mathcal{X}$), we have that the infimum is attained and $D_\psi(\mathcal{X}^\star, \cdot)$ is continuous. By continuity of $D_\psi(\mathcal{X}^\star, \cdot)$ and compactness of $\mathcal{X}$, we have $x^{(t)} \to \mathcal{X}^\star$ if and only if $D_\psi(\mathcal{X}^\star, x^{(t)}) \to 0$.

We start by showing that $D_\psi(\mathcal{X}^\star, x^{(t)})$ converges almost surely, using a semi martingale convergence theorem. From inequality (5.4) in Proposition 18, summing over k and letting μ be the harmonic mean of (μ_k), we have

$$D_\psi(x, x^{(\tau+1)}) \le D_\psi(x, x^{(\tau)}) - \eta^{(\tau)} \left\langle \hat{\ell}^{(\tau)}, x^{(\tau)} - x \right\rangle + \frac{(\eta^{(\tau)})^2}{2\mu} \|\hat{\ell}^{(\tau)}\|_*^2.$$

In particular, taking x to be equal to $x^{\star(\tau)} := \arg\min_{x^\star \in \mathcal{X}^\star} D_\psi(x^\star, x^{(\tau)})$, we have

$$\begin{aligned} D_\psi(\mathcal{X}^\star, x^{(\tau+1)}) &\leq D_\psi(x^{\star(\tau)}, x^{(\tau+1)}) \\ &\leq D_\psi(x^{\star(\tau)}, x^{(\tau)}) - \eta^{(\tau)} \left\langle \hat{\ell}^{(\tau)}, x^{(\tau)} - x^{\star(\tau)} \right\rangle + \frac{(\eta^{(\tau)})^2}{2\mu} \|\hat{\ell}^{(\tau)}\|_*^2 \\ &= D_\psi(\mathcal{X}^\star, x^{(\tau)}) - \eta^{(\tau)} \left\langle \hat{\ell}^{(\tau)}, x^{(\tau)} - x^{\star(\tau)} \right\rangle + \frac{(\eta^{(\tau)})^2}{2\mu} \|\hat{\ell}^{(\tau)}\|_*^2. \end{aligned}$$

Then, we take conditional expectations with respect to $\mathcal{F}_{\tau-1}$, and observe that since $x^{(\tau)}$ and $x^{\star(\tau)}$ are $\mathcal{F}_{\tau-1}$-measurable,

$$\begin{aligned} \mathbb{E}\left[\left\langle \hat{\ell}^{(\tau)}, x^{(\tau)} - x^{\star(\tau)} \right\rangle | \mathcal{F}_{\tau-1} \right] &= \left\langle \mathbb{E}\left[\hat{\ell}^{(\tau)} | \mathcal{F}_{\tau-1} \right], x^{(\tau)} - x^{\star(\tau)} \right\rangle \\ &\overset{a.s.}{=} \left\langle \nabla f(x^{(\tau)}), x^{(\tau)} - x^{\star(\tau)} \right\rangle \\ &\geq f(x^{(\tau)}) - f^\star. \end{aligned}$$

Therefore, we have a.s.

$$\mathbb{E}\left[D_\psi(\mathcal{X}^\star, x^{(\tau+1)}) | \mathcal{F}_{\tau-1} \right] \leq D_\psi(\mathcal{X}^\star, x^{(\tau)}) - \eta^{(\tau)}(f(x^{(\tau)}) - f^\star) + \frac{(\eta^{(\tau)})^2}{2\mu} \mathbb{E}\left[\|\hat{\ell}^{(\tau)}\|_*^2 | \mathcal{F}_{\tau-1} \right].$$

By the previous inequality, and the fact that

(i) $\eta^{(\tau)}(f(x^{(\tau)}) - f^\star) \geq 0$, and

(ii) $\sum_{\tau=1}^\infty \frac{(\eta^{(\tau)})^2}{2\mu} \|\hat{\ell}^{(\tau)}\|_*^2$ is a.s. finite, since $(\eta^{(t)})$ is square summable and $\mathbb{E}[\|\hat{\ell}^{(\tau)}\|_*^2]$ is finite,

the process $\left(D_\psi(\mathcal{X}^\star, x^{(t)})\right)$ is an almost super-martingale. Therefore, by the Robbins and Siegmund [116] convergence theorem, $D_\psi(\mathcal{X}^\star, x^{(\tau)})$ converges a.s., and the sum

$$\sum_{\tau=1}^\infty \eta^{(\tau)}(f(x^{(\tau)}) - f^\star) < \infty \text{ a.s.}$$

To show that the limit of $D_\psi(\mathcal{X}^\star, x^{(t)})$ is a.s. 0, suppose that for some realization, $D_\psi(\mathcal{X}^\star, x^{(t)})$ converges to $d > 0$, then there exists $T > 0$ such that for all $t \geq T$, $D_\psi(\mathcal{X}^\star, x^{(t)}) > d/2$. Let $\delta \triangleq \inf_{\{x \in \mathcal{X} : D_\psi(\mathcal{X}^\star, x) > \frac{d}{2}\}} f(x) - f^\star$. By continuity of f, we have that $\delta > 0$, thus

$$\sum_{\tau=1}^\infty \eta^{(\tau)}(f(x^{(\tau)}) - f^\star) \geq \delta \sum_{t \geq T} \eta^{(\tau)} = \infty.$$

since $(\eta^{(\tau)})$ is assumed summable. Therefore the event $\lim_{t\to\infty} D_\psi(\mathcal{X}^\star, x^{(t)}) > 0$ is a subset of the event $\sum_\tau \eta^{(\tau)}(f(x^{(\tau)}) - f^\star) = \infty$, which proves that $D_\psi(\mathcal{X}^\star, x^{(\tau)}) \overset{a.s.}{\to} 0$. □

Strongly convex case

In this section, we assume that f is μ_f-strongly convex with respect to D_ψ, in the following sense: for all $x, y \in \mathcal{X}$,

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \mu_f \max(D_\psi(x, y), D_\psi(y, x)).$$

We show that under this assumption, the variance of the iterates converges to 0. First, we observe that by strong convexity of ψ, we have $\mathbb{E}\left[\|x - x^\star\|^2\right] \leq \frac{2}{\mu}\mathbb{E}\left[D_\psi(x^\star, x)\right]$, thus it suffices to show the convergence of $\mathbb{E}\left[D_\psi(x^\star, x^{(t)})\right]$. First, we show the following Lemma.

Lemma 8. *Suppose f is μ_f-strongly convex with respect to D_ψ, and let $x^\star$ be the minimizer of f over $\mathcal{X}$. Then for all $y \in \mathcal{X}$, $\langle \nabla f(y), y - x^\star \rangle \geq 2\mu_f D_\psi(x^\star, y)$.*

Proof. By strong convexity of f, we have

$$\begin{aligned} f(x^\star) &\geq f(y) + \langle \nabla f(y), x^\star - y \rangle + \mu_f D_\psi(x^\star, y) \\ f(y) &\geq f(x^\star) + \mu_f D_\psi(x^\star, y), \end{aligned}$$

and we conclude by summing the two inequalities. □

Proposition 19. *Suppose that f is μ_f-strongly convex with respect to D_ψ, where ψ is defined as the sum of ψ_k, as in Equation (5.2). Then the homogeneous DSMD algorithm with homogeneous learning rates $(\eta^{(t)})$ guarantees*

$$\mathbb{E}\left[D_\psi(x^\star, x^{(t+1)})\right] \leq (1 - 2\mu_f \eta^{(t)})\,\mathbb{E}\left[D_\psi(x^\star, x^{(t)})\right] + \frac{G^2}{2\mu}(\eta^{(t)})^2.$$

Proof. We start from inequality (5.4) in Proposition 18. Taking expectation with $x_k = x_k^\star$, and summing over k, it follows that

$$\mathbb{E}[D_\psi(x^\star, x^{(t+1)})] \leq \mathbb{E}[D_\psi(x^\star, x^{(t)})] - \eta^{(t)}\mathbb{E}[\langle \hat{\ell}^{(t)}, x^{(t)} - x^* \rangle] + \frac{\mathbb{E}\,\|\hat{\ell}^{(t)}\|_*^2}{2\mu}(\eta^{(t)})^2,$$

and since $\mathbb{E}\left[\hat{\ell}^{(t)} | \mathcal{F}_{t-1}\right] = \nabla f(x^{(t)})$ a.s., we have by Lemma 8

$$-\mathbb{E}\left[\langle \hat{\ell}^{(t)}, x^{(t)} - x^\star \rangle\right] \leq -2\mu_f\,\mathbb{E}\left[D_\psi(x^\star, x^{(t)})\right].$$

Combining the two inequalities, we have the claim. □

Theorem 13 (Convergence of variance for $\eta^{(t)} = \Theta(t^{-\alpha})$)**.** *Suppose that f is μ_f strongly convex with respect to D_ψ, and consider the homogeneous DSMD with learning rates $\eta^{(t)} = \frac{\theta}{2\mu_f t^\alpha}$, $\alpha \in (0, 1)$. Then for all $t \geq t_0$*

$$\mathbb{E}\left[D_\psi(x^\star, x^{(t)})\right] \leq \frac{C}{t^\alpha}, \tag{5.11}$$

where $t_0 = \left\lceil \left(\frac{2\alpha}{\theta}\right)^{\frac{1}{1-\alpha}} \right\rceil$ and $C = \max(D t_0^\alpha, \frac{G^2\theta}{4\mu\mu_f^2})$.

Proof. We show the claim by induction on $t \geq t_0$. For $t = t_0$, we have by assumption on D_ψ

$$\mathbb{E}\left[D_\psi(x^\star, x^{(t_0)})\right] \leq D \leq \frac{C}{t_0^\alpha}.$$

Now suppose by induction that $\mathbb{E}\left[D_\psi(x^\star, x^{(t)})\right] \leq \frac{C}{t^\alpha}$. Then by Proposition 19,

$$\begin{aligned}
\mathbb{E}\left[D_\psi(x^\star, x^{(t+1)})\right] &\leq \left(1 - \frac{\theta}{t^\alpha}\right) \mathbb{E}\left[D_\psi(x^\star, x^{(t)})\right] + \frac{G^2\theta^2}{8\mu\mu_f^2} \frac{1}{t^{2\alpha}} \\
&\leq \left(1 - \frac{\theta}{t^\alpha}\right) \frac{C}{t^\alpha} + \frac{G^2\theta^2}{8\mu\mu_f^2} \frac{1}{t^{2\alpha}} \\
&= \frac{C}{(t+1)^\alpha} \left[\left(\frac{t+1}{t}\right)^\alpha \left(1 + \frac{1}{t^\alpha}\left(-\theta + \frac{G^2\theta^2}{8\mu\mu_f^2 C}\right)\right)\right] \\
&\leq \frac{C}{(t+1)^\alpha} \exp\left[\frac{\alpha}{t} + \frac{1}{t^\alpha}\left(\frac{G^2\theta^2}{8\mu\mu_f^2 C} - \theta\right)\right].
\end{aligned}$$

To conclude, it suffices to prove that the exponential term is less than one. By definition of C, $\frac{G^2\theta^2}{8\mu\mu_f^2 C} - \theta \leq -\frac{\theta}{2}$, thus the exponential term is less than one if $\frac{\alpha}{t} - \frac{\theta}{2t^\alpha} \leq 0$, i.e. $t \geq \left(\frac{2\alpha}{\theta}\right)^{\frac{1}{1-\alpha}}$, which is true if $t \geq t_0$. Therefore we have

$$\mathbb{E}\left[D_\psi(x^\star, x^{(t+1)})\right] \leq \frac{C}{(t+1)^\alpha},$$

which completes the induction. □

We observe that when $\alpha = 1$, the inequality $\frac{1}{t} - \frac{\theta}{2t} \leq 0$ holds whenever $\theta \geq 2$, in which case $t_0 = 1$, and we recover the $\mathcal{O}(\frac{1}{t})$ bound of Shamir and Zhang [127] for the Euclidean case with $\eta^{(t)} = \frac{1}{\mu_f t}$.

In fact, we can show that $\mathbb{E}\left[D_\psi(x^\star, x^{(t)})\right]$ converges to 0 for any sequence of learning rates such that $\eta^{(t)} \to 0$ and $\sum_t \eta^{(t)} = \infty$.

Lemma 9. *Let $(d^{(t)})$ be a sequence of non-negative numbers that satisfy the following inequality*

$$d^{(t+1)} \leq (1 - \nu_t) d^{(t)} + \Gamma \nu_t^2,$$

for some $\Gamma > 0$ and a positive decreasing sequence ν_t with $\sum_t \nu_t = \infty$. Then for all T with $\nu_T \leq 1$, and all $t > T$,

$$d^{(t)} \leq \nu_T \Gamma + d^{(T)} e^{-\sum_{\tau=T}^{t-1} \nu_\tau}.$$

Proof. Let T be fixed in $\mathbb{N}$, and such that $\nu_T \leq 1$. Then

$$\begin{aligned} d^{(t+1)} - \Gamma\nu_T &\leq (1-\nu_t)d^{(t)} + \Gamma\nu_t^2 - \Gamma\nu_T \\ &\leq (1-\nu_t)d^{(t)} + \Gamma\nu_T\nu_t - \Gamma\nu_T \qquad \text{since } \nu_t \leq \nu_T \\ &= (1-\nu_t)(d^{(t)} - \Gamma\nu_T) \end{aligned}$$

And since $(1-\nu_\tau) \geq 0$ for all $t \geq T$, we have by induction on $t > T$

$$d^{(t)} - \Gamma\nu_T \leq \Pi_{\tau=T}^{t-1}(1-\nu_\tau)(d^{(T)} - \nu_T\Gamma)$$

And we conclude by bounding the product $\Pi_{\tau=T}^{t-1}(1-\nu_\tau) \leq \Pi_{\tau=T}^{t-1}e^{-\nu_\tau} = e^{-\sum_{\tau=T}^{t-1}\nu_\tau}$ □

Combining Proposition 19 and Lemma 9, we can take $\nu_t = \mu_f\eta^{(t)}$ and $\Gamma = \frac{G^2}{2\mu\mu_f^2}$ to obtain

$$\mathbb{E}\left[D_\psi(x^\star, x^{(t)})\right] \leq \frac{G^2}{2\mu\mu_f}\eta^{(T)} + De^{-\sum_{\tau=T}^{t-1}\mu_f\eta^{(\tau)}},$$

for any $t > T$ such that $\mu_f\eta^{(T)} \leq 1$. In particular, this proves that $\mathbb{E}\left[D_\psi(x^\star, x^{(t)})\right] \to 0$.

5.6 Numerical examples

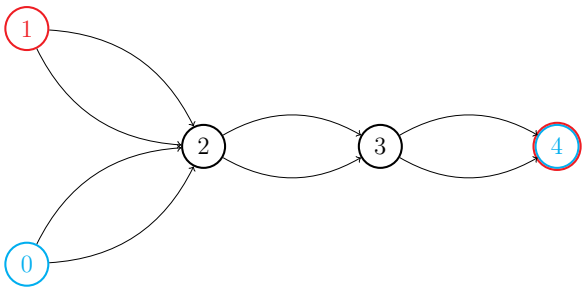


Figure 5.2: Example routing game network, with a weakly convex Rosenthal potential.

To illustrate the convergence results of this chapter, we consider a routing game example, as defined in Section 2.3, on the network given in Figure 5.2. The game involves two populations of players, with origin nodes v_0 and v_1, respectively, and a common destination node v_4. The resulting Rosenthal potential function f (as defined in (2.7)) is not strongly convex. To simulate the stochastic learning model, we add, to each path, a centered Gaussian noise with standard deviation σ, which results in stochastic loss vectors with a bounded second moment. For the population dynamics, we implement the DSMD given by Algorithm 5, with the smoothed KL divergence defined in Section B.8 in the appendix, and given by

$$D_{\mathrm{KL},\epsilon}(x,y) = \sum_a (x_a + \epsilon)\ln\frac{x_a+\epsilon}{y_a+\epsilon}.$$

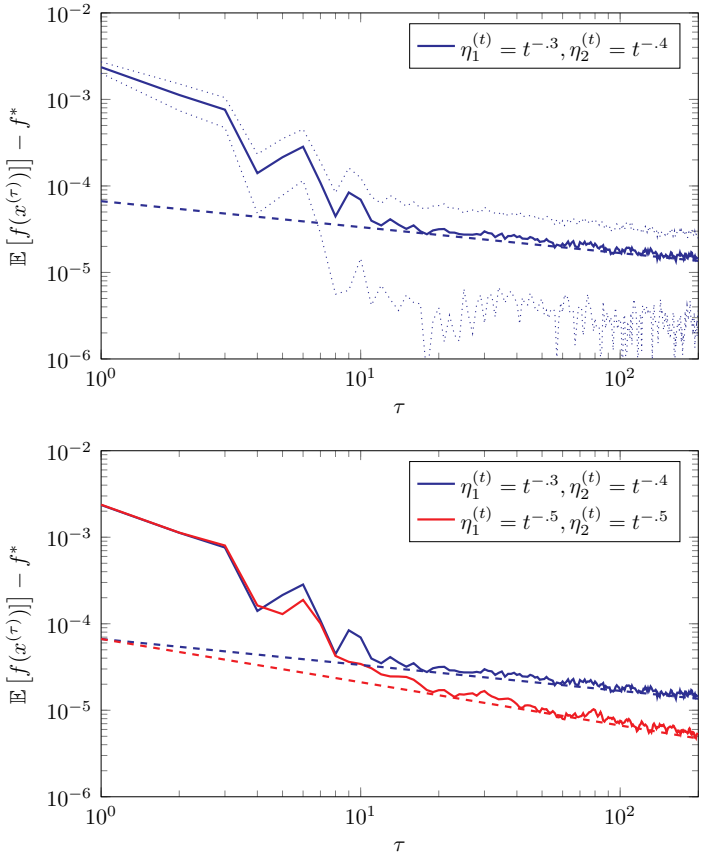


Figure 5.3: Potential values $f(x^{(\tau)}) - f^\star$, averaged across 100 simulations (with a 1 standard deviation in dotted lines), for different choices of learning rate sequences. The dashed lines show the $\mathcal{O}(t^{-\min_k \min(\alpha_k, 1-\alpha_k)} \ln t)$ rate predicted by Theorem 10.

for a positive parameter $\epsilon > 0$. When $\epsilon = 0$, this reduces to the KL divergence, and the mirror descent method reduces to the Hedge algorithm, as discussed in Section B.6. However, the KL divergence is not bounded on the simplex ($D_{\mathrm{KL}}(x, y)$ diverges when y_a vanishes for a in the support of x), which violates condition (i) in Assumption 3. Taking $\epsilon > 0$ ensures that the Bregman divergence remains bounded on the simplex by Proposition 25. And although the mirror descent update (5.3) does not have a closed form solution, it can be computed efficiently using the algorithms developed in Appendix C. If the action set has size $|\mathcal{A}| = n$, then the solution can be computed in $\mathcal{O}(n \ln n)$ time using a deterministic sorting method, given in Algorithm 15, and in expected linear time using a randomized sorting method given in Algorithm 16.

The results of the simulations are given in Figure 5.3, in which we show, in log-log scale, the potential values averaged over 100 realizations, for two different choices of (heterogeneous) learning rates. The empirical convergence rates observed in simulation are consistent with those predicted by Theorem 10.

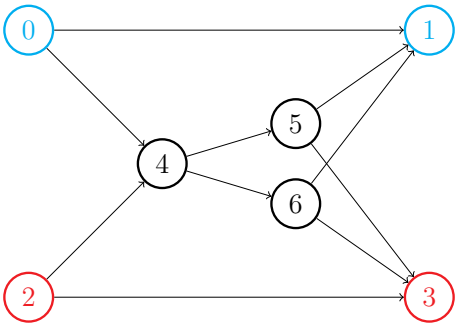


Figure 5.4: Example routing game network, with a strongly convex Rosenthal potential.

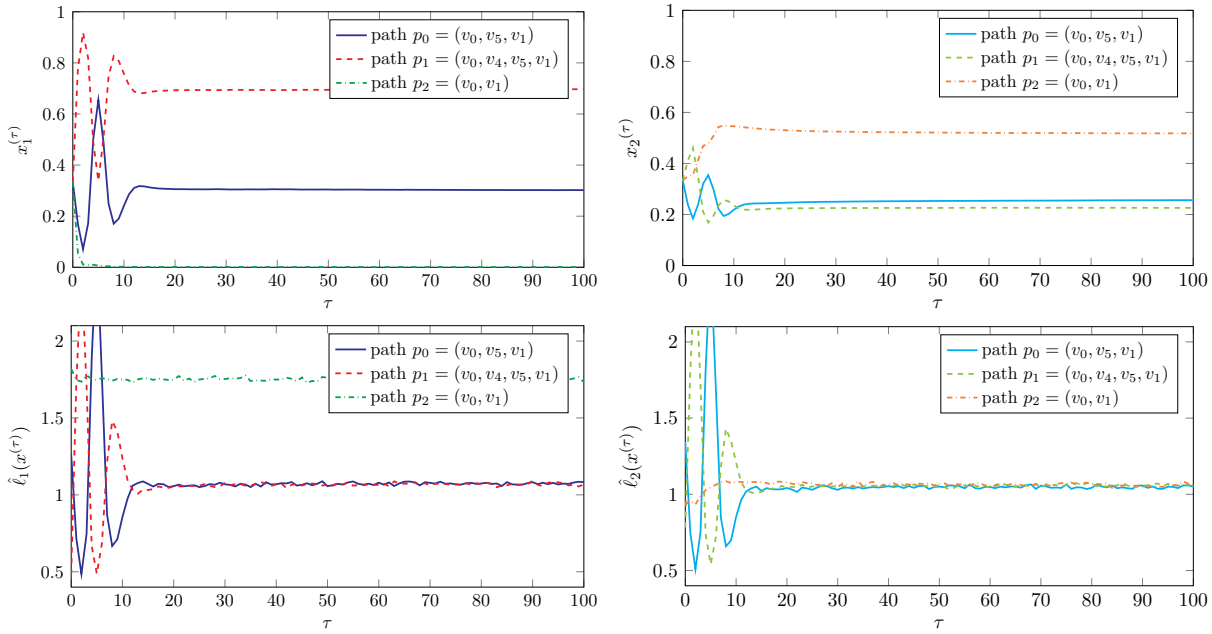


Figure 5.5: Sequences of mass distributions $x_k^{(\tau)}$ (top) and noisy loss functions $\hat{\ell}_k^{(\tau)}$, averaged across 100 simulations.

In addition to the convergence of $\mathbb{E}\left[f(x^{(t)})\right]$, Theorem 13 provides a bound on $\mathbb{E}\left[D_\psi(x^\star, x^{(t)})\right]$ if the potential f is strongly convex and the learning rates are homogeneous. To illustrate this result, we simulate the stochastic routing game on a second network, given in Figure 5.4, for which the Rosenthal potential is strongly convex. We show in Figure 5.5 the sequence of mass distributions $(x^{(\tau)})$ and noisy losses $\hat{\ell}^{(\tau)}$, averaged across 100 simulations. In expectation $x^{(\tau)}$ converges to a Nash equilibrium, such that for each population, all paths with positive mass have the same loss (see Definition 2). Finally, we show the sequence of Bregman divergence $D_\psi(x^\star, x^{(t)})$ in Figure 5.6. The empirical convergence rate is consistent with the $\mathcal{O}(1/t)$ bound predicted by Theorem 13.

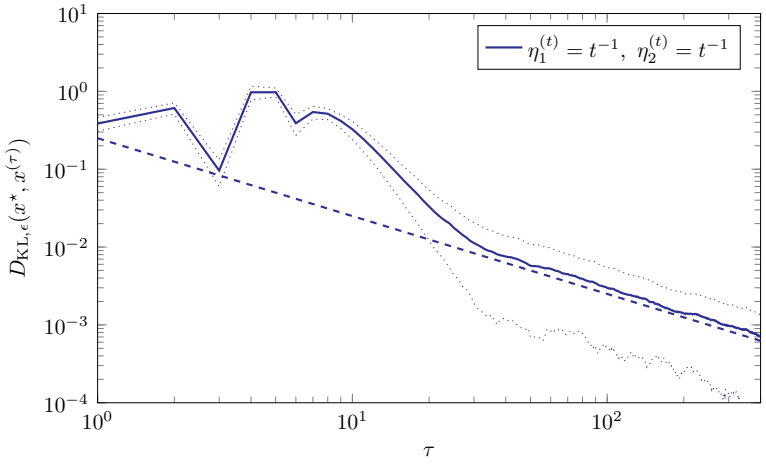


Figure 5.6: Bregman divergence to equilibrium , averaged across 100 simulations. The dashed line shows the $\mathcal{O}(t^{-1})$ convergence rate predicted by Theorem 13.

Conclusion

The stochastic mirror descent method provides a broad family of methods for convex optimization and online learning in convex potential games. We showed that we can provide convergence guarantees on the sequence of iterates $(x^{(\tau)})$, even in the heterogeneous model in which different learning rates are used in the different subproblems in the mirror update. This provides a powerful and flexible model of distributed learning, which is not only useful for solving distributed learning problems, but can also be used as a tool to model, and perhaps alter the decision dynamics of players who face an online learning problem, as we will see in the final two chapters of this first part.

Chapter 6

Estimation of Learning Dynamics: On Learning How Players Learn

The mirror descent dynamics developed in Chapter 5 provides a family of distributed algorithms for solving convex optimization problems, as well as coupled online learning problems in nonatomic convex potential games, as defined in Chapter 2. They provide convergence guarantees even in the heterogeneous case in which different players use different sequences of learning rates ($\eta_k^{(t)}$), as long as these sequences have appropriate decay rates.

Besides prescribing dynamics for learning, the mirror descent method can be used as a *model to describe the behavior* of a decision maker who faces an online learning problem. Many cyber-physical systems have human decision makers who face online learning problems, such as in transportation networks (drivers who make decisions on which path to take to drive from their origin to their destination) and communication networks (routers who make decisions on which path to route packets). In such systems, a central coordinator can be responsible for actuating the system, for example by setting tolls or incentives to reduce congestion [106], by allocating capacity on the network [74], or by routing a fraction of the flow in order to reduce the total delay [120]. Thus, it is necessary for the system coordinator to have a model of the player decisions. Such problems are usually solved by assuming that the selfish players respond to the action of the coordinator by playing a Nash equilibrium, and there is an extensive literature on mechanism design (see, e.g. [104] and the references therein) and Stackelberg games (see [12] and the references therein), that studies such problems, which in many cases are hard to solve (e.g., the Stackleberg routing problem [120] is proved to be NP-hard).

In this chapter, we propose a different approach: We consider an online learning setting, in which each player (or population) faces a sequential decision problem, and is assumed to follow an online learning algorithm with unknown parameters. By observing the sequence of decisions, the central coordinator can estimate these parameters to fit the model of decision dynamics to the observations. More precisely, we will assume that each population follows the distributed mirror descent dynamics of Algorithm 5, with a known Bregman divergence (the KL divergence), but with unknown learning rates. We pose a simple problem of estimating

the learning rates from observations from the model, and show that the estimation problem is convex in the case of the KL divergence.

In order to demonstrate our approach, and to evaluate whether the mirror descent dynamics are descriptive of actual decision dynamics, we developed a web application that allows players to participate in a distributed, online routing game, as defined in Section 2.3. When players log in, they are assigned an origin and destination on a shared network. They can choose, at each iteration, a distribution over their available routes, and each player seeks to minimize her own cost. We collect a data set using this platform, then apply the proposed method to estimate the learning rates of each player. We observe in particular that after an exploration phase, the joint decision of the players remains within a small distance of the set of Nash equilibria. We also use the estimated model parameters to predict future mass distributions, and compare our predictions to the actual distributions, showing that the online learning model can be used as a predictive model over short horizons.

6.1 Learning rate estimation in Hedge dynamics

Consider the distributed mirror descent model proposed in Algorithm 5, in which K populations of players face an online learning problem. At iteration t, population k has mass distribution $x_k^{(t)} \in \Delta^{\mathcal{A}_k}$, observes a loss vector $\ell_k^{(t)}$, and updates its mass distribution by following the mirror descent update (5.3). We assume that we can observe the sequence of decisions $(x_k^{(t)})$ and the sequence of loss functions $(\ell_k^{(t)})$. These quantities are effectively measured in our experimental setting using the routing game web application, and can be measured on transportation networks using many existing traffic monitoring and forecasting systems, such as the Mobile Millennium system [13] or the Grenoble Traffic Lab [37].

Given the current mass distribution $x_k^{(t)}$ and the current loss vector $\ell_k^{(t)}$, the mirror descent model prescribes that the next distribution $x_k^{(t+1)}$ is given by the update equation (5.3). We assume that the Bregman divergence D_{ψ_k} is given by the KL divergence, but that the sequence of learning rates $(\eta_k^{(t)})$ is an unknown, positive decreasing sequence. The learning model under Hedge dynamics is summarized in Algorithm 6 below.

Therefore, we can define a predictive model $\hat{x}_k^{(t+1)}(\eta)$, parameterized by a non-negative learning rate η, defined as follows:

$$\hat{x}_k^{(t+1)}(\eta) := \operatorname*{arg\,min}_{x_k \in \Delta^{\mathcal{A}_k}} \left\langle \ell_k^{(t)}, x_k \right\rangle + \frac{1}{\eta} D_{\mathrm{KL}}(x_k, x_k^{(t)}). \tag{6.1}$$

The solution of the mirror descent update (6.1) is given by the Hedge update rule (see Section B.6 in the appendix),

$$\hat{x}_{k,a}^{(t+1)}(\eta) = \frac{x_{k,a}^{(t)} e^{-\eta \ell_{k,a}^{(t)}}}{Z_k^{(t)}(\eta)} \tag{6.2}$$

where $Z_k^{(t)}(\eta)$ is the normalization constant $Z_k^{(t)}(\eta) = \sum_{a'} x_{k,a'}^{(t)} e^{-\eta \ell_{k,a'}^{(t)}}$.

Algorithm 6 Distributed Hedge algorithm with learning rates $(\eta_k^{(t)})$.

1: **for** $t \in \mathbb{N}$ **do**
2: **for** each $k \in \{1, \dots, K\}$ **do**
3: Observe $\ell_k^{(t)}$
4: Update

$$x_k^{(t+1)} = \arg\min_{x_k \in \Delta^{\mathcal{A}_k}} \left\langle \ell_k^{(t)}, x_k \right\rangle + \frac{1}{\eta_k^{(t)}} D_{\mathrm{KL}}(x_k, x_k^{(t)}) = \left(\frac{x_{k,a}^{(t)} e^{-\eta_k^{(t)} \ell_{k,a}^{(t)}}}{\sum_{a'} x_{k,a'}^{(t)} e^{-\eta_k^{(t)} \ell_{k,a'}^{(t)}}} \right)_{a \in \mathcal{A}_k}$$

5: **end for**
6: **end for**

Given the next mass distribution $x_k^{(t+1)}$, we can measure the discrepancy between the model prediction and the observation using the KL divergence between $x_k^{(t+1)}$ and $\hat{x}^{(t+1)}(\eta)$. Thus, let

$$d_k^{(t)}(\eta) := D_{\mathrm{KL}}(x_k^{(t+1)}, \hat{x}_k^{(t+1)}(\eta)). \tag{6.3}$$

Estimating a single term of the learning rates sequence

Fix t, and suppose that we are given the tuple of observations $(x_k^{(t)}, \ell_k^{(t)}, x_k^{(t+1)})$. We define the estimate of the learning rate $\hat{\eta}_k^{(t)}$ to be the minimizer of the KL divergence,

$$\hat{\eta}_k^{(t)} = \arg\min_{\eta \geq 0} d_k^{(t)}(\eta). \tag{6.4}$$

Note that we impose the constraint that $\eta \geq 0$. This is an assumption of the model, and in our experiments, this turns out to be an important constraint, as we will see that $d_k^{(t)}(\eta)$ can, in some rare cases, be minimal for negative values of η if the problem were solved without the non-negativity constraint. This corresponds to rare instances in which the players exhibit irrational behavior, by shifting probability mass to actions with higher losses, and will be further discussed in Section 6.3. In the next theorem, we show that the minimization problem (6.4) is convex.

Theorem 14. *$d_k^{(t)}(\eta) := D_{KL}(x_k^{(t+1)}, \hat{x}_k^{(t+1)}(\eta))$ is a convex function of η, and its gradient with respect to η is given by*

$$\frac{d}{d\eta} d_k^{(t)}(\eta) = \left\langle \ell_k^{(t)}, x_k^{(t+1)} - \hat{x}_k^{(t+1)}(\eta) \right\rangle.$$

Proof. Given the expression (6.2) of $\hat{x}_k^{(t+1)}(\eta)$, we can explicitly compute

$$\begin{aligned} d_k(\eta) &= D_{\mathrm{KL}}(x_k^{(t+1)}, \hat{x}_k^{(t+1)}(\eta)) \\ &= \sum_{a \in \mathcal{A}_k} x_{k,a}^{(t+1)} \ln \frac{x_{k,a}^{(t+1)}}{\hat{x}_{k,a}^{(t+1)}(\eta)} \\ &= \sum_{a \in \mathcal{A}_k} x_{k,a}^{(t+1)} \left(\ln \frac{x_{k,a}^{(t+1)}}{x_{k,a}^{(t)}} + \eta \ell_{k,a}^{(t)} + \ln Z_k^{(t)}(\eta) \right) \\ &= D_{\mathrm{KL}}(x_k^{(t+1)}, x_k^{(t)}) + \eta \left\langle \ell_k^{(t)}, x_k^{(t+1)} \right\rangle + \ln Z_k^{(t)}(\eta), \end{aligned} \tag{6.5}$$

where we used the explicit form (6.2) of $\hat{x}_k^{(t+1)}(\eta)$ in the third equality, and the fact that $\sum_a x_{k,a}^{(t+1)} = 1$ in the last equality. In this expression, the first term does not depend on η, the second term is linear in η, and the last term is the function $\eta \mapsto \ln Z_k^{(t)}(\eta) = \ln \sum_a x_{k,a}^{(t)} e^{-\eta \ell_{k,a}^{(t)}}$, which is known to be convex in η (see for example Section 3.1.5 in [30]). Therefore $d_k^{(t)}(\eta)$ is convex, and its gradient can be obtained by differentiating each term

$$\begin{aligned} \frac{d}{d\eta} d_k^{(t)}(\eta) &= \left\langle \ell_k^{(t)}, x_k^{(t+1)} \right\rangle + \frac{\frac{d}{d\eta} Z_k^{(t)}(\eta)}{Z_k^{(t)}(\eta)} \\ &= \left\langle \ell_k^{(t)}, x_k^{(t+1)} \right\rangle + \frac{\sum_a -\ell_{k,a}^{(t)} x_{k,a}^{(t)} e^{-\eta \ell_{k,a}^{(t)}}}{Z_k^{(t)}(\eta)} \\ &= \left\langle \ell_k^{(t)}, x_k^{(t+1)} \right\rangle - \left\langle \ell_k^{(t)}, \hat{x}_k^{(t+1)}(\eta) \right\rangle, \end{aligned}$$

which proves the claim. □

On the support of the distributions

According to the entropy update and its explicit solution (6.6), the support of $\hat{x}_k^{(t+1)}(\eta)$ always coincides with the support of $x_k^{(t)}$ (due to the multiplicative form of the Hedge solution). As a consequence, if we observe a tuple $(x_k^{(t)}, \ell_k^{(t)}, x_k^{(t+1)})$ such that some a is in the support of $x_k^{(t+1)}$ but not in the support of $x_k^{(t)}$, the KL divergence $D_{\mathrm{KL}}(x_k^{(t+1)}, \hat{x}_k^{(t+1)}(\eta))$ is infinite for all η, since $\mathrm{support}(x^{(t+1)}) \not\subset \mathrm{support}(\hat{x}_k^{(t+1)}(\eta))$. This is problematic, as the estimation problem is ill-posed in such cases (which did occur in the routing game experiment). However, observe that from Equation (6.5), the KL divergence can be decomposed into two terms:

$$d_k^{(t)}(\eta) = D_{\mathrm{KL}}(x_k^{(t+1)}, x_k^{(t)}) + \eta \left\langle \ell_k^{(t)}, x_k^{(t+1)} \right\rangle + \ln Z_k^{(t)}(\eta),$$

where the first term, $D_{\mathrm{KL}}(x_k^{(t+1)}, x_k^{(t)})$ may be infinite (if $\mathrm{support}(x^{(t+1)}) \not\subset \mathrm{support}(x_k^{(t)})$), but does not depend on η, while the second term, $\eta \left\langle \ell_k^{(t)}, x_k^{(t+1)} \right\rangle + \ln Z_k^{(t)}(\eta)$ is finite for all

values of $\eta \geq 0$, regardless of the supports of the observations. Thus, instead of minimizing $d_k^{(t)}(\eta)$, we can minimize

$$\tilde{d}_k^{(t)}(\eta) := \eta \left\langle \ell_k^{(t)}, x_k^{(t+1)} \right\rangle + \ln Z_k^{(t)}(\eta),$$

and the problem becomes well-posed regardless of the supports.

Estimating the decay rate of the learning rate sequence

In the previous section, we proposed a method to estimate a single term of the learning rate sequence. One can of course repeat this procedure at every iteration, thus generating a sequence of estimated learning rates. However, the resulting sequence may not be decreasing. In order to be consistent with the assumptions of the mirror descent model, we can assume a parameterized sequence of learning rates, $\eta_k^{(t)} = \theta_k t^{-\alpha_k}$, with parameters $\theta_k > 0$ and $\alpha_k \in (0, 1)$. This polynomial decay rate is motivated by the convergence guarantees provided in Chapter 5, in Theorem 13 and Theorem 10.

Given the observations $(x_k^{(t)})$ and $(\ell_k^{(t)})$, we can define the cumulative divergence,

$$D_k^{(t)}(\alpha_k, \theta_k) := \sum_{\tau=1}^{t} d_k^{(\tau)}(\theta_k \tau^{-\alpha_k}),$$

where each term of the sum is as defined in Equation (6.3), then estimate (α_k, θ_k) by solving the problem

$$(\alpha_k, \theta_k) = \underset{\alpha_k \in (0,1), \theta_k \geq 0}{\arg\min} \; D_k^{(t)}(\alpha_k, \theta_k). \tag{6.6}$$

This problem is non-convex in general, however, since it is low-dimensional (two parameters to estimate), it can be solved approximately using non-convex optimization techniques.

6.2 The routing game web application

We have implemented a web application based on the routing game, using the Python Django Framework. The code is available on Github: `www.github.com/walidk/routing`. The application has been deployed on the Heroku service at the url: `routing-game.herokuapp.com`. In this section, we will describe the architecture of the web application.

Web Application Architecture

The web application implements the repeated routing game described in Section 2.3. In our implementation, each player represents a population k, and chooses a mass distribution $x_k^{(t)} \in \Delta^{\mathcal{A}_k}$ at each iteration.

The general architecture of the system is summarized in Figure 6.1. It consists of two different client interfaces, that are used respectively by the administrator of the game and

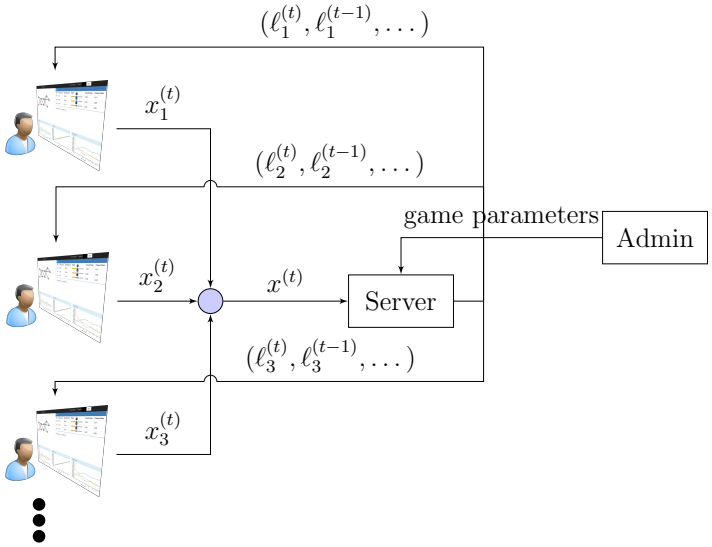


Figure 6.1: General architecture of the system. The administrator sets up the game. During iteration t, the clients input the current values of the distributions $x_k^{(t)}$ and send them to the server. At the end of the iteration, the server uses these values to compute the loss functions $\ell_k^{(t)}$ and sends them back to the clients.

the players, shown in Figures 6.2 and 6.3, and a backend server that is responsible for collecting inputs from the clients, updating the state of the game, then broadcasting current information to each player.

Admin Interface

The administrator can set up the game using the admin interface shown in Figure 6.2 by:

1. Creating a graph and defining the cost functions on each edge.
2. Creating player models. A player model is defined by its origin, destination and total mass. When a player connects to the game, she is randomly assigned to one of the player models (note that multiple players can have the same player model).
3. Setting additional parameters of the game, such as the total number of iterations and the duration of each iteration.

Once the game is set up, players can log in to the client interface. During the game, the administrator can monitor, for each player, her expected cost and the learning rate estimates, computed as described in Section 6.3.

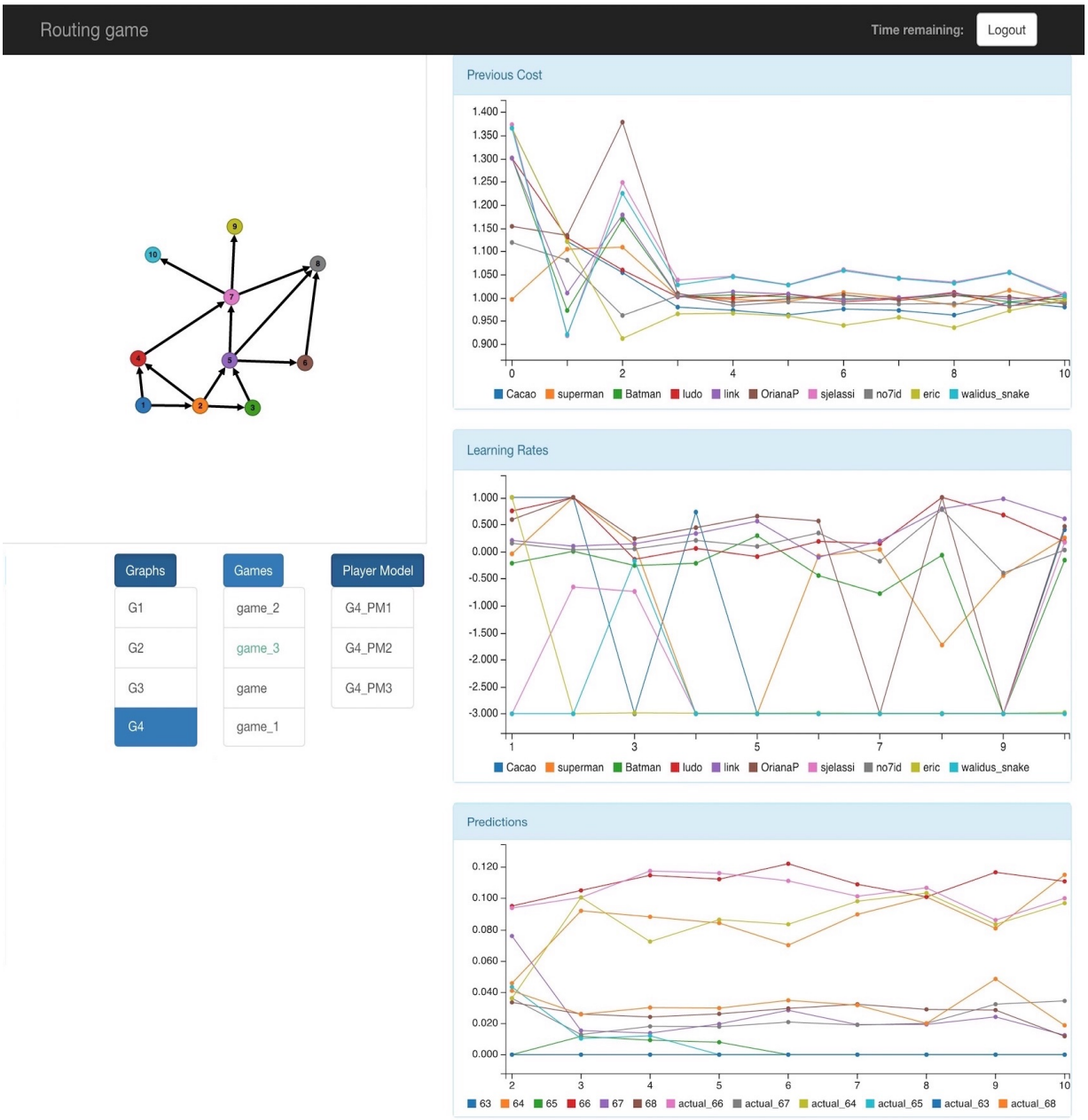


Figure 6.2: Admin interface

Player Interface

Figure 6.3 shows a screenshot of the client interface for the players. The table is the main element of the graphical user interface, and can be used by the player to set weights on the different paths, using the sliders. The weights determine the mass distribution $x_k^{(t)}$ for the current iteration. The table also shows the previous mass distribution ($x_k^{(t-1)}$), and the

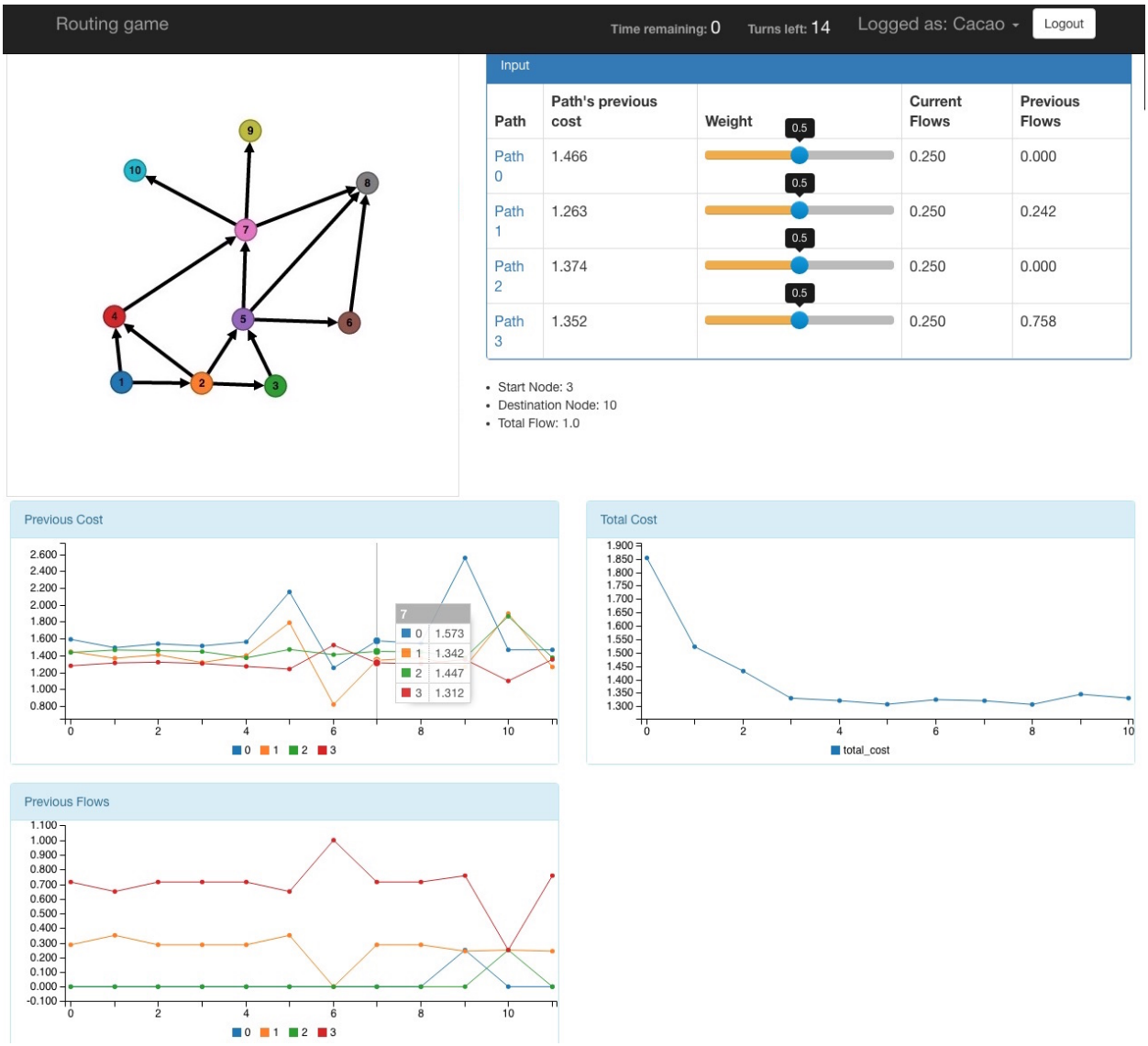


Figure 6.3: User interface

previous costs ($\ell_k^{(t-1)}$). Clicking a path on the table will also highlight that path on the graph. The bottom charts show the full history of distributions $x_k^{(\tau)}$, costs $\ell_k^{(\tau)}$, and expected costs given by the inner product $\left\langle x_k^{(\tau)}, \ell_k^{(\tau)} \right\rangle$, for $\tau \leq t$. The top navigation bar shows the time left until the end of the current iteration, and the number of iterations left until the end of the game.

At the end of the iteration, the server uses the values of $x_k^{(t)}$ for all players $k \in \{1, \ldots, K\}$ to compute the costs $\ell_k(x^{(t)})$, then sends this information to the clients, which then update the charts and the table with the last value of the cost.

6.3 Experimental results

To illustrate the methods proposed in this chapter, we ran the experiment on the example network (shown in Figure 6.4), with 10 anonymous players. The game is played over a horizon of 25 iterations. The edge cost functions are taken to be linear increasing.

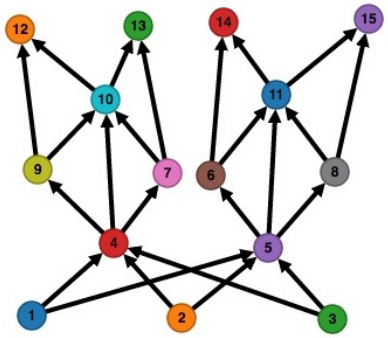


Figure 6.4: Network of the routing game experiment.

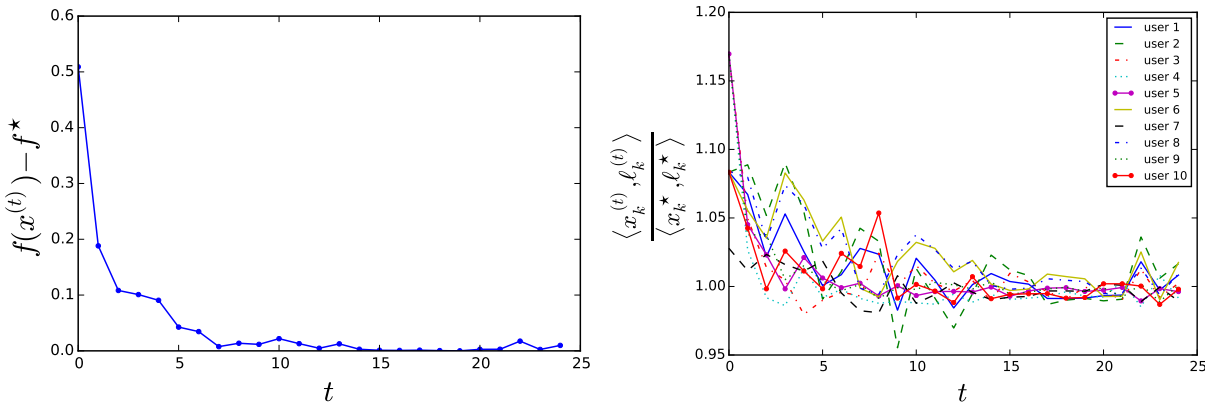


Figure 6.5: Exploration and convergence to equilibrium. The left figure shows the distance to equilibrium, measured by the Rosenthal potential $f(x^{(t)}) - f^\star$ as a function of iteration t. The right figure shows the costs of each player, normalized by the equilibrium costs $\left\langle x_k^{(t)}, \ell_k^{(t)} \right\rangle / \langle x_k^\star, \ell_k(x^\star) \rangle$.

Convergence to equilibrium

First, we evaluate whether the distributed decisions of the players converge to the set of Nash equilibria of the game. The distance to equilibrium can be measured simply by the Rosenthal potential defined in (2.7). Figure 6.5 shows the potential $f(x^{(t)}) - f^\star$ as a function

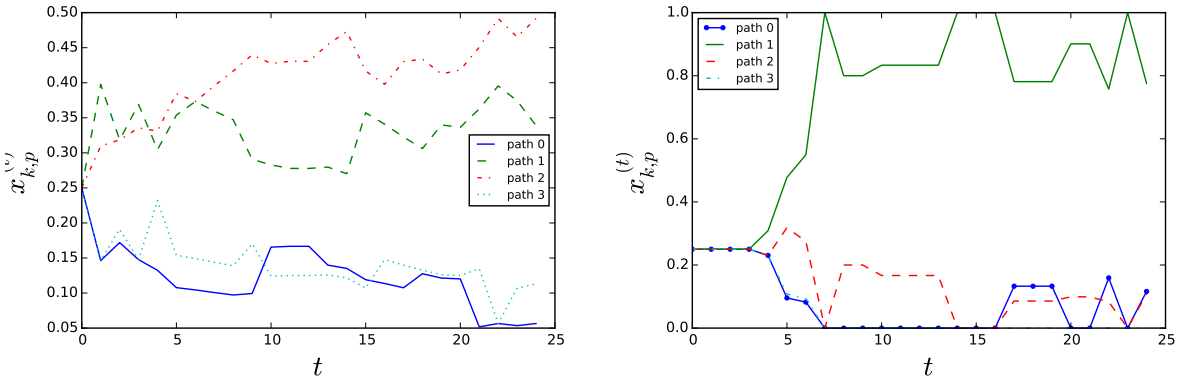


Figure 6.6: Sample mass distributions $(x_k^{(t)})$ for two different players.

of iteration t, as well as the corresponding costs $\left\langle x_k^{(t)}, \ell_k^{(t)} \right\rangle$ of the players, normalized by the equilibrium costs $\langle x_k^\star, \ell_k(x^\star) \rangle$ (so that, close to equilibrium, the normalized costs are close to one). Figure 6.6 shows the mass distributions $x_k^{(t)}$ for two different players. We can observe that at the beginning of the game, there is an exploration phase in which players tend to make aggressive adjustments in their distributions (observe the oscillations in the early iterations on Figure 6.6), while during later turns, the joint distribution $x^{(t)}$ remains close to equilibrium (as measured by the potential function f on Figure 6.5). However, joint distribution moves away from equilibrium close to the end of the game, on iteration 22 (due to a player performing an aggressive update), which results in a sharp increase in the potential value, and we can observe that the players react to this sudden change by significantly changing their distribution during the next iteration.

Learning rate estimation

We now apply the methods proposed in Section 6.1 to estimate the learning rates of each player, then use the estimated rates to predict the decision of the players over a short horizon.

First, we solve Problem (6.4) to estimate the learning rate sequence one term at a time. Figure 6.7 compares the estimated distributions $\hat{x}_k^{(t+1)}(\hat{\eta}_k^{(t)})$, to the actual distributions $x_k^{(t+1)}$, for one of the players. The figure shows that the estimated distributions are close to the actual distributions, which indicates that the mirror descent model is expressive enough to describe the observed behavior of the players.

In addition to estimating one term of the learning rate sequence at a time, we also use the parameterized form $\eta_k^{(t)} = \theta_k t^{-\alpha_k}$, and estimate θ_k and α_k by solving problem (6.6).

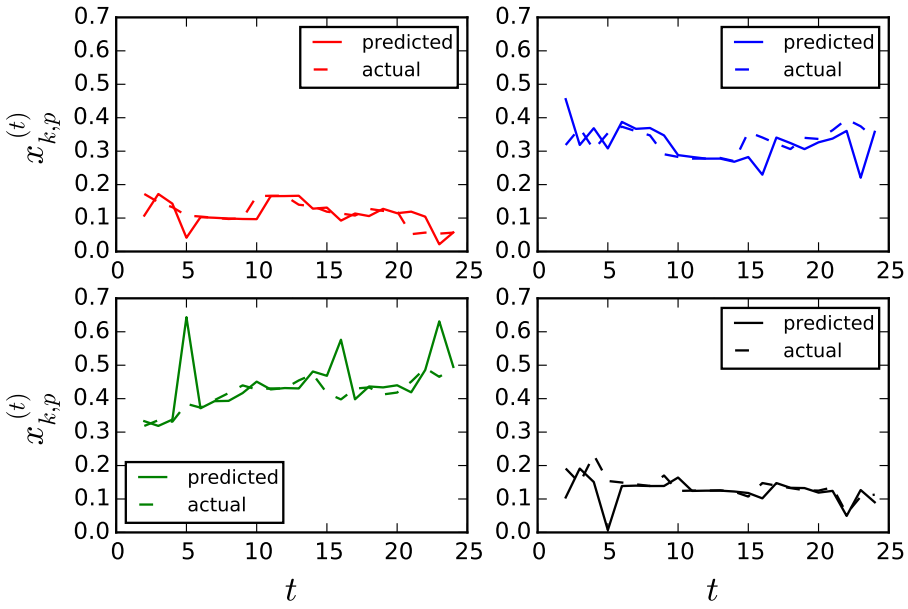


Figure 6.7: Comparison of the distributions $\hat{x}_k^{(t)}$ of the estimated model to the actual distributions $x_k^{(t)}$, for player $k = 2$. Each subplot corresponds to a path.

Predicting future mass distributions

We discuss one possible application of the proposed estimation problem. Once we have computed an estimate of the learning rate sequence, we can propagate the model forward in order to predict the distributions of the players for the next time step. More precisely, if we have computed, at iteration t_0, an estimate of the learning rate sequence given by $(\hat{\eta}_k^{(t)})$, the mirror descent model predicts that the mass distributions obey the update rule

$$\hat{x}^{(t_0+i+1)} = \arg\min_{x_k \in \Delta^{\mathcal{A}_k}} \left\langle x_k, \ell_k(\hat{x}^{(t_0+i)}) \right\rangle + D_{\mathrm{KL}}(x_k, \hat{x}_k^{(t_0+i)}).$$

So starting from the current observation $\hat{x}^{(t_0)} = x^{(t_0)}$, we can propagate the model forward, over a horizon h, by inductively applying the update rule.

Here, we assume that we have an estimate of the entire sequence of learning rates, not just terms up to t_0 (we need the future terms $\hat{\eta}_k^{(t_0+i)}$ to be able to propagate the model). To obtain such an estimate in our experiment, we tested the following simple methods:

1. For the single term estimates (obtained by solving problem (6.4)), we use a stationary sequence, $\hat{\eta}_k^{t_0+i}$, either equal to the last estimate $\hat{\eta}_k^{(t_0)}$, or the average of the last N estimates $\frac{1}{N}\sum_{n=0}^{N-1} \hat{\eta}_k^{(t_0-n)}$.

2. For the parameterized model (obtained by solving problem (6.6)), we can simply use the current estimate of θ_k, α_k and set $\hat{\eta}_k^{(t_0+i)} = \theta_k(t_0+i)^{-\alpha_k}$.

We numerically test these methods to predict the mass distributions of the players over a short horizon $h \in \{1, \ldots, 7\}$. We evaluate each method by computing the average Bregman divergence (per player and per iteration) between the predicted distribution $\hat{x}_k^{(t_0+h)}$ and the actual distribution $x_k^{(t_0+h)}$, i.e.

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{t_{\max}-t_{\min}}\sum_{t_0=t_{\min}}^{t_{\max}-1} D_{\mathrm{KL}}(x_k^{(t_0+h)}, \hat{x}_k^{(t_0+h)}),$$

where $t_{\min}$ is taken to be equal to 5 (so that there is always a minimal history of observations to estimate the parameters). The results are given in Figure 6.8. One can observe that for all methods, as the horizon h increases, the average divergence tends to increase, since the modeling errors propagate and the quality of the predictions degrade. The best overall performance is obtained with the parameterized model $\hat{\eta}_k^{(t)} = \theta_k t^{-\alpha_k}$, although for $h = 1$, the best prediction is achieved using the single term estimates (since this model has as many parameters as time steps, it allows for a much better fit of the observed data, but has poor generalization performance, i.e. its prediction quickly degrades beyond the first iteration).

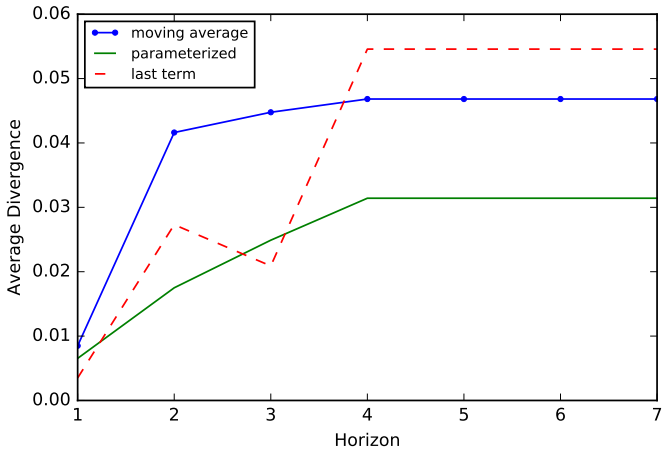
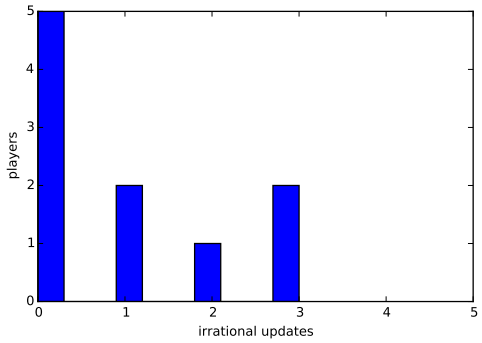


Figure 6.8: Average Bregman divergence per player and per iteration, between the predicted distributions $\hat{x}_k^{(t_0+h)}$ and the actual distributions $x_k^{(t_0+h)}$, as a function of the prediction horizon h.

Irrational updates

It was interesting and perhaps surprising to observe that when estimating learning rates one term at a time, in some rare instances, the objective $d_k^{(t)}(\eta)$, as defined in Equation (6.4),



Path	$x^{(t)}$	$\ell^{(t)}$	$x^{(t+1)}$
p_1	0.197	2.349	0.251
p_2	0.314	1.856	0.285
p_3	0.266	2.435	0.242
p_4	0.223	2.575	0.222

Figure 6.9: Histogram of irrational updates (left), corresponding to iterations t such that the inner product $\left\langle \ell_k^{(t)}, x_k^{(t+1)} - x_k^{(t)} \right\rangle > 0$, which means that the player shifts probability mass to paths with higher costs, which is hard to predict by the model. Example of an irrational update (right), corresponding to iteration $t = 2$ for one of the players. In particular, this player decreased the mass on path p_2 even though this is the best path).

is minimal at a negative η (if we ignore the constraint $\eta \geq 0$), which means that the player shifted the probability mass towards paths with *higher* costs. Fig. 6.9 shows the histogram of the number of irrational updates. In particular, 50% of the players performed at least one irrational update, and a total of 10 irrational updates were observed across all players (corresponding to 4.17% of the total number of updates).

Such behavior is hard to interpret or justify (at least within our framework which models players as sequential decision makers). A negative learning rate does not make sense in our model, since the mirror descent update (6.1), would encourage shifting mass towards paths with higher cost. Thus we add the constraint $\eta \geq 0$ when solving the estimation problem (6.4).

Conclusion

We proposed a problem of learning rate estimation in the mirror descent model, given a sequence of observations of player decisions, and we tested this method on data collected from our routing game experiment. The experimental results suggest that the mirror descent model can be a good descriptive model of player behavior, although in some rare cases, a player decision can be hard to model (e.g. when a player increases the probability mass on previously bad routes).

Chapter 7

Optimal Control Under Hedge Dynamics

As we discussed in the previous chapter, if we consider a system with decision makers who face an online learning problem, we can model their decision dynamics using the mirror descent method studied in Chapter 5. We considered in particular the Hedge dynamics (i.e. mirror descent with the KL divergence) with unknown learning rates, and showed that given a sequence of observations of the player decisions and their losses, we can estimate the learning rates, and tested the method experimentally.

In this chapter, we consider the online learning model of the nonatomic potential game defined in Chapter 2. We suppose that a central authority can control a fraction of the population mass, by deciding their mass distributions, and seeks to improve an objective function over a given horizon, while the remaining mass obeys an online learning algorithm, given by the Hedge dynamics, with known learning rates (these learning rates can be estimated using the methods discussed in the previous chapter).

This results in a non convex, optimal control problem under Hedge dynamics, defined in Section 7.1. We propose two different methods for approximately solving this problem: The first method, presented in Section 7.2 is a greedy algorithm, which sequentially minimizes one term of the objective at a time. In the second method, presented in Section 7.3, we use the adjoint method [87, 56] to perform a local search using the gradient of the objective function by locally linearizing the Hedge constraints. In particular, we derive the adjoint system equations of the Hedge dynamics and show that they can be solved efficiently.

We illustrate these methods on the routing game example from Chapter 2. We first present a simple example on a parallel network in Section 7.4, and discuss the qualitative behavior of each method. Finally, we perform a test on a model of the Los Angeles highway network in Section 7.5, and show the improvement in the total travel time that could be achieved, for various proportions of controlled traffic.

Optimal control problems in routing and transportation

In the one-shot routing game, the partial control problem (controlling a fraction of the flow while the remaining flow responds selfishly), is known as the Stackelberg routing problem, and was proved to be NP-hard even in the simple case of parallel networks with linear latencies [120]. Stackelberg equilibria provide a theoretical framework for understanding the inefficiencies of a network and how much they can be alleviated, but they do not capture *route choice dynamics*.

In other approaches, e.g. [31, 109, 114], one can model and control the dynamics of traffic, by using a macroscopic model based on conservation laws, such as the cell transmission model, that can be obtained as a Godunov discretization of a PDE modeling traffic dynamics known as the Lighthill-Whitham-Richards equation, due to [86, 115], and studied for example in [55]. Our approach is different in that we do not explicitly model the flow dynamics (time scale of minutes or seconds). Instead, we model route choice dynamics (time scale of days), by modeling the players as sequential decision makers. We consider the Hedge dynamics in particular, since it is an instance of both the AREP class and the mirror descent class for which we provided convergence guarantees, and since we have a method for estimating the learning rates for the Hedge algorithm, as discussed in Chapter 6.

7.1 Problem formulation

Consider the model of nonatomic, convex potential games defined in Chapter 2, given by K populations, $\mathcal{S}_1, \dots, \mathcal{S}_K$, such that each population $\mathcal{S}_k$ has an action set $\mathcal{A}_k$. The loss vector of population $\mathcal{S}_k$ is given by a loss function

$$\tilde{\ell}_k : \Delta \to \mathbb{R}^{\mathcal{A}_k},$$

which is a function of the joint mass distribution $x \in \Delta = \Delta^{\mathcal{A}_1} \times \cdots \times \Delta^{\mathcal{A}_K}$, defined in Equation (2.1).

Suppose that a central controller has the task of assigning the actions of a subset of the population $\mathcal{S}_k$, and the rest follows an online learning algorithm given by the Hedge dynamics described in Algorithm 6, with known learning rates. In other words, we partition $\mathcal{S}_k = \mathcal{U}_k \cup \mathcal{X}_k$, where $\mathcal{U}_k$ is the subset of the population controlled by the coordinator, and $\mathcal{X}_k$ is the subset which follows the Hedge dynamics. Let $u_k \in \Delta^{\mathcal{A}_k}$ denote the mass distribution of sub-population $\mathcal{U}_k$, and $x_k \in \Delta^{\mathcal{A}_k}$ the mass distribution of sub-population $\mathcal{X}_k$. Then the total mass distribution of $\mathcal{S}_k$ is simply

$$\tilde{x}_k = \frac{m(\mathcal{U}_k)u_k + m(\mathcal{X}_k)x_k}{m(\mathcal{S}_k)} \in \Delta^{\mathcal{A}_k}$$

and letting $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_K)$, we can redefine the loss vector as a function of (x, u) as follows:

$$\begin{aligned} \ell_k : \Delta \times \Delta &\to \mathbb{R}^{\mathcal{A}_k} \\ (x, u) &\mapsto \tilde{\ell}_k(\tilde{x}). \end{aligned}$$

Now suppose that the coordinator seeks to minimize an objective function over a fixed horizon T, given by

$$\begin{aligned} &J : \Delta^T \times \Delta^T \to \mathbb{R} \\ &(x^{(1:T)}, u^{(1:T)}) \mapsto J(x^{(1:T)}, u^{(1:T)}) = \sum_{t=1}^{T} J^{(t)}(x^{(t)}, u^{(t)}). \end{aligned} \tag{7.1}$$

where $x^{(1:T)}$ denotes the tuple $(x^{(t)})_{1 \le t \le T}$, and similarly for $u^{(1:T)}$. Each function $J^{(t)}$ is assumed to be jointly convex on $\Delta \times \Delta$.

The Hedge dynamics for populations $\mathcal{X}_1, \ldots, \mathcal{X}_K$ is given by Algorithm 6, with a known initial distribution $x^{(1)} = x^{\text{init}}$, and known learning rates $(\eta_k^{(1:T)})$. The loss vectors $\ell_k^{(t)}$ are given by the loss functions defined above

$$\ell_k^{(t)} := \ell_k(x^{(t)}, u^{(t)}).$$

This defines an optimal control problem, in which one seeks to minimize the objective function subject to the Hedge dynamics, summarized below:

$$\text{minimize} \quad \sum_{t=1}^{T} J^{(t)}(x^{(t)}, u^{(t)}) \tag{7.2}$$

$$\text{subject to } u^{(t)} \in \Delta \ \forall 1 \le t \le T, \tag{7.3}$$

$$x^{(1)} = x^{\text{init}}, \tag{7.4}$$

$$x_{k,a}^{(t+1)} = \frac{x_{k,a}^{(t)} e^{-\eta_k^{(t)} \ell_{k,a}(u^{(t)}, x^{(t)})}}{\sum_{a' \in \mathcal{A}_k} x_{k,a'}^{(t)} e^{-\eta_k^{(t)} \ell_{k,a'}(u^{(t)}, x^{(t)})}}. \tag{7.5}$$

We refer to $u^{(t)}$ as the control vectors or control distributions, and $x^{(t)}$ as the selfish distributions (since they correspond to the mass distributions of selfish online learners).

This problem is non-convex in general, due to the equality constraints (7.5) corresponding to the Hedge dynamics. However, we propose two methods for efficiently finding approximate solutions to this optimal control problem.

Example: Minimizing total delay in the routing game

Although the proposed methods apply to general cost functions $J^{(t)}$, we will focus, in our numerical examples, on minimizing total delay in the routing game defined in Section 2.3. In this example, the action set of each population is a set of paths connecting a common origin o_k to a common destination d_k on a given graph. The loss vector corresponds to the delays on the paths, and are given by Equation (2.6). The total delay on the network is

$$J^{(t)}(x, u) = \sum_{k=1}^{K} \left\langle m(\mathcal{U}_k) u_k + m(\mathcal{X}_k) x_k, \ell_k^{(t)} \right\rangle, \tag{7.6}$$

where $m(\mathcal{U}_k)u_k + m(\mathcal{X}_k)x_k$ is the vector of total mass along paths.

7.2 A greedy method

Optimizing terms successively in the cost function

In this method, we minimize the objective function one term at a time, given the state on the previous time steps. That is, we minimize $J^{(t)}(x^{(t)}, u^{(t)})$ given the state and control vectors $(x^{(\tau)}, u^{(\tau)})_{1\leq\tau\leq t-1}$. Since the state at time $t-1$ completely determines $x^{(t)}$ by the Hedge update equation (7.5), the subproblem becomes

$$\text{minimize}_{u\in\Delta} J^{(t+1)}(x^{(t+1)}, u). \tag{7.7}$$

In words, the controller anticipates the move of the selfish players and myopically optimizes the objective on the next iteration. This is a convex optimization problem since $J^{(t)}$ is convex by assumption, so it can be solved using mirror descent on the product of simplices Δ. The greedy algorithm can then be summarized as follows:

Algorithm 7 Greedy method for optimal control under Hedge dynamics

1: Input: x^{init} and $\eta_k^{(1:T)}$ are given.
2: $x^{(1)} = x^{\text{init}}$.
3: **for** each time step $1 \leq t \leq T$ **do**
4: Solve $u^{(t)} = \arg\min_{u\in\Delta} J^{(t)}(x^{(t)}, u)$.
5: Compute $x^{(t+1)}$ according to Equation (7.5).
6: **end for**
7: **return** $u^{(1:T)}$.

Computational complexity

Since the optimal control problem is non-convex in general, we cannot provide guarantees on convergence to a desired precision. However, studying the computational complexity of a single iteration, as a function of the problem size, can be useful in evaluating how well the method scales. Therefore, in analyzing the computational complexity of the two proposed methods, we will only consider the dependence on the size of the problem $|\mathcal{A}_k|$, and not on the desired precision ϵ.

The greedy method solves T convex optimization problems on the product of simplices Δ, where each problem is followed by a Hedge update of the mass distribution. Each iteration of mirror descent requires computing the gradient of the objective $J^{(t)}$, then updating the distribution $u^{(t)}$, which has a linear cost $\mathcal{O}(|\mathcal{A}_k|)$. Thus the total complexity of each problem is $\mathcal{O}(|\mathcal{A}_k|)$, and the total complexity of the greedy method is $\mathcal{O}(T\sum_{k=1}^{K}|\mathcal{A}_k|)$.

7.3 The adjoint method

In this section, we propose to use the adjoint method to find a local minimum of the non-convex problem (7.2). First, we can reformulate problem (7.2) as follows:

$$\begin{aligned} \text{minimize} \quad & J(x,u) \\ \text{subject to} \quad & H(x,u) = 0 \\ & x \in \times_{t=1}^{T}\Delta, u \in \times_{t=1}^{T}\Delta \end{aligned} \tag{7.8}$$

where we used x to denote the entire tuple $x^{(1:T)}$ (we drop the time indices to simplify notation), and we define a function H on $\mathbb{R}^{NT} \times \mathbb{R}^{NT}$, with values in $\mathbb{R}^{NT}$, to encode the constraints on the selfish distributions x, where $N = \sum_{k=1}^{K} |\mathcal{A}_k|$ is the total number of actions. The constraint function H can be obtained simply from the initial condition constraint (7.4) and the Hedge update equation (7.5), as follows: $\forall k, \forall a \in \mathcal{A}_k$,

$$H_{k,a}^{(1)}(x,u) = x_{k,a}^{(1)} - x_{k,a}^{(init)}, \tag{7.9}$$

$$H_{k,a}^{(t)}(x,u) = x_{k,a}^{(t)} - \frac{x_{k,a}^{(t-1)} e^{-\eta_k^{(t-1)} \ell_{k,a}(x^{(t-1)}, u^{(t-1)})}}{\sum_{a' \in \mathcal{A}_k} x_{k,a'}^{(t-1)} e^{-\eta_k^{(t-1)} \ell_{k,a'}(x^{(t-1)}, u^{(t-1)})}} \qquad \text{for } 2 \le t \le T. \tag{7.10}$$

The adjoint method is a general local search method for solving optimal control problems under non-linear constraints, of the form given in problem (7.8). It is derived using the stationarity conditions of Pontryagin's maximum principle. For an introduction to the adjoint method in optimal control, see for example [48, 110] and references therein. A complete exposition of the adjoint method is beyond the scope of this chapter, but we give below a formal derivation and intuitive interpretation of the adjoint system equations.

Since the control distributions $u^{(1:T)}$ entirely determine the selfish distributions $x^{(1:T)}$, let us assume, for the sake of discussion, that $x^{(1:T)}$ can be written as $x^{(1:T)} = X(u^{(1:T)})$ for some differentiable function $X : \times_{t=1}^{T}\Delta \to \times_{t=1}^{T}\Delta$. The optimal control problem would then be equivalent to minimizing the function $J(X(u^{(1:T)}), u^{(1:T)})$ over the feasible set $\times_{t=1}^{T}\Delta$, and we can use the mirror descent algorithm to solve this problem, since the constraint set is a product of simplices. To apply mirror descent, we need to compute, at each iteration, the gradient of the function $u \mapsto J(X(u), u)$, which we denote $\nabla_u J(x,u)$. Using the chain rule, we have the following expression of the gradient

$$\nabla_u J(x,u) = \frac{\partial J}{\partial x}(x,u) \nabla_u X(u) + \frac{\partial J}{\partial u}(x,u), \tag{7.11}$$

where the Jacobian term $\nabla_u X(u)$, which represents the dependence of $x^{(1:T)}$ on the control $u^{(1:T)}$, can be expensive to compute. The adjoint method provides a different approach to computing the gradient (7.11) without explicitly computing the Jacobian $\nabla_u X(u)$: Since $H(X(u), u) = 0$, we have, taking derivatives,

$$\frac{\partial H}{\partial x}(x,u) \nabla_u X(u) + \frac{\partial H}{\partial u}(x,u) = 0. \tag{7.12}$$

Algorithm 8 Adjoint method for optimal control under Hedge dynamics

1: Initialize $i = 0$, $u^{[0]} \in \times_{t=1}^{T} \Delta^u(\alpha, F)$, and $x^{[0]}$ by solving $H(x^{[0]}, u^{[0]}) = 0$.
2: **while** stopping criterion not satisfied **do**
3: Solve the adjoint system

$$\left[\frac{\partial H}{\partial x}(x^{[i]}, u^{[i]})\right]^T \lambda^{[i]} = -\left[\frac{\partial J}{\partial x^{[i]}}(x^{[i]}, u^{[i]})\right]^T.$$

4: Compute the gradient

$$g^{[i]} = \nabla_u J(x^{[i]}, u^{(i)}) = \lambda^{[i]^T} \frac{\partial H}{\partial u}(x^{[i]}, u^{[i]}) + \frac{\partial J}{\partial u}(x^{[i]}, u^{[i]}).$$

5: Perform one mirror descent step in the direction $-g^{[i]}$: $\forall k$, $\forall t \in \{1, \dots, T\}$, $\forall a \in \mathcal{A}_k$,

$$u_{k,a}^{(t)[i+1]} \propto u_{k,a}^{(t)[i]} \exp(-\beta_i g_{k,a}^{(t)[i]}).$$

6: Update $x^{[i+1]}$ by solving $H(x^{[i+1]}, u^{[i+1]}) = 0$.
7: Update $i \leftarrow i + 1$.
8: **end while**
9: **return** Control solution $u^{[i_{\text{best}}]}$.

Therefore if we let λ be a solution to the system

$$\left[\frac{\partial H}{\partial x}(x, u)\right]^T \lambda = -\left[\frac{\partial J}{\partial x}(x, u)\right]^T, \tag{7.13}$$

called the adjoint system, then

$$\lambda^T \frac{\partial H}{\partial u}(x, u) = -\lambda^T \frac{\partial H}{\partial x}(x, u) \nabla_u X(u) = \frac{\partial J}{\partial x}(x, u) \nabla_u X(u),$$

where we used (7.12) in the first equality and (7.13) in the second. Plugging this expression in (7.11), we obtain the following expression of the gradient

$$\nabla_u J(x, u) = \lambda^T \frac{\partial H}{\partial u}(x, u) + \frac{\partial J}{\partial u}(x, u). \tag{7.14}$$

We apply the adjoint method as follows: at each iteration i, we solves the adjoint system equations (7.13) and compute the gradient using (7.14), then perform one mirror descent step in the direction of the gradient. This is summarized in Algorithm 8, where we use the superscript $[i]$ to denote step i in the algorithm, not to be confused with superscript

(t), which denotes time t (corresponding to one term of the objective function). In the experiments, we can run the method from multiple random initial points $u^{[0]}$ and keep the best local minimum.

Derivation of the adjoint system equations for the Hedge dynamics

In this section, we explicitly derive, for the Hedge learning dynamics, the adjoint system equations (7.13). First, since $H^{(t)}$ only depends on $x^{(t)}$, $x^{(t-1)}$ and $u^{(t-1)}$, we have for all k, k', $a \in \mathcal{A}_k$, $a' \in \mathcal{A}_{k'}$,

$$\frac{\partial H^{(t)}_{k,a}}{\partial x^{(s)}_{k',a'}}(x,u) = 0 \qquad \forall s \notin \{t-1, t\},$$
$$\frac{\partial H^{(t)}_{k,a}}{\partial u^{(s)}_{k',a'}} = 0 \qquad \forall s \neq t-1.$$

Then, to simplify the derivation, let $A^{(t-1)}, B^{(t-1)}$ denote the Jacobian of the loss function $\ell^{(t-1)}$ with respect to $x^{(t-1)}$ (respectively $u^{(t-1)}$), so that

$$A^{(t-1)}_{k,k',a,a'} = \frac{\partial \ell_{k,a}(x^{(t-1)}, u^{(t-1)})}{\partial x^{(t-1)}_{k',a'}},$$
$$B^{(t-1)}_{k,k',a,a'} = \frac{\partial \ell_{k,a}(x^{(t-1)}, u^{(t-1)})}{\partial u^{(t-1)}_{k',a'}},$$

and let

$$w^{(t-1)}_{k,a} = \exp(-\eta^{(t-1)}_k \ell_{k,a}(x^{(t-1)}, u^{(t-1)})),$$
$$W^{(t-1)}_k = \sum_{a \in \mathcal{A}_k} w^{(t-1)}_{k,a}.$$

Using this notation, and the Kronecker delta $\delta^{k'}_k = 1$ if $k = k'$ and 0 otherwise, we have

$$\frac{\partial H^{(t)}_{k,a}}{\partial x^{(t-1)}_{k',a'}} = -w^{(t-1)}_{k,a} \left[x^{(t-1)}_{k,a} \eta^{(t-1)}_k \left(\frac{A^{(t-1)}_{k,k',a,a'}}{W^{(t-1)}_k} - \frac{\sum_{b \in \mathcal{A}_k} {x_{k,b}}^{(t-1)} w^{(t-1)}_{k,b} A^{(t-1)}_{k,k',b,a'}}{(W^{(t-1)}_k)^2} \right) + \delta^{k'}_k x^{(t-1)}_{k,a} \frac{w^{(t-1)}_{k',a'}}{(W^{(t-1)}_k)^2} - \delta_{a,a'} \frac{1}{W^{(t-1)}_k} \right]. \quad (7.15)$$

And

$$\frac{\partial H^{(t)}_{k,a}}{\partial u^{(t-1)}_{k,a'}} = -w^{(t-1)}_{k,a} x^{(t-1)}_{k,a} \eta^{(t-1)}_k \left(\frac{B^{(t-1)}_{k,k',a,a'}}{W^{(t-1)}_k} - \frac{\sum_{b \in \mathcal{A}_k} x^{(t-1)}_{k,b} w^{(t-1)}_{k,b} B^{(t-1)}_{k,k',b,a'}}{(W^{(t-1)}_k)^2} \right). \quad (7.16)$$

Complexity analysis

The method performs a mirror descent on the product of simplices $\times_{t=1}^{T}\Delta$. Now, each gradient evaluation requires solving the adjoint system (7.13) then using the expression (7.14) to compute the full gradient. One gradient evaluation thus requires calculating the partial derivatives $\frac{\partial J}{\partial u}(x,u), \frac{\partial J}{\partial x}(x,u) \in \mathbb{R}^{NT}$, and $\frac{\partial H}{\partial x}(x,u), \frac{\partial H}{\partial u}(x,u) \in \mathbb{R}^{NT\times NT}$, then solving the adjoint system (7.13) for $\lambda \in \mathbb{R}^{NT}$. Since the cost function $J^{(t)}$ at time-step t only depends on $x^{(t)}$ and $u^{(t)}$, the matrix $\frac{\partial H}{\partial x}(x,u)$ is banded lower-triangular, and contains $\mathcal{O}(TN^2)$ non-zero terms. Therefore solving the adjoint system can be done in $\mathcal{O}(TN^2)$ using Gaussian elimination, as discussed for example in [30] Appendix C-2. Therefore, the total computational complexity of the adjoint method scales as $\mathcal{O}(TN^2)$. It is linear in T similarly to the greedy method, but scales quadratically in the number of actions, N.

7.4 Optimal routing on the Pigou network

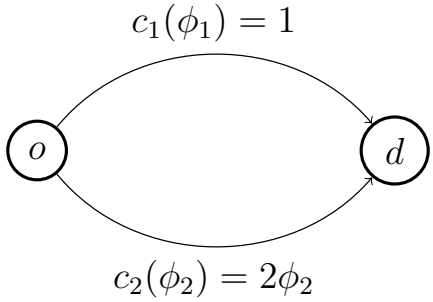


Figure 7.1: Pigou network used in the numerical experiment.

To illustrate the qualitative difference between the greedy and the adjoint solutions, we consider a simple example known as the Pigou network, given in Figure 7.1. Consider a single population of players $\mathcal{S}_1$, with total mass $m(\mathcal{S}_1) = 1$, and with two paths connecting the origin to the destination, each consisting of a single edge. The congestion function on the top edge is constant, $c_1(\phi_1) = 1$, and the congestion function on the second edge is linear, $c_2(\phi_2) = 2\phi_2$.

This routing game has a unique Nash equilibrium, given by $x^{\text{Nash}} = (\frac{1}{2}, \frac{1}{2})$ (under this equilibrium, both edges have the same loss). It can be obtained by minimizing the Rosenthal potential function defined in (2.7), which in this case is given by

$$f(x) = \int_0^{x_1} c_1(u)du + \int_0^{x_2} c_2(u)du = x_1 + x_2^2 = x_1 + (1-x_1)^2,$$

which is minimal at $x_1 = \frac{1}{2}$. The total delay of the network is

$$\langle x, \ell(x)\rangle = x_1 + 2x_2^2 = x_1 + 2(1-x_1)^2,$$

which is minimal at $x^{(\text{social})} = (\frac{1}{4}, \frac{3}{4})$, usually referred to as the social optimum of the routing game, since it minimizes the total (social) cost across the entire population (but x^{social} is clearly not a Nash equilibrium).

Now suppose that a coordinator controls a fraction α of the total mass, and let u be the mass distribution of the controlled population, and x that of the selfish mass. Then the total delay of the network, defined in (7.6), is given by

$$J(u, x) = \alpha u_1 + (1 - \alpha)x_1 + 2[\alpha u_2 + (1 - \alpha)x_2]^2.$$

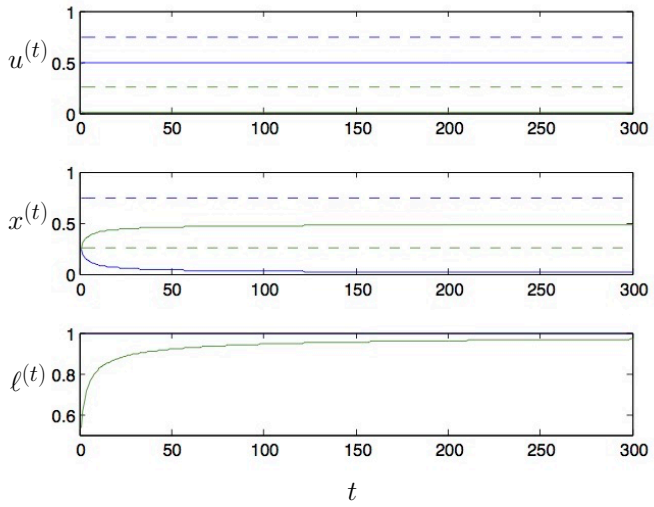
The particularity of this network is that if we consider the Stackelberg game, in which the controller chooses u, and the rest of the population plays a Nash equilibrium x induced by u, no Stackelberg strategy can improve the total delay if $\alpha \leq \frac{1}{2}$. Indeed, any allocation u will induce the same total mass distribution $\alpha u + (1 - \alpha)x$, equal to $(\frac{1}{2}, \frac{1}{2})$. So for the one-shot game, the total delay at equilibrium cannot be improved. However, in our online learning model, one may take advantage of the learning dynamics of the selfish population to reduce the cost on a finite horizon.

We assume that the selfish population obeys the Hedge dynamics, starting from the uniform distribution, and with learning rates $\eta_1^{(t)} = \frac{1}{\sqrt{t}}$. Without control (i.e. $\alpha = 0$), the selfish distribution is stationary, $x^{(t)} = (\frac{1}{2}, \frac{1}{2})$ for all t, since it starts at the Nash equilibrium. We simulate the greedy method and the adjoint method on a horizon $T = 300$, with $\alpha = \frac{1}{2}$. The value of the total delay obtained for each solution is given in the table below, where we compare the solutions of the greedy method and the adjoint method to the selfish solution (obtained by setting $\alpha = 0$, i.e. we simply let the selfish population follow the Hedge dynamics), and the social optimum, obtained by setting $\alpha = 1$, (i.e. we control the entire population).

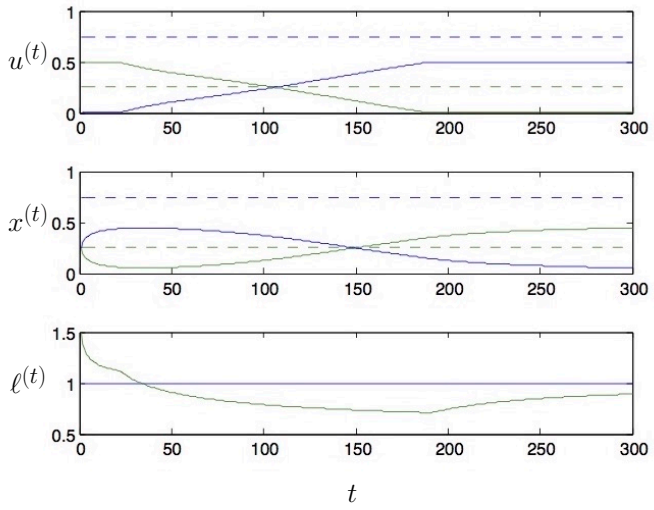
J^{social}	J^{greedy}	J^{adjoint}	J^{selfish}
262.5	291.7	283.0	300.0

The greedy and the adjoint solutions are illustrated in Figure 7.2, and are quite different qualitatively: The greedy solution assigns all the controlled flow $u^{(t)}$ to the upper path for all t, since this is the best myopic decision at any time (while the selfish mass keeps shifting to the second path). The adjoint solution, however, first allocates mass $u^{(t)}$ to the lower path, which results in a decrease of the selfish mass on that path. Then, $u^{(t)}$ is moved to the upper path, which results in decreasing the cost on the lower path.

The per-time-step costs $J^{(t)}$ for both solutions are given in Figure 7.3, where we can observe that the adjoint solution sacrifices the cost on the first few time-steps for a better cost on later time steps. In particular, this example illustrates the limitations of the greedy approach, since, by definition, it does not anticipate the dynamics of the selfish population over several time steps.



(a) Greedy solution.



(b) Adjoint solution.

Figure 7.2: Control solution on the Pigou network. Controlled mass $u^{(t)}$ (top), selfish mass $x^{(t)}$ (middle) and corresponding path losses $\ell^{(t)} = \ell(\alpha u^{(t)} + (1-\alpha)x^{(t)})$ (bottom). The green lines correspond to the top path, and the blue lines to the bottom path. The dashed lines show the social optimum $x^{\text{social}} = (\frac{1}{4}, \frac{3}{4})$.

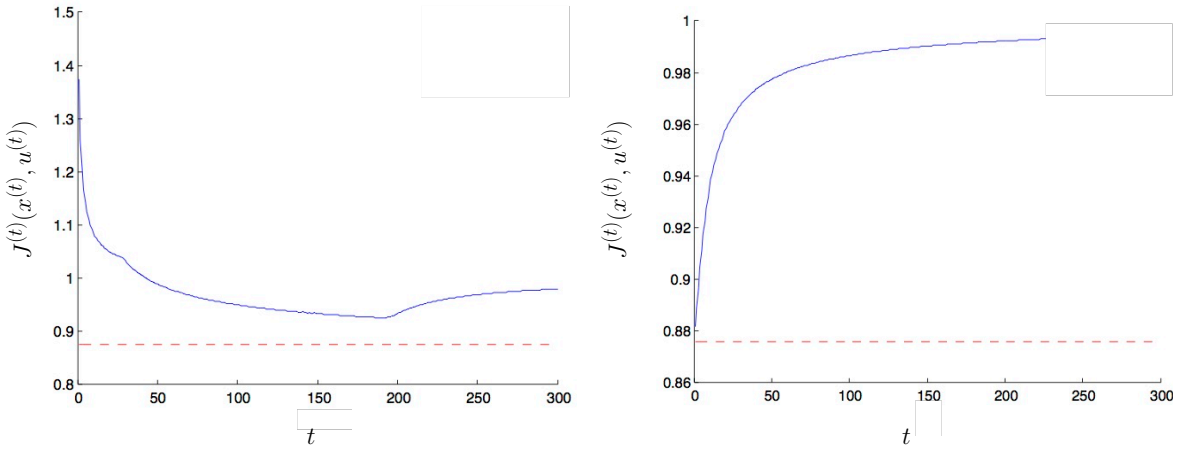


Figure 7.3: Profile of network delays $J^{(t)}$ over time, induced by adjoint solution (left) and the greedy solution (right). The dashed line shows the cost of the social optimum.

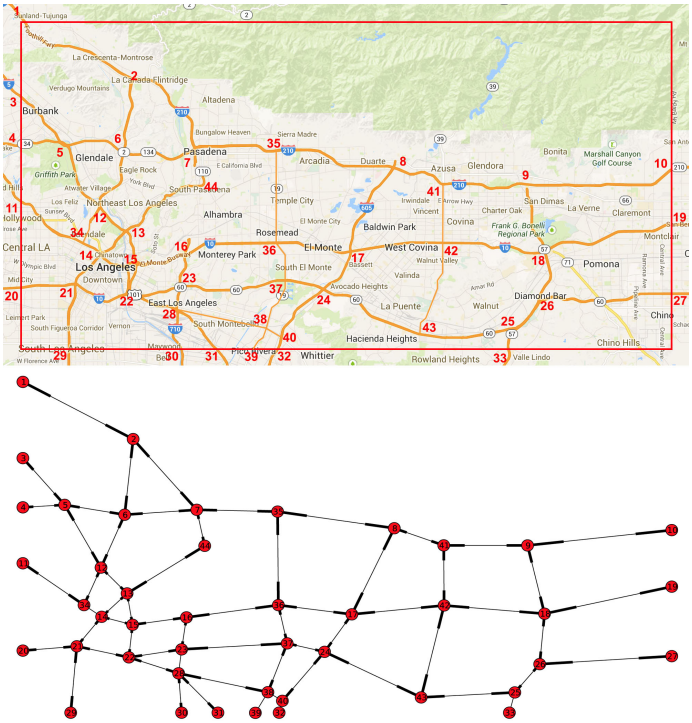


Figure 7.4: Los Angeles highway network and its graph model.

7.5 Numerical experiment on the Los Angeles highway network

In this section, we consider a model of the Los Angeles highway road network, used in [133], and illustrated in Figure 7.4. The network topology is obtained from OpenStreetMap data,

by keeping highways that contain five lanes or more. We consider $K = 42$ origin-destination pairs, illustrated in Figure 7.5, for the following destinations: Hollywood (node 5), Santa Monica (node 20), Central L.A. (node 22). The congestion functions are those estimated by the Bureau of Public Roads for a network in quasi-static equilibrium [133]. More precisely, the congestion function is assumed to be of the form $c_e(\phi_e) = d_e D(\phi_e/\xi_e)$, where $D(x) = 1 + 0.15x^4$, ϕ_e is the edge mass, d_e a minimal delay on the edge, and ξ_e is called the capacity on edge e. The simulations are run with a time horizon $T = 20$, learning rates $\eta_k^{(t)} = \frac{1}{\sqrt{t}}$, with values of $\alpha \in \{.1, .3, .5, .7, .9, 1\}$. The results are given in Figure 7.6, and discussed bellow. In addition to the adjoint method described in Section 7.3, which uses mirror descent with a fixed sequence of learning rates (β_i), we also implement a version of the adjoint method with backtracking line search, which uses the Armijo rule to set β_i.

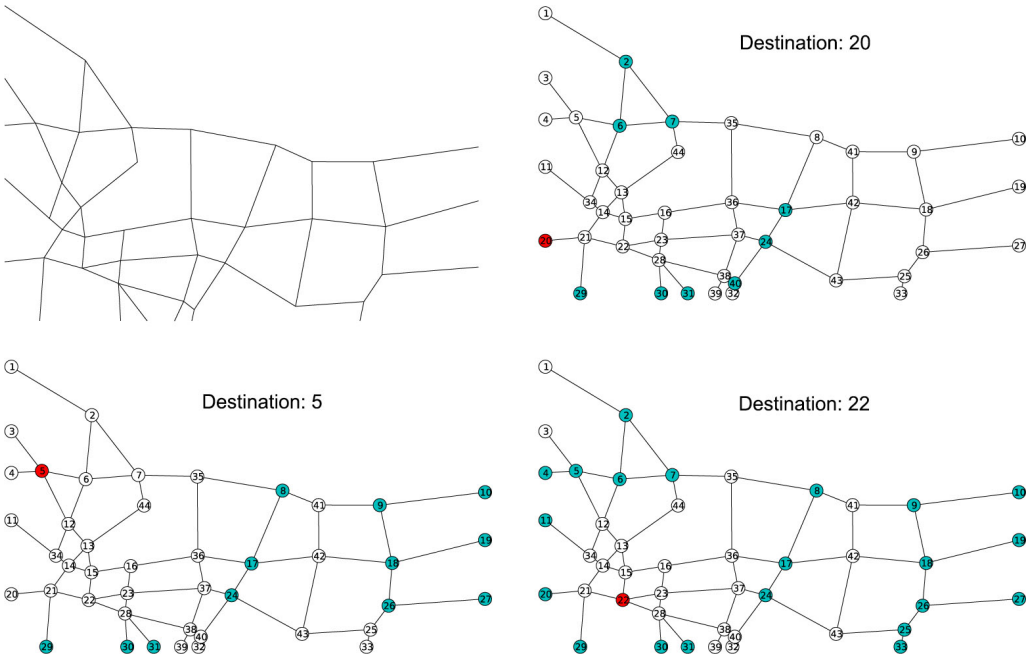
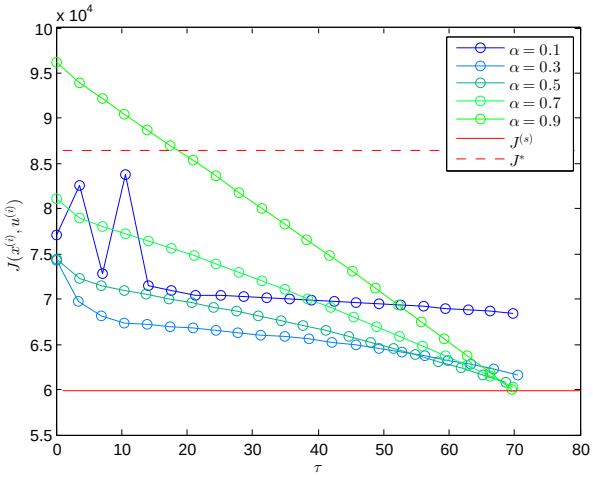


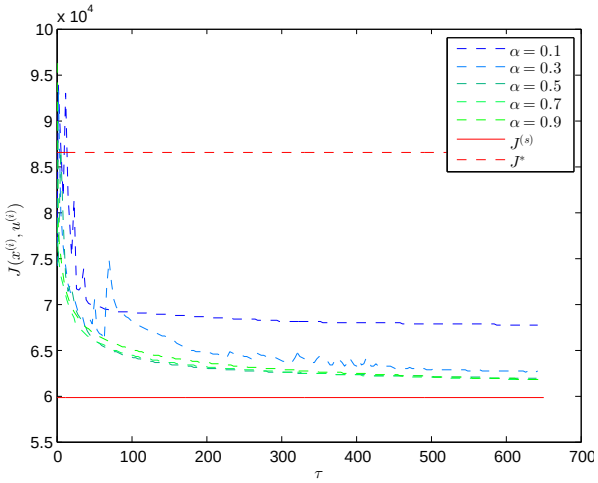
Figure 7.5: Selected origins (blue) and destinations (red) on the Los Angeles highway network.

Effect of increasing control

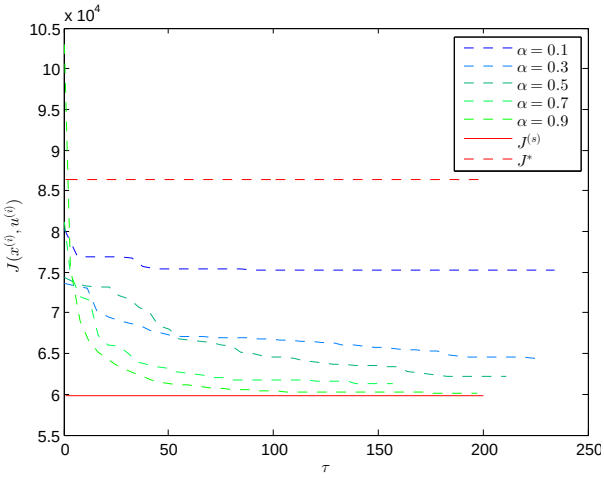
First, we observe that increasing the control parameter α results in a decrease in the total delay, and for higher values of α, the value of the objective becomes close to that of the social optimum. Although intuitive, this cannot be guaranteed in general, since the problem is non-convex for all α strictly between 0 and 1, so the problem may converge to a worse local minimum for a lower value of α.



(a) Greedy method.



(b) Adjoint method.



(c) Adjoint method with backtracking line search.

Figure 7.6: Total delay $J(x^{[i]}, u^{[i]})$, as a function of iteration number i, for the proposed methods, with different control proportion α. The red solid and dotted lines represent, respectively, the social optimum ($\alpha = 1$) and the selfish response without control ($\alpha = 0$).

Numerical Results for $\alpha = 0.1$

We now have a more detailed look at the performance of each method with a fixed $\alpha = 0.1$. The values of the objective function for each method are reported in the following table, and the average delay per vehicle per day is reported as a function of iteration number in Figure 7.6.

J^{social}	J^{greedy}	J^{adjoint}	$J^{\text{adjoint\_ls}}$	J^{selfish}
25.4	29.0	28.7	32.0	36.7

We observe that most of the methods do not guarantee a decrease in the value of the objective from one iteration to the next, except the adjoint method with line search, which, by definition, searches for a step size which guarantees a descent (by Armijo's rule). Nevertheless, the adjoint method without line search performs best, and converges to a local minimum with lower objective value than that with line search. This may be a result of line search being too conservative: Requiring the Armijo rule to be satisfied at each iteration may prevent the method from exploring the search space. The greedy method performs surprisingly well, and is within 3% of the (normalized) objective value of the best method. The convex penalization is the worst performing among all methods, although it still results in a 38% decrease in the distance to social optimum. Finally, it is worth observing that even when controlling a fraction of the population as small as $\alpha = 0.1$, the improvement in the total delay can be significant (70% reduction in the distance to social optimum).

7.6 Conclusion

In the first part of the thesis, we studied decision dynamics for online learning in nonatomic convex potential games. We studied several classes of dynamics using different techniques, each leading to different convergence guarantees of the sequence of mass distributions $(x^{(\tau)})$ to the set of Nash equilibria, i.e. the set of minimizers of the potential function. We first showed that algorithms with sublinear regret guarantee convergence in the sense of Cesàro. Then we showed that approximate replicator (AREP) algorithms, obtained by taking a discrete time approximation of the replicator ODE, guarantee almost sure convergence. Then using results from stochastic optimization, we gave a more detailed analysis of stochastic mirror descent dynamics, and derived convergence rates, both in the homogeneous and heterogeneous models of learning. In particular, we used connections between discrete and continuous time dynamics both to motivate the study of the replicator ODE (by showing that it can be obtained as a continuous-time limit of the Hedge algorithm), and to relate the asymptotic properties of the discrete process to the those of the solution trajectories of the ODE (by using the notion of asymptotic pseudo trajectories).

We showed that the Hedge algorithm is both an instance of approximate replicator algorithms, studied in Chapter 4, and of mirror descent algorithms, studied in Chapter 5. Using Hedge as a model of decision dynamics, we studied an estimation problem in Chapter 6 and an optimal control problem in Chapter 7. First, assuming we can observe a sequence of decisions of a player who follows the Hedge dynamics with unknown learning rates, we proposed a method to estimate the learning rates to fit the model to the observations. We demonstrated this approach on field data collected using a web application that simulates

the routing game, and showed that the model of Hedge dynamics is descriptive of actual player decisions, and can even be used to predict future decisions over short horizons. Second, assuming that players follow Hedge dynamics, we posed an optimal control problem in which we control the decisions of a small fraction of players. We used the adjoint method to compute a local minimizer of the problem, and demonstrated this approach on different examples of the routing game, showing that in a realistic model of congestion in transportation networks, control over a small fraction of traffic could potentially lead to significant improvements of the network-wide efficiency.

While we focused on the routing game as an application of our estimation and optimal control problem, there are several additional examples of systems which involve sequential decision makers (either humans or automated computer systems), and in which a central coordinator has control over a fraction of the players, or over some parameters of the system, e.g. through pricing or tolling. This is for example the case in power networks, and auction platforms for online advertising. In such systems, the dynamics of the decision makers can be similarly modeled using the Hedge algorithm or other instances of the mirror descent family, and one can use the approach proposed in Chapters 6 and 7 respectively to estimate the learning rates, and to optimally control the system.

Part II

Accelerated Dynamics for Constrained Convex Optimization

Chapter 8

Accelerated Mirror Descent in Continuous Time

In the second part of the thesis, we study dynamics for constrained convex optimization. Similarly to the first part, we will exhibit connections between discrete and continuous-time dynamics, by first designing dynamics in the continuous-time domain, using a Lyapunov approach, then discretizing the resulting ODE to obtain a discrete algorithm. We start from two important families of methods: the mirror descent method, due to Nemirovski and Yudin [98], and Nesterov's original accelerated method for unconstrained convex optimization [102]. Both methods can be interpreted as a discretization of a continuous-time dynamics, and their continuous-time ODE can be analyzed using simple Lyapunov functions. Combining ideas from both methods, we show that for constrained convex problems, one can define a natural energy function which encodes the constraints and the desired convergence rate, then design the dynamics to make that function a Lyapunov function. This is different from the usual approach in which one starts from a given dynamics then looks for an appropriate Lyapunov function.

We give different interpretations of the resulting dynamics. In particular, we show that it can be interpreted as the equations of motion of a particle in a potential field with viscous friction, with a time-varying friction coefficient, which sheds light on some of the qualitative behavior typically observed in accelerated descent methods. We also show that the dynamics can be interpreted as coupled ODEs of a dual variable $Z(t)$ cumulating gradients at a rate $\eta(t)$, and a primal variable $X(t)$ obtained as the weighted average of the mirror of the dual trajectory, with weights $w(t)$. This interpretation makes a rigorous connection between acceleration and averaging, which was previously observed in the special case of quadratic functions in [50]. As an example, we show that the replicator dynamics studied in Part I can be accelerated using averaging.

This also motivates the study of a more general averaging scheme in Chapter 9, in which we give sufficient conditions on the primal and dual weight functions w and η to guarantee a given convergence rate. We also propose an adaptive averaging heuristic, which intuitively works by increasing weights on portions of the trajectory which make the most progress.

This heuristic empirically gives faster convergence and alleviates the oscillations typically observed in accelerated methods. It also compares favorably to other popular heuristics, such as restarting, and gives significant improvements in many cases. All these heuristics (adaptive averaging and the different restarting conditions) have been developed as a result of the different continuous-time interpretations, which shows some of the advantages of studying the continuous-time dynamics. This not only helps simplify and guide the analysis (although the discrete-time analysis is usually more involved than its continuous-time counterpart, as discussed in Chapter 10), but it also results in new insights and interpretations, which can lead to a better understanding of the dynamics, and to heuristics to adaptively improve the speed of convergence.

8.1 Introduction

We consider a constrained convex optimization problem,

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{X}, \end{array}$$

where $\mathcal{X} \subseteq \mathbb{R}^n$ is the feasible set, assumed to be convex and closed, f is a C^1 convex function, and its gradient, ∇f is assumed to be L_f-Lipschitz with respect to a pair of dual norms ($\|\cdot\|$, $\|\cdot\|_*$), i.e. $\|\nabla f(x) - \nabla f(y)\|_* \leq L_f \|x - y\|$ for all $x, y \in \mathcal{X}$. Let $S \subset \mathcal{X}$ be the set of minimizers of f on $\mathcal{X}$, and suppose that S is non-empty. Let $f^\star$ be the value of f on S. Many convex optimization methods can be interpreted as the discretization of an ordinary differential equation, the solutions of which are guaranteed to converge to S. Perhaps the simplest such method is gradient descent for the unconstrained problem (when $\mathcal{X}$ is all of $\mathbb{R}^n$), given by the iteration $x^{(k+1)} = x^{(k)} - s\nabla f(x^{(k)})$ for some step size $s > 0$, which can be interpreted as the discretization of the ODE $\dot{X}(t) = -\nabla f(X(t))$, with discretization step s. The well-established theory of ordinary differential equations can provide guidance in the design and analysis of optimization algorithms, and has been used for unconstrained optimization [32, 26, 64], constrained optimization [125] and stochastic optimization [112]. It has also been applied to second-order methods, for example the Hessian-driven damping method in [7], and to more general problems, such as finding a zero of a monotone operator [3]. In particular, proving convergence of the solution trajectories of an ODE can often be achieved using simple and elegant Lyapunov arguments. The ODE can then be carefully discretized to obtain an optimization algorithm for which the convergence rate can be analyzed by using an analogous Lyapunov argument in discrete time.

In this chapter, we focus on two families of first-order methods: Nesterov's accelerated method [102], and Nemirovski's mirror descent method [98]. First-order methods have become increasingly important for large-scale optimization problems that arise in machine learning applications. Nesterov's accelerated method [102] has been applied to many problems and extended in a number of ways, see for example [103, 101, 100, 14]. The mirror descent method also provides an important generalization of the gradient descent method

to constrained, non-Euclidean geometries, as discussed in [98, 47, 131, 15], and has many applications in convex optimization [24, 23, 43, 70], as well as online learning [34, 40]. An intuitive understanding of these methods is of particular importance for the design and analysis of optimization algorithms. Although Nesterov's method has been notoriously hard to explain intuitively [69], progress has been made recently: in [130], Su et al. give an ODE interpretation of Nesterov's method. However, this interpretation is restricted to the original method [102], and does not apply to constrained, non-Euclidean geometries. In [2], Allen-Zhu and Orecchia give another interpretation of Nesterov's method, as performing, at each iteration, a convex combination of a mirror step and a gradient step. Although it covers a broader family of algorithms (including non-Euclidean geometries), this interpretation still requires an involved analysis, and lacks the simplicity and elegance of ODEs. We provide a new interpretation which has the benefits of both approaches: we show that a broad family of accelerated methods (which includes those studied in [130] and [2]) can be obtained as a discretization of a simple ODE, which is guaranteed to converge in $\mathcal{O}(1/t^2)$.

The continuous-time analysis of Nesterov's method [130] and that of mirror descent [98] both rely on a Lyapunov argument. They are reviewed in Section 8.2. By combining these ideas, we propose, in Section 8.3, a candidate Lyapunov function $\mathcal{V}(t) := V(X(t), Z(t), t)$ that depends on two state variables: $X(t)$, which evolves in the primal space $E = \mathbb{R}^n$ (more precisely, $X(t)$ evolves in the feasible set $\mathcal{X} \subset E$), and $Z(t)$, which evolves in the dual space E^*, and we design coupled dynamics of (X, Z) to guarantee that $\frac{d}{dt}\mathcal{V}(t) \leq 0$. Such a function is said to be a Lyapunov function in reference to Aleksandr Mikhailovich Lyapunov [88]; see also [71] for an introduction to Lyapunov theory in the context of modern control theory. This derivation leads us to a new family of ODE systems, given by

$$\text{AMD} \begin{cases} \dot{Z} = -\dfrac{t}{r}\nabla f(X) \\ \dot{X} = \dfrac{r}{t}(\nabla\psi^*(Z) - X) \\ X(0) = x_0,\ Z(0) = z_0 \text{ with } \nabla\psi^*(z_0) = x_0 \end{cases} \tag{8.1}$$

where r is a positive parameter, and ψ^* is a distance generating function on E^* with Lipschitz gradient, and such that its gradient, $\nabla\psi^*$, is a mapping from the dual space E^* to the feasible set $\mathcal{X}$; it is usually referred to as the mirror operator, and such a function can be constructed using standard results from convex analysis, by taking the convex conjugate of a strongly convex function ψ with effective domain $\mathcal{X}$; see Chapter B in the appendix for a brief review of the definition and basic properties of mirror operators.

We prove the existence and uniqueness of the solution to (8.11) in Theorem 15, and Section 8.4 is dedicated to proving the theorem. In Section 8.5, we prove, using the Lyapunov function V, that the solution trajectories are such that $f(X(t)) - f^\star = \mathcal{O}(1/t^2)$. In Section 8.6, we derive an equivalent formulation of the ODE: we show that the second equation is equivalent, in integral form, to $X(t) = \int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau / \int_0^t w(\tau)d\tau$, where $w(\tau) = \tau^{r-1}$, so that the primal variable X can be interpreted as a weighted average of the mirrored dual trajectory $\nabla\psi^*(Z(\tau))$, $\tau \in [0, t]$. Motivated by this averaging interpretation, we will study,

in the next chapter, the ODE with generalized averaging and give sufficient conditions on the weight function to achieve a given convergence rate. We close this chapter with a discussion of possible extensions to non-differentiable objective functions in Section 8.8.

8.2 Nemirovski's mirror descent and Nesterov's accelerated method

Proving convergence of the solution trajectories of an ODE often involves a Lyapunov argument. For example, to prove convergence of the solutions of the unconstrained gradient descent ODE, $\dot{X}(t) = -\nabla f(X(t))$, consider the candidate Lyapunov function $D(x^\star, X) = \frac{1}{2}\|X - x^\star\|_2^2$ for some minimizer $x^\star \in S$ (such a minimizer exists since S is supposed nonempty). Then the time derivative of $D(x^\star, X(t))$ along a solution trajectory $X(t)$ is given by

$$\begin{aligned}\frac{d}{dt}D(x^\star, X(t))(t) &= \left\langle \dot{X}(t), X(t) - x^\star \right\rangle \\ &= \langle -\nabla f(X(t)), X(t) - x^\star \rangle \\ &\leq -(f(X(t)) - f^\star),\end{aligned}$$

where the last inequality is by convexity of f. Integrating the inequality, we have $D(x^\star, X(t)) - D(x^\star, X(0)) \leq tf^\star - \int_0^t f(X(\tau))d\tau$, thus by Jensen's inequality, $f\left(\frac{1}{t}\int_0^t X(\tau)d\tau\right) - f^\star \leq \frac{1}{t}\int_0^t f(X(\tau))d\tau - f^\star \leq \frac{D(x^\star, X(0))}{t}$, which proves that $f\left(\frac{1}{t}\int_0^t X(\tau)d\tau\right)$ converges to the optimum at a $\mathcal{O}(1/t)$ rate. Additionally, if the set of minimizers S is compact, then we can prove that $X(t)$ converges to S. Indeed, let us define the distance to the set of minimizers, $D(S, x) = \inf_{x^\star \in S} D(x^\star, x)$ (this is a continuous function of x since S is compact). We have shown that $D(x^\star, X(t))$ is a nonincreasing function of t for all $x^\star \in S$. Since $t \mapsto D(S, X(t))$ is the pointwise infimum of non-negative, nonincreasing functions, it is also non-negative nonincreasing, therefore it has a limit as $t \to \infty$, and its limit is necessarily 0: By contradiction, suppose that its limit is strictly positive. Then there exists $d > 0$ and $T \geq 0$ such that for all $t \geq T$, $D(S, X(t)) > d$, and by continuity of f and $D(S, \cdot)$, $\delta \triangleq \inf_{\{x : D(S,x) > d\}} f(x) - f^* > 0$. Thus for all $t \geq T$, and for all $x^\star \in S$,

$$\frac{d}{dt}D(x^\star, X(t)) \leq f^\star - f(X(t)) \leq -\delta.$$

Integrating, we would have $D(x^\star, X(t)) \leq D(x^\star, X(T)) - (t - T)\delta$ for all $t \geq T$, which contradicts the fact that D is non-negative. This proves that $D(S, X(t))$ converges to 0.

Mirror descent ODE

The previous argument was extended by Nemirovski and Yudin in [98] to a family of methods called mirror descent, to solve general constrained convex optimization problems. The idea

is to start from a non-negative function, then to design dynamics for which that function is a Lyapunov function. Nemirovski and Yudin argue that one can replace the Lyapunov function $D(x^\star, X(t)) = \frac{1}{2}\|X(t) - x^\star\|_2^2$ (used in gradient descent) by a function defined on the dual space, $D_{\psi^*}(Z(t), z^\star)$, where $Z(t) \in E^*$ is a dual variable for which we will design the dynamics, and the corresponding trajectory in the primal space is $X(t) = \nabla\psi^*(Z(t))$ and $x^\star = \nabla\psi^*(z^\star)$. Here, E^* is the dual space, i.e. the space of linear functionals on E (in our case, since $E = \mathbb{R}^n$, E^* can also be identified with $\mathbb{R}^n$, but we make this distinction since, conceptually, the spaces E and E^* are different), and ψ^* is a convex function assumed to be finite and differentiable on all of E^*, and such that $\nabla\psi^*$ is a Lipschitz function that maps from E^* to $\mathcal{X}$. We will refer to ψ^* as the distance generating function on E^*, and to $\nabla\psi^*$ as the mirror operator. Such a function ψ^* can be obtained by taking the convex conjugate of a strongly convex function ψ with effective domain $\mathcal{X}$ (hence our choice of notation for ψ^*); See Chapter B in the appendix for a more detailed discussion on the duality properties of ψ and ψ^*, and the operator $\nabla\psi^*$.

The function $D_{\psi^*}(\cdot,\cdot)$ is the Bregman divergence associated with ψ^*, given as follows: for all $z, y \in E^*$,

$$D_{\psi^*}(z, y) = \psi^*(z) - \psi^*(y) - \langle \nabla\psi^*(y), z - y\rangle .$$

By definition of the Bregman divergence, we have

$$\begin{aligned}\frac{d}{dt}D_{\psi^*}(Z(t), z^\star) &= \frac{d}{dt}\left(\psi^*(Z(t)) - \psi^*(z^\star) - \langle \nabla\psi^*(z^\star), Z(t) - z^\star\rangle\right)\\ &= \left\langle \nabla\psi^*(Z(t)) - \nabla\psi^*(z^\star), \dot{Z}(t)\right\rangle\\ &= \left\langle X(t) - x^\star, \dot{Z}(t)\right\rangle .\end{aligned}$$

Therefore, if the dual variable Z obeys the dynamics $\dot{Z} = -\nabla f(X)$, then

$$\frac{d}{dt}D_{\psi^*}(Z(t), z^\star) = -\langle \nabla f(X(t)), X(t) - x^\star\rangle \le -(f(X(t)) - f^\star)$$

and by the same argument as in the gradient descent ODE, $D_{\psi^*}(Z(t), z^\star)$ is a Lyapunov function and $f\left(\frac{1}{t}\int_0^t X(\tau)d\tau\right) - f^\star$ converges to 0 at a $\mathcal{O}(1/t)$ rate. The mirror descent ODE system can be summarized by

$$\text{MD}\begin{cases} X = \nabla\psi^*(Z) \\ \dot{Z} = -\nabla f(X) \\ X(0) = x_0,\ Z(0) = z_0 \text{ with } \nabla\psi^*(z_0) = x_0 \end{cases} \tag{8.2}$$

This is illustrated in Figure 8.1.

Note that ODE (8.2) can be rewritten as $\dot{Z} = -\nabla f(\nabla\psi^*(Z))$, and since by assumption, ∇f and $\nabla\psi^*$ are Lipschitz functions, we can invoke the Cauchy-Lipschitz theorem (Theorem 2.5 in [132]) to prove existence and uniqueness of a solution $Z(t)$ defined on $[0, +\infty)$. In

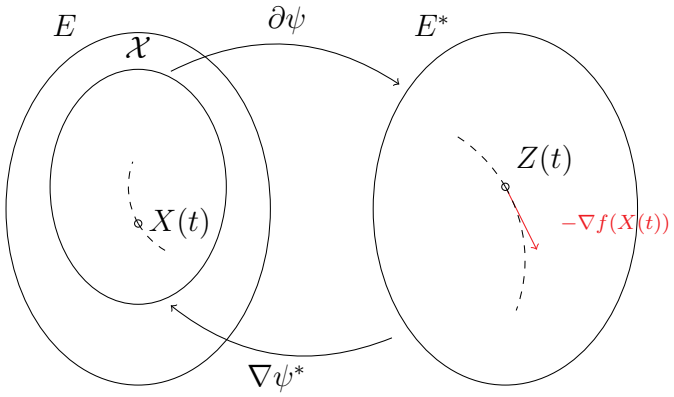


Figure 8.1: Illustration of the mirror descent ODE. The dual variable Z evolves in the (unconstrained) dual space E^*, and follows the flow of $-\nabla f(X(t))$. The primal trajectory $X(t)$ is obtained by applying the mirror operator $\nabla\psi^*$ to the dual trajectory $Z(t)$.

addition to the convergence of the time average $f(\frac{1}{t}\int_0^t X(\tau)d\tau)$, one can show the following stronger convergence result under the additional assumption that ψ^* is twice differentiable: Consider the energy function

$$V^{\mathrm{MD}}(X, Z, t) := t(f(X) - f^\star) + D_{\psi^*}(Z, z^\star). \tag{8.3}$$

and let $\mathcal{V}^{\mathrm{MD}}(t) := V^{\mathrm{MD}}(X(t), Z(t), t)$. Taking the time derivative of $\mathcal{V}^{\mathrm{MD}}(t)$, we have that

$$\begin{aligned}
\frac{d}{dt}\mathcal{V}^{\mathrm{MD}}(t) &= \frac{d}{dt}V^{\mathrm{MD}}(X(t), Z(t), t) \\
&= f(X(t)) - f^\star + t\left\langle \nabla f(X(t)), \dot{X}\right\rangle + \left\langle \dot{Z}(t), \nabla\psi^*(Z(t)) - \nabla\psi^*(z^\star)\right\rangle \\
&= f(X(t)) - f^\star - t\left\langle \nabla f(X(t)), \nabla^2\psi^*(Z(t))\nabla f(X(t))\right\rangle + \left\langle \dot{Z}(t), X(t) - x^\star\right\rangle \\
&\le f(X(t)) - f^\star - \langle \nabla f(X(t)), X(t) - x^\star\rangle \\
&\le 0,
\end{aligned}$$

where we used the fact that $\dot{X}(t) = \frac{d}{dt}\nabla\psi^*(Z(t)) = \nabla^2\psi^*(Z(t))\dot{Z}(t) = -\nabla^2\psi^*(Z(t))\nabla f(X(t))$ in the second equality (here $\nabla^2\psi^*(Z)$ is the Hessian of ψ^* at Z); the fact that $\nabla^2\psi^*(Z)$ is positive semi-definite (by convexity of ψ^*) in the third inequality; and convexity of f in the last inequality.

This proves that the energy $\mathcal{V}^{\mathrm{MD}}$ is a nonincreasing function of time, thus

$$f(X(t)) - f^\star \le \frac{\mathcal{V}^{\mathrm{MD}}(t)}{t} \le \frac{\mathcal{V}^{\mathrm{MD}}(0)}{t} = \frac{D_\psi(z_0, z^\star)}{t},$$

which proves that $f(X(t))$ converges to $f^\star$ at a $\mathcal{O}(1/t)$ rate.

Note that since $X = \nabla\psi^*(Z)$, and the mirror operator $\nabla\psi^*$ maps into $\mathcal{X}$ by assumption, the solution trajectory $X(t)$ remains in $\mathcal{X}$. Therefore, the mirror descent ODE is a natural generalization of gradient descent to constrained optimization problems: if one can construct a mirror operator $\nabla\psi^*$ which maps into $\mathcal{X}$, the solution is guaranteed to remain in $\mathcal{X}$. We also observe that the unconstrained gradient descent ODE can be obtained as a special case of the mirror descent ODE (8.2) by taking $\psi^*(z) = \frac{1}{2}\|z\|_2^2$, for which $\nabla\psi^*$ is the identity, in which case X and Z coincide.

The family of mirror descent methods can then be obtained by discretizing the ODE (8.2), and can be analyzed by using an analogous Lyapunov function in discrete time [98]. The mirror descent method is of particular importance in convex optimization, since the appropriate choice of Bregman divergence D_{ψ^*} can lead to improving the dependence of the convergence rate on the dimension of the space, see for example Chapter 3 in [98] and [24].

ODE interpretation of Nesterov's accelerated method

In [130], Su et al. show that Nesterov's accelerated method [102] can be interpreted as a discretization of a second-order differential equation, given by

$$\begin{cases} \ddot{X} + \dfrac{r+1}{t}\dot{X} + \nabla f(X) = 0, \\ X(0) = x_0,\ \dot{X}(0) = 0. \end{cases} \tag{8.4}$$

The analysis of the ODE uses the following candidate Lyapunov function (up to reparameterization)

$$\mathcal{V}^{\text{Nesterov}}(t) := \frac{t^2}{r^2}(f(X) - f^\star) + \frac{1}{2}\|X + \frac{t}{r}\dot{X} - x^\star\|^2,$$

which is proved to be a Lyapunov function for the ODE (8.4) whenever $r \geq 2$. This can be viewed by taking the time derivative of $\mathcal{V}^{\text{Nesterov}}(t)$ and plugging in the dynamics:

$$\begin{aligned} \frac{d}{dt}\mathcal{V}^{\text{Nesterov}}(t) &= \frac{2t}{r^2}(f(X) - f^\star) + \frac{t^2}{r^2}\left\langle \nabla f(X), \dot{X}\right\rangle + \left\langle X + \frac{t}{r}\dot{X} - x^\star, \dot{X}\frac{r+1}{r} + \frac{t}{r}\ddot{X}\right\rangle \\ &= \frac{2t}{r^2}(f(X) - f^\star) + \frac{t^2}{r^2}\left\langle \nabla f(X), \dot{X}\right\rangle + \left\langle X + \frac{t}{r}\dot{X} - x^\star, -\frac{t}{r}\nabla f(X)\right\rangle \\ &= \frac{2t}{r^2}(f(X) - f^\star) - \frac{t}{r}\left\langle X - x^\star, \nabla f(X)\right\rangle \\ &\leq \left(\frac{2t}{r^2} - \frac{t}{r}\right)(f(X) - f^\star), \end{aligned}$$

where we used convexity of f in the last inequality.

Since $\mathcal{V}^{\text{Nesterov}}$ is a non-increasing function of time, it follows that for all $t > 0$,

$$f(X(t)) - f^\star \leq \frac{r^2}{t^2}\mathcal{V}^{\text{Nesterov}}(t) \leq \frac{r^2}{t^2}\mathcal{V}^{\text{Nesterov}}(0) \leq \frac{r^2}{t^2}\frac{\|x_0 - x^\star\|^2}{2},$$

which proves that $f(X(t))$ converges to $f^\star$ at a $\mathcal{O}(1/t^2)$ rate.

One should note in particular that the dynamics is unconstrained, and the Euclidean distance is used in the definition of the Lyapunov function. As a consequence, discretizing the ODE (8.4) leads to a family of unconstrained, Euclidean accelerated methods. In the next section, we show that by combining elements from the Lyapunov analysis of Nesterov's accelerated method and Nemirovski's mirror descent, we can construct a much more general family of ODE systems which have the same $\mathcal{O}(1/t^2)$ convergence guarantee, and which apply to constrained optimization with non-Euclidean geometries.

8.3 Lyapunov design of the dynamics

Let $\|\cdot\|_*$ be a reference norm on the dual space E^*, and let ψ^* be a distance generating function on E^*, assumed to be L_{ψ^*}-smooth with respect to $\|\cdot\|_*$. Consider the function

$$V(X, Z, t) = \frac{t^2}{r^2}(f(X) - f^\star) + D_{\psi^*}(Z, z^\star) \tag{8.5}$$

where Z is a dual variable for which we will design the dynamics, and $z^\star$ is its value at equilibrium. Given a C^1 trajectory $(X(t), Z(t))$, let

$$\mathcal{V}(t) := V(X(t), Z(t), t).$$

This function combines the Bregman divergence term of the mirror descent Lyapunov function $\mathcal{V}^{\mathrm{MD}}$ (which encodes the constraint set) and the first term of the Nesterov Lyapunov function $\mathcal{V}^{\mathrm{Nesterov}}$ (which encodes the desired quadratic convergence rate). We will now design the dynamics of (X, Z) to make this candidate function a Lyapunov function. Taking the time-derivative of $\mathcal{V}(t)$, we have

$$\begin{aligned}\frac{d}{dt}\mathcal{V}(t) &= \frac{d}{dt}V(X(t), Z(t), t)\\ &= \frac{2t}{r^2}(f(X) - f^\star) + \frac{t^2}{r^2}\left\langle \nabla f(X), \dot{X}\right\rangle + \left\langle \dot{Z}, \nabla\psi^*(Z) - \nabla\psi^*(z^\star)\right\rangle\end{aligned}$$

Assume that $\dot{Z} = -\frac{t}{r}\nabla f(X)$. Then, the time-derivative becomes

$$\frac{d}{dt}\mathcal{V}(t) = \frac{2t}{r^2}(f(X) - f^\star) - \frac{t}{r}\left\langle \nabla f(X), -\frac{t}{r}\dot{X} + \nabla\psi^*(Z) - \nabla\psi^*(z^\star)\right\rangle.$$

Therefore, if X satisfies $X + \frac{t}{r}\dot{X} = \nabla\psi^*(Z)$, and $\nabla\psi^*(z^\star) = x^\star$, then,

$$\begin{aligned}\frac{d}{dt}\mathcal{V}(t) &= \frac{2t}{r^2}(f(X) - f^\star) - \frac{t}{r}\left\langle \nabla f(X), X - x^\star\right\rangle\\ &\le \frac{2t}{r^2}(f(X) - f^\star) - \frac{t}{r}(f(X) - f^\star)\\ &= -t\frac{r-2}{r^2}(f(X) - f^\star)\end{aligned} \tag{8.6}$$

and it follows that V is a Lyapunov function whenever $r \geq 2$. The proposed ODE system is then given by the system (8.11), copied below:

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X), \\ \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X), \\ X(0) = x_0,\ Z(0) = z_0, \text{ with } \nabla\psi^*(z_0) = x_0. \end{cases}$$

In the Euclidean case, taking $\psi^*(z) = \frac{1}{2}\|z\|_2^2$, we have $\nabla\psi^*(z) = z$, thus $Z = X + \frac{t}{r}\dot{X}$, and the ODE system is equivalent to $\frac{d}{dt}\left(X + \frac{t}{r}\dot{X}\right) = -\frac{t}{r}\nabla f(X)$, i.e. $\frac{t}{r}\ddot{X} + \frac{r+1}{r}\dot{X} + \frac{t}{r}\nabla f(X) = 0$, which is equivalent to the ODE (8.4) studied in [130], which we recover as a special case.

It is also important to observe that since $\nabla\psi^*$ maps into $\mathcal{X}$, then any primal solution $X(t)$ is viable (i.e. remains in the feasible set $\mathcal{X}$). Intuitively, since $\dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X)$, then $\dot{X}(t)$ always points inside the feasible set $\mathcal{X}$. In particular, whenever $X(t)$ is on the boundary of E, $\dot{X}(t)$ towards the interior of $\mathcal{X}$, thus guaranteeing that X remains in $\mathcal{X}$. This argument is made more precise in the proof of Theorem 15 in the next section.

8.4 Existence, uniqueness and viability of the solution

First, we prove existence and uniqueness of a solution to the ODE system (8.11), defined for all $t > 0$. By assumption, both ∇f and $\nabla\psi^*$ are Lipschitz-continuous functions. Unfortunately, due to the $\frac{r}{t}$ term in the expression of $\dot{X}$, the function $(X, Z, t) \mapsto (\dot{X}, \dot{Z})$ is not Lipschitz at $t = 0$. However, one can work around this by considering a sequence of approximating ODEs, similarly to the argument used in [130]. We observe that, alternatively, we could have instead initialized the ODE at a positive time t_0, which avoids the degeneracy at $t = 0$. This will be indeed our approach for proving existence and uniqueness of the solution for the more general ODE studied in Chapter 9, see Theorem 18. We present this more elaborate proof in this section merely to satisfy mathematical curiosity of the interested reader, since for practical purposes, it does not matter at which time the ODE is initialized.

Theorem 15. *Suppose f is C^1, ∇f is L_f-Lipschitz and $\nabla\psi^*$ is L_{ψ^*}-Lipschitz. Let $x_0 \in \mathcal{X}$ and $z_0 \in E^*$ such that $\nabla\psi^*(z_0) = x_0$. Then the accelerated mirror descent ODE system* (8.11) *with initial condition (x_0, z_0) has a unique maximal solution (X, Z) (i.e. defined on a maximal interval) in $C^1([0,\infty), \mathbb{R}^n)$. Furthermore, the primal solution X is viable, that is $X(t) \in \mathcal{X}$ for all $t \geq 0$.*

By a solution to (8.11), we mean a pair of functions (X, Z) that are C^1 on $[0, \infty)$, and which satisfy the differential equations for all $t > 0$. We first show existence and uniqueness

of a solution on any given interval $[0, T]$. Let $\delta > 0$, and consider the smoothed ODE system

$$\mathrm{AMD}^{\delta} \begin{cases} \dot{Z} = -\dfrac{t}{r}\nabla f(X), \\ \dot{X} = \dfrac{r}{\max(t,\delta)}(\nabla\psi^*(Z) - X), \\ X(0) = x_0,\ Z(0) = z_0 \text{ with } \nabla\psi^*(z_0) = x_0. \end{cases} \tag{8.7}$$

Since the functions $(X, Z) \mapsto -\frac{t}{r}\nabla f(X)$ and $(X, Z) \mapsto \frac{r}{\max(t,\delta)}(\nabla\psi^*(Z) - X)$ are Lipschitz for all $t \in [0, T]$, by the Cauchy-Lipschitz theorem (Theorem 2.5 in [132]), the system (8.7) has a unique solution (X_δ, Z_δ) in $C^1([0, T])$. In order to show the existence of a solution to the original ODE, we use the following property of the solution to the smoothed ODE. The proof of Lemma 10 is deferred to the end of the chapter.

Lemma 10. *Let $t_0 = \frac{2}{\sqrt{L_f L_{\psi^*}}}$. Then the family of solutions $\left((X_\delta, Z_\delta)|_{[0,t_0]}\right)_{\delta \le t_0}$ is equi-Lipschitz-continuous and uniformly bounded. More precisely,*

$$\|\dot{Z}_\delta(t)\| \le \frac{3t}{r}\|\nabla f(x_0)\|,$$
$$\|\dot{X}_\delta(t)\| \le \frac{(3+r)L_{\psi^*}t}{2}\|\nabla f(x_0)\|.$$

Proof of existence. Consider the family of solutions $((X_{\delta_i}, Z_{\delta_i}), \delta_i = t_0 2^{-i})_{i\in\mathbb{N}}$ restricted to $[0, t_0]$. By Lemma 10, this family is equi-Lipschitz-continuous and uniformly bounded, thus by the Arzelà-Ascoli theorem, there exists a subsequence $((X_{\delta_i}, Z_{\delta_i}))_{i\in\mathcal{I}}$ that converges uniformly on $[0, t_0]$. Let $(\bar{X}, \bar{Z})$ be its limit. Then we prove that $(\bar{X}, \bar{Z})$ is a solution to the original ODE (8.11) on $[0, t_0]$.

First, since for all $i \in \mathcal{I}$, $X_{\delta_i}(0) = x_0$ and $Z_{\delta_i}(0) = z_0$, it follows that

$$\bar{X}(0) = \lim_{i\to\infty, i\in\mathcal{I}} X_{\delta_i}(0) = x_0,$$
$$\bar{Z}(0) = \lim_{i\to\infty, i\in\mathcal{I}} Z_{\delta_i}(0) = z_0,$$

thus $(\bar{X}, \bar{Z})$ satisfies the initial conditions. Next, let $t_1 \in (0, t_0)$, and let $(\tilde{X}, \tilde{Z})$ be the solution of the ODE (8.11) on $t \ge t_1$, with initial condition $(\bar{X}(t_1), \bar{Z}(t_1))$. Since $(X_{\delta_i}(t_1), Z_{\delta_i}(t_1))_{i\in\mathcal{I}} \to (\bar{X}(t_1), \bar{Z}(t_1))$ as $i \to \infty$, then by continuity of the solution w.r.t. initial conditions, we have that for some $\epsilon > 0$, $X_{\delta_i} \to \tilde{X}$ uniformly on $[t_1, t_1 + \epsilon)$. But we also have $X_{\delta_i} \to \bar{X}$ uniformly on $[0, t_0]$, therefore $\bar{X}$ and $\tilde{X}$ coincide on $[t_1, t_1 + \epsilon)$, therefore $\bar{X}$ satisfies the ODE on $[t_1, t_1 + \epsilon)$. And since t_1 is arbitrary in $(0, t_0)$, this concludes the proof of existence. □

Proof of uniqueness. It suffices to prove uniqueness on an open neighborhood of 0, since away from 0, uniqueness is guaranteed by the Cauchy-Lipschitz theorem.

Let (X, Z) and $(\bar{X}, \bar{Z})$ be two solutions of the ODE (8.11), and let $\Delta_Z = Z - \bar{Z}$ and $\Delta_X = X - \bar{X}$. Then Δ_X, Δ_Z are C^1, and we have

$$\begin{cases} \dot{\Delta}_Z = -\dfrac{t}{r}\left(\nabla f(X) - \nabla f(\bar{X})\right) \\ \dot{\Delta}_X = \dfrac{r}{t}\left(\nabla \psi^*(Z) - \nabla \psi^*(\bar{Z}) - \Delta_X\right) \\ \Delta_Z(0) = \Delta_X(0) = 0 \end{cases}$$

Let $A(t) = \sup_{u\in(0,t]} \frac{\|\dot{\Delta}_Z(u)\|}{u}$, and $B(t) = \sup_{u\in[0,t]} \|\Delta_X(u)\|$. Note that $B(t)$ is finite since Δ_X is continuous on $[0, t]$. The finiteness of $A(t)$ will be established below. We have

$$\|\dot{\Delta}_Z(t)\| = \frac{t}{r}\|\nabla f(X(t)) - \nabla f(\bar{X}(t))\| \leq \frac{L_f t}{r}\|\Delta_X(t)\| \leq \frac{L_f t}{r} B(t).$$

Dividing by t and taking the supremum, we have

$$A(t) \leq \frac{L_f}{r} B(t). \tag{8.8}$$

Next, since $t^r \dot{\Delta}_X + r t^{r-1}\Delta_X = r t^{r-1}\left(\nabla \psi^*(Z) - \nabla \psi^*(\bar{Z})\right)$, we have

$$\frac{d}{dt}(t^r \Delta_X) = r t^{r-1}\left(\nabla \psi^*(Z) - \nabla \psi^*(\bar{Z})\right).$$

Therefore, integrating and taking norms

$$\begin{aligned} t^r\|\Delta_X(t)\| &\leq \int_0^t r\tau^{r-1}\|\nabla \psi^*(Z(\tau)) - \nabla \psi^*(\bar{Z}(\tau))\| d\tau \\ &\leq r t^{r-1}\int_0^t L_{\psi^*}\|\Delta_Z(\tau)\| d\tau \\ &\leq L_{\psi^*} r t^{r-1} A(t)\int_0^t \frac{\tau^2}{2} d\tau \\ &= \frac{L_{\psi^*} r t^{r-1} t^3 A(t)}{6}, \end{aligned}$$

where we used the fact that $\|\Delta_Z(\tau)\| = \|\int_0^\tau \dot{\Delta}_Z(u) du\| \leq \int_0^\tau u A(t) du = A(t)\frac{\tau^2}{2}$. Dividing by t^r and taking the supremum,

$$B(t) \leq \frac{L_{\psi^*} r t^2}{6} A(t). \tag{8.9}$$

Combining (8.8) and (8.9), we have $A(t) \leq \frac{L_f L_{\psi^*} t^2}{6} A(t)$. It follows that $A(t) = 0$ for $0 \leq t < \sqrt{\frac{6}{L_f L_{\psi^*}}}$, which in turn implies that $B(t) = 0$ on the same interval. This concludes the proof. □

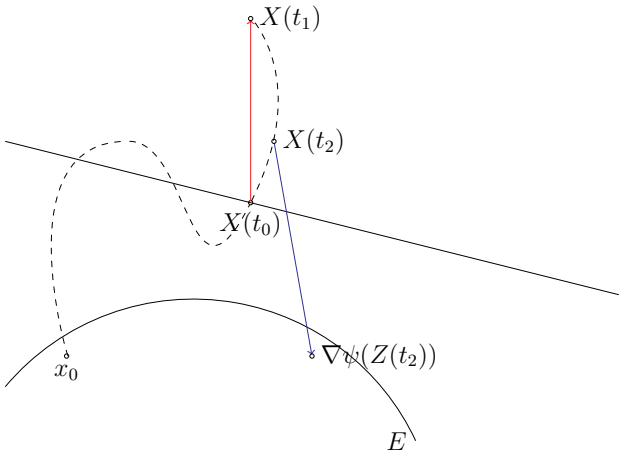


Figure 8.2: Illustration of the proof of viability.

Proof of viability. We now prove that the primal solution X remains in $\mathcal{X}$ for all t. Intuitively, since $\dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X)$, the derivative $\dot{X}$ will point towards $\mathcal{X}$, keeping $X(t)$ inside the feasible set.

Suppose by contradiction that there exists $t_1 > 0$ such that $x_1 = X(t_1) \notin \mathcal{X}$. Let $D = \sup_{t\in[0,t_1]} \|\nabla\psi^*(Z(t))\|$ (finite by continuity of the solution), and consider the restriction of the feasible set to the ball of radius D, $\bar{\mathcal{X}} = \mathcal{X} \cap \{x : \|x\| \le D\}$. Then $\bar{\mathcal{X}}$ is convex and compact and does not contain x_1, so by the separation theorem, there exists a hyperplane that strictly separates x_1 and $\bar{\mathcal{X}}$. That is, there exists an affine functional $\ell(\cdot) = \langle u, \cdot\rangle - \alpha$, $u \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, such that $\ell(x_1) > 0$ and $\ell(x) < 0$ for all $x \in \bar{\mathcal{X}}$. Since the solution trajectory $X(t)$ is C^1, $t \mapsto \ell(X(t))$ is also C^1, and its time-derivative is $\dot{\ell}(X(t)) = \frac{d}{dt}\langle u, X(t)\rangle - \alpha = \left\langle u, \dot{X}(t)\right\rangle$.

We have $\ell(X(0)) < 0$ (since $x_0 \in \bar{\mathcal{X}}$) and $\ell(X(t_1)) > 0$, thus there exists t_0 such that $\ell(X(t_0)) = 0$ and $\ell(X(t)) > 0$ for all $t \in (t_0, t_1]$, that is, t_0 is the last time $X(t)$ crosses the separating hyperplane (t_0 is simply $\sup\{t \le t_1 : \ell(X(t)) \le 0\}$). Then by definition, $\ell(X(t_1)) - \ell(X(t_0)) > 0$, but by the mean value theorem, there exists $t_2 \in [t_0, t_1]$ such that

$$\begin{aligned}\frac{\ell(X(t_1)) - \ell(X(t_0))}{t_1 - t_0} &= \dot{\ell}(X(t_2)) = \left\langle u, \dot{X}(t_2)\right\rangle \\ &= \frac{r}{t}\langle u, \nabla\psi^*(Z(t_2)) - X(t_2)\rangle \\ &= \frac{r}{t}\left(\ell(\nabla\psi^*(Z(t_2))) - \ell(X(t_2))\right) < 0\end{aligned}$$

since $\nabla\psi^*(Z(t_2)) \in \bar{\mathcal{X}}$. This is a contradiction, which concludes the proof. □

8.5 Convergence rate

Now that we have proved the existence and uniqueness of the solution, it becomes straightforward to establish the convergence rate of the function values, using the Lyapunov function which motivated the dynamics.

Theorem 16. *Suppose that ∇f and $\nabla \psi^*$ are Lipschitz. Let $(X(t), Z(t))$ be the solution to the accelerated mirror descent ODE* (8.11) *with $r \geq 2$. Then for all $t > 0$,*

$$f(X(t)) - f^\star \leq \frac{r^2 D_{\psi^*}(z_0, z^\star)}{t^2} \tag{8.10}$$

Furthermore, if $r > 2$, then $\int_0^\infty t(f(X(t)) - f^\star)dt \leq \frac{r^2}{r-2} D_{\psi^*}(z_0, z^\star)$.

Proof. By construction of the ODE, we have $V(X(t), Z(t), t) = \frac{t^2}{r^2}(f(X(t)) - f^\star) + D_{\psi^*}(Z(t), z^\star)$ is a Lyapunov function. It follows that for all $t > 0$,

$$\frac{t^2}{r^2}(f(X(t)) - f^\star) \leq V(X(t), Z(t), t) \leq V(x_0, z_0, 0) = D_{\psi^*}(z_0, z^\star),$$

which proves the first inequality. Furthermore, we have that

$$\frac{d}{dt} V(X(t), Z(t), t) \leq -\frac{r-2}{r^2} t(f(X(t)) - f^\star),$$

thus, integrating from 0 to T and rearranging, we have

$$\int_0^T t(f(X(t)) - f^\star)dt \leq \frac{r^2}{r-2} V(x_0, z_0, 0) = \frac{r^2}{r-2} D_{\psi^*}(z_0, z^\star),$$

which proves the second part of the claim. □

Remark 2. *The second part of the theorem indicates that the convergence rate is in fact better than $\Omega(1/t^2)$. Indeed, if $f(X(t)) - f^\star \geq \frac{c}{t^2}$ for some positive constant c, then $\int_1^T t(f(X(t)) - f^\star)dt \geq c \ln T$, which would contradict the theorem. We also observe that, although it seems from the bound* (8.10) *that smaller values of the parameter r are better, the upper bound on the integral diverges as r approaches 2, which indicates that smaller values of r are not necessarily better. In Section 8.7, we will give another interpretation of the parameter r as a damping coefficient, and we will further discuss its effect on convergence.*

Remark 3 (On scaling time)**.** *Note that in continuous time, a faster convergence rate can be obtained by rescaling time. In other words, if $X(t)$ converges to the set of minimizers at the rate $r_1(t)$ in the sense that $f(X(t)) - f^\star = \mathcal{O}(1/r_1(t))$ (where r_1 is an increasing function on $\mathbb{R}_+$), then given am increasing function $h : \mathbb{R}_+ \to \mathbb{R}_+$ such that $h(t) \geq t$, the trajectory $\tilde{X}(t) := X(h(t))$ satisfies $f(\tilde{X}(t)) - f^\star = \mathcal{O}(1/r_2(t))$ where $r_2 = r_1 \circ h$ is a faster rate (i.e. $r_2(t) \geq r_1(t)$ for all t). Of course the spatial trajectories $\{X(t), t \in \mathbb{R}_+\}$ and $\{\tilde{X}(t), t \in \mathbb{R}_+\}$*

coincide, but rescaling time seems to lead to faster convergence. Such a transformation is not possible in discrete time, as scaling time by a superlinear function would correspond to scaling up the step sizes, which would eventually violate the upper bounds on step sizes. Thus convergence rates do not have the same interpretation in continuous and discrete time. Since the quadratic rate of convergence is the optimal rate for first-order methods for convex optimization (according the to the lower bounds derived by Nemirovski and Yudin [98]), it is natural to consider continuous dynamics with a time scale that gives a quadratic convergence rate, and to seek a discretization which preserves this rate.

8.6 Averaging interpretation

Starting from the equation $\dot{X} = \frac{r}{t}(\nabla\psi^*(Z(t)) - X(t))$, we can multiply both sides by $\frac{t^r}{r}$ and rearrange to obtain $\frac{t^r}{r}\dot{X}(t) + t^{r-1}X(t) = t^{r-1}\nabla\psi^*(Z(t))$. Integrating from 0 to t, and observing that $\frac{t^r}{r}\dot{X}(t) + t^{r-1}X(t)$ is the time derivative of $\frac{t^r}{r}X(t)$, we have

$$\frac{t^r}{r}X(t) = \int_0^t \tau^{r-1}\nabla\psi^*(Z(\tau))d\tau.$$

Finally, dividing by $\frac{t^r}{r}$, we have

$$X(t) = \frac{r}{t^r}\int_0^t \tau^{r-1}\nabla\psi^*(Z(\tau))d\tau = \frac{\int_0^t \tau^{r-1}\nabla\psi^*(Z(\tau))d\tau}{\int_0^t \tau^{r-1}d\tau}.$$

Therefore the primal variable $X(t)$ can be interpreted as a weighted average of the trajectory $\nabla\psi^*(Z(\tau))$, $\tau \in [0,t]$, with time-varying weights $w(\tau) = \tau^{r-1}$. This interpretation formalizes a connection between acceleration and averaging, as observed in [50] for the unconstrained quadratic case. This also provides an intuitive interpretation of the parameter r: it controls the weights in the expression of X. A higher value of r puts larger weights on the recent points $\nabla\psi^*(Z(t))$.

The accelerated mirror descent ODE can then be written in the equivalent form:

$$\begin{cases} \dot{Z} = -\eta(t)\nabla f\left(X(t)\right), \ \eta(t) = \dfrac{t}{r} \\ X(t) = \dfrac{\int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau}{\int_0^t w(\tau)d\tau}, \ w(\tau) = \tau^{r-1} \\ X(0) = x_0, \ Z(0) = z_0 \text{ with } \nabla\psi^*(z_0) = x_0 \end{cases} \tag{8.11}$$

Here Z is a dual variable which accumulates the negative gradient of f, at a rate $\eta(t) = \frac{t}{r}$, and X is a weighted average of the "mirrored" dual trajectory $\nabla\psi^*(Z(\tau))$, $\tau \in [0,t]$, with weight function $w(\tau) = \tau^{r-1}$. We also note that since $\nabla\psi^*(Z(\tau))$ remains in $\mathcal{X}$ for all τ, so does X, by convexity of the feasible set $\mathcal{X}$. This provides an alternate, simple proof of the viability of the solution (last part of Theorem 15).

8.7 Damped nonlinear oscillator interpretation

In this section, we will assume that ψ^* is twice differentiable on all of E^*, and we will denote its Hessian at a point $z \in E^*$ by $\nabla^2\psi^*(z)$, defined as $\nabla^2\psi^*(z)_{i,j} = \frac{\partial^2\psi^*(z)}{\partial z_i \partial z_j}$. This assumption is not particularly restrictive, see Chapter B in the Appendix for examples. Writing $\frac{t}{r}\dot{X} + X = \nabla\psi^*(Z)$ and taking the time-derivative, we have

$$\frac{t}{r}\ddot{X} + \frac{1}{r}\dot{X} + \dot{X} = \nabla^2\psi^*(Z)\dot{Z} = -\frac{t}{r}\nabla^2\psi^*(Z)\nabla f(X).$$

Multiplying both sides by $\frac{r}{t}$, we have

$$\ddot{X} + \frac{r+1}{t}\dot{X} + \nabla^2\psi^*(Z)\nabla f(X) = 0. \tag{8.12}$$

The initial condition for $\dot{X}$ is $\dot{X}(0) = 0$. To prove this, one can argue that for all $\delta > 0$, the solution to the smoothed ODE (8.7) satisfies $\dot{X}_\delta(0) = \frac{r}{\delta}(\nabla\psi^*(z_0) - x_0) = 0$, thus $\dot{X}(0)$ is also equal to zero since the solution X is a limit point of the equi-Lipschitz family of solutions (X_δ).

The ODE (8.12) can be interpreted as a generalization of a damped nonlinear oscillator: In the unconstrained Euclidean case, we can take $\psi^*(z) = \frac{1}{2}\|z\|_2^2$, in which case $\nabla^2\psi^*(Z)$ is the identity, then the ODE becomes $\ddot{X} + \frac{r+1}{t}\dot{X} + \nabla f(X) = 0$, and we recover ODE 8.4 studied in [130]. It can be interpreted as describing the evolution of a particle with position X, velocity $\dot{X}$ and acceleration $\ddot{X} = -\nabla f(X) - \frac{r+1}{t}\dot{X}$. The first term is a conservative force due to the scalar potential f, and the second term is a dissipative force proportional to the velocity, which can be thought of as a viscous friction term. Some properties of this system have been recently studied in [6]. Note that the damping constant $\frac{r+1}{t}$ is time-dependent, and vanishes as time tends to infinity. The parameter r can then be interpreted as a damping coefficient. Intuitively, the larger r, the more energy is dissipated. This is illustrated in Figure 8.3, which shows the solution trajectory of the ODE on a finite time interval, in a simplex-constrained example, with different values of r. In this case, a natural measure of the energy of the system is given by the mechanical energy, the sum of the potential energy $f(X)$ and the kinetic energy $\frac{1}{2}\|\dot{X}\|_2^2$,

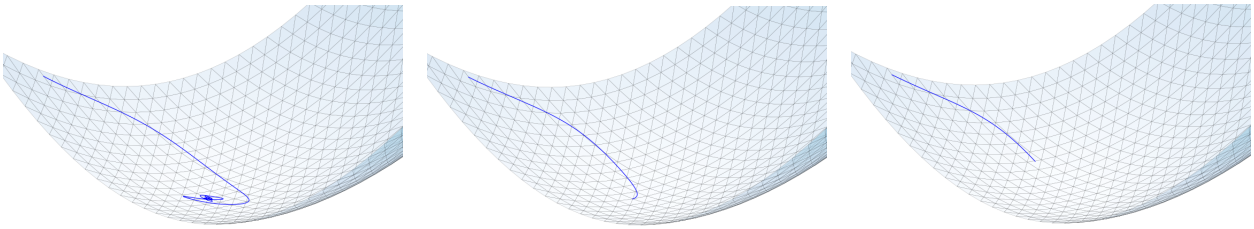
$$\mathcal{E}(t) = f(X(t)) + \frac{1}{2}\|\dot{X}(t)\|_2^2. \tag{8.13}$$

Taking the time-derivative of the energy, we have

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(t) &= \left\langle \nabla f(X(t)) + \ddot{X}(t), \dot{X}(t) \right\rangle \\ &= \left\langle -\frac{r+1}{t}\dot{X}(t), \dot{X}(t) \right\rangle \\ &= -\frac{r+1}{t}\|\dot{X}(t)\|_2^2\end{aligned}$$

which proves that the energy is non-increasing (and strictly decreasing as long as the particle is not at rest), as expected due to the presence of the dissipative friction term. In the next chapter, we will define a generalization of this energy function that is suited to the constrained case, see Section 9.4.

In the constrained case, the Hessian term $\nabla^2\psi^*(Z)$ which appears in ODE (8.12) is a non-linear transformation that applies to the gradient, in order to keep the trajectory in the feasible set. Remarkably, this transformation is not static, it depends on the value of the dual variable, hence varies with time. Intuitively, whenever $\nabla\psi^*(Z)$ approaches the (relative) boundary of the feasible set, the term $\nabla^2\psi^*(Z)$ should transform the gradient so that it points inside the feasible set. The role of the Hessian term will be further discussed in Section 9.5 in the next chapter.



(a) $r = 2$ (b) $r = 20$ (c) $r = 200$

Figure 8.3: Solution trajectories of the accelerated mirror descent ODE on a finite time interval $t \in [0, T]$, for simplex-constrained quadratic minimization, with different values of the parameter r. Larger values of r result in more energy dissipation, and suppress oscillations, but because the time-horizon is finite, too much energy dissipation means that the trajectory does not make enough progress within $[0, T]$, as can be seen in plot (c). This example shows that the "best damping" is not necessarily obtained for smaller values of r, as one could think from the bound of Theorem 16.

8.8 On extending the dynamics to non-differentiable objective functions

In this section, we consider the case in which the objective function is non-differentiable. One such case of particular interest is composite optimization, in which the objective function can be decomposed into the sum of two terms $f = f_1 + f_2$ where f_1 is differentiable with Lipschitz gradient, and f_2 is a general convex function; this model covers many problems in machine learning, such as ℓ_1-regularized regression, and many algorithms have been developed for composite optimization in discrete time, for example [100], as well as continuous time, for example [7]. In this section, we discuss how the Lyapunov argument can be extended to

non-differentiable functions. More precisely, assume that f is a closed and proper convex function (not necessarily differentiable), and denote by $\partial f(x)$ the subdifferential of f at x (a closed and convex set). A natural way to extend the ODE (8.11) to this non-differentiable case is to replace the dual differential equation $\dot{Z}(t) = -\frac{t}{r}\nabla f(X(t))$ by the differential inclusion $\dot{Z}(t) \in -\frac{t}{r}\partial f(X(t))$. As we will see, this may not suffice to guarantee that the energy function V decreases along continuous solution trajectories. As observed by [130], the directional derivative $f'(X; \dot{X})$ plays a central role in deriving the correct dynamics in the non-differentiable case.

The (one-sided) directional derivative of f at x in the direction y is defined by

$$f'(x; y) = \lim_{\epsilon \to 0,\ \epsilon > 0} \frac{f(x + \epsilon y) - f(x)}{\epsilon},$$

where the limit can be $+\infty$. It exists at any point x in the domain of f (i.e. where f is finite), and is a positively homogeneous convex function of y, see Theorem 23.1 in [117]. Additionally, we have the following connection between the directional derivative and the subdifferential: By Theorem 23.4 in [117] we have that for all x in the interior of the domain of f (denoted $\operatorname{int} \operatorname{dom} f$), $\partial f(x)$ is a non-empty compact set, and

$$f'(x; y) = \sup_{g \in \partial f(x)} \langle g, y \rangle . \tag{8.14}$$

Thus we can associate to $f'(x; y)$ the set of subgradients which achieve the maximum (the supremum is attained since $\partial f(x)$ is a compact set in this case). We will denote this set

$$d(x; y) = \arg\max_{g \in \partial f(x)} \langle g, y \rangle .$$

Theorem 17. *Consider the energy function $t \mapsto V(X(t), Z(t), t) = \frac{t^2}{r^2}(f(X(t)) - f^\star) + D_{\psi^*}(Z(t), z^\star)$ where f is a proper closed convex function, and suppose that $(X(t), Z(t))$ is a continuous and differentiable solution trajectory of the ODE*

$$\begin{cases} \dot{Z} \in -\dfrac{t}{r} d(X, \dot{X}) \\ \dot{X} = \dfrac{r}{t}(\nabla \psi^*(Z) - X), \end{cases}$$

such that $X(t)$ remains in the relative interior of the domain of f. Then the energy function is differentiable and $\frac{d}{dt} V(X(t), Z(t), t) \leq 0$.

Since the energy function is decreasing, any continuous and differentiable solution will satisfy $f(X(t)) - f^\star = \mathcal{O}(1/t^2)$ by a similar argument to Theorem 16. Note however that we do not discuss existence of such solutions in this case.

Proof. To prove that the energy function is differentiable, consider the difference quotient, defined for $\epsilon > 0$,

$$\begin{aligned}\Delta_t(\epsilon) &= \frac{V(t+\epsilon) - V(t)}{\epsilon} \\ &= \frac{t^2}{r^2}\frac{f(X(t+\epsilon)) - f(X(t))}{\epsilon} + \frac{2t+\epsilon}{r^2}(f(X(t+\epsilon)) - f^\star) + \frac{D_{\psi^*}(Z(t+\epsilon), z^\star) - D_{\psi^*}(Z(t), z^\star)}{\epsilon}.\end{aligned}$$

Using the fact that a convex function is locally Lipschitz on the relative interior of its domain (so that $f(x + o(\epsilon)) = f(x) + o(\epsilon)$), and that $D_{\psi^*}(Z(t), z^\star)$ is differentiable, we have

$$\begin{aligned}\Delta_t(\epsilon) &= \frac{t^2}{r^2}\frac{f(X(t) + \epsilon\dot{X}(t)) + o(\epsilon) - f(X(t))}{\epsilon} + \frac{2t+\epsilon}{r^2}(f(X(t)) + o(1) - f^\star) + \frac{d}{dt}D_{\psi^*}(Z(t), z^\star) + o(1) \\ &= \frac{t^2}{r^2}\frac{f(X(t) + \epsilon\dot{X}(t)) - f(X(t))}{\epsilon} + \frac{2t}{r^2}(f(X(t)) - f^\star) + \frac{d}{dt}D_{\psi^*}(Z(t), z^\star) + o(1). \qquad (8.15)\end{aligned}$$

The derivative of the Bregman divergence in (8.15) is

$$\frac{d}{dt}D_{\psi^*}(Z(t), z^\star) = \left\langle \dot{Z}(t), \nabla\psi^*(Z(t)) - \nabla\psi^*(z^\star) \right\rangle = \left\langle \dot{Z}(t), X(t) + \frac{t}{r}\dot{X}(t) - x^\star \right\rangle.$$

The first term in (8.15) converges, as $\epsilon \to 0$, to $f'(X; \dot{X})$. Combining the two limits, we have that the limit of $\Delta_t(\epsilon)$ exists and

$$\lim_{\epsilon\to 0,\ \epsilon>0} \Delta_t(\epsilon) = \frac{t^2}{r^2}f'(X(t); \dot{X}(t)) + \frac{2t}{r^2}(f(X(t)) - f^\star) + \left\langle \dot{Z}(t), X(t) + \frac{t}{r}\dot{X}(t) - x^\star \right\rangle,$$

and if we let $\dot{Z}(t) = -\frac{t}{r}g(t)$, then

$$\lim_{\epsilon\to 0,\ \epsilon>0} \Delta_t(\epsilon) \le \frac{t^2}{r^2}\left(f'(X; \dot{X}) - \left\langle g, \dot{X} \right\rangle\right) + \frac{t}{r}(f(X) - f^\star - \langle g, X - x^\star\rangle), \qquad (8.16)$$

where we used the assumption that $r \ge 2$. Note that if $\dot{Z}$ satisfies the differential inclusion $\dot{Z}(t) \in -\frac{t}{r}\partial f(X(t))$ (in other words, $g(t)$ is a subgradient of f at $X(t)$), then the second term in inequality (8.16) is non-positive by definition of a subgradient, but the first term $f'(X; \dot{X}) - \left\langle g, \dot{X} \right\rangle$ is non-negative by (8.14), and one cannot conclude that the energy is decreasing. This motivates our choice of the subgradient. Indeed, when $\dot{Z}(t) \in -\frac{t}{r}d(X; \dot{X})$ (in other words, $g(t)$ is a subgradient of f at $X(t)$ *that maximizes the linear functional* $\langle \cdot, \dot{X}(t)\rangle$), the first term in inequality (8.16) is non-positive, therefore $\lim_{\epsilon\to 0,\ \epsilon>0} \Delta_t(\epsilon) \le 0$, which concludes the proof. □

Proof of Lemma 10

Let us rewrite the smoothed accelerated mirror descent ODE system

$$\text{AMD}^\delta \begin{cases} \dot{Z} = -\dfrac{t}{r}\nabla f(X) \\ \dot{X} = \dfrac{r}{\max(t,\delta)}(\nabla\psi^*(Z) - X) \\ X(0) = x_0,\ Z(0) = z_0 \text{ with } \nabla\psi^*(z_0) = x_0. \end{cases}$$

By the Cauchy-Lipschitz theorem, there exists a unique solution (X_δ, Z_δ) defined on $[0,\infty)$, and the solution is C^1. Define, for $t > 0$,

$$\begin{aligned} A_\delta(t) &= \sup_{u\in[0,t]} \frac{\|\dot{Z}_\delta(u)\|}{u} \\ B_\delta(t) &= \sup_{u\in[0,t]} \frac{\|X_\delta(u) - x_0\|}{u} \\ C_\delta(t) &= \sup_{u\in[0,t]} \|\dot{X}_\delta(u)\| \end{aligned}$$

These quantities are finite for the following reasons:

- $\frac{\|X_\delta(u)-x_0\|}{u} = \|\dot{X}_\delta(0)\| + o(1)$ near 0, thus B_δ is finite.
- $\|\dot{X}_\delta\|$ is continuous thus bounded on $[0,t]$, thus C_δ is finite.
- Finiteness of A_δ is a consequence of the following lemma.

To prove Lemma 10, we first need the auxiliary lemma below, that provides bounds on $A_\delta, B_\delta, C_\delta$.

Lemma 11. *For all t,*

$$rA_\delta(t) \le \|\nabla f(x_0)\| + L_f t B_\delta(t), \tag{8.17}$$

$$B_\delta(t) \le \frac{L_{\psi^*} r t}{6} A_\delta(t), \tag{8.18}$$

$$C_\delta(t) \le r\left(\frac{tL_{\psi^*}}{2}A_\delta(t) + B_\delta(t)\right). \tag{8.19}$$

Proof. By definition of A_δ and B_δ, we have

$$\begin{aligned} &\|Z_\delta(t) - z_0\| \le \int_0^t \|\dot{Z}_\delta(v)\|dv \le A_\delta(t)\int_0^t v dv = \frac{t^2}{2}A_\delta(t), \qquad (8.20) \\ &\|X_\delta(t) - x_0\| \le tB_\delta(t). \end{aligned}$$

Now, from the first equation in (8.7), we have for all $0 < t \leq t_0$

$$\begin{aligned}\frac{r}{t}\|\dot{Z}_\delta(t)\| &= \|\nabla f(X_\delta(t))\| \\ &\leq \|\nabla f(x_0)\| + \|\nabla f(X_\delta(t)) - \nabla f(x_0)\| \\ &\leq \|\nabla f(x_0)\| + L_f\|X_\delta(t) - x_0\| \qquad \nabla f \text{ is } L_f\text{-Lipschitz} \\ &\leq \|\nabla f(x_0)\| + L_f t B_\delta(t).\end{aligned}$$

Thus,

$$rA_\delta(t) \leq \|\nabla f(x_0)\| + L_f t B_\delta(t).$$

To prove inequality (8.18), we show that $\|X_\delta(t) - x_0\| \leq \frac{r}{\max(\delta,t)}\int_0^t \|\nabla\psi^*(Z_\delta(s)) - \nabla\psi^*(z_0)\|ds$. We consider the two cases $t \leq \delta$ and $t \geq \delta$.

- Let $t \leq \delta$. From the second equation in (8.7), we have

$$e^{\frac{rt}{\delta}}\left(\dot{X}_\delta + \frac{r}{\delta}(X_\delta - x_0)\right) = \frac{r}{\delta}e^{\frac{rt}{\delta}}(\nabla\psi^*(Z_\delta) - \nabla\psi^*(z_0)),$$

  i.e.,

$$\frac{d}{dt}\left((X_\delta(t) - x_0)e^{\frac{rt}{\delta}}\right) = \frac{r}{\delta}e^{\frac{rt}{\delta}}(\nabla\psi^*(Z_\delta(t)) - \nabla\psi^*(z_0)),$$

  thus integrating

$$(X_\delta(t) - x_0)e^{\frac{rt}{\delta}} = \frac{r}{\delta}\int_0^t e^{\frac{rs}{\delta}}(\nabla\psi^*(Z_\delta(s)) - \nabla\psi^*(z_0))ds$$

  dividing by $e^{\frac{rt}{\delta}}$ and taking norms we obtain the desired inequality.

- Let $t \geq \delta$. From the second equation in (8.7), we have

$$t^r\left(\dot{X}_\delta + \frac{r}{t}(X_\delta - x_0)\right) = rt^{r-1}(\nabla\psi^*(Z_\delta) - \nabla\psi^*(z_0)),$$

  i.e.

$$\frac{d}{dt}\left(t^r(X_\delta(t) - x_0)\right) = rt^{r-1}(\nabla\psi^*(Z_\delta) - \nabla\psi^*(z_0)),$$

  thus integrating

$$t^r(X_\delta(t) - x_0) = \int_0^t rs^{r-1}(\nabla\psi^*(Z_\delta(s)) - \nabla\psi^*(z_0))ds$$

  dividing by t^r and taking norms, we obtain the desired inequality.

Now we have

$$\begin{aligned}
\|X_\delta(t) - x_0\| &\le \frac{r}{\max(\delta, t)} \int_0^t \|\nabla\psi^*(Z_\delta(s)) - \nabla\psi^*(z_0)\| ds \\
&\le \frac{L_{\psi^*} r}{\max(\delta, t)} \int_0^t \|Z_\delta(s) - z_0\| ds && \nabla\psi^* \text{ is } L_{\psi^*}\text{-Lipschitz} \\
&\le \frac{L_{\psi^*} r}{\max(\delta, t)} \int_0^t \frac{s^2}{2} A_\delta(t) ds && \text{by (8.20)} \\
&= \frac{L_{\psi^*} r}{\max(\delta, t)} A_\delta(t) \frac{t^3}{6} \\
&\le \frac{L_{\psi^*} r t^2 A_\delta(t)}{6}.
\end{aligned}$$

Dividing by t and taking the supremum, we have (8.18).

Finally, to bound C_δ, we have from the second equation in (8.7), for all $0 < t \le t_0$,

$$\begin{aligned}
\|\dot{X}_\delta(t)\| &= \frac{r}{\max(\delta, t)} \|\nabla\psi^*(Z_\delta(t)) - X_\delta(t)\| \\
&\le \frac{r}{\max(\delta, t)} (\|\nabla\psi^*(Z_\delta(t)) - \nabla\psi^*(z_0)\| + \|X_\delta(t) - x_0\|) \\
&\le \frac{r}{\max(\delta, t)} (L_{\psi^*} \|Z_\delta(t) - z_0\| + \|X_\delta(t) - x_0\|) \\
&\le \frac{r}{\max(\delta, t)} \left(\frac{t^2}{2} L_{\psi^*} A_\delta(t) + t B_\delta(t) \right) \\
&\le r \left(\frac{L_{\psi^*} t}{2} A_\delta(t) + B_\delta(t) \right),
\end{aligned}$$

which conclude the proof. □

Proof of Lemma 10. First, we show that $A_\delta, B_\delta, C_\delta$ are bounded on $[0, t_0]$, uniformly in δ. Combining (8.17) and (8.18), we have

$$r A_\delta(t) \le \|\nabla f(x_0)\| + L_f t B_\delta(t) \le \|\nabla f(x_0)\| + L_f t \frac{L_{\psi^*} r t}{6} A_\delta(t).$$

Thus $A_\delta(t) \left(1 - \frac{L_{\psi^*} L_f}{6} t^2\right) \le \frac{\|\nabla f(x_0)\|}{r}$. And when $t \le \frac{2}{\sqrt{L_f L_{\psi^*}}}$, $1 - \frac{L_{\psi^*} L_f}{6} t^2 \ge \frac{1}{3}$, thus

$$A_\delta(t) \le \frac{3}{r} \|\nabla f(x_0)\|. \tag{8.21}$$

Next, we have

$$\begin{aligned} C_\delta(t) &\le r\left(\frac{tL_{\psi^*}}{2}A_\delta(t) + B_\delta(t)\right) && \text{by (8.19)} \\ &\le r\left(\frac{tL_{\psi^*}}{2}A_\delta(t) + \frac{L_{\psi^*}rt}{6}A_\delta(t)\right) && \text{by (8.18)} \\ &\le \frac{(3+r)L_{\psi^*}t}{2}\|\nabla f(x_0)\| && \text{by (8.21)} \end{aligned}$$

To conclude, we have for all $t \in [0, t_0]$

$$\begin{aligned} \|\dot{Z}_\delta(t)\| &\le tA_\delta(t) \le \frac{3t}{r}\|\nabla f(x_0)\|, \\ \|\dot{X}_\delta(t)\| &\le C_\delta(t) \le \frac{(3+r)L_{\psi^*}t}{2}\|\nabla f(x_0)\|, \end{aligned}$$

which are bounded uniformly in δ on $[0, t_0]$, thus the family is equi-Lipschitz-continuous on $[0, t_0]$. It also follows that it is uniformly bounded on the same interval. □

Chapter 9

Generalized and Adaptive Averaging

9.1 Accelerated mirror descent with generalized averaging

In Chapter 8, we proposed an accelerated mirror descent ODE for constrained, smooth convex optimization, which was motivated by a simple Lyapunov argument. In particular, we showed in Section 8.6 that the ODE can be interpreted as coupled dynamics of a dual variable $Z(t)$ which evolves in the dual space E^*, and a primal variable $X(t)$ which is obtained as the weighted average of a non-linear transformation of the dual trajectory. More precisely,

$$\begin{cases} \dot{Z}(t) = -\eta(t)\nabla f(X(t)),\ \eta(t) = \dfrac{t}{r} \\ X(t) = \dfrac{\int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau}{\int_0^t w(\tau)d\tau},\ w(\tau) = \tau^{r-1} \\ X(0) = \nabla\psi^*(Z(0)) = x_0, \end{cases}$$

where $r \geq 2$ is a fixed parameter, the initial condition x_0 is a point in the feasible set $\mathcal{X}$, and $\nabla\psi^*$ is a Lipschitz function, called the mirror operator, that maps from the dual space E^* to the feasible set $\mathcal{X}$. We showed in Theorem 16 that the solution trajectories of this ODE exhibit a quadratic convergence rate, i.e. if $f^\star$ is the minimum of f over the feasible set, then $f(X(t)) - f^\star \leq C/t^2$ for a constant C which depends on the initial conditions. This formalized an interesting connection between acceleration and averaging, which had been observed in [50] in the special case of unconstrained quadratic minimization.

A natural question that arises is whether different averaging schemes can be used to achieve the same rate, or perhaps faster rates. In this chapter, we provide a positive answer.

We study a broader family of accelerated mirror descent dynamics, given by

$$\mathrm{AMD}_{w,\eta}\begin{cases} \dot{Z}(t) = -\eta(t)\nabla f(X(t)) \\ X(t) = \dfrac{X(t_0)W(t_0) + \int_{t_0}^{t} w(\tau)\nabla\psi^*(Z(\tau))d\tau}{W(t)}, \text{ with } W(t) = \displaystyle\int_0^t w(\tau)d\tau \\ X(t_0) = \nabla\psi^*(Z(t_0)) = x_0, \end{cases} \tag{9.1}$$

parameterized by two positive, continuous weight functions w and η, where w is used in the averaging and η determines the rate at which Z accumulates gradients. This is illustrated in Figure 9.1. In this generalization, we choose to initialize the ODE at $t_0 > 0$ instead of 0 (to guarantee existence and uniqueness of a solution, as discussed in Section 9.2).

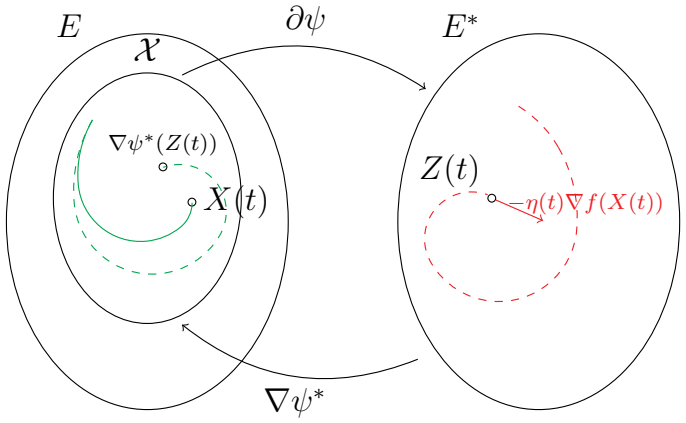


Figure 9.1: Illustration of $\mathrm{AMD}_{w,\eta}$. The dual variable Z (red dashed line) evolves in the dual space E^*, and accumulates negative gradients at a rate $\eta(t)$, and the primal variable $X(t)$ (green solid line) is obtained by averaging the mirrored trajectory $\nabla\psi^*(Z(\tau))$, $\tau \in [t_0, t]$ (green dashed line), with weights $w(\tau)$.

We give a unified study of this ODE using a parameterized Lyapunov function given by

$$V_r(X, Z, t) = r(t)(f(X) - f^\star) + D_{\psi^*}(Z, z^\star), \tag{9.2}$$

where D_{ψ^*} is the Bregman divergence associated with ψ^*, and $r(t)$ is a desired convergence rate (a non-negative function defined on $\mathbb{R}_+$). This function is a generalization of the Lyapunov function used in Chapter 8. We give in Section 9.3 a sufficient condition on η, w and r for V_r to be a Lyapunov function for $\mathrm{AMD}_{w,\eta}$. As an immediate consequence, we obtain that $f(X(t))$ converges to $f^\star$ at the rate $1/r(t)$ (Theorem 19), since whenever V_r is a Lyapunov function,

$$f(x(t)) - f^\star \leq \frac{1}{r(t)} V_r(X(t), Z(t), t) \leq \frac{1}{r(t)} V_r(x_0, z_0, t_0).$$

In Section 9.4, we exhibit a natural energy function of the system and show that under the same conditions on w, η and r, the energy is decreasing (Theorem 20). This provides further physical intuition on the dynamics. In Section 9.5, we give an equivalent formulation of $\text{AMD}_{w,\eta}$ written purely in the primal space. We give several examples of these dynamics for simple constraint sets. In particular, when the feasible set is the probability simplex, we derive in Section 9.6 an accelerated version of the replicator dynamics, an ODE that plays an important role in evolutionary game theory [135], viability theory [8], and has many applications including traffic systems [49], as discussed in Chapters 3.

Many heuristics have been developed to speed up the convergence of accelerated methods. Most of these heuristics consist in restarting the ODE (or the algorithm in discrete time) whenever a simple condition is met. When the function is strongly convex, and the strong convexity constant is known, then restarting the ODE at fixed intervals (which depend on the parameters of the function) can provably lead to exponential convergence. This is discussed in Section 9.7. For general, non-strongly convex functions, some heuristics have been proposed based on restarting. For example, a gradient restart heuristic is proposed in [105], in which the algorithm is restarted whenever the trajectory forms an acute angle with the gradient (which intuitively indicates that the trajectory is not making progress), and a speed restarting heuristic is proposed in [130], in which the ODE is restarted whenever the speed $\|\dot{X}(t)\|$ decreases (which intuitively indicates that progress is slowing). These heuristics are known to empirically improve the speed of convergence, but provide few guarantees. For example, the gradient restart in [105] is only studied for unconstrained quadratic problems, and the speed restart in [130] is only studied for unconstrained strongly convex problems. In particular, it is not guaranteed (to our knowledge) that these heuristics preserve the original convergence rate of the non-restarted method, when the objective function is not strongly convex. In Section 9.8, we propose a new heuristic that provides such guarantees, and that is based on a simple idea for adaptively computing the weights $w(t)$ along the solution trajectories (thus the weights become effectively a function of X and t, and not a predefined function of time). The heuristic simply decreases the time derivative of the Lyapunov function $L_r(X(t), Z(t), t)$ whenever possible. Thus it preserves the $1/r(t)$ convergence rate.

In the next chapter, we will derive a discretization of the accelerated mirror descent dynamics and of our adaptive averaging heuristic which guarantees a quadratic rate of convergence. We will also give numerical experiments in which we compare the performance of these heuristics. The experiments indicate that adaptive averaging compares favorably to the restarting heuristics in all of the examples, and gives a significant improvement in many cases.

9.2 Existence, uniqueness and viability of the solution

We start by giving an equivalent form of $\text{AMD}_{w,\eta}$, which we use to briefly discuss existence and uniqueness of the solution. Writing the second equation as $X(t)W(t) - X(t_0)W(t_0) =$

$\int_{t_0}^{t} w(\tau)\nabla\psi^*(Z(\tau))d\tau$, then taking the time-derivative, we have

$$\dot{X}(t)W(t) + X(t)w(t) = w(t)\nabla\psi^*(Z(t)).$$

Thus the ODE is equivalent to

$$\text{AMD}'_{w,\eta}\begin{cases} \dot{Z}(t) = -\eta(t)\nabla f(X(t)) \\ \dot{X}(t) = \dfrac{w(t)}{W(t)}(\nabla\psi^*(Z(t)) - X(t)) \\ X(t_0) = \nabla\psi^*(Z(t_0)) = x_0. \end{cases} \tag{9.3}$$

The following theorem guarantees existence, uniqueness and viability of the solution.

Theorem 18. *Suppose that ∇f and $\nabla\psi^*$ are Lipschitz, and that w, η are continuous functions. Furthermore suppose that $W(t_0) > 0$. Then $\text{AMD}_{w,\eta}$ has a unique maximal solution (i.e. defined on a maximal interval) $(X(t), Z(t))$ in $C^1([t_0, +\infty))$. Furthermore, the solution is viable, i.e. for all $t \geq t_0$, $X(t)$ belongs to the feasible set $\mathcal{X}$.*

Proof. By assumption, ∇f and $\nabla\psi^*$ are both Lipschitz, and w, η are continuous. Furthermore, $W(t)$ is non-decreasing and continuous, as the integral of a non-negative function, thus $w(t)/W(t) \leq w(t)/W(t_0)$. This guarantees that on any finite interval $[t_0, T)$, the functions $\eta(t)$ and $w(t)/W(t)$ are bounded. Therefore, $-\eta(t)\nabla f(X)$ and $\frac{w(t)}{W(t)}(\nabla\psi^*(Z) - X)$ are Lipschitz functions of (X, Z), uniformly in $t \in [t_0, T)$. By the Cauchy-Lipschitz theorem (e.g. Theorem 2.5 in [132]), there exists a unique C^1 solution defined on $[t_0, T)$. Since T is arbitrary, this defines a unique solution on all of $[t_0, +\infty)$. Indeed, any two solutions defined on $[t_0, T_1)$ and $[t_0, T_2)$ with $T_2 > T_1$ coincide on $[t_0, T_1)$. Finally, viability of the solution follows from the fact that $\mathcal{X}$ is convex and $X(t)$ is the weighted average of points in $\mathcal{X}$, specifically, x_0 and the set $\{\nabla\psi^*(Z(\tau)), \tau \in [t_0, t]\}$. □

Note that in general, it is important to initialize the ODE at t_0 and not at 0, since $W(0) = 0$ and $w(t)/W(t)$ can diverge at 0, in which case one cannot apply the Cauchy-Lipschitz theorem. It is possible however to prove existence and uniqueness with $t_0 = 0$ for some choices of w, by taking a sequence of Lipschitz ODEs that approximate the original one, as is done in the proof of Theorem 15. This is a technicality and does not matter for practical purposes, since the ODE can be initialized at any point in time.

9.3 Convergence guarantees

We now move to our main convergence result. Suppose that r is an increasing, positive differentiable function on $[t_0, +\infty)$, and consider the candidate Lyapunov function V_r defined in (9.2), where the Bregman divergence term is given by

$$D_{\psi^*}(z, y) := \psi^*(z) - \psi^*(y) - \langle\nabla\psi^*(y), z - y\rangle,$$

and $z^\star$ is a point in the dual space such that $\nabla\psi^*(z^\star) = x^\star$ belongs to the set of minimizers S. Let $(X(t), Z(t))$ be the unique maximal solution trajectory of $\mathrm{AMD}_{w,\eta}$, and let

$$\mathcal{V}_r(t) := V_r(X(t), Z(t), t) = r(t)(f(X(t)) - f^\star) + D_{\psi^*}(Z(t), z^\star).$$

Taking the time-derivative of $\mathcal{V}_r(t)$, we have

$$\begin{aligned}
\frac{d}{dt}\mathcal{V}_r(t) &= \frac{d}{dt}V_r(X(t), Z(t), t) \\
&= r'(t)(f(X(t)) - f^\star) + r(t)\left\langle \nabla f(X(t)), \dot{X}(t)\right\rangle + \left\langle \dot{Z}(t), \nabla\psi^*(Z(t)) - \nabla\psi^*(z^\star)\right\rangle \\
&= r'(t)(f(X(t)) - f^\star) + r(t)\left\langle \nabla f(X(t)), \dot{X}(t)\right\rangle + \left\langle -\eta(t)\nabla f(X(t)), X(t) + \frac{W(t)}{w(t)}\dot{X}(t) - x^\star\right\rangle \\
&\le (f(X(t)) - f^\star)(r'(t) - \eta(t)) + \left\langle \nabla f(X(t)), \dot{X}(t)\right\rangle \left(r(t) - \frac{\eta(t)W(t)}{w(t)}\right), \qquad (9.4)
\end{aligned}$$

where we used the expressions for $\dot{Z}$ and $\nabla\psi^*(Z)$ from $\mathrm{AMD}'_{w,\eta}$ in the second equality, and convexity of f in the last inequality. Equipped with this bound, it becomes straightforward to give sufficient conditions for V_r to be a Lyapunov function.

Theorem 19. *Suppose that for all $t \in [t_0, +\infty)$,*

1. *$\eta(t) \ge r'(t)$ and*
2. *$\left\langle \nabla f(X(t)), \dot{X}(t)\right\rangle \left(r(t) - \frac{\eta(t)W(t)}{w(t)}\right) \le 0$.*

Then V_r is a Lyapunov function for $\mathrm{AMD}_{w,\eta}$, and for all $t \ge t_0$,

$$f(X(t)) - f^\star \le \frac{V_r(X(t_0), Z(t_0), t_0)}{r(t)}.$$

Proof. The two conditions, combined with inequality (9.4), imply that $\frac{d}{dt}V_r(X(t), Z(t), t) \le 0$, thus V_r is a Lyapunov function. Finally, since D_{ψ^*} is non-negative, and the Lyapunov function is decreasing, we have

$$f(X(t)) - f^\star \le \frac{V_r(X(t), Z(t), t)}{r(t)} \le \frac{V_r(X(t_0), Z(t_0), t_0)}{r(t)}.$$

which proves the claim. □

Note that the second condition depends on the solution trajectory $X(t)$, and may be hard to check a priori. However, we give one special case in which the condition trivially holds.

Corollary 3. *Suppose that for all $t \in [t_0, +\infty)$, $\eta(t) = \frac{w(t)r(t)}{W(t)}$, and $\frac{w(t)}{W(t)} \ge \frac{r'(t)}{r(t)}$. Then V_r is a Lyapunov function for $\mathrm{AMD}_{w,\eta}$, and for all $t \ge t_0$, $f(X(t)) - f^\star \le \frac{V_r(X(t_0), Z(t_0), t_0)}{r(t)}$.*

Next, we describe a method to construct weight functions w, η that satisfy the conditions of Corollary 3, given a desired rate r. Of course, it suffices to construct w that satisfies $\frac{w(t)}{W(t)} \geq \frac{r'(t)}{r(t)}$ for all t, then to set $\eta(t) = \frac{w(t)r(t)}{W(t)}$. We can reparameterize the weight function by writing $\frac{w(t)}{W(t)} = a(t)$. We will refer to $a(t)$ as the normalized weight function. Then integrating from t_0 to t, we have $\frac{W(t)}{W(t_0)} = e^{\int_{t_0}^{t} a(\tau)d\tau}$, and

$$w(t) = w(t_0)\frac{a(t)}{a(t_0)}e^{\int_{t_0}^{t} a(\tau)d\tau}. \tag{9.5}$$

Therefore the conditions of the corollary are satisfied whenever $w(t)$ is of the form (9.5) and $a : \mathbb{R}_+ \to \mathbb{R}_+$ is a continuous, positive function with $a(t) \geq \frac{r'(t)}{r(t)}$. Note that the expression of w is defined up to the constant $w(t_0)$, which reflects the fact that the condition of the corollary is scale-invariant (if the condition holds for a function w, then it holds for αw for all $\alpha > 0$).

Example 1. *Let $r(t) = \frac{t^2}{r^2}$ for some positive constant r. Then $r'(t)/r(t) = 2/t$, and we can take $a(t) = \frac{\beta}{t}$ with $\beta \geq 2$. Then $w(t) = \frac{a(t)}{a(t_0)}e^{\int_{t_0}^{t} a(\tau)d\tau} = \frac{\beta/t}{\beta/t_0}e^{\beta \ln(t/t_0)} = (t/t_0)^{\beta-1}$ and $\eta(t) = a(t)r(t) = \frac{\beta}{r^2}t$, and for $\beta = r$, we recover the weighting scheme used in Chapter 8.*

Example 2. *More generally, if $r(t) = \frac{t^p}{r^p}$ for some positive constant r, then $r'(t)/r(t) = p/t$, and we can take $a(t) = \frac{\beta}{t}$ with $\beta \geq p$. Then $w(t) = (t/t_0)^{\beta-1}$, and $\eta(t) = a(t)r(t) = \frac{\beta}{r^p}t^{p-1}$.*

9.4 Energy of the system

In this section, we exhibit a second energy function that is guaranteed to decrease under the same conditions of Theorem 19. This energy function, unlike the Lyapunov function V_r, does not guarantee a specific convergence rate. However, it captures a natural measure of energy in the system, and generalizes the mechanical energy in the damped oscillator interpretation of Section 8.7.

To define this energy function, we will use the following characterization of the inverse mirror map: By duality of the subdifferentials (e.g. Theorem 23.5 in [117]), we have for a pair of convex conjugate functions ψ and ψ^* that $x \in \partial\psi^*(x^*)$ if and only if $x^* \in \partial\psi(x)$. To simplify the discussion, we will assume that ψ is also differentiable, so that $(\nabla\psi^*)^{-1} = \nabla\psi$ (this assumption can be relaxed). In what follows, we will denote by $\check{X} = \nabla\psi(X)$ and $\check{Z} = \nabla\psi^*(Z)$.

Theorem 20. *Let $(X(t), Z(t))$ be the unique maximal solution of* $\mathrm{AMD}_{w,\eta}$*, and let $\check{X} = \nabla\psi(X)$. Consider the energy function*

$$E_r(X, Z, t) = f(X) + \frac{1}{r(t)}D_{\psi^*}(Z, \check{X}), \tag{9.6}$$

and let $\mathcal{E}_r(t) := E_r(X(t), Z(t), t)$. Then if w, η satisfy condition (2) of Theorem 19, E_r is a decreasing function of time.

Proof. To make the notation more concise, we omit the explicit dependence on time in this proof. We have $D_{\psi^*}(Z, \check{X}) = \psi^*(Z) - \psi^*(\check{X}) - \left\langle X, Z - \check{X} \right\rangle$. Taking the time-derivative , we have

$$\begin{aligned}\frac{d}{dt} D_{\psi^*}(Z, \check{X}) &= \left\langle \nabla\psi^*(Z), \dot{Z} \right\rangle - \left\langle \nabla\psi^*(\check{X}), \dot{\check{X}} \right\rangle - \left\langle \dot{X}, Z - \check{X} \right\rangle - \left\langle X, \dot{Z} - \dot{\check{X}} \right\rangle \\ &= \left\langle \nabla\psi^*(Z) - X, \dot{Z} \right\rangle - \left\langle \dot{X}, Z - \check{X} \right\rangle\end{aligned}$$

Using the second equation in $\text{AMD}'_{w,\eta}$, we have $\nabla\psi^*(Z) - X = \frac{1}{a}\dot{X}$, and $\left\langle \dot{X}, Z - \check{X} \right\rangle = a \left\langle \nabla\psi^*(Z) - \nabla\psi^*(\check{X}), Z - \check{X} \right\rangle \geq 0$ by monotonicity of $\nabla\psi^*$. Combining, we have

$$\frac{d}{dt} D_{\psi^*}(Z, \check{X}) \leq -\frac{\eta}{a} \left\langle \dot{X}, \nabla f(X) \right\rangle,$$

and we can finally bound the derivative of E_r: since r is increasing, r'/r^2 is positive, and

$$\begin{aligned}\frac{d}{dt}\mathcal{E}_r(t) &= \left\langle \nabla f(X), \dot{X} \right\rangle + \frac{1}{r}\frac{d}{dt} D_{\psi^*}(Z, \check{X}) - \frac{r'}{r^2} D_{\psi^*}(Z, \check{X}) \\ &\leq \left\langle \nabla f(X), \dot{X} \right\rangle \left(1 - \frac{\eta}{ar}\right).\end{aligned}$$

Therefore condition (2) of Theorem 19 implies that $\frac{d}{dt} E_r(t) \leq 0$. □

This energy function can be interpreted, loosely speaking, as the sum of a potential energy given by $f(X)$, and a generalization of the kinetic energy, given by $\frac{1}{r(t)} D_{\psi^*}(Z, \check{X})$. Indeed, when the problem is unconstrained, then one can take $\psi^*(z) = \frac{1}{2}\|z\|^2$, in which case $\nabla\psi^* = \nabla\psi = I$, the identity, and

$$\frac{1}{r(t)} D_{\psi^*}(Z, \check{X}) = \frac{1}{2r(t)} \|\check{Z} - X\|^2 = \frac{1}{2r(t)a^2(t)} \|\dot{X}\|^2,$$

a quantity proportional to the kinetic energy (in the quadratic case given in Example 1, we have $r(t) = \frac{t^2}{r^2}$ and we can take $a(t) = \frac{r}{t}$, so that $\frac{1}{2r(t)a^2(t)} = \frac{1}{2}$, and the energy function reduces to the mechanical energy (8.13) studied in the oscillator interpretation in Chapter 8.

9.5 Primal Representation

An equivalent primal representation of the ODE $\text{AMD}_{w,\eta}$ can be obtained by rewriting the equations in terms of $\check{Z} = \nabla\psi^*(Z)$ and its derivatives ($\check{Z}$ is a primal variable that remains in $\mathcal{X}$, since $\nabla\psi^*$ maps into $\mathcal{X}$). Taking the time derivative of $\check{Z}(t) = \nabla\psi^*(Z(t))$, we have

$$\dot{\check{Z}}(t) = \nabla^2\psi^*(Z(t))\dot{Z}(t) = -\eta(t)\nabla^2\psi^* \circ \nabla\psi(\check{Z}(t))\nabla f(X(t)),$$

where $\nabla^2\psi^*(z)$ is the Hessian of ψ^* at z, defined as $\nabla^2\psi^*(z)_{ij} = \frac{\partial^2\psi^*(z)}{\partial z_j \partial z_i}$. Then using the averaging expression for X, we can write $\mathrm{AMD}_{w,\eta}$ in the following primal form

$$\mathrm{AMD}^p_{w,\eta} \begin{cases} \dot{\check{Z}}(t) = -\eta(t)\nabla^2\psi^* \circ \nabla\psi(\check{Z}(t))\nabla f\left(\dfrac{x_0 W(t_0) + \int_{t_0}^t w(\tau)\check{Z}(\tau)d\tau}{W(t)}\right) \\ \check{Z}(t_0) = x_0. \end{cases} \tag{9.7}$$

A similar derivation can be made for the mirror descent ODE without acceleration, which can be written as follows (see Section 8.2, and the original derivation of Nemirovski and Yudin in Chapter 3 in [98])

$$\mathrm{MD} \begin{cases} \dot{Z}(t) = -\nabla f(X(t)) \\ X(t) = \nabla\psi^*(Z(t)) \\ X(t_0) = x_0. \end{cases}$$

Note that this can be interpreted as a limit case of $\mathrm{AMD}_{\eta,w}$ with $\eta(t) \equiv 1$ and $w(t)$ a Dirac function at t. Taking the time derivative of $X(t) = \nabla\psi^*(Z(t))$, we have $\dot{X}(t) = \nabla^2\psi^*(Z(t))\dot{Z}(t)$, which leads to the primal form of the mirror descent ODE

$$\mathrm{MD}^p \begin{cases} \dot{X}(t) = -\nabla^2\psi^* \circ \nabla\psi(X(t))\nabla f(X(t)) \\ X(t_0) = x_0. \end{cases} \tag{9.8}$$

For some choices of ψ, $\nabla^2\psi^* \circ \nabla\psi$ has a simple expression. We give some examples below. The operator $\nabla^2\psi^* \circ \nabla\psi$ appears in both primal representations (9.7) and (9.8), and multiplies the gradient of f. It can be thought of as a transformation of the gradient which ensures that the primal trajectory remains in the feasible set, which will be illustrated in the examples below.

We also observe that in its primal form, $\mathrm{AMD}^p_{w,\eta}$ is a generalization of the ODE family studied by Wibisono et al. in [136], which can be written as $\frac{d}{dt}\nabla\psi(X(t) + e^{-\alpha(t)}\dot{X}(t)) = -e^{\alpha(t)+\beta(t)}\nabla f(X(t))$, for which they prove the convergence rate $\mathcal{O}(e^{-\beta(t)})$. This corresponds to setting, in our notation, $a(t) = e^{\alpha(t)}$, $r(t) = e^{\beta(t)}$ and taking $\eta(t) = a(t)r(t)$ (which corresponds to the condition of Corollary 3). The ODE studied in this section, $\mathrm{AMD}_{w,\eta}$, is more general in that it does not assume the condition of Corollary 3, which will be essential in deriving the adaptive averagin heuristic in Section 9.8.

Positive-orthant-constrained dynamics Suppose that $\mathcal{X}$ is the positive orthant $\mathbb{R}^n_+$, and consider the negative entropy function $\psi(x) = \sum_{i=1}^n x_i \ln x_i$. Then its dual is $\psi^*(z) = \sum_{i=1}^n e^{z_i - 1}$, and we have $\nabla\psi(x)_i = 1 + \ln x_i$ and $\nabla^2\psi^*(z)_{i,j} = \delta_i^j e^{z_i - 1}$, where δ_i^j is 1 if $i = j$ and 0 otherwise (see Section B.4 in the appendix). Thus for all $x \in \mathbb{R}^n_+$,

$$\nabla^2\psi^* \circ \nabla\psi(x) = \mathrm{diag}(x).$$

Therefore, the primal forms (9.8) and (9.7), reduce to, respectively,

$$\begin{cases} \forall i, \dot{X}_i = -X_i \nabla f(X)_i \\ X(0) = x_0 \end{cases}$$
$$\begin{cases} \forall i, \ \dot{\check{Z}}_i = -\eta(t) \check{Z}_i \nabla f(X(\check{Z}))_i \\ \check{Z}(t_0) = x_0 \end{cases}$$

where for the second ODE we write $X(\check{Z})$ compactly to denote the weighted average given by the second equation of $\mathrm{AMD}_{w,\eta}$,

$$X(\check{Z}) := \frac{X(t_0)W(t_0) + \int_{t_0}^{t} w(\tau)\check{Z}(\tau)d\tau}{W(t)}.$$

When f is affine, the mirror descent ODE lead to Lotka-Volterra equation which has applications in economics and ecology. For the mirror descent ODE, one can verify that the solution remains in the positive orthant since $\dot{X}$ tends to 0 as X_i approaches the boundary of the feasible set. Similarly for the accelerated version, $\dot{\check{Z}}$ tends to 0 as $\check{Z}$ approaches the boundary, thus $\check{Z}$ remains feasible, and so does X by convexity.

9.6 The accelerated replicator dynamics

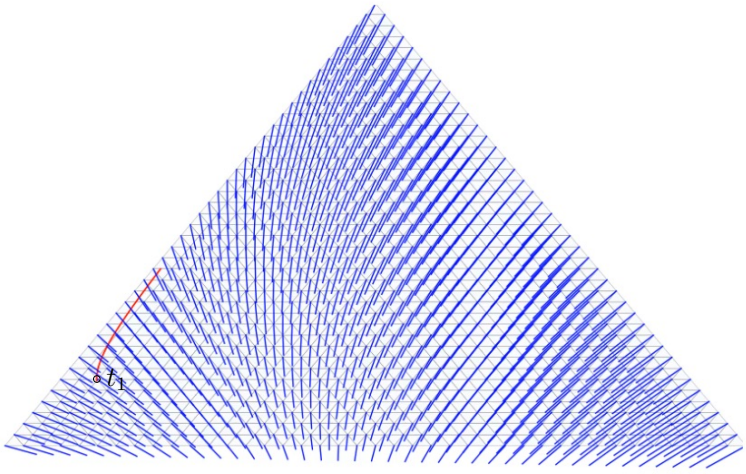
Now suppose that $\mathcal{X}$ is the n-simplex, $\mathcal{X} = \Delta = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$. Consider the distance-generating function $\psi(x) = \sum_{i=1}^n x_i \ln x_i + \delta_{\mathcal{X}}(x)$, where $\delta_{\mathcal{X}}(\cdot)$ is the convex indicator function of the feasible set (see Section B.6 in the appendix). Then its conjugate is $\psi^*(z) = \ln\left(\sum_{i=1}^n e^{z_i}\right)$, defined on E^*, and we have $\nabla\psi(x)_i = 1 + \ln x_i$, $\nabla\psi^*(z)_i = \frac{e^{z_i}}{\sum_k e^{z_k}}$, and $\nabla^2\psi^*(z)_{ij} = \frac{\delta_i^j e^{z_i}}{\sum_k e^{z_k}} - \frac{e^{z_i}e^{z_j}}{\left(\sum_k e^{z_k}\right)^2}$. Then it is simple to calculate

$$\nabla^2\psi^* \circ \nabla\psi(x)_{ij} = \frac{\delta_i^j x_i}{\sum_k x_k} - \frac{x_i x_j}{\left(\sum_k x_k\right)^2} = \delta_i^j x_i - x_i x_j.$$

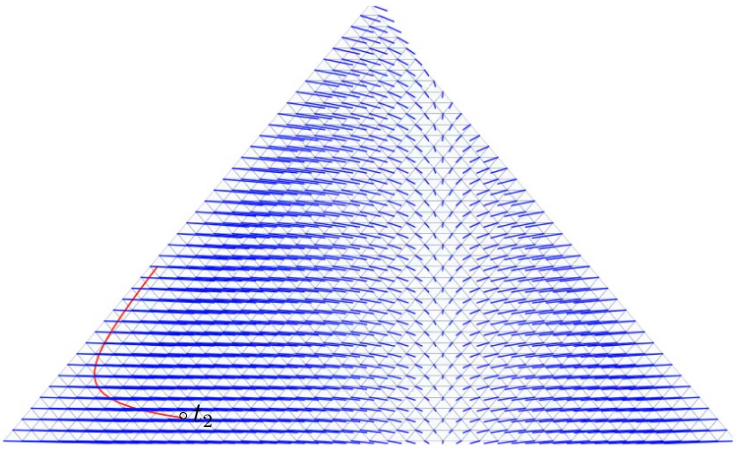
Therefore, the primal forms (9.8) and (9.7) reduce to, respectively,

$$\begin{cases} \forall i, \ \dot{X}_i + X_i \left(\nabla f(X)_i - \langle X, \nabla f(X)\rangle\right) = 0 \\ X(0) = x_0 \end{cases}$$
$$\begin{cases} \forall i, \dot{\check{Z}}_i + \eta(t)\check{Z}_i \left(\nabla f(X)_i - \left\langle \check{Z}, \nabla f(X)\right\rangle\right) = 0 \\ \check{Z}(0) = x_0. \end{cases}$$

The first ODE is the replicator dynamics analyzed in Chapter 3 for the congestion game. It has been studied for a long time, see [128] for a survey, and has many applications ranging from evolutionary game theory [135] and viability theory [8] to traffic networks and routing [49].



(a) Vector field $\nabla^2\psi^*(Z(t_1))\nabla f(\cdot)$.



(b) Vector field $\nabla^2\psi^*(Z(t_2))\nabla f(\cdot)$.

Figure 9.2: Vector field $X \mapsto \nabla^2\psi^*(Z(t))\nabla f(X)$ for different values of $Z(t)$ (taken along a solution trajectory for an example problem with solution on the relative boundary of the simplex). As $\nabla\psi^*(Z(t))$ approaches the relative boundary, the vector field changes in such a way that the component that is orthogonal to the boundary vanishes.

It is used to study large population dynamics, where one considers a population of players and a finite action set $\{1, \ldots, n\}$, such that at time t, $X_i(t)$ is the proportion of players who adopt action i. Then $\nabla f_i(X)$ is the cost (or the negative fitness) of action i given the distribution X. The ODE is called replicator as it can be obtained using a simple model of adaptive play as follows: at time t, players are randomly matched in pairs, and if their

current actions are, respectively, i and j, then the first player will switch to j (i.e. replicate the action of the second player) with a rate proportional to $\nabla_j f(X) - \nabla_i f(X)$, and similarly for the second player. As a consequence, the rate of increase of X_i is simply the sum over all actions j of $X_i X_j$ (the probability of the match (i, j)) multiplied by the difference in costs $\nabla_j f(X) - \nabla_i f(X)$, i.e.

$$\begin{aligned}\dot{X}_i &= \sum_{j=1}^{n} X_i X_j (\nabla_j f(X) - \nabla_i f(X)) \\ &= X_i \left(\sum_{j=1}^{n} X_j (\nabla_j f(X) - \nabla_i f(X)) \right) \\ &= X_i \left(\langle X, \nabla f(X) \rangle - \nabla_i f(X) \right).\end{aligned}$$

This example shows that the replicator dynamics can be accelerated simply by performing the original replicator update on the variable $\check{Z}$, in which (i) the gradient of the objective function is scaled by $\eta(t)$ at time t, and (ii) the gradient is evaluated at $X(t)$, the weighted average of the $\check{Z}$ trajectory.

Illustration of the operator $\nabla^2\psi^* \circ \nabla\psi(Z)$

Consider the accelerated replicator dynamics given above. This example can be used to illustrate the role of the Hessian term in Equation (9.7). Suppose that $\nabla\psi^*(Z)$ approaches the relative boundary of the feasible set, say $e^{Z_{i_0}}$ approaches 0. Then $(\nabla^2\psi^*(Z)\nabla f(X))_{i_0} = \frac{e^{Z_{i_0}}}{\sum_k e^{Z_k}} \left(\nabla_{i_0} f(X) - \left\langle \nabla f(X), \frac{e^Z}{\sum_{e^{Z_k}}} \right\rangle \right)$, also approaches 0. Figure 9.2 displays the vector field $\nabla^2\psi^*(Z)\nabla f(X)$ for different values of Z, to illustrate this phenomenon.

9.7 Restarting the ODE in the strongly convex case

When the objective function is strongly convex with a known parameter, faster convergence can be obtained by restarting the ODE at fixed intervals. That is, for some period T which depends on the function parameters, if we call $T_k = t_0 + kT$, we can define a trajectory (X, Z) to be the union of the solutions, on each interval $[T_k, T_{k+1})$, of the ODE

$$\begin{cases} \dot{Z}(t) = -\eta(t - T_k)\nabla f(X(t)) \\ X(t) = \dfrac{X(T_k)W(t_0) + \int_{T_k}^{t} w(\tau - kT)\nabla\psi^*(Z(\tau))d\tau}{W(t - kT)} \\ X(T_k) = \nabla\psi^*(Z(T_k)) = x_{T_k}, \text{ where } x_{T_k} = X(T_k^-). \end{cases} \tag{9.9}$$

That is, we solve a sequence of ODEs, one on each interval, and choose the initial conditions at the start of each interval so that the primal trajectory is continuous. The dual variable

Z is reinitialized at T_k in order to satisfy $\nabla\psi^*(Z(T_k)) = X(T_k^-)$. Since $X(t)$ is the weighted average of the $\nabla\psi^*(Z(\tau))$ on the interval $[T_k, t]$, setting $\nabla\psi^*(Z(T_k)) = X(T_k^-)$ ensures continuity of the primal trajectory (but the dual trajectory is in general, discontinuous at T_k). We now present a simple restarting strategy in the strongly convex case.

Theorem 21. *Suppose that the objective function f is ℓ_f-strongly convex, and that the distance generating function ψ^* is ℓ_{ψ^*}-strongly convex. Let $r(t)$ be a positive increasing rate function, and consider the ODE* $\mathrm{AMD}_{w,\eta}$*, with w, η, r satisfying the conditions of Theorem 19. Then the restarted ODE with period $T = r^{-1}\left(c\left(r(t_0) + \frac{1}{\ell_{\psi^*}\ell_f}\right)\right)$, with $c > 1$, guarantees that for all $k \geq 0$ and all $t \in [T_k + 1, T_{k+1}]$,*

$$f(X(t)) - f^\star \leq r(T)c^{-\frac{t}{T}}(f(x_0) - f^\star).$$

In other words, $f(X(t))$ converges exponentially to $f^\star$, at a rate $\frac{1}{T}$. Note, however, that this method requires previous knowledge of the strong convexity parameter ℓ_f.

Proof. By Theorem 19, we have for all $t \in [T_k, T_{k+1}]$,

$$\begin{aligned} f(X(t)) - f^\star &\leq \frac{1}{r(t - T_k)} V(X(T_k), Z(T_k), t_0) \\ &= \frac{1}{r(t - T_k)}(r(t_0)(f(X(T_k)) - f^\star) + D_{\psi^*}(Z(T_k), z^\star)). \end{aligned}$$

But since ψ^* is strongly convex,

$$\begin{aligned} D_{\psi^*}(Z(T_k), z^\star) &\leq \frac{1}{\ell_{\psi^*}}\|\nabla\psi^*(Z(T_k)) - \nabla\psi^*(z^\star)\|^2 && \text{by Proposition 23,} \\ &= \frac{1}{\ell_{\psi^*}}\|X_{T_k} - x^\star\|^2 \\ &\leq \frac{1}{\ell_{\psi^*}\ell_f}(f(X_k) - f^\star) && \text{by strong convexity of } f. \end{aligned}$$

Therefore,

$$f(X(t)) - f^\star \leq \frac{1}{r(t - T_k)}\left(r(t_0) + \frac{1}{\ell_{\psi^*}\ell_f}\right)(f(X(T_k)) - f^\star)$$

Thus using a restarting interval $T = r^{-1}\left(c\left(r(t_0) + \frac{1}{\ell_{\psi^*}\ell_f}\right)\right)$, and evaluating the last inequality at $t = T_{k+1}$, we have

$$f(X(T_{k+1})) - f^\star \leq \frac{f(X(T_k)) - f^\star}{c},$$

so by induction $f(X(T_k)) - f^\star \leq \frac{f(x_0)-f^\star}{c^k}$, and for $T_k + 1 \leq t \leq T_{k+1}$, we have

$$\begin{aligned} f(X(t)) - f^\star &\leq \frac{r(T)}{c}(f(X(t_k)) - f^\star) \\ &\leq \frac{r(T)}{c}\frac{f(x_0) - f^\star)}{c^k} \\ &= r(T)c^{-\frac{t}{T}}(f(x_0) - f^\star) \qquad \text{since } k+1 \geq \frac{t}{T} \end{aligned}$$

This concludes the proof. □

9.8 Adaptive averaging

In this section, we propose an adaptive averaging heuristic for adaptively computing the weights w. Note that in Corollary 3, we simply set $a(t) = \frac{\eta(t)}{r(t)}$ so that

$$\left\langle \nabla f(X(t)), \dot{X}(t) \right\rangle \left(r(t) - \frac{\eta(t)}{a(t)} \right)$$

is identically zero (thus trivially satisfying condition (2) of Theorem 19). However, from the bound (9.4), if this term is negative, then this helps further decrease the Lyapunov function V_r (as well as the energy function E_r). A simple strategy is then to adaptively define $a(t)$ as follows.

Definition 13. *Consider the accelerated mirror descent ODE with generalized averaging* $\mathrm{AMD}_{w,\eta}$*, and suppose the weight function w is given by Equation* (9.5)*. We say that the averaging is adaptive if the following conditions hold:*

$$\begin{cases} a(t) = \dfrac{\eta(t)}{r(t)} & \text{if } \left\langle \nabla f(X(t)), \dot{X}(t) \right\rangle > 0, \\ a(t) \geq \dfrac{\eta(t)}{r(t)} & \text{otherwise.} \end{cases}$$

If we further suppose that $\eta(t) \geq r'(t)$, then the conditions of Theorem 19 and Theorem 20 are satisfied, which guarantees the decrease of the Lyapunov function V_r and the energy function E_r.

Theorem 22. *Let $r(t)$ be a positive increasing rate function, and consider accelerated mirror descent* $\mathrm{AMD}_{w,\eta}$ *with adaptive averaging. Suppose that $\eta(t) \geq r'(t)$. Then V_r is a Lyapunov function and the energy E_r is non-increasing. In particular, we have for all $t \geq t_0$*

$$f(X(t)) - f^\star \leq \frac{V_r(X(t_0), Z(t_0), t_0)}{r(t)}$$

and adaptive averaging preserves the rate $r(t)$.

In the next chapter, we will propose a discretization of the accelerated mirror descent ODE, both for adaptive and non-adaptive averaging schemes, which results in a family of accelerated first-order methods. We show that when $r(t)$ is a quadratic, the discretization preserves the Lyapunov function, and as a consequence, it preserves the convergence rate. We also find that empirically, adaptive averaging significantly improves the speed of convergence.

Example 3 (Adaptive averaging for quadratic rates)**.** *Let $r(t) = \frac{t^2}{r^2}$ for some positive constant r, and let $\eta(t) = \frac{\beta t}{r^2}$ so that $\eta(t) \geq \frac{r'(t)}{r(t)}$. Then $a(t)$ is adaptive if*

$$\begin{cases} a(t) = \dfrac{\beta}{t} & \text{if } \left\langle \nabla f(X(t)), \dot{X}(t) \right\rangle > 0, \\ a(t) \geq \dfrac{\beta}{t} & \text{otherwise.} \end{cases}$$

This defines an adaptive version of Example 1, the discrete version of which will be studied in further detail in the next chapter. One limiting example is to set $a(t) = \frac{\beta}{t}$ if $\left\langle \nabla f(X(t)), \dot{X}(t) \right\rangle > 0$ and set $a(t)$ to be constant otherwise (this is a limit case because it would violate continuity of a). It is worth observing that a constant $a(t)$ over an interval corresponds to an exponential increase in the weight $w(t)$ by Equation (9.5), while $a(t) = \frac{\beta}{t}$ corresponds to the polynomial increase $w(t) = w(t_0)(t/t_0)^{\beta-1}$. In other words, the adaptive averaging scheme would increase the weights polynomially by default, and exponentially whenever the trajectory is moving in a good direction, i.e. $\left\langle \nabla f(X), \dot{X} \right\rangle \leq 0$.

Chapter 10

Discretizing the Accelerated Dynamics

In this chapter, we show that with a careful discretization of the continuous-time accelerated mirror descent dynamics developed in Chapters 8 and 9, we can obtain a general family of accelerated first-order methods for constrained optimization, which have a quadratic rate of convergence. We start with a naive discretization using a forward-backward Euler scheme in Section 10.1, and discuss why such a discretization does not, in general, preserve the Lyapunov function associated to the ODE. In Section 10.2, we give a modification of the discretization, and prove in Section 10.3 that the modified scheme is consistent with the ODE (i.e. the continuous-time limit of the discrete difference equations still correspond to the original ODE). In Section 10.4, we prove that the proposed family of accelerated mirror descent (as well as adaptive averaging) preserve the Lyapunov function associated to the ODE, and as a consequence, these methods are guaranteed to have a quadratic convergence rate. We give a detailed example in Section 10.5, which corresponds to the discretization of the accelerated replicator dynamics studied in the previous chapter. We review and discuss different restarting heuristics in Section 10.6, and test these different methods on several numerical examples in Section 10.7. In particular, we compare the performance of restarting and adaptive averaging. The results indicate that adaptive averaging compares favorably to the best known heuristics, with significant improvements in some cases. Finally, we conclude this part of the thesis in Section 10.8 with a summary and discussion of our results, and with directions for future research.

10.1 Forward-backward Euler discretization

Using a mixed Euler scheme (forward in the Z variable, and backward in the X variable), see e.g. Chapter 2 in [36], we can discretize the ODE system (8.11) using a step size $\sqrt{s}$ as follows (the choice of the step size as $\sqrt{s}$ instead of s will become clear in Section 10.2). Given a solution (X, Z) of the ODE (8.11), consider the correspondence between discrete

and continuous time, $t_k = k\sqrt{s}$, and let

$$x^{(k)} = X(t_k) = X(k\sqrt{s}).$$

Starting from the second form of the accelerated mirror descent ODE with generalized averaging (9.3),

$$\mathrm{AMD}'_{w,\eta} \begin{cases} \dot{Z}(t) = -\eta(t)\nabla f(X(t)) \\ \dot{X}(t) = a(t)(\nabla\psi^*(Z(t)) - X(t)) \\ X(t_0) = \nabla\psi^*(Z(t_0)) = x_0, \end{cases}$$

and approximating $\dot{X}(t_k)$ with $\frac{X(t_k+\sqrt{s})-X(t_k)}{\sqrt{s}}$, and, similarly, $\dot{Z}(t_k)$ with $\frac{Z(t_k+\sqrt{s})-Z(t_k)}{\sqrt{s}}$, consider the Euler discretization, forward in X and backward in Z,

$$\begin{cases} \dfrac{z^{(k+1)} - z^{(k)}}{\sqrt{s}} = -\eta_k \nabla f(x^{(k)}), \\ \dfrac{x^{(k+1)} - x^{(k)}}{\sqrt{s}} = a_{k+1}\left(\nabla\psi^*(z^{(k+1)}) - x^{(k+1)}\right), \end{cases} \tag{10.1}$$

where we have defined $\eta_k := \eta(k\sqrt{s})$ and $a_k := a(k\sqrt{s})$. The second equation can be rewritten as

$$x^{(k+1)} = \frac{x^{(k)} + a_{k+1}\sqrt{s}\nabla\psi^*(z^{(k+1)})}{1 + a_{k+1}\sqrt{s}}.$$

In other words, $x^{(k+1)}$ is a convex combination of $\nabla\psi^*(z^{(k)})$ and $x^{(k)}$ with coefficients $\lambda_{k+1} = \frac{a_{k+1}\sqrt{s}}{1+a_{k+1}\sqrt{s}}$ and $1-\lambda_{k+1} = \frac{1}{1+a_{k+1}\sqrt{s}}$. Note that since $\nabla\psi^*$ maps into $\mathcal{X}$, starting from $x^{(0)} \in \mathcal{X}$ guarantees that $x^{(k)}$ remains in $\mathcal{X}$ for all k. To summarize, our first discrete scheme can be written as

$$\begin{cases} z^{(k+1)} = z^{(k)} - \eta_k\sqrt{s}\nabla f(x^{(k)}), \\ x^{(k+1)} = \lambda_{k+1}\nabla\psi^*(z^{(k+1)}) + (1-\lambda_{k+1})x^{(k)}, \ \lambda_k = \dfrac{a_k\sqrt{s}}{1+a_k\sqrt{s}}. \end{cases} \tag{10.2}$$

An equivalent form of the mirror descent update

When the primal distance generating function can be written as the restriction to $\mathcal{X}$ of a differentiable function Ψ (Assumption 5 in the appendix), the mirror update $z^{(k+1)} = z^{(k)} - \eta_k\sqrt{s}\nabla f(x^{(k)})$ can be written in terms of the primal variable $\check{z}^{(k)} = \nabla\psi^*(z^{(k)})$, see the discussion in Section B.3 in the appendix.

In this case, the discretization can be written purely in terms of the primal variables $x^{(k)}$ and $\check{z}^{(k)}$ as follows

$$\begin{cases} x^{(k+1)} = \lambda_{k+1}\check{z}^{(k+1)} + (1-\lambda_{k+1})x^{(k)}, \ \lambda_k = \dfrac{a_k\sqrt{s}}{1+a_k\sqrt{s}}, \\ \check{z}^{(k+1)} = \arg\min\limits_{x\in\mathcal{X}} \eta_k\sqrt{s}\left\langle \nabla f(x^{(k+1)}), x\right\rangle + D_\psi(x, \check{z}^{(k)}). \end{cases} \tag{10.3}$$

We will eventually modify this scheme in order to be able to prove the quadratic convergence rate. However, we start by analyzing this version of the discretization to show that, in general, it fails to preserve the Lyapunov function from continuous time.

A candidate Lyapunov sequence

Motivated by the continuous-time Lyapunov function (9.2), and using the correspondence $t_k = k\sqrt{s}$, consider the candidate Lyapunov sequence, defined for $k \geq 1$,

$$v_r^{(k)} = V_r(x_{k-1}, z_k, t_k) = r_k(f(x^{(k-1)}) - f^\star) + D_{\psi^*}(z^{(k)}, z^\star).$$

where $r_k := r(k\sqrt{s})$. Then we have

$$\begin{aligned} v_r^{(k+1)} - v_r^{(k)} &= r_{k+1}(f(x^{(k)}) - f^\star) - r_k(f(x^{(k-1)}) - f^\star) + D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\ &= r_k(f(x^{(k)}) - f(x^{(k-1)})) + (r_{k+1} - r_k)(f(x^{(k)}) - f^\star) + D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star). \end{aligned}$$

The difference of the Bregman divergences in the last equality can be bounded as follows

$$\begin{aligned} &D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\ &= D_{\psi^*}(z^{(k+1)}, z^{(k)}) + \left\langle \nabla\psi^*(z^{(k)}) - \nabla\psi^*(z^\star), z^{(k+1)} - z^{(k)} \right\rangle \\ &= D_{\psi^*}(z^{(k+1)}, z^{(k)}) + \left\langle \frac{1}{a_k\sqrt{s}}(x^{(k)} - x^{(k-1)}) + x^{(k)} - x^\star, -\eta_k\sqrt{s}\nabla f(x^{(k)}) \right\rangle \\ &\leq D_{\psi^*}(z^{(k+1)}, z^{(k)}) + \eta_k\sqrt{s}(f^\star - f(x^{(k)})) + \frac{\eta_k}{a_k}(f(x^{(k-1)}) - f(x^{(k)})). \end{aligned}$$

where the first equality follows from the Bregman identity in Lemma 15, the second equality is by the discretization (10.1), and the last inequality is by convexity of f. Combining the last inequality with the expression of $v_r^{(k+1)} - v_r^{(k)}$, we have

$$\begin{aligned} &v_r^{(k+1)} - v_r^{(k)} \leq \\ &\sqrt{s}(f(x^{(k)}) - f^\star)\left(\frac{r_{k+1} - r_k}{\sqrt{s}} - \eta_k\right) + (f(x^{(k)}) - f(x^{(k-1)}))\left(r_k - \frac{\eta_k}{a_k}\right) + D_{\psi^*}(z^{(k+1)}, z^{(k)}). \end{aligned} \tag{10.4}$$

Let us compare this expression with the bound (9.4) on $\frac{d}{dt}V_r(X(t), Z(t), t)$ in the continuous-time case, copied below.

$$\frac{d}{dt}V_r(X(t), Z(t), t) \leq (f(X(t)) - f^\star)(r'(t) - \eta(t)) + \left\langle \nabla f(X(t)), \dot{X}(t) \right\rangle \left(r(t) - \frac{\eta(t)}{a(t)}\right).$$

We see that we obtain an analogous bound (where $r'(t)$ is approximated by $\frac{r_{k+1}-r_k}{\sqrt{s}}$), except for the additional Bregman divergence term $D_{\psi^*}(z^{(k+1)}, z^{(k)})$. Using a discrete counterpart of the conditions of Theorem 19, we can guarantee that the first two terms in the bound (10.4) are non-positive, but due to the additional Bregman divergence term, we cannot immediately conclude that $v^{(k)}$ is a Lyapunov sequence. This can be remedied by a modification of the discretization, described next.

10.2 Discrete-time accelerated mirror descent and adaptive averaging

In this section, we propose a modification of the Euler discretization, which preserves quadratic convergence rates, both for adaptive and non-adaptive averaging schemes. This results in a family of accelerated first-order methods, that generalizes Nesterov's accelerated proximal method [101]. For faster rates $r(t) = t^p$, $p > 2$, it is also possible to discretize the ODE and preserve the convergence rate, as proposed by Wibisono et al. [136], at the expense of using higher-order methods. For example, for cubic convergence rates, their discretization results in Nesterov's acceleration of regularized Newton method [99]. We focus on first-order methods since they are better suited to large-scale optimization, given the size and dimensionality of the data sets typically encountered in machine learning and modern data analysis applications.

First, we specialize the Euler discretization to the quadratic case. Following Example 1 and Example 3 (for the adaptive and non-adaptive version of the ODE), let $r(t) = \frac{t^2}{r^2}$ for some positive parameter r, and let $\eta(t) = \frac{\beta t}{r^2}$ with $\beta \geq 2$, so that $\eta(t) \geq r'(t)$ to satisfy the first condition of Theorem 19. We will keep a general normalized weight function $a(t)$, to be able to analyze both the adaptive and non-adaptive versions of the algorithm. As a result, we have

$$\begin{aligned} r_k &= r(k\sqrt{s}) = \frac{k^2 s}{r^2}, \\ \eta_k &= \eta(k\sqrt{s}) = \frac{\beta k\sqrt{s}}{r^2}, \end{aligned} \tag{10.5}$$

and the discretization (10.2) becomes

$$\begin{cases} x^{(k+1)} = \lambda_{k+1}\tilde{z}^{(k+1)} + (1-\lambda_{k+1})x^{(k)}, \ \lambda_k = \dfrac{a_k\sqrt{s}}{1+a_k\sqrt{s}}, \\ \tilde{z}^{(k+1)} = \arg\min\limits_{x\in\mathcal{X}} \dfrac{\beta k s}{r^2}\left\langle \nabla f(x^{(k+1)}), x\right\rangle + D_\psi(x, \tilde{z}^{(k)}). \end{cases}$$

Next, in the expression of $x^{(k+1)} = \lambda_k \nabla\psi^*(z^{(k+1)}) + (1-\lambda_k)x^{(k)}$, we propose to replace $x^{(k)}$ with $\tilde{x}^{(k+1)}$, obtained as the solution to the minimization problem

$$\tilde{x}^{(k+1)} = \arg\min_{x\in\mathcal{X}} \gamma s\left\langle \nabla f(x^{(k)}), x\right\rangle + R(x, x^{(k)}),$$

where γ is a positive constant that scales the step size (the appropriate conditions on γ will be derived in Theorem 23), and R is regularization function that satisfies the following assumption:

Assumption 4. *There exist $0 < \ell_R \leq L_R$ such that for all $x, x' \in E$, $\frac{\ell_R}{2}\|x-x'\|^2 \leq R(x,x') \leq \frac{L_R}{2}\|x-x'\|^2$.*

In the Euclidean case, one can take R to be the squared Euclidean distance, $R(x, x') = \frac{\|x-x'\|_2^2}{2}$, in which case $\ell_R = L_R = 1$ and the $\tilde{x}$ update becomes a prox-update. In the general case, one can take $R(x, x') = D_\phi(x, x')$ for some distance generating function ϕ which is ℓ_R-strongly convex and L_R-smooth, in which case the $\tilde{x}$ update becomes a mirror update.

The resulting method is illustrated in Figure 10.1. This algorithm is a generalization of Allen-Zhu and Orecchia's interpretation of Nesterov's method in [2], where $x^{(k+1)}$ is a convex combination of a mirror descent update and a gradient descent update.

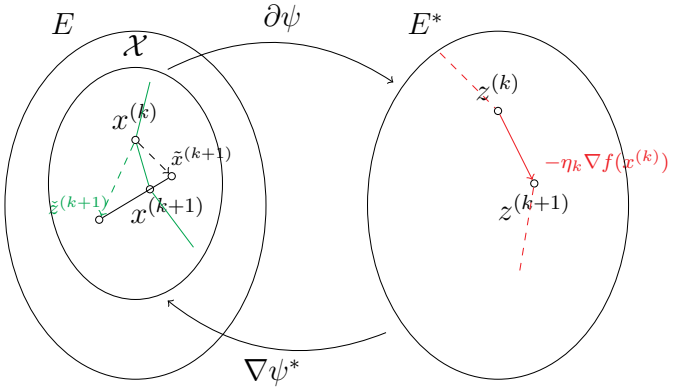


Figure 10.1: Illustration of the accelerated mirror descent method in discrete time. The dual variable $z^{(k)}$ is updated by taking a step in the direction of the negative gradient $-\nabla f(x^{(k)})$, with a rate $\eta_k\sqrt{s}$. The corresponding primal variable is $\check{z}^{(k+1)} = \nabla\psi^*(z^{(k+1)})$. The variable $\tilde{x}^{(k+1)}$ is obtained by performing a prox update from $x^{(k)}$, then $x^{(k+1)}$ is updated by taking a convex combination of $\check{z}^{(k+1)}$ and $\tilde{x}^{(k+1)}$. The weights used in the convex combination can be adaptive, depending on which averaging scheme is used.

Adaptive and non-adaptive averaging

In order to fully specify the algorithm, we need to define the sequence of normalized weights a_k. In the non-adaptive version, using a discrete counterpart of Example 1 in which $a(t) = \frac{\beta}{t}$, we simply use the correspondance $t = k\sqrt{s}$ and set

$$a_k = \frac{\beta}{k\sqrt{s}}, \tag{10.6}$$

In the adaptive version, we propose a rule which is based on the continuous-time adaptive averaging heuristic in Example 3. Let

$$a_k \in \left\{ \frac{\beta}{k\sqrt{s}}, \min\left(a_{k-1}, \frac{\beta^{\max}}{k\sqrt{s}}\right) \right\}, \text{ with } f(\tilde{x}^{(k+1)}) - f(\tilde{x}^{(k)}) > 0 \text{ only if } a_k = \frac{\beta}{k\sqrt{s}}. \tag{10.7}$$

This rule should be interpreted as follows: At iteration k, we try setting $a_k = \min\left(a_{k-1}, \frac{\beta^{\max}}{k\sqrt{s}}\right)$, compute $x^{(k)}$, then evaluate $f(\tilde{x}^{(k+1)}) - f(\tilde{x}^{(k)})$. If this quantity is non-positive, we move to

the next iteration, otherwise we set $a_k = \frac{\beta}{k\sqrt{s}}$ then recompute $x^{(k)}$ and move to the next iteration. This guarantees the following properties, which will be essential in proving that the discretization preserves the Lyapunov function in Lemma 13, and the quadratic convergence rate in Theorem 23.

Lemma 12. *The sequences a_k defined following the non-adaptive and the adaptive schemes* (10.6) *and* (10.7)*, both satisfy the following properties:*

1. For all k, $\frac{\beta}{k\sqrt{s}} \leq a_k \leq \frac{\beta^{\max}}{k\sqrt{s}}$;
2. For all k, $\frac{1}{a_k\sqrt{s}}(f(\tilde{x}^{(k+1)}) - f(\tilde{x}^{(k)})) \leq \frac{k}{\beta}(f(\tilde{x}^{(k+1)}) - f(\tilde{x}^{(k)}))$.

The resulting methods are summarized below, in Algorithm 9 and Algorithm 10. To simplify the presentation, we chose to write all of the updates in the primal space. In both algorithms, the mirror descent update line 3 can be equivalently written in the dual space as

$$z^{(k+1)} = z^{(k)} - \frac{\beta k s}{r^2}\nabla f(x^{(k)})$$
$$\check{z}^{(k+1)} = \nabla\psi^*(z^{(k+1)})$$

Algorithm 9 Accelerated mirror descent with non-adaptive averaging

Parameters: Distance generating function ψ^*,
Regularizer R,
Step size s,
Initial point x_0,
Weight rate $\beta \geq 2$.

1: Initialize $\tilde{x}^{(0)} = \check{z}^{(0)} = x_0$.
2: **for** $k \in \mathbb{N}$ **do**
3: $\quad \check{z}^{(k+1)} = \arg\min_{\check{z}\in\mathcal{X}} \frac{\beta k s}{r^2}\left\langle \nabla f(x^{(k)}), \check{z}\right\rangle + D_\psi(\check{z}, \check{z}^{(k)})$
4: $\quad \tilde{x}^{(k+1)} = \arg\min_{\tilde{x}\in\mathcal{X}} \gamma s\left\langle \nabla f(x^{(k)}), \tilde{x}\right\rangle + R(\tilde{x}, x^{(k)})$
5: $\quad x^{(k+1)} = \lambda_{k+1}\check{z}^{(k+1)} + (1-\lambda_{k+1})\tilde{x}^{(k+1)}$, with $\lambda_k = \frac{\sqrt{s}a_k}{1+\sqrt{s}a_k} = \frac{\beta}{\beta+k}$.
6: **end for**

10.3 Consistency of the discretization

One can show that given our assumptions on R, $\tilde{x}^{(k+1)} = x^{(k)} + \mathcal{O}(s)$. Indeed, we have

$$\begin{aligned}\frac{\ell_R}{2}\|\tilde{x}^{(k+1)} - x^{(k)}\|^2 &\leq R(\tilde{x}^{(k+1)}, x^{(k)}) \leq R(x^{(k)}, x^{(k)}) + \gamma s\left\langle \nabla f(x^{(k)}), x^{(k)} - \tilde{x}^{(k+1)}\right\rangle \\ &\leq \gamma s\|\nabla f(x^{(k)})\|_*\|\tilde{x}^{(k+1)} - x^{(k)}\|\end{aligned}$$

Algorithm 10 Accelerated mirror descent with adaptive averaging

Parameters: Distance generating function ψ^*,
Regularizer R,
Step size s,
Initial point x_0,
Weight rates $\beta^{\max} \geq \beta \geq 2$.

1: Initialize $\tilde{x}^{(0)} = x_0$, $\check{z}^{(0)} = x_0$, $a_1 = \frac{\beta}{\sqrt{s}}$
2: **for** $k \in \mathbb{N}$ **do**
3: $\quad \check{z}^{(k+1)} = \arg\min_{\check{z} \in \mathcal{X}} \frac{\beta k s}{r^2} \left\langle \nabla f(x^{(k)}), \check{z} \right\rangle + D_\psi(\check{z}, \check{z}^{(k)})$.
4: $\quad \tilde{x}^{(k+1)} = \arg\min_{\tilde{x} \in \mathcal{X}} \gamma s \left\langle \nabla f(x^{(k)}), \tilde{x} \right\rangle + R(\tilde{x}, x^{(k)})$
5: $\quad x^{(k+1)} = \lambda_{k+1} \check{z}^{(k+1)} + (1 - \lambda_{k+1}) \tilde{x}^{(k+1)}$, with $\lambda_k = \frac{\sqrt{s} a_k}{1 + \sqrt{s} a_k}$.
6: $\quad a_k \in \left\{ \frac{\beta}{k\sqrt{s}}, \min\left(a_{k-1}, \frac{\beta^{\max}}{k\sqrt{s}} \right) \right\}$, with $f(\tilde{x}^{(k+1)}) - f(\tilde{x}^{(k)}) > 0$ only if $a_k = \frac{\beta}{k\sqrt{s}}$.
7: **end for**

therefore $\|\tilde{x}^{(k+1)} - x^{(k)}\| \leq s \frac{2\gamma \|\nabla f(x^{(k)})\|_*}{\ell_R}$, which proves the claim. Using this observation, we can show that the modified discretization scheme is consistent with the original ODE (8.11), that is, the difference equations defining $x^{(k)}$ and $z^{(k)}$ converge, as s tends to 0, to the ordinary differential equations of the continuous-time system (8.11). The difference equations of Algorithm 9 are equivalent to (10.1) in which $x^{(k)}$ is replaced by $\tilde{x}^{(k+1)}$, i.e.

$$\begin{cases} \dfrac{z^{(k+1)} - z^{(k)}}{\sqrt{s}} = -\eta_k \nabla f(x^{(k)}) \\ \dfrac{x^{(k+1)} - \tilde{x}^{(k+1)}}{\sqrt{s}} = a_{k+1}(\nabla \psi^*(z^{(k+1)}) - x^{(k+1)}). \end{cases}$$

Now suppose there exist C^1 functions (X, Z), defined on $\mathbb{R}^+$, such that $X(t_k) \approx x^{(k)}$ and $Z(t_k) \approx z^{(k)}$ for $t_k = k\sqrt{s}$. Then, using the fact that $\tilde{x}^{(k)} = x^{(k)} + \mathcal{O}(s)$, we have $\frac{x^{(k+1)} - \tilde{x}^{(k+1)}}{\sqrt{s}} = \frac{x^{(k+1)} - x^{(k)}}{\sqrt{s}} + \mathcal{O}(\sqrt{s}) \approx \frac{X(t_k + \sqrt{s}) - X(t_k)}{\sqrt{s}} + \mathcal{O}(\sqrt{s}) = \dot{X}(t_k) + \mathcal{O}(\sqrt{s})$, and similarly, $\frac{z^{(k+1)} - z^{(k)}}{\sqrt{s}} \approx \dot{Z}(t_k) + o(1)$, therefore the difference equation system can be written as

$$\begin{cases} \dot{Z}(t_k) + o(1) = -\eta(t_k) \nabla f(X(t_k)) \\ \dot{X}(t_k) + \mathcal{O}(\sqrt{s}) = a(t_k + \sqrt{s})(\nabla \psi^*(Z(t_k + \sqrt{s})) - X(t_k + \sqrt{s})) \end{cases}$$

which converges to the ODE (8.11) as $s \to 0$.

10.4 Convergence guarantees

To prove convergence of the discrete accelerated mirror descent algorithms, we will show that the following sequence is a Lyapunov sequence,

$$\tilde{v}^{(k)} = V_r(\tilde{x}^{(k)}, z^{(k)}, t_k) = \frac{k^2 s}{r^2}(f(\tilde{x}^{(k)}) - f^\star) + D_{\psi^*}(z^{(k)}, z^\star).$$

In the following, we will suppose that $\beta^{\max} \geq \beta \geq 2$, that ψ^* is L_{ψ^*} smooth, and that R is ℓ_R strongly convex and L_R smooth.

Lemma 13. *Consider the sequence of iterates generated by Algorithm 9 or Algorithm 10. If* $\gamma \geq \frac{\beta\beta^{\max} L_R L_{\psi^*}}{r^2}$ *and* $s \leq \frac{\ell_R}{2L_f\gamma}$*, then*

$$\tilde{v}^{(k+1)} - \tilde{v}^{(k)} \leq \frac{(2k+1-k\beta)s}{r^2} f(\tilde{x}^{(k+1)} - f^\star).$$

It follows that if $\beta \geq 3$*,* $\tilde{v}^{(k)}$ *is non-increasing for* $k \geq 1$*.*

As a consequence, we can prove that the discrete accelerated mirror descent algorithms exhibit a quadratic convergence rate.

Theorem 23. *Accelerated mirror descent with adaptive and non-adaptive averaging, given in Algorithm 9 and Algorithm 10, with weight rate* $\beta^{\max} \geq \beta \geq 3$*, and step sizes* $\gamma \geq \frac{\beta\beta^{\max} L_R L_{\psi^*}}{r^2}$ *and* $s \leq \frac{\ell_R}{2L_f\gamma}$*, guarantees that for all* $k > 0$*,*

$$f(\tilde{x}^{(k)}) - f^\star \leq \frac{r^2 \tilde{v}^{(1)}}{sk^2}.$$

Additionally, we can bound $\tilde{v}^{(1)}$ *in terms of the initial conditions of the algorithm as follows:*

$$\tilde{v}^{(1)} \leq D_{\psi^*}(z^{(0)}, z^\star) + \frac{s}{r^2}(f(x^{(0)}) - f^\star).$$

Proof. Since $\tilde{v}^{(k)}$ is a Lyapunov sequence for $k \geq 1$ by Lemma 13, we have

$$f(\tilde{x}^{(k)}) - f^\star \leq \frac{r^2}{sk^2}\tilde{v}^{(k)} \leq \frac{r^2}{sk^2}\tilde{v}^{(1)}.$$

It remains to prove the bound on $\tilde{v}^{(1)}$. By Lemma 13, we have

$$\begin{aligned}\tilde{v}^{(1)} &\leq \tilde{v}^{(0)} + \frac{s}{r^2}(f(\tilde{x}^{(1)}) - f^\star) \\ &= D_{\psi^*}(z^{(0)}, z^\star) + \frac{s}{r^2}(f(\tilde{x}^{(1)}) - f^\star).\end{aligned}$$

To conclude, it suffices to show that $f(\tilde{x}^{(1)}) \leq f(x^{(0)})$. By definition, we have

$$\tilde{x}^{(1)} = \arg\min_{\tilde{x}\in\mathcal{X}} \gamma s \left\langle \nabla f(x^{(0)}), \tilde{x} \right\rangle + R(\tilde{x}, x^{(0)}),$$

thus

$$\gamma s \left\langle \nabla f(x^{(0)}), \tilde{x}^{(1)} \right\rangle + R(\tilde{x}^{(1)}, x^{(0)}) \leq \gamma s \left\langle \nabla f(x^{(0)}), x^{(0)} \right\rangle. \tag{10.8}$$

Therefore,

$$\begin{aligned}
&f(\tilde{x}^{(1)}) - f(x^{(0)}) \\
&\quad \leq \left\langle \nabla f(x^{(0)}), \tilde{x}^{(1)} - x^{(0)} \right\rangle + \frac{L_f}{2}\|\tilde{x}^{(1)} - x^{(0)}\|^2 && \text{by Lemma 14} \\
&\quad \leq \left\langle \nabla f(x^{(0)}), \tilde{x}^{(1)} - x^{(0)} \right\rangle + \frac{L_f}{\ell_R} R(\tilde{x}^{(1)}, x^{(0)}) && \text{by assumption on } R \\
&\quad \leq \left\langle \nabla f(x^{(0)}), \tilde{x}^{(1)} - x^{(0)} \right\rangle + \frac{1}{\gamma s} R(\tilde{x}^{(1)}, x^{(0)}) && \text{using that } \frac{L_f}{\ell_R} \leq \frac{1}{\gamma s} \\
&\quad \leq 0 && \text{by (10.8).}
\end{aligned}$$

This concludes the proof. □

Proof of Lemma 13. The difference $\tilde{v}^{(k+1)} - \tilde{v}^{(k)}$ is given by

$$\begin{aligned}
&\tilde{v}^{(k+1)} - \tilde{v}^{(k)} \\
&= \frac{(k+1)^2 s}{r^2}(f(\tilde{x}^{(k+1)}) - f^\star) - \frac{k^2 s}{r^2}(f(\tilde{x}^{(k)}) - f^\star) + D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star)
\end{aligned}$$

and we start by bounding the difference in Bregman divergences.

$$\begin{aligned}
&D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\
&\quad = -D_{\psi^*}(z^{(k)}, z^{(k+1)}) + \left\langle \nabla\psi^*(z^{(k+1)}) - \nabla\psi^*(z^\star), z^{(k+1)} - z^{(k)} \right\rangle && \text{by Lemma 15,} \\
&\quad \leq -\frac{1}{2L_{\psi^*}}\|\check{z}^{(k+1)} - \check{z}^{(k)}\|^2 + \left\langle \check{z}^{(k+1)} - x^\star, -\frac{\beta k s}{r^2}\nabla f(x^{(k)}) \right\rangle && \text{by Lemma 16.} \quad (10.9)
\end{aligned}$$

Now using the step from $x^{(k)}$ to $\tilde{x}^{(k+1)}$, we have

$$\tilde{x}^{(k+1)} = \arg\min_{x \in \mathcal{X}} \gamma s \left\langle \nabla f(x^{(k)}), x \right\rangle + R(x, x^{(k)})$$

with $\frac{\ell_R}{2}\|x-y\|^2 \leq R(x,y) \leq \frac{L_R}{2}\|x-y\|^2$. Therefore, for any x, $R(x, x^{(k)}) \geq R(\tilde{x}^{(k+1)}, x^{(k)}) + \gamma s \left\langle \nabla f(x^{(k)}), \tilde{x}^{(k+1)} - x \right\rangle$. We can write

$$\check{z}^{(k+1)} - \check{z}^{(k)} = \frac{1}{\lambda_k}\left(\lambda_k \check{z}^{(k+1)} + (1-\lambda_k)\tilde{x}^{(k)} - x^{(k)}\right) = \frac{1}{\lambda_k}\left(d^{(k+1)} - x^{(k)}\right),$$

where we have defined $d^{(k+1)}$ in the obvious way. Thus

$$\begin{aligned}
&\|\check{z}^{(k+1)} - \check{z}^{(k)}\|^2 \\
&= \frac{1}{\lambda_k^2}\|d^{(k+1)} - x^{(k)}\|^2 \\
&\geq \frac{1}{\lambda_k^2}\frac{2}{L_R}R(d^{(k+1)}, x^{(k)}) \\
&\geq \frac{1}{\lambda_k^2}\frac{2}{L_R}\left(R(\tilde{x}^{(k+1)}, x^{(k)}) + \gamma s\left\langle \nabla f(x^{(k)}), \tilde{x}^{(k+1)} - d^{(k+1)}\right\rangle\right) \\
&\geq \frac{1}{\lambda_k^2}\frac{2}{L_R}\left(\frac{\ell_R}{2}\|\tilde{x}^{(k+1)} - x^{(k)}\|^2 + \gamma s\left\langle \nabla f(x^{(k)}), \tilde{x}^{(k+1)} - \lambda_k\check{z}^{(k+1)} - (1-\lambda_k)\tilde{x}^{(k)}\right\rangle\right).
\end{aligned}$$

Thus, multiplying by $\frac{\lambda_k\beta k L_R}{2\gamma r^2}$,

$$\begin{aligned}
&\frac{\lambda_k\beta k L_R}{2\gamma r^2}\|\check{z}^{(k+1)} - \check{z}^{(k)}\|^2 \\
&\geq \frac{\beta k\ell_R}{2\lambda_k\gamma r^2}\|\tilde{x}^{(k+1)} - x^{(k)}\|^2 + \left\langle \frac{\beta k s}{r^2}\nabla f(x^{(k)}), \frac{1}{\lambda_k}\tilde{x}^{(k+1)} - \check{z}^{(k+1)} - \frac{1-\lambda_k}{\lambda_k}\tilde{x}^{(k)}\right\rangle. \qquad (10.10)
\end{aligned}$$

Subtracting (10.10) from (10.9),

$$\begin{aligned}
D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\
&\leq -\alpha_k\|\check{z}^{(k+1)} - \check{z}^{(k)}\|^2 - \frac{\beta k\ell_R}{2\lambda_k\gamma r^2}\|\tilde{x}^{(k+1)} - x^{(k)}\|^2 \\
&\quad + \left\langle -\frac{\beta k s}{r^2}\nabla f(x^{(k)}), -x^\star + \frac{1}{\lambda_k}\tilde{x}^{(k+1)} - \frac{1-\lambda_k}{\lambda_k}\tilde{x}^{(k)}\right\rangle,
\end{aligned}$$

where

$$\alpha_k = \frac{1}{2L_{\psi^*}} - \frac{\beta k\lambda_k L_R}{2\gamma r^2}.$$

Defining $D_1^{(k+1)} = \|\tilde{x}^{(k+1)} - x^{(k)}\|^2$ and $D_2^{(k+1)} = \|\check{z}^{(k+1)} - \check{z}^{(k)}\|^2$, we can rewrite the last inequality as

$$\begin{aligned}
D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\
&\leq -\alpha_k D_2^{(k+1)} - \frac{\beta k\ell_R}{2\lambda_k\gamma r^2}D_1^{(k+1)} \\
&\quad + \frac{\beta k s}{r^2}\left\langle -\nabla f(x^{(k)}), \tilde{x}^{(k+1)} - x^\star\right\rangle + \frac{1-\lambda_k}{\lambda_k}\frac{\beta k s}{r^2}\left\langle -\nabla f(x^{(k)}), \tilde{x}^{(k+1)} - \tilde{x}^{(k)}\right\rangle.
\end{aligned}$$

By Lemma 14, we can bound the inner products as follows

$$\begin{aligned}
\left\langle \tilde{x}^{(k+1)} - \tilde{x}^{(k)}, -\nabla f(x^{(k)})\right\rangle &\leq f(\tilde{x}^{(k)}) - f(\tilde{x}^{(k+1)}) + \frac{L_f}{2}D_1^{(k+1)}, \\
\left\langle \tilde{x}^{(k+1)} - x^\star, -\nabla f(x^{(k)})\right\rangle &\leq f^* - f(\tilde{x}^{(k+1)}) + \frac{L_f}{2}D_1^{(k+1)}.
\end{aligned}$$

Combining the last inequalities,

$$\begin{aligned}
&D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\
&\leq -\alpha_k D_2^{(k+1)} - \frac{\beta k \ell_R}{2\lambda_k \gamma r^2} D_1^{(k+1)} + \frac{\beta k s}{r^2}\left(f^\star - f(\tilde{x}^{(k+1)}) + \frac{L_f}{2} D_1^{(k+1)}\right) \\
&\quad + \frac{\beta k s}{r^2}\frac{1-\lambda_k}{\lambda_k}\left(f(\tilde{x}^{(k)}) - f(\tilde{x}^{(k+1)}) + \frac{L_f}{2} D_1^{(k+1)}\right) \\
&= \frac{\beta k s}{r^2}\frac{1-\lambda_k}{\lambda_k}\left(f(\tilde{x}^{(k)}) - f(\tilde{x}^{(k+1)})\right) + \frac{\beta k s}{r^2}\left(f^* - f(\tilde{x}^{(k+1)})\right) - \alpha_k D_2^{(k+1)} - \beta_k D_1^{(k+1)},
\end{aligned}$$

where

$$\beta_k = \frac{\beta k \ell_R}{2\lambda_k \gamma r^2} - \frac{\beta k s L_f}{2r^2} - \frac{\beta k s L_f}{2r^2}\frac{1-\lambda_k}{\lambda_k} = \frac{\beta k}{2r^2 \lambda_k}\left(\frac{\ell_R}{\gamma} - L_f s\right).$$

Next, observe that $\frac{1-\lambda_k}{\lambda_k} = \frac{1}{a_k\sqrt{s}}$, and by Lemma 12, $\frac{1}{a_k\sqrt{s}}(f(\tilde{x}^{(k+1)}) - f(\tilde{x}^{(k)})) \leq \frac{k}{\beta}(f(\tilde{x}^{(k+1)}) - f(\tilde{x}^{(k)}))$, therefore

$$\frac{1-\lambda_k}{\lambda_k}(f(\tilde{x}^{(k+1)}) - f(\tilde{x}^{(k)})) \leq \frac{k}{\beta}(f(\tilde{x}^{(k+1)}) - f(\tilde{x}^{(k)})).$$

Combining with the previous inequality, we have

$$\begin{aligned}
&D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\
&\leq \frac{k^2 s}{r^2}\left(f(\tilde{x}^{(k)}) - f(\tilde{x}^{(k+1)})\right) + \frac{\beta k s}{r^2}\left(f^* - f(\tilde{x}^{(k+1)})\right) - \alpha_k D_2^{(k+1)} - \beta_k D_1^{(k+1)},
\end{aligned}$$

Finally, we obtain a bound on the difference $\tilde{v}^{(k+1)} - \tilde{v}^{(k)}$:

$$\begin{aligned}
&\tilde{v}^{(k+1)} - \tilde{v}^{(k)} \\
&= \frac{(k+1)^2 s}{r^2}(f(\tilde{x}^{(k+1)}) - f^\star) - \frac{k^2 s}{r^2}(f(\tilde{x}^{(k)}) - f^\star) + D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\
&= \frac{k^2 s}{r^2}(f(\tilde{x}^{(k+1)}) - f(\tilde{x}^{(k)})) + \frac{(2k+1)s}{r^2}(f(\tilde{x}^{(k+1)}) - f^\star) + D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\
&\leq \frac{(2k+1-\beta k)s}{r^2}(f(\tilde{x}^{(k+1)}) - f^\star) - \alpha_k D_2^{(k+1)} - \beta_k D_1^{(k+1)}
\end{aligned}$$

For the desired inequality to hold, it suffices that $\alpha_k, \beta_k \geq 0$, i.e.

$$\begin{cases}
\frac{1}{2L_{\psi^*}} - \frac{\beta k \lambda_k L_R}{2\gamma r^2} \geq 0 \\
\frac{\beta k}{2r^2\lambda_k}\left(\frac{\ell_R}{\gamma} - L_f s\right) \geq 0,
\end{cases}$$

i.e.

$$\begin{cases}
\gamma \geq \frac{\beta k \lambda_k L_R L_{\psi^*}}{r^2} \\
s \leq \frac{\ell_R}{L_f \gamma}.
\end{cases}$$

To simplify the condition on γ, we observe that $\lambda_k = \frac{1}{1+\frac{1}{\sqrt{s}a_k}}$, and since $a_k \leq \frac{\beta^{\max}}{k\sqrt{s}}$ by Lemma 12, we have

$$\beta k \lambda_k \leq \frac{\beta k}{1 + \frac{k}{\beta^{\max}}} \leq \beta \beta^{\max}.$$

So it is sufficient that

$$\begin{cases} \gamma \geq \frac{\beta \beta^{\max} L_R L_{\psi^*}}{r^2}, \\ s \leq \frac{\ell_R}{L_f \gamma}, \end{cases}$$

which concludes the proof. □

10.5 Accelerated entropic descent

We give an instance of Algorithms 9 and Algorithm 10 for simplex-constrained problems, which corresponds to a discretization of the accelerated replicator dynamics studied in Section 9.6. Suppose that $\mathcal{X} = \Delta^n = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$ is the probability simplex in $\mathbb{R}^n$. Taking ψ to be the negative entropy on Δ, we have for $x \in \mathcal{X}$, $z \in E^*$,

$$\psi(x) = \sum_{i=1}^n x_i \ln x_i, \quad \psi^*(z) = \ln\left(\sum_{i=1}^n e^{z_i}\right), \quad \nabla\psi(x)_i = 1 + \ln x_i, \quad \nabla\psi^*(z)_i = \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}}.$$

The resulting mirror descent update is a simple entropy projection and can be computed exactly in $\mathcal{O}(n)$ operations (since the projection is given in closed form by the expression of $\nabla\psi^*$), and ψ^* can be shown to be 1-smooth w.r.t. $\|\cdot\|_\infty$, see for example [15, 24]. For the second update, we take $R(x, y) = D_\phi(x, y)$ where ϕ is a smoothed negative entropy function defined as follows: let $\epsilon > 0$, and let

$$\phi(x) = \epsilon \sum_{i=1}^n (x_i + \epsilon) \ln(x_i + \epsilon) + \delta_\Delta(x).$$

Although no simple expression is known for the mirror operator $\nabla\phi^*(z) = \arg\max_x \langle z, x\rangle - \phi(x)$, it can be solved efficiently, in $\mathcal{O}(n \log n)$ time using a deterministic algorithm, or $\mathcal{O}(n)$ expected time using a randomized algorithm, see Appendix C. Additionally, D_ϕ satisfies our assumptions: ϕ is $\frac{\epsilon}{1+n\epsilon}$-strongly convex w.r.t. $\|\cdot\|_1$, and 1-smooth w.r.t. $\|\cdot\|_\infty$. The resulting accelerated mirror descent method on the simplex can then be implemented efficiently, and by Theorem 23, it is guaranteed to converge in $\mathcal{O}(1/k^2)$ whenever $\gamma \geq 1$ and $s \leq \frac{\epsilon}{2(1+n\epsilon)L_f\gamma}$.

10.6 Restarting in discrete time

In this section, we adapt the restarting heuristics proposed by O'Donoghue and Candès in [105], and Su et al. in [130]. In Section 9.7, we motivated restarting in continuous time

for strongly convex functions, by observing that restarting at fixed intervals (determined by the strong convexity parameter of the objective), allows us to recover linear convergence. Even when the function is not strongly convex, restarting can be intuitively motivated by the observation that because of the "memory" in the solution (both in the dual variable $Z(t) = Z(0) + \int_0^t -\tau \nabla f(X(\tau))$, which accumulates gradients, and the primal variable due to averaging), the trajectory may point in a bad direction at a given time t. Thus, one can restart the ODE whenever a given condition is met, by resetting time to zero and reinitializing it at the current point, effectively wiping the memory of the solution.

Recall that in continuous-time, the algorithm is restarted at a given time T_k, by solving a new ODE given by (9.9), in which time is shifted by $-T_k$, and the dual variable is reinitialized to have $\nabla\psi^*(Z(T_k)) = X(T_k^-)$ (to ensure continuity of the primal trajectory).

We define restarting in discrete time similarly to the continuous time: The algorithm is restarted at time K simply by shifting future time by $-K$, and setting the dual variable $z^{(k+1)}$ such that $\nabla\psi^*(z^{(k+1)})$ coincides with the current iterate $x^{(k+1)}$. This is summarized in Algorithm 11, where we give a general template for the restarted version of Algorithm 9; specific restarting conditions are discussed below.

Algorithm 11 Accelerated mirror descent with restarting

Parameters: Distance generating function ψ^*,
Regularizer R,
Step size s,
Initial point x_0,
Weight rate $\beta \geq 2$.

Initialize $K = 0$, $\tilde{x}^{(0)} = \check{z}^{(0)} = x_0$.
for $k \in \mathbb{N}$ **do**
  $\check{z}^{(k+1)} = \arg\min_{\check{z}\in\mathcal{X}} \frac{\beta(k-K)s}{r^2} \left\langle \nabla f(x^{(k)}), \check{z} \right\rangle + D_\psi(\check{z}, \check{z}^{(k)})$
  $\tilde{x}^{(k+1)} = \arg\min_{\tilde{x}\in\mathcal{X}} \gamma s \left\langle \nabla f(x^{(k)}), \tilde{x} \right\rangle + R(\tilde{x}, x^{(k)})$
  $x^{(k+1)} = \lambda_{k-K+1}\check{z}^{(k+1)} + (1 - \lambda_{k-K+1})\tilde{x}^{(k+1)}$, with $\lambda_k = \frac{\sqrt{s}a_k}{1+\sqrt{s}a_k} = \frac{\beta}{\beta+k}$.
  if Restart condition **then**
    $K \leftarrow k$
    $\check{z}^{(k+1)} \leftarrow x^{(k+1)}$
  end if
end for

Many restarting conditions have been proposed in recent literature, motivated by unconstrained continuous-time optimization. We review some of these conditions below.

(i) Gradient restart condition [105]: $\left\langle x^{(k+1)} - x^{(k)}, \nabla f(x^{(k)}) \right\rangle > 0$. Intuitively, the algorithm is restarted whenever the trajectory makes an acute angle with the gradient, i.e. the trajectory is moving in a bad direction.

(ii) Function restart condition [105]: $f(x^{(k+1)}) \geq f(x^{(k)})$. This condition is similar to the gradient condition, since one has $f(x^{(k+1)}) \geq f(x^{(k)}) + \left\langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \right\rangle$ by convexity of f, thus the second condition is implied by the first.

(iii) Speed restart condition [130]: $\|x^{(k+1)} - x^{(k)}\| < \|x^{(k)} - x^{(k-1)}\|$. This condition was proposed by Su et al. in [130], and is motivated by the unconstrained Euclidean case: intuitively, the speed starts to decrease whenever the trajectory points in a bad direction.

We test these restarting heuristics numerically in the next section, compare them to our adaptive averaging heuristic, and discuss their empirical performance and qualitative differences.

10.7 Numerical experiments

To illustrate our results, we implement the accelerated mirror descent methods developed in this chapter, in Algorithms 9, 10 and 11, on simplex-constrained problems in $\mathbb{R}^n$, first for $n = 3$, to be able to visualize the probability simplex and the solution trajectories, then in higher dimension to better evaluate the performance of the method. We run the algorithm on the following objective functions:

1. A quadratic $f(x) = \langle x - x^\star, Q(x - x^\star) \rangle$ for a random positive definite matrix Q.
2. A weakly convex function given by $f(x) = g(x_1 - x_1^\star)^2 + (x_2 - x_3)^2$, where $g(x) = \min(x + \alpha, \max(0, x - \alpha))$. The set of minimizers of the second problem is the segment given by $\{x \in \Delta : x_1 \in [x_1^\star - \alpha, x_1^\star + \alpha] \text{ and } x_2 = x_3\}$. In the plots, the set of minimizers is visualized as a solid black segment.
3. A linear function $f(x) = \langle c, x \rangle$.

We compare the following methods:

1. The mirror descent method without acceleration.
2. The accelerated mirror descent method with non-adaptive weights given in Algorithm 9 (in which the normalized weights follow a predetermined schedule given by $a_k = \frac{\eta_k}{r_k} = \frac{\beta}{k\sqrt{s}}$),
3. Accelerated mirror descent with adaptive averaging, given in Algorithm 10,
4. The gradient restarting heuristic in [105], in which the algorithm is restarted from the current point whenever $\left\langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \right\rangle > 0$,
5. The speed restarting heuristic in [130], in which the algorithm is restarted from the current point whenever $\|x^{(k+1)} - x^{(k)}\| \leq \|x^{(k)} - x^{(k-1)}\|$.

We omit the function restarting heuristic, since in practice, it has a very similar behavior and performance to the gradient restarting heuristic. We implement the accelerated entropic descent algorithm proposed in Section 10.5, with parameter $\beta = r$, so the Lyapunov rate and the dual weight function are given by $r_k = \frac{k^2 s}{r^2}$, $\eta_k = \frac{ks}{r}$, according to Equation (10.5). The corresponding code and videos are available at the following url: `http://www.github.com/walidk/AcceleratedMirrorDescent`.

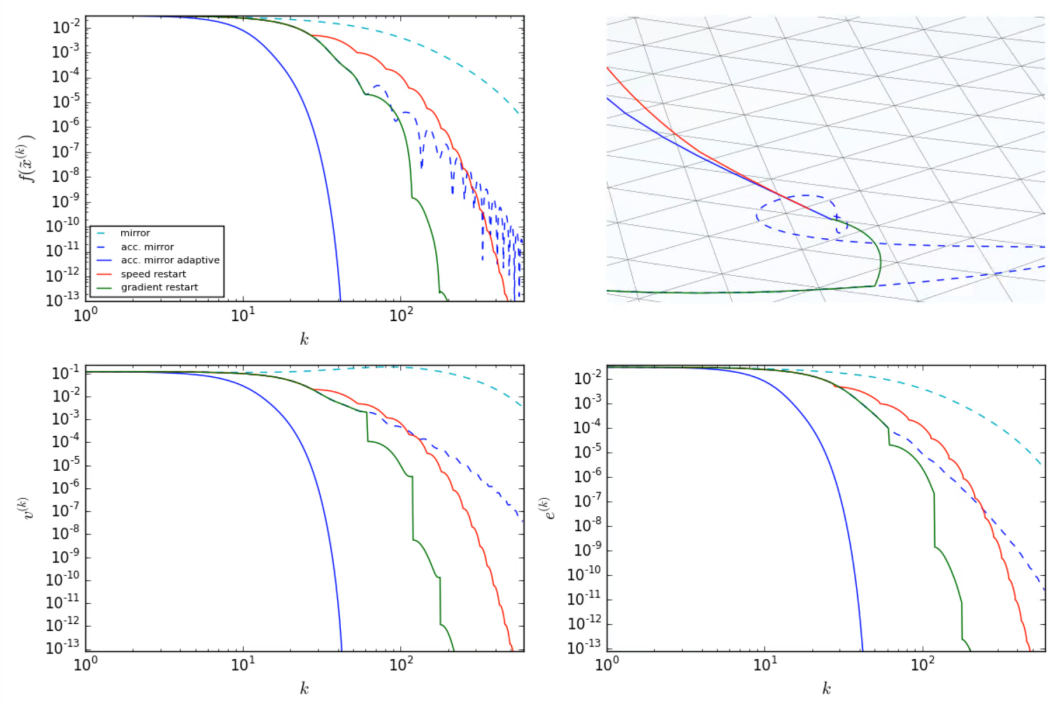
The results are shown in Figures 10.2, 10.3 and 10.4 for the experiments in $\mathbb{R}^3$. Each subfigure is divided into four plots: Clockwise from the top left, we show the value of the objective function $f(\tilde{x}^{(k)})$, the trajectory on the simplex, viewed as a subset of $\mathbb{R}^2$, the value of the energy function $e_r^{(k)} := E_r(\tilde{x}^{(k)}, z^{(k)}, k\sqrt{s})$, and the value of the Lyapunov function $v_r^{(k)} := V_r(\tilde{x}^{(k)}, z^{(k)}, k\sqrt{s})$. They are given by the following expressions:

$$\begin{aligned} e_r^{(k)} &= E_r(\tilde{x}^{(k)}, z^{(k)}, k\sqrt{s}) \\ &= f(\tilde{x}^{(k)}) + \frac{r^2}{k^2 s} D_{KL}(\tilde{x}^{(k)}, \tilde{z}^{(k)}), \\ v_r^{(k)} &= V_r(\tilde{x}^{(k)}, z^{(k)}, k\sqrt{s}) \\ &= \frac{k^2 s}{r^2}(f(\tilde{x}^{(k)}) - f^\star) + D_{KL}(\tilde{x}^{(k)}, \tilde{z}^{(k)}). \end{aligned}$$

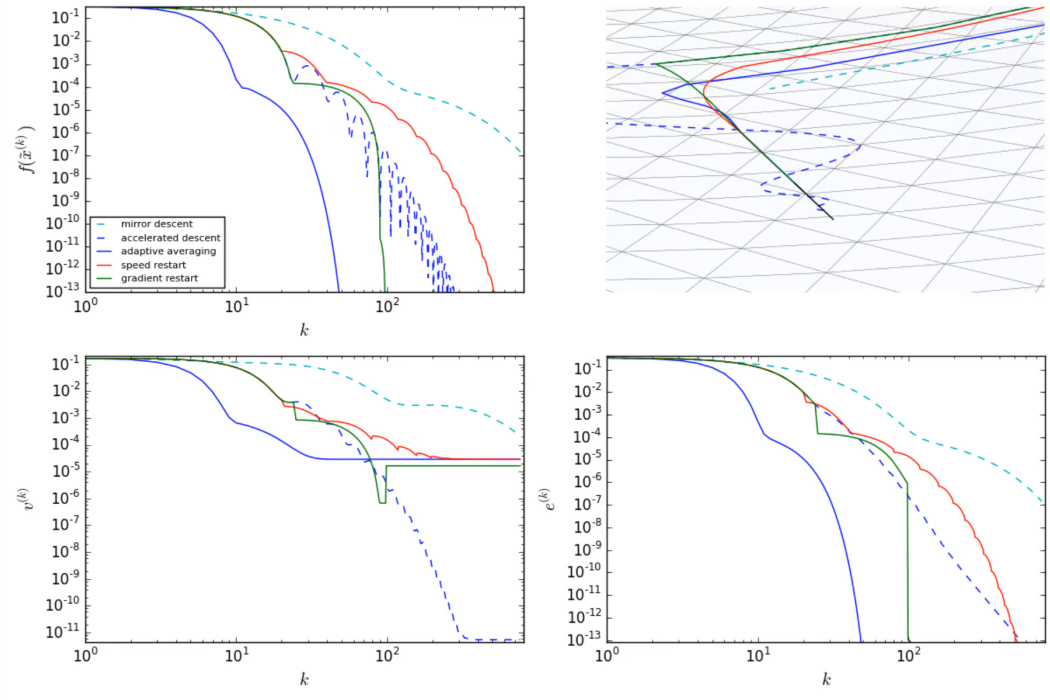
Effect of acceleration

The accelerated mirror descent method exhibits a polynomial convergence rate, which is empirically faster than the $\mathcal{O}(1/k^2)$ rate predicted by Theorem 23, both in the strongly and weakly convex cases. The experiments confirm that the Lyapunov function is decreasing for the accelerated method, but not for the original mirror descent method.

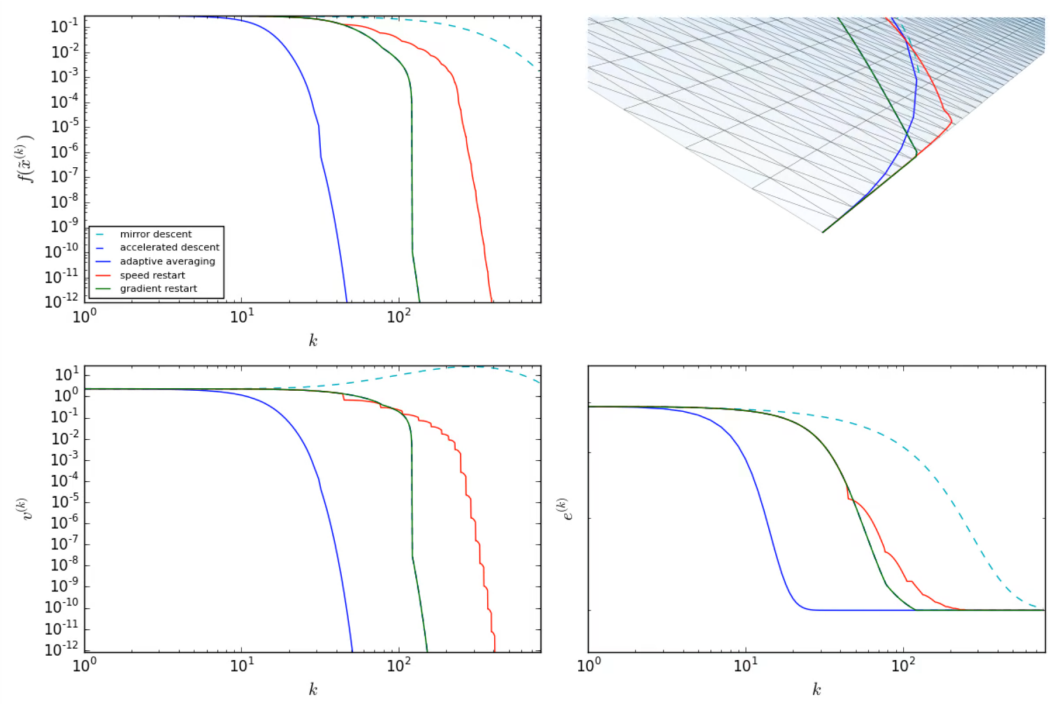
The method also exhibits oscillations around the set of minimizers. We observe that increasing the parameter r seems to reduce the period of the oscillations, and results in a trajectory that is initially slower, but faster for large k, see Figure 10.3. The restarting heuristics and the adaptive averaging heuristic alleviate the oscillation and empirically speed up the convergence. This observation also holds when the solution is on the boundary of the feasible set, see Figure 10.4 for an example.



(a) Strongly convex quadratic.



(b) Weakly convex function.



(a) Linear function.

Figure 10.2: Accelerated mirror descent on the simplex, adaptive averaging, and restarting heuristics. Each figure is divided into four subplots. Clockwise from the top right: function values $f(x^{(k)})$, trajectory of the simplex, Lyapunov values $v_r^{(k)} = V_r(\tilde{x}^{(k)}, z^{(k)}, k\sqrt{s})$, and energy values $e_r^{(k)} := E_r(\tilde{x}^{(k)}, z^{(k)}, k\sqrt{s})$.

Effect of adaptive averaging and restarting heuristics

The results in Figure 10.2 show that adaptive averaging compares favorably to the restarting heuristics on all these examples, with a significant improvement in the strongly convex case. Additionally, the experiments confirm that under the adaptive averaging heuristic, the Lyapunov function is decreasing. This is not the case for the restarting heuristics as can be seen on the weakly convex example. It is interesting to observe, however, that the energy function E_r is non-increasing for all the methods in our experiments. If we interpret the energy as the sum of a potential and a kinetic term, then this could be explained intuitively by the fact that restarting preserves the potential energy and can only decrease the kinetic energy. It is also worth observing that even though the Lyapunov function is non-decreasing, it will not necessarily converge to 0 when there is more than one minimizer (in particular, its limit will depend on the choice of $z^\star$ in the definition of V_r).

Finally, we observe that these heuristics have a different qualitative behavior. The accelerated method exhibits oscillations around the set of minimizers, and the heuristics alleviate

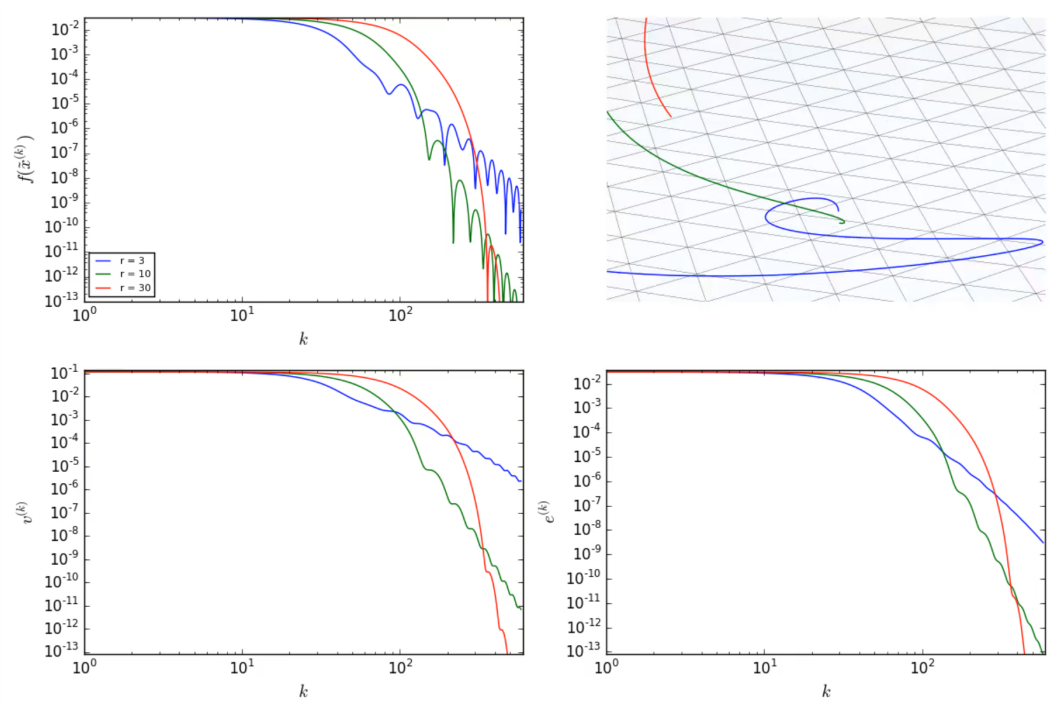


Figure 10.3: Effect of the parameter r.

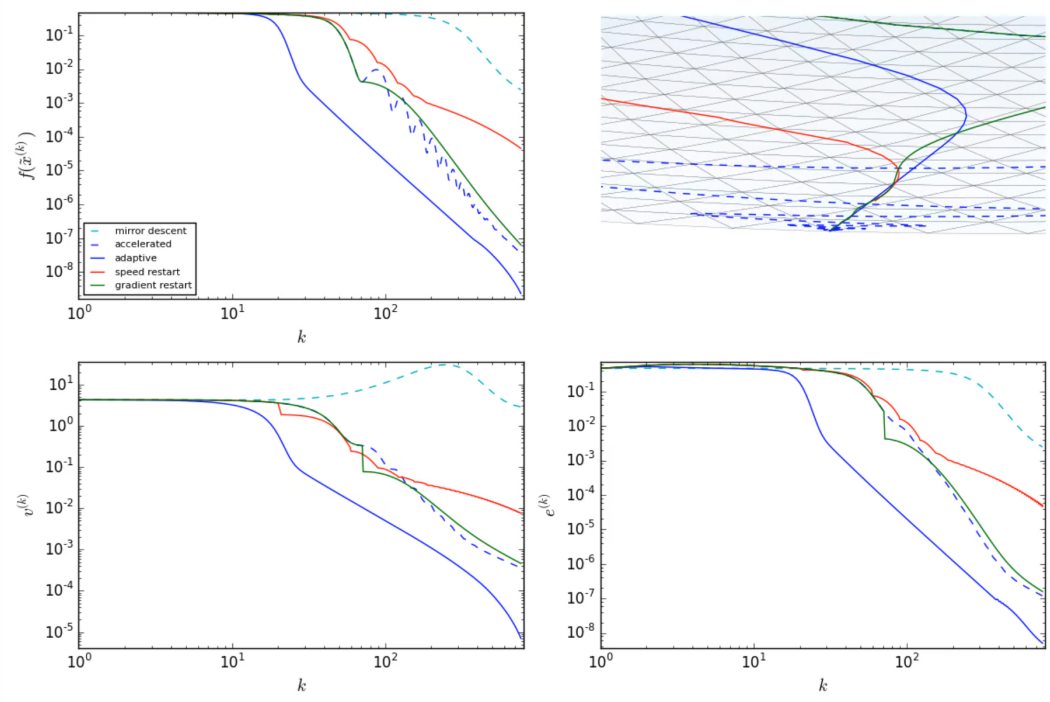
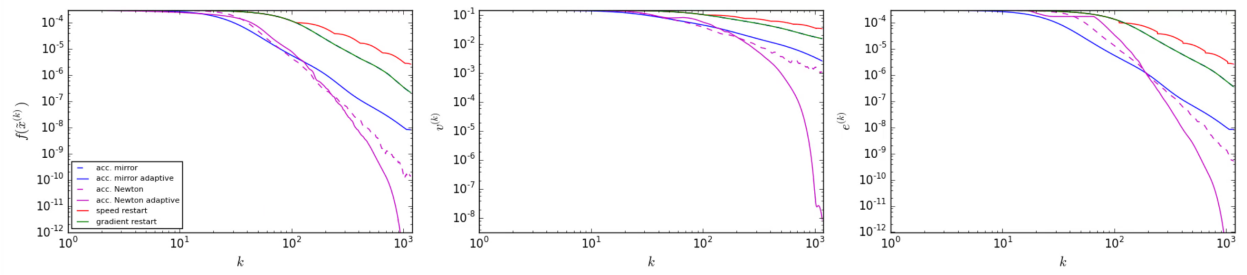
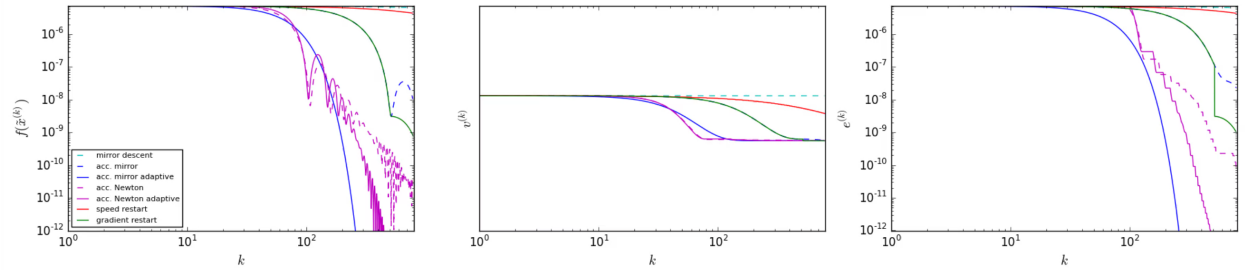


Figure 10.4: Example with the solution on the relative boundary of the simplex.

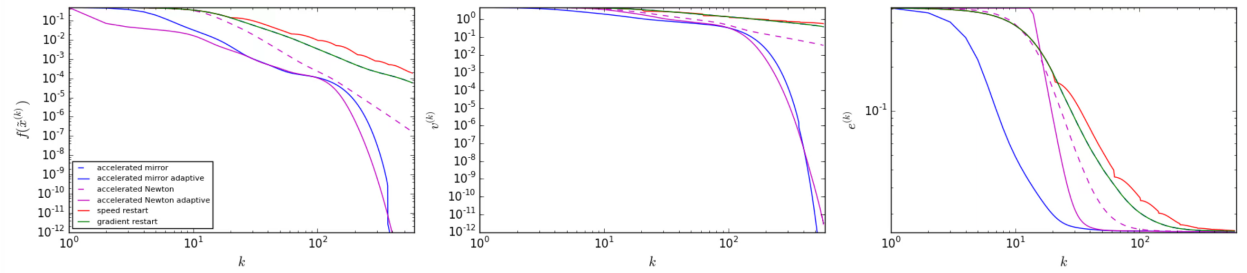
these oscillations in different ways: Intuitively, the adaptive averaging acts by increasing the weights on portions of the trajectory which make the most progress, while the restarting heuristics reset the velocity of the solution trajectory to zero whenever the algorithm detects that the trajectory is moving in a bad direction. The speed restarting heuristic seems to be more conservative in that it restarts more frequently.



(a) Strongly convex quadratic.



(b) Weakly convex function.



(c) Linear function.

Figure 10.5: Adaptive averaging for accelerated mirror descent and cubic-regularized Newton method.

Adaptive averaging in higher-order methods

In this section, we implement a discrete version of adaptive averaging for Nesterov's cubic-regularized Newton method [99]. Wibisono et al. show in [136] that this method can be

interpreted as a discretization of the ODE with cubic rate $r(t) = t^3$. Although we do not provide convergence guarantees in the cubic case for the discrete adaptive averaging, the continuous-time version is guaranteed to converge by Theorem 22. In order to implement adaptive averaging in discrete time, we can take, following Example 2, $r(t) = \frac{t^3}{r^3}$ and $\eta(t) = \frac{\beta t^2}{r^3}$, with $\beta \geq 3$, i.e.

$$r_k = \frac{k^3 s^{3/2}}{r^3}$$
$$\eta_k = \frac{\beta k^2 s}{r^3}.$$

Like in the quadratic case, we have that $\eta_k / r_k = \frac{\beta}{k\sqrt{s}}$, so we use the same adaptive rule (10.7) for setting the normalized weights a_k.

We provide additional numerical experiments in higher dimension $n = 100$, to illustrate the performance of the adaptive averaging compared to the restarting heuristics, both in the quadratic and the cubic case. We test the algorithm on simplex-constrained problems, with quadratic objective functions $f(x) = (x - s)^T A(x - s)$ with a positive definite matrix A in the first example, and a positive semidefinite matrix in the second example (with rank 10), and use a linear function in the last example. The results are reported in Figure 10.5. Each subfigure has three plots: From left to right, we show the value of objective function, the Lyapunov function and the energy function. We observe similar results to those in dimension 3. Adaptive averaging speeds up the convergence, both in the quadratic and cubic case, and performs as well as the restarting heuristics, with a significant improvement in one of the examples (in this case the linear example).

10.8 Conclusion

By combining the Lyapunov argument that motivated mirror descent, and a recent ODE interpretation [130] of Nesterov's method, we proposed a method to construct an energy function tailored to a given constrained convex optimization problem. The energy function combines a term that encodes the desired convergence rate, and a term that encodes the constraints (the Bregman divergence term). We then derived an ODE which is tailored to that energy function, and showed existence, uniqueness and viability of its solutions. It turns out that this ODE also has a simple interpretation as a coupling between a dual variable $Z(t)$ which cumulates gradients (similar to the original mirror descent method, but with an *increasing* rate $\eta(t)$), and a primal variable $X(t)$ obtained by averaging the mirrored dual trajectory $\nabla\psi^*(Z(\tau))$, $\tau \in [0, t]$ with weights $w(t)$. Motivated by this averaging interpretation, we studied a family of ODEs with a generalized averaging scheme, and gave sufficient conditions on the weight functions w, η to guarantee a given convergence rate in continuous time. We showed as an example how the replicator ODE can be accelerated by averaging. Our adaptive averaging heuristic preserves the convergence rate in continuous time (since it preserves the Lyapunov function). We proved that a careful discretization of

the ODE gives a quadratic convergence rate, both for adaptive and non-adaptive averaging. Empirically, adaptive averaging performs at least as well as other known heuristics for accelerated first-order methods, and in some cases considerably better. This encourages further investigation into the performance of adaptive averaging, both theoretically (by attempting to prove faster rates, e.g. for strongly convex functions), and numerically, by testing it on other, higher-order accelerated methods.

This approach can also be extended to more general classes of problems, such as maximal monotone operators. Continuous and discrete dynamics for finding a zero of a maximal monotone operator are derived for example in [108], and a promising direction is to develop a Lyapunov approach to these classes of dynamics, and extend them to the constrained case using the formalism of mirror descent and Bregman divergences, such as in [131, 97].

The main tool we used for proving the convergence of the accelerated methods obtained after discretization is to use a discrete counterpart of the continuous-time Lyapunov function. A promising avenue for research is to use the theory of variational time integrators, which studies the question of discretizing continuous dynamics while preserving natural quantities, such as the mechanical energy in mechanical systems, see e.g. [93, 85] and the references therein. The idea of the method is to discretize Hamilton's principle of critical action, associated to the system, rather than the ODE of the dynamics. And while these methods have been designed mainly for conservative mechanical systems, they are known to have good empirical performance for dissipative systems. This is important since the accelerated mirror descent dynamics are dissipative by design (otherwise the trajectories would not settle at the bottom of the potential field). In [136], Wibisono et al. give a Lagrangian interpretation (and its dual Hamiltonian interpretation) of the family of accelerated mirror descent dynamics. Starting from this interpretation and applying variational time integrators can lead to numerical methods to discretize the continuous-time dynamics, while preserving the Lyapunov function, hence the convergence rates.

Part III

Appendices

Appendix A

Results from convex analysis

In this chapter, we review some standard definitions and results from convex analysis which are used throughout the thesis. Most standard results are given without proof, but proofs can be found for example in [117].

A.1 Convex functions and convex conjugates

A function $f : E \to \mathbb{R} \cup \{+\infty\}$ defined on $E = \mathbb{R}^n$ is said to be a proper convex function if and only if it satisfies, for all x, y and all $\lambda \in [0, 1]$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

It is strictly convex if the inequality is strict for all $x \neq y$ and $\lambda \in (0, 1)$.

The effective domain of f is the set of points where f is finite,

$$\operatorname{dom} f = \{x \in E : f(x) < \infty\}.$$

The effective domain is a convex set, and one can define its relative interior, denoted by $\operatorname{ri} \operatorname{dom} f$, as its interior relative to the smallest affine set containing it. The relative boundary is defined similarly.

A proper convex function is said to be closed if its epigraph is closed. We will mainly consider closed proper convex functions, since they have nicer properties in general than functions that are not closed, and taking the closure of a convex function f (i.e. the infimum of all closed functions that dominate f) only changes f on the relative boundary of its effective domain.

Convex conjugate

Most duality results in convex analysis stem from the following simple idea: that a closed convex function can be described either as a locus of points $(x, f(x))$, or as the supremum of linear functions that lower-bound f (the same idea applies to a closed convex set, which can be described as the intersection of all half-spaces that contain it). More precisely:

Definition 14 (Convex conjugate)**.** *Given a convex function f, the convex conjugate of f is given by: for all $x^* \in E^*$ (the dual space of E),*

$$f^*(x^*) = \sup_{x \in E} \langle x^*, x \rangle - f(x).$$

As a consequence of the definition, we have

$$f(x) \geq \langle x^*, x \rangle - f^*(x^*) \ \forall x,$$

and this defines a linear function $x \mapsto \langle x^*, x \rangle - f^*(x^*)$ that lower-bounds f. Fenchel's duality theorem simply states that when f is closed and convex, it is the supremum of all these linear functions. In other words,

$$f(x) = \sup_{x^* \in \mathbb{R}^n} \langle x^*, x \rangle - f^*(x^*),$$

which is, by definition, the conjugate of f^*.

Theorem 24 (Fenchel's duality theorem)**.** *If f is a closed proper convex function, then*

$$f^{**} = f.$$

A.2 Duality of subdifferentials

Definition 15 (Subgradient and subdifferential)**.** *A vector $x^* \in E^*$ is called a subgradient of f at x if*

$$f(y) \geq f(x) + \langle x^*, y - x \rangle \ \ \forall y,$$

that is, if the linear function $y \mapsto f(x) + \langle x^*, y - x \rangle$ lower-bounds f. The set of all such vectors x^* is called the subdifferential of f at x and denoted $\partial f(x)$.

By definition of the subgradient, we have the following characterization in terms of the convex conjugate of f,

$$\begin{aligned} x^* \in \partial f(x) &\Leftrightarrow \langle x^*, x \rangle - f(x) \geq \langle x^*, y \rangle - f(y) \, \forall y \in E \\ &\Leftrightarrow x \in \arg\max_{y \in E} \langle x^*, y \rangle - f(y) \\ &\Leftrightarrow f^*(x^*) = \langle x^*, x \rangle - f(x) \end{aligned}$$

and switching the roles of f and f^* (and using the fact that $f^{**} = f$), we have the following theorem (which can be obtained as a special case of Theorem 23.5 in [117])

Theorem 25. *Given a proper convex closed function f, we have the following equivalence for all $(x, x^*) \in E \times E^*$,*

$$\begin{aligned} f^*(x^*) + f(x) = \langle x^*, x\rangle &\Leftrightarrow x^* \in \partial f(x) \\ &\Leftrightarrow x \in \partial f^*(x^*) \\ &\Leftrightarrow x \in \arg\max_{y\in E} \langle x^*, y\rangle - f(y) \\ &\Leftrightarrow x^* \in \arg\max_{y^*\in E^*} \langle y^*, x\rangle - f^*(y^*). \end{aligned}$$

In particular, this proves that ∂f and ∂f^* are inverses of each other.

A.3 Duality of strict convexity and differentiability

Definition 16 (Essential smoothness)**.** *A convex function f is essentially smooth if it is differentiable on the interior of its domain, and $\|\nabla f(x)\| \to \infty$ as x tends to the boundary of the domain.*

Definition 17 (Essential strict convexity)**.** *A convex function f is essentially strictly convex if it is strictly convex on all convex subsets where it is subdifferentiable.*

By Theorem 25.3 in [117], we have the following duality result:

Theorem 26. *Let f be closed proper convex. Then f is essentially strictly convex if and only if f^* is essentially smooth.*

A.4 Strong convexity and smoothness

Definition 18 (Smoothness)**.** *A convex function is L-smooth with respect to the norm $\|\cdot\|$ if it is differentiable and for all $x, y \in E$*

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

Definition 19 (Strong convexity)**.** *A convex function is ℓ-strongly convex with respect to the norm $\|\cdot\|$ if it is differentiable and for all $x, y \in E$*

$$f(x) \ge f(y) + \langle \nabla f(y), x - y\rangle + \frac{\ell}{2}\|y - x\|^2.$$

Next, we prove some properties which are used throughout the thesis.

Lemma 14. *Let f be an convex function, and suppose that f is L-smooth w.r.t. $\|\cdot\|$. Then for all x, x', x^+,*

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

Proof. Since f is L-smooth, we have

$$f(x^+) \leq f(x) + \left\langle \nabla f(x), x^+ - x \right\rangle + \frac{L}{2}\|x^+ - x\|^2$$

and by convexity of f,

$$f(x') \geq f(x) + \langle \nabla f(x), x' - x \rangle$$

Summing the two inequalities, we obtain the result. □

Appendix B

Mirror Operators and Bregman divergences

B.1 Dual distance generating functions and the mirror operator $\nabla\psi^*$

The dynamics studied in Part II of the thesis rely on the construction of a mirror operator $\nabla\psi^*$ which satisfies a certain number of properties. Consider a constrained optimization problem

$$\min_{x\in\mathcal{X}} f(x),$$

where $\mathcal{X}$ is a closed, convex subset of $E=\mathbb{R}^n$, and f is a differentiable, closed proper convex function. We defined Nemirovski's mirror descent dynamics (8.2) following Chapter III in [98] as follows:

$$\text{MD}\begin{cases}\dot{Z}(t) = -\nabla f(X(t)) \\ X(t) = \nabla\psi^*(Z(t)) \\ \nabla\psi^*(Z(0)) = x_0\end{cases}$$

where $X\in E$ is a primal variable, and $Z\in E^*$ is a dual variable, which are related using the mirror operator $X=\nabla\psi^*(Z)$. For the dynamics to be well-posed, we require the following properties:

1. ψ^* is differentiable on all of $E^*=\mathbb{R}^n$.
2. $\nabla\psi^*$ maps to $\mathcal{X}$.

This ensures that the dual variable Z can evolve in the unconstrained dual space, and that $X=\nabla\psi^*(Z)$ is well-defined, and remains in the primal feasible set $\mathcal{X}$.

We now discuss how to obtain such an operator $\nabla\psi^*$. Consider a pair of conjugate convex functions ψ,ψ^* such that ψ is closed and proper, and the effective domain of ψ is $\mathcal{X}$. We denote $\mathcal{X}^*$ the effective domain of ψ^*. Since ψ and ψ^* are proper convex functions, each is

subdifferentiable on the relative interior of its effective domain (Theorem 23.4 in [117]). And if we denote $\partial\psi(x)$ the subdifferential of ψ at x, then we have, by Theorem 25

$$\partial\psi^*(z) = \arg\max_{x\in\mathbb{R}^n} \langle z, x\rangle - \psi(x),$$

and since $\operatorname{dom}\psi = \mathcal{X}$, we have that $\partial\psi^*(z) \subset \mathcal{X}$. Thus we have a set-valued function $\partial\psi^*(\cdot)$ which maps E^* into $\mathcal{X}$, and whenever $\partial\psi^*(x^*)$ contains a single point, it reduces to $\nabla\psi^*(x^*)$. Thus, to satisfy the mirror operator properties, it is sufficient for ψ^* to be differentiable on all of E^*. The following proposition gives a necessary and sufficient condition in terms of properties of ψ.

Definition 20. *A convex function ψ is cofinite if its epigraph does not contain any non-vertical half-line.*

Proposition 20. *Let ψ, ψ^* be a pair of proper convex closed functions which are conjugates of each other. Then ψ^* is finite and differentiable on all of E^* if and only if ψ is essentially strictly convex and cofinite.*

Proof. By Theorem 13.3 in [117], $\operatorname{dom}\psi^* = E^*$ if and only if ψ is cofinite. And by Theorem 26, ψ^* is essentially smooth if and only if ψ is essentially strictly convex. But when $\operatorname{dom}\psi^* = E^*$, essential smoothness and differentiability are equivalent. Therefore,

$$\begin{aligned}\psi^* \text{ is finite and differentiable on } E^* &\Leftrightarrow \operatorname{dom}\psi^* = E^* \text{ and } \psi^* \text{ is essentially smooth}\\ &\Leftrightarrow \psi \text{ is cofinite and } \psi \text{ is essentially strictly convex.}\end{aligned}$$

□

This defines a general way to construct mirror operators which satisfy the desired properties: given a closed convex set $\mathcal{X}$, choose a function ψ that is essentially strictly convex and cofinite, and take ψ^* to be its convex conjugate. Then ψ^* is differentiable and we have an explicit characterization of the mirror operator: for all $z \in E^*$,

$$\nabla\psi^*(z) = \arg\max_{x\in\mathcal{X}} \langle z, x\rangle - \psi(x). \tag{B.1}$$

Note that the conditions of Proposition 20 are satisfied whenever ψ is strongly convex, or when ψ is strictly convex and $\mathcal{X}$ is bounded. In general, ψ need not be differentiable, even though this assumption is often made to simplify the discussion (for example in Chapter 4 in [33], Chapter 11 in [40], and Section 2.7 in [126]). In fact, differentiability of ψ is restrictive: By definition, ψ is differentiable at x if and only if there exists z such that $\lim_{\|x'-x\|\to 0} \frac{\psi(x')-\psi(x)-\langle z, x'-x\rangle}{\|x'-x\|} = 0$; in particular, ψ can only be differentiable on the interior of $\mathcal{X}$ since ψ needs to be finite in a neighborhood of x for the limit to be 0. Therefore, if $\mathcal{X}$ has empty interior, ψ is nowhere differentiable. This was previously observed for example by [126], who argues that the negative entropy function restricted to the simplex is

non-differentiable, and that the usual mirror descent analysis does not apply to this case. We will see that differentiability of ψ is not required, and that one can define Bregman divergences for non-differentiable ψ.

Finally, note that we required ψ^* to be differentiable on all of E^* since in the general case, the dynamics of the dual variable $\dot{Z} = -\nabla f(X)$ can evolve anywhere in E^*. However, for some problems, one may have a particular structure of ∇f which guarantees that Z remains in a subset of E^*. For example, suppose that there exists a convex cone $\mathcal{K}$ such that $\nabla f(x) \in \mathcal{K}$ for all $x \in \mathcal{X}$. Then Z remains in $-\mathcal{K}$, and it suffices that ψ^* is differentiable on $-\mathcal{K}$, not necessarily all of E^*.

B.2 Bregman divergences

Next, we define the Bregman divergences generated by the functions ψ and ψ^*. Suppose that ψ^* is differentiable on E^*. Then for $(z, z') \in E^* \times E^*$, let

$$D_{\psi^*}(z, z') = \psi^*(z) - \psi^*(z') - \langle \nabla\psi^*(z'), z - z' \rangle .$$

In words, $D_{\psi^*}(z, z')$ measures the distance between the convex function $\psi^*(z)$ and its linear approximation around z', given by $\psi^*(z') + \langle \nabla\psi^*(z'), z - z' \rangle$, and by convexity of ψ^*, the Bregman divergence is non-negative, and convex in its first argument.

One can similarly define $D_\psi(x, x')$ for $x \in E$ and points x' at which ψ is differentiable. However, as noted in the previous section, if $\mathcal{X}$ (the effective domain of ψ) has empty interior, ψ is, strictly speaking, nowhere differentiable. However, as we will see in the next proposition, if ψ can be written as the restriction to $\mathcal{X}$ of a function that is differentiable on $\operatorname{ri}\mathcal{X}$, then the Bregman divergence can still be unambiguously defined whenever $x' \in \operatorname{ri}\mathcal{X}$. Recall that the convex indicator of a convex set $\mathcal{X}$, denoted by $\delta_\mathcal{X}$, is defined as follows:

$$\delta_\mathcal{X}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{X}, \\ +\infty & \text{otherwise.} \end{cases}$$

In what follows, we will make the following assumption:

Assumption 5. *ψ can be written as $\psi = \Psi + \delta_\mathcal{X}$, where $\delta_\mathcal{X}$ is the convex indicator of $\mathcal{X}$, and Ψ is convex and differentiable on* $\operatorname{ri}\mathcal{X}$.

We will denote A the affine hull of $\mathcal{X}$ (i.e. the smallest affine space containing $\mathcal{X}$), and N the subspace of normal vectors to A.

$$N = \{n \in E : \langle n, x - x' \rangle = 0 \ \forall x, x' \in A\}.$$

Proposition 21 (Characterization of $\partial\psi$)**.** *Suppose that ψ is of the form given in Assumption 5. Then ψ is subdifferentiable on* $\operatorname{ri}\mathcal{X}$*, and $\forall x \in \operatorname{ri}\mathcal{X}$,*

$$\partial\psi(x) = \nabla\Psi(x) + N.$$

Furthermore, for all $(x, x') \in \mathcal{X} \times \operatorname{ri} \mathcal{X}$ *and all* $z \in \partial\psi(x')$,

$$\psi(x') - \psi(x) - \langle z, x' - x \rangle = \psi(x') - \psi(x) - \langle \nabla\Psi(x'), x' - x \rangle , \tag{B.2}$$

Since the expression (B.2) does not depend on the choice of $z \in \partial\psi(x')$, we can define the Bregman divergence for all $(x, x') \in \mathcal{X} \times \operatorname{ri} \mathcal{X}$ as

$$D_\psi(x, x') = \psi(x') - \psi(x) - \langle \partial\psi(x'), x' - x \rangle ,$$

which is defined unambiguously.

Proof. First, by additivity of the subdifferentials (Theorem 23.8 in [117]), we have for $x \in \operatorname{ri} \mathcal{X}$,

$$\begin{aligned} \partial\psi(x) &= \partial\Psi(x) + \partial\delta_{\mathcal{X}}(x) \\ &= \nabla\Psi(x) + \partial\delta_{\mathcal{X}}(x), \end{aligned}$$

since Ψ is differentiable on $\operatorname{ri} \mathcal{X}$. The subdifferential of $\delta_{\mathcal{X}}$ at $x \in \operatorname{ri} \mathcal{X}$ is simply the subspace N of vectors that are normal to the affine hull of $\mathcal{X}$, which proves the claim.

To prove that the Bregman divergence is defined unambiguously, let $z \in \partial\psi(x)$. Then $\exists u \in N$ such that $z = \nabla\Psi(x) + u$, and

$$\begin{aligned} \psi(x') - \psi(x) - \langle z, x' - x \rangle &= \psi(x') - \psi(x) - \langle \nabla\Psi(x) + u, x' - x \rangle \\ &= \psi(x') - \psi(x) - \langle \nabla\Psi(x), x' - x \rangle \end{aligned}$$

since u is normal to the affine hull of $\mathcal{X}$. □

Next, we have the following characterization of the subdifferential of ψ^*.

Proposition 22 (Characterization of $\nabla\psi^*$ and $\nabla^2\psi^*$)**.** *Suppose that* ψ *satisfies Assumption 5, and that* ψ^* *is differentiable on* E^**. Then for all* $z \in E^*$*, and for all* $u \in N$*,*

$$\nabla\psi^*(z + u) = \nabla\psi^*(z).$$

Furthermore, if ψ^* *is twice differentiable on* E^**, then for all* $x \in \operatorname{ri} \mathcal{X}$ *and all* $z \in \partial\psi(x)$*,*

$$\nabla^2\psi^*(z) = \nabla^2\psi^*(\nabla\Psi(x)).$$

As a consequence, it follows that ψ^* is linear in directions that are normal to the affine hull of $\mathcal{X}$. And if ψ^* is twice differentiable, then the operator $\nabla^2\psi^* \circ \partial\psi(x)$ is defined unambiguously for $x \in \operatorname{ri} \mathcal{X}$ and does not depend on the choice of $z \in \partial\psi(x)$.

Proof. To prove the first part, let $z \in E^*$ and $u \in N$. Then

$$\begin{aligned} x = \nabla\psi^*(z) &\Leftrightarrow z \in \partial\psi(x) && \text{by Theorem 25} \\ &\Leftrightarrow z \in \partial\psi(x) - u && \text{since } \partial\psi(x) = \nabla\Psi(x) + N \\ &\Leftrightarrow z + u \in \partial\psi(x) \\ &\Leftrightarrow x = \partial\psi^*(z + u) && \text{by Theorem 25 again.} \end{aligned}$$

If ψ^* is twice differentiable, and $z \in \partial\psi(x)$, then $z = \nabla\Psi(x) + u$ for some $u \in N$, and

$$\nabla^2\psi^*(z) = \nabla^2\psi^*(\nabla\Psi(x) + u) = \nabla^2\psi^*(\nabla\Psi(x)),$$

where the last equality follows simply from the fact that $\nabla\psi^*(\nabla\Psi(x) + u)$ coincides with $\nabla\psi^*(\nabla\Psi(x))$, thus so do their differentials. □

Finally, we have the following identity that relates the dual Bregman divergences.

Proposition 23 (Duality of Bregman divergences)**.** *Suppose that ψ satisfies Assumption 5 and that ψ^* is differentiable on E^*. Then for all $u, v \in E^*$, we have*

$$D_{\psi^*}(u, v) = D_\psi(\check{v}, \check{u})$$

where $\check{u} = \nabla\psi^*(u)$ and $\check{v} = \nabla\psi^*(v)$.

Proof. Using the characterization of subdifferentials in Theorem 25, we have

$$\check{u} \in \nabla\psi^*(u) \Leftrightarrow u \in \partial\psi(\check{u}) \Leftrightarrow \psi(\check{u}) + \psi^*(u) = \langle u, \check{u}\rangle,$$

so that

$$\begin{aligned} D_{\psi^*}(u, v) &= \psi^*(u) - \psi^*(v) - \langle\nabla\psi^*(v), u - v\rangle \\ &= [\langle u, \check{u}\rangle - \psi(\check{u})] - [\langle v, \check{v}\rangle - \psi(\check{v})] - \langle\check{v}, u - v\rangle \\ &= \psi(\check{v}) - \psi(\check{u}) - \langle u, \check{v} - \check{u}\rangle \\ &= D_\psi(\check{v}, \check{u}), \end{aligned}$$

where the last inequality uses the definition of the primal Bregman divergence and the fact that $u \in \partial\psi(\check{u})$. □

Lemma 15 (Bregman identity)**.** *For all u, v, w*

$$D_{\psi^*}(u, v) - D_{\psi^*}(w, v) = -D_{\psi^*}(w, u) + \langle\nabla\psi^*(u) - \nabla\psi^*(v), u - w\rangle.$$

Proof. By definition of the Bregman divergence, we have

$$\begin{aligned} &D_{\psi^*}(u, v) - D_{\psi^*}(w, v) \\ &= \psi^*(u) - \psi^*(v) - \langle\nabla\psi^*(v), u - v\rangle - (\psi^*(w) - \psi^*(v) - \langle\nabla\psi^*(v), w - v\rangle) \\ &= \psi^*(u) - \psi^*(w) - \langle\nabla\psi^*(v), u - w\rangle \\ &= -(\psi^*(w) - \psi^*(u) - \langle\nabla\psi^*(u), w - u\rangle) + \langle\nabla\psi^*(u) - \nabla\psi^*(v), u - w\rangle \\ &= -D_{\psi^*}(w, u) + \langle\nabla\psi^*(u) - \nabla\psi^*(v), u - w\rangle \end{aligned}$$

□

Lemma 16 (Bounds on a smooth Bregman divergence)**.** *Suppose that ψ^* is a dual distance generating function that is differentiable on E^*, and such that $\nabla\psi^*$ is L_{ψ^*} Lipschitz. Then for all $u, v \in E^*$,*

$$\frac{1}{2L_{\psi^*}}\|\check{u} - \check{v}\|^2 \leq D_{\psi^*}(u, v) \leq \frac{L_{\psi^*}}{2}\|u - v\|_*^2$$

where $\check{u} = \nabla\psi^*(u)$ and $\check{v} = \nabla\psi^*(v)$.

Proof. We have

$$\begin{aligned}
D_{\psi^*}(u, v) &= \psi^*(u) - \psi^*(v) - \langle \nabla\psi^*(v), u - v\rangle \\
&= \int_0^1 \nabla \langle \psi^*(v + t(u - v)) - \nabla\psi^*(v), u - v\rangle \, dt \\
&\leq \|u - v\|_* \int_0^1 \|\psi^*(v + t(u - v)) - \nabla\psi^*(v)\| dt \quad \text{by the Cauchy-Schwartz inequality} \\
&\leq L_{\psi^*}\|u - v\|_* \int_0^1 \|v + t(u - v) - v\|_* dt \qquad \text{since } \psi^* \text{ is } L_{\psi^*}\textit{Lipschitz} \\
&= L_{\psi^*}\|u - v\|_*^2 \int_0^1 t dt
\end{aligned}$$

which proves the second inequality. The first inequality will be proved by dualizing this inequality. Fix $v \in E^*$ and define

$$\begin{aligned}
h(u) &= D_{\psi^*}(u, v) = \psi^*(u) - \psi^*(v) - \langle \nabla\psi^*(v), u - v\rangle \,, \\
d(u) &= \frac{L_{\psi^*}}{2}\|u - v\|_*^2.
\end{aligned}$$

Then by the previous inequality, $h(u) \leq d(u)$ for all $u \in E^*$, and taking duals, we have $h^*(u^*) \geq d^*(u^*)$ for all u^*. We now derive the duals. Let $\check{v} = \psi^*(v)$. Then,

$$\begin{aligned}
h^*(u^*) &= \sup_u \langle u^*, u\rangle - h(u) \\
&= \sup_u \langle u^*, u\rangle - \psi^*(u) + \psi^*(v) + \langle \check{v}, u - v\rangle \\
&= \psi^*(v) - \langle v, \check{v}\rangle + \sup_u \langle u^* + \check{v}, u\rangle - \psi^*(u) \\
&= \psi^*(v) - \langle v, \check{v}\rangle + \psi(u^* + \check{v}),
\end{aligned}$$

and

$$\begin{aligned} d^*(u^*) &= \sup_u \langle u^*, u\rangle - d(u) \\ &= \sup_u \langle u^*, u\rangle - \frac{L_{\psi^*}}{2}\|u-v\|_*^2 \\ &= \sup_w \langle u^*, v+w\rangle - \frac{L_{\psi^*}}{2}\|w\|_*^2 \\ &= \langle u^*, v\rangle + \sup_w \langle u^*, w\rangle - \frac{L_{\psi^*}}{2}\|w\|_*^2 \\ &= \langle u^*, v\rangle + \frac{1}{2L_{\psi^*}}\|u^*\|^2, \end{aligned}$$

where the last equality uses Cauchy-Schwartz. Therefore combining the two inequalities,

$$\psi^*(v) - \langle v, u^* + \check{v}\rangle + \psi(u^* + \check{v}) \geq \frac{1}{2L_{\psi^*}}\|u^*\|^2.$$

In particular, for all $u \in E^*$, if we call $\check{u} = \nabla\psi^*(u)$, and take $u^* = \check{u} - \check{v}$, then

$$\psi^*(v) - \langle v, \check{u}\rangle + \psi(\check{u}) \geq \frac{1}{2L_{\psi^*}}\|\check{u} - \check{v}\|^2,$$

and by Theorem 23.5 in Rockafellar, $\psi(\check{u}) = \langle u, \check{u}\rangle - \psi^*(\check{u})$, thus

$$\psi^*(v) - \psi^*(u) - \langle \check{u}, v - u\rangle \geq \frac{1}{2L_{\psi^*}}\|\check{u} - \check{v}\|^2.$$

which proves the claim. □

B.3 Mirror update and Bregman projection

In this section, we draw a connection between the mirror operator and the Bregman divergence. Many first-order methods for convex optimization and online learning can be formulated as a sequence of mirror updates, which maintains a dual variable $z^{(k)}$ that accumulates dual vectors $\ell^{(k)}$ (in convex optimization, $\ell^{(k)}$ is a subgradient of the objective function at the current point, and in online learning, $\ell^{(k)}$ is the loss function at the current iteration), and a primal variable $x^{(k)} = \nabla\psi^*(z^{(k)})$, obtained by applying the mirror operator to the dual variable. Consider for example the constrained convex minimization problem, $\min_{x\in\mathcal{X}} f(x)$. By discretizing the mirror descent dynamics given in ODE (8.2), using a sequence of step sizes (η_k), we obtain the discrete mirror descent method proposed by Nemirovski and Yudin in [98] (see also [47, 41, 23]) summarized in Algorithm 12.

Algorithm 12 Mirror descent method with learning rates (η_k) and mirror operator ∇_{ψ^*}.

1: **for** $\tau \in \mathbb{N}$ **do**
2: Query a dual vector $\ell^{(k)}$
3: Perform a mirror update

$$\begin{cases} z^{(k+1)} = z^{(k)} - \eta_k \ell^{(k)}, \\ x^{(k+1)} = \nabla\psi^*(z^{(k+1)}). \end{cases} \tag{B.3}$$

4: **end for**

And using the characterization of $\nabla\psi^*$, given in (B.1), we can write the primal update as follows:

$$\begin{aligned} x^{(k+1)} &= \nabla\psi^*(z^{(k+1)}) \\ &= \nabla\psi^*\left(z^{(k)} - \eta_k \ell^{(k)}\right) \\ &= \arg\min_{x\in\mathcal{X}} \psi(x) - \left\langle z^{(k)} - \eta_k \ell^{(k)}, x \right\rangle \\ &= \arg\min_{x\in\mathcal{X}} \eta_k \left\langle \ell^{(k)}, x \right\rangle + \psi(x) - \psi(x^{(k)}) - \left\langle z^{(k)}, x - x^{(k)} \right\rangle, \end{aligned}$$

where in the last equality, we only added constant terms which do not depend on x. As a result, we see that the primal update can be written as a minimization problem involving a Bregman divergence. Indeed, if ψ is differentiable, then $x^{(k)} = \nabla\psi^*(z^{(k)})$ if and only $z^{(k)} = \nabla\psi(x^{(k)})$, thus

$$\psi(x) - \psi(x^{(k)}) - \left\langle z^{(k)}, x - x^{(k)} \right\rangle = D_\psi(x, x^{(k)}).$$

More generally, if the primal distance generating function can be written in the form of Assumption 5, as the restriction to $\mathcal{X}$ of a differentiable function Ψ, then $x^{(k)} = \nabla\psi^*(z^{(k)})$ if and only if $z^{(k)} \in \partial\psi(x^{(k)})$, and we have by Proposition 21 that the Bregman divergence is unambiguously defined, and

$$\psi(x) - \psi(x^{(k)}) - \left\langle z^{(k)}, x - x^{(k)} \right\rangle = D_\psi(x, x^{(k)}).$$

In both cases, the mirror descent method can be written equivalently in Algorithm 13

Algorithm 13 Primal form of the mirror descent method

1: **for** $\tau \in \mathbb{N}$ **do**
2: Query a dual vector $\ell^{(k)}$
3: Perform a mirror update (in the primal space only)

$$x^{(k+1)} = \arg\min_{x\in\mathcal{X}} \eta_k \left\langle \ell^{(k)}, x \right\rangle + D_\psi(x, x^{(k)}). \tag{B.4}$$

4: **end for**

This primal form of the mirror descent update can be interpreted as performing a Bregman projection, since it minimizes the sum of a linear function and a Bregman divergence. In the case of convex optimization problems, $\ell^{(k)}$ is a subgradient of f at $x^{(k)}$, and the mirror update (B.4) can be interpreted, as observed by Beck and Teboulle in [15], as minimizing a local approximation of the function around the current point, since

$$\arg\min_{x\in\mathcal{X}} \eta_k \left\langle \ell^{(k)}, x \right\rangle + D_\psi(x, x^{(k)} = \arg\min_{x\in\mathcal{X}} \eta_k (f(x^{(k)}) + \left\langle \ell^{(k)}, x - x^{(k)} \right\rangle) + D_\psi(x, x^{(k)}),$$

where the first term $f(x^{(k)}) + \left\langle \ell^{(k)}, x - x^{(k)} \right\rangle$ is a linear approximation of the function around the current iterate, and the Bregman divergence term $D_\psi(x, x^{(k)})$ penalizes deviations from the current iterate. The step size η_k trades-off the two terms.

Example: projected gradient descent

Projected gradient descent can be obtained as a special case of mirror descent: Let $\Psi(x) = \frac{1}{2}\|x\|_2^2$, and suppose that ψ is the restriction of Ψ to the feasible set $\mathcal{X}$, i.e. $\psi(x) = \Psi(x) + \delta_{\mathcal{X}}(x)$. The Bregman divergence associated to ψ is

$$\begin{aligned} D_\psi(x, y) &= \psi(x) - \psi(y) - \langle \nabla\Psi(y), x - y \rangle \\ &= \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y, x - y \rangle \\ &= \frac{1}{2}\|x - y\|_2^2, \end{aligned}$$

and the mirror descent update (B.4) becomes

$$\begin{aligned} x^{(k+1)} &= \arg\min_{x\in\mathcal{X}} \eta_k \left\langle \ell^{(k)}, x \right\rangle + \frac{1}{2}\|x - x^{(k)})\|^2 \\ &= \arg\min_{x\in\mathcal{X}} \frac{1}{2}\|x - (x^{(k)} - \eta_k \left\langle \ell^{(k)}, x \right\rangle)\|^2 \end{aligned}$$

which is the projection, in the Euclidean norm, of the vector $x^{(k)} - \eta_k \ell^{(k)}$, which corresponds to the projected gradient descent update.

B.4 Entropy projection on the positive orthant

Let $\mathcal{X}$ be the positive orthant $\mathcal{X} = \mathbb{R}_+^n$, and consider the negative (generalized) entropy $\psi(x) = -H(x) = \sum_{i=1}^n x_i \ln x_i$. Then ψ is differentiable on the interior of $\mathcal{X}$, $\nabla\psi(x) = (1 + \ln x_i)_i$, and a simple calculation shows that $D_\psi(x, x') = \sum_{i=1}^n x_i \ln \frac{x_i}{x'_i} - \sum_{i=1}^n (x_i - x'_i)$, the generalized I-divergence of x to x'.

Writing the definition of ψ^*, we have

$$\psi^*(z) = \sup_{x\in\mathbb{R}_+^n} \langle z, x \rangle - \sum_i x_i \ln x_i.$$

The maximization can be solved explicitly by writing the Lagrangian of the problem: for $\lambda \in \mathbb{R}^n_+$, let $L(x, \lambda) = \langle z, x\rangle - \sum_i x_i \ln x_i + \sum_{i=1}^n \lambda_i x_i$. Its gradient with respect to x is $z - (1 + \ln x_i)_i + \lambda$. Then by the KKT optimality conditions, x is optimal if and only if there exist $\lambda \in \mathbb{R}^n_+$ such that

$$\begin{cases} z - (1 + \ln x_i)_i + \lambda = 0 \\ x \geq 0, \\ x_i \lambda_i = 0 \ \forall i. \end{cases}$$

The first condition is equivalent to $x_i = e^{z_i + \lambda_i - 1}$, and since any solution of this form is strictly positive, the complementary slackness condition requires that $\lambda = 0$, thus the solution is simply

$$\nabla \psi^*(z) = x = \left(e^{z_i - 1}\right)_i$$

and $\psi^*(z) = \langle z, x\rangle - \psi(x) = \sum_{i=1}^n e^{z_i - 1}$, which is finite and differentiable on all of E^*.

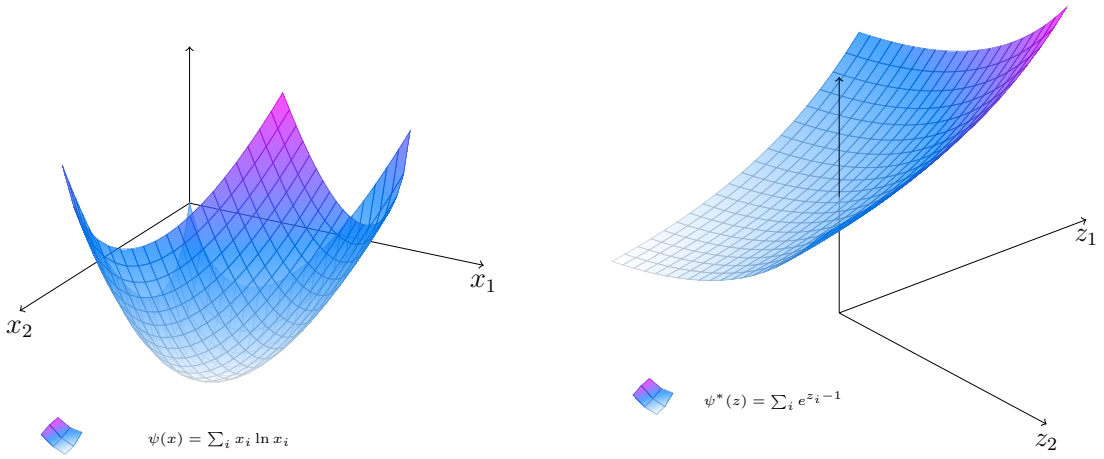


Figure B.1: Illustration of the generalized negative entropy function $\psi(x) = -H(x)$, and its conjugate $\psi^*(z) = \sum_{i=1}^n e^{z_i - 1}$.

B.5 Itakura-Saito divergence on the positive orthant

Let $\mathcal{X}$ be the positive orthant $\mathcal{X} = \mathbb{R}^n_+$, and let $\psi(x) = -\sum_{i=1}^n \ln x_i$. Then $\nabla \psi(x) = \left(-\frac{1}{x_i}\right)_i$, and a simple calculation shows that $D_\psi(x, y) = \sum_{i=1}^n \left(\frac{x_i}{y_i} - \ln \frac{x_i}{y_i} - 1\right)$, the Itakura-Saito divergence of x and y.

Writing the expression of ψ^*, we have

$$\psi^*(z) = \sup_{x \in \mathbb{R}^n_+} \langle z, x\rangle + \sum_{i=1}^n \ln x_i,$$

it is finite on $\mathcal{X}^* = \mathbb{R}^n_-$. The maximization can be solved using the same approach as in the previous example. Define the Lagrangian, for $\lambda \in \mathbb{R}^n_+$, $L(x, \lambda) = \langle z, x\rangle + \sum_{i=1}^n \ln x_i + \sum_{i=1}^n \lambda_i x_i$. Its gradient with respect to x is $z + \left(\frac{1}{x_i}\right)_i + \lambda$, and x is optimal if and only if there exists $\lambda \in \mathbb{R}^n_+$ such that

$$\begin{cases} z + \left(\frac{1}{x_i}\right)_i + \lambda = 0 \\ x \geq 0 \\ \lambda_i x_i = 0, \end{cases}$$

and the first condition can be rewritten as $x_i = \frac{1}{z_i + \lambda_i}$. Since any solution of this form is non-zero, the complementary slackness condition requires that $\lambda = 0$, and the first condition becomes $x_i = -\frac{1}{z_i}$. Therefore

$$\nabla \psi^*(z) = \left(-\frac{1}{z_i}\right)_i$$

and simple calculation shows that $\psi^*(z) = \langle z, x\rangle + \sum_{i=1}^n \ln x_i = -\sum_{i=1}^n [1 + \ln(-z_i)]$, defined on $\mathcal{X}^* = \mathbb{R}^n_-$.

B.6 Entropy projection on the simplex and the Hedge algorithm

Let $\mathcal{X}$ be the probability simplex on $\mathbb{R}^n$, i.e. $\mathcal{X} = \Delta = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$, and let ψ be the negative entropy $-H$ restricted to Δ. Formally, $\psi(x) = -H(x) + \delta_\Delta(x)$.

We have $-\nabla H(x) = (1 + \ln x_i)_i$, and by Proposition 21, for all $x, x' \in \Delta \times \operatorname{ri} \Delta$,

$$D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \Psi(y), x - y\rangle = \sum_i x_i \ln \frac{x_i}{y_i}$$

i.e. the Kullback Leibler divergence between the distribution vectors x, x'. Similarly to the previous section, we can write the definition of ψ^*,

$$\psi^*(z) = \max_{x \in \Delta} \langle x, z\rangle - \sum_{i=1}^n x_i \ln x_i,$$

and solve the maximization problem by writing the Lagrangian: for $\mu \in \mathbb{R}$ and $\lambda \in \mathbb{R}^n_+$, let $L(x, \nu, \mu) = \langle x, z\rangle - \sum_{i=1}^n x_i \ln x_i + \nu(\sum_{i=1}^n x_i - 1) + \sum_i \lambda_i x_i$. Its gradient with respect to x is $z - (1 + \ln x_i)_i - \boldsymbol{\nu} + \lambda$. Then by the KKT optimality conditions, x is optimal if and only if there exist $\lambda \in \mathbb{R}^n_+$ and ν such that

$$\begin{cases} z - (1 + \ln x_i)_i - \boldsymbol{\nu} + \lambda = 0 \\ x \geq 0,\ \sum_{i=1}^n x_i = 1 \\ x_i \lambda_i = 0\ \forall i. \end{cases}$$

The first condition can be rewritten $x_i = e^{z_i+\lambda_i}/e^{\nu+1}$. Thus the third condition (complementary slackness), requires λ_i to be 0, and the expression of x simplifies to $x_i = e^{z_i}/e^{\nu+1}$. Finally, the primal feasibility condition $\sum_{i=1}^n x_i = 1$ requires that $\sum_{i=1}^n e^{z_i}/e^{\nu+1} = 1$. Therefore, the unique solution of the maximization problem is

$$\nabla\psi^*(z)_i = x_i = \frac{e^{z_i}}{\sum_j e^{z_j}},$$

and simple algebra shows that $\psi^*(z) = \langle z, x\rangle - \psi(x) = \ln\sum_{i=1}^n e^{z_i}$, differentiable on all of $E^* = \mathbb{R}^n$. Note that we can verify the invariance of $\nabla\psi^*$ in Proposition 22: if u is a normal vector to the affine hull of Δ, i.e. $u = \alpha\mathbf{1}$ for some scalar α, then $\nabla\psi^*(z) = \nabla\psi^*(z+u)$.

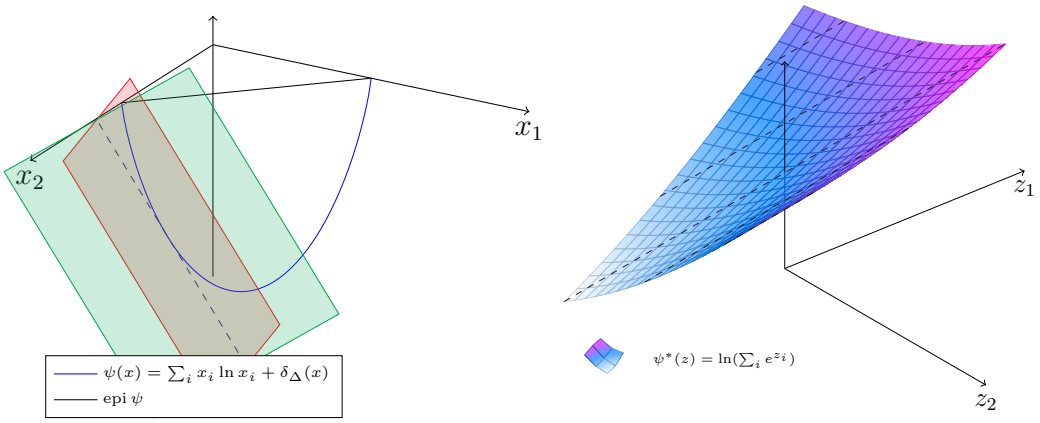


Figure B.2: Illustration of the negative entropy function restricted to the simplex $\psi(x) = -H(x) + \delta_\Delta(x)$, and its conjugate $\psi^*(z) = \ln(\sum_{i=1}^n e^{z_i})$. The function ψ is subdifferentiable on the interior of Δ, but nowhere differentiable. The figure illustrates this fact by showing two supporting hyperplanes at the same point. The conjugate function ψ^* is linear in the direction normal to the simplex (shown in dashed lines on the right).

B.7 Csiszár potentials on the simplex

Let $\mathcal{X}$ be the probability simplex on $\mathbb{R}^n$. We define a class of distance generating functions which exhibit a certain symmetry, as follows:

Definition 21 (Csiszàr potential)**.** *Let $\omega \le 0$. An increasing, C^1-diffeomorphism $\phi : \mathbb{R} \to (\omega, +\infty)$ such that $\int_0^1 \phi^{-1}(u)du < \infty$ is called a Csiszàr potential.*

Note that Audibert et al. [9] introduce a similar definition, which they call ω-potential, in which the domain of ϕ could be a subset of $\mathbb{R}$. We require, in our definition, that ϕ be defined on all of $\mathbb{R}$ to ensure that ψ^* is differentiable, as discussed below.

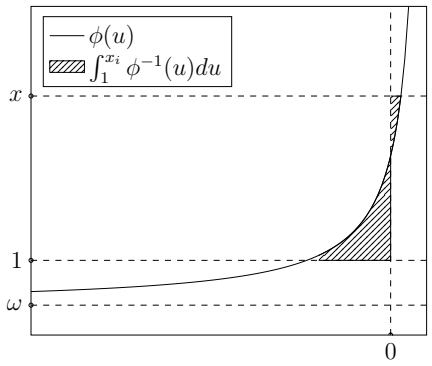


Figure B.3: Illustration of a Csiszàr potential

Definition 22 (Csiszàr distance generating function)**.** *Let ϕ be a Csiszàr potential. The distance generating function associated to ϕ is the function defined on the simplex*

$$\psi(x) = \Psi(x) + \delta_\Delta(x)$$

where

$$\begin{aligned} \Psi : \mathbb{R}^n_+ &\to \mathbb{R} \\ x &\mapsto \sum_{i=1}^n \int_1^{x_i} \phi^{-1}(u)du. \end{aligned}$$

This can be viewed as a generalization of the entropy function of the previous section. The entropy can be obtained as a special case by taking the exponential potential $\phi(u) = e^{u-1}$, then $\Psi(x) = \sum_{i=1}^n \int_1^{x_i} 1 + \ln u du = \sum_{i=1}^n x_i \ln x_i$.

By definition, Ψ is finite (in particular, the condition $\int_1^0 \phi^{-1}(u)du < \infty$ on the potential ensures that Ψ is finite on the boundary of the simplex), differentiable on $\mathbb{R}^n_{++}$, and its gradient is given by

$$\begin{aligned} \nabla\Psi : \mathbb{R}^n_{++} &\to \mathbb{R}^n \\ x &\mapsto \nabla\Psi(x) = (\phi^{-1}(x_i))_{i=1,\dots,n}. \end{aligned}$$

Since ϕ^{-1} in strictly increasing, ψ is strictly convex. We have by Proposition 21, for all $x, x' \in \Delta \times \operatorname{ri}\Delta$,

$$\begin{aligned} D_\psi(x,y) &= \psi(x) - \psi(y) - \langle \nabla\Psi(y), x - y\rangle \\ &= \sum_{i=1}^n \int_1^{x_i} \phi^{-1}(u)du - \int_1^{y_i} \phi^{-1}(u)du - \phi^{-1}(y_i)(x_i - y_i) \\ &= \sum_{i=1}^n \int_{y_i}^{x_i} (\phi^{-1}(u) - \phi^{-1}(y_i))du. \end{aligned} \tag{B.5}$$

By definition of ψ^*,

$$\psi^*(z) = \max_{x\in\Delta} \langle x, z\rangle - \sum_{i=1}^n \int_1^{x_i} \phi^{-1}(u)du.$$

Unlike in the previous example, there is no closed-form solution in general. However, we can give a simple necessary and sufficient condition of optimality.

Proposition 24 (Optimality conditions for Csiszàr mirror operators)**.** *Let ψ be the distance generating function associated to a Csiszàr potential ϕ. Then ψ^* is differentiable on $\mathbb{R}^n$, and for all $z \in E^*$,*

$$x = \nabla\psi^*(z) \Leftrightarrow \begin{cases} \exists \nu \in \mathbb{R} : \forall i,\ x_i = (\phi(z_i + \nu))_+ \\ \sum_{i=1}^n x_i = 1. \end{cases} \tag{B.6}$$

Proof. Since ϕ^{-1} is strictly increasing, ψ is strictly convex. And since Δ, the effective domain of ψ is bounded, the epigraph of ψ does not contain any non-vertical half-lines, thus ψ is cofinite. Therefore, by Proposition 20, ψ^* is differentiable on all of E^*.

Let $z \in E^*$. Recall from equation B.1 that $\nabla\psi^*(z) = \arg\max_{x\in\Delta} \langle x, z\rangle - \sum_{i=1}^n \int_1^{x_i} \phi^{-1}(u)du$. First, define the Lagrangian of the problem: for $\mu \in \mathbb{R}$ and $\lambda \in \mathbb{R}^n_+$, let

$$L(x, \nu, \mu) = \langle x, z\rangle - \sum_{i=1}^n \int_1^{x_i} \phi^{-1}(u)du - \nu\left(\sum_{i=1}^n x_i - 1\right) + \sum_i \lambda_i x_i.$$

Its gradient with respect to x is $z - (\phi^{-1}(x_i))_i - \boldsymbol{\nu} + \lambda$. Then by the KKT optimality conditions (together with Slater's condition for constraint qualification), x is optimal if and only if there exist $\lambda \in \mathbb{R}^n_+$ and ν such that

$$\begin{cases} z - (\phi^{-1}(x_i))_i - \boldsymbol{\nu} + \lambda = 0, \\ x \geq 0,\ \sum_{i=1}^n x_i = 1, \\ x_i\lambda_i = 0\ \forall i. \end{cases}$$

The first condition is equivalent to $x_i = \phi(z_i + \lambda_i - \nu)$. Let $\mathcal{I} = \{i : x_i > 0\}$ be the support of an optimal point. Then by the complementary slackness condition $x_i\lambda_i = 0$, we have for all $i \in \mathcal{I}$, $\lambda_i = 0$, thus $x_i = \phi(z_i + \nu)$, and for all $i \notin \mathcal{I}$,

$$\begin{aligned} \phi(z_i + \nu) &\leq \phi(z_i + \nu + \lambda_i) && \text{since } \phi \text{ is increasing} \\ &= x_i = 0. \end{aligned}$$

Therefore x_i can be simply written $x_i = (\phi(z_i + \nu^\star))_+$ which proves the claim. □

The optimality conditions of Proposition 24 will be useful in developing efficient algorithms for computing approximate and exact Bregman projections, as discussed in the next chapter. Next, we give one family of Csiszár potentials which lead to a generalization of the entropy projection that enjoys some desirable properties.

B.8 Generalized entropy projection on the simplex and the smoothed KL divergence

Definition 23 (Exponential potential)**.** *Let $\epsilon \geq 0$. We call the function*

$$\begin{aligned} \phi_\epsilon : \mathbb{R} &\to (-\epsilon, +\infty) \\ u &\mapsto e^{u-1} - \epsilon, \end{aligned}$$

the exponential potential with parameter ϵ. It is a Csiszàr potential.

The distance generating function induced by ϕ_ϵ is given by

$$\begin{aligned} \psi(x) &= \sum_{i=1}^n \int_1^{x_i} \phi_\epsilon^{-1}(u)du \\ &= \sum_{i=1}^n \int_1^{x_i} 1 + \ln(u+\epsilon)du \\ &= \sum_{i=1}^n (x_i+\epsilon)\ln(x_i+\epsilon) - (1+\epsilon)\ln(1+\epsilon) \\ &= H(x+\boldsymbol{\epsilon}) - H(\mathbf{1}+\boldsymbol{\epsilon}), \end{aligned}$$

where $\boldsymbol{\epsilon}$ is the vector whose entries are all equal to ϵ, and H is the negative entropy function $H(x) = \sum_{i=1}^n x_i \ln x_i$. The corresponding Bregman divergence is

$$\begin{aligned} D_\psi(x,y) &= H(x+\boldsymbol{\epsilon}) - H(y+\boldsymbol{\epsilon}) - \langle \nabla H(y+\boldsymbol{\epsilon}), x-y\rangle \\ &= D_{\mathrm{KL}}(x+\boldsymbol{\epsilon}, y+\boldsymbol{\epsilon}) \\ &= \sum_{i=1}^n (x_i+\epsilon)\ln\frac{x_i+\epsilon}{y_i+\epsilon}, \end{aligned}$$

and will be denoted $D_{\mathrm{KL},\epsilon}(x,y)$. It corresponds to the KL divergence between the vectors $x+\boldsymbol{\epsilon}$ and $y+\boldsymbol{\epsilon}$, and can be thought of as a regularized KL divergence. In particular, when $\epsilon > 0$, $D_{\mathrm{KL},\epsilon}$ is finite for all $x, y \in \Delta$, unlike the KL divergence which is infinite if the support of y is not a subset of the support of x. We show some additional properties below.

Properties of the generalized KL divergence

Proposition 25. *For all $\epsilon > 0$, $D_{KL,\epsilon}$ is $\frac{1}{1+n\epsilon}$-strongly convex and $\frac{1}{\epsilon}$-smooth w.r.t. $\|\cdot\|_1$. Furthermore, it is bounded on the simplex by $\ln\frac{1+\epsilon}{\epsilon}$.*

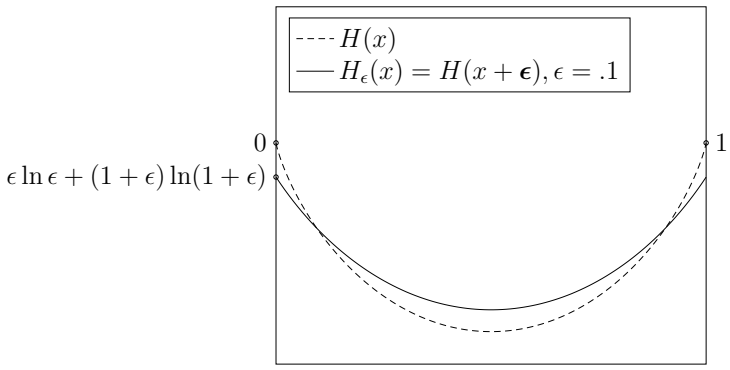


Figure B.4: Illustration of the distance generating function induced by exponential potentials with parameter ϵ, for $n = 2$: $H(x) = x_1 \ln(x_1) + (1 - x_1) \ln(1 - x_1)$.

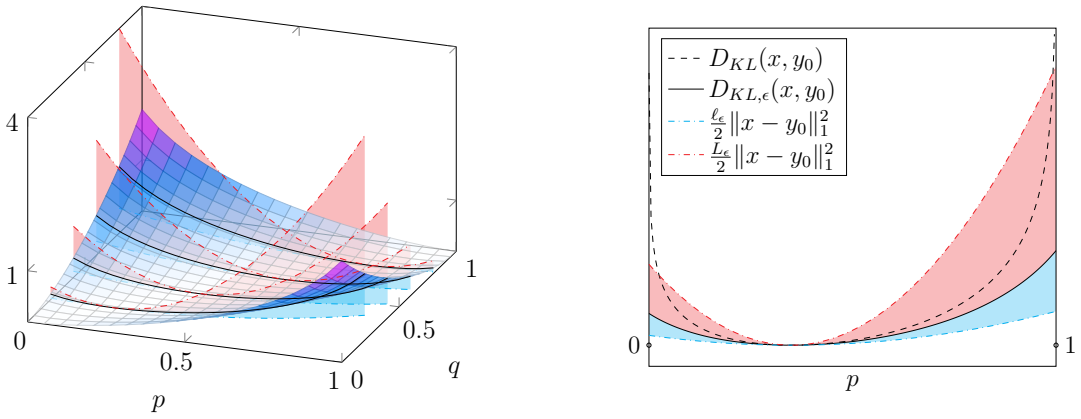


Figure B.5: Illustration of Proposition 25, when $d = 2$. The distributions x and y are parameterized as follows: $x = (p, 1 - p)$ and $y = (q, 1 - q)$. The surface plot (left) shows the generalized KL divergence for $\epsilon = .1$, with, in dashed lines, the quadratic upper and lower bounds, $\frac{\ell_\epsilon}{2}\|y - x\|_1^2$ and $\frac{L_\epsilon}{2}\|x - y\|_1^2$. The second plot (right) compares $D_{\mathrm{KL},.1}(x, y_0)$ and $D_{\mathrm{KL}}(x, y_0)$ for a fixed $y_0 = (.35, .65)$.

Proof. First, we show strong convexity and smoothness. Let $x, y \in \Delta$. By Taylor's theorem, $\exists z \in (x + \boldsymbol{\epsilon}, y + \boldsymbol{\epsilon})$ such that

$$\begin{aligned} D_{\mathrm{KL},\epsilon}(x, y) &= H(x + \boldsymbol{\epsilon}) - H(y + \boldsymbol{\epsilon}) - \langle \nabla H(y + \boldsymbol{\epsilon}), x - y \rangle \\ &= \frac{1}{2} \left\langle x - y, \nabla^2 H(z)(x - y) \right\rangle \\ &= \frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{z_i}, \end{aligned}$$

where we used the fact that the Hessian of the negative entropy function is $\nabla^2 H(z) =$

$\mathrm{diag}(\frac{1}{z_i})$. And since $\forall i,\ z_i \geq \epsilon$ (z belongs to the segment $(x+\boldsymbol{\epsilon}, y+\boldsymbol{\epsilon})$), it follows that

$$D_{\mathrm{KL},\epsilon}(x,y) \leq \frac{1}{2\epsilon}\sum_{i=1}^{n}(x_i-y_i)^2 \leq \frac{1}{2\epsilon}\|x-y\|_1^2,$$

which proves smoothness. Next, by the Cauchy-Schwartz inequality, we have

$$\left(\sum_{i=1}^{n}|x_i-y_i|\right)^2 \leq \sum_{i=1}^{n}\frac{(x_i-y_i)^2}{z_i}\sum_{i=1}^{n}z_i,$$

thus

$$D_{\mathrm{KL},\epsilon}(x,y) \geq \frac{1}{2}\frac{\|x-y\|_1^2}{\|z\|_1} = \frac{1}{2}\frac{1}{1+d\epsilon}\|x-y\|_1^2.$$

To compute the upper bound on $D_{\mathrm{KL},\epsilon}$, we observe that $D_{\mathrm{KL},\epsilon}(x,y)$ is jointly-convex in (x,y) (by joint-convexity of the KL divergence), therefore, its maximum on $\Delta^d \times \Delta^d$ is attained on a vertex of the feasible set, that is, for $(x,y) = (\delta^{i_0}, \delta^{j_0})$, for some (i_0, j_0), where δ^{i_0} is the Dirac distribution on i_0. Finally, simple calculation shows that

$$D_{\mathrm{KL},\epsilon}(\delta^{i_0}, \delta^{j_0}) = \begin{cases} 0 & \text{if } i_0 = j_0, \\ \ln\frac{1+\epsilon}{\epsilon} & \text{otherwise.} \end{cases}$$

□

Projecting on the simplex with the KL divergence plays a central role in many applications in online learning and convex optimization [34, 40, 126]. Some applications include non-parametric statistical estimation, e.g. Section 7.2 in [30], multi-commodity flow problems, e.g. Chapter 12 in [38], tomography image reconstruction [24] and learning dynamics in repeated games as discussed in Chapter 5. However, some variants of mirror descent require the Bregman divergence to be bounded on the simplex in order to have guarantees on the convergence rate, see for example [45], as well as in Chapter 5. The accelerated mirror descent algorithm that we develop in Chapter 10 also uses a mirror update using a Bregman divergence which is required to be both strongly convex and smooth. These examples motivate the use of the smoothed KL divergence.

Although the mirror operator $\nabla\psi^*$ has no closed-form solution, we will develop, in the next chapter, efficient algorithms for computing the exact projection in $\mathcal{O}(n \ln n)$ time using a deterministic sorting method, given in Algorithm 15, and in expected linear time using a randomized sorting method given in Algorithm 16.

$\mathcal{X}$	$\mathcal{X}^*$	$\psi^*(z)$	$\Psi(x)$	$\nabla\psi^*(z)$	$\nabla\Psi(x)$	$D_{\psi^*}(x,y)$
$\mathbb{R}^n_+$	$\mathbb{R}^n$	$\sum_{i=1}^n e^{z_i-1}$	$\sum_{i=1}^n x_i \ln x_i$	$(e^{z_i-1})_i$	$(1+\ln x_i)_i$	$\sum_{i=1}^n x_i \ln \frac{x_i}{y_i} - \sum_{i=1}^n (x_i - y_i)$
$\mathbb{R}^n_+$	$\mathbb{R}^n_-$	$-\sum_{i=1}^n [1+\ln(-z_i)]$	$-\sum_{i=1}^n \ln x_i$	$\left(-\frac{1}{z_i}\right)_i$	$\left(-\frac{1}{x_i}\right)_i$	$\sum_{i=1}^n \left(\frac{x_i}{y_i} - \ln \frac{x_i}{y_i} - 1\right)$
Δ	$\mathbb{R}^n$	$\ln \sum_{i=1}^n e^{z_i}$	$\sum_{i=1}^n x_i \ln x_i$	$\left(\frac{e^{z_i}}{\sum_j e^{z_j}}\right)_i$	$(1+\ln x_i)_i$	$\sum_{i=1}^n x_i \ln \frac{x_i}{y_i}$
Δ	$\mathbb{R}^n$	–	$\sum_{i=1}^n (x_i+\epsilon)\ln(x_i+\epsilon)$	*Alg.* 15, 16	$(1+\ln(x_i+\epsilon))_i$	$\sum_{i=1}^n (x_i+\epsilon)\ln \frac{x_i+\epsilon}{y_i+\epsilon}$

Table B.1: Examples of dual distance generating functions and the corresponding mirror operators and Bregman divergences.

Appendix C

Efficient Bregman Pojections on the Simplex

In Section B.3, we discussed that many algorithms for online learning and convex optimization involve repeatedly solving the mirror update equation of Algorithm 13 on the simplex, i.e.

$$x^\star = \arg\min_{x \in \Delta} \left\langle \bar{\ell}, x \right\rangle + D_\psi(x, \bar{x}). \tag{C.1}$$

where Δ is the probability simplex on $\mathbb{R}^n$, D_ψ is the Bregman divergence associated to a distance generating function ψ, $\bar{x}$ is the current iterate in the primal space, and $\bar{\ell}$ is a given dual vector, which can be either a loss function (in online learning problems) or a subgradient of the objective function at the current point (in convex optimization), scaled by a step size. We refer to problem (C.1) as the Bregman projection or the mirror update on the simplex.

Some instances of Bregman projections on the simplex are known to have an exact solution which can be computed efficiently. For example, the solution of the KL divergence projection on the simplex is given by the exponential weights update [98, 15], and the Euclidean projection on the simplex can be computed efficiently either by sorting and thresholding in $\mathcal{O}(n \log n)$, or by using a randomized pivot method in $\mathcal{O}(n)$, see [46].

In this chapter, we show that for the Csiszár potentials defined in Section B.7, the solution of the Bregman projection on the simplex can be approximated efficiently: an ϵ-approximate solution can be computed in $\mathcal{O}(n \log \frac{1}{\epsilon})$ operations. Finally, we show that for the exponential potentials defined in Section B.8, the exact solution can be computed using a deterministic algorithm with $\mathcal{O}(n \log n)$ complexity, or a randomized algorithm with expected linear complexity.

C.1 Efficient approximate projection with Csiszàr potentials

First, we consider the problem of projecting on a simplex with a Csiszàr potential, as defined in Section B.7. We derive optimality conditions for the Bregman projection problem (C.1), similar to Proposition 24. Using the expression (B.5) of the Bregman divergence, we can rewrite the problem as

$$\begin{aligned} x^\star &= \arg\min_{x\in\Delta} \left\langle \bar{\ell}, x \right\rangle + D_\psi(x, \bar{x}) \\ &= \arg\min_{x\in\Delta} \left\langle \bar{\ell}, x \right\rangle + \sum_{i=1}^n \int_{\bar{x}_i}^{x_i} (\phi^{-1}(u) - \phi^{-1}(\bar{x}_i)) du \end{aligned}$$

Proposition 26. *Let ψ be a Bregman divergence associated to Csiszár potential ϕ. Consider the Bregman projection onto the simplex given in Problem* (B.7)*. Then $x^\star$ is a solution if and only if there exists $\nu^\star \in \mathbb{R}$ such that*

$$\begin{cases} \forall i, \quad x_i^\star = \left(\phi(\phi^{-1}(\bar{x}_i) - \bar{\ell}_i + \nu^\star)\right)_+, \\ \sum_{i=1}^n x_i^\star = 1, \end{cases}$$

where x_+ denotes the positive part of x, $x_+ = \max(x, 0)$.

Proof. Define the Lagrangian of the problem: for $\mu \in \mathbb{R}$ and $\lambda \in \mathbb{R}_+^n$, let

$$L(x, \nu, \mu) = \left\langle x, \bar{\ell} \right\rangle + \sum_{i=1}^n \int_{\bar{x}_i}^{x_i} (\phi^{-1}(u) - \phi^{-1}(\bar{x}_i)) du - \nu \left(\sum_{i=1}^n x_i - 1 \right) - \sum_i \lambda_i x_i.$$

Its gradient with respect to x is $\bar{\ell} - (\phi^{-1}(x_i) - \phi^{-1}(\bar{x}_i))_i - \boldsymbol{\nu} - \lambda$. Then by the KKT optimality conditions (together with Slater's condition for constraint qualification), $x^\star$ is optimal if and only if there exist $\lambda^\star \in \mathbb{R}_+^n$ and $\nu^\star$ such that

$$\begin{cases} \bar{\ell} + (\phi^{-1}(x_i^\star) - \phi^{-1}(\bar{x}_i))_i - \boldsymbol{\nu}^\star - \lambda^\star = 0, \\ x^\star \geq 0, \ \sum_{i=1}^n x_i^\star = 1, \\ x_i^\star \lambda_i^\star = 0 \ \forall i. \end{cases}$$

The first condition is equivalent to $x_i^\star = \phi(\phi^{-1}(\bar{x}_i) - \bar{\ell}_i + \lambda_i^\star + \nu^\star)$. Let $\mathcal{I} = \{i : x_i^\star > 0\}$ be the support of $x^\star$. Then by the complementary slackness condition, we have for all $i \in \mathcal{I}$, $\lambda_i^\star = 0$, thus $x_i^\star = \phi(\phi^{-1}(\bar{x}_i) + \bar{\ell}_i + \nu)$, and for all $i \notin \mathcal{I}$,

$$\begin{aligned} \phi(\phi^{-1}(\bar{x}_i) - \bar{\ell}_i + \nu^\star) &\leq \phi(\phi^{-1}(\bar{x}_i) - \bar{\ell}_i + \nu^\star + \lambda_i^\star) \qquad \text{since } \phi \text{ is increasing} \\ &= x_i^\star = 0. \end{aligned}$$

Therefore $x_i^\star$ can be simply written $x_i^\star = \left(\phi(\phi^{-1}(\bar{x}_i) - \bar{\ell}_i + \nu^\star)\right)_+$ which proves the claim. □

Next, we make the following observation regarding the support of the solution:

Proposition 27. *Let $x^\star$ be the solution to the projection problem* (C.1)*, and let $\mathcal{I}$ be its support. Then for all i, j, if $i \in \mathcal{I}$ and $\phi^{-1}(\bar{x}_i) - \bar{\ell}_i \leq \phi^{-1}(\bar{x}_j) - \bar{\ell}_j$, then $j \in \mathcal{I}$.*

Proof. Follows from Proposition 26 and the fact that ϕ is increasing. □

As a consequence of the previous propositions, computing the projection reduces to computing the optimal dual variable $\nu^\star$, and since the potential is increasing, one can iteratively approximate $\nu^\star$ using a bisection method, given in Algorithm 14: we start by defining a bound on the optimal $\nu^\star$, $\underline{\nu} \leq \nu^\star \leq \bar{\nu}$, then we iteratively halve the size of the interval by inspecting the value of a carefully defined criterion function.

Algorithm 14 Bisection method to approximate the Bregman projection with precision ϵ.

Input: $\bar{x}, \bar{\ell}, \epsilon$.
Initialize
  $\bar{\nu} = \phi^{-1}(1) - \max_i \phi^{-1}(\bar{x}_i) - \bar{\ell}_i$
  $\underline{\nu} = \phi^{-1}(1/n) - \max_i \phi^{-1}(\bar{x}_i) - \bar{\ell}_i$
Define $\tilde{x}(\nu) = \left(\phi(\phi^{-1}(\bar{x}_i) - \bar{\ell}_i + \nu)_+\right)_{i=1,\dots,n}$
while $\|\tilde{x}(\bar{\nu}) - \tilde{x}(\underline{\nu})\|_1 > \epsilon$ **do**
  Let $\nu^+ \leftarrow \frac{\bar{\nu}+\underline{\nu}}{2}$
  if $\sum_{i=1}^n \tilde{x}_i(\nu^+) > 1$ **then**
    $\bar{\nu} \leftarrow \nu^+$
  else
    $\underline{\nu} \leftarrow \nu^+$
  end if
end while
Return $\tilde{x}(\bar{\nu})$

Theorem 27. *Consider the Bregman projection onto the simplex given in Problem* (C.1)*, with Csiszár potential ϕ. Let $\epsilon > 0$, and consider the bisection method given in Algorithm 14. Then the Algorithm terminates after $T = \mathcal{O}(\log \frac{1}{\epsilon})$ steps, and its output $\tilde{x}(\bar{\nu}^{(T)})$ is such that*

$$\|\tilde{x}(\bar{\nu}^{(T)}) - x^\star\|_1 \leq \epsilon.$$

Each step of the algorithm has complexity $\mathcal{O}(n)$, thus the total complexity is $\mathcal{O}\left(d \log \frac{1}{\epsilon}\right)$.

Proof. Define, as in Algorithm 14, the function

$$\tilde{x}(\nu) = \left(\phi(\phi^{-1}(\bar{x}_i) - \bar{\ell}_i + \nu)_+\right)_{i=1,\dots,n}.$$

Since ϕ is, by assumption, increasing, so is $\nu \mapsto \tilde{x}_i(\nu)$, which is the key fact that allows us to use a bisection.

We will denote by a superscript (t) the value of each variable at iteration t of the loop. To prove the claim, we show the following invariant for t:

(i) $0 \leq \bar{\nu}^{(t)} - \underline{\nu}^{(t)} \leq \frac{\bar{\nu}^{(0)} - \underline{\nu}^{(0)}}{2^t}$,

(ii) $\forall i,\ 0 \leq \tilde{x}_i(\underline{\nu}^{(t)}) \leq \tilde{x}_i(\underline{\nu}^{(t)}) \leq 1$,

(iii) $\sum_{i=1}^n \tilde{x}_i(\underline{\nu}^{(t)}) \leq 1 \leq \sum_{i=1}^n \tilde{x}_i(\bar{\nu}^{(t)})$.

We first prove the invariant for $t = 0$. Let $i_0 = \arg\max_i \phi^{-1}(\bar{x}_i) - \bar{\ell}_i$. By definition of $\bar{\nu}^{(0)}$ and $\underline{\nu}^{(0)}$, we have

$$\phi^{-1}(1/n) - \underline{\nu} = \phi^{-1}(\bar{x}_{i_0}) - \bar{\ell}_{i_0} = \phi^{-1}(1) - \bar{\nu}, \tag{C.2}$$

and it follows that $\tilde{x}_{i_0}(\underline{\nu}^{(0)}) = \frac{1}{n}$ and $\tilde{x}_{i_0}(\bar{\nu}^{(0)}) = 1$. By (C.2), $\bar{\nu}^{(0)} - \underline{\nu}^{(0)} = \phi^{-1}(1) - \phi^{-1}(1/n) \geq 0$ (since ϕ^{-1} is increasing), which proves (i). Next, since $\nu \mapsto \tilde{x}_i(\nu)$ is increasing, we have

$$0 \leq \tilde{x}_i(\underline{\nu}^{(0)}) \leq \tilde{x}_i(\bar{\nu}^{(0)}) \leq \tilde{x}_{i_0}(\bar{\nu}^{(0)}) = 1,$$

which proves (ii). Finally, we have

$$\sum_{i=1}^n \tilde{x}_i(\underline{\nu}^{(0)}) \leq n\tilde{x}_{i_0}(\underline{\nu}^{(0)}) = 1,$$
$$\sum_{i=1}^n \tilde{x}_i(\bar{\nu}^{(0)}) \geq \tilde{x}_{i_0}(\bar{\nu}^{(0)}) = 1,$$

which proves (iii). This proves the invariant for $t = 0$. Now suppose it holds at iteration t, and let us prove it still holds at $t + 1$. By definition of the bisection (lines 5–10), we immediately have

$$\bar{\nu}^{(t+1)} - \underline{\nu}^{(t+1)} = \frac{\bar{\nu}^{(t)} - \underline{\nu}^{(t)}}{2} = \frac{1}{2}\frac{\bar{\nu}^{(0)} - \underline{\nu}^{(0)}}{2^t},$$

which proves (i). We also have that $\underline{\nu}^{(t)} \leq \underline{\nu}^{(t+1)} \leq \bar{\nu}^{(t+1)} \leq \bar{\nu}^{(t)}$, which proves (ii) since $\nu \mapsto \tilde{x}_i(\nu)$ is increasing. Finally, (iii) follows from the condition of the bisection (line 6).

To conclude the proof, we simply observe that since the distance $|\bar{\nu} - \underline{\nu}|$ decreases exponentially, the algorithm will terminate after a number of steps logarithmic in $1/\epsilon$. Indeed, since ϕ is C^1, it is Lipschitz-continuous on $[\phi^{-1}(0), \phi^{-1}(1)]$. Let L be its Lipschitz constant, then

$$\begin{aligned}
\|\tilde{x}(\underline{\nu}^{(t)}) - \tilde{x}(\bar{\nu}^{(t)})\|_1 &= \sum_{i=1}^n |\tilde{x}_i(\underline{\nu}^{(t)}) - \tilde{x}_i(\bar{\nu}^{(t)})| \\
&\leq nL|\underline{\nu}^{(t)} - \bar{\nu}^{(t)}| \\
&= \frac{nL|\underline{\nu}^{(0)} - \bar{\nu}^{(0)}|}{2^t} && \text{by (ii)},
\end{aligned}$$

thus the algorithm terminates after $T = \log_2 \frac{|\underline{\nu}^{(0)} - \bar{\nu}^{(0)}|}{\epsilon nL}$ iterations, and the last iterate satisfies

$$\begin{aligned}
&\|\tilde{x}(\nu^\star) - \tilde{x}(\bar{\nu}^{(T)})\|_1 \\
&\leq \|\tilde{x}(\underline{\nu}^{(T)}) - \tilde{x}(\bar{\nu}^{(t)})\|_1 \quad \text{by (iii) and since } \tilde{x}_i \text{ are increasing} \\
&\leq \epsilon,
\end{aligned}$$

which concludes the proof. □

C.2 Efficient exact projection with exponential potentials

We now consider a special case of Csiszár potentials, given by the exponential potentials defined in Section B.8, and show that the exact solution can be computed using a sorting algorithm. We first apply the optimality conditions of Proposition 26 to this special class, and show that the solution is entirely determined by its support.

Proposition 28. *Consider the Bregman projection onto the simplex given in Problem* (C.1)*, with Bregman divergence $D_{KL,\epsilon}$. Let $x^\star$ be the solution and $\mathcal{I} = \{i : x_i^\star > 0\}$ its support. Then*

$$\begin{cases} \forall i \in \mathcal{I}, \quad x_i^\star = -\epsilon + \dfrac{(\bar{x}_i + \epsilon)e^{-\bar{\ell}_i}}{Z^\star}, \\ Z^\star = \dfrac{\sum_{i\in\mathcal{I}}(\bar{x}_i + \epsilon)e^{-\bar{\ell}_i}}{1 + |\mathcal{I}|\epsilon}. \end{cases} \tag{C.3}$$

Proof. Applying Proposition 26 with the expression $\phi(u) = e^{u-1}+\epsilon$ and $\phi^{-1}(u) = 1+\ln(u+\epsilon)$, $x^\star$ is a solution if and only if there exists $\nu^\star \in \mathbb{R}$ such that $\forall i$, $x_i^\star = \left(-\epsilon + (\bar{x}_i + \epsilon)e^{-\bar{\ell}_i}e^{\nu^\star}\right)_+$, and $\sum_i x_i^\star = 1$. Thus, if $\mathcal{I}$ is the support of $x^\star$, then these optimality conditions are equivalent to

$$\begin{cases} \forall i \in \mathcal{I}, \quad x_i^\star = -\epsilon + (\bar{x}_i + \epsilon)e^{-\bar{\ell}_i}e^{\nu^\star}, \\ \displaystyle\sum_{i\in\mathcal{I}} -\epsilon + (\bar{x}_i + \epsilon)e^{-\bar{\ell}_i}e^{\nu^\star} = 1, \end{cases}$$

and the second equation can be rewritten as

$$1 + \epsilon|\mathcal{I}| = e^{\nu^\star}\sum_{i\in\mathcal{I}}(\bar{x}_i + \epsilon)e^{-\bar{\ell}_i},$$

which proves the claim, with $Z^\star = e^{-\nu^\star}$. □

Proposition 28 shows that solving the Bregman projection with smoothed KL divergence reduces to finding the support of the solution. Next, we show that the support has a simple characterization. To this end, we associate to $(\bar{x}, \bar{\ell})$ the vector $\bar{y}$ defined as follows

$$\forall i, \ \bar{y}_i = (\bar{x}_i + \epsilon)e^{-\bar{\ell}_i},$$

and we denote by $\bar{y}_{\sigma(i)}$ the i-th largest element of $\bar{y}$.

Algorithm 15 ExpProject: Sort based method to compute the Bregman projection with smoothed KL divergence $D_{\mathrm{KL},\epsilon}$

1: Input: $\bar{x}$, $\bar{\ell}$
2: Output: $x^\star$
3: Form the vector $\bar{y}_i = (\bar{x}_i + \epsilon)e^{-\bar{\ell}_i}$
4: Sort y, let $\bar{y}_{\sigma(i)}$ be the i-th smallest element of y.
5: Let $j^\star$ be the smallest index for which $c(j) := (1+\epsilon(n-j+1))\bar{y}_{\sigma(j)} - \epsilon\sum_{i\geq j}\bar{y}_{\sigma(i)} > 0$.
6: Set

$$\begin{cases} Z(j^\star) = \frac{\sum_{i\geq j^\star}\bar{y}_{\sigma(i)}}{1+\epsilon(n-j^\star+1)} \\ x_i^\star = \left(-\epsilon + \frac{\bar{y}_i}{Z(j^\star)}\right)_+ \end{cases}$$

Lemma 17. *The function*

$$c(j) \mapsto (1+\epsilon(n-j+1))\bar{y}_{\sigma(j)} - \epsilon\sum_{i\geq j}\bar{y}_{\sigma(i)}$$

is increasing, and the support of $x^\star$ is $\{\sigma(j^\star),\dots,\sigma(n)\}$, where $j^\star = \min\{j : c(j) > 0\}$.

Proof. First, straightforward algebra shows that

$$c(j+1) - c(j) = (1+\epsilon(n-j))(\bar{y}_{\sigma(j+1)} - \bar{y}_{\sigma(j)}) \geq 0.$$

Thus c is increasing. To prove the second part of the claim, we know by Proposition 27 that the support is $\{\sigma(i^\star),\dots,\sigma(n)\}$ for some $i^\star$, and to show that $i^\star = j^\star = \min\{j : c(j) > 0\}$, it suffices to show that $c(i^\star) > 0$ and $c(j) \leq 0$ for all $j < i^\star$. First, by the expression (C.3) of $x^\star$, we have

$$x^\star_{\sigma(i^\star)} = -\epsilon + \frac{\bar{y}_{\sigma(i^\star)}}{\frac{\sum_{i\geq i^\star}\bar{y}_{\sigma(i)}}{1+\epsilon(n-i^\star+1)}} > 0,$$

which is equivalent to $c(i^\star) > 0$. And if $j < i^\star$ (i.e. $\sigma(j)$ is outside the support), then by the expression (C.3) again,

$$0 = x^\star_{\sigma(j)} \geq -\epsilon + \frac{\bar{y}_{\sigma(j)}}{\frac{\sum_{i\geq i^\star}\bar{y}_{\sigma(i)}}{1+\epsilon(n-i^\star+1)}}$$

which is equivalent to

$$(1+\epsilon(n-i^\star-1))\bar{y}_{\sigma(j)} - \epsilon\sum_{i\geq i^\star}\bar{y}_{\sigma(i)} \leq 0,$$

but $c(j)$ is smaller than the LHS, since

$$c(j) - (1+\epsilon(n-i^\star-1))\bar{y}_{\sigma(j)} - \epsilon\sum_{i\geq i^\star}\bar{y}_{\sigma(i)} = \epsilon\sum_{j\leq i<i^\star}\bar{y}_{\sigma(j)} - \bar{y}_{\sigma(i)} \leq 0,$$

which concludes the proof. □

Theorem 28. *The ExpProject Algorithm 15 terminates after $\mathcal{O}(n \log n)$ iterations and outputs the solution $x^\star$ to the Bregman projection problem* (C.1) *with smoothed KL divergence $D_{KL,\epsilon}$.*

Proof. Correctness of the algorithm follows from the characterization of the support of $x^\star$ in Lemma 17 and the expression of $x^\star$ in Proposition 28. The complexity of the sort operation (step 4) is $\mathcal{O}(n \log n)$, and finding $j^\star$ (step 5) can be done in linear time since the criterion function $c(\cdot)$ is such that $c(j+1) - c(j) = (1 + \epsilon(n-j))(\bar{y}_{\sigma(j+1)} - \bar{y}_{\sigma(j)})$, so each criterion evaluation costs $\mathcal{O}(1)$. Therefore, the overall complexity of Algorithm 15 is $\mathcal{O}(n \log n)$. □

C.3 A randomized pivot algorithm with expected linear time

We now propose a randomized version of Algorithm 15, which selects a random pivot at each iteration, instead of sorting the full vector. The resulting algorithm, which we call QuickProject, is an extension of the QuickSelect algorithm due to Hoare [65]. A similar idea is used in the randomized version of the ℓ_2 projection on the simplex in [46].

Theorem 29. *In expectation, the QuickProject Algorithm terminates after $\mathcal{O}(n)$ operations, and outputs the solution $x^\star$ of the Bregman projection problem* (C.1) *with the smoothed KL divergence $D_{KL,\epsilon}$.*

Proof. First, we prove that the algorithm has expected linear complexity. Let $T(n)$ be the expected complexity of the while loop when $|\mathcal{J}| = n$.

The partition and compute step (7) takes $3n$ operations, then we recursively apply the loop to $\mathcal{J}^-$ or $\mathcal{J}^+$, which have sizes $(m, n-m)$ for any $m \in \{1, \dots, n\}$, with uniform probability. Thus we can bound $T(n)$ as follows

$$\begin{aligned} T(n) &\leq 3n + \frac{1}{n} \sum_{m=1}^{n} T(\max(m, n-m)) \\ &\leq 3n + \frac{2}{n} \sum_{m=\frac{n}{2}}^{n} T(m), \end{aligned}$$

and we can show by induction that $T(n) \leq 12n$, since $T(0) = 0$ and

$$3n + \frac{2}{n} \sum_{m=\frac{n}{2}}^{n} 12m \leq 3n + 12\frac{3n}{4} = 12n.$$

To prove the correctness of the algorithm, we will prove that once the while loop terminates, $s^\star = \sigma(j^\star)$, and S, C are respectively the sum and the cardinality of $\{\bar{y}_{\sigma(i)} : i \geq j^\star\}$, then by Proposition 28, we have the correct expression of $x^\star$. We start by showing the following invariants:

Algorithm 16 QuickExpProject: Randomized pivot based method to compute the Bregman projection with $D_{\mathrm{KL},\epsilon}$.

1: Input: $\bar{x}$, $\bar{\ell}$
2: Output: $x^\star$
3: Form the vector $\bar{y}_i = (\bar{x}_i + \epsilon)e^{-\bar{\ell}_i}$
4: Initialize $\mathcal{J} = \{1, \ldots, n\}$, $S = 0$, $C = 0$, $s^\star = n + 1$
5: **while** $\mathcal{J} \neq \emptyset$ **do**
6: Select a random pivot index $j \in \mathcal{J}$
7: Partition $\mathcal{J}$ into $\mathcal{J}^+ = \{i \in \mathcal{J} : \bar{y}_i \geq \bar{y}_j\}$ and $\mathcal{J}^- = \{i \in \mathcal{J} : \bar{y}_i < \bar{y}_j\}$.
8: Compute $S^+ = \sum_{i \in \mathcal{J}^+} \bar{y}_i$ and $C^+ = |\mathcal{J}^+|$.
9: Let $\gamma = (1 + \epsilon(C + C^+))\bar{y}_j - \epsilon(S + S^+)$
10: **if** $\gamma > 0$ **then**
11: $\mathcal{J} \leftarrow \mathcal{J}^-$, $s^\star = j$
12: $S \leftarrow S + S^+$, $C \leftarrow C + C^+$
13: **else**
14: $\mathcal{J} \leftarrow \mathcal{J}^+$
15: **end if**
16: **end while**
17: Set

$$\begin{cases} Z = \frac{S}{1+\epsilon C} \\ x_i^\star = \left(-\epsilon + \frac{\bar{y}_i}{Z}\right)_+ \end{cases}$$

(i) If $\bar{y}_{\sigma(m_t)}$, is the largest element in $\mathcal{J}^{(t)}$, then $\sigma(m_t + 1) = (s^\star)^{(t)}$.

(ii) $\mathcal{J}^{(t)}$ contains $\sigma(j^\star)$ or $\sigma(j^\star - 1)$.

(iii) S and C are the sum and cardinality of $\{i : \sigma(i) \geq s^\star\}$.

(iv) $\gamma^{(t)} = c(j^{(t)})$, where c is the criterion function defined in Lemma 17.

The invariant holds for the first iteration since $\mathcal{J}^{(1)} = \{1, \ldots, n\}$, $m_t = d$, and $S^{(1)} = C^{(1)} = 0$. Suppose the invariant is true at iteration t of the loop. Then two cases are possible:

1. If $\gamma^{(t)} \leq 0$, then $\mathcal{J}^{(t+1)} = (\mathcal{J}^{(t)})^+$ and $m^{(t+1)} = m^{(t)}$, and the invariant still holds.
2. If $\gamma^{(t)} > 0$, then $\mathcal{J}^{(t+1)} = (\mathcal{J}^{(t)})^-$ and $(s^\star)^{(t+1)} = j^{(t)}$, thus

$$\begin{aligned} \{i : \sigma(i) \geq (s^\star)^{t+1}\} &= \{i : \sigma(i) \geq (s^\star)^{(t)}\} \cup \{i : (s^\star)^{t+1} \leq \sigma(i) \leq (s^\star)^{(t)} - 1\} \\ &= \{i : \sigma(i) \geq (s^\star)^{(t)}\} \cup (\mathcal{J}^{(t)})^+, \end{aligned}$$

and by the update step (lines 10–11), the invariant still holds.

To finish the proof, suppose the while loop terminates after T iterations, i.e. $\mathcal{J}^{(T+1)} = \emptyset$. We claim that $(s^\star)^{(T+1)} = \sigma(j^\star)$. During the last update, two cases are possible:

1. If $\gamma^{(T)} > 0$, then $\bar{y}_{j^{(T)}}$ is the smallest element of $\mathcal{J}^{(T)}$. In this case, since $c(i) \leq 0$ for $i < j^\star$, and $\mathcal{J}^{(T)}$ contains $\sigma(j^\star)$ or $\sigma(j^\star - 1)$, it must be that $j^{(T)} = \sigma(j^\star)$, thus
$$(s^\star)^{T+1} = j^{(T)} = \sigma(j^\star).$$
2. If $\gamma^{(T)} \leq 0$, then $\bar{y}_{j^{(T)}}$ is the largest element of $\mathcal{J}^{(T)}$, in this case, since $c(j^\star) > 0$, it must be that $j^{(T)} = \sigma(j^\star - 1)$, so $m^{(t)} = j^\star - 1$ and
$$(s^\star)^{(T+1)} = (s^\star)^{(T)} = \sigma(m^{(t)} + 1) = \sigma(j^\star).$$

This concludes the proof. □

This proves that the Bregman projection problem (C.1) with smoothed KL divergence can be solved exactly in expected $\mathcal{O}(n)$ time. A question which remains open is whether it can be solved in $\mathcal{O}(n)$ time using a *deterministic* algorithm, akin to the "median of medians" algorithm due to Blum et al. [27] which solves the selection problem in deterministic linear time.

C.4 Numerical experiments

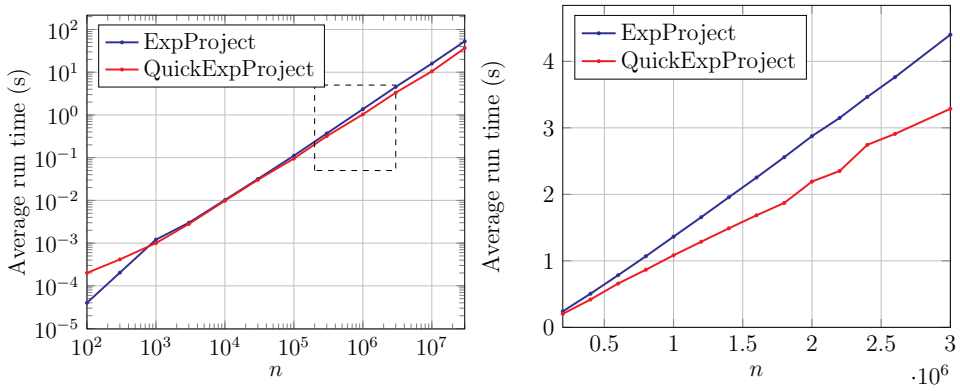


Figure C.1: Run time as a function of the dimension n, with $\epsilon = .1$, in log-log scale (left). The highlighted region is zoomed-in in linear scale on the right.

We provide a simple python implementation of the projection algorithms at `github.com/walidk/BregmanProjection`. The implementation of Algorithm 14 is generic and can be instantiated for any Csiszár potential by providing the function ϕ and its inverse. The implementation of Algorithm 15 and QuickProject are specific to the exponential potential. Finally, we report in Figure C.1 the run times of both algorithms as the dimension n grows, averaged over 50 runs, for randomly generated, normally distributed vectors $\bar{x}$ and $\bar{g}$. The numerical simulations are also available on the same repository.

Bibliography

[1] J. Abernethy, P. Bartlett, and E. Hazan. "Blackwell Approachability and no-regret learning are equivalent". In: *Journal of Machine Learning Research* 19 (2011), pp. 27–46.

[2] Z. Allen-Zhu and L. Orecchia. "Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent". In: *CoRR* abs/1407.1537 (2014).

[3] F. Alvarez and H. Attouch. "An Inertial Proximal Method for Maximal Monotone Operators via Discretization of a Nonlinear Oscillator with Damping". In: *Set-Valued Analysis* 9.1-2 (2001), pp. 3–11.

[4] S. Arora, E. Hazan, and S. Kale. "The Multiplicative Weights Update Method: a Meta-Algorithm and Applications." In: *Theory of Computing* 8.1 (2012), pp. 121–164.

[5] G. Arslan and J. Shamma. "Distributed convergence to Nash equilibria with local utility measurements". In: *43rd IEEE Conference on Decision and Control*. Vol. 2. Dec. 2004, 1538–1543 Vol.2.

[6] H. Attouch, J. Peypouquet, and P. Redont. "Fast Convergence of an Inertial Gradient-like System with Vanishing Viscosity". In: *CoRR* abs/1507.04782 (2015).

[7] H. Attouch, J. Peypouquet, and P. Redont. "A Dynamical Approach to an Inertial Forward-Backward Algorithm for Convex Minimization". In: *SIAM Journal on Optimization* 24.1 (2014), pp. 232–256.

[8] J.-P. Aubin. *Viability Theory*. Cambridge, MA, USA: Birkhauser Boston Inc., 1991.

[9] J.-Y. Audibert, S. Bubeck, and G. Lugosi. "Regret in Online Combinatorial Optimization". In: *Mathematics of Operations Research* 39.1 (2014), pp. 31–45.

[10] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. "Clustering with Bregman Divergences". In: *J. Mach. Learn. Res.* 6 (Dec. 2005), pp. 1705–1749.

[11] P. Bartlett, E. Hazan, and A. Rakhlin. "Adaptive online gradient descent". In: *Advances in Neural Information Processing Systems (NIPS)*. Montreal, Canada, 2008.

[12] T. Basar and G. Olsder. *Dynamic Noncooperative Game Theory: Second Edition*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1999.

[13] A. Bayen, J. Butler, A. Patire, CCIT, UC Berkeley ITS, and California Dpartment of Transportation, Division of Research and Innovation. *Mobile Millennium Final Report*. 2011.

[14] A. Beck and M. Teboulle. "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems". In: *SIAM Journal on Imaging Sciences* 2.1 (2009), pp. 183–202.

[15] A. Beck and M. Teboulle. "Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization". In: *Operations Research Letters* 31.3 (May 2003), pp. 167–175.

[16] M. J. Beckmann, C. B. McGuire, and C. B. Winsten. *Studies in the Economics of Transportation*. Yale University Press, 1955.

[17] M. Benaïm. "A Dynamical System Approach to Stochastic Approximations". In: *SIAM Journal on Control and Optimization* 34.2 (1996), pp. 437–472.

[18] M. Benaïm. "Dynamics of stochastic approximation algorithms". In: *Séminaire de probabilités XXXIII*. Springer, 1999, pp. 1–68.

[19] M. Benaïm. "On Gradient Like Properties of Population Games, Learning Models and Self Reinforced Processes". In: *Dynamics, Games and Science: International Conference and Advanced School Planet Earth, DGS II, Portugal, August 28–September 6, 2013*. Ed. by J.-P. Bourguignon, R. Jeltsch, A. A. Pinto, and M. Viana. Cham: Springer International Publishing, 2015, pp. 117–152.

[20] M. Benaïm and M. W. Hirsch. "Asymptotic pseudotrajectories and chain recurrent flows, with applications". In: *Journal of Dynamics and Differential Equations* 8.1 (1996), pp. 141–176.

[21] M. Benaïm, J. Hofbauer, and S. Sorin. "Stochastic Approximations and Differential Inclusions". In: *SIAM Journal on Control and Optimization* 44.1 (2005), pp. 328–348.

[22] M. Benaïm, J. Hofbauer, and S. Sorin. "Stochastic Approximations and Differential Inclusions, Part II: Applications". In: *Mathematics of Operations Research* 31.4 (2006), pp. 673–695.

[23] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization*. Society for Industrial and Applied Mathematics, 2001.

[24] A. Ben-Tal, T. Margalit, and A. Nemirovski. "The Ordered Subsets Mirror Descent Optimization Method with Applications to Tomography". In: *SIAM Journal on Optimization* 12.1 (Jan. 2001), pp. 79–108.

[25] D. Blackwell. "An analog of the minimax theorem for vector payoffs." In: *Pacific Journal of Mathematics* 6.1 (1956), pp. 1–8.

[26] A. Bloch, ed. *Hamiltonian and Gradient Flows, Algorithms, and Control*. American Mathematical Society, 1994.

[27] M. Blum, R. W. Floyd, V. Pratt, R. L. Rivest, and R. E. Tarjan. "Time Bounds for Selection". In: *Journal of Computer and System Sciences* 7.4 (Aug. 1973), pp. 448–461.

[28] L. E. Blume. "The Statistical Mechanics of Strategic Interaction". In: *Games and Economic Behavior* 5.3 (1993), pp. 387–424.

[29] L. Bottou. "Online Algorithms and Stochastic Approximations". In: *Online Learning and Neural Networks*. Ed. by D. Saad. revised, oct 2012. Cambridge, UK: Cambridge University Press, 1998.

[30] S. Boyd and L. Vandenberghe. *Convex Optimization*. Vol. 25. Cambridge University Press, 2010.

[31] A. Bressan and K. Han. "Optima and Equilibria for a Model of Traffic Flow". In: *SIAM Journal on Mathematical Analysis* 43.5 (2011), pp. 2384–2417.

[32] A. A. Brown and M. C. Bartholomew-Biggs. "Some Effective Methods for Unconstrained Optimization Based on the Solution of Systems of Ordinary Differential Equations". In: *Journal of Optimization Theory and Applications* 62.2 (1989), pp. 211–224.

[33] S. Bubeck. "Convex Optimization: Algorithms and Complexity". In: *Foundations and Trends in Machine Learning* 8.3-4 (2015), pp. 231–357.

[34] S. Bubeck and N. Cesa-Bianchi. "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems". In: *Foundations and Trends in Machine Learning* 5.1 (2012), pp. 1–122.

[35] S. Bubeck, V. Perchet, and P. Rigollet. "Bounded regret in stochastic multi-armed bandits". In: *CoRR* abs/1302.1611 (2013).

[36] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, Ltd, 2008.

[37] C. Canudas De Wit, F. Morbidi, L. Leon Ojeda, A. Y. Kibangou, I. Bellicot, and P. Bellemain. "Grenoble Traffic Lab: An experimental platform for advanced traffic monitoring and forecasting". In: *IEEE Control Systems* 35.3 (June 2015), pp. 23–39.

[38] Y. Censor and S. Zenios. *Parallel Optimization: Theory, Algorithms and Applications*. Oxford University Press, 1997.

[39] N. Cesa-Bianchi and G. Lugosi. "Potential-Based Algorithms in On-Line Prediction and Game Theory". In: *Machine Learning* 51.3 (2003), pp. 239–261.

[40] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[41] G. Chen and M. Teboulle. "Convergence Analysis of a Proximal-Like Minimization Algorithm Using Bregman Functions". In: *SIAM Journal on Optimization* 3.3 (1993), pp. 538–543.

[42] I. Csiszár. "Information-type measures of difference of probability distributions and indirect observations". In: *Studia Scientiarum Mathematicarum Hungarica* 2 (1967), pp. 299–318.

[43] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. "Optimal Distributed Online Prediction". In: *Proceedings of the 28th International Conference on Machine Learning (ICML)*. June 2011.

[44] B. Drighès, W. Krichene, and A. Bayen. "Stability of Nash Equilibria in the Congestion Game under Replicator Dynamics". In: *53rd IEEE Conference on Decision and Control (CDC)*. Los Angeles, CA, 2014, pp. 1923–1929.

[45] J. C. Duchi, A. Agarwal, M. Johansson, and M. Jordan. "Ergodic Mirror Descent". In: *SIAM Journal on Optimization (SIOPT)* 22.4 (2010), pp. 1549–1578.

[46] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. "Efficient Projections Onto the L1-ball for Learning in High Dimensions". In: *25th International Conference on Machine Learning (ICML)*. Helsinki, Finland: ACM, 2008, pp. 272–279.

[47] J. Eckstein. "Nonlinear Proximal Point Algorithms Using Bregman Functions, with Applications to Convex Programming". In: *Mathematics of Operations Research* 18.1 (1993), pp. 202–226.

[48] L. C. Evans. *An Introduction to Mathematical Optimal Control Theory*.

[49] S. Fischer and B. Vöcking. "On the Evolution of Selfish Routing". In: *Algorithms–ESA 2004*. Springer, 2004, pp. 323–334.

[50] N. Flammarion and F. R. Bach. "From Averaging to Acceleration, There is Only a Step-size". In: *Proceedings of The 28th Conference on Learning Theory (COLT), Paris, France*. 2015, pp. 658–695.

[51] M. J. Fox and J. S. Shamma. "Population Games, Stable Games, and Passivity". In: *Games* 4.4 (2013), pp. 561–583.

[52] D. H. Fremlin. *Measure theory*. Vol. 4. Torres Fremlin, 2000.

[53] Y. Freund and R. E. Schapire. "Adaptive Game Playing Using Multiplicative Weights". In: *Games and Economic Behavior* 29.1 (1999), pp. 79–103.

[54] D. Fudenberg and D. K. Levine. *The theory of learning in games*. Vol. 2. MIT press, 1998.

[55] M. Garavello and B. Piccoli. *Traffic Flow on Networks: Conservation Laws Models*. AIMS series on applied mathematics. American Institute of Mathematical Sciences, 2006.

[56] M. B. Giles and N. A. Pierce. "An Introduction to the Adjoint Approach to Design". In: *Flow, Turbulence and Combustion* 65.3-4 (2000), pp. 393–415.

[57] J. Hannan. "Approximation to Bayes risk in repeated plays". In: *Contributions to the Theory of Games* 3 (1957). Ed. by M. Dresher, A. W. Tucker, and P. Wolfe, pp. 97–139.

[58] S. Hart. "Adaptive Heuristics". In: *Econometrica* 73.5 (2005), pp. 1401–1430.

[59] S. Hart and A. Mas-Colell. "A General Class of Adaptive Strategies". In: *Journal of Economic Theory* 98.1 (2001), pp. 26–54.

[60] S. Hart and A. Mas-Colell. "A Simple Adaptive Procedure Leading to Correlated Equilibrium". In: *Econometrica* 68.5 (2000), pp. 1127–1150.

[61] S. Hart and A. Mas-Colell. "Regret-based continuous-time dynamics". In: *Games and Economic Behavior* 45.2 (2003). Special Issue in Honor of Robert W. Rosenthal, pp. 375–394.

[62] S. Hart and A. Mas-Colell. *Simple Adaptive Strategies: From Regret-matching to Uncoupled Dynamics*. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2012.

[63] E. Hazan, A. Agarwal, and S. Kale. "Logarithmic regret algorithms for online convex optimization". In: *Machine Learning* 69.2-3 (2007), pp. 169–192.

[64] U. Helmke and J. Moore. *Optimization and dynamical systems*. Communications and control engineering series. Springer-Verlag, 1994.

[65] C. A. R. Hoare. "Algorithm 65: Find". In: *Communications of the ACM* 4.7 (July 1961), pp. 321–322.

[66] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.

[67] J. Hofbauer and K. Sigmund. *The Theory of Evolution and Dynamical Systems: Mathematical Aspects of Selection*. London Mathematical Society Student Texts. Cambridge University Press, 1988.

[68] J. Hofbauer and W. H. Sandholm. "Stable games and their dynamics". In: *Journal of Economic Theory* 144.4 (2009), 1665–1693.e4.

[69] A. Juditsky. *Convex Optimization II: Algorithms, Lecture Notes*. 2013.

[70] A. Juditsky, A. Nemirovski, and C. Tauvel. "Solving Variational Inequalities with Stochastic Mirror-Prox Algorithm". In: *Stochastic Systems* 1.1 (2011), pp. 17–58.

[71] H. Khalil. *Nonlinear systems*. Macmillan Pub. Co., 1992.

[72] J. Kivinen and M. K. Warmuth. "Exponentiated Gradient versus Gradient Descent for Linear Predictors". In: *Information and Computation* 132.1 (1997), pp. 1–63.

[73] R. Kleinberg, G. Piliouras, and E. Tardos. "Multiplicative Updates Outperform Generic No-Regret Learning in Congestion Games". In: *Proceedings of the 41st annual ACM Symposium on Theory of Computing*. ACM. 2009, pp. 533–542.

[74] Y. A. Korilis, A. A. Lazar, and A. Orda. "Capacity Allocation under Noncooperative Routing". In: *IEEE Transactions on Automatic Control* 42 (1997), pp. 309–325.

[75] E. Koutsoupias and C. Papadimitriou. "Worst-Case Equilibria". In: *In proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science*. 1999, pp. 404–413.

[76] S. Krichene, W. Krichene, R. Dong, and A. Bayen. "Convergence of Heterogeneous Distributed Learning in Stochastic Routing Games". In: *53rd Annual Allerton Conference on Communication, Control and Computing*. Monticello, IL, 2015.

[77] W. Krichene, A. Bayen, and P. Bartlett. "Accelerated Mirror Descent in Continuous and Discrete Time". In: *29th Annual Conference on Neural Information Processing Systems (NIPS)*. Montreal, Canada, 2015.

[78] W. Krichene, A. Bayen, and P. Bartlett. "Adaptive Averaging in Accelerated Descent Dynamics". In: *30th Annual Conference on Neural Information Processing Systems (NIPS)*. Barcelona, Spain, 2016.

[79] W. Krichene, B. Drighès, and A. Bayen. "On the Convergence of No-Regret Learning in Selfish Routing". In: *31st International Conference on Machine Learning (ICML)*. Beijing, China, 2014, pp. 163–171.

[80] W. Krichene, B. Drighès, and A. Bayen. "Online Learning of Nash Equilibria in Congestion Games". In: *SIAM Journal on Control and Optimization (SICON)* 53.2 (2015), pp. 1056–1081.

[81] W. Krichene, S. Krichene, and A. Bayen. "Convergence of Mirror Descent dynamics in the Routing Game". In: *European Control Conference (ECC)*. Linz, Austria, 2015.

[82] W. Krichene, S. Krichene, and A. Bayen. "Efficient Bregman Projections onto the Simplex". In: *54th IEEE Conference on Decision and Control (CDC)*. Osaka, Japan, 2015.

[83] P. R. Kumar and P. Varaiya. *Stochastic Systems: Estimation, Identification and Adaptive Control*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1986.

[84] K. Lam, W. Krichene, and A. Bayen. "On Learning How Players Learn: Estimation of Learning Dynamics in the Routing Game". In: *7th International Conference on Cyber-Physical Systems (ICCPS)*. 2016.

[85] A. Lew, J. E. Marsden, M. Ortiz, and M. West. "Variational Time Integrators". In: *International Journal for Numerical Methods in Engineering* 60.1 (2004), pp. 153–212.

[86] M. J. Lighthill and G. B. Whitham. "On Kinematic Waves. II. A Theory of Traffic Flow on Long Crowded Roads". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 229.1178 (1955), pp. 317–345.

[87] J. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, 1971.

[88] A. Lyapunov. *General Problem of the Stability Of Motion*. Control Theory and Applications Series. Taylor & Francis, 1992.

[89] J. R. Marden. "Regret Based Dynamics: Convergence in Weakly Acyclic Games". In: *In Proceedings of the 2007 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 2007.

[90] J. R. Marden, S. D. Ruben, and L. Y. Pao. "A Model-Free Approach to Wind Farm Control Using Game Theoretic Methods". In: *Control Systems Technology, IEEE Transactions on* 21.4 (2013), pp. 1207–1214.

[91] J. R. Marden and J. S. Shamma. "Revisiting Log-Linear Learning: Asynchrony, Completeness and Payoff-Based Implementation". In: *Games and Economic Behavior* 75.2 (2012), pp. 788–808.

[92] J. Marden and J. Shamma. "Game Theory and Distributed Control". In: *Handbook of Game Theory Vol. 4*. Ed. by H. Young and S. Zamir. Elsevier Science, 2013.

[93] J. E. Marsden and M. West. "Discrete Mechanics and Variational Integrators". In: *Acta Numerica* 10 (2001), pp. 357–514.

[94] M. Muresan. *A Concrete Approach to Classical Analysis*. Springer, 2009, pp. 85–86.

[95] J. Nash. "Non-Cooperative Games". In: *The Annals of Mathematics* 54.2 (1951), pp. 286–295.

[96] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. "Robust Stochastic Approximation Approach to Stochastic Programming". In: *SIAM Journal on Optimization* 19.4 (2009), pp. 1574–1609.

[97] A. Nemirovski. "Prox-Method with Rate of Convergence O(1/t) for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-Concave Saddle Point Problems". In: *SIAM Journal on Optimization* 15.1 (2004), pp. 229–251.

[98] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience series in discrete mathematics. Wiley, 1983.

[99] Y. Nesterov. "Accelerating the cubic regularization of Newton's method on convex problems". In: *Mathematical Programming* 112.1 (2008), pp. 159–181.

[100] Y. Nesterov. "Gradient methods for minimizing composite functions". In: *Mathematical Programming* 140.1 (2013), pp. 125–161.

[101] Y. Nesterov. "Smooth minimization of non-smooth functions". In: *Mathematical Programming* 103.1 (2005), pp. 127–152.

[102] Y. Nesterov. "A method of solving a convex programming problem with convergence rate $O(1/k^2)$". In: *Soviet Mathematics Doklady* 27.2 (1983), pp. 372–376.

[103] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Vol. 87. Springer Science & Business Media, 2004.

[104] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory*. New York, NY, USA: Cambridge University Press, 2007.

[105] B. O'Donoghue and E. Candès. "Adaptive Restart for Accelerated Gradient Schemes". In: *Foundations of Computational Mathematics* 15.3 (2015), pp. 715–732.

[106] A. Ozdaglar and R. Srikant. "Incentives and pricing in communication networks". In: *Algorithmic Game Theory* (2007), pp. 571–591.

[107] V. Perchet. "Exponential weight approachability, applications to calibration and regret minimization". In: *Dynamic Games and Applications* 5.1 (2015), pp. 136–153.

[108] J. Peypouquet and S. Sorin. "Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time". In: *Journal of Convex Analysis* 17 (2010), pp. 1113–1163.

[109] D. Pisarski and C. Canudas-de-Wit. "Optimal balancing of road traffic density distributions for the Cell Transmission Model". In: *51st IEEE Conference on Decision and Control (CDC)*. 2012, pp. 6969–6974.

[110] E. Polak. *Optimization: Algorithms and Consistent Approximations*. New York, NY, USA: Springer-Verlag New York, Inc., 1997.

[111] B. T. Polyak and A. B. Juditsky. "Acceleration of Stochastic Approximation by Averaging". In: *SIAM Journal on Control and Optimization* 30.4 (July 1992), pp. 838–855.

[112] M. Raginsky and J. Bouvrie. "Continuous-time stochastic Mirror Descent on a network: Variance reduction, consensus, convergence". In: *IEEE Conference on Decision and Control (CDC)*. 2012, pp. 6793–6800.

[113] A. Rakhlin, O. Shamir, and K. Sridharan. "Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization". In: *CoRR* (2011).

[114] J. Reilly, S. Samaranayake, M. Delle Monache, W. Krichene, P. Goatin, and A. Bayen. "Adjoint-Based Optimization on a Network of Discretized Scalar Conservation Laws with Applications to Coordinated Ramp Metering". In: *Journal of Optimization Theory and Applications (JOTA)* 167.2 (2015), pp. 733–760.

[115] P. I. Richards. "Shock Waves on the Highway". In: *Operations Research* 4.1 (1956), pp. 42–51.

[116] H. Robbins and D. Siegmund. "A Convergence Theorem for Non Negative Almost Supermartingales and Some Applications". In: *Optimizing Methods in Statistics* (1971).

[117] R. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[118] R. W. Rosenthal. "A class of games possessing pure-strategy Nash equilibria". In: *International Journal of Game Theory* 2.1 (1973), pp. 65–67.

[119] T. Roughgarden. "Routing games". In: *Algorithmic Game Theory*. Cambridge University Press, 2007. Chap. 18, pp. 461–486.

[120] T. Roughgarden. "Stackelberg scheduling strategies". In: *SIAM Journal on Computing* 33.2 (2004), pp. 332–350.

[121] T. Roughgarden and E. Tardos. "Bounding the inefficiency of equilibria in nonatomic congestion games". In: *Games and Economic Behavior* 47.2 (2004), pp. 389–403.

[122] T. Roughgarden and É. Tardos. "How bad is selfish routing?" In: *Journal of the ACM (JACM)* 49.2 (2002), pp. 236–259.

[123] W. H. Sandholm. *Population Games and Evolutionary Dynamics*. MIT Press, 2010.

[124] W. H. Sandholm. "Potential games with continuous player sets". In: *Journal of Economic Theory* 97.1 (2001), pp. 81–108.

[125] J. Schropp and I. Singer. "A dynamical systems approach to constrained minimization". In: *Numerical Functional Analysis and Optimization* 21.3-4 (2000), pp. 537–551.

[126] S. Shalev-Shwartz. "Online Learning and Online Convex Optimization". In: *Foundations and Trends in Machine Learning* 4.2 (Feb. 2012), pp. 107–194.

[127] O. Shamir and T. Zhang. "Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes." In: *30th International Conference on Machine Learning (ICML)*. 2013, pp. 71–79.

[128] K. Sigmund. "Complexity, Language, and Life: Mathematical Approaches". In: ed. by J. L. Casti and A. Karlqvist. Berlin, Heidelberg: Springer Berlin Heidelberg, 1986. Chap. A Survey of Replicator Equations, pp. 88–104.

[129] H. A. Simon. "A Behavioral Model of Rational Choice". In: *The Quarterly Journal of Economics* 69.1 (1955), pp. 99–118.

[130] W. Su, S. Boyd, and E. Candès. "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights". In: *28th Annual Conference on Neural Information Processing Systems (NIPS)*. 2014.

[131] M. Teboulle. "Convergence of Proximal-Like Algorithms". In: *SIAM Journal on Optimization* 7.4 (Apr. 1997), pp. 1069–1083.

[132] G. Teschl. *Ordinary differential equations and dynamical systems*. Vol. 140. American Mathematical Society, 2012.

[133] J. Thai, R. Hariss, and A. Bayen. "A multi-convex approach to latency inference and control in traffic equilibria from sparse data". In: *American Control Conference (ACC)*. July 2015, pp. 689–695.

[134] J. G. Wardrop. "Some Theoretical Aspects of Road Traffic Research". In: *ICE Proceedings: Engineering Divisions*. Vol. 1. 3. 1952, pp. 325–362.

[135] J. W. Weibull. *Evolutionary Game Theory*. MIT press, 1997.

[136] A. Wibisono, A. C. Wilson, and M. I. Jordan. "A Variational Perspective on Accelerated Methods in Optimization". In: *CoRR* abs/1603.04245 (2016).

[137] M. Zinkevich. "Online Convex Programming and Generalized Infinitesimal Gradient Ascent". In: *20th International Conference on Machine Learning (ICML)*. 2003, pp. 928–936.