Designing Algorithms for Learning and Decision-Making in Societal Systems

Eric Mazumdar

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2021-166
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-166.html

July 22, 2021

Acknowledgement

This work was supported in part by HICON-LEARN (Design of HIgh CONfidence LEARNing-Enabled Systems), Defense Advanced Research Projects Agency award number FA8750-18-C-0101.

Designing Algorithms for Learning and Decision-Making in Societal Systems

by

Eric Mazumdar

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael I. Jordan, Co-chair
Professor S. Shankar Sastry, Co-chair
Professor Pravin Varaiya
Professor Shachar Kariv

Summer 2021

The dissertation of Eric Mazumdar, titled Designing Algorithms for Learning and Decision-Making in Societal Systems , is approved:

Co-chair _____ Date _____

Co-chair _____ Date _____

_____ Date _____

_____ Date _____

University of California, Berkeley

Designing Algorithms for Learning and Decision-Making in Societal Systems

Abstract

Designing Algorithms for Learning and Decision-Making in Societal Systems

by

Eric Mazumdar

Doctor of Philosophy in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Michael I. Jordan, Co-chair

Professor S. Shankar Sastry, Co-chair

The ability to learn from data and make decisions in real-time has led to the rapid deployment of machine learning algorithms across many aspects of everyday life. Despite their potential to enable new services and address persistent societal issues, the widespread use of these algorithms has led to unintended consequences like flash crashes in financial markets or price collusion on e-commerce platforms. These consequences are the inevitable result of deploying algorithms— that were designed to operate in isolation— in uncertain dynamic environments in which they interact with other autonomous agents, algorithms, and human decision makers.

To address these issues, it is necessary to develop an understanding of the fundamental limits of learning algorithms in societal-scale systems. The work in this thesis is divided into three parts, each addressing a different aspect of learning and decision-making in societal-scale systems: (i) learning in the presence of strategic agents, (ii) learning and decision-making in uncertain and dynamic environments, and (iii) learning models of human decision-making from data.

In the first part, we blend ideas from game theory and optimization to demonstrate both theoretically and empirically how current machine learning approaches fail in multi-agent settings. We then leverage our understanding of the underlying structure of competitive settings to design efficient algorithms with provable guarantees of performance.

In the second part of this thesis, we combine ideas from statistics— namely the analysis of Langevin Markov Chain Monte Carlo Algorithms (MCMC)— and machine learning to design a versatile and computationally efficient model-based algorithm for the multi-armed bandit problem that has guarantees of optimal performance. In, particular, we develop new characterizations of posteriors in log-concave families of likelihoods and priors and finite-

time convergence rates for Langevin MCMC algorithms and use these theoretical results to show that approximate sampling algorithms like Langevin MCMC can be integrated into Thompson Sampling (the original multi-armed bandit algorithm) without sacrificing performance.

In the final part of this thesis we bring together ideas from behavioral economics and reinforcement learning to develop a method for inverse risk-sensitive reinforcement learning. We first develop a forward model that combines ideas from prospect theory with reinforcement learning to capture the nuances of risk-sensitive decision-making in dynamic environments. We then propose an algorithm for solving the inverse problem of learning a model of an agents' decision-making process from observations of their sequential decisions in dynamic environments.

Altogether, this thesis represents a small step in an emerging research area at the intersection of economics, statistics, machine learning, and control. We conclude with a short discussion of emerging problems and themes in this wider area.

For my family. À ma famille.

Contents

Acknowledgments

During my six years at Berkeley I have had the great fortune of interacting with, learning from, and being mentored by a great number of people who have shaped my graduate school experiences and research interests. The work in this thesis would not have been possible with them.

First and foremost I owe a great debt of gratitude to my two advisors Michael Jordan and Shankar Sastry for their support and mentorship over the course of my studies. When I joined Berkeley in 2015, Shankar was my main advisor and his encouragement and positivity throughout the past six years gave me the confidence to investigate new research directions and helped me grow as a researcher. From day one of my PhD I was— and remain— impressed by his ability to ask big, insightful, and thought-provoking research questions while still being able to dive deep into the weeds on any problem. It is something I really look up to and hope to emulate as I continue on in academia.

Joining Mike's group, SAIL, later in my PhD was a real privilege that introduced me to a number of new ideas and problems and ultimately broadened my horizons as a researcher. Mike has a wonderful way of distilling problems to their simplest forms and communicating ideas clearly without striping away their complexity— something that I really admire. Further, his remarkable breadth of knowledge and insightful advice was an invaluable resource throughout the final years of graduate school and his excitement about the intersection of machine learning and economics re-kindled my interest in the research area, for which I am extremely grateful.

I would be remiss if I did not mention Lillian Ratliff in the same breadth as Mike and Shankar. Lillian mentored and guided me from day one of my PhD, and has a creativity and tenacity in solving problems that I really admire. I was fortunate to have Lillian to throw around research ideas and brainstorm with throughout my PhD and she was really instrumental in setting me on the research path that I am on today.

I also would like to thank professors Ruzena Bajcsy, Shachar Kariv, Sanjoy Mitter, Claire Tomlin, and Pravin Varaiya for taking the time to read my papers, discuss research, and advise me throughout my PhD, their advice and encouragement was invaluable throughout my time at Berkeley.

My time as a graduate student would not have been nearly so enjoyable if it wasn't for all my friends and lab-mates in Shankar's group, Semi-autonomous, and SAIL. From early on in my PhD, I had wonderful thought-provoking chats in 337 Cory with Oladapo Afolabi, Frank Chiu, David Friedovitch-Keil, Chinmay Maheshwari, David McPherson, Kamil Nar, and Dexter Scobee that really influenced my research interests. I give a lot of credit to Dan Calderone and Roy Dong for helping me develop intuition and a mathematical background and curiosity that has served me again and again during grad school. I also have to thank Tyler Westenbroek for entertaining my many control theory questions and ideas over the past couple years and for the engaging discussions on learning and control. Later in my PhD, I was fortunate to find a new set of wonderful collaborators and friends in SAIL and

in the creation of the class DS102 that I also need to thank: Karl Krauth, Tianyi Lin, Horia Mania, Esther Rolf, Yaodong Yu, and Tijana Zrnic.

In the past year I also was extremely privileged to be able to turn to Jaime Fisac, Dorsa Sadigh, and Sylvia Herbert for advice as I navigated the job market in the age of covid. Their advice and encouragement was really invaluable.

Finally, I have to thank a number of good friends from Berkeley who I was fortunate to share my graduate school experience with: Andrea Bajcsy, Somil Bansal, Niladri Chatterjee, Ghassen Jerfel, Anusha Nagabandi, Aldo Pachianno, and Nilesh Tripuraneni.

Last, but most certainly not least, this thesis would not have been possible without the unwavering support, guidance, and love of my family. In particular, I have to thank my parents Ravi Mazumdar and Catherine Rosenberg who are my biggest inspirations and sources of strength, nothing would have been possible without them.

Chapter 1

Introduction

From matching drivers and riders on ride-sharing platforms to approving loans, the ability to learn from data and make decisions in real-time has led to the rapid deployment of machine learning algorithms *at scale* across many aspects of everyday life. The use of these algorithms on such a massive scale, however, has highlighted key failings in our current approach to algorithm design in machine learning. Indeed, the traditional machine learning paradigm largely treats algorithms as operating *in isolation*, but they are in fact increasingly deployed in *uncertain dynamic* environments in which they have to interact with other autonomous agents, algorithms, and human decision makers. These interactions can give rise to surprising and undesired behaviors, like flash crashes in financial markets[1] or price collusion on platforms[2].

To confidently deploy machine learning algorithms in real world settings, it is imperative to view them as parts of complex *societal-scale* systems, and to take this complexity into account in our analysis and algorithm design. Indeed, we need to design algorithms that take into account not only their own objectives, but also their impacts on –and interactions with– humans and other autonomous agents in the larger system.

This dissertation lays the groundwork for developing an understanding of the fundamental limits of learning algorithms in dynamic and multi-agent environments, and designing practical algorithms with provable guarantees of performance for societal-scale systems. In particular, this dissertation addresses three of the core themes of this emerging research agenda:

1. **Learning in games** – where we demonstrate both *theoretically and empirically* how current machine learning approaches fail in multi-agent settings and then leverage an understanding of the underlying structure of competitive settings to design efficient algorithms with provable guarantees of performance [37, 113, 114, 116, 119].

[1] The stock market is now run by computers, algorithms and passive managers, *The Economist*, 2019.
[2] When Bots Collude, *New Yorker*, 2015.

2. **Model-based learning in uncertain dynamic environments** – where we investigate how to design efficient and provably optimal model-based algorithms which can adapt online to uncertainty in dynamic environments [117, 118, 194, 195, 196].

3. **Learning interpretable and expressive models of human decision-making** – where we combine ideas from game theory and behavioral economics with machine learning to learn models of human decision-making in dynamic environments from data [32, 115, 153, 154].

Throughout this dissertation, we combine techniques and analysis tools from dynamical systems theory, stochastic processes, machine learning, and statistics with ideas from game theory and behavioral economics to understand the *dynamics* and inherent *uncertainty* of systems where multiple agents *interact*. We apply our results to problems in transport systems [32, 115, 153, 154] and more traditional machine learning domains like the training of generative adversarial networks (GANs) [113] and robotics [117, 195].

1.1 Dissertation Overview

This dissertation expands upon several recent publications which span three broad aspects of learning and decision-making in societal-scale systems: (i) learning and decision-making in the presence of other autonomous agents and algorithms, (ii) learning and decision-making in uncertain and dynamic environments, and (iii) learning models of human decision-making from data. Each of these areas are considered separately in parts I,II, and III respectively, and brief discussion on future work and combining these themes is presented in Part IV. We now give a broad overview of the high-level takeaways and results of each part.

Part I: Learning in Games

With machine learning algorithms increasingly being deployed in the real world, it is crucial that we improve our understanding of how algorithms interact, and the dynamics that can arise from their interactions. In Part I of this dissertation, we present a sequence of results that expand upon [37, 113, 114, 116, 119].

After introducing the class of games we consider and discussing related work on learning in continuous games in Chapter 1, in Chapter 2, we analyze the equilibria of continuous games using tools and ideas from dynamical systems theory. We make several connections between the equilibria of the game and dynamical systems characterizations of equilibria for gradient-play and discuss their implications for gradient-based learning algorithms. In Chapter 3 we expand upon these connections and present convergence rates and non-convergence guarantees for (stochastic) gradient-play in several classes of games. To emphasize the relevance of these results for practitioners we present in Chapter 4 strong negative results showing that existing multi-agent reinforcement learning algorithms have no guarantees of convergence even in simple Markov games. We conclude this part in Chapter 5 by using our

understanding of the optimization landscape of non-convex-non-concave zero-sum games to design an efficiently-implementable gradient-based algorithm that does not suffer the failures of gradient-play.

Part II: Model-Based Decision-Making Under Uncertainty

Algorithms which are deployed into societal-scale systems need to make decisions in an *online* manner (i.e. make decisions as data arrives) and *adapt to uncertainty in dynamic environments* while having performance and safety guarantees. The dominant paradigm for decision-making under uncertainty is currently that of model-free learning, whereby a large number of general-purpose algorithms with non-asymptotic guarantees have been developed in recent years. Despite this, there is growing evidence that model-based algorithms are more efficient in terms of data (largely because they allow for the incorporation of prior knowledge about the problem structure), and are more amenable to safety and performance guarantees. As such there is a pressing need to design versatile model-based algorithms for decision-making under uncertainty.

To understand how to optimally adapt a model to uncertainty, in Part II we analyze model-based approaches in the simplest dynamic environment: a multi-armed bandit setting. Despite its widespread use in industry and the fact that it is nearly 100 years old, the main model-based algorithm, Thompson Sampling, is still not well understood theoretically. In particular, the fact that the algorithm requires samples from posterior distributions makes it difficult to analyze outside of *particularly nice* problems, and it remains an open question whether using approximate posteriors can give an optimal algorithm. In this dissertation we present recent work that makes use of ideas from statistical machine learning and the continuous-time analysis of Langevin Markov Chain Monte Carlo algorithms to design an approximate Thompson Sampling algorithm with provably optimal performance guarantees for the multi-armed bandit problem. Unlike prior work, our proof techniques were also able to accurately capture the dependence of prior information on performance. The resulting algorithm is the first provably optimal approximate Thompson Sampling algorithm.

Part III: Learning interpretable and expressive models of human decision-making

As autonomous agents interact more with humans as economic agents, there is an increasing need to design algorithms that can reason about the impact of their decisions on people in real-time. Doing so requires efficiently making decisions in regimes with little data and using interpretable and actionable models of human decision-making. Motivated by the observation that firms are collecting increasingly fine-grained data on people's sequential decisions in dynamic environments, in Part III we focus on incorporating ideas from behavioral economics into inverse reinforcement learning.

Most work on inverse reinforcement learning assumes that humans are simple utility maximizers. Following the Nobel prize winning findings of economists Daniel Kahneman

and Amos Tversky, we combine ideas from *Prospect Theory* and reinforcement learning to model humans as *risk-sensitive* decision makers in dynamic environments. We then derive a procedure to learn prospect-theoretic utility functions from data. In particular, we expand upon prior work on risk-sensitive reinforcement learning to provably incorporate prospect-theoretic utility functions into risk-sensitive Q-learning and then demonstrate how to solve the inverse problem of finding a utility function that could best explain observed behaviors [115, 153]. Unlike prior work, the learned utilities capture key aspects of human decision-making like risk preferences and reference points. To validate our algorithm we implemented it on data from ride-sharing platforms where our method was able to capture agents' risk-attitudes towards surge-pricing [115].

Part I

Learning in Games

Chapter 1

Understanding the Optimization Landscape of Continuous Games

With machine learning algorithms increasingly being deployed in real world settings, it is crucial that we understand how the algorithms can interact, and the dynamics that can arise from their interactions. In recent years, there has been a resurgence in research efforts on multi-agent learning, and learning in games. The recent interest in adversarial learning techniques also serves to show how game theoretic tools can be being used to *robustify* and improve the performance of machine learning algorithms. Despite this activity, however, machine learning algorithms are still being treated as black-box approaches and being naïvely deployed in settings where other algorithms are actively changing the environment.

In general, outside of highly structured settings, there exists no guarantees on the performance or limiting behaviors of learning algorithms in such settings. Indeed, previous work on understanding the collective behavior of coupled learning algorithms, either in competitive or cooperative settings, has mainly looked at games where the global structure is well understood like bilinear games [76, 99, 121, 176], convex games [123, 164], or potential games [129], among many others. Such games are more conducive to the statement of global convergence guarantees since the assumed global structure can be exploited.

In games with fewer assumptions on the players' costs, however, there remains a lack of understanding of the dynamics and limiting behaviors of learning algorithms. Such settings are becoming increasingly prevalent as deep learning is increasingly being used in game theoretic settings [2, 55, 68, 203].

Gradient-based learning algorithms are extremely popular in a variety of these multi-agent settings due to their versatility, ease of implementation, and dependence on local information. There are numerous recent papers in multi-agent reinforcement learning that employ gradient-based methods (see, e.g.[2, 55, 203]), yet even within this well-studied class of learning algorithms, a thorough understanding of their convergence and limiting behaviors in general continuous games is still lacking.

Generally speaking, in both the game theory and the machine learning communities, two of the central questions when analyzing the dynamics of learning in games are the following:

Q1. *Do learning algorithms employed by agents find equilibria that are relevant to the underlying game?*

Q2. *Can all equilibria relevant to the game be found by the learning algorithms that agents employ?*

In the following chapters, we provide some answers to the above questions for the class of gradient-based learning algorithms by analyzing their limiting behavior in general continuous games. In particular, we leverage the continuous time limit of the more naturally discrete multi-agent learning algorithms. This allows us to draw on the extensive theory of dynamical systems and stochastic approximation to make statements about the limiting behaviors of these algorithms in both deterministic and stochastic settings. The latter is particularly relevant since it is common for stochastic gradient methods to be used in multi-agent machine learning contexts.

Analyzing gradient-based algorithms through the lens of dynamical systems theory has recently yielded new insights into their behavior in the classical optimization setting [98, 170, 197]. We show that a similar type of analysis can also help understand the limiting behaviors of gradient-based algorithms in games. We remark, however, that there is a *fundamental difference* between the dynamics that are analyzed in much of the single-agent, gradient-based learning and optimization literature and the ones we analyze in the competitive multi-agent case: the combined dynamics of gradient-based learning schemes in games *do not necessarily correspond to a gradient flow*. This may seem a subtle point, but it it turns out to be extremely important.

Gradient flows admit desirable convergence guarantees—e.g., almost sure convergence to local minimizers—due to the fact that they preclude flows with the *worst geometries* [146]. In particular, they do not exhibit non-equilibrium limiting behavior such as periodic orbits. Gradient-based learning in games, on the other hand, does not preclude such behavior. Moreover, as we show, asymmetry in the dynamics of gradient-play in games can lead to surprising behaviors such as non-relevant limiting behaviors being attracting under the flow of the game dynamics and relevant limiting behaviors, such as a subset of the Nash equilibria being almost surely avoided.

In the next section we introduce the general framework for modeling competitive gradient-based learning that we analyze throughout Part I and then discuss some recent work on gradient-based learning in games.

1.1 Gradient-Play in Continuous Games

Consider N agents indexed by $\mathcal{I} = \{1, \dots, N\}$. Each agent $i \in \mathcal{I}$ has its own decision variable $x_i \in X_i$, where X_i is its finite-dimensional strategy space of dimension d_i. Define $X = X_1 \times \cdots \times X_N$ to be the finite-dimensional joint strategy space with dimension $d = \sum_{i \in \mathcal{I}} d_i$. Each agent is endowed with a cost function $f_i \in C^s(X, \mathbb{R})$ with $s \geq 2$ and such that $f_i : (x_i, x_{-i}) \mapsto f_i(x_i, x_{-i})$ where we use the notation $x = (x_i, x_{-i})$ to make the dependence on the action of

the agent x_i, and the actions of all agents excluding agent i, $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ explicit. The agents seek to minimize their own cost, but only have control over their own decision variable x_i. In this setup, agents' costs are not necessarily aligned with one another, meaning they are competing.

Given the game $\mathcal{G} = (f_1, \ldots, f_N)$, agents are assumed to update their strategies *simultaneously* according to a gradient-based learning algorithm of the form

$$x_{i,t+1} = x_{i,t} - \gamma_{i,t} h_i(x_{i,t}, x_{-i,t}), \tag{1.1}$$

where $\gamma_{i,t}$ is agent i's step-size at iteration t.

We analyze the following two settings:

1. Agents have *oracle access* to the gradient of their cost with respect to their own choice variable:
$$h_i(x_{i,t}, x_{-i,t}) = D_i f_i(x_{i,t}, x_{-i,t}),$$
where $D_i f_i \equiv \partial f_i / \partial x_i$ denotes the derivative of f_i with respect to x_i.

2. Agents have an *unbiased estimator* of their gradient:
$$h_i(x_{i,t}, x_{-i,t}) = D_i f_i(x_{i,t}, x_{-i,t}) + w_{i,t+1},$$
where $\{w_{i,t}\}$ is a zero mean, finite variance stochastic process.

We refer to the former setting as *deterministic* gradient-based learning and the latter setting as *stochastic* gradient-based learning. To simplify notation, we define:

$$\omega(x) = (D_1 f_1(x), \ldots, D_N f_N(x)),$$

to be the vector of player derivatives of their own cost functions with respect to their own choice variables. This is the core object of our analysis throughout the following sections.

Assuming that all agents are employing such algorithms, we aim to analyze the limiting behavior of the agents' strategies. To do so we leverage the following notion of a Nash equilibrium from game theory.

Definition 1 (Local Nash equilibrium)**.** *A strategy $x \in X$ is a local Nash equilibrium for the game $(f_1, \ldots, f_N)$ if, for each $i \in \mathcal{I}$, there exists an open set $W_i \subset X_i$ such that that $x_i \in W_i$ and $f_i(x_i, x_{-i}) \leq f_i(x_i', x_{-i})$ for all $x_i' \in W_i$. If the above inequalities are strict, then we say x is a strict local Nash equilibrium.*

The focus on *local* Nash equilibria is due to our lack of assumptions on the agents' cost functions and an implicit assumption that agents do not cooperate and that there is no order-of-play in the game. If $W_i = X_i$ for each i, then a local Nash equilibrium x is a global Nash equilibrium. Depending on the agents' costs, a game $(f_1, \ldots, f_N)$ may admit anywhere from one to a continuum of local or global Nash equilibria; or none at all.

We now introduce several broad classes of games which are of particular interest in various application domains. We call games in which we make no particular assumptions on the players' losses $f_1, ..., f_N$ beyond $f_i \in C^s(X, \mathbb{R})$ with $s \geq 2$ for $i \in \mathcal{I}$, N-player general-sum games or general-sum games for short.

To begin, we first define a class of games that admits particularly strong structure guarantees: potential games [129].

Definition 2 (Potential Games)**.** *An N-player potential game is a game in which ω corresponds to a gradient flow under a coordinate transformation—that is, there exists a function ϕ (commonly referred to as the potential function) such that for each $i \in \mathcal{I}$, $D_i f_i \equiv D_i \phi$ for $f_i \in C^s(X, \mathbb{R})$ where $s \geq 2$.*

We remark that due to the equivalence this class of games is sometimes referred to as an *exact* potential game, and that such games enjoys many particularly nice properties owing to the fact that the entire N-player game game be characterized by one function ϕ.

A second important class of games are two-player zero-sum games, which often arise when training GANs [68], adversarial learning [140], multi-agent reinforcement learning[39], and more broadly in the context of min-max optimization [101, 135, 151].

Definition 3 (Zero-sum game)**.** *A zero-sum game is a 2-player game where one player seeks to minimize a function f with respect to their decision-variable and the second seeks to maximize f (or equivalently they seek to minimize $-f$) with respect to theirs:*

$$f_1 = f \qquad f_2 = -f \qquad \text{for} \qquad f \in C^s(X, \mathbb{R}),$$

where $s \geq 2$.

The fact that zero-sum games are fully characterized by a single function gives zero-sum games a particularly nice structure, that, while considerably more difficult to analyze than simple minimization problems (and therefore potential games), allows us to make stronger statements than in the more arbitrary class of 2-player general-sum games.

To conclude, we introduce several classes of structured games that have been well analyzed in the literature. We start with the class of convex games which have been the focus of a large body of work stretching back to Rosen in 1965[164].

Definition 4 (Convex game)**.** *A N-player convex game is a N-player game where, for each $i \in \mathcal{I}$, $f_i(x_i, x_{-i})$ is convex in x_i for fixed $x_{-i} \in X_{-i}$:*

Convex games are know to admit either a unique global Nash equilibrium or a continuum of Nash equilibria. A subset of convex games of particular interest in recent years is the class of monotone (or diagonally strictly convex) games:

Definition 5 (Monotone game)**.** *A monotone game is a N-player convex game where ω satisfies:*

$$\langle \omega(x) - \omega(x'), x - x' \rangle \geq 0 \qquad \forall\ x, x' \in X.$$

If the inequality is strict, the game is strictly monotone.

While monotone games are in general a strict subset of convex games, in the class zero-sum games assuming convexity of the players' costs is equivalent to assuming monotoncity. That is, convex zero-sum games (alternatively known as convex-concave zero-sum games) are equivalent to monotone zero-sum games [120]. A particularly common form of monotone zero-sum game which has garnered much attention in recent years [44, 63] is that of bilinear games where $f(x_1, x_2) = x_1^T A x_2$ for a given matrix $A \in \mathbb{R}^{d_1 \times d_2}$.

The last class of structured games of particular interest is the class of strongly-monotone games.

Definition 6 (Strongly monotone game)**.** *A strongly monotone game is a N-player convex game where ω satisfies:*

$$\langle \omega(x) - \omega(x'), x - x' \rangle \geq \mu \|x - x'\|^2 \qquad \forall\ x, x' \in X$$

for $\mu > 0$.

Once again, it is clear that this is a subset of monotone games. In zero-sum games, this is equivalent to assuming that f is strongly convex in its first argument and strongly concave in its second.

1.2 Related Work

The study of continuous games is quite extensive (see e.g. [19, 141]), though in large part the focus has been on games admitting a fair amount of structure. The behavior of learning algorithms in games is also well-studied (see e.g. [57]). In this section, we comment on the most relevant prior work.

Characterizing the properties of Nash equilibria in games has been a topic or interest going back to Nash in his seminal work [134]. In continuous games—the area of interest of this dissertation— the literature goes back to work by Rosen [164], in which n–player concave games are shown to either admit a unique global Nash equilibrium or a continuum of Nash equilibria, all of which (under the additional assumption of monotonicity) are attracting under gradient-play. Understanding how to efficiently compute Nash equilibria in such games continues to attract interest in machine learning (see e.g., [14, 29] and the references therein) and theoretical computer science more generally [43].

The majority of the existing work on learning in games has focused on understanding and designing learning rules that be used to find Nash equilibria i.e., attempting to answer **Q1**. Proceeding under various structural assumptions on the players' costs and strategies— most of which preclude the existence of non-Nash equilibria— answering **Q1** reduces to analyzing the convergence of various learning algorithms (including gradient-play) to the unique Nash equilibrium or the set of Nash equilibria of the game. Examples of classes of structured games amenable to such strong guarantees are potential games [129], concave or monotone games [29, 123, 164], and gradient-play over the space of stochastic policies in two-player

finite-action bilinear games [176]. In [176], the authors investigate the convergence of the gradient dynamics in such games. Additionally, the dynamics of other (non gradient-based) algorithms like multiplicative weights have been studied in [76] among many others. In these classes of structured games, a large line of recent work in machine learning has begun to use tools developed in the analysis of gradient descent in optimization to understand how to efficiently compute Nash equilibria [14, 29, 72].

In more general classes of continuous games, recent works have also looked at characterizing when Nash equilibria are attracting for gradient-based approaches. Sufficient conditions for this to occur are the conditions for stable differential Nash equilibria introduced in [155, 156, 157] (which we adopt in our treatment) and the condition for variational stability later analyzed in [123]. We remark that these conditions are equivalent for the classes of games we consider and that the results in these papers only characterize the *local* properties of these equilibria. In the work presented in this dissertation we focus on giving global non-convergence results and comment on other non-equilibrium attracting behaviors (i.e., answering **Q2**).

Another line of recent work worth mentioning introduces new equilibrium concepts. Whereas the focus of the work presented in the subsequent sections is simultaneous-play games, two recent papers [78] and [52] analyze Stackelberg (or leader-follower) continuous games and introduce the notion of a local Stackelberg equilibrium (the analogous equilibrium concept to local Nash equilibria). While the focus is mainly on zero-sum games due to their links with min-max optimization, these local Stackelberg equilibria have a number of desirable properties (e.g., existence under weaker assumptions) and warrant investigation in their own right. However, since the focus of this dissertation is on simultaneous-play games these equilibria are of unknown relevance to the underlying games. For a more in depth discussion on local Stackelberg equilibria, we refer the reader to [52] and [78].

The last line of relevant work to the results presented in subsequent chapters is the emerging research area of efficient algorithms for min-max optimization. Driven by the interest in adversarial and robust learning paradigms in machine learning, a large number of recent work have leveraged tools from the analysis of gradient descent in convex optimization to understand how to compute saddle points (or Nash equilibria) in zero-sum games (see. e.g., [44, 78, 121, 128]). A number of works have shown that gradient-play in such games can converge to cycles (see, e.g., [121, 193]) or even diverge completely [128]. Several papers (including the one this dissertation draws from [116]) pointed our concurrently that there also exist non-Nash attracting equilibria for gradient-play in zero-sum games [6, 44], meaning that not all equilibria for gradient-play in zero-sum games are necessarily relevant to the game.

Expanding on this rich body of literature (only the most relevant of which is covered in our short review), in Part I of this dissertation we provide answers to **Q1** without imposing structure on the game outside regularity conditions on the cost functions by exploiting the observation that gradient-based learning dynamics are not gradient flows. We also provide answers to **Q2** by demonstrating that a non-trivial set of games admit Nash equilibria that are almost surely avoided by gradient-play. We give explicit conditions for when this occurs. Using similar analysis tools, we also provide new insights into the behavior of gradient-based

learning in structured classes of games such as zero-sum and potential games.

1.3 Overview of Part I

In Chapter 2, we draw connections between the limiting behavior of this class of algorithms and game-theoretic and dynamical systems notions of equilibria. In particular, we construct general-sum and zero-sum games that admit non-Nash attracting equilibria of the gradient dynamics. Such points are attracting under the learning dynamics, yet at least one player—*and potentially all of them*—has a direction in which they could unilaterally deviate to decrease their cost. Thus, these non-Nash equilibria are of questionable game theoretic relevance and can be seen as artifacts of the players' algorithms. In the context of zero-sum games we also present strong characterizations of Nash equilibria that hold for 'almost all' zero-sum games in a formal mathematical sense.

In Chapter 3, we show that policy gradient multi-agent reinforcement learning (MARL), generative adversarial networks (GANs), and gradient-based multi-agent multi-armed bandits, among several other common multi-agent learning settings all fit into the framework of gradient-based learning algorithms we analyze. We then proceed to prove that a subset of the local Nash equilibria in general-sum games and potential games is avoided almost surely when each player employs a gradient-based algorithm. We show that this holds in two broad settings: the full information setting when each player has oracle access to their gradient but randomly initializes their first action, and a partial information setting where each player has access to an unbiased estimate of their gradient.

Altogether, in Chapters 2 and 3 we provide a negative answer to both **Q1** and **Q2** for N–player general-sum games, and highlight the nuances present in zero-sum and potential games. We also show that the dynamics formed from the individual gradients of agents' costs are *not gradient flows*. This in turn implies that competitive gradient-based learning in general-sum games may converge to periodic orbits, and other complex ω limit sets, possibly including chaotic dynamics.

To highlight the fact that these issues are not of merely theoretical interest, in Chapter 4 we expand upon our non-convergence results to show that policy gradient algorithms have no guarantees of convergence in even the simplest continuous action and state multi-agent reinforcement learning problems— linear quadratic games.

In Chapter 5 we move away from gradient-play and develop an efficiently implementable algorithm for zero-sum games that, unlike gradient-play, has provable guarantees of convergence to Nash equilibria even in non-convex-non-concave zero-sum games. In doing so, we highlight how many recently proposed algorithms for min-max optimization have no guarantees of finding game relevant equilibria.

Chapter 2

Linking Games and Dynamical Systems

With machine learning algorithms increasingly being placed in more complex, real world settings, there has been a renewed interest in continuous games [52, 123, 203], and particularly zero-sum continuous games [44, 68, 78, 116] because of their links to adversarial learning [45, 121], robust reinforcement learning [101, 151], min-max optimization [135], and generative adversarial networks [68].

Despite this, a systematic characterization of the equilibria of such games and of even the limiting behaviors of simple learning algorithms in these games is sorely lacking. In this chapter, we draw links between the limiting behavior of dynamical systems and game-theoretic notions of equilibria in three broad classes of continuous games. A high-level summary of the links we draw is shown in Figure 2.1.

Key to our analysis is the fact that, when each player is employing a gradient-based learning algorithm, the joint strategy of the players—in the limit as the agents' step-sizes go to zero— follows the differential equation

$$\dot{x} = -\omega(x). \tag{2.1}$$

Thus, by analyzing properties of (2.1) we are able to say something about the dynamics of players using gradient-based algorithms in games.

Chapter Overview

This chapter is organized as follows. In Section 2.1 we highlight connections between critical points of (2.1) and the underlying game and further characterize the behavior of gradient-play in neighborhoods of Nash equilibria in broad classes of games. In doing so we uncover previously unknown phenomena like the existence of spurious stationary points in the gradient dynamics which have no relevance to the underlying games.

In Section 2.2, we also present a series of strong characterizations of the differential topology of local Nash equilibria in zero-sum continuous games. These results expand upon characterizations of local Nash equilibria in general-sum games that first appeared in [157]. The results presented in this chapter are a combination of results that first appeared in [116] and [114] that have been combined for clarity of exposition.

2.1 Equilibrium Notions in Games and Dynamical Systems

To begin, we draw links between equilibria in dynamical systems theory and equilibria of the game.

A point $x \in X$ is said to be an equilibrium, critical point, or stationary point of the dynamics if $\omega(x) = 0$. Stationary points of $\dot{x} = -\omega(x)$ are joint strategies from which, under gradient-play, the agents do not move. We note that $\omega(x) = 0$ is a necessary condition for a point $x \in X$ to be a local Nash equilibrium [155]. Hence, all local Nash equilibria are critical points of the joint dynamics $\dot{x} = -\omega(x)$.

Central to dynamical systems theory is the study of limiting behavior and its stability properties. A classical result in dynamical systems theory allows us to characterize the stability properties of an equilibrium x^* by analyzing the Jacobian of the dynamics at x^*. The Jacobian of ω is defined by:

$$J(x) = \begin{bmatrix} D_1^2 f_1(x) & \cdots & D_{N1} f_1(x) \\ \vdots & \ddots & \vdots \\ D_{1N} f_N(x) & \cdots & D_N^2 f_N(x) \end{bmatrix}.$$

Since J is a matrix of second derivatives, it is sometimes referred to as the 'game Hessian'. Similar to the Hessian matrix of a gradient flow, J allows us to further characterize the critical points of ω by their properties under the flow of $\dot{x} = -\omega(x)$. Let $\lambda_i(x) \in \text{spec}(J(x))$ for $i \in \{1, \ldots, m\}$ denote the eigenvalues of J at x where $\text{Re}(\lambda_1(x)) \leq \cdots \leq \text{Re}(\lambda_m(x))$—that is, $\lambda_1(x)$ is the eigenvalue with the smallest real part. Given these definitions, we first define important class of critical points known as hyperbolic critical points.

Definition 7. *A critical point x is hyperbolic if $J(x)$ has no eigenvalues with zero real part.*

Hyperbolic critical points are of particular importance from the point of view of convergence since they are either exponential stable or exponentially unstable under the joint dynamics [168].

A further important property of a critical point of ω is non-degeneracy.

Definition 8. *A critical point x is non-degenerate if* $\det(J(x)) \neq 0$ *(i.e. x is isolated).*

Non-degenerate equilibria are isolated, meaning there exists an open neighborhood around them devoid of critical points [157]. We remark that by definition, all hyperbolic critical points are non-degenerate, but not all non-degenerate critical points are hyperbolic.

Given these building blocks, we now define different classes of equilibria which have different characteristics under the flow $\dot{x} = -\omega(x)$. Of particular interest are asymptotically stable equilibria.

Definition 9. *A point $x \in X$ is a locally asymptotically stable equilibrium of the continuous time dynamics $\dot{x} = -\omega(x)$ if $\omega(x) = 0$ and $\mathrm{Re}(\lambda) > 0$ for all $\lambda \in \mathrm{spec}(J(x))$.*

Locally asymptotically stable equilibria have two properties of interest. First, they are non-degenerate critical points, meaning that there exists a neighborhood around them in which no other equilibria exist. Second, they are hyperbolic, and in fact exponentially attracting under the flow of $\dot{x} = -\omega(x)$, meaning that if agents initialize in a neighborhood of a locally asymptotically stable equilibrium x^* and follow the dynamics described by $\dot{x} = -\omega(x)$, they will converge to x^* exponentially fast [168]. This, in turn, implies that a discretized version of $\dot{x} = -\omega(x)$, namely

$$x_{t+1} = x_t - \gamma\omega(x_t), \tag{2.2}$$

converges locally for appropriately selected step size γ at a rate of $O(1/t)$. Such results motivate the study of the continuous time dynamical system $\dot{x} = -\omega(x)$ in order to understand convergence properties of gradient-based learning algorithms of the form (1.1).

Another important class of critical points of a dynamical system are saddle points.

Definition 10. *A point $x \in X$ is a saddle point of the dynamics $\dot{x} = -\omega(x)$ if $\omega(x) = 0$ and $\lambda_1(x) \in \mathrm{spec}(J(x))$ is such that $\mathrm{Re}(\lambda_1(x)) \leq 0$. A saddle point such that $\mathrm{Re}(\lambda_i) < 0$ for $i \in \{1, \ldots, \ell\}$ and $\mathrm{Re}(\lambda_j) > 0$ for $j \in \{\ell + 1, \ldots, m\}$ with $0 < \ell < m$ is a strict saddle point of the continuous time dynamics $\dot{x} = -\omega(x)$.*

Strict saddle points are especially relevant to our analysis since they are hyperbolic critical points whose neighborhoods are characterized by stable and unstable manifolds [168]. When the agents evolve according to the dynamics solely on the stable manifold, they converge exponentially fast to the critical point. However, when they evolve solely on the unstable manifold, they diverge from the equilibrium exponentially fast. Agents whose strategies lie on the union of the two manifolds asymptotically avoid the equilibrium. We make use of this general fact in Section 3.2.

To better understand the links between the critical points of the gradient dynamics and the Nash equilibria of the game, we make use of an equivalent characterization of strict local Nash that leverages first and second order conditions on player cost functions. This makes them simpler objects to link to the various dynamical systems notions of equilibria than local Nash equilibria.

Definition 11 ([155, 156])**.** *A point $x \in X$ is a differential Nash equilibrium for the game defined by $(f_1, \ldots, f_N)$ if $\omega(x) = 0$ and $D_i^2 f_i(x) \succ 0$ for each $i \in \mathcal{I}$.*

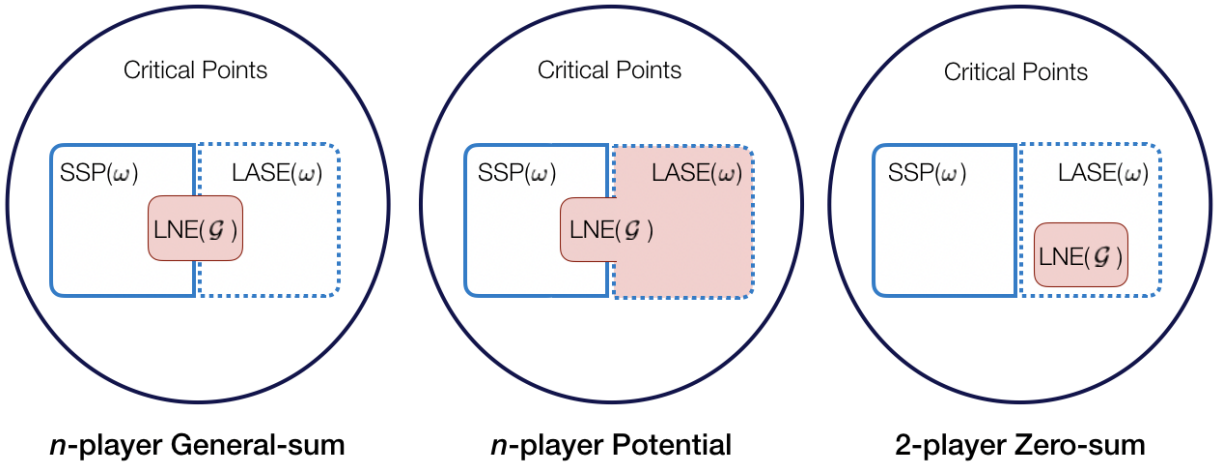


Figure 2.1: Links between the equilibria of generic continuous games $\mathcal{G}$ and their properties under the gradient dynamics $\dot{x} = -\omega(x)$.

If $x \in X$ is a differential Nash equilibrium where $\det(J(x)) \neq 0$ (i.e., x is non-degenerate), then x is referred to a non-degenerate Nash equilibrium. Non-degenerate differential Nash equilibria are necessarily isolated critical points of gradient-play [157].

Given these different equilibrium notions of the learning dynamics and the underlying game, let us define the following sets which will be useful in stating the results in the following sections. For a game $\mathcal{G} = (f_1, \ldots, f_N)$, denote the sets of strict saddle points and locally asymptotically stable equilibria of the gradient dynamics, $\dot{x} = -\omega(x)$, as $\texttt{SSP}(\omega)$ and $\texttt{LASE}(\omega)$, respectively, where we recall that $\omega(x) = (D_1 f_1(x), \ldots, D_N f_N(x))$. Similarly, denote the set of local Nash equilibria, differential Nash equilibria, and non-degenerate differential Nash equilibria of $\mathcal{G}$ as $\texttt{LNE}(\mathcal{G})$, $\texttt{DNE}(\mathcal{G})$, and $\texttt{NDDNE}(\mathcal{G})$, respectively. As we show in the subsequent section, $\texttt{NDDNE}(\mathcal{G}) = \texttt{LNE}(\mathcal{G})$ in 'almost all' continuous games. The key takeaways of this section are summarized in Figure 2.1.

Equilibria in General-Sum Games

We first characterize the properties of local Nash equilibria under the joint gradient dynamics in N-player general-sum games.

Proposition 1. *A non-degenerate differential Nash equilibrium is either a locally asymptotically stable equilibrium or a strict saddle point of $\dot{x} = -\omega(x)$—i.e., $\texttt{NDDNE}(\mathcal{G}) \subset \texttt{SSP}(\omega) \cup \texttt{LASE}(\omega)$.*

Proof of Proposition 1. Suppose that $x \in X$ is a non-degenerate differential Nash equilibrium. We claim that $\mathrm{tr}(J(x)) > 0$. Since x is a differential Nash equilibrium, $D_i^2 f_i(x) \succ 0$ for each $i \in \mathcal{I}$; these are the diagonal blocks of $J(x)$. Further $D_i^2 f_i(x) \succ 0$ implies that $\mathrm{tr}(D_i^2 f_i(x)) > 0$. Since $\mathrm{tr}(J) = \sum_{i=1}^n \mathrm{tr}(D_i^2 f_i(x))$, $\mathrm{tr}(J(x)) > 0$. Thus, it is not possible

for all the eigenvalues to have negative real part. Since x is non-degenerate, $\det(J(x)) \neq 0$ so that none of the eigenvalues can have zero real part. Hence, at least one eigenvalue has strictly positive real part.

To complete the proof, we show that the conditions for non-degenerate differential Nash equilibrium are not sufficient to guarantee that x is locally asymptotically stable for the gradient dynamics—that is, not all eigenvalues of $J(x)$ have strictly positive real part. We do this by constructing a class of games with the strict saddle point property. Consider a class of two player games $\mathcal{G} = (f_1, f_2)$ on $\mathbb{R} \times \mathbb{R}$ defined as follows:

$$(f_1(x_1, x_2), f_2(x_1, x_2)) = \Big(\frac{a}{2}x_1^2 + bx_1x_2, \frac{d}{2}x_2^2 + cx_1x_2\Big).$$

In this game, the Jacobian of the gradient dynamics is given by

$$J(x) = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{2.3}$$

with $a, b, c, d \in \mathbb{R}$. If x is a non-degenerate differential Nash equilibria, $a, d > 0$ and $\det(J(x)) \neq 0$ which implies that $ad \neq cb$. Choosing c, d such that $ad < cb$ will guarantee that one of the eigenvalues of $J(x)$ is negative and the other is positive, making x a strict saddle point. This shows that non-degenerate differential Nash equilibria can be strict saddle points of the combined gradient dynamics.

Hence, for any game $(f_1, \ldots, f_n)$, a non-degenerate differential Nash equilibrium is either a locally asymptotically stable equilibrium or a strict saddle point, but it not strictly unstable or strictly marginally stable (i.e. having eigenvalues all on the imaginary axis). □

Locally asymptotically stable differential Nash equilibria satisfy the notion of variational stability introduced in [123]. In fact, a simple analysis shows that the definitions of variationally stable equilibria and locally asymptotically stable differential Nash equilibria [156] are equivalent in the games we consider—i.e., games where each players' cost is at least twice continuously differentiable. We remark that, from the definition of asymptotic stability, the gradient dynamics have an *exponential* convergence rate in the neighborhood of such equilibria.

An important point to make is that not every locally asymptotically stable equilibrium of $\dot{x} = -\omega(x)$ is a non-degenerate differential Nash equilibrium. Indeed, the following proposition provides an entire class of games whose corresponding gradient dynamics admit locally asymptotically stable equilibria that are not local Nash equilibria.

Proposition 2. *In the class of general-sum continuous games, there exists a continuum of games containing games $\mathcal{G}$ such that* `LASE`$(\omega) \not\subset$ `NDDNE`$(\mathcal{G})$*, and moreover,* `LASE`$(\omega) \not\subset$ `LNE`$(\mathcal{G})$*.*

Proof. Consider a two player game $\mathcal{G} = (f_1, f_2)$ on $\mathbb{R}^2$ where

$$f_1(x_1, x_2) = \frac{a}{2}x_1^2 + bx_1x_2, \quad \text{and} \quad f_2(x_1, x_2) = \frac{d}{2}x_2^2 + cx_1x_2$$

for constants $a, b, c, d \in \mathbb{R}$. The Jacobian of ω is given by

$$J(x_1, x_2) = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad \forall (x_1, x_2) \in \mathbb{R}^2. \tag{2.4}$$

If $a > 0$ and $d < 0$, then the unique stationary point $x = (0, 0)$ is neither a differential Nash nor a local Nash equilibria since the necessary conditions are violated (i.e., $d < 0$). However, if $a > -d$ and $ad > cb$, the eigenvalues of J have positive real parts and $(0, 0)$ is asymptotically stable. Further, this clearly holds for a continuum of games. Thus, the set of locally asymptotically stable equilibria that are not Nash equilibria may be arbitrarily large. □

The, preceding proposition shows that there exists attracting critical points of the gradient dynamics in general-sum continuous games that are not Nash equilibria and may not be even relevant to the game. Thus, this provides a negative answer to **Q2** (whether all attracting equilibria in general-games are game-relevant for the learning dynamics).

Remark 1. *We note that, by definition, the non-Nash locally asymptotically stable equilibria (or non-Nash equilibria) do not satisfy the second-order conditions for Nash equilibria. Thus, at these joint strategies, at least one player – and maybe all of them – has a direction in which they would unilaterally deviate if they were not using gradient descent. Some— but as shown in both [78] and [52], not all— non-Nash attracting equilibria might be local Stackelberg equilibria under different orders-of-play in the game. Since the focus of this dissertation is on* simultaneous*-play games, these non-Nash local Stackelberg equilibria are not relevant to the game. As such, we view convergence to these points to be undesirable.*

Equilibria in Potential Games

We now specialize our results to a class of games with interesting connections between the Nash equilibria and the critical points of the gradient dynamics: *potential games*. Note that a necessary and sufficient condition for $(f_1, \ldots, f_N)$ to be a potential game is that J is *symmetric* [129]—that is, $D_{ij} f_j \equiv D_{ji} f_i$. This gives potential games the desirable property that the only locally asymptotically stable equilibria of the gradient dynamics are local Nash equilibria.

Proposition 3. *For an arbitrary potential game, $\mathcal{G} = (f_1, \ldots, f_N)$ on $\mathbb{R}^m$, if x is a locally asymptotically stable equilibrium of $\dot{x} = -\omega(x)$ (i.e., $x \in$ `LASE`(ω)), then x is a non-degenerate differential Nash equilibrium (i.e., $x \in$ `NDDNE`$(\mathcal{G})$).*

The proof that all locally asymptotically stable equilibria in potential games are differential Nash equilibria relies on the symmetry of J in potential games.

Proof of Proposition 3. The proof follows from the definition of a potential game. Since $(f_1, \ldots, f_n)$ is a potential game, it admits a potential function ϕ such that $D_i f_i(x) = D_i \phi(x)$ for

all x. This, in turn, implies that at a locally asymptotically stable equilibrium of $\dot{x} = -\omega(x)$, $J(x) = D^2\phi(x)$, where $D^2\phi$ is the Hessian matrix of the function ϕ. Further $D^2\phi(x)$ must have strictly positive eigenvalues for x to be a locally asymptotically stable equilibrium of $\dot{x} = -\omega(x)$. Since the Hessian matrix of a function must be symmetric, $D^2\phi(x)$, must be positive definite, which through Sylvester's criterion ensures that each of the diagonal blocks of $D^2\phi(x)$ is positive definite. Thus, we have that the existence of a potential function guarantees that the only locally asymptotically stable equilibria of $\dot{x} = -\omega(x)$, are differential Nash equilibria. □

The preceding proposition rules out non-Nash locally asymptotically stable equilibria of the gradient dynamics in potential games, and implies that every local minimum of a potential game must be a local Nash equilibrium. Thus, in potential games, unlike in general-sum and zero-sum games, the answer to **Q2** is positive. However, the following proposition shows that the existence of a potential function is not enough to rule out local Nash equilibria that are saddle points of the dynamics.

Proposition 4. *In the class of continuous games, there exist a continuum of potential games containing games $\mathcal{G}$ that admit Nash equilibria that are saddle points of the dynamics $\dot{x} = -\omega(x)$—i.e., $\exists\ \mathcal{G}$ such that for some $x \in \texttt{LNE}(\mathcal{G})$, $x \in \texttt{SSP}(\omega)$.*

Proof. Consider the game (f, f) on $X = \mathbb{R}^2$ described by

$$f(x_1, x_2) = \frac{a}{2}x_1^2 + bx_1x_2 + \frac{c}{2}x_2^2$$

where $a, b, d \in \mathbb{R}$. The Jacobian of ω is given by

$$J(x_1, x_2) = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \quad \forall\ (x_1, x_2) \in \mathbb{R}^2.$$

If $a, c > 0$, then $x = (0, 0)$ is a local Nash equilibrium. However, if $ac < b^2$, $J(x)$ has one positive and one negative eigenvalue and $(0, 0)$ is a saddle point of the gradient dynamics. Thus, there exists a continuum of potential games where a large set of differential Nash equilibria are strict saddle points of $\dot{x} = -\omega(x)$. □

Proposition 4 demonstrates a surprising fact about potential games. Even though all minimizers of the potential function must be local Nash equilibria, *not all local Nash equilibria are minimizers of the potential function.*

Equilibria in Zero-Sum Games

The following proposition shows that all differential Nash equilibria in two-player zero-sum games are locally asymptotically stable equilibria under the flow of $\dot{x} = -\omega(x)$.

Proposition 5. *For an arbitrary two-player zero-sum game, $(f, -f)$ on $\mathbb{R}^m$, if x is a differential Nash equilibrium, then x is both a non-degenerate differential Nash equilibrium and a locally asymptotically stable equilibrium of $\dot{x} = -\omega(x)$—that is, $\texttt{DNE}(\mathcal{G}) \equiv \texttt{NDDNE}(\mathcal{G}) \subset \texttt{LASE}(\omega)$.*

The proof of Proposition 5, which claims that all differential Nash equilibria in zero-sum games are locally asymptotically stable, again just relies on basic linear algebra and the definition of a differential Nash equilibrium.

Proof of Proposition 5. Consider a two player game $(f, -f)$ on $X_1 \times X_2 = \mathbb{R}^m$ with $X_i = \mathbb{R}^{m_i}$. For such a game,

$$J(x) = \begin{bmatrix} D_1^2 f(x) & D_{21} f(x) \\ -D_{12} f(x) & -D_2^2 f(x) \end{bmatrix}.$$

Note that $D_{21} f(x) = (D_{12} f(x))^T$. Suppose that $x = (x_1, x_2)$ is a differential Nash equilibrium and let $v = [v_1, v_2] \in \mathbb{R}^m$ with $v_1 \in \mathbb{R}^{m_1}$ and $v_2 \in \mathbb{R}^{m_2}$. Then, $v^T J(x) v = v_1^T D_1^2 f(x) v_1 - v_2^T D_2^2 f(x) v_2 > 0$ since $D_1^2 f(x) \succ 0$ and $-D_2^2 f(x) \succ 0$ for x, a differential Nash equilibrium. Since v is arbitrary, this implies that $J(x)$ is positive definite and hence, clearly non-degenerate. Thus, for two-player zero-sum games, all differential Nash equilibria are both non-degenerate differential Nash equilibria and locally asymptotically stable equilibria of $\dot{x} = -\omega(x)$ □

This result guarantees that the differential Nash equilibria of zero-sum games are isolated and exponentially attracting under the flow of $\dot{x} = -\omega(x)$. This in turn guarantees that simultaneous gradient-play has a local linear rate of convergence to all local Nash equilibria in all zero-sum continuous games. Thus, the answer to **Q1** is the context of zero-sum games is "yes", since all Nash equilibria are attracting for the gradient dynamics.

In fact, in the next result we strengthen Proposition 5 to show that not only are differential Nash equilibria attracting, but they are hyperbolic.

Proposition 6. *Consider a two-player, zero-sum continuous game $(f, -f)$. If x is a differential Nash equilibrium, it is non-degenerate, and furthermore, it is hyperbolic.*

Proof. It is enough to show that all differential Nash are hyperbolic since all hyperbolic equilibria correspond to a non-degenerate J. Further, just as we noted in the proof of Lemma 2, stationarity, definiteness, and non-degeneracy are coordinate-invariant properties. Thus, we simply treat the Euclidean case here.

By definition, at a differential Nash equilibrium x of a zero-sum game, $\omega(x) = 0$, $D_1^2 f(x) > 0$, and $-D_2^2 f(x) > 0$. Further, in zero-sum games, $D_{12}^2 f = (D_{21}^2 f)^T$. Thus, the bilinear map J, takes the form

$$\begin{aligned} J(x) &= \begin{bmatrix} D_1^2 f(x) & D_{12} f(x) \\ -D_{21} f(x) & -D_2^2 f(x) \end{bmatrix} \\ &= \begin{bmatrix} D_1^2 f(x) & D_{12} f(x) \\ -D_{12}^T f(x) & -D_2^2 f(x) \end{bmatrix}. \end{aligned}$$

Let (λ, v) be an eigenpair of $J(x)$. The real part of λ, denoted $\mathrm{Re}(\lambda)$, can be written as

$$\begin{aligned}\mathrm{Re}(\lambda) &= \tfrac{1}{2}(\lambda + \bar{\lambda}) = \tfrac{1}{2}(v^* J^T(x)v + v^* J(x)v) \\ &= \tfrac{1}{2} v^*(J^T(x) + J(x))v \\ &= \frac{1}{2} v^* \begin{bmatrix} D_1^2 f(x) & 0 \\ 0 & -D_2^2 f(x) \end{bmatrix} v > 0\end{aligned}$$

where the last line follows from the positive definiteness of $\mathrm{diag}(D_1^2 f(x), -D_2^2 f(x))$ at a differential Nash equilibrium. Hence, x is hyperbolic therefore $\det(J(x)) \neq 0$. □

The above proposition provides a strong result for the class of zero-sum games. In particular, simply due to the structure of J, all differential Nash have the nice property of being hyperbolic, and hence, exponentially attracting under gradient-play dynamics—that is, $\dot{x} = -\omega(x)$ or its discrete time variant $x^+ = x - \gamma\omega(x)$ for appropriately chosen stepsize γ. Note that numerous learning algorithms in machine learning applications of zero-sum games take this form (see, e.g., [68, 113, 116]).

Unfortunately, it is not true that every locally asymptotically stable equilibrium in two-player zero-sum games are non-degenerate differential Nash equilibria. Indeed, there may be many locally asymptotically stable equilibria in a zero-sum game that are not local Nash equilibria. The following proposition highlights this fact.

Proposition 7. *In the class of zero-sum continuous games, there exists a continuum of games such that for each game $\mathcal{G}$, $\texttt{LASE}(\omega) \not\subset \texttt{DNE}(\mathcal{G}) \subset \texttt{LNE}(\mathcal{G})$.*

Proof. Consider the two-player zero-sum game $(f, -f)$ on $\mathbb{R}^2$ where

$$f(x_1, x_2) = \frac{a}{2}x_1^2 + bx_1x_2 + \frac{c}{2}x_2^2;$$

and $a, b, c \in \mathbb{R}$. The Jacobian of ω is given by

$$J(x_1, x_2) = \begin{bmatrix} a & b \\ -b & -c \end{bmatrix}, \quad \forall\ (x_1, x_2) \in \mathbb{R}^2.$$

If $a > c > 0$ and $b^2 > ac$, then $J(x_1, x_2)$ has eigenvalues with strictly positive real part, but the unique stationary point is not a differential Nash equilibrium—since $-c < 0$—and, in fact, is not even a Nash equilibrium. Indeed,

$$-f(0,0) > -f(0, x_2) = -\frac{c}{2}x_2^2, \quad \forall\ x_2 \neq 0.$$

Thus, there exists a continuum of zero-sum games with a large set of locally asymptotically stable equilibria of the corresponding dynamics $\dot{x} = -\omega(x)$ that are not differential Nash. □

The, preceding proposition again shows that there exists non-Nash equilibria of the gradient dynamics in zero-sum continuous games. Thus, this proposition also provides a negative answer to **Q2** in the context of zero-sum games.

2.2 Differential Topology of Local Nash Equilibria

In this section we analyze local Nash equilibria using tools from differential topology. A line of previous work [155, 156, 157] characterized properties of local Nash equilibria in 'generic' general-sum continuous games. In particular they showed that in 'almost all' general-sum games, local Nash equilibria are in fact differential Nash equilibria (meaning they are isolated and strict local Nash equilibria), and that these equilibria satisfy robustness in the sense of 'structural stability'.

In this section, we take a similar approach in the context of zero-sum games where we show that local Nash equilibria have several important properties which make them particularly amenable to computation. Since the set of zero-sum games is of zero measure in the space of general-sum continuous games the results of this section are not a direct implication of previous results. Furthermore, with the growing interest in gradient-play in zero-sum games from the machine learning community [14, 44, 78], such an understanding of Nash equilibria is becoming increasingly crucial.

In particular, most recent convergence analysis on gradient-play in min-max optimization depends on an assumption of *local convexity in the game space* around a local Nash equilibrium—that is, nearby Nash equilibria the Jacobian of the gradient-based learning rule is assumed to be locally positive definite (e.g., [14, 45, 62]).

This implicit structural assumption gives rise two natural questions:

- Is this assumption satisfied for 'almost all zero-sum games' in the sense of *genericity*—i.e., how restrictive is this assumption?;
- Is this a 'robust' assumption in the sense of *structural stability*—i.e., does the property persist under smooth perturbations to the game?

Building on the work in [155, 156, 157] on understanding these questions in general-sum games, this section addresses these two questions in the context of zero-sum games where the additional structure allows for stronger results.

In particular, the results of this section are summarized as follows:

1. We prove that differential Nash equilibria are generic amongst local Nash equilibria in continuous zero-sum games (Theorem 2). This implies that almost all zero-sum games played on continuous functions admit local Nash equilibria that are strict and isolated.
2. Combining this fact with the fact that all differential Nash equilibria are hyperbolic in zero-sum games (Proposition 6 from Section 2.1), we show that local Nash equilibria are generically hyperbolic (Corollary 1) which implies strong local convergence guarantees in 'almost all' zero-sum games.
3. We prove that zero-sum games are structurally stable (Theorem 3); that is, the structure of the game—and hence, its equilibria—is robust to smooth perturbations within the space of zero-sum games.

We remark that 2. is a much stronger statement than the genericity of differential Nash equilibria shown in [157]. This is only possible by exploiting the additional structure of zero-sum games, namely the fact that they are defined completely in terms of a single sufficiently smooth function—i.e., given $f \in C^r(X, \mathbb{R})$, the corresponding zero-sum game is $(f, -f)$.

Remark 2. *many recent works have made the assumption of hyperbolicity of local Nash equilibria without a thorough understanding of whether or not such an assumption is restrictive (see e.g. [15, 44, 63, 72, 78]). The results in this section show that this assumption simply* rules out a measure zero set of zero-sum games.

Preliminaries

Before developing the main results of this section, we present our general setup, as well as some preliminary game theoretic and mathematical definitions. To deal with the situation where players' action spaces are more general manifolds we re-introduce many of the objects from Chapter 1 and Section 2.1 with the necessary mathematical rigor. We first introduce the necessary mathematical preliminaries. An interested reader should see standard references for a more detailed introduction [5, 97].

Mathematical Preliminaries

Throughout this section we consider full information continuous, two-player zero-sum games. Each player $i \in \mathcal{I} = \{1, 2\}$ selects an action x_i from a *topological space* X_i in order to minimize its cost $f_i : X \to \mathbb{R}$ where $X = X_1 \times X_2$ is the *joint strategy space* of all the agents. Note that f_i depends on x_{-i} which is the collection of actions of all other agents excluding agent i—that is, $f_i : (x_i, x_{-i}) \mapsto f_i(x_i, x_{-i}) \in \mathbb{R}$. Furthermore, each X_i can be finite-dimensional smooth manifolds or infinite-dimensional Banach manifolds.

A *smooth manifold* is a topological manifold with a smooth atlas. In particular, we use the term *manifold* generally; we specify whether it is a finite– or infinite–dimensional manifold only when necessary. If a covering by charts takes their values in a Banach space E, then E is called the *model space* and we say that X is a *C^r–Banach manifold*. For a vector space E, we define the vector space of continuous $(r + s)$–multilinear maps $T^r_s(E) = L^{r+s}(E^*, \dots, E^*, E, \dots, E; \mathbb{R})$ with s copies of E and r copes of E^* and where E^* denotes the dual. Elements of $T^r_s(E)$ are *tensors* on E, and $T^r_s(X)$ denotes the vector bundle of such tensors [5, Definition 5.2.9].

Suppose $f : X \to M$ is a mapping of one manifold X into another M. Then, we can interpret the derivative of f on each chart at x as a linear mapping $df(x) : T_xX \to T_{f(x)}M$. When $M = \mathbb{R}$, the collection of such maps defines a *1–form* $df : X \to T^*X$. Indeed, a 1–form is a continuous map $\omega : X \to T^*X$ satisfying $\pi \circ \omega = \mathrm{Id}_X$ where $\pi : T^*X \to X$ is the natural projection mapping $\omega(x) \in T^*_xX$ to $x \in X$.

At a critical point $x \in X$ (i.e., where $df(x) = 0$), there is a uniquely determined continuous, symmetric bilinear form $d^2f(x) \in T^0_2(X)$ such that $d^2f(x)$ is defined for all $v, w \in T_xX$

by $d^2(f \circ \varphi^{-1})(\varphi(x))(v_\varphi, w_\varphi)$ where φ is any product chart at x and v_φ, w_φ are the local representations of v, w respectively [142, Proposition in §7]. We say $d^2 f(x)$ is *positive semi–definite* if there exists $\alpha \geq 0$ such that for any chart φ,

$$d^2(f \circ \varphi^{-1})(\varphi(x))(v, v) \geq \alpha \|v\|^2, \quad \forall\ v \in T_{\varphi(x)}E. \tag{2.5}$$

If $\alpha > 0$, then we say $d^2 f(x)$ is *positive–definite*. Both critical points and positive definiteness are invariant with respect to the choice of coordinate chart.

Now, consider smooth manifolds X_1, X_2. The product space $X_1 \times X_2$ is naturally a smooth manifold [5, Definition 3.2.4]. Further, there is a canonical isomorphism at each point such that the cotangent bundle of the product manifold splits:

$$T^*_{(x_1,x_2)}(X_1 \times X_2) \cong T^*_{x_1}X_1 \oplus T^*_{x_2}X_2 \tag{2.6}$$

where $\oplus$ denotes the direct sum of vector spaces. There are natural bundle maps $\psi_{X_1} : T^*(X_1 \times X_2) \to T^*(X_1 \times X_2)$ annihilating the all the components other than those corresponding to X_i of an element in the cotangent bundle. In particular, $\psi_{X_1}(\omega_1, \omega_2) = (0_1, \omega_2)$ and $\psi_{X_2}(\omega_1, \omega_2) = (\omega_1, 0_2)$ where $\omega = (\omega_1, \omega_2) \in T^*_x(X_1 \times X_2)$ and $0_j \in T^*_{x_j}X_j$ for each $j \in \{1, 2\}$ is the zero functional.

For smooth manifolds X and Y of dimension n_x and n_y respectively, an *k–jet* from X to Y is an equivalence class $[x, f, U]_k$ of triples (x, f, U) where $U \subset X$ is an open set, $x \in U$, and $f : U \to Y$ is a C^k map. The equivalence relation satisfies $[x, f, U]_k = [y, g, V]_k$ if $x = y$ and in some pair of charts adapted to f at x, f and g have the same derivatives up to order k. We use the notation $[x, f, U]_k = j^k f(x)$ to denote the k–jet of f at x. The set of all k–jets from X to Y is denoted by $J^k(X, Y)$. The jet bundle $J^k(X, Y)$ is a smooth manifold (see [168] Chapter 2 for the construction). For each C^k map $f : X \to Y$ we define a map $j^k f : X \to J^k(X, Y)$ by $x \mapsto j^k f(x)$ and refer to it as the *k–jet extension*.

Definition 12. *Let X, Y be smooth manifolds and $f : X \to Y$ be a smooth mapping. Let Z be a smooth submanifold of Y and p a point in X. Then f* intersects Z transversally at p *(denoted $f \pitchfork Z$ at p) if either $f(p) \notin Z$ or $f(p) \in Z$ and $T_{f(p)}Y = T_{f(p)}Z + (f_*)_p(T_pX)$.*

For $1 \leq k < s \leq \infty$ consider the jet map $j^k : C^s(X, Y) \to C^{s-k}(X, J^k(X, Y))$ and let $Z \subset J^k(X, Y)$ be a submanifold. Define

$$\bigpitchfork^s(X, Y; j^k, Z) = \{h \in C^s(X, Y) |\ j^k h \pitchfork Z\}. \tag{2.7}$$

A subset of a topological space X is *residual* if it contains the intersection of countably many open–dense sets. We say a property is *generic* if the set of all points of X which possess this property is residual [30]. Given these definitions, we now present two key results from differential geometry that are crucial to our results.

Theorem 1. *(Jet Transversality Theorem, Chap. 2 [168]). Let X, Y be C^∞ manifolds without boundary, and let $Z \subset J^k(X, Y)$ be a C^∞ submanifold. Suppose that $1 \leq k < s \leq \infty$. Then, $\bigpitchfork^s(X, Y; j^k, Z)$ is residual and thus dense in $C^s(X, Y)$ endowed with the strong topology, and open if Z is closed.*

Proposition 8. *(Chap. II.4, Proposition 4.2 [65]). Let X, Y be smooth manifolds and $Z \subset Y$ a submanifold. Suppose that $\dim X < \text{codim} Z$. Let $f : X \to Y$ be smooth and suppose that $f \pitchfork Z$. Then, $f(X) \cap Z = \emptyset$.*

The Jet Transversality Theorem and Proposition 8 can be used to show a subset of a jet bundle having a particular set of desired properties is generic. Indeed, consider the jet bundle $J^k(X, Y)$ and recall that it is a manifold that contains jets $j^k f : X \to J^k(X, Y)$ as its elements where $f \in C^k(X, Y)$. Let $Z \subset J^k(X, Y)$ be the submanifold of the jet bundle that *does not* possess the desired properties. If $\dim X < \text{codim } Z$, then for a generic function $f \in C^k(X, Y)$ the image of the k–jet extension is disjoint from Z implying that there is an open–dense set of functions having the desired properties. It is exactly this approach we use to show the genericity of non-degenerate differential Nash equilibria of zero-sum continuous games.

Setup

Given the preliminaries we now re-define many of the objects from previous sections with the necessary mathematical rigor. To begin, we formally define the object ω as a differential 1-form $\omega : X \to T^*X$ defined by

$$\omega = \psi_{X_1} \circ df - \psi_{X_2} \circ df$$

where ψ_{X_i} are the natural bundle maps $\psi_{X_1} : T^*X \to T^*X$ that annihilate those components of the co-vector field df corresponding to X_1 and analogously for ψ_{X_2}. Note that when each X_i is a finite dimensional manifold of dimensions m_i (e.g., Euclidean space $\mathbb{R}^{m_i}$), then the differential game form has the coordinate representation,

$$\omega_\psi = \sum_{j=1}^{m_1} \frac{\partial(f \circ \psi^{-1})}{\partial y_1^j} dy_1^j + \sum_{j=1}^{m_2} \frac{\partial(-f \circ \psi^{-1})}{\partial y_2^j} dy_2^j,$$

for product chart (U, ψ) in X at $x = (x_1, \ldots, x_n)$ with local coordinates $(y_1^1, \ldots, y_1^{m_1}, y_2^1, \ldots, y_2^{m_2})$ and where $U = U_1 \times U_2$ and $\psi = \psi_1 \times \psi_2$. This form captures a differential view of the strategic interaction between the players. Note that each player's cost function depends on its own choice variable as well as all the other players' choice variables. However, each player can only affect their payoff by adjusting their own strategy.

We also define, J the generalization of the Jacobian, as the bilinear map induced by $d\omega$ which is composed of the partial derivatives of components of ω, where $d\omega = d(\psi_{X_1} \circ df) - d(\psi_{X_2} \circ df)$. Intrinsically, $d\omega \in T_2^0(X)$ is a tensor field; at a point x where $\omega(x) = 0$, it determines a bilinear form constructed from the uniquely determined continuous, symmetric, bilinear forms $\{d^2 f_i(x)\}_{i=1}^n$. For example, consider a two-player, zero-sum game $(f_1, f_2) = (f, -f)$ on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. In this case the matrix representation of this bilinear map is given by

$$J(x) = \begin{bmatrix} D_1^2 f(x) & D_{12} f(x) \\ -D_{12}^T f(x) & -D_2^2 f(x) \end{bmatrix},$$

which is exactly the Jacobian we analyzed previously.

Given these definitions the definition of critical points for the game is as before. Further the definitions of hyperbolicity, and non-degeneracy of critical points are the same as in Section 2.1. We note that even in the more general manifold setting, these definitions are invariant with respect to the coordinate chart [30, 155].

Theoretical Results

Given this setup, we specialize the results in [156] and [155] on genericity and structural stability of differential Nash equilibria to the class of zero-sum games.

Genericity

To develop the proof that local Nash equilibria of zero-sum games are generically non-degenerate differential Nash equilibria, we leverage the fact that it is a generic property of sufficiently smooth functions that all critical points are non-degenerate.

Lemma 1 ([30, Chapter 1])**.** *For C^r functions, $r \geq 2$ on $\mathbb{R}^n$, or on a manifold, it is a generic property that all the critical points are non-degenerate.*

The above lemma implies that for a generic function $f \in C^r(X, \mathbb{R})$ on an m–dimensional manifold X, the Hessian

$$H(x) = \begin{bmatrix} D_1^2 f(x) & \cdots & D_{1m} f(x) \\ \vdots & \ddots & \vdots \\ D_{m1} f(x) & \cdots & D_m^2 f(x) \end{bmatrix}$$

is non-degenerate at critical points—that is, $\det(H(x)) \neq 0$. Since a zero-sum game is completely characterized b a function f, and the critical points of f are the same as critical points of the game, we then show that all critical points of gradient-play in zero-sum games are generically non-degenerate.

Lemma 2. *Consider $f \in C^r(X, \mathbb{R})$ and the zero-sum game $(f, -f)$. For any critical point $x \in X$ (i.e., $x \in \{x \in X |\ \omega(x) = 0\}$), $\det(H(x)) \neq 0 \iff \det(D\omega(x)) \neq 0$.*

Proof. Before proceeding we note that in the case that X is a smooth manifold, the stationarity of critical points and definiteness of H and $D\omega$ are coordinate-invariant properties and hence, independent of coordinate chart [30, 155, 156, 157]. Thus, to shorten the presentation of proofs, we simply treat the Euclidean case here; showing the more general case simply requires selecting a coordinate chart defined on a neighborhood of the differential Nash, showing the results with respect to this chart, and then invoking coordinate invariance.

Let $x = (x_1, x_2)$ where $X = X_1 \times X_2$ and X_i is m_i–dimensional. Note that $H(x)$ is equal to $D\omega(x)$ with the last m_2 rows scaled each by -1. Indeed,

$$D\omega(x) = \begin{bmatrix} D_1^2 f(x) & D_{12} f(x) \\ -D_{12}^T f(x) & -D_2^2 f(x) \end{bmatrix}$$

where $D_i^2 f(x)$ is $m_i \times m_i$ dimensional for each $i \in \{1, 2\}$ and $D_{12} f(x)$ is $m_1 \times m_2$ dimensional. Clearly, $D\omega(x) = PH(x)$ where $P = \text{blockdiag}(I_{m_1}, -I_{m_2})$ with each I_{m_i} the $m_i \times m_i$ identity matrix, so that $\det(H(x)) = (-1)^{m_2} \det(D\omega(x))$. Hence, the result holds. □

This equivalence between the non-degeneracy of the Hessian and the game Jacobian $D\omega$ allows us to lift the fact that non-degeneracy of critical points is a generic property to zero-sum games. This leads to the following, strong result that non-degenerate differential Nash equilibria are generic in zero-sum games.

Theorem 2. *For two-player, zero-sum continuous games, non-degenerate differential Nash are generic amongst local Nash equilibria. That is, given a generic $f \in C^r(X, \mathbb{R})$, all local Nash equilibria of the game $(f, -f)$ are (non-degenerate) differential Nash equilibria.*

Proof. First, critical points of a function f are those such that $(D_1 f_1(x)\ D_2 f_2(x)) = 0$ and hence they coincide with critical points of the zero-sum game—i.e., those points x such that $\omega(x) = (D_1 f(x), -D_2 f(x)) = 0$. By Lemma 2, for any critical point x, $\det(H(x)) = 0$ if and only if $\det(D\omega(x)) = 0$. Hence, critical points of f are non-degenerate if and only if critical points of the zero-sum game are non-degenerate.

Consider a generic function f and the corresponding zero-sum game $(f, -f)$. If X is a smooth manifold, let (U, φ) be a product chart on $X_1 \times X_2$ that contains x. Suppose that x is a local Nash equilibrium so that $\omega(x) = 0$ and $D_1^2 f(x) \geq 0$ and $-D_2^2 f(x) \geq 0$. By the above argument, since f is generic and the critical points of f coincide with those of the zero-sum game, $\det(D\omega(x)) \neq 0$. By Lemma 1, critical points of a generic zero-sum game are non-degenerate. That is, there exists an open-dense set of functions f in $C^r(X, \mathbb{R})$ such that critical points of the corresponding game are non-degenerate.

Let $J^2(X, \mathbb{R})$ denote the second-order jet bundle containing 2–jets $j^2 f$ such that $f : X \to \mathbb{R}$. Then, $J^2(X, \mathbb{R})$ is locally diffeomorphic to

$$\mathbb{R}^m \times \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^{\frac{m(m+1)}{2}}$$

and the 2–jet extension of f at any point $x \in X$ in coordinates is given by

$$(\varphi(x), (f \circ \varphi^{-1})(\varphi(x)), D^\varphi f(x), (D^\varphi)^2 f(x))$$

where $D^\varphi f = [D_1^\varphi f\ D_2^\varphi f]$ with $D_j^\varphi = [\partial(f \circ \varphi^{-1})/(\partial y_j^1)\ \cdots\ \partial(f \circ \varphi^{-1})/(\partial y_j^{m_i})]$ and similarly for $(D^\varphi)^2 f$. Again, we note that the properties of interest (stationarity, definiteness, and non-degeneracy) are known to be coordinate invariant.

Consider a subset of $J^2(X, \mathbb{R})$ defined by

$$\mathcal{D} = \mathbb{R}^m \times \mathbb{R} \times \{0_m\} \times Z(m_1) \times \mathbb{R}^{m_1 \times m_2} \times Z(m_2)$$

where $Z(m_i)$ is the subset of symmetric $m_i \times m_i$ matrices such that for $A \in Z(m_i)$, $\det(A) = 0$. Each $Z(m_i)$ is algebraic and has no interior points; hence, we can use the Whitney stratification theorem [61, Chapter 1, Theorem 2.7] to get that each $Z(m_i)$ is the

union of submanifolds of co-dimension at least 1. Hence $\mathcal{D}$ is the union of submanifolds and has co-dimension at least $m+2$. Applying the Jet Transversality Theorem (Theorem 1) and Proposition 8 yields an open-dense set of functions f such that when $\omega(x)=0$, $\det(D_i^2 f(x)) \neq 0$, for $i=1,2$.

Now, the intersection of two open-dense sets is open-dense so that we have an open-dense set of functions f in $C^r(X,\mathbb{R})$ such that when $\omega(x)=0$, $\det(D_i^2 f(x)) \neq 0$ for each $i \in \{1,2\}$ and $\det(D\omega(x)) \neq 0$. This, in turn, implies that there is an open-dense set $\mathcal{F}$ of functions f in $C^r(X,\mathbb{R})$ such that for zero-sum games constructed from these functions, local Nash equilibria are non-degenerate differential Nash equilibria. Indeed, consider an $f \in \mathcal{F}$ in this set such that x is a local Nash equilibrium of $(f,-f)$. Then necessary conditions for Nash imply that $\omega(x)=0$, $D_1^2 f(x) \geq 0$ and $-D_2^2 f(x) \geq 0$. However, since $f \in \mathcal{F}$, $\det(D_1^2 f(x)) \neq 0$ and $\det(-D_2^2 f(x)) = (-1)^{m_2}\det(D_2^2 f(x)) \neq 0$. Hence, x is a differential Nash equilibrium. Moreover, since $f \in \mathcal{F}$, $\det(H(x)) \neq 0$ which is equivalent to $\det(D\omega(x)) \neq 0$ (by Lemma 2). Thus, x is a non-degenerate differential Nash. □

As shown in Proposition 6, all differential Nash for zero-sum games are non-degenerate simply by the structure of $D\omega$. This further implies that local Nash equilibria are generically hyperbolic critical points, meaning there are no eigenvalues of $D\omega$ with zero real part.

Corollary 1. *Within the class of two-player zero-sum continuous games, local Nash equilibria are generically hyperbolic critical points.*

Proof. Consider a two-player, zero-sum game $(f,-f)$ for some generic sufficiently smooth $f \in C^r(X,\mathbb{R})$. Then, by Theorem 2, a local Nash equilibria x is a differential Nash equilibria. Moreover, by Proposition 6, x is hyperbolic so that all eigenvalues of $D\omega(x)$ must have strictly positive real parts. This implies that all such points are hyperbolic critical points of the gradient dynamics $\dot{x} = -\omega(x)$. □

Remark 3. *Genericity gives a formal mathematical rigor to the term 'almost all' for a certain property—in this case, non-degeneracy and further hyperbolicity of local Nash. Thus, corollary 1 implies that in 'almost all' zero-sum games, local Nash equilibria are locally exponentially attracting for gradient-play (or gradient-descent-ascent as it is known in the min-max optimization literature).*

Structural Stability

In this section we show that local Nash equilibria in generic zero-sum games are structurally stable, meaning that they persist under smooth perturbations within the class of zero-sum games.

Theorem 3. *For zero-sum games, differential Nash equilibria are structurally stable: given $f \in C^r(X_1 \times X_2, \mathbb{R})$, $g \in C^r(X_1 \times X_2, \mathbb{R})$, and a differential Nash equilibrium $(x_1,x_2) \in X_1 \times X_2$, there exists a neighborhoods $U \subset \mathbb{R}$ of zero and $V \subset X_1 \times X_2$ such that for all*

$t \in U$ there exists a unique differential Nash equilibrium $(\tilde{x}_1, \tilde{x}_2) \in V$ for the zero-sum game $(f + tg, -f - tg)$.

Proof. Define the smoothly perturbed cost function $\tilde{f} : X_1 \times X_2 \times \mathbb{R} \to \mathbb{R}$ by $\tilde{f}(x, y, t) = f(x, y) + tg(x, y)$, and its differential game form $\tilde{\omega} : X_1 \times X_2 \times \mathbb{R} \to T^*(X_1 \times X_2)$ by

$$\tilde{\omega}(x, y, t) = (D_1(\tilde{f}(x, y) + tg(x, y), -D_2(\tilde{f}(x, y) + tg(x, y)),$$

for all $t \in \mathbb{R}$ and $(x, y) \in X_1 \times X_2$.

Since (x_1, x_2) is a differential Nash equilibrium, $D\tilde{\omega}(x, y, 0)$ is necessarily non-degenerate (see the proof of Corollary 1). Invoking the implicit function theorem [97], there exists neighborhoods $V \subset \mathbb{R}$ of zero and $W \subset X_1 \times X_2$ and a smooth function $\sigma \in C^r(V, W)$ such that for all $t \in V$ and $(x_1, x_2) \in W$,

$$\tilde{\omega}(x_1, x_2, s) = 0 \iff (x_1, x_2) = \sigma(t).$$

Since $\tilde{\omega}$ is continuously differentiable, there exists a neighborhood $U \subset W$ of zero such that $D\tilde{\omega}(\sigma(t), t)$ is invertible for all $t \in U$. Thus, for all $t \in U$, $\sigma(t)$ must be the unique Nash equilibrium of $(f + tg|_W, -f - tg|_W)$. □

We note that both the genericity and structural stability results follow largely from the fact that the class of two-player zero-sum games are defined completely in terms of a single (sufficiently) smooth function $f \in C^r(X, \mathbb{R})$, so that its fairly straightforward to lift the properties of genericity and structural stability to the class of zero-sum games from the class of smooth functions. We also remark that the perturbations considered here are those such that the game remains in the class of zero-sum games; that is, the function f is smoothly perturbed and this induces the perturbed zero sum game $(f + tg, -f - tg)$.

Examples

To illustrate the implications of structural stability, we provide a simple example. Consider a classic set of zero-sum continuous games known as biliear games. Such games have similar characteristics as bimatrix games played on the simplex; in particular, bimatrix games have the same cost structure as bilinear games where the stratgegy space of the former is considered to be a probability distribution over the finite set of pure strategies. This is particularly interesting since it demonstrates that interior equilibria of such games can be altered arbitrarily small perturbations.

Example 1. *Consider two-players with decision variables $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$ respectively, playing a zero-sum game on the function:*

$$f(x, y) = x^T A y$$

Where $A \in \mathbb{R}^{d_x \times d_y}$. The x player would like to minimize f while the y player would like to maximize it. Looking at ω for this game, we can see that the local Nash equilibria live in $\mathcal{N}(A) \times \mathcal{N}(A^T)$, where $\mathcal{N}(A)$ and $\mathcal{N}(A^T)$ denote the nullspaces of A and A^T respectively:

$$\omega(x, y) = \begin{bmatrix} Ay \\ -A^T x \end{bmatrix}$$

We note that the local Nash equilibria are not differential Nash equilibria, and that $D\omega$ has purely imaginary eigenvalues everywhere since it is skew-symmetric. Thus the local Nash equilibria are non-hyperbolic and this a non-generic case. Letting $f_\epsilon = f(x, y) - \frac{\epsilon}{2}||x||^2$, we see that ω for this perturbed game (denoted ω_ϵ) has the form:

$$\omega_\epsilon(x, y) = \begin{bmatrix} Ay - \epsilon x \\ -A^T x \end{bmatrix}$$

This perturbation fundamentally changes the critical points, and looking at $D\omega_\epsilon$, we can see that for any $\epsilon > 0$, there are no more local Nash equilibria:

$$D\omega_\epsilon(0, 0) = \begin{bmatrix} -\epsilon I_{d_x} & A \\ -A^T & 0 \end{bmatrix}$$

Since any arbitrarily small perturbation of this form can cause all of the local Nash equilibria to change, these games cannot be structurally stable.

We now show how this behavior extends to more complicated settings. Specifically we present an example of a game of rock-paper-scissors where both players have stochastic policies over the three actions which are parametrized by weights. The following example highlights how this classic problem is non-generic and the behavior changes drastically when the loss is perturbed in a small way.

Example 2. *Consider the game of rock-paper-scissors where each player has three actions $\{0, 1, 2\}$, with payoff matrix:*

$$M = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix}$$

Each player $i \in \{1, 2\}$ has a policy or mixed strategy π_i parametrized by a set of weights $\{w_{ij}\}_{j \in \{0,1,2\}}$ of the form:

$$\pi_i(j) = \frac{\exp(-\beta_i w_{ij})}{\sum_{k=0}^{2} \exp(-\beta_i w_{ij})}$$

Where β_i is a hyper-parameter for player i that determines the 'greediness' of their policy with respect to their set of weights. For simplicity, we treat π_i as a vector in $\mathbb{R}^3$. Each player would like to maximize their expected reward given by

$$f(w_1, w_2) = \pi_1^T M \pi_2.$$

We note that there is a continuum of local Nash equilibria for the policies $\pi_i = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ for $i \in \{1, 2\}$ and that this is achieved whenever each player has all of their weights equal.

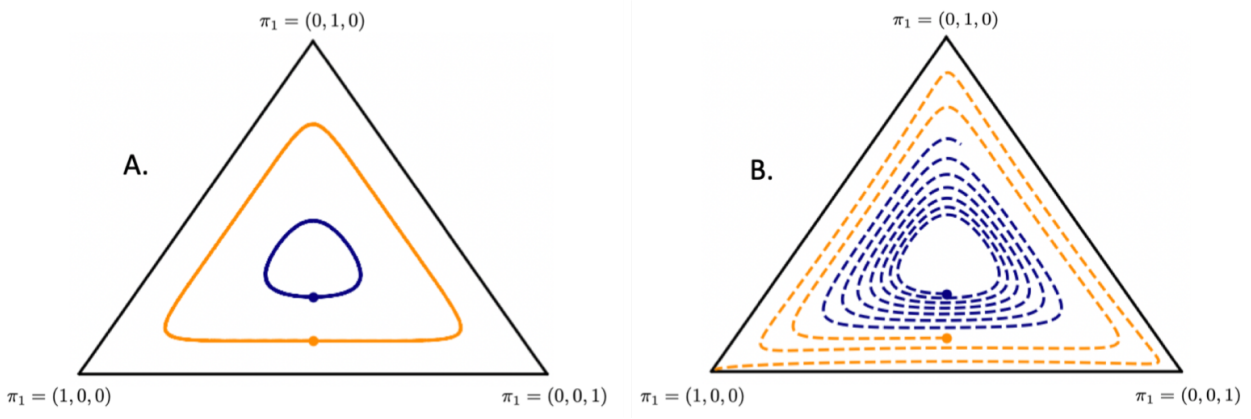


Figure 2.2: The trajectory of the policy of player 1 under gradient-play for A. rock-paper scissors and B. a perturbed version of rock-paper-scissors. A. Player 1 cycles around the local Nash equilibrium of $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ from either initialization (shown with circles). We remark that player 1's time average policy is in fact $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. B. Player 1 diverges from the local Nash equilibrium from either initialization for the perturbed game given by (2.8).

In Fig. 2.2 we show the trajectories of the policy of player 1, when $\beta_1 = \beta_2 = 1$ and both players use gradient descent to update their weights at each iteration. In Figure 2.2A. we see that player 1 cycles around the local Nash equilibrium in policy space. In Figure 2.2B. we show the trajectories of the policy of player 1, starting from the same initializations, but for a perturbed version of the game defined by

$$f_\epsilon(w_1, w_2) = \pi_1^T M \pi_2 + \epsilon g(w_1, w_2) \tag{2.8}$$

where $\epsilon =$ 1e-3 and $g(x, y) = ||y||^2 - ||x||^2$. Here we can see that this relatively small perturbation causes a drastic change in the behavior where player 1 diverges from the Nash of the original game and converges to the sub-optimal policy of always playing action zero.

2.3 Chapter Summary

Most general-purpose learning algorithms are based on local information such as gradient updates, and as such, representations of Nash equilibria that are amenable to computation such as the differential Nash concept studied in this section are extremely relevant. Much of the existing convergence analysis for machine learning algorithms proceeds under the structural assumptions implicit in the definition of the differential Nash equilibrium concept. In this chapter, we show that characterizations such as these are generic and structurally stable and we further investigate the implications of their structure for gradient-play. In particular, in zero-sum games our results show that local Nash equilibria have very strong local guarantees of convergence for gradient-play in 'almost all' zero-sum games.

More generally, the main takeaways of this chapter are summarized in Figure 2.1. In particular, the results from both Sections 2.1 and 2.2 allow us to answer **Q2** from Chapter 1 for the class of gradient-based learning algorithms. For zero-sum games and general-sum games, Propositions 7 and 2 (combined with the genericity results from Section 2.2 and [157] respectively) shows that $\texttt{LNE}(\mathcal{G}) \subset \texttt{LASE}(\omega)$, meaning that there exist attracting, non-Nash equilibria for gradient-play. Thus, in these classes of games the answer to **Q2** (whether all attractors for gradient-play are relevant to the game) is "no". In potential games, however, since $\texttt{LASE}(\omega) \subset \texttt{LNE}(\mathcal{G})$ the answer is "yes".

In the next chapter, we provide answers to **Q1** (whether all game relevant equilibria can be reliably found using gradient-based algorithms) by showing that all local Nash equilibria in $\texttt{LNE}(\mathcal{G}) \cap \texttt{SSP}(\omega)$ are avoided almost surely by gradient-based algorithms. In particular, since $\texttt{LNE}(\mathcal{G}) \cap \texttt{SSP}(\omega) \neq \emptyset$ in potential and general-sum games, one cannot give a positive answer to **Q1** in either of these classes of games.

Chapter 3

Gradient-Based Learning in Continuous Games

In this chapter, we provide convergence and non-convergence results for gradient-based algorithms. Recall that agents are assumed to update their strategies *simultaneously* according to a gradient-based learning algorithm of the form

$$x_{i,t+1} = x_{i,t} - \gamma_{i,t} h_i(x_{i,t}, x_{-i,t}), \tag{3.1}$$

where $\gamma_{i,t}$ is agent i's step-size at iteration t, and the map h reflects one of two settings:

1. Each agent has *oracle access* to the gradient of its cost with respect to its own choice variable:
$$h_i(x_{i,t}, x_{-i,t}) = D_i f_i(x_{i,t}, x_{-i,t}),$$
where $D_i f_i \equiv \partial f_i / \partial x_i$ denotes the derivative of f_i with respect to x_i.

2. Each agent has an *unbiased estimator* of their gradient:
$$h_i(x_{i,t}, x_{-i,t}) = D_i f_i(x_{i,t}, x_{-i,t}) + w_{i,t+1},$$
where $\{w_{i,t}\}$ is a zero mean, finite variance stochastic process.

Throughout this section we refer to the former setting as *deterministic* gradient-based learning and the latter setting as *stochastic* gradient-based learning. In Section 3.1 we show that a large number of multi-agent learning algorithms and machine learning paradigms fit into one of these two settings.

Given the characterizations of Nash equilibria and their behaviors under the continuous-time flow $\dot{x} = -\omega(x)$ for

$$\omega(x) = (D_1 f_1(x), \ldots, D_N f_N(x)),$$

from Chapter 2, in this chapter we build upon the intuition that both settings outlined above should have the same limiting behavior as the limiting continuous-time ODE. Indeed, the

first can simply be seen as the forward Euler discretization of the ODE while the second is simply a stochastic approximation of the first. We make this intuition formal in subsequent sections.

Before doing so, we comment briefly on the large body of recent work on understanding gradient-play (and variants of gradient-play) in continuous games. As mentioned in Chapter 1, the machine learning community has recently begun adapting tools and techniques first developed in optimization to understanding how to solve min-max optimization problems (or equivalently zero-sum games) [44, 45, 62, 128], and more general classes of games [14, 29]. The vast majority of these works proceed under strong structural assumptions on the games (like e.g,., monotonicity or convexity) and analyze (sometimes stochastic) gradient-play. In the context of zero-sum games, many works in recent years have also analyzed the proximal point algorithm [53], and its approximations like extra-gradient algorithms [135] or optimistic gradient descent-ascent [45, 128] . We remark that this class of algorithms can simply be seen as the *backward* Euler discretization of $\dot{x} = -\omega(x)$, and as such, it retains the same essential behaviors as gradient-play with respect to Nash equilibria. Differences only appear in degenerate classes of games like bilinear games where the discretization has a large impact on the convergence of the algorithms.

In this chapter we proceed without making strong structural assumptions on the players losses and give global non-convergence guarantees to a subset of Nash equilibria for gradient-play. These results also extend to the analysis of proximal point algorithms as well.

3.1 Classes of Gradient-Based Learning Algorithms

The stochastic gradient-based learning setting we study is general enough to include a variety of commonly used multi-agent learning algorithms. The classes of algorithms we include is hardly an exhaustive list, and indeed many extensions and altogether different algorithms exist that can be considered members of this class. In this section, we provide a detailed analysis of these different algorithms including the derivation of the gradient-based update rules. In each of these cases, one can view an agent employing the given algorithm as building an unbiased estimate of their gradient from their observation of the environment. We note that the derivation of gradient-based approaches for multi-armed bandits can be found in [179] among other classic references on reinforcement learning.

The takeaways of this section are shown in Table 3.1 in which we provide the gradient-based update rule for six different example classes of learning problems: (i) gradient-play in non-cooperative continuous games, (ii) GANs, (iii) multi-agent policy gradient, (iv) individual Q-learning, (v) multi-agent gradient bandits, and (vi) multi-agent experts.

Online Optimization: Gradient Play in Non-Cooperative Games

We first show that classical online optimization algorithms fit into the framework we describe. In this case, each agent is directly trying to minimize its own function $f_i(x_i, x_{-i})$, which

Class	Gradient Learning Rule
Gradient-Play	$x_i^+ = x_i - \gamma_i D_i f_i(x_i, x_{-i})$
Training GANs	$\theta^+ = \theta - \gamma \mathbb{E}[D_\theta L(\theta, w)]$ $w^+ = w + \gamma \mathbb{E}[D_w L(\theta, w)]$
MA Policy Gradient	$x_i^+ = x_i - \gamma_i \mathbb{E}[D_i J_i(x_i, x_{-i})]$
Individual Q-learning	$q_i^+(u_i) = q_i(u_i) + \gamma_i(r_i(u_i, \pi_{-i}(q_i, q_{-i})) - q_i(u_i))$
MA Gradient Bandits	$x_{i,\ell}^+ = x_{i,\ell} + \gamma_i \mathbb{E}[\beta_i R_i(u_i, u_{-i}) \vert u_i = \ell]$, $\ell = 1, \ldots, m_i$
MA Experts	$x_{i,\ell}^+ = x_{i,\ell} + \gamma_i \mathbb{E}[R_i(u_i, u_{-i}) \vert u_i = \ell]$, $\ell = 1, \ldots, m_i$

Table 3.1: Example problem classes that fit into competitive gradient-based learning rules. Details on the derivation of these update rules as gradient-based learning schemes is provided in Section 3.1.

can depend on the current iterate of the other agents. There are many examples in the optimization literature of this type of setup. We note that in the full information case, the competitive gradient-based learning framework we describe here is simply *gradient play* [57], a very well-studied game-theoretic learning rule.

Of more interest are some gradient-free online optimization algorithms that also fit into the framework we describe. The game can be described as follows. At each iteration, t of the game, every player publishes its current iterate $x_{i,t}$. Player i, implementing this algorithm, then updates its iterate by taking a random unit vector u, and querying $f_i(x_i + \delta_i u, x_{-i})$. The update map is given by $x_{i,t+1} = x_{i,t} - \gamma_i f_i(x_i + \delta_i u, x_{-i})u$. It is shown in [54] that $f_i(x_i + \delta_i u, x_{-i})u$ is an unbiased estimate of the gradient of a smoothed version of f_i—i.e. $\hat{f}_i(x_i, x_{-i}) = \mathbb{E}_v[f_i(x + \delta v, x_{-i})]$. Thus the loss function being minimized by the agent is $\hat{f}_i$. In this case, the results on characterizing limiting behavior presented in Section 3.3 apply.

Training of Generative Adversarial Networks

Generative adversarial networks take a game theoretic approach to fitting a generative model in complex structured spaces. Specifically, they approach the problem of fitting a generative model from a data set of samples from some distribution $Q \in \Delta(Y)$ as a zero-sum game between a *generator* and a *discriminator*. In general, both the generator and the discriminator are modeled as deep neural networks. The generator network outputs a sample $G_\theta(z) \in Y$ in the same space Y as the sampled data set given a random noise signal $z \sim F$ as an input. The discriminator $D_w(y)$ tries to discriminate between a true sample and a sample generated by the generator—that is, it takes as input a sample y drawn from Q or the generator and tries to determine if its *real* or *fake*. The goal, is to find a Nash equilibrium of the zero-sum game

under which the generator will learn to generate samples that are indistinguishable from the true samples—i.e. in equilibrium, the generator has learned the underlying distribution.

To prevent instabilities in the training of GANs with zero-one discriminators, the Wasserstein GAN attempts to approximate the Wasserstein-1 metric between the true distribution and the distribution of the generator. In this setting, $D_w(\cdot)$ is a 1-Lipschitz function leading to the problem

$$\inf_\theta \sup_w \mathbb{E}_{y\sim Q}[D_w(y)] - \mathbb{E}_{z\sim F}[D_w(G_\theta(z))]$$

which has corresponding dynamics $w_{t+1} = w_t + \gamma \nabla_w L(\theta_t, w_t)$ and $\theta_{t+1} = \theta_t - \gamma \nabla_\theta L(\theta_t, w_t)$ where $L(\theta, w) = \mathbb{E}_{y\sim Q}[D_w(y)] - \mathbb{E}_{z\sim F}[D_w(G_\theta(z))]$ and where γ is the learning rate.

GANs are notoriously difficult to train. The typical approach is to allow each player to perform (stochastic) gradient descent on the derivative of their cost with respect to their own choice variable. There are two important observations about gradient-based learning approaches to training GANs that are relevant to this chapter. First, the equilibrium that is sought is generally a saddle point and second, the dynamics of GANs are complex enough to admit limit cycles [121]. None-the-less, training GANs with gradient descent is still very common. We note that our results suggest that, on top of periodic orbits and oscillations, training GANs with gradient descent can result in convergence to non-Nash equilibria.

Multi-Agent Reinforcement Learning Algorithms

Consider a setting in which all agents are operating in an MDP. There is a shared state space $\mathcal{S}$. Each agent, indexed by $\mathcal{I} = \{1, \dots, n\}$ has its own action space U_i and reward function $R_i : \mathcal{S} \times U \to \Delta_{\mathbb{R}}$ where $U = U_1 \times \cdots \times U_n$. We note the reward functions could themselves be random, but for illustrative purposes we suppose they are deterministic. Finally, the dynamics of the MDP are described by a state transition kernel $P : \mathcal{S} \times U \to \Delta_{\mathcal{S}}$ and an initial state distribution P_0. Each agent i also has a policy, π_i, that returns a distribution over U_i for each state $s \in \mathcal{S}$. We define a trajectory of the MDP, τ as $\tau = \{(s_t, u_{i,t}, u_{-i,u})\}_{t=0}^{T-1}$. Thus, a trajectory is a finite sequence of states, the actions of each player in that state, and the reward agent i received in that state, where T is the time horizon. Given fixed policies we can define a distribution over the space of all trajectories Γ, namely $P_\Gamma(\pi)$, by

$$P_\Gamma(\tau; \pi) = P_0(s_0) \prod_{i\in\mathcal{I}} \pi_i(u_{i,0}|s_0) \cdots P(s_t|s_{t-1}, u_{t-1}) \prod_{i\in\mathcal{I}} \pi_i(u_{i,t}|s_t) \cdots$$

The goal of each single agent in this setup is to maximize its cumulative expected reward over a time horizon T. That is, the agent is trying to find a policy π_i so as to maximize some function, which in keeping with our general formulation in Section 1.1, we write as $-f_i$ since this problem is a maximization. When an agent is employing policy gradient in this MARL setup, we assume that its policy comes from a parametric class of policies parametrized by $x_i \in X_i \subset \mathbb{R}^{m_i}$. To simplify notation, we write the parametric policy as $\pi_i(x_i)$ where for each x_i, given an state s, $\pi_i(x_i)$ is a probability distribution on actions u_i which we denote by $\pi_i(x_i)(\cdot|s)$.

The policy gradient MARL algorithm can be reformulated in the competitive gradient-based learning framework. An agent i using policy gradient is trying to tune the parameters x_i of their policy to maximize their expected reward over a trajectory of length T. We define the reward of agent i over a trajectory of the MDP, $\tau \in \Gamma$, to be $\boldsymbol{R}_i(\tau) = \sum_{t=0}^{T-1} R_i(s_{t,i,t}, u_{-i,t})$. Thus, each agent's loss function f_i, in keeping with our notation, is given by $f_i(x_i, x_{-i}) = -J_i(\pi_i(x_i), \pi_{-i}) = -\mathbb{E}_{\tau \sim P_\Gamma(\pi)}[\boldsymbol{R}_i(\tau))]$. The actions of agent i in the continuous game framework described in previous sections are the parameters of its policy, and thus their action space is $X_i \subset \mathbb{R}^{m_i}$. We note that we have made no assumptions on the other player's actions x_{-i}. That is, they do not need to be employing the same parameterized policy class or exactly the same gradient-based update procedure; the only requirement is that they also be using a gradient based multi-agent learning algorithm, and that their actions give rise to a set of policies π_{-i} that govern the way they choose their actions in the MDP.

In the full information case, at each round, t of the game, a player plays according to $\pi_i(x_{i,t})$ for a time horizon T, and then performs a gradient update on their parameters where $D_i f_i(x_i, x_{-i}) = D_i J_i(\pi_i(x_i), \pi_{-i,t})$ is given by

$$D_i J_i(\pi_i(x_i), \pi_{-i}) = \mathbb{E}_{\tau \sim P_\Gamma(\pi)}\Big[\sum_{t=0}^{T-1} R_i(s_t, u_t) \sum_{j=0}^{t} \nabla_{x_i} \log \pi_i(x_i)(u_{i,j}|s_j)\Big] \tag{3.2}$$

The derivation of this gradient is exactly the same as that of classic policy gradient. From (3.2) it is clear that an unbiased estimate of the gradient can be constructed. At each time t in the policy gradient update procedure, agent i receives a T horizon roll-out, say $z_{i,t} = \{(s_k, u_{i,k}, r_{i,k})\}_{k=0}^{T-1}$, and constructs the unbiased estimate of the gradient—i.e. $\widehat{D_i J_i} = \sum_{k=0}^{T-1} r_{i,k}\big(\sum_{j=0}^{k} \nabla_{x_i} \log \pi_i(x_{i,t})(u_{i,j}|s_j)\big)$. We note that in this case, the agent does not need to know the policies of the other agents, or anything about the dynamics of the MDP. The agent can construct the estimator solely from the sequence of states, the reward they received in those states, and their own actions. With these two derivations of the gradient for the full information and gradient-free cases, policy gradient for MARL conforms to the competitive gradient-based learning framework and hence, the results of Section 3 apply under appropriate assumptions.

3.2 Convergence and Non-Convergence of Deterministic Gradient-Play

We first address convergence to equilibria in the *deterministic* setting in which agents have oracle access to their gradients at each time step. This includes the case where agents know their own cost functions f_i and observe their own actions as well as their competitors' actions—and hence, can compute the gradient of their cost with respect to their own choice variable.

Since we have assumed that each agent $i \in \mathcal{I}$ has their own *learning rate* (i.e. step sizes γ_i), the joint dynamics of all the players are given by

$$x_{t+1} = g(x_t) \tag{3.3}$$

where $g : x \mapsto x - \gamma \odot \omega(x)$ with $\gamma = (\gamma_i)_{i \in \mathcal{I}}$ and $\gamma > 0$ element-wise. By a slight abuse of notation, $\gamma \odot \omega(x_t)$ is defined to be element-wise multiplication of γ and $\omega(\cdot)$ where γ_1 is multiplied by the first m_1 components of $\omega(\cdot)$, γ_2 is multiplied by the next m_2 components, and so on.

We remark that this update rule immediately distinguishes gradient-based learning in games from gradient descent. By definition, the dynamics of gradient descent in single-agent settings always correspond to gradient flows —i.e x evolves according to an ordinary differential equation of the form $\dot{x} = -\nabla\phi(x)$ for some function $\phi : \mathbb{R}^d \to \mathbb{R}$. Outside of the class of *exact* potential games we defined in Chapter 2, the dynamics of players' actions in games are not afforded this luxury—indeed, J is not in general symmetric (which is a necessary condition for a gradient flow). This makes the potential limiting behaviors of $\dot{x} = -\omega(x)$ highly non-trivial to characterize in general-sum games.

The structure present in a gradient-flow implies strong properties on the limiting behaviors of x. In particular, it precludes the existence of limit cycles or periodic orbits (limiting behaviors of dynamical systems where the state of system cycles infinitely through a set of states with a finite period) and chaos (an attribute of nonlinear dynamical systems where the system's behavior can vary extremely due to slight changes in initial position) [168]. We note that both of these behaviors can occur in the dynamics of gradient-based learning algorithms in games[1].

Despite the wide breadth of behaviors that gradient dynamics can exhibit in competitive settings, we are still make statements about convergence (and non-convergence) to certain types of equilibria. To do so, we first make the following standard assumptions on the smoothness of the cost functions f_i and the magnitude of the agents' learning rates γ_i.

Assumption 1. *For each $i \in \mathcal{I}$, $f_i \in C^s(X, \mathbb{R})$ with $s \geq 2$, $\sup_{x \in X} \|J(x)\|_2 \leq L < \infty$, and $0 < \gamma_i < 1/L$ where $\|\cdot\|_2$ is the induced 2-norm.*

Given these assumptions, the following result rules out converging to strict saddle points.

Theorem 4. *Let $f_i : X \to \mathbb{R}$ and γ satisfy Assumption 1. Suppose that $X = X_1 \times \cdots \times X_N \subseteq \mathbb{R}^m$ is open and convex. If $g(X) \subset X$, the set of initial conditions $x \in X$ from which competitive gradient-based learning converges to strict saddle points is of measure zero.*

We remark that the above theorem holds for $X = X_1 \times \cdots \times X_N = \mathbb{R}^m$ in particular, since $g(X) \subset X$ holds trivially in this case. It is also important to note that, as we point out in

[1]The Van der Pol oscillator and Lorenz system (see e.g [168]) can be seen as the resulting gradient dynamics in a 2-player and 3-player general-sum game respectively. The first is a classic example of a system where players converge to cycles and the second is an example of a chaotic system.

Chapter 2, local Nash equilibria can be strict saddle points. Thus, all local Nash equilibria that are strict saddle points for $\dot{x} = -\omega(x)$ are avoided almost surely by gradient-play even with oracle gradient access and random initializations. This holds even when players randomly initialize uniformly in an arbitrarily small ball around such Nash equilibria. In Chapter 4, we show that many linear quadratic dynamic games have a strict saddle point as their global Nash equilibrium.

The core of the proof of Theorem 4 is the celebrated stable manifold theorem from dynamical systems theory. Using the notation $\phi^t = \phi \circ \cdots \circ \phi$ to denote the t–times composition of a function $\phi : \mathbb{R}^d \to \mathbb{R}^d$, we present the theorem below for completeness.

Theorem 5 (Center and Stable Manifolds [175, Theorem III.7], [177])**.** *Let x_0 be a fixed point for the C^r local diffeomorphism $\phi : U \to \mathbb{R}^d$ where $U \subset \mathbb{R}^d$ is an open neighborhood of x_0 in $\mathbb{R}^d$ and $r \geq 1$. Let $E^s \oplus E^c \oplus E^u$ be the invariant splitting of $\mathbb{R}^d$ into generalized eigenspaces of $D\phi(x_0)$ corresponding to eigenvalues of absolute value less than one, equal to one, and greater than one. To the $D\phi(x_0)$ invariant subspace $E^s \oplus E^c$ there is an associated local ϕ–invariant C^r embedded disc W^{cs}_{loc} called the local stable center manifold of dimension $\dim(E^s \oplus E^c)$ and ball B around x_0 such that $\phi(W^{cs}_{loc}) \cap B \subset W^{cs}_{loc}$, and if $\phi^t(x) \in B$ for all $t \geq 0$, then $x \in W^{sc}_{loc}$.*

Importantly, this theorem allows us to characterize— locally— the set of initial conditions under which dynamics converge to equilibria. This is a crucial step in the proof of Theorem 4.

Proof of Theorem 4. The proof is composed of two parts: (a) the map g is a diffeomorphism, and (b) application of the stable manifold theorem to conclude that the set of initial conditions is measure zero.

(a) g is diffeomorphism We claim the mapping $g : \mathbb{R}^m \to \mathbb{R}^m$ is a diffeomorphism. If we can show that g is invertible and a local diffeomorphism, then the claim follows. Consider $x \neq y$ and suppose $g(y) = g(x)$ so that $y - x = \gamma \cdot (\omega(y) - \omega(x))$. The assumption $\sup_{x \in \mathbb{R}^m} \|J(x)\|_2 \leq L < \infty$ implies that ω satisfies the Lipschitz condition on $\mathbb{R}^m$. Hence, $\|\omega(y) - \omega(x)\|_2 \leq L\|y - x\|_2$. Let $\Gamma = \operatorname{diag}(\Gamma_1, \ldots, \Gamma_n)$ where $\Gamma_i = \operatorname{diag}((\gamma_i)_{j=1}^{m_i})$—that is, Γ_i is an $m_i \times m_i$ diagonal matrix with γ_i repeated on the diagonal m_i times. Then, $\|x - y\|_2 \leq L\|\Gamma\|_2\|y - x\|_2 < \|y - x\|_2$ since $\|\Gamma\|_2 = \max_i |\gamma_i| < 1/L$.

Now, observe that $Dg = I - \Gamma J(x)$. If Dg is invertible, then the implicit function theorem [97, Theorem C.40] implies that g is a local diffeomorphism. Hence, it suffices to show that $\Gamma J(x)$ does not have an eigenvalue of 1. Indeed, letting $\rho(A)$ be the spectral radius of a matrix A, we know in general that $\rho(A) \leq \|A\|$ for any square matrix A and induced operator norm $\|\cdot\|$ so that $\rho(\Gamma J(x)) \leq \|\Gamma J(x)\|_2 \leq \|\Gamma\|_2 \sup_{x \in \mathbb{R}^m} \|J(x)\|_2 < \max_i |\gamma_i| L < 1$ Of course, the spectral radius is the maximum absolute value of the eigenvalues, so that the above implies that all eigenvalues of $\Gamma J(x)$ have absolute value less than 1.

Since g is injective by the preceding argument, its inverse is well-defined and since g is a local diffeomorphism on $\mathbb{R}^m$, it follows that g^{-1} is smooth on $\mathbb{R}^m$. Thus, g is a diffeomorphism.

(b) Application of the stable manifold theorem Consider all critical points to the game—i.e. $\mathcal{X}_c = \{x \in X |\ \omega(x) = 0\}$. For each $p \in \mathcal{X}_c$, let B_p be the open ball derived from Theorem 5 and let $\mathcal{B} = \cup_p B_p$. Since $X \subseteq \mathbb{R}^m$, Lindelõf's lemma [84]—every open cover has a countable subcover—gives a countable subcover of $\mathcal{B}$. That is, for a countable set of critical points $\{p_i\}_{i=1}^{\infty}$ with $p_i \in \mathcal{X}_c$, we have that $\mathcal{B} = \cup_{i=1}^{\infty} B_{p_i}$.

Starting from some point $x_0 \in X$, if gradient-based learning converges to a strict saddle point, then there exists a t_0 and index i such that $g^t(x_0) \in B_{p_i}$ for all $t \geq t_0$. Again, applying Theorem 5 and using that $g(X) \subset X$—which we note is obviously true if $X = \mathbb{R}^m$—we get that $g^t(x_0) \in W^{cs}_{\text{loc}} \cap X$.

Using the fact that g is invertible, we can iteratively construct the sequence of sets defined by $W_1(p_i) = g^{-1}(W^{cs}_{\text{loc}} \cap X)$ and $W_{k+1}(p_i) = g^{-1}(W_k(p_i) \cap X)$. Then we have that $x_0 \in W_t(p_i)$ for all $t \geq t_0$. The set $\mathcal{X}_0 = \cup_{i=1}^{\infty} \cup_{t=0}^{\infty} W_t(p_i)$ contains all the initial points in X such that gradient-based learning converges to a strict saddle. Since p_i is a strict saddle, $I - \Gamma J(p_i)$ has an eigenvalue greater than 1. This implies that the co-dimension of E^u is strictly less than m. (i.e. $\dim(W^{cs}_{\text{loc}}) < m$). Hence, $W^{cs}_{\text{loc}} \cap X$ has Lebesgue measure zero in $\mathbb{R}^m$.

Again since. g is a diffeomorphism, $g^{-1} \in C^1$ and is further locally Lipschitz and null set preserving. Hence, $W_k(p_i)$ has measure zero for all k by induction so that $\mathcal{X}_0$ is a measure zero set since it is a countable union of measure zero sets. □

In potential games we can strengthen the above non-convergence result and give convergence guarantees.

Corollary 2. *Consider a potential game $(f_1, \ldots, f_N)$ on open, convex $X = X_1 \times \cdots \times X_N \subseteq \mathbb{R}^m$ and where each $f_i \in C^s(X, \mathbb{R})$ for $s \geq 3$. Let ν be a prior measure with support X which is absolutely continuous with respect to the Lebesgue measure and assume $\lim_{t\to\infty} g^t(x)$ exists. Then, under Assumption 1, competitive gradient-based learning converges to non-degenerate differential Nash equilibria almost surely. Moreover, the non-degenerate differential Nash to which it converges is generically a local Nash equilibrium.*

Corollary 2 guarantees that in potential games, gradient-play will converge to a differential Nash equilibrium. The proof of follows from the symmetry of J in potential games, and our observations in Chapter 2.

Proof of Corollary 2. Since the game admits a potential function ϕ, there is a transformation of coordinates such that agents following the dynamics $x_{t+1} = x_t - \gamma \odot \omega(x_t)$ converge to the same equilibria as the gradient dynamics $x_{t+1} = x_t - \gamma \odot D\phi(x_t)$. Hence, the analysis of the gradient-based learning scheme reduces to analyzing gradient-based optimization of ϕ. Moreover, existence of a potential function also implies that $D_{ij} f_j \equiv D_{ji} f_i$ so that J is symmetric. Indeed, writing $\omega(x)$ as the differential form $\sum_{i=1}^n D_i f_i(x) \mathrm{d}x_i$ and noting that $\mathrm{d} \circ \mathrm{d} = 0$ for the differential operator d, we have that $\mathrm{d}(\omega) = \sum_i \mathrm{d}(D_i f_i) \wedge \mathrm{d}x_i = \sum_{i,j:j>i} (D_{ij} f_j - D_{ji} f_i)\, \mathrm{d}x_i \wedge \mathrm{d}x_j = 0$ where $\wedge$ is the standard exterior product [97]. Symmetry of J implies that all periodic orbits are equilibria—i.e. the dynamics do not possess any limit cycles. By Theorem 4, the set of initial points that converge to strict saddle points is

of measure zero. Since all the stable critical points of the dynamics are equilibria, with the assumption that $\lim_{t\to\infty} g^t(x)$ exists for all $x \in X$, we have that $P_\nu\left[\lim_{t\to\infty} g^t(x) = x^*\right] = 1$ where x^* is a non-degenerate differential Nash equilibrium which is generically a local Nash equilibrium [157].

□

Combining this with Theorem 4 and the insights from Chapter 2 guarantees that the differential Nash equilibrium it converges to is a local minimizer of the potential function. A simple implication of this result is that gradient-based learning in potential games cannot exhibit limit cycles or chaos.

Of note is the fact that the agents *do not* need to be performing gradient-based learning on ϕ to converge to Nash almost surely. That is, they do not need to know the function ϕ; they simply need to follow the derivative of their own cost with respect to their own choice variable, and they are guaranteed to converge to a local Nash equilibrium that is a local minimizer of the potential function.

Remark 4. *We note that convergence to Nash equilibria is a known characteristic of gradient-play in potential games. However, our analysis also highlights that gradient-play will avoid a subset of the Nash equilibria of the game (namely local Nash equilibria that are saddle points of the potential function as shown in Proposition 4). This is surprising given the particularly strong structural properties of such games.*

Implications and Interpretation of Convergence Analysis

Both Theorem 4 and Corollary 2 show that gradient-play in multi-agent settings avoids strict saddles almost surely even in the deterministic setting. Combined with the analysis in Chapter 2 which shows that (local) Nash equilibria can be strict saddles of the dynamics for general-sum games, this implies that a subset of the Nash equilibria are almost surely avoided by individual gradient-play, a potentially undesirable outcome in view of **Q1** (whether all Nash equilibria are attracting for the learning dynamics). In Chapter 4, we show that the global Nash equilibrium is a saddle point of the gradient dynamics in a large number of randomly sampled LQ dynamic games. This suggests that policy gradient algorithms may fail to converge in such games, which is highly undesired. This is in stark contrast to the single agent setting where policy gradient has been shown to converge to the unique solution of LQR problems [51].

In Chapter 2, we also showed that local Nash equilibria of potential games can be strict saddles points of the potential function. Non-convergence to such points in potential games is not necessarily a bad result since this in turn implies convergence to a local minimizer of the potential function (as shown in [98, 144]) which are guaranteed to be local Nash equilibria of the game. However, these results do imply that *one cannot answer "yes" to **Q1** in potential games* since some of the Nash equilibria are not attracting under gradient-play.

In zero-sum games, where local Nash equilibria cannot be strict saddle points of the gradient dynamics, our result suggests that *eventually* gradient-based learning algorithms

will escape saddle points of the dynamics, though it is impossible to rule out converging to cycles or non-Nash equilibria without making additional structural assumptions.

The almost sure avoidance of all equilibria that are saddle points of the dynamics further implies that if (2.1) converges to a critical point x, then $x \in \texttt{LASE}(\omega)$—i.e., x is locally asymptotically stable for $\dot{x} = -\omega(x)$. This may not be a desired property however, since we showed in Chapter 2 that zero-sum and general-sum games both admit non-Nash LASE.

Since gradient-play in games generally does not result in a gradient flow, other types of limiting behaviors such as limit cycles can occur in gradient-based learning dynamics. Theorem 4 says nothing about convergence to other limiting behaviors. In the following sections we prove that the results described in this section extend to the stochastic gradient setting. We also formally define periodic orbits in the context of dynamical systems and state stronger results on avoidance of some more complex limiting behaviors like linearly unstable limit cycles.

3.3 Convergence and Non-Convergence of Stochastic Gradient-Play

We now analyze the stochastic case in which agents are assumed to have an unbiased estimator for their gradient. The results in this section allow us to extend the results from the deterministic setting to a setting where each agent builds an estimate of the gradient of their loss at the current set of strategies from potentially noisy observations of the environment. Thus, we are able to analyze the limiting behavior of a class of commonly used machine learning algorithms for competitive, multi-agent settings. In particular, we show that agents will almost surely not converge to strict saddle points. In Theorem 8, we show that the gradient dynamics will actually avoid more general limiting behaviors called linearly unstable cycles which we define formally.

To perform our analysis, we make use of tools and ideas from the literature on stochastic approximations (see e.g [26]). We note that the convergence of stochastic gradient schemes in the single-agent setting has been extensively studied [28, 122, 147, 161]. We extend this analysis to the behavior of stochastic gradient algorithms in games.

We assume that each agent updates their strategy using the update rule

$$x_{i,t+1} = x_{i,t} - \gamma_{i,t}(D_i f_i(x_{i,t}, x_{-i,t}) + w_{i,t+1}) \tag{3.4}$$

for some zero-mean, finite-variance stochastic process $\{w_{i,t}\}$.

In particular, we make the following standard assumptions on the noise processes [161, 163].

Assumption 2. *The stochastic process $\{w_{i,t+1}\}$ satisfies the assumptions $\mathbb{E}[w_{i,t+1}|\ \mathcal{F}_i^t] = 0$, $t \geq 0$ and $\mathbb{E}[\|w_{i,t+1}\|^2|\ \mathcal{F}_i^t] \leq \sigma^2 < \infty$ a.s., for $t \geq 0$, where $\mathcal{F}_{i,t}$ is an increasing family of σ_i-fields—i.e. filtration, or history generated by the sequence of random variables—given by $\mathcal{F}_{i,t} = \sigma_i(x_{i,k}, w_{i,k}, k \leq t),\ t \geq 0$.*

We also make new assumptions on the players' step-sizes. These are standard assumptions in the stochastic approximation literature and are needed to ensure that the noise processes are asymptotically controlled.

Assumption 3. *For each $i \in \mathcal{I}$, $f_i \in C^s(X, \mathbb{R})$ with $s \geq 2$, $D_i f_i$ is L_i–Lipschitz with $0 < L_i < \infty$, the step-sizes satisfy $\gamma_{i,t} \equiv \gamma_t$ for all $i \in \mathcal{I}$ and $\sum_t \gamma_t = \infty$ and $\sum_t (\gamma_t)^2 < \infty$, and $\sup_t \|x_t\| < \infty$ a.s.*

Once again, our results make use classical results from dynamical systems theory, namely Theorem 1 from [147] which guarantees that stochastic approximation schemes avoid unstable equilibria of the limiting ode.

Theorem 6 (Theorem 1 [147])**.** *Consider a general stochastic approximation framework of the form:*

$$x_{t+1} = x_t + \gamma_t(h(x_t)) + \epsilon_t$$

for $h : X \to TX$ with $h \in C^2$ where $X \subset \mathbb{R}^d$ and TX denotes the tangent space. Suppose γ_t is $\mathcal{F}_t$–measurable and $\mathbb{E}[\epsilon_t | \mathcal{F}_t] = 0$.

Let the stochastic process $\{x_t\}_{t \geq 0}$ be defined as above for some sequence of random variables $\{\epsilon_t\}$ and $\{\gamma_t\}$ and let $p \in X$ with $h(p) = 0$ with W a neighborhood of p. Assume that there are constants $\eta \in (1/2, 1]$ and $c_1, c_2, c_3, c_4 > 0$ for which the following conditions are satisfied whenever $x_t \in W$ and t sufficiently large:

a. p is a strict saddle point of $\dot{x} = -h(x)$,

b. $c_1/t^\eta \leq \gamma_t \leq c_2/t^\eta$,

c. $\mathbb{E}[(w_t \cdot v)^+ | \mathcal{F}_t] \geq c_3/t^\eta$ for every unit vector $v \in TX$,

d. $\|w_t\|_2 \leq c_4/t^\eta$.

Then $P(x_t \to p) = 0$.

Given this result, the following theorem extends the results of Theorem 4 to the stochastic gradient dynamics in games.

Theorem 7. *Consider a game $(f_1, \ldots, f_N)$ on $X = X_1 \times \cdots \times X_n = \mathbb{R}^m$. Suppose each agent $i \in \mathcal{I}$ adopts a stochastic gradient algorithm that satisfies Assumptions 2 and 3. Further, suppose that for each $i \in \mathcal{I}$, there exists a constant $b_i > 0$ such that $\mathbb{E}[(w_{i,t} \cdot v)^+ | \mathcal{F}_{i,t}] \geq b_i$ for every unit vector $v \in \mathbb{R}^{m_i}$. Then, competitive stochastic gradient-based learning converges to strict saddle points of the game on a set of measure zero.*

The proof follows directly from showing that (3.4) satisfies Theorem 6. The assumption that $\mathbb{E}[(w_{i,t} \cdot v)^+ | \mathcal{F}_{i,t}] \geq b_i$ rules out degenerate cases where the noise forces the stochastic dynamics onto the stable manifold of strict saddle points.

Theorem 7 implies that the dynamics of stochastic gradient-based learning defined in (3.4), have the same limiting properties as the deterministic dynamics vis-à-vis saddle points. Thus, the implications described in Section 3.2 extend to the stochastic gradient setting. In particular, stochastic gradient-based algorithms will avoid a non-negligible subset of the Nash equilibria in general-sum and potential games. Further, in zero-sum and general-sum games, if the players fo converge to a critical point, that point may be a non-Nash equilibrium.

Further Convergence Results for Stochastic Gradient-Play in Games

As we demonstrated in Section 3.2, outside of potential games, the dynamics of gradient-based learning algorithms in games are not gradient flows. As such, the players' actions can converge to more complex sets than simple equilibria. A particularly prominent class of limiting behaviors for dynamical systems are known as limit cycles (see e.g [168]). Limit cycles (or periodic orbits) are sets of states $\mathcal{S}$ such that each state $x \in \mathcal{S}$ is visited at periodic intervals *ad infinitum* under the dynamics. Thus, if the gradient-based algorithms converge to a limit cycle they will cycle infinitely through the same sequence of actions. Like equilibria, limit cycles can be stable or unstable under the dynamics $\dot{x} = -\omega(x)$, meaning that the dynamics can either converge to or diverge from them depending on their initializations.

We remark that the existence of oscillatory behaviors and limit cycles has been observed in the dynamics of of gradient-based learning in various settings like the training of Generative Adversarial Networks [45], and multiplicative weights in finite action games [121]. We simply emphasize that the existence of such limiting behaviors is due to the fact that the dynamics are no longer gradient flows. This fact also allows for other complex limiting behaviors like chaos[2] to exist in the dynamics of gradient-based learning in games.

In the following subsections we formalize the notion of a limit cycle and its stability in the stochastic setting. Using these concepts, we then provide an analogous theorem to Theorem 7 which states that competitive stochastic gradient-based learning converges to linearly unstable limit cycles—a parallel notion to strict saddle points but pertaining to more general limit sets—on a set of measure zero, provided that analogous assumptions to those in the statement of Theorem 7 hold. Providing such guarantees requires a bit more mathematical formalism which we develop in the next subsection.

Avoidance of Repelling Sets

To show that stochastic gradient-based learning avoid more general limiting behaviors than saddle points, we need further assumptions on our underlying space—i.e. we need the underlying decision spaces of each agent—i.e. X_i for each $i \in \mathcal{I}$—to be *smooth, compact manifolds*

[2] A general term used to characterize dynamical systems where arbitrarily small perturbations in the initial conditions lead to drastically different solutions to the differential equations

without boundary[3]. The stochastic process $\{x_n\}$ which follows (3.4) is *defined on* X—that is, $x_n \in X$ for all $n \geq 0$. As before, it is natural to compare sample points $\{x_n\}$ to solutions of $\dot{x} = -\omega(x)$ where we think of (3.4) as a noisy approximation. The asymptotic behavior of $\{x_n\}$ can indeed be described by the asymptotic behavior of the flow generated by ω.

We also need a formal notion of *cycles*. A non-stationary periodic orbit of ω is called a *cycle*. Let $\xi \subset X$ be a cycle of period $T > 0$. Denote by Φ_T the flow corresponding to ω. For any $x \in \xi$, $\text{spec}(D\Phi_T(x)) = \{1\} \cup C(\xi)$ where $C(\xi)$ is the set of characteristic multipliers. We say ξ is *hyperbolic* if no element of $C(\xi)$ is on the complex unit circle. Further, if $C(\xi)$ is strictly inside the unit circle, ξ is called *linearly stable* and, on the other hand, if $C(\xi)$ has at least one element on the outside of the unit circle—that is, $D\Phi_T(x)$ for $x \in \xi$ has an eigenvalue with real part strictly greater than 1—then ξ is called *linearly unstable*. The latter is the analog of strict saddle points in the context of periodic orbits. We denote by $\{x_t\}$ sample paths of the process (3.4) and $L(\{x_t\})$ is the *limit set* of any sequence $\{x_t\}_{t\geq 0}$ which is defined in the usual way as all $p \in X$ such that $\lim_{k\to\infty} x_{t_k} = p$ for some sequence $t_k \to \infty$. It was shown in [22] that under less restrictive assumptions than Assumptions 2 and 3, $L(\{x_t\})$ is contained in the *chain recurrent set* of ω and $L(\{x_t\})$ is a non-empty, compact and connected set invariant under the flow of ω.

Theorem 8. *Consider a game $(f_1, \ldots, f_n)$ where each X_i is a smooth, compact manifold without boundary. Suppose each agent $i \in \mathcal{I}$ adopts a stochastic gradient-based learning algorithm that satisfies Assumptions 2 and 3 and is such that sample points $x_t \in X$ for all $t \geq 0$. Further, suppose that for each $i \in \mathcal{I}$, there exist a constant $b_i > 0$ such that $\mathbb{E}[(w_{i,t} \cdot v)^+ | \mathcal{F}_{i,t}] \geq b_i$ for every unit vector $v \in \mathbb{R}^{m_i}$. Then competitive stochastic gradient-based learning converges to linearly unstable cycles on a set of measure zero—i.e. $P(L(x_t) = \xi) = 0$ where $\{x_t\}$ is a sample path.*

As we noted, periodic orbits are not necessarily excluded from the limiting behavior of gradient-based learning in games. We leave out the proof of Theorem 8 since after some algebraic manipulation, it is a direct application of Theorem 2.1 in [24] which is re-stated below.

Theorem 9 (Theorem 2.1 [24])**.** *Consider a general stochastic approximation framework of the form:*

$$x_{t+1} = x_t + \gamma_t(h(x_t)) + \epsilon_t$$

for $h : X \to TX$ with $h \in C^2$ where $X \subset \mathbb{R}^d$ and TX denotes the tangent space. Suppose γ_t is $\mathcal{F}_t$–measurable and $\mathbb{E}[\epsilon_t | \mathcal{F}_t] = 0$.

Let $\xi \subset X$ be a hyperbolic linearly unstable cycle of h. Assume the following:

a $h \in C^2$;

b $c_1/t^\eta \leq \gamma_t \leq c_2/t^\eta$ with $0 < c_1 \leq c_2$ and $0 < \eta \leq 1$;

[3]The torus $\mathbb{T} = \mathbb{S}^1 \times \mathbb{S}^1$ is an example. The interested reader can consult, e.g., [97] for more details on differential geometry.

c there exists $b \geq 0$ *such that for all unit vectors* $v \in \mathbb{R}^m$, $\mathbb{E}[(w_t \cdot v)^+ | \mathcal{F}_t] \geq b$.

Then $P(L(\{x_t\}) = \xi) = 0$.

Morse-Smale Games

In pursuit of a more general class of games with desirable convergence properties, we now introduce a generalization of potential games, namely Morse-Smale games, for which the combined gradient dynamics correspond to a Morse-Smale vector field [74, 143]. In such games players are guaranteed to converge to only (linearly stable) cycles or equilibria. In such games, however, players may still converge to non-Nash equilibria and avoid a subset of the Nash equilibria.

For a class of games admitting *gradient-like* vector fields we can go beyond non-convergence results and give convergence guarantees. Following [24], we introduce a new class of games, which we call *Morse-Smale games*, that are a generalization of potential games. Such games represent an important class since the vector field of ω corresponds to Morse-Smale vector field which is known to be generic in $\mathbb{R}^2$ and are otherwise structurally stable [74, 143].

Definition 13. *A game* $(f_1, \ldots, f_n)$ *with* $f_i \in C^r$ *for some* $r \geq 3$ *and where strategy spaces* X_i *is a smooth, compact manifold without boundary for each* $i \in \mathcal{I}$ *is a Morse-Smale game if the vector field corresponding to the differential* ω *is Morse-Smale—that is, the following hold: (i) all periodic orbits* ξ *(i.e. equilibria and cycles) are hyperbolic and the stable and unstable manifolds of any two periodic orbits* ξ *and* ξ' *intersect transversally:* $W^s(\xi) \pitchfork W^u(\xi')$*, (ii) every forward and backward omega limit set is a periodic orbit, (iii) and* ω *has a global attractor.*

The conditions of Morse-Smale in the above definition ensure that there are only finitely many periodic orbits. The dynamics of games with more general vector fields, on the other hand, can admit chaos (e.g. the classic Lorentz attractor can be cast as gradient-play in a 3-player game). Hyperbolic equilibria and periodic orbits are the only types of limiting behavior that have been shown to correspond to strategies relevant to the underlying game [21]. The simplest example of a Morse-Smale vector field is a gradient flow. However, not all Morse-Smale vector fields are gradient flows and hence, not all Morse-Smale games are potential games.

Example 3. *Consider the n-player game with* $X_i = \mathbb{R}$ *for each* $i \in \mathcal{I}$ *and* $f_n(x) = x_n(x_1^2 - 1)$, $f_i(x) = x_i x_{i+1}$, $\forall i \in \mathcal{I}/\{n\}$ *This is a Morse-Smale game that is not a potential game. Indeed,* $\dot{x} = -\omega(x)$ *where* $\omega = [x_2, x_3, \ldots, x_{n-1}, x_1^2 - 1]$ *is a dynamical system with a Morse-Smale vector field that is not a gradient vector field [40].*

Essentially, in a neighborhood of a critical point for a Morse-Smale game, the game behavior can be described by a Morse function ϕ such that near critical points ω can be written as $D\phi$ and away from critical points ω points in the same direction as $D\phi$—i.e. $\omega \cdot D\phi > 0$. Specializing the class of Morse-Smale games, we have stronger convergence guarantees.

Theorem 10. *Consider a Morse-Smale game $(f_1, \ldots, f_n)$ on smooth boundaryless compact manifold X. Suppose Assumptions 2 and 3 hold and that $\{x_t\}$ is defined on X. Let $\{\xi_i,\ i = 1, \ldots, l\}$ denote the set of periodic orbits in X. Then $\sum_{i=1}^{l} P(L(\{x_t\}) = \xi_i) = 1$ and $P(L(\{x_t\}) = \xi_i) > 0$ implies ξ_i is linearly stable. Moreover, if the periodic orbit ξ_i with $P(L(\{x_t\}) = \xi_i) > 0$ is an equilibrium, then it is either a non-degenerate differential Nash equilibrium—which is generically a local Nash—or a non-Nash locally asymptotically stable equilibrium.*

The proof of Theorem 10 follows by invoking Corollary 3 which is stated below.

Corollary 3 (Corollary 2.2 [24])**.** *Assume that there exists $\delta \geq 1$ such that $\sum_{n\geq 0} \gamma_n^{1+\delta} < \infty$ and that h is a Morse-Smale vector field. If we denote by $\{\xi_i,\ i = 1, \ldots, l\}$ the set of periodic orbits in X, then $\sum_{i=1}^{l} P(L(\{x_t\}) = \xi_i) = 1$. Further, if conditions (i)–(iii) of Theorem 9 hold, then $P(L(\{x_t\}) = \xi_i) > 0$ implies ξ_i is linearly stable.*

Thus, in Morse-Smale games, with probability one, the limit sets of competitive gradient-based learning with stochastic updates are attractors (i.e., periodic orbits, which includes limit cycles and equilibria) of $\dot{x} = -\omega(x)$ and if any attractor has positive probability of being a limit set of the players' collective update rule, then it is (linearly) stable. Moreover, attractors that are equilibria are either non-degenerate differential Nash equilibria (generically local Nash equilibria) or non-Nash locally asymptotically stable equilibria, but not saddle points.

If we further restrict the class of games to potential games, the results for Morse-Smale games imply convergence to Nash almost surely, a particularly strong convergence guarantee.

Corollary 4. *Consider the game $(f_1, \ldots, f_n)$ on smooth boundaryless compact manifold $X = X_1 \times \cdots \times X_n$ admitting potential function ϕ. Suppose each agent $i \in \mathcal{I}$ adopts a stochastic gradient-based learning algorithm that satisfies Assumptions 2 and 3 and such that $\{x_t\}$ evolves on X. Further, suppose that for each $i \in \mathcal{I}$, there exist a constant $b_i > 0$ such that $\mathbb{E}[(w_{i,t} \cdot v)^+ | \mathcal{F}_{i,t}] \geq b_i$ for every unit vector $v \in \mathbb{R}^{m_i}$. Then, competitive stochastic gradient-based learning converges to a non-degenerate differential Nash equilibrium almost surely.*

The proof of Corollary 4 follows from the fact that potential games are trivially Morse-Smale games that admit no periodic cycles as we showed in the proof of Corollary 2.

Proof of Corollary 4. Consider a potential game $(f_1, \ldots, f_n)$ where each X_i is a smooth, compact boundaryless manifold. Then $\omega = D\phi$ for some $\phi \in C^r$ which implies that ω is a gradient flow and hence, does not admit limit cycles. Let $\{\xi_i,\ i = 1, \ldots, l\}$ be the set of equilibrium points in X. Under the assumptions of Theorem 10, $\sum_{i=1}^{l} P(L(\{x_t\}) = \xi_i) = 1$ and, if $P(L(\{x_t\}) = \xi_i) > 0$, then ξ_i is a linearly stable equilibrium point which is a non-degenerate differential Nash equilibrium of the game due to the fact that $D\omega(x)$ is symmetric in potential games. Hence, a sample path $\{x_t\}$ converges to a non-degenerate differential

Nash equilibrium with probability one. Moreover, by [157], we know it is generically a local Nash. □

We note, that even though a potential function is enough to guarantee convergence to a local Nash equilibrium, potential games can still admit local Nash equilibria that are strict saddle points as shown in Section 2. Thus, even this relatively well-behaved class of games has problems when applying a gradient-based learning scheme.

3.4 Chapter Summary

Our results suggest that gradient-play in multi-agent settings has fundamental problems. Depending on the players' costs, in general games and even potential games, which have a particularly *nice* structure, a subset of the Nash equilibria will be almost surely avoided by gradient-based learning when the agents randomly initialize their first action. In zero-sum and general-sum games, even if the algorithms do converge, they may have converged to a point that has no game theoretic relevance, namely a non-Nash locally asymptotically stable equilibrium.

Lastly, these results show that limit cycles persist even under a stochastic update scheme. This explains the empirical observations of limit cycles in gradient dynamics presented in [45, 76, 99]. It also implies that gradient-based learning in multi-agent reinforcement learning, multi-armed bandits, generative adversarial networks, and online optimization all admit limit cycles under certain loss functions. Our empirical results show that these problems are not merely of theoretical interest, but also have great relevance in practice.

Which classes of games have all Nash being attracting for gradient-play and which classes preclude the existence of non-Nash equilibria is an open and particularly interesting question. Further, the question of whether gradient-based algorithms can be constructed for which only game-theoretically relevant equilibria are attracting is of particular importance as gradient-based learning is increasingly implemented in game theoretic settings. Indeed, more generally, as learning algorithms are increasingly deployed in markets and other competitive environments understanding and dealing with such theoretical issues will become increasingly important.

Chapter 4

Gradient-Based Learning in Multi-Agent Reinforcement Learning

Interest in multi-agent reinforcement learning has seen a recent surge of late, and policy-gradient algorithms are championed due to their potential scalability. Indeed, recent impressive successes of multi-agent reinforcement learning have made use of policy optimization algorithms such as multi-agent actor-critic [77, 103, 178], multi-agent proximal policy optimization [17], and even simple multi-agent policy-gradients [92] in problems where the various agents have high-dimensional continuous state and action spaces like StarCraft [190].

Despite these successes, a theoretical understanding of these algorithms in multi-agent settings is still lacking. Missing perhaps, is a tractable yet sufficiently complex setting in which to study these algorithms. Recently, there has been much interest in analyzing the convergence and sample complexity of policy-gradient algorithms in the classic linear quadratic regulator (LQR) problem from optimal control [82]. The LQR problem is a particularly apt setting to study the properties of reinforcement learning algorithms due to the existence of an optimal policy which is a linear function of the state and which can be found by solving a Ricatti equation. Indeed, the relative simplicity of the problem has allowed for new insights into the behavior of reinforcement learning algorithms in continuous action and state spaces [46, 51, 110].

An extension of the LQR problem to the setting with multiple agents, known as a *linear quadratic (LQ) game*, has also been well studied in the literature on dynamic games and optimal control [19]. As the name suggests, an LQ game is a game in which multiple agents attempt to control a shared linear dynamical system subject to quadratic costs. Since the players have their own costs, the notion of 'optimality' in such games is a Nash equilibrium, properties of which have been well analyzed in the literature [18, 50, 105, 152].

Like LQR for the classical single-agent setting, LQ games are an appealing setting in which to analyze the behavior of multi-agent reinforcement learning algorithms in continuous action and state spaces since they admit global Nash equilibria in the space of linear feedback policies. Moreover, these equilibria can be found by solving a coupled set of Ricatti equations. As such, LQ games are a natural benchmark problem on which to test policy-

gradient algorithms in multi-agent settings. Furthermore, policy gradient methods open up the possibility to new scalable approaches to finding solutions to control problems even with constraints. In the single-agent setting, it was recently shown that policy-gradient has global convergence guarantees for the LQR problem [51]. These results have recently been extended to projected policy-gradient algorithms in zero-sum LQ games [204].

Chapter Overview

In this chapter we present a *negative* result, showing that policy-gradient in general-sum LQ games does not enjoy *even local* convergence guarantees, unlike in LQR. In particular, we show that, if each player randomly initializes their policy and then uses a policy-gradient algorithm, there exists an LQ game in which the players would almost surely avoid a Nash equilibrium. Further, our numerical experiments indicate that LQ games in which this occurs may be quite common. We also observe empirically that when players fail to converge to the Nash equilibrium they do converge to stable limit cycles. These cycles do not seem to have any readily apparent relationship to the Nash equilibria of the game.

We note that non-convergence to Nash equilibria is not in itself a new phenomenon (see e.g. [35, 37, 45, 113, 116]) and that the existence of cycles in the dynamics of learning dynamics in games has also been repeatedly observed in various contexts [114, 116, 121, 145]. However, such phenomena have not yet been shown to occur in the dynamics of multi-agent reinforcement learning algorithms in continuous action and state spaces. Since such algorithms have had such striking successes in recent years, we believe a theoretical understanding of their behaviors can lay the groundwork for the development of more efficient and theoretically sound multi-agent learning algorithms.

Organization.

Section 4.1 introduces n-player general-sum LQ games and presents previous results on the existence of the Nash equilibrium in such games. In Section 4.2, we show that these games are *not* convex games and that all the stationary points of the joint policy-gradient dynamics are Nash equilibria. Following this, we give sufficient conditions under which policy-gradient almost surely avoids a Nash equilibrium in Section 4.3. Given these theoretical results, in Section 4.4 we present empirical results demonstrating that a large number of 2-player LQ games satisfy these sufficient conditions. Numerical experiments showing the existence of limit cycles in the gradient dynamics of general-sum LQ games are also presented. The paper is concluded with a discussion in Section 4.5.

4.1 Linear Quadratic Games

We consider n-player LQ games subject to a discrete-time dynamical system defined by

$$z(t+1) = Az(t) + \sum_{i=1}^{n} B_i u_i(t) \quad z(0) = z_0 \sim \mathcal{D}_0, \tag{4.1}$$

where $z(t) \in \mathbb{R}^m$ is the state at time t, $\mathcal{D}_0$ is the initial state distribution, and $u_i(t) \in \mathbb{R}^{d_i}$ is the control input of player $i \in 1, \ldots, n$. For LQ games, it is known that under reasonable assumptions, linear feedback policies for each player that constitute a Nash equilibrium exist and are unique if a set of coupled Ricatti equations admit a unique solution [19]. Thus, we consider that each player i searches for a linear feedback policy of the form $u_i(t) = -K_i z(t)$ that minimizes their loss, where $K_i \in \mathbb{R}^{d_i \times m}$. We use the notation $d = \sum_{i=1}^{n} d_i$ for the combined dimension of the players' parameterized policies.

As the name of the game implies, the players' loss functions are quadratic functions given by

$$f_i(u_1, \ldots, u_n) = \mathbb{E}_{z_0 \sim \mathcal{D}_0} \left[\sum_{t=0}^{\infty} z(t)^T Q_i z(t) + u_i(t)^T R_i u_i(t) \right],$$

where Q_i and R_i are the cost matrices for the state and input, respectively.

Assumption 4. *For each player $i \in \{1, \ldots, n\}$, the state and control cost matrices satisfy $Q_i \succ 0$ and $R_i \succ 0$.*

We note that the players are coupled through the dynamics since $z(t)$ is constrained to obey the update equation given in (4.1). We focus on a setting in which all players randomly initialize their strategy and then perform gradient descent simultaneously on their own cost functions with respect to their individual control inputs. That is, the players use policy-gradient algorithms of the following form:

$$K_{i,n+1} = K_{i,n} - \gamma_i D_i f_i(K_{1,n}, \ldots, K_{n,n}) \tag{4.2}$$

where $D_i f_i(\cdot, \cdot)$ denotes the derivatives of f_i with respect to the i–th argument, and $\{\gamma_i\}_{i=1}^n$ are the step-sizes of the players. We note that there is a slight abuse of notation here in the expression of $D_i f_i$ as functions of the parameters K_i as opposed to the control inputs u_i. To ensure there is no confusion between t and n, we also point out that n indexes the policy-gradient algorithm iterations while t indexes the time of the dynamical system.

To simplify notation, define

$$\Sigma_K = \mathbb{E}_{z_0 \sim \mathcal{D}_0} \left[\sum_{t=0}^{\infty} z(t) z(t)^T \right],$$

where we use the subscript notation to denote the dependence on the collection of controllers $K = (K_1, \ldots, K_n)$. Define also the initial state covariance matrix

$$\Sigma_0 = \mathbb{E}_{z_0 \sim D_0}[z_0 z_0^T].$$

Direct computation verifies that for player i, $D_i f_i$ is given by:

$$D_i f_i(K_1, \ldots, K_n) = 2(R_i K_i - B_i^T P_i \bar{A}) \Sigma_K, \tag{4.3}$$

where $\bar{A} = A - \sum_{i=1}^{n} B_i K_i$, is the closed–loop dynamics given all players' control inputs and, for given $(K_1, \ldots, K_n)$, the matrix P_i is the unique positive definite solution to the Bellman equation:

$$P_i = \bar{A}^T P_i \bar{A} + K_i^T R_i K_i + Q_i, \quad i \in \{1, \ldots, n\}. \tag{4.4}$$

Given that the players may have different control objectives and do not engage in coordination or cooperation, the best they can hope to achieve is a Nash equilibrium.

Definition 14. *A* feedback Nash equilibrium *is a collection of policies* $(K_1^*, \ldots, K_n^*)$ *such that:*

$$f_i(K_1^*, \ldots, K_i^*, \ldots, K_n^*) \leq f_i(K_1^*, \ldots, K_i, \ldots, K_n^*), \quad \forall\ K_i \in \mathbb{R}^{d_i \times m}.$$

for each $i \in \{1, \ldots, n\}$.

Under suitable assumptions on the cost matrices, the Nash equilibrium of an LQ game is known to exist in the space of linear policies [19, 102]. However, this Nash equilibrium may not be unique. To the best of our knowledge, there are no general set of conditions under which the Nash equilibrium is unique in general-sum LQ games outside of the scalar dynamics setting [50]. There are, however, algebraic geometry methods to compute all Nash equilibria in LQ games [152]. We make use of a simpler algorithm to find Nash equilibria which solves coupled Ricatti equations using the method of Lyapunov iterations. The method is outlined in [102] for continuous time LQ games, and an analogous procedure can be followed for discrete time. Convergence of this method requires the following assumption.

Assumption 5. *For at least one player* $i \in \{1, \ldots, n\}$*, the system* (A, B_i) *is stabilizable.*

Assumption 5 is a necessary condition for the players to be able to stabilize the system. Indeed, the player's costs are finite only if the closed loop system $\bar{A}$ is asymptotically stable, meaning that $|\mathbb{R}(\lambda)| < 1$ for all $\lambda \in \text{spec}(\bar{A})$, where $\mathbb{R}(\lambda)$ denotes the real part of λ and $\text{spec}(M)$ is the spectrum of a matrix M.

4.2 Analyzing the Optimization Landscape of LQ Games

Having introduced the class of games we consider we now analyze the optimization landscape in general-sum LQ games. Letting $x = (K_1, \ldots, K_n)$, the object of interest is the map $\omega : \mathbb{R}^{md} \to \mathbb{R}^{md}$ defined as follows:

$$\omega(x) = \begin{bmatrix} D_1 f_1(K_1, \ldots, K_n) \\ \vdots \\ D_n f_n(K_1, \ldots, K_n) \end{bmatrix}.$$

Note that $D_i f_i = \partial f_i / \partial K_i$ has been converted to an md_i dimensional vector and each K_i has also been vectorized. This is a slight abuse of notation and throughout we treat the K_i's as

both vectors and matrices; in general, the shape should be clear from context, and otherwise we make comments where necessary to clarify.

Before analyzing the stationary points of policy-gradient in LQ games, we show that the class of LQ games we consider are *not* convex games. This holds despite the linearity of the dynamics and the positive definiteness of the cost matrices. This fact makes the analysis of such games non-trivial since the lack of strong structural guarantees on the players' costs allows for non-trivial limiting behaviors like cycles, non-Nash equilibria, and chaos in the joint gradient dynamics. [116].

Proposition 9. *There exists a n-player LQ game satisfying assumptions 4 and 5 that is not a convex game.*

Proof. The proof of Proposition 9 follows directly from the non-convexity of the set of stabilizing policies for the single-agent LQR problem which was shown in [51]. Holding every other players' actions fixed, a player i is faced with a simple LQR problem. Since this problem is non-convex, LQ games are not convex games. □

In the absence of strong structural guarantees on the players' costs, simultaneous gradient-play in general-sum games can converge to strategies that are not Nash equilibria [116]. The following theorem shows that, despite the fact that LQ games are not convex for each player, such non-Nash equilibria cannot exist in the gradient dynamics of general-sum LQ games. Indeed, we show that a point x is a critical point of the policy gradient dynamics in a n-player LQ game if and only if it is a Nash equilibrium. We note that critical points of gradient-play are strategies $x = (K_1, \dots, K_n)$ such that $\omega(x) = 0$. Such points are of particular importance since a necessary condition for a point x to be a Nash equilibrium is that it is a critical point.

Theorem 11. *Consider the set $x^* = (K_1^*, \dots, K_n^*)$ of stabilizing policies such that $\Sigma_{K^*} > 0$. $D_i f_i(K_1^*, \dots, K_n^*) = 0$ for each $i \in \{1, \dots, n\}$, if and only if x^* is a Nash equilibrium.*

Proof. We first prove the forward direction and show that if $D_i f_i(x^*) = 0$ for each $i \in \{1, \dots, n\}$, then x^* is a Nash equilibrium. We show this by contradiction. Suppose the claim does not hold so that $\Sigma_{K^*} > 0$ and $D_i f_i(K_1^*, \dots, K_n^*) = 0$ for each $i \in \{1, \dots, n\}$, yet $(K_1^*, \dots, K_n^*)$ is not a Nash equilibrium. That is, without loss of generality, there exists a $\bar{K}_1$ such that

$$f_1(\bar{K}_1, K_2^*, \dots, K_n^*) < f_1(K_1^*, \dots, K_n^*).$$

Now, fixing $(K_2^*, \dots, K_n^*)$, player 1 can be seen as facing an LQR problem. Indeed, letting $(K_2^*, \dots, K_n^*)$ be fixed, player 1 aims to find a 'best response' in the space of linear feedback policies of the form $u_1(t) = Kz(t)$ with $K \in \mathbb{R}^{d_i \times m}$ that minimizes $f_1(\cdot, K_2^*, \dots, K_n^*)$ subject to the dynamics defined by

$$z(t+1) = \left(A - \textstyle\sum_{i=2}^n B_i K_i\right) z(t) + B_1 u_1(t).$$

Note that this system is necessarily stabilizable since $\bar{A}$ is stable. Hence, the discrete algebraic Riccati equation for player 1's LQR problem has a positive definite solution P such that

$R_1 + B_1^T P B_1 > 0$ since $R_1 > 0$ by assumption. Since $\Sigma_{K^*} > 0$ and $D_1 f_1(x^*) = 0$, applying Corollary 4 of [51], we have that K_1^* must be optimal for player 1's LQR problem so that

$$f_1(K_1^*, \ldots, K_n^*) \leq f_1(K, K_2^*, \ldots, K_n^*), \quad \forall\ K \in \mathbb{R}^{d_1 \times m}.$$

In particular, the above inequality holds for $\bar{K}_1$, which leads to a contradiction.

To prove the reverse direction, we note that a necessary condition for a point x to be a Nash equilibrium for each player, is that $D_i f_i(x^*) = 0$ for each $i \in \{1, \ldots, n\}$ [156]. □

Theorem 11 shows that, just as in the single-player LQR setting and zero-sum LQ games, the critical points of gradient-play in n–player general-sum LQ games are all Nash equilibria. We note that the condition $\Sigma_K > 0$ can be satisfied by choosing an initial state distribution $\mathcal{D}_0$ with a full-rank covariance matrix.

A simple consequence of Theorem 11 is that when the coupled Ricatti equations characterizing the Nash equilibria of the game have a unique positive definite solution and Assumptions 4 and 5 hold, the gradient dynamics admit a unique critical point.

Corollary 5. *Under Assumption 4 and 5, if the coupled Ricatti equations admit a unique solution and $\Sigma_0 \succ 0$, then the map ω has a unique critical point.*

Given that the critical points of the gradient dynamics in LQ games are Nash equilibria, the aim is to show, via constructing counter-examples, that games in which the gradient dynamics avoid the Nash equilibria do in fact exist. A sufficient condition for this would be to find a game in which gradient-play diverges from neighborhoods of Nash equilibria.

It is demonstrated in [116] that there may be Nash equilibria that are not even *locally attracting* under the gradient dynamics in n–player general-sum games in which the players' costs are sufficiently smooth (i.e., at least twice continuously differentiable). In games that admit such Nash equilibria, the agents could initialize arbitrarily close to the Nash equilibrium, simultaneously perform individual gradient descent with arbitrarily small step sizes, and still diverge.

The class of n–player LQ games we consider does not, however, satisfy the smoothness assumptions necessary to simply invoke the results in [116]. Indeed, the cost functions are non-smooth and, in fact, are infinite whenever the players have strategies that do not stabilize the dynamics. Further, the set of stabilizing policies for a dynamical system is not even convex [51]. Despite these challenges, in the sequel we show that the negative convergence results in [116] extend to the general-sum LQ setting. In particular, we show that even with arbitrarily small step sizes, players using policy-gradient in LQ games may still diverge from neighborhoods of a Nash equilibrium.

4.3 Sufficient Conditions for Policy-Gradient to Avoid Nash

We now give sufficient conditions under which gradient-play has no guarantees of even *local*, much less global, convergence to a Nash equilibrium. Towards this end, we first show that ω is sufficiently smooth on the set of stabilizing policies.

Let $\mathcal{S}^{md} \subset \mathbb{R}^{md}$ be the subset of stabilizing md–dimensional matrices.

Proposition 10. *Consider an n–player LQ game. The vector-valued map ω associated with the game is twice continuously differentiable on $\mathcal{S}^{md}$—i.e., $\omega \in C^2(\mathcal{S}^{md}, \mathcal{S}^{md})$.*

Using our notation, Lemma 6.5 in [204] shows for two-player zero-sum LQ games that (P_1, P_2), and Σ_K are continuously differentiable with respect to K_1 and K_2 when $A - B_1K_1 - B_2K_2$ is stable. This, in turn, implies that $\omega(K_1, K_2)$ is continuously differentiable with respect to K_1 and K_2 when the closed loop system $A - B_1K_1 - B_2K_2$ is stable. The result follows by a straightforward application of the implicit function theorem [5]. We make use of the same proof technique here in extending the result to n–player general-sum LQ games and, in fact, the proof implies that ω has even stronger regularity properties.

Proof. Following the proof technique of [204], we show the regularity of ω using the implicit function theorem [5]. In particular, we show that $\Sigma_K = \mathbb{E}_{z_0 \sim \mathcal{D}_0}\left[\sum_{t=0}^{\infty} z(t)z(t)^T\right]$ and P_i for $i \in \{1, \ldots, n\}$ are C^1 with respect to each K_i on the space of stabilizing matrices.

For any stabilizing $(K_1, \ldots, K_n)$, Σ_K is the unique solution to the following discrete-time Lyapunov equation:

$$\bar{A}\Sigma_K\bar{A}^T + \Sigma_0 = \Sigma_K, \tag{4.5}$$

where $\Sigma_0 = \mathbb{E}_{z_0 \sim \mathcal{D}_0}[z(0)z(0)^T] > 0$ and $\bar{A} = A - \sum_{i=1}^{n} B_iK_i$. Both sides of this expression can be vectorized. Indeed, using the same notation as in [204], let $\mathrm{vect}(\cdot)$ be the map that vectorizes its argument and let $\Psi : \mathbb{R}^{m^2} \times \mathbb{R}^{d_1 \times m} \times \cdots \times \mathbb{R}^{d_n \times m} \to \mathbb{R}^{m^2}$ be defined by

$$\Psi(\mathrm{vect}(\Sigma_K), K_1, \ldots, K_n) = \left[\bar{A} \otimes \bar{A}\right] \cdot \mathrm{vect}(\Sigma_K) + \mathrm{vect}(\Sigma_0).$$

Then, (4.5) can be written as

$$\begin{aligned} F(\mathrm{vect}(\Sigma_K), K_1, \ldots, K_n) &= \Psi(\mathrm{vect}(\Sigma_K), K_1, \ldots, K_n) - \mathrm{vect}(\Sigma_K) \\ &= 0. \end{aligned}$$

The map F implicitly defines Σ_K. Moreover, letting I denote the appropriately sized identity matrix, we have that

$$\frac{\partial F(\mathrm{vect}(\Sigma_K), K_1, \ldots, K_n)}{\partial \mathrm{vect}^T(\Sigma_K)} = \left[\bar{A} \otimes \bar{A}\right] - I.$$

For stabilizing $(K_1, \ldots, K_n)$, this matrix is an isomorphism since $\mathrm{spec}(\bar{A})$ is inside the unit circle. Thus, using the implicit function theorem, we conclude that $\mathrm{vect}(\Sigma_K) \in C^1$. As noted in [204], the proof for each P_i, $i \in \{1, \ldots, n\}$ is completely analogous. Since Σ_K and P_i are C^1 and ω is linear in these terms, the result of the proposition follows. □

Given that ω is continuously differentiable over the set of stabilizing joint policies, the following result gives sufficient conditions such that the set of initial conditions in a neighborhood of the Nash equilibrium from which gradient-play converges to the Nash equilibrium is of measure zero. This implies that the players will almost surely avoid the Nash equilibrium even if they randomly initialize in a uniformly small ball around it.

Let the Jacobian of the vector field ω be denoted by $D\omega$. Given a critical point x^*, let λ_j be the eigenvalues of $D\omega(x^*)$, for $j \in \{1, \ldots, md\}$, where $d = \sum_{i=1}^{n} d_i$. Recall that the state $z(t)$ is dimension m.

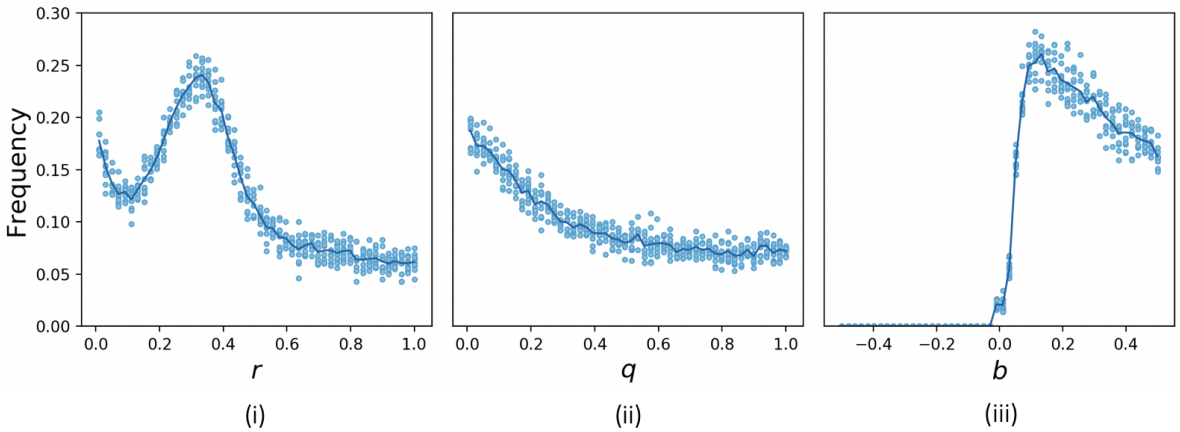


Figure 4.1: Frequency (out of 1000) of randomly sampled LQ games with global Nash equilibria that are avoided by policy-gradient. Each point represents, for the given parameter value, the frequency of such games out of 1000 randomly sampled A matrices. The solid line shows the average frequency of these games. (i) r is varied in $(0, 1)$, $b = 0$, $q = 0.01$. (ii) q is varied in $(0, 1)$, $b = 0$, $r = 0.1$. (iii) b is varied in $(-0.5, 0.5)$, $q = 0.01$, $r = 0.1$.

Theorem 12. *Suppose that $\Sigma_0 > 0$. Consider any n–player LQ game satisfying Assumptions 4 and 5 that admits a Nash equilibrium that is a saddle point of the policy-gradient dynamics—i.e., LQ games for which the Jacobian of ω evaluated at the Nash equilibrium $x^* = (K_1^*, \ldots, K_n^*)$ has eigenvalues λ_j such that $\mathbb{R}(\lambda_j) < 0$ for $j \in \{1, \ldots, \ell\}$ and $\mathbb{R}(\lambda_j) > 0$ for $j \in \{\ell + 1, \ldots, md\}$ for some ℓ such that $0 < \ell < md$. Then there exists a neighborhood U of x^* such that policy-gradient converges on a set of measure zero.*

Proof. At a high level, we characterize the set of initializations from which the players converge to Nash equilibria. The proof makes use of classic results in dynamical systems theory and topology to iteratively construct this set of initial conditions and characterize its "size". When the Nash equilibrium satisfies the strict saddle condition we show that the set is vanishingly small such that even if the players randomly initialize uniformly in an arbitrarily small ball around such solutions, the players will (almost surely) end up diverging from the equilibrium.

The proof is made up of three parts: (i) we show the existence of an open-convex neighborhood U of x^* on which ω is locally Lipschitz with constant L; (ii) we show that the map

$g(x) = x - \Gamma\omega(x)$ is a diffeomorphism on U; and, (iii) we invoke the stable manifold theorem to show that the set of initializations in U on which policy-gradient converges is measure zero. The proof of

(i) ω is locally Lipschitz. Proposition 10 shows that ω is continuously differentiable on the set of stabilizing policies $\mathcal{S}^{md}$. Given Assumptions 4 and 5, the Nash equilibrium exists and $x^* \in \mathcal{S}^{md}$. Thus, there must exist an open convex neighborhood U of x^* such that $||D\omega||_2 < L$ for some $L > 0$.

(ii) g is a diffeomorphism. By the preceding argument, ω is locally Lipschitz on U with Lipschitz constant L. Consider the policy-gradient algorithm with $\gamma_i < 1/L$ for each $i \in \{1, \dots, n\}$. Let $\Gamma = \text{diag}(\Gamma_1, \dots, \Gamma_n)$ where $\Gamma_i = \text{diag}((\gamma_i)_{j=1}^{md_i})$—that is, Γ_i is an $md_i \times md_i$ diagonal matrix with γ_i repeated on the diagonal md_i times. Now, we claim the mapping $g : \mathbb{R}^{md} \to \mathbb{R}^{md} : x \mapsto x - \Gamma\omega(x)$ is a diffeomorphism on U. If we can show that g is invertible on U and a local diffeomorphism, then the claim follows. Let us first prove that g is invertible.

Consider $x \neq y$ and suppose $g(y) = g(x)$ so that $y - x = \gamma \cdot (\omega(y) - \omega(x))$. Since $\|\omega(y) - \omega(x)\|_2 \leq L\|y - x\|_2$ on U, $\|x - y\|_2 \leq L\|\Gamma\|_2\|y - x\|_2 < \|y - x\|_2$ since $\|\Gamma\|_2 = \max_i |\gamma_i| < 1/L$.

Now, observe that $Dg = I - \Gamma D\omega(x)$. If Dg is invertible, then the implicit function theorem [5] implies that g is a local diffeomorphism. Hence, it suffices to show that $\Gamma D\omega(x)$ does not have an eigenvalue equal to one. Indeed, letting $\rho(A)$ be the spectral radius of a matrix A, we know in general that $\rho(A) \leq \|A\|$ for any square matrix A and induced operator norm $\|\cdot\|$ so that $\rho(\Gamma D\omega(x)) \leq \|\Gamma D\omega(x)\|_2 \leq \|\Gamma\|_2 \sup_{x \in U} \|D\omega(x)\|_2 < \max_i |\gamma_i| L < 1$. Of course, the spectral radius is the maximum absolute value of the eigenvalues, so that the above implies that all eigenvalues of $\Gamma D\omega(x))$ have absolute value less than one.

Since g is injective by the preceding argument, its inverse is well-defined and since g is a local diffeomorphism on U, it follows that g^{-1} is smooth on U. Thus, g is a diffeomorphism.

(iii) Local convergence occurs on a set of measure zero. Let B be the open ball derived from the central manifold theorem 5.

Starting from $x_0 \in U$, if gradient-based learning converges to a strict saddle point, then there exists an n_0 such that $g^n(x_0) \in B$ for all $n \geq n_0$. Applying Theorem 5, we get that $g^n(x_0) \in W^{cs}_{\text{loc}} \cap B$. Now, using the fact that g is invertible, we can iteratively construct the sequence of sets defined by $W_1(x^*) = g^{-1}(W^{cs}_{\text{loc}} \cap B) \cap U$ and $W_{k+1}(x^*) = g^{-1}(W_k(x^*) \cap B) \cap U$. Then we have that $x_0 \in W_n(x^*)$ for all $n \geq n_0$. The set $U_0 = \cup_{k=1}^{\infty} W_k(x^*)$ contains all the initial points in U such that gradient-based learning converges to a strict saddle.

Since x^* is a strict saddle, $I - \Gamma D\omega(x^*)$ has an eigenvalue greater than one. This implies that the co-dimension of the unstable manifold is strictly less than md so that $\dim(W^{cs}_{\text{loc}}) < md$. Hence, $W^{cs}_{\text{loc}} \cap B$ has Lebesgue measure zero in $\mathbb{R}^{md}$. Using again that g is a diffeomorphism, $g^{-1} \in C^1$ so that it is locally Lipschitz and locally Lipschitz maps are

null-set preserving. Hence, $W_k(x^*)$ has measure zero for all k by induction so that U_0 is a measure-zero set since it is a countable union of measure-zero sets. □

Theorem 12 gives sufficient conditions under which, with random initializations of K_i, policy-gradient methods would almost surely avoid the critical point. Let each players' initial strategy $K_{i,0}$ be sampled from a distribution $p_{i,0}$ for $i \in \{1, ..., n\}$, and let p_0 be the resulting the joint distribution of $(K_{1,0}, \ldots, K_{n,0})$.

Corollary 6. *Suppose $\mathcal{D}_0$ is chosen such that $\Sigma_0 \succ 0$, and consider an n–player LQ game satisfying Assumptions 4 and 5 in which there is a Nash equilibrium which is a saddle point of the policy-gradient dynamics. If each player $i \in \{1, \ldots, n\}$ performs policy-gradient with a random initial strategy $K_{i,0} \sim p_{i,0}$ such that the support of p_0 is U, they will almost surely avoid the Nash equilibrium.*

Corollary 6 shows that even if the players randomly initialize in a neighborhood of a Nash equilibrium that is a saddle point of the joint gradient dynamics they will almost surely avoid it. The proof follows trivially from the fact that the set of initializations that converge to the Nash equilibrium is of measure zero in U.

In the next section, we generate a large number of LQ games that satisfy the conditions of Corollary 6. Taken together, these theoretical and numerical results imply that policy-gradient algorithms have no guarantees of local, and consequently global, convergence in general-sum LQ games.

Remark 5. *Theorem 12 gives us sufficient conditions under which policy-gradient in general-sum LQ games does not even have* local convergence guarantees*, much less global convergence guarantees. We remark that this is very different from the single-player LQR setting, where policy-gradient will converge from any initialization in a neighborhood of the optimal solution [51]. In zero-sum LQ games, the structure of the game also precludes any Nash equilibrium from satisfying the conditions of Theorem 12 [116], meaning that local convergence is always guaranteed. In [204], the guarantee of local convergence is strengthened to that of global convergence for a class of projected policy-gradient algorithms in zero-sum LQ games.*

We conclude by noting that the non-convergence results we present extend to a stochastic setting in which the players have access to *unbiased* estimates of their gradients and the step-sizes are monotonically decreasing as time progresses. Indeed, classical results from the stochastic approximations literature (see e.g. [26, 147]) guarantee under mild assumptions on the estimators that such stochastic dynamics will have the same asymptotic behavior vis-a-vis saddle points as the deterministic dynamics they seek to follow.

4.4 Generating Counterexamples

Since it is difficult to find a simple closed form for the Jacobian of ω due to the fact that the matrices P_i implicitly depend on all the K_i, we perform random search to find instances of

LQ games in which the Nash equilibrium is a strict saddle point of the gradient dynamics. For each LQ game we generate, we use the method of Lyapunov iterations to find a global Nash equilibrium of the LQ game and numerically approximate the Jacobian to machine precision. We then check whether the Nash equilibrium is a strict saddle. Surprisingly, such a simple search procedure finds a large number of LQ games in which policy-gradient avoids Nash equilibria.

For simplicity, we focus on two-player LQ games where $z \in \mathbb{R}^2$ and $d_1 = d_2 = 1$. Thus, each player $i = 1, 2$ has two parameters to learn, which we denote $K_{i,j}$, $j = 1, 2$.

In the remainder of this section, we detail our experimental setup and then present our findings.

Experimental setup

To search for examples of LQ games in which policy-gradient avoids Nash equilibria, we fix B_1, Q_1, and R_1 and parametrize B_2, Q_2, and R_2 by b, q, and r, respectively. For various values of the parameters b, q, and r, we uniformly sample 1000 different dynamics matrices $A \in \mathbb{R}^{2\times 2}$ such that A, B_1, Q_1 satisfies Assumption 5. Then, for each of the 1000 different LQ games we find the optimal feedback matrices (K_1^*, K_2^*) using the method of Lyapunov iterations (i.e., a discrete time variant of the algorithm outlined in [102]), and then numerically approximate $D\omega(K_1^*, K_2^*)$ using auto-differentiation[1] tools and check its eigenvalues.

The exact values of the matrices are defined as follows:

$$A \in \mathbb{R}^{2\times 2} : a_{i,j} \sim \text{Uniform}(0, 1) \quad i, j = 1, 2,$$
$$B_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \; B_2 = \begin{bmatrix} b \\ 1 \end{bmatrix}, \; Q_1 = \begin{bmatrix} 0.01 & 0 \\ 0 & 1 \end{bmatrix}, \; Q_2 = \begin{bmatrix} 1 & 0 \\ 0 & q \end{bmatrix},$$
$$R_1 = 0.01, \; R_2 = r.$$

Numerical results

Using the setup outlined in the previous section we randomly generated LQ games to search for counterexamples. We first present results that show that these counterexamples may be quite common. We then use policy-gradient in two of the LQ games we generated and highlight the existence of limit cycles and the fact that the players' time-averaged strategies do not converge to the Nash equilibrium.

Avoidance of Nash in a nontrivial class of LQ games. As can be seen in Figure 4.1, across the different parameter values we considered, we found that anywhere from 0% to

[1]We use auto-differentiation due to the fact that finding an analytical expression for $D\omega$ is unduly arduous even in low dimensions due to the dependence of P_i and Σ_{K_1,K_2} on (K_1, K_2), both of which are implicitly defined.

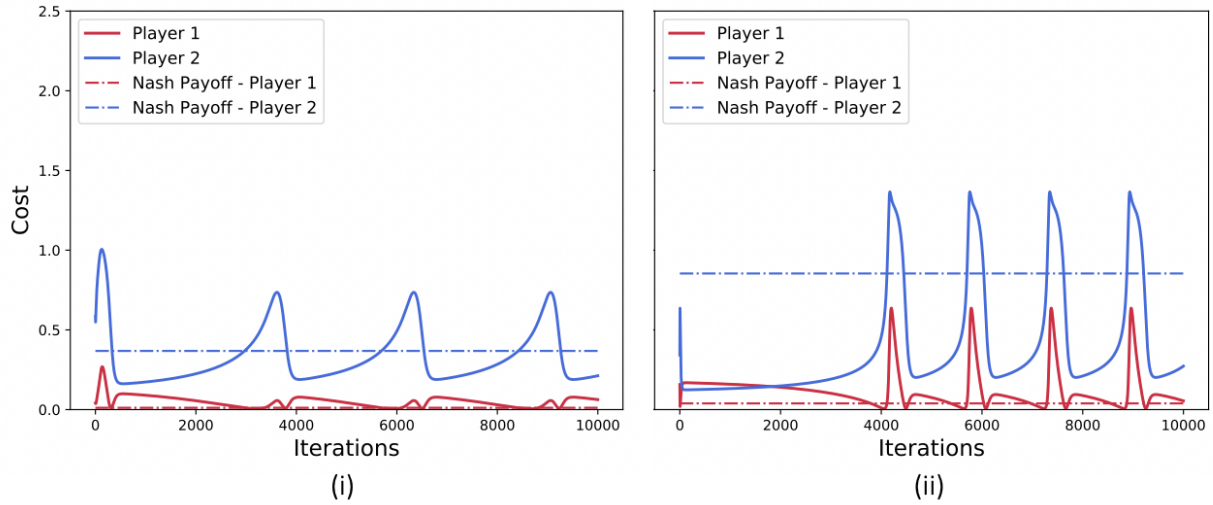


Figure 4.2: Payoffs of the two players in two general-sum LQ game where there is a Nash equilibrium that is avoided by the gradient dynamics. We observe empirically that in both games the two players diverge from the local Nash equilibrium and converge to a limit cycle around the Nash equilibrium.

25% of randomly sampled LQ games, had Nash equilibria that are strict saddle points of the gradient dynamics. Therefore, in up to 25% of the LQ games we generated policy-gradient would almost surely avoid a Nash solution. Of particular interest, for all values of q and r that we tested, when $b = 0$ at least 5% of the LQ games had a global Nash equilibrium with the strict saddle property.

These empirical observations imply that policy-gradient in competitive settings, even in the relatively straightforward setting of linear dynamics, linear policies, and quadratic costs, could fail to converge to a Nash equilibrium in up to one out of four such problems. This suggests that for more complicated cost functions, policy classes, and dynamics, Nash equilibria may often be avoided by policy-gradient.

We remark that each point in Figure 4.1 represents the number of counterexamples found (out of 1000) for each parameter value, meaning that for $r \approx 0.35, b = 0$, and $q = 0.01$ we were able to consistently generate around 250 different examples of games where policy-gradient almost surely avoids the only stationary point of the dynamics.

Note also that we were unable to find any counterexamples when b was varied in $(-0.5, 0.5)$ and $q = 0.01$, $r = 0.1$. This suggests that depending on the structure of the dynamical system it may be possible to give stronger convergence guarantees.

Convergence to Cycles. Figures 4.2–4.3 show the payoffs and parameter values of the two players when they use policy-gradient in two general-sum LQ games we identified as being counterexamples for convergence to the Nash equilibrium.

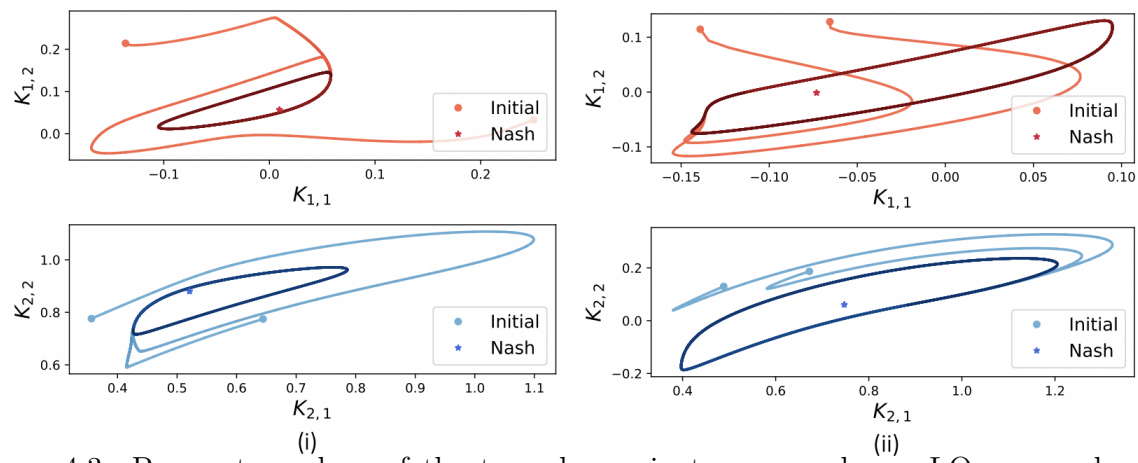


Figure 4.3: Parameter values of the two players in two general-sum LQ game where the Nash equilibrium is avoided by the gradient dynamics. We empirically observe in both games described in (4.6) that players converge to the same cycle from different initializations. Time is shown by the progressive darkening of the players' strategies.

In the two games, we initialize both players in a ball of radius 0.25 around their Nash equilibrium strategies and let them perform policy-gradient with step size 0.05. We observe that in both games the players diverge from the Nash equilibrium and converge to limit cycles.

For the two games in Figures 4.2–4.4, the game parameters are such that $b = 0$, $r = 0.01$, and $q = 0.147$. The two A matrices are defined as follows:

$$\text{(i):}\quad A = \begin{bmatrix} 0.588 & 0.028 \\ 0.570 & 0.056 \end{bmatrix}, \qquad \text{(ii):}\quad A = \begin{bmatrix} 0.511 & 0.064 \\ 0.533 & 0.993 \end{bmatrix}. \tag{4.6}$$

We also chose the initial state distribution to be $[1, 1]^T$ or $[1, 1.1]^T$ with probability 0.5 each.

The eigenvalues of the corresponding game Jacobian $D\omega$ evaluated at the Nash equilibrium are as follows:

$$\begin{aligned} \text{(i):}\quad & \text{spec}(D\omega(K_1^*, K_2^*)) = \{10.88, 2.02, -0.21, -0.06\} \\ \text{(ii):}\quad & \text{spec}(D\omega(K_1^*, K_2^*)) = \{9.76, 0.54, -0.01 \pm 0.08j\}. \end{aligned}$$

Thus, these games do satisfy the conditions of Corollary 6 for the avoidance of Nash equilibria. We conclude this section by noting that, as shown in Figure 4.4, the players' average payoffs do not necessarily converge to the Nash equilibrium payoffs.

Cycles in 3-player games. We conclude our numerical results by noting that these observations extend to $N > 2$ numbers of players. In particular, we use policy gradients in a 3-player LQ game and (as shown in Figure 4.5) we again observe convergence to limit cycles.

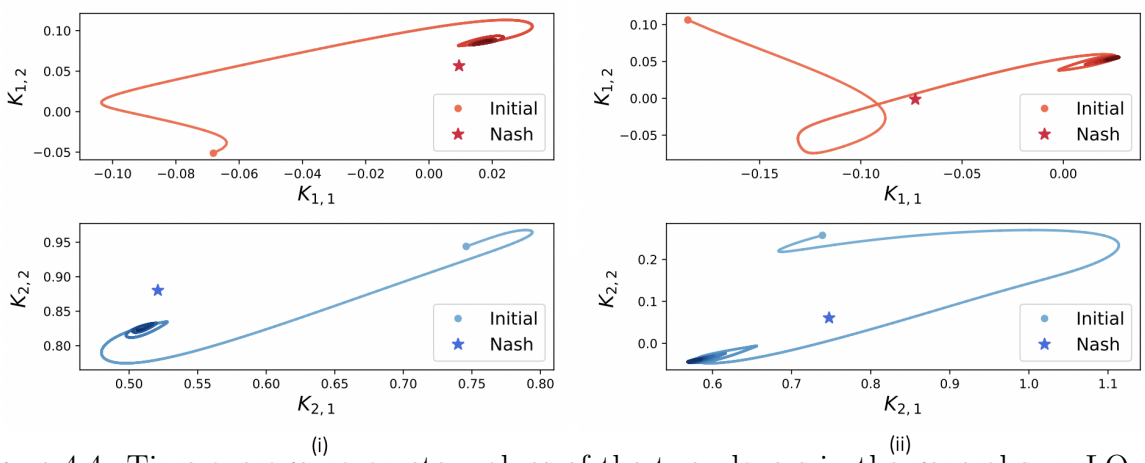


Figure 4.4: Time average parameter values of the two players in the general-sum LQ game with dynamics given in (4.6). We empirically observe that in both games the players' time average strategy does not converge to the Nash equilibrium strategy. Time is shown by progressive darkening of the players' strategies.

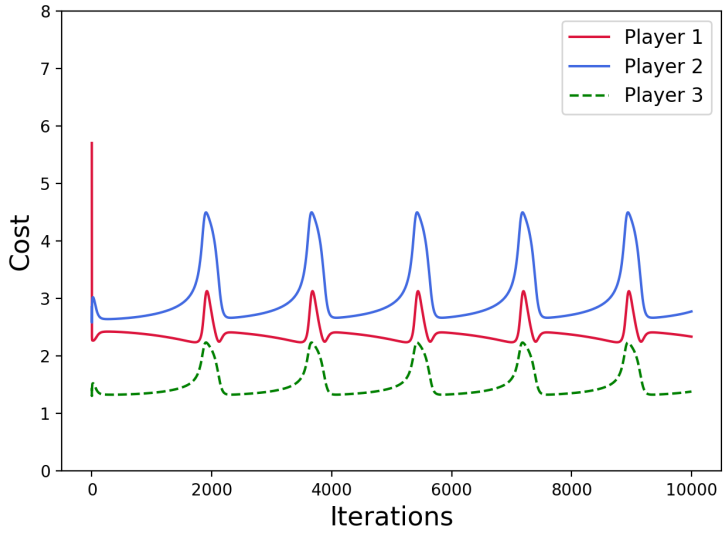


Figure 4.5: Payoffs of the three players in a LQ game. We observe empirically that the three players converge to a limit cycle instead of to a Nash equilibrium.

For this game A is the same as in example (ii) above and the first two players $i = 1, 2$ have the same cost matrices Q_i, R_i as in example (ii), and the third player has $B_3 = [0, 1]^T$, $R_3 = 0.01$, and $Q_3 = \frac{1}{2}Q_2$.

4.5 Chapter Summary

We have shown that in the relatively straightforward setting of n–player LQ games, agents performing policy-gradient have no guarantees of local, and therefore global, convergence to the Nash equilibria of the game even if they randomly initialize their first policies in a small neighborhood of the Nash equilibrium. Since we also showed that the Nash equilibria are the only critical points of the gradient dynamics, this means that, for this class of games, policy-gradient algorithms may have no guarantees of convergence to *any* set of stationary policies.

Since linear dynamics, quadratic costs, and linear policies are a relatively simple setup compared to many recent deep multi-agent reinforcement learning problems [17, 77], we believe that the issues of non-convergence are likely to be present in more complex scenarios involving more complex dynamics and parametrizations of the policies. This can be viewed as a cautionary note, but it also suggests that the algorithms that have yielded impressive results in multi-agent settings can be further improved by leveraging the underlying game-theoretic structure.

We remark that we only analyzed the deterministic policy–gradient setting, though the findings extend to settings in which players construct unbiased estimates of their gradients [179] and even actor-critic methods [178]. Indeed all of these algorithms will suffer the same problems since they all seek to track the same limiting continuous–time dynamical system and straightforward application of the results in Section 3 (or equivalently [116]) would yield the same non-convergence results for stochastic algorithms.

Our numerical experiments also highlight the existence of limit cycles in the policy-gradient dynamics. Unlike in classical optimization settings in which oscillations are normally caused by the choice of step sizes, the cycles we highlight are behaviors that can occur even with arbitrarily small step sizes. They are a fundamental feature of learning in multi-agent settings and have been observed in the dynamics of many learning algorithms [75, 116, 121, 145]. We remark, however, that there is no obvious link between the limit cycles that arise in the gradient dynamics of the LQ games and the Nash equilibrium of the game. Indeed, unlike with other game dynamics in more simple games, such as the well-studied replicator dynamics in bilinear games [121] or multiplicative weights in rock-paper-scissors [75], the time average of the players' strategies does not coincide with the Nash equilibrium. This may be due to the fact that the Nash equilibrium is a saddle point of the gradient dynamics and not simply marginally stable, though the issue warrants further investigation.

This chapter highlights how algorithms developed for classical optimization or single-agent optimal control settings may not behave as expected in multi-agent and competitive environments. Algorithms and approaches that have provable convergence guarantees and performance in competitive settings, while retaining the scalability and ease of implementation of simple policy-gradient methods, are therefore a crucial and promising open area of research. In the next chapter we develop one such algorithm in the context of zero-sum continuous games.

Chapter 5

Finding Local Nash Equilibria (and Only Local Nash Equilibria) in Zero-sum Games

The classical problem of finding Nash equilibria in multi-player games has been a focus of intense research in computer science, control theory, economics and mathematics [19, 43, 138]. Some connections have been made between this extensive literature and machine learning [see, e.g., 16, 35, 55], but these connections have focused principally on decision-making by single agents and multiple agents, and not on the burgeoning pattern-recognition side of machine learning, with its focus on large data sets and simple gradient-based algorithms for prediction and inference. This gap has begun to close in recent years, due to several new directions of research: formulations of learning problems as involving competition between subsystems that are construed as adversaries [68], concern over mismatch between assumptions and data-generating mechanisms [64, 201], the need to robustify learning systems with regard to against actual adversaries [199], and an increasing awareness that real-world machine-learning systems are often embedded in larger economic systems or networks [80].

These emerging connections bring significant algorithmic and conceptual challenges to the fore. Indeed, while gradient-based learning has been a major success in machine learning, both in theory and in practice, work on gradient-based algorithms in game theory has often highlighted their limitations. For example, gradient-based approaches are known to be difficult to tune and train [15, 45, 76, 125], and recent work has shown that gradient-based learning will almost surely avoid a subset of the local Nash equilibria in general-sum games [116]. Moreover, there is no shortage of work showing that gradient-based algorithms can converge to limit cycles or even diverge in game-theoretic settings [25, 45, 76, 121].

These drawbacks have led to a renewed interest in new approaches to finding the Nash equilibria of non-convex-concave zero-sum games, or equivalently, to solving saddle point problems. Recent work has attempted to use second-order information to reduce oscillations around equilibria and speed up convergence to fixed points of the gradient dynamics [15, 125]. Other recent approaches have attempted to tackle the problem from the variational

inequality perspective but also with an eye on reducing oscillatory behaviors [63, 124].

None of these approaches, however, address a fundamental issue that arises in non-convex-concave zero-sum games. As we will discuss, the set of attracting fixed points for the gradient dynamics in such games can include critical points that are not Nash equilibria. In fact, any saddle point of the underlying function that does not satisfy a particular alignment condition of a Nash equilibrium is a candidate attracting equilibrium for the gradient dynamics. Further, as we show, these points are attracting for a variety of recently proposed adjustments to gradient-based algorithms, including consensus optimization [125], the symplectic gradient adjustment [15], and a two-timescale version of simultaneous gradient descent [73]. Moreover, we show by counterexample that these algorithms can all converge to non-Nash stationary points.

We present a new gradient-based algorithm for finding the local Nash equilibria of two-player zero-sum games and prove that the only stationary points to which the algorithm can converge are local Nash equilibria. Our algorithm makes essential use of the underlying structure of zero-sum games. To obtain our theoretical results we work in continuous time—via an ordinary differential equation (ODE)—and our algorithm is obtained via a discretization of the ODE. While a naive discretization would require a matrix inversion and would be computationally burdensome, our discretization is a two-timescale discretization that avoids matrix inversion entirely and is of a similar computational complexity as that of other gradient-based algorithms. The resulting algorithm has stronger guarantees both with respect to the critical points to which it can converge and the manner in which it converges to them as compared to previous work on finding local Nash equilibria in non-convex-concave games. We note that these are *local* properties of the algorithm and that stronger guarantees can only be given by assuming more structure in the game.

This chapter is organized as follows. In Section 5.1 we define our notation, the problem we address, and highlight relevant related work. In Section 5.2 we define the limiting ODE that we would like our algorithm to follow and show that it has the desirable property that its only limit points are local Nash equilibria of the game. In Section 5.4 we introduce local symplectic surgery, a two-timescale procedure that asymptotically tracks the limiting ODE and show that it can be implemented efficiently. Finally, in Section 5.5 we present two numerical examples to validate the algorithm.

5.1 Preliminaries

We consider a two-player game in which one player tries to minimize a function, $f : \mathbb{R}^d \to \mathbb{R}$, with respect to their decision variable $x \in \mathbb{R}^{d_x}$, and the other player aims to maximize f with respect to their decision variable $y \in \mathbb{R}^{d_y}$, where $d = d_y + d_x$. We write such a game as $\mathcal{G} = \{(f, -f), \mathbb{R}^d\}$, since the second player can be seen as minimizing $-f$. We assume that neither player knows anything about the critical points of f, but that both players follow the rules of the game. Such a situation arises naturally when training machine learning algorithms (e.g., training generative adversarial networks or in multi-agent reinforcement

learning). Without restricting f, and assuming both players are non-cooperative, the best they can hope to achieve is a *local Nash equilibrium*; i.e., a point (x^*, y^*) that satisfies

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*),$$

for all x and y in neighborhoods of x^* and y^* respectively. Such equilibria are locally optimal for both players with respect to their own decision variable, meaning that neither player has an incentive to unilaterally deviate from such a point. As discussed in Chapter 2, generically, local Nash equilibria will satisfy slightly stronger conditions, namely they will be differential Nash equilibria (DNE).

Both differential and local Nash equilibria in two-player zero-sum games are, by definition, special saddle points of the function f that satisfy a particular *alignment condition* with respect to the player's decision variables. Indeed, the definition of differential Nash equilibria, which holds for almost all local Nash equilibria in a formal mathematical sense, makes this condition clear: the directions of positive and negative curvature of the function f at a local Nash equilibria must be *aligned* with the minimizing and maximizing player's decision variables respectively.

Issues with gradient-based algorithms in zero-sum games

Having introduced local Nash equilibria as the solution concept of interest, we now consider how to find such solutions, and in particular we highlight some issues with gradient-based algorithms in zero-sum continuous games. The most common method of finding local Nash equilibria in such games is to have both players randomly initialize their variables (x_0, y_0) and then follow their respective gradients. That is, at each step $n = 1, 2, ...$, each agent updates their variable as follows:

$$x_{n+1} = x_n - \gamma_n D_x f(x_n, y_n), \qquad y_{n+1} = y_n + \gamma_n D_y f(x_n, y_n),$$

where $\{\gamma_n\}_{n=0}^{\infty}$ is a sequence of step sizes. The minimizing player performs gradient descent on their cost while the maximizing player ascends their gradient. We refer to this algorithm as *simultaneous gradient descent* (GDA). To simplify the notation, we let $z = (x, y)$, and define the vector-valued function $\omega : \mathbb{R}^d \to \mathbb{R}^d$ as:

$$\omega(z) = \begin{bmatrix} D_x f(x, y) \\ -D_y f(x, y) \end{bmatrix}.$$

In this notation, the GDA update is given by:

$$z_{n+1} = z_n - \gamma_n \omega(z_n). \tag{5.1}$$

Since (5.1) is in the form of a discrete-time dynamical system, we proceed as in previous Chapters and examine its limiting behavior through the lens of dynamical systems theory.

Intuitively, given a properly chosen sequence of step sizes, (5.1) should have the same limiting behavior as the continuous-time flow:

$$\dot{z} = -\omega(z). \tag{5.2}$$

We can analyze this flow in neighborhoods of equilibria by studying the Jacobian matrix of ω, denoted $J : \mathbb{R}^d \to \mathbb{R}^{d\times d}$:

$$J(z) = \begin{bmatrix} D^2_{xx} f(x,y) & D^2_{yx} f(x,y) \\ -D^2_{xy} f(x,y) & -D^2_{yy} f(x,y) \end{bmatrix}. \tag{5.3}$$

We remark that the diagonal blocks of $J(z)$ are always symmetric and $D^2_{xy} f = (D^2_{yx} f)^T$. Thus $J(z)$ can be written as the sum of a block symmetric matrix $S(z)$ and a block anti-symmetric matrix $A(z)$, where:

$$S(z) = \begin{bmatrix} D^2_{xx} f(z) & 0 \\ 0 & -D^2_{yy} f(z) \end{bmatrix}, \qquad A(z) = \begin{bmatrix} 0 & D^2_{yx} f(z) \\ -D^2_{xy} f(z) & 0 \end{bmatrix}.$$

Given the structure of the Jacobian, we can now draw links between differential Nash equilibria and equilibrium concepts in dynamical systems theory.

As shown in Chapter 2, the fact that all differential Nash equilibria are critical points of ω coupled with the structure of J in zero-sum games guarantees that all differential Nash equilibria of the game are LASE of the gradient dynamics. However the converse is not true. The structure present in zero-sum games is not enough to ensure that the differential Nash equilibria are the only LASE of the gradient dynamics. When either $D^2_{xx} f$ or $D^2_{yy} f$ is indefinite at a critical point of $\omega(z)$, the Jacobian can still have eigenvalues with strictly positive real parts as shown below.

Example 4. *Consider a matrix $M \in \mathbb{R}^{2\times 2}$ having the form:*

$$M = \begin{bmatrix} a & c \\ -c & -b \end{bmatrix},$$

where $a, b \in \mathbb{R}$ and $a, b > 0$. These conditions imply that M cannot be the Jacobian of ω at an local Nash equilibria. However, if $b < a$ and $c^2 > ab$, both of the eigenvalues of M will have strictly positive real parts, and such a point could still be a LASE of the gradient dynamics.

Such points, which we refer to as non-Nash LASE of (5.2), are what makes having guarantees on the convergence of algorithms in zero-sum games particularly difficult. By definition, at least one of the two players has a direction in which they would move to unilaterally decrease their cost, meaning that such points are not locally optimal for at least one of the players. These points arise solely due to the gradient dynamics, and persist even in other gradient-based dynamics suggested in the literature. In Appendix 5.1, we show that three recent algorithms for finding local Nash equilibria in zero-sum continuous games—consensus

optimization, symplectic gradient adjustment, and a two-time scale version of GDA—are susceptible to converge to such points and therefore have no guarantees of convergence to local Nash equilibria. We note that such points can be very common since every saddle point of f that is not a local Nash equilibrium is a candidate non-Nash LASE of the gradient dynamics. Further, local minima or maxima of f could also be non-Nash LASE of the gradient dynamics.

To understand how non-Nash equilibria can be attracting under the flow of $-\omega$, we again analyze the Jacobian of ω. At such points, the symmetric matrix $S(z)$ must have both positive and negative eigenvalues. The sum of S with A, however, has eigenvalues with strictly positive real part. Thus, the anti-symmetric matrix $A(z)$ can be seen as stabilizing such points. Previous gradient-based algorithms for zero-sum games have also pinpointed the matrix A as the source of problems in zero-sum games, however they focus on a different issue. Consensus optimization [125] and the symplectic gradient adjustment [15] both seek to adjust the gradient dynamics to reduce oscillatory behaviors in neighborhoods of stable equilibria. Since the matrix $A(z)$ is anti-symmetric, it has only imaginary eigenvalues. If it dominates S, then the eigenvalues of J can have a large imaginary component. This leads to oscillations around equilibria that have been shown empirically to slow down convergence [125]. As shown in next subsection neither of the adjustments are able to rule out convergence to non-Nash equilibria.

Counter-examples for other algorithms

We now show that three state-of-the art algorithms for finding the local Nash equilibria of zero-sum games are all attracted to non-Nash equilibria. This implies that these algorithms do not have guarantees on the local optimality of their limit points and cannot give guarantees on convergence to local Nash equilibria. To do this we construct a simple counter-example in $\mathbb{R}^2$ that demonstrates that all of the algorithms are attracted to some non-Nash limit point. The particular game we use is described below.

Example 5. *Consider the game $\mathcal{G} = \{(f, -f), \mathbb{R}^2\}$ where:*

$$f(x, y) = \frac{1}{2} \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} 1 & 1 \\ 1 & 0.1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

Staying with our earlier notation, the player with variable x seeks to minimize f, and the player with variable y minimizes $-f$.

This game has only one critical point, $(x, y) = (0, 0)$, and the combined gradient dynamics for this game are linear and are given by:

$$\omega(x, y) = \begin{bmatrix} 1 & 1 \\ -1 & -0.1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

For all $x, y \in \mathbb{R}$ the Jacobian of ω is given by:

$$J(x, y) = \begin{bmatrix} 1 & 1 \\ -1 & -0.1 \end{bmatrix}.$$

Since the diagonal is not strictly positive, $(x, y) = (0, 0)$ *is not an LNE. However, since the eigenvalues of* $J(x, y)$ *are* $0.45 + 0.835i$ *and* $0.45 - 0.835i$, $(x, y) = (0, 0)$ *is a LASE of the gradient dynamics.*

We first show that even running SGA on two timescales as suggested by [73], cannot guarantee convergence to only local Nash equilibria. We assume that the maximizing player is on the slower timescale, and, in the following proposition, show that SGA on two timescales can still converge to non-Nash fixed points.

Proposition 11. *Simultaneous gradient ascent on two timescales can converge to non-Nash equilibria.*

Proof. Consider the game introduced in Example 5, and the following dynamics:

$$x_{n+1} = x_n - a_n(\omega_1(x_n, y_n))$$
$$y_{n+1} = y_n - b_n(\omega_2(x_n, y_n)),$$

where ω_i denotes the ith component of ω as described in Example 5 and a_n, b_n are sequences of step sizes satisfying Assumption 9.

Since the dynamics $\dot{x} = -x - y$ have a unique equilibrium $x = -y$ for a fixed $y \in \mathbb{R}$ and the x process is on the faster timescale, Chapter 4 in [26], assures us that $x_n \to -y_n$ asymptotically. Now, assuming that x_n has converged, we analyze the slower timescale. Plugging in for x_n, we get that, asymptotically, the dynamics will track:

$$\dot{y} = -y + 0.1y = -0.9y.$$

Since these dynamics have a unique (exponentially) stable equilibrium at $y = 0$, $(x_n, y_n) \to (0, 0)$. As we showed, the origin is non-Nash, so we are done. □

The previous proposition shows that a two-timescale version of simultaneous gradient ascent will still be susceptible to non-Nash equilibria. This implies that such an approach cannot guarantee convergence to local Nash equilibria. We now show that the consensus optimization approach introduced in [125], can also converge to non-Nash points.

Proposition 12. *Consensus optimization can converge to non-Nash equilibria.*

Proof. The update in consensus optimization is given by:

$$z_{n+1} = z_n - \gamma \left(\omega(z_n) + \lambda J^T(z_n)\omega(z_n)\right),$$

where $\lambda > 0$ is a hyperparameter to be chosen. We note that the signs are different than in [125] because we consider simultaneous gradient descent as the base algorithm while they consider simultaneous gradient ascent. The limiting dynamics of this approach is given by:

$$\dot{z} = -\left(\omega(z_n) + \lambda J^T(z_n)\omega(z_n)\right).$$

At a critical point z where $\omega(z) = 0$, the Jacobian of this dynamics is given by:

$$J_{CO}(z) = (J(z) + \lambda J^T(z)J(z)).$$

Now, in the game described in Example 5, the Jacobian of the consensus optimization dynamics would be:

$$J_{CO}(x, y) = \begin{bmatrix} 1 + 2\lambda & 1 + 1.1\lambda \\ -1 + 1.1\lambda & -0.1 + 1.01\lambda \end{bmatrix}.$$

For $\lambda > 0$, $J_{CO}(x, y)$ has eigenvalues with strictly positive real parts, which implies that, with any choice of the hyper-parameter λ, consensus optimization will converge to the non-Nash equilibrium at (0,0) in Example 5. □

The preceding proposition shows that the adjustment to the gradient dynamics proposed in [125] does not solve the issue of non-Nash fixed points. This is not surprising since the primary goal for the algorithm is to reduce oscillation around equilibria and speed up convergence to the stable equilibria of the gradient dynamics. We note that as shown in Theorem 13, the proposed algorithm also achieves this goal.

The final algorithm we consider is the symplectic gradient adjustment proposed in [15]. In the text, the authors do remark that the adjustment is not enough to guarantee convergence to only Nash equilibrium. For completeness, we make this assertion concrete by showing that the adjustment term is not enough to avoid the non-Nash equilibrium in Example 5.

Proposition 13. *The symplectic gradient adjustment to the gradient dynamics can converge to non-Nash equilibria.*

Proof. The dynamics resulting from the symplectic gradient adjustment is given by:

$$z_{n+1} = z_n - \gamma \left(\omega(z_n) + \frac{\lambda}{2}(J(z_n) - J^T(z_n))^T \omega(z_n) \right),$$

where $\lambda > 0$ is a hyperparameter to be chosen. The limiting dynamics of this algorithm is given by:

$$\dot{z} = - \left(\omega(z_n) + \frac{\lambda}{2}(J(z_n) - J^T(z_n))^T \omega(z_n) \right).$$

At a critical point z where $\omega(z) = 0$, the Jacobian of the above dynamics is given by:

$$J_{SGA}(z) = (J(z) + \frac{\lambda}{2}(J(z) - J^T(z))^T J(z)).$$

Now, in the game described in Example 5, the Jacobian of the SGA dynamics is given by:

$$J_{SGA}(x, y) = \begin{bmatrix} 1 + \lambda & 1 + 0.1\lambda \\ 1 + \lambda & -0.1 + \lambda \end{bmatrix}.$$

This has eigenvalues with strictly positive real parts for all values of $\lambda > 0$, which implies that the symplectic gradient adjustment will converge to the non-Nash equilibrium at (0,0) in Example 5. □

Other proposals have approached the issue of finding local Nash equilibria from the perspective of variational inequalities [63, 124]. In [124] and [63] extragradient methods were used to solve coherent saddle point problems and reduce oscillations when converging to saddle points. In such problems, however, the theoretical treatment focuses on settings in which all saddle points of the function f are local Nash equilibria or the cost function is implicitly assumed to be convex-concave. This in turn implies that the Jacobian satisfies the conditions for a Nash equilibrium everywhere. Thus the issue of converging to non-Nash equilibria is assumed away. The behavior of these approaches in more general zero-sum games with less structure (like the training of GANs) is therefore not clear. Moreover, since these approaches rely on averaging the gradients or are derived from implicit discretizations of the same limiting o.d.e. as GDA, they do not fundamentally change the nature of the critical points of GDA.

The problem of non-Nash equilibria has recently been addressed in the saddle point optimization literature [6]. The proposed algorithm, extreme curvature exploitation, only converges to saddle points satisfying the alignment conditions of local Nash equilibria but requires accurately computing the minimum eigenvalues of the diagonal blocks of the Jacobian, as well as their associated eigenvectors at each iteration. This makes the algorithm hard to implement in high-dimensional settings and slow to converge in practice as can be seen in Section 5.5. Further, this algorithm still exhibits the oscillatory behaviors of GDA around Nash equilibria.

In the following sections we propose an algorithm for which the only LASE are the differential Nash equilibria of the game. We also show that, regardless of the choice of hyperparameter, the Jacobian of the new dynamics at LASE has real eigenvalues, which means that the dynamics cannot exhibit oscillatory behaviors around differential Nash equilibria. Finally, we show that the algorithm can be efficiently implemented and that its speed and performance are comparable to GDA and consensus optimization when training a small GAN.

5.2 The limiting differential equation

In this section we define the continuous-time flow that our discrete-time algorithm should ideally follow.

Assumption 6 (Lipschitz assumptions on f and J)**.** *Assume that $f \in \mathcal{C}^3(\mathbb{R}^d, \mathbb{R})$ and ω is L_ω-Lipschitz. Finally assume that all critical points of ω are hyperbolic.*

Note that we do not require $J(z)$ to be invertible everywhere, but only at the critical points of f, and the assumption on the hyperbolicity of critical points holds generically for zero-sum games on continuously differentiable functions f from the results in Chapter 2 and [114].

Now, consider the continuous-time flow:

$$\dot{z} = -\frac{1}{2}\left(\omega(z) + J^{-1}(z)J^T(z)\omega(z)\right). \tag{5.4}$$

The dynamics introduced in (5.4) can be seen as an adjusted version of the gradient dynamics where the adjustment term only allows trajectories to approach critical points of ω along the players' axes. If a critical point is not locally optimal for one of the players (i.e., it is a non-Nash critical point) then that player can push the dynamics out of a neighborhood of that point.

As we show in the following theorem the Jacobian of the adjustment is similar to J^T when $||\omega(z)||_2$ is small. This approximation is exact at critical points of ω. Adding this adjustment term to ω exactly cancels out the rotational part of the vector field contributed by the antisymmetric matrix $A(z)$ in a neighborhood of critical points. Since we identified $A(z)$ as the source of oscillatory behavior and non-Nash equilibria in Section 5.1, this adjustment addresses both of these issues. The following theorem establishes this formally.

Theorem 13. *Under Assumption 6 and if for all $z \in \mathbb{R}^d$ such that $\omega(z) \neq 0$ we have:*

$$S(z)\omega(z) \neq 0,$$

then the continuous-time dynamical system in(5.4) *satisfies:*

- *z is a LASE of $\dot{z} = -h(z) \iff$ z is a differential Nash equilibrium of $\mathcal{G} = \{(f, -f), \mathbb{R}^d\}$.*
- *If z is a critical point of $h(z)$, then the Jacobian of h at z has real eigenvalues.*

Proof. To prove Theorem 13, we first show that:

$$h(z) = 0 \iff \omega(z) = 0.$$

Clearly, $\omega(z) = 0 \implies h(z) = 0$. To show the converse, we assume that $h(z) = 0$ but $\omega(z) \neq 0$. This implies that:

$$J^T(z)\omega(z) = -J(z)\omega(z) \implies 2S(z)\omega(z) = 0$$

Since we assumed that this cannot be true, we must have that $h(z) = 0 \implies \omega(z) = 0$.

Having shown that under our assumptions, the critical points of h are the same as those of ω, we now note that the Jacobian of $h(z)$ at a critical point must have the form:

$$\begin{aligned} J_h(z) &= \frac{1}{2}\left(J(z) + J^{-1}(z)J^T(z)J(z)\right) \\ &= J^{-1}(z)S(z)J(z). \end{aligned}$$

At critical points, $J(z)$ is invertible and $\lambda(z) = 0$ by assumption. Given that $\omega(z) = 0$, terms that include $\omega(z)$ disappear, and the adjustment term contributes only a factor of $J^T(z)$ to

the Jacobian of h at a critical point. This exactly cancels out the antisymmetric part of the Jacobian of ω. The Jacobian of h is therefore similar to a symmetric matrix at critical points of ω and has positive eigenvalues only when $D^2_{xx} f(z) \succ 0$ and $D^2_{yy} f(z) \prec 0$.

Since these are also the conditions for differential Nash equilibria, all differential Nash equilibria of $\mathcal{G}$ must be LASE of $\dot{z} = -h(z)$. Further, non-Nash LASE of $\dot{z} = -\omega(z)$ cannot be LASE of $\dot{z} = -h(z)$, since by definition either $D^2_{xx} f(z)$ or $D^2_{yy} f(z)$ is indefinite at such points. To show the second part of the theorem, we simply note that J_h must be symmetric at all critical points which in turn implies that it has only real eigenvalues. □

Theorem 13 shows that the only attracting hyperbolic equilibria of the limiting ordinary differential equation (ODE) are the differential Nash equilibria of the game. Also, since $J_h(z)$ is symmetric at critical points of ω, if either $D^2_{xx} f(z)$ or $-D^2_{yy} f(z)$ has at least one negative eigenvalue then such a point would be a linearly unstable equilibrium of $\dot{z} = -h(z)$. Such points are linearly unstable and are therefore almost surely avoided when the algorithm is randomly initialized [24, 168]. Theorem 13 also guarantees that the continuous-time solutions do not oscillate near critical points. Reducing oscillations near critical points is the main goal of consensus optimization [125] and the symplectic gradient adjustment [15]. However, for both algorithms, the extent to which they are able to reduce the oscillations depends on the choice of hyperparameter.

Remark 6. *The assumption that $\omega(z)$ is not in the nullspace of $S(z)$ when $\omega(z)$ is nonzero ensures that the adjustment does not create new critical points. This can cause problems in settings with little to no second order curvature like e.g., in bilinear games, but in the next section we show how to alleviate this problem by introducing some regularization. We also further expand on this theorem in full generality in Section 5.6.*

.

5.3 Rates in Structured Games

In this section, we derive rates of convergence for local symplectic surgery in structured classes of games. In particular, we show how, by design, local symplectic surgery achieves the optimal rate in strongly-monotone games and how adding regularization to symplectic surgery gives efficient, last-iterate convergence guarantees in bilinear, and convex-concave games.

Remark 7. *We remark that other algorithms such as extra-gradient algorithms or proximal-point methods can also achieve such rates in structured games, but, unlike LSS, have no guarantees of convergence to game-relevant equilibria in less structured settings.*

Last-iterate convergence in strongly monotone games

To begin, we focus on convergence in strongly monotone zero-sum games, which have been analyzed in [14, 29]. Recall, that a zero-sum game is μ-strongly monotone if ω satisfies:

$$\left(\omega(z) - \omega(z')\right)^T (z - z') \geq \mu \|z - z'\|^2 \quad \forall z \in \mathbb{R}^d.$$

In this class of games—also known as strongly-convex, strongly concave games—there is a unique critical point of the gradient dynamics which is also a global Nash equilibrium. Further, by definition, we are guaranteed that the matrix $S(z)$ is positive definite for all z, and consequently that $J(z)$ is invertible everywhere.

Given this assumption, we show that the simple Euler discretization of (5.4) given by:

$$z_{t+1} = z_t - \gamma \left(\omega(z_t) + J^{-1}(z_t) J^T(z_t) \omega(z_t)\right), \tag{5.5}$$

converges at a fast rate. The proof relies on the intuition that the function,

$$V(z) = \frac{1}{2} \|\omega(z)\|^2,$$

serves as a Lyapunov function for the continuous-time dynamics since:

$$\dot{V}(z) < -\omega(z)^T S(z) \omega(z) < -\mu \|\omega(z)\|^2$$

We make this concrete in the following theorem, and doing so requires a smoothness assumption on the Lyapunov function, which is common in the analysis of other second-order algorithms in the literature (see e.g., [4, 15, 52]):

Assumption 7. *$V(x) = \frac{1}{2}\|\omega(z)\|^2$ is L-smooth, meaning that:*

$$\|J^T(z)\omega(z) - J^T(z')\omega(z')\| \leq L\|z - z'\| \quad \forall z, z'$$

Given this assumption, let $\kappa = \frac{L}{\mu}$ we now prove that our discretized process enjoys fast convergence in strongly monotone games.

Theorem 14. *Under Assumption 7, suppose the game is μ-strongly monotone, then for $0 < \gamma <$, the iterates of (5.5) satisfy, for all $t > 0$:*

$$\|\omega(z_{t+1})\|^2 \leq \left(1 - \mu\gamma\left(1 - 2\gamma\frac{2\kappa^3}{\mu^2}\right)\right)^t \|\omega(z_0)\|^2$$

and choosing $\gamma < \frac{\mu^2}{2\kappa^3}$ results in exponentially fast convergence.

Proof. To begin, we take a Taylor expansion of $\|\omega(z_{t+1})\|^2$, where for simplicity we let, $h(z) = \omega(z) + J^{-1}(z)J^T(z)\omega(z)$, and z^* the unique critical point of ω.

$$\begin{aligned}
\frac{1}{2}\|\omega(z_{t+1})\|^2 &\leq \frac{1}{2}\|\omega(z_t)\|^2 - \gamma\omega(z_t)^T J(z_t)\left[\omega(z_t) + J^{-1}(z_t)J^T(z_t)\omega(z_t)\right] + \frac{L\gamma^2}{2}\|h(z_t)\|^2 \\
&\leq (\frac{1}{2} - \mu\gamma)\|\omega(z_t)\|^2 + L\gamma^2\|\omega(z_t)\|^2 + L\gamma^2\|J^{-1}(z_t)J^T(z_t)\omega(z_t)\|^2 \\
&\leq (\frac{1}{2} - \mu\gamma + L\gamma^2)\|\omega(z_t)\|^2 + \frac{L^3}{\mu^4}\gamma^2\|\omega(z_t)\|^2 \\
\|\omega(z_{t+1})\|^2 &\leq \left(1 - 2\mu\gamma\left(1 - \gamma\frac{2\kappa^3}{\mu^2}\right)\right)\|\omega(z_t)\|^2
\end{aligned}$$

where we used the fact that $J^TJ > \mu^2 I$ (see e.g., [4].), the smoothness of $J^T(z)\omega(z)$, $\kappa > 1$ and the fact that strong monotonicity implies that: $\mu\|z - z'\| \leq \|\omega(z) - \omega(z')\|$. We also for simplicity, assumed that $\mu < 1$ to help us simplify. Thus, choosing $\gamma < \frac{\mu^2}{2\kappa^3}$, makes the leading term on the right hand side above strictly less than 1. By recursion, we find our result. □

.

Last-iterate convergence in convex-concave games

A key class of games where the approach of cancelling out the symplectic part of the vector field poses problems is in bilinear games. To rectify this, we introduce a regularized version of the discrete-time system:

$$z_{t+1} = z_t - \gamma_t(\omega(z_t) + J^{-1}(z_t)(J^T(z_t) + \lambda I)\omega(z_t)) \tag{5.6}$$

with the introduction of the parameter λ. This version of the algorithm enjoys fast, last-iterate convergence in bilinear games. To see this, we note that in bilinear games we have:

$$\omega(z) = Jz \qquad J(z) = J = -J^T(z) \forall z \in \mathbb{R}^d$$

Thus, in this case the dynamics above simplify to:

$$\begin{aligned}
z_{t+1} &= z_t - \gamma(Jz_t + J^{-1}(J^T + \lambda I)Jz_t) \\
&= z_t - \gamma\lambda z_t,
\end{aligned}$$

which results in the following bound:

$$\begin{aligned}
\|z_T\| &\leq (1 - \gamma\lambda)^2\|z_{T-1}\| \\
&\leq \prod_{t=1}^{T-1}(1 - \gamma\lambda)\,\|z_0\|.
\end{aligned}$$

Choosing $\gamma < 1$ and $\lambda < 1$ gives a problem-independent linear rate of convergence:

$$\|z_T\| \leq (1 - \gamma\lambda)^{T-1}\|z_0\|$$

This highlights the need for some form of regularization to allow us to obtain rates in convex-concave games where we do not have the additional structure of strong monotonicity. In the next theorem we give rates for this regularized process without assuming strong monotonicity. To do so, we need another layer of smoothness on ω:

Assumption 8. *f is L_ω-smooth, meaning that:*

$$\|\omega(z) - \omega(z')\| \leq L_\omega\|z - z'\| \quad \forall z, z'$$

Letting $\kappa = \frac{L_\omega}{\sigma}$, we now prove the following theorem:

Theorem 15. *Under Assumptions 7 and 8, suppose f is strictly convex-concave and L-smooth in x and y respectively. Further assume that $J^T(z)J(z) > \sigma^2$ for all z. Then the process in (5.6) with stepsize $\gamma < \min\left(\frac{1}{8\kappa^2}\lambda, \frac{1}{4L}\lambda, \frac{\sigma^2}{8}\right)$, and $\lambda < 1$ satisfies:*

$$\|\omega(z_{t+1})\|^2 \leq \left(1 - \frac{\gamma\lambda}{2}\right)\|\omega(z_t)\|^2$$

Further assuming that $\kappa^2 > L$ and $\sigma^2 > \frac{1}{\kappa^2}$ the process satisfies:

$$\|\omega(z_{t+1})\|^2 \leq \left(1 - \frac{\lambda^2}{16\kappa^2}\right)\|\omega(z_t)\|^2$$

We remark that. the assumption that $J^T(z)J(z) > \sigma^2$ does not imply strong monotoncity, and is in fact a form of the *sufficiently bilinear* assumption which is common in the analysis of second-order dynamics in zero-sum games(see e.g., [4]). Indeed, in bilinear games this is equivalent to assuming that the anti symmetric matrix $A(z)$ satisfies $A(z)^T A(z) \succ \sigma^2 I \ \forall z$.

Proof. Once again, we have that $\frac{1}{2}\|\omega(z)\|^2$ is L-smooth, and the loss function is convex-concave and L-smooth in x and y respectively. Denote $h_\lambda(z) = \omega(z) + J^{-1}(z)(J^T(z)_\lambda(t))\omega(z)$, then taking the Taylor expansion of $\frac{1}{2}\|\omega(z)\|^2$ gives:

$$\frac{1}{2}\|\omega(z_{t+1})\|^2 \leq \frac{1}{2}\|\omega(z_t)\|^2 - \gamma\omega(z_t)^T S(z_t)\omega(z_t) - \gamma\lambda\|\omega(z_t)\|^2 + \frac{\gamma^2 L}{2}\|h_\lambda(z_t)\|^2.$$

Since f is convex-concave, $S \succeq 0$. Thus, this simplifies to:

$$\|\omega(z_{t+1})\|^2 \leq \|\omega(z_t)\|^2 - 2\gamma\lambda\|\omega(z_t)\|^2 + \gamma^2 L\|h_\lambda(z_t)\|^2.$$

Algorithm 1 Local Symplectic Surgery

Input Functions f, ω, J, λ; Step sizes a_n, b_n; Initial values (x_0, y_0, v_0)
Initialize $(x, y, v, n) \leftarrow (x_0, y_0, v_0, 0)$
while not converged **do**
  $g_x \leftarrow D_x \left[f(x, y) + v \right]$
  $g_y \leftarrow D_y \left[-f(x, y) + v \right]$
  $g_v \leftarrow D_v \left[||J(x, y)v - J(x, y)^T \omega(x, y) + \lambda \omega(x, y)||_2^2 \right]$
  $(x, y, v) \leftarrow (x - a_n g_x, y - a_n g_y, v - b_n g_v)$
  $n \leftarrow n + 1$
end while
Output $(x^*, y^*) \leftarrow (x, y)$

Expanding the last term gives us that:

$$\begin{aligned}
\|\omega(z_{t+1})\|^2 &\leq (1 - 2\gamma\lambda + 2\gamma^2 L)\|\omega(z_t)\|^2 + 4\gamma^2 \omega(z_t)^T J^T(z_t)(J^T(z_t)J(z_t))^{-1} J(z_t)\omega(z_t) \\
&\quad + 4\gamma^2\lambda^2 \omega(z_t)^T (J^T(z_t)J(z_t))^{-1}\omega(z_t) \\
&\leq \left(1 - 2\gamma\lambda + 2\gamma^2 L + \frac{4\gamma^2 L_\omega^2 + 4\gamma^2\lambda^2}{\sigma^2}\right) \|\omega(z_t)\|^2 \\
&= \left(1 - 2\gamma\left(\lambda - \gamma L - \frac{2\gamma L_\omega^2 + 2\gamma\lambda^2}{\sigma^2}\right)\right) \|\omega(z_t)\|^2
\end{aligned}$$

Choosing $\gamma < \min\left(\frac{\sigma^2}{8L_\omega^2}\lambda, \frac{1}{4L}\lambda, \frac{\sigma^2}{8}\right)$, and $\lambda < 1$ gives us that:

$$\|\omega(z_{t+1})\|^2 \leq \left(1 - \gamma\lambda + \frac{\gamma\lambda^2}{2}\right) \|\omega(z_t)\|^2$$

Since $\lambda < 1$ we can further simplify to find that:

$$\|\omega(z_{t+1})\|^2 \leq \left(1 - \frac{\gamma\lambda}{2}\right) \|\omega(z_t)\|^2$$

□

5.4 Efficient Implementation through a Two-Timescale Approximation

Given the limiting ODE, we could perform a straightforward Euler discretization to obtain a discrete-time update having the form: $z_{n+1} = z_n - \gamma h(z_n)$.

However, due to the matrix inversion, such a discrete-time update would be prohibitively expensive to implement in high-dimensional parameter spaces such as those encountered when training GANs. To solve this problem, we now introduce a two-timescale approximation to the continuous-time dynamics that has the same limiting behavior, but is much faster to compute at each iteration than the simple discretization. Since this procedure serves to exactly remove $A(z)$, the symplectic part of the Jacobian, in neighborhoods of hyperbolic critical points, we refer to this two-timescale procedure as local symplectic surgery (LSS). In Appendix 5.6 we derive the two-timescale update rule for the time-varying version of the limiting ODE and show that it also has the same properties.

The two-timescale approximation to (5.4) is given by:

$$z_{n+1} = z_n - a_n h_1(z_n, v_n) \qquad v_{n+1} = v_n - b_n h_2(z_n, v_n), \tag{5.7}$$

where h_1 and h_2 are defined as:

$$h_1(z, v) = \frac{1}{2}\left(\omega(z) + v\right)$$
$$h_2(z, v) = J^T(z)J(z)v - \left(J^T(z)\right)^2 \omega(z),$$

and the sequences of step sizes $\{a_n\}_{n=0}^{\infty}$,$\{b_n\}_{n=0}^{\infty}$ satisfy the following assumptions:

Assumption 9 (Assumptions on the step sizes)**.** *The sequences* $\{a_n\}_{n=0}^{\infty}$ *and* $\{b_n\}_{n=0}^{\infty}$ *satisfy:*

- $\sum_{i=1}^{\infty} a_i = \infty$*, and* $\sum_{i=1}^{\infty} b_i = \infty$*;*
- $\sum_{i=1}^{\infty} a_i^2 < \infty$*, and* $\sum_{i=1}^{\infty} b_i^2 < \infty$*;*
- $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$.

We note that h_2 is Lipschitz continuous in v uniformly in z under Assumption 6.

The v process performs gradient descent on a regularized version of least squares, where the regularization is governed by $\lambda(z)$. If the v_n process is on a faster time scale, the intuition is that it will first converge to $J(z_n)^{-1}J^T(z_n)\omega(z_n)$, and then z_n will track the limiting ODE in (5.4). In the next subsection we prove that this behavior holds even in the presence of noise.

Before showing this, however, we briefly note that the key benefit to the two-timescale process is that z_{n+1} and v_{n+1} can be computed efficiently since neither require a matrix inversion. In fact, the computation can be done with auto-differentiation tools with the same order of complexity as that of GDA, consensus optimization, and the symplectic gradient adjustment. In particular, using Jacobian-vector products, the computation of h_1 and h_2 can be done relatively quickly.

We first note that a Jacobian-vector product calculates $J^T(z)u$ for a constant vector $u \in \mathbb{R}^d$, by calculating the gradient of $\omega^T(z)u$ with respect to v. This allows us to write the x and y updates in Algorithm 1 in a clean form. The next proposition shows that, using

the structure of J that was discussed in Section 5.1, the term $J(z)v$ can also be efficiently computed. Together these results show that the per-iteration complexity of LSS is on the same order as that of consensus optimization and the symplectic gradient adjustment.

Proposition 14. *The computation of $J(z)v$ requires two Jacobian-vector products, for $v \in \mathbb{R}^d$ a constant vector.*

Proof. Let $v = (v_1, v_2)^T$ where $v_1 \in \mathbb{R}^{d_x}$ and $v_2 \in \mathbb{R}^{d_y}$. Then $J(z)v$ can be written as:

$$J(z)v = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} J^T(z) \begin{bmatrix} v_1 \\ 0 \end{bmatrix} + \begin{bmatrix} -I & 0 \\ 0 & I \end{bmatrix} J^T(z) \begin{bmatrix} 0 \\ v_2 \end{bmatrix}.$$

We note that the above expression is only possible due to the structure of the Jacobian in two-player zero-sum games. Having written $J(z)$ as above, it is now clear that computing $J(z)v$ will require two Jacobian-vector products. □

This simple technique gives rise to the procedure outlined in Algorithm 1.

Long-term behavior of the two-timescale approximation

We now show that LSS asymptotically tracks the limiting ODE even in the presence of noise. This implies that the algorithm has the same limiting behavior as (5.4). In particular, our setup allows us to treat the case where one only has access to unbiased estimates of h_1 and h_2 at each iteration. This is the setting most likely to be encountered in practice, for example in the case of training GANs in a mini-batch setting. We assume that we have access to estimators $\hat{h}_1$ and $\hat{h}_2$ such that:

$$\mathbb{E}\left[\hat{h}_1(z, v)\right] = \omega(z) + v$$
$$\mathbb{E}\left[\hat{h}_2(z, v)\right] = J^T(z)J(z)v + (J^T(z))^2\omega(z).$$

To place this in the form of classical two-timescale stochastic approximation processes, we write each estimator $\hat{h}_1$ and $\hat{h}_2$ as the sum of its mean and zero-mean noise processes M^z and M^v respectively. This results in the following two timescale process:

$$z_{n+1} = z_n - a_n[\omega(z_n) + v_n + M^z_{n+1}] \tag{5.8}$$
$$v_{n+1} = v_n - b_n[J^T(z_n)J(z_n)v_n + (J^T(z_n))^2\omega(z_n) + M^v_{n+1}]. \tag{5.9}$$

We assume that the noise processes satisfy the following standard conditions [23, 26]:

Assumption 10. *Assumptions on the noise: Define the filtration $\mathcal{F}_n$:*

$$\mathcal{F}_n = \sigma(z_0, v_0, M^v_1, M^z_1, ..., M^z_n, M^v_n),$$

for $n \geq 0$. Given $\mathcal{F}_n$, we assume that:

- *M^v_{n+1} and M^z_{n+1} are conditionally independent given $\mathcal{F}_n$ for $n \geq 0$.*
- *$\mathbb{E}[M^v_{n+1}|\mathcal{F}_n] = 0$ and $\mathbb{E}[M^z_{n+1}|\mathcal{F}_n] = 0$ for $n \geq 0$.*
- *$\mathbb{E}[||M^z_{n+1}|||\mathcal{F}_n] \leq c_z(1 + ||z_n||)$ and $\mathbb{E}[||M^v_{n+1}|||\mathcal{F}_n] \leq c_v(1 + ||z_n||)$ almost surely for some positive constants c_z and c_v.*

Given our assumptions on the estimator, cost function, and step sizes we now show that (5.8) asymptotically tracks a trajectory of the continuous-time dynamics almost surely. Since h_1, h_2, and $v^*(z) = J(z)^{-1}J^T(z)\omega(z)$ are not uniformly Lipschitz continuous in both z and v, we cannot directly invoke results from the literature. Instead, we adapt the proof of Theorem 2 in Chapter 6 of [26] to show that $v_n \to v^*(z_n)$ almost surely. We then invoke Proposition 4.1 from [23] to show that z_n asymptotically tracks (5.4). We note that this approach only holds on the event $\{\sup_n ||z_n|| + ||v_n|| < \infty\}$. Thus, if the stochastic approximation process remains bounded, then under our assumptions we are sure to track a trajectory of the limiting ODE.

Lemma 3. *Under Assumptions 6-10, and on the event $\{\sup_n ||z_n||_2 + ||v_n||_2 < \infty\}$:*

$$(z_n, v_n) \to \{(z, v^*(z)) : z \in \mathbb{R}^d\},$$

almost surely.

Proof. We first rewrite (5.8) as:

$$\begin{aligned} z_{n+1} &= z_n - b_n \left[\frac{a_n}{b_n} h_1(z_n, v_n) + \bar{M}^z_{n+1}\right] \\ v_{n+1} &= v_n - b_n[h_2(z_n, v_n) + M^v_{n+1}], \end{aligned}$$

where $\bar{M}^z_{n+1} = \frac{a_n}{b_n} M^z_{n+1}$. By assumption, $\frac{a_n}{b_n} \to 0$. Since h_1 is locally Lipschitz continuous, it is bounded on the event $\{\sup_n ||z_n||_2 + ||v_n||_2 < \infty\}$. Thus, $\frac{a_n}{b_n} h_1(z_n, v_n) \to 0$ almost surely.

From Lemma 1 in Chapter 6 of [26] , the above processes, on the event $\{\sup_n ||z_n||_2 + ||v_n||_2 < \infty\}$, converge almost surely to internally chain-transitive invariant sets of $\dot{v} = -h_2(z, v)$ and $\dot{z} = 0$. Since, for a fixed z, $h_2(z, v)$ is a Lipschitz continuous function of v with a globally asymptotically stable equilibrium at $(J^T(z)J(z) + \lambda(z)I)^{-1}\omega(z)$, the claim follows. □

Having shown that $||v_n - v^*(z_n)||_2 \to 0$ almost surely, we now show that z_n will asymptotically track a trajectory of the limiting ODE. Let us first define $z(t, s, z_s)$ for $t \geq s$ to be the trajectory of $\dot{z} = -h(z)$ starting at z_s at time s.

Theorem 16. *Given Assumptions 6-10, let $t_n = \sum_{i=0}^{n-1} a_i$. On the event $\{sup_n ||z_n||_2 + ||v_n||_2 < \infty\}$, for any integer $K > 0$ we have:*

$$\lim_{n\to\infty} \sup_{0 \leq h \leq K} \|z_{n+h} - z(t_{n+h}, t_n, z_n)\|_2 = 0.$$

The proof of Theorem 16 relies on a combination of Proposition 4.1 and 4.2 from [23] which gives us conditions under which a stochastic approximation process is an asymptotic pseudo-trajectory of the underlying ODE:

Proposition 15. *Let h be a continuous globally integrable vector field. Further, let $x(t, s, x_s)$ for $t \geq s$ be a trajectory of the dynamical system $\dot{x} = -h(x)$ starting from state x_s at time s. Finally let the stochastic approximation process be given by:*

$$x_{n+1} = x_n + a_n(h(x_n) + \chi_n + M_{n+1}),$$

where:

1. $\sup_n ||x_n|| < \infty$
2. $\sup_n \mathbb{E}[||M_n||^2 < \infty$
3. $\sum_{i=0}^{\infty} a_n^2 < \infty$
4. $\sum_{i=0}^{\infty} a_n = \infty$
5. $\lim_{n\to\infty} \chi_n = 0$ *almost surely.*

Then:

$$\lim_{n\to\infty} \sup_{0\leq h\leq K} ||x_{n+h} - x(t_{n+h}, t_n, x_n)|| = 0.$$

Given this proposition, we now develop the proof.

Proof. We first rewrite the z_n process as:

$$z_{n+1} = z_n - a_n \left[h(z) - J^T(z_n)\left(v^*(z_n) - v_n\right) + M_{n+1}^z\right].$$

We note that, from Lemma 3, $(v^*(z_n) - v_n) \to 0$ almost surely. Since $||J^T(z_n)||_2 < L_\omega$, we can write this process as:

$$z_{n+1} = z_n - a_n \left[h(z) - \chi_n + M_{n+1}^z\right],$$

where $\chi_n \to 0$ almost surely. Since h is continuously differentiable, it is locally Lipschitz, and on the event $\{sup_n||z_n|| + ||v_n|| < \infty\}$ it is bounded. It thus induces a continuous globally integrable vector field, and therefore satisfies the assumptions for Propositions 4.1 in [23]. Further, by assumption the sequence of step sizes and martingale difference sequences satisfy the assumptions of Proposition 4.2 in [23]. Invoking Proposition 4.1 and 4.2 in [23] gives us the desired result. □

Theorem 16 guarantees that LSS asymptotically tracks a trajectory of the limiting ODE. Coupled with standard assumptions on the noise, classic results from the stochastic approximation literature (e.g., [147]) guarantees that the process will avoid non-Nash equilibria of the gradient dynamics which are saddle points of the dynamics. Thus the only locally asymptotically stable points for LSS must be the differential Nash equilibria of the game.

5.5 Numerical Examples

We now present two numerical examples that illustrate the performance of both the limiting ODE and LSS. The first is a zero-sum game played over a function in $\mathbb{R}^2$ that allows us to observe the behavior of both the limiting ODE around both local Nash and non-Nash equilibria. In the second example we use LSS to train a small generative adversarial network (GAN).

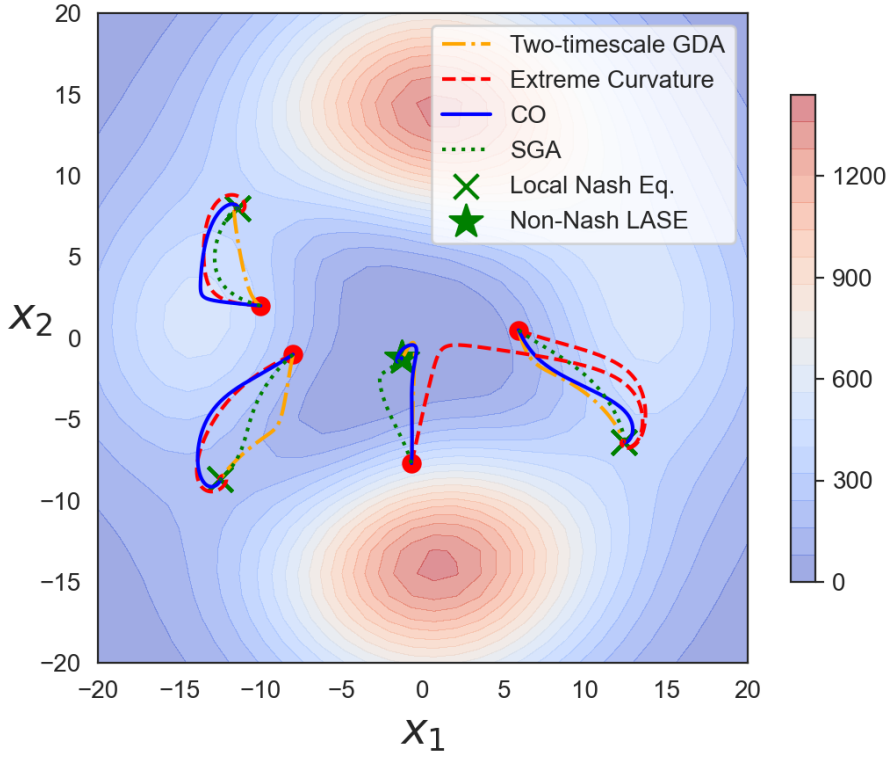


Figure 5.1: Convergence of extreme curvature exploitation, consensus optimization, the symplectic gradient adjustment and GDA to different equilibria in the zero-sum game played on (5.10). The local Nash equilibria are denoted by 'x' and the non-Nash equilibrium by a star. Consensus optimization, GDA, and the symplectic gradient adjustment both converge to non-Nash equilibria. Extreme curvature exploitation avoids the non-Nash equilibrium.

2-D example

For the first example, we consider the game based on the following function f in $\mathbb{R}^2$:

$$f(x, y) = e^{-0.01(x^2+y^2)}((0.3x^2 + y)^2 + (0.5y^2 + x)^2). \tag{5.10}$$

This function is a fourth-order polynomial that is scaled by an exponential to ensure that it is bounded. The gradient dynamics $\dot{z} = -\omega(z)$ associated with function have four LASE, three of the which are local Nash equilibria. In Figure 5.1, we plot the sample paths of extreme curvature exploitation, consensus optimization, the symplectic gradient adjustment (SGA), simultaneous gradient descent (GDA) and our limiting ODE from the same initial positions, shown with red dots. We clearly see that GDA, consensus optimization, and the symplectic gradient adjustment converge to all four LASE, depending on the initialization. Extreme curvature exploitation, on the other hand, only converges to the local Nash equilibria.

In Figure 5.2 we empirically validate that LSS asymptotically tracks the limiting ODE. When the fast timescale has not converged, the process tracks the gradient dynamics. Once it has converged however, we see that it closely tracks the limiting ODE which leads it to converge to only the local Nash equilibria. Figure 5.2. also highlights how, unlike extreme curvature exploitation, our algorithm does not oscillate around equilibria, which can lead to faster convergence. This also empirically validates the second part of Theorem 13.

The discretized full information process is calculated as:

$$z_{t+1} = z_t - \gamma(\omega(z) + v),$$

where $v = J^{-1}(z)J^T(z)\omega(z)$. For the two-timescale process, since there is no noise we use constant step sizes and the following update:

$$\begin{aligned} z_{n+1} &= z_n - \gamma_1(\omega(z_n) + v_n) \\ v_{n+1} &= v_n - \gamma_2(J^T(z_n)J(z_n)v_n - \left(J^T(z_n)\right)^2\omega(z_n)), \end{aligned}$$

where $\gamma = 0.0015$,$\gamma_1 = 0.0001$, and $\gamma_2 = 0.0005$.

For consensus optimization, the symplectic gradient adjustment, and extreme curvature exploitation the updates are given by:

$$\begin{aligned} z_{n+1} &= z_n - \gamma(\omega(z_n) - \gamma_{CO}J^T(z_n)\omega(z_n)), \\ z_{n+1} &= z_n - \gamma(\omega(z_n) - \gamma_{SGA}A^T(z_n)\omega(z_n)), \\ z_{n+1} &= z_n - \gamma(\omega(z_n) - \gamma_{CURV}u(z_n)), \end{aligned}$$

respectively, where $\gamma = 0.002$, $\gamma_{CO} = \gamma_{SGA} = 0.1$, $\gamma_{CURV} = 1.0$, and $u(z_n)$ is the adjustment as described in [6].

Generative adversarial network

We now train a generative adversarial network with LSS, GDA, the symplectic gradient adjustment (SGA), consensus optimization (CO), and extreme curvature (EC) exploitation. Both the discriminator and generator are fully connected neural networks with four hidden

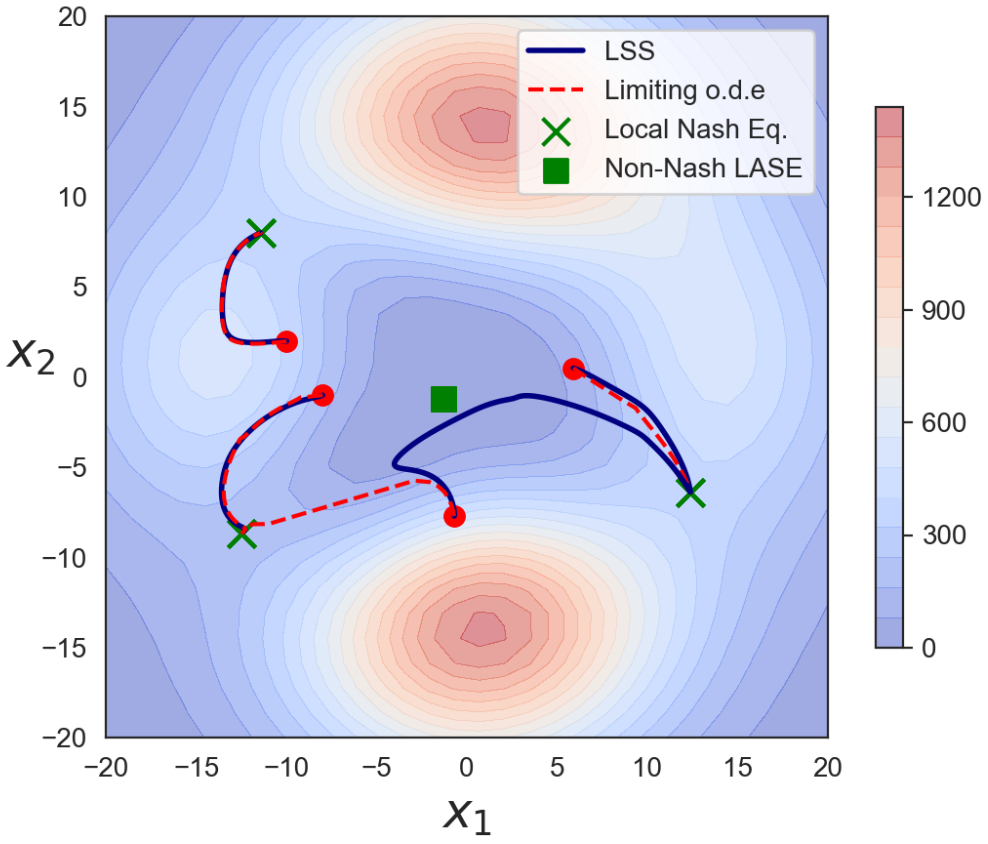


Figure 5.2: Behavior of the two-timescale approximation with respect to the limiting ODE. The two-timescale procedure closely tracks the limiting ODE from three of the four initializations, but converges to a different Nash equilibrium in the fourth due to approximation errors early in the process.

layers of 64 neurons each. The tanh activation function is used since it satisfies the smoothness assumptions for our functions. For the latent space, we use a 64-dimensional Gaussian with mean zero and covariance $\Sigma = 0.01 I_{64}$. The ground truth distribution is a mixture of 16 Gaussians used in [15] to test for mode collapse, each with covariance $\Sigma = 0.02 I_2$. In Figure 5.3, we show the generator after 30000 iterations of the various algorithms, initialized with the same weights and biases. We observe that LSS converges to the true distribution while the other algorithms have worse performance, showing how the adjusted dynamics can lead to convergence to better equilibria.

We caution that convergence rate per se is not necessarily a reasonable metric on which to compare performance in the GAN setting or in other game-theoretic settings. Indeed, competing algorithms may converge faster than our method when used to train GANs, but

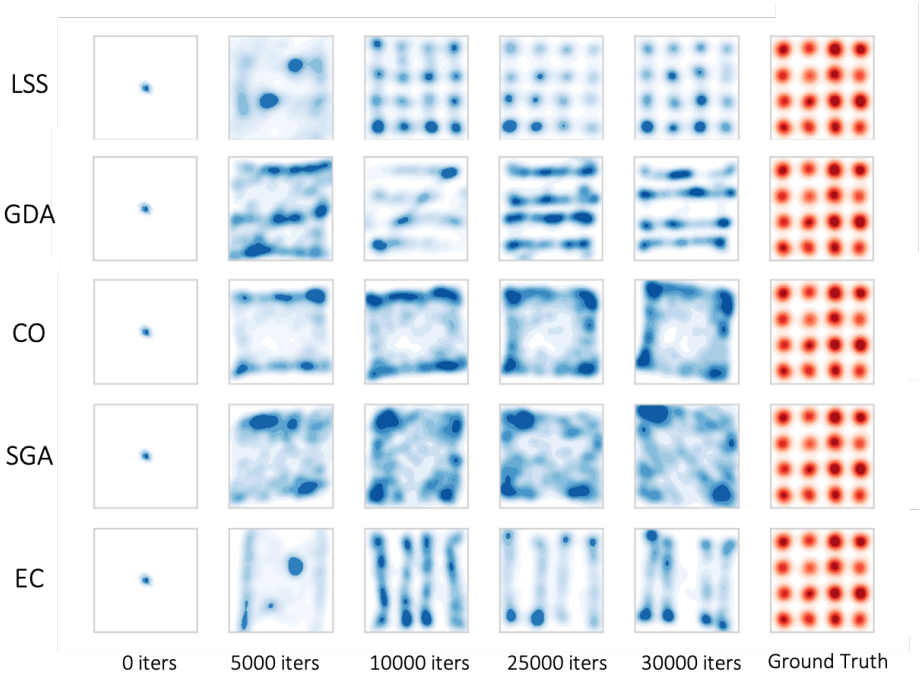


Figure 5.3: The output of a GAN trained with LSS, GDA, Consensus optimization (CO), the symplectic gradient adjustment (SGA), and Extreme curvature exploitation (EC) after 30000 iterations of training.

since the competitors could be converging quickly to a non-Nash equilibrium, which is not desirable. Indeed, the optimal solution is known to be a local Nash equilibrium for GANs [68, 132]. LSS may initially move towards a non-Nash equilibrium, while subsequently escaping the neighborhood of such points before converging. This will lead to a slower convergence rate, but a better quality solution.

In Figure 5.4 we show further numerical experiments that show the training of the same generative adversarial network described in Section 5.5 in wall-clock time, starting with the same initialization. This is to account for the fact that the different algorithms have different per-iteration complexities. In 300 seconds, LSS runs for around 12,000 iterations while GDA, symplectic gradient adjustment (SGA), and consensus optimization (CO), and extreme curvature (EC) complete 70,000, 40,000, 25,000, and 6000 iterations respectively. In Figure 5.4, we observe that LSS correctly recovers the ground truth distribution in around 200 seconds while GDA takes longer to converge. This suggests that LSS, despite having a slower per-iteration complexity may be faster at converging in neighborhoods of equilibria. Figure 5.4 also highlights the benefits of LSS over extreme curvature exploitation. LSS is able to perform twice the number of iterations in the same amount of time as extreme curvature exploitation completes, all the while successfully converging to the correct distribution.

In Figure 5.5 we see that LSS recovers the ground truth distribution, when GDA quickly converges to a suboptimal equilibrium. While the other algorithms are slow to converge in

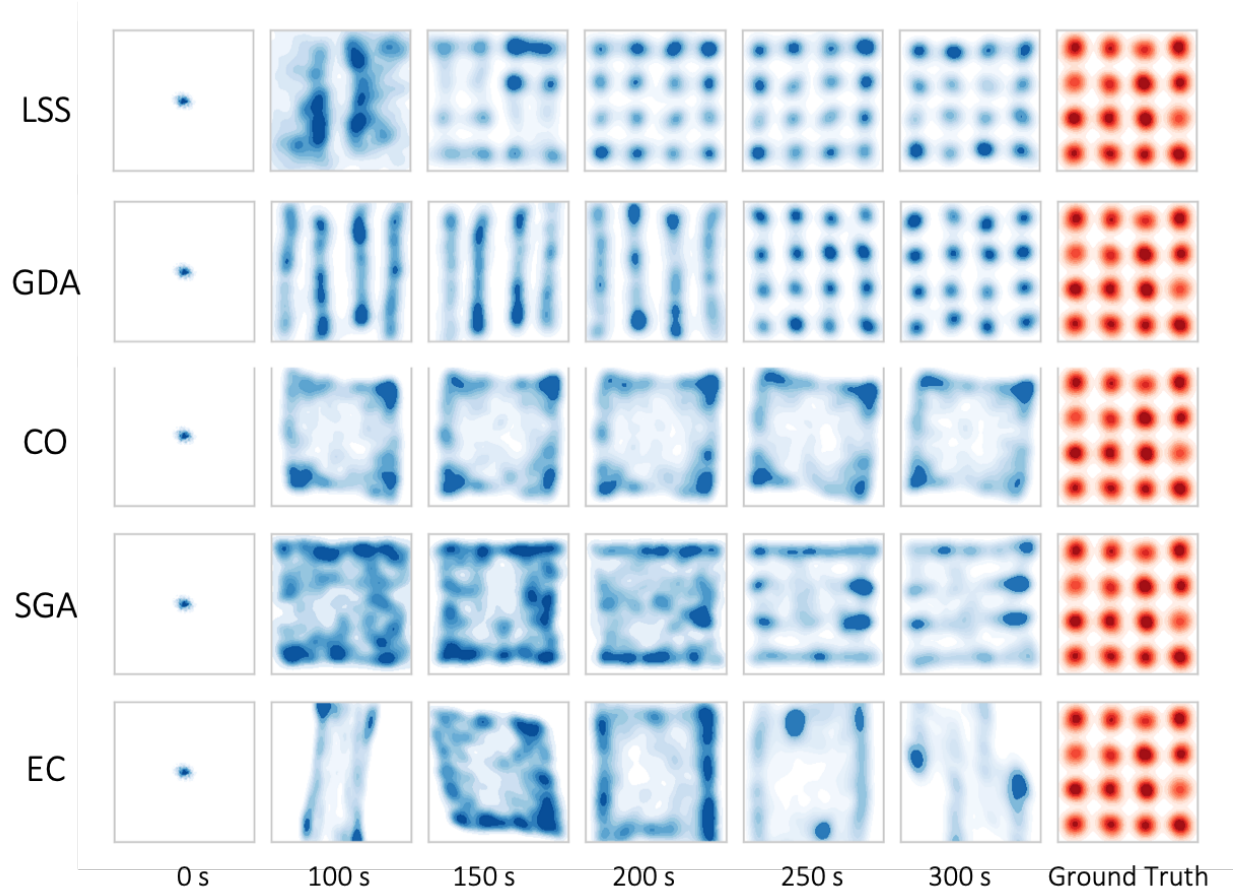


Figure 5.4: The output of the generator trained with A. LSS, B. GDA, C. extreme curvature exploitation, and D. consensus optimization over training in wall-clock time. We see that LSS converges to the correct distribution, while GDA converges quickly to an incorrect distribution. Extreme curvature exploitation is slow to converge in wall-clock time while consensus optimization fails to properly identify the individual Gaussians.

a similar amount of time (without carefully tuning hyper-parameters)

These experiments highlight the benefits of LSS over both GDA, the symplectic gradient adjustment, consensus optimization, and extreme curvature exploitation. The performance of GDA seems highly reliant on the initialization, and the algorithm can quickly converge to undesired equilibria and remain stuck there. LSS, on the other hand, is more computationally intensive to compute than GDA, but seems to consistently find better quality solutions in fewer iterations than the other benchmark algorithms. Extreme curvature exploitation moves towards better quality solutions, but seems prohibitively expensive to implement efficiently in complex settings.

For the training of the generative adversarial network, we randomly initialize the weights and biases using the standard TensorFlow initializations. For both the implementation of

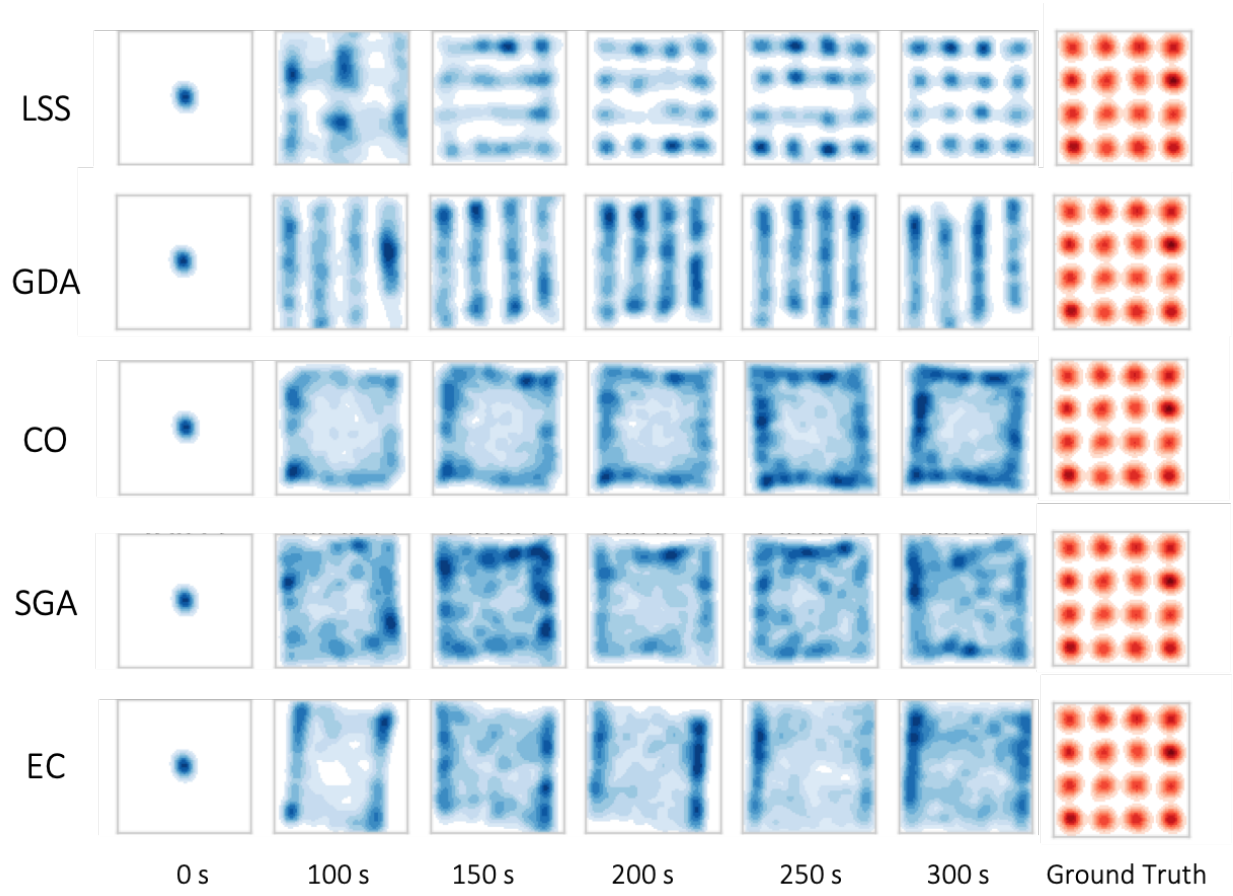


Figure 5.5: The output of the generator trained with A. LSS, B. GDA, C. extreme curvature exploitation, and D. consensus optimization over training in wall-clock time. We see that GDA converges quickly to the correct distribution faster than LSS. Extreme curvature exploitation is slow to converge in wall-clock time while consensus optimization fails to properly identify the individual Gaussians.

GDA and LSS used to generate Figures 5.3,5.4, and 5.5 we used the RMSProp optimizer with step size $2e-4$ for the x and y processes. For the v process in LSS, we used the RMSProp optimizer with step size $2e-5$. We note that these do not satisfy our assumptions on the step size sequences for the two-timescale process, but are meant to show that the approach still works in practical settings. Lastly, we use a batch size of 128. For consensus optimization we set the hyper-parameter to 1.0 and for curvature exploitation we performed 5 iterations of the power method per gradient step to find the suitable eigenvalue-eigenvector pairs and used a hyper-parameter of 1.0.

5.6 Time-varying adjustment

In this section we analyze a slightly different version of (5.4) that allows us to remove the assumption that $\omega(z)$ is never in the nullspace of $S(z)$. Though this assumption is relatively mild we can remove it entirely while retaining our theoretical guarantees. The new dynamics are constructed by adding a time-varying term to the dynamics that goes to zero only when $\omega(z)$ is zero, and it even serves the same purpose as the regularizer in the bilinear game example. This guarantees that the only critical points of the limiting dynamics are the critical points of ω. The analysis of these dynamics is slightly more involved and requires generalizations of the definition of a LASE to handle time-varying dynamics. We first define an equilibrium of a potentially time-varying dynamical system $\dot{\theta} = g(\theta, t)$ as a point θ^* such that $g(\theta^*, t) \equiv 0$ for all $t \geq 0$. We can now generalize the definition of a LASE to the time-varying setting.

Definition 15. *A strategy $\theta^* \in \mathbb{R}^d$ is a* locally uniformly asymptotically stable equilibrium *of the time-varying continuous time dynamics $\dot{\theta} = -f(\theta, t)$ if θ^* is an equilibrium of $\dot{\theta} = -f(\theta, t)$, and $D_\theta f(\theta^*, t) \equiv J(\theta^*)$ and $Re(\lambda) > 0$ for all $\lambda \in \text{spec}(J(z^*))$.*

Locally uniformly asymptotically stable equilibria, under this definition, also have the property that they are locally exponentially attracting under the flow, $\dot{\theta} = -f(\theta, t)$. Further, since the linearization around a locally uniformly asymptotically stable equilibrium is time-invariant, we can still invoke converse Lyapunov theorems like those presented in [168] when deriving the non-asymptotic bounds.

Having defined equilibria and a generalization of LASE for time-varying systems, we now introduce a time-varying version of the continuous-time ODE presented in Section 5.2 which allows us to remove the assumption that $\omega(z)$ is never in the nullspace of $S(z)$. The limiting ODE is given by:

$$\dot{z} = -h_{TV}(z, t) = -(h(z, t), \tag{5.11}$$

where $h(z, t)$ is given by:

$$h(z, t) = \omega(z) + J^{-1}(z) \left(J^T(z) + \lambda(z, t)I\right) \omega(z)$$

and $\lambda(z, t)$ satisfies:

- $0 \leq \lambda_1(z, t) \leq \xi_2$ for all $z \in \mathbb{R}^d$ and $t > 0$.
- $\lambda_1(z, t) = 0 \iff \omega(z) = 0$.
- $\omega(z) = 0 \implies D_z\lambda_1(z) = 0$,
- if $\omega(z) \neq 0$, $\lambda_1(z, t) \to 0$ as $t \to \infty$.

Thus we require that the time-varying adjustment term g_{TV} must be bounded and is equal to zero only when $\omega(z) = 0$. Most importantly, we require that for any z that is not a critical point of ω, g_{TV} must be changing in time. An example of a g_{TV} that satisfies these requirements is:

$$g_{TV}(z,t) = \xi_1 \left(1 - e^{-\xi_2 ||\omega(z)||^2}\right)/t, \tag{5.12}$$

These conditions, as the next theorem shows, allow us to guarantee that the only locally asymptotically stable equilibria are the differential Nash equilibria of the game.

Theorem 17. *Under Assumption 6 the continuous-time dynamical system $\dot{z} = -h_{TV}(z,t)$ satisfies:*

- *z is a locally uniformly asymptotically stable equilibrium of $\dot{z} = -h_{TV}(z,t) \iff$ z is a DNE of the game $\{(f,-f),\mathbb{R}^d\}$.*
- *If z is an equilibrium point of $\dot{z} = -h_{TV}(z,t)$, then the Jacobian of h_{TV} at z is time-invariant and has real eigenvalues.*

Proof. We first show that:

$$h_{TV}(z,t) \equiv 0 \ \ \forall t \geq 0 \iff \omega(z) = 0.$$

By construction $\omega(z) = 0 \implies h_{TV}(z,t) \equiv 0 \ \ \forall t \geq 0$. To show the converse, we assume that there exists a z such that $h_{TV}(z,t) \equiv 0 \ \ \forall \ \ t \geq 0$ but $\omega(z) \neq 0$. This implies that:

$$2S(z)\omega(z) = -\lambda(z,t)\omega(z) \quad \forall t \geq 0.$$

Since z is a constant and $\omega(z) \neq 0$, $\lambda(z,t) > 0$, and is changing for all t, this cannot hold since $\lambda(z,t)$ cannot be an eigenvalue of $S(z)$ for all t.

$$D_t u(t) = 0 \ \ \forall t \geq 0.$$

Thus, we have a contradiction and $h_{TV}(z,t) \equiv 0 \ \ \forall t \geq 0 \implies \omega(z) = 0$.

Having shown that the critical points of $\dot{z} = -h_{TV}(z,t)$ are the same as that of $\dot{z} = -\omega(z)$, we now note that the Jacobian of $h_{TV}(z,t)$, at critical points, must be $S(z)$. Under the same development as the proof of Theorem 13 the Jacobian of h_{TV} is given by:

$$J_{TV}(z) = J(z) + J^T(z) + (D_z\lambda(z,t)).$$

Again, by construction $D_z\lambda(z,t) = 0$ when $\omega(z) = 0$. The third term therefore disappears and we have that $J_{TV}(z) = S(z)$. The proof now follows from that of Theorem 13. □

5.7 Chapter Summary

We have introduced local symplectic surgery, a two-timescale algorithm for finding the local Nash equilibria of zero-sum continuous games. We have shown that the only hyperbolic critical points to which it can converge are the local Nash equilibria of the underlying game. This significantly improves upon previous efficient methods for finding such points which, as shown in Appendix 5.1, cannot give such guarantees.

We emphasize that our analysis has been limited to neighborhoods of equilibria; the proposed algorithm can converge in principle to limit cycles at other locations of the space. These are hard to rule out completely. Moreover, some of these limit cycles may actually have some game-theoretic relevance [21, 76]. Another limitation of our analysis is that we have assumed the existence of local Nash equilibria in games. Showing that they exist and finding them is very hard to do in general. Our algorithm will converge to local Nash equilibria, but may diverge when the game does not admit equilibria or when the algorithm does not approach any equilibria in its region of attraction. Thus, divergence of our algorithm is not a certificate that no equilibria exist. Such caveats, however, are the same as those for other gradient-based approaches for finding local Nash equilibria.

Another drawback to our approach is the use of second-order information. Though the two-timescale approximation does not need access to the full Jacobian of the gradient dynamics, the update does involve computing Jacobian-vector products. This is similar to other recently proposed approaches but will be inherently slower to compute than pure first- or zeroth-order methods. Bridging this gap while retaining similar theoretical properties remains an interesting avenue of further research. In summary, we have shown that some of the inherent flaws of gradient-based methods in zero-sum games can be overcome by designing algorithms to take advantage of structural aspects of the game-theoretic setting.

Part II

Decision-Making under Uncertainty

Chapter 6

Model-Based Approaches for Multi-Armed Bandits

Sequential decision making under uncertainty has become one of the fastest developing fields of machine learning, with applications from healthcare, to social-networks, and recommendation systems. A central theme in such problems is addressing exploration-exploitation tradeoffs [12, 94], wherein an algorithm must balance between exploiting its current knowledge and exploring previously unexplored options.

The classic stochastic multi-armed bandit problem has provided a theoretical laboratory for the study of exploration/exploitation tradeoffs [91]. A vast literature has emerged that provides algorithms, insights, and matching upper and lower bounds in many cases. The dominant paradigm in this literature has been that of *frequentist analysis*; cf. in particular the analyses devoted to the celebrated upper confidence bound (UCB) algorithm [12]. Interestingly, however, Thompson sampling, a Bayesian approach first introduced almost a century ago [180] has been shown to be competitive and sometimes outperform UCB algorithms in practice [36, 171]. Further, the fact that Thompson sampling, being a Bayesian method, explicitly makes use of prior information, has made it particularly popular in industrial applications [see, e.g., 166, and the references therein]— reflecting the need for explainable algorithms which can readily incorporate domain knowledge (i.e., model-based algorithms).

Although most theory in the bandit literature is focused on non-Bayesian methods, there is a smaller, but nontrivial, theory associated with Thompson sampling. In particular, Thompson sampling has been shown to achieve optimal risk bounds in multi-armed bandit settings with Bernoulli rewards and beta priors [8, 83], Gaussian rewards with Gaussian priors [8], one-dimensional exponential family models with uninformative priors [86], and finitely-supported priors and observations [69]. Thompson sampling has further been shown to asymptotically achieve optimal instance-independent performance [165].

Despite these appealing foundational results, the deployment of Thompson sampling in complex problems is often constrained by its use of samples from posterior distributions, which are often difficult to generate in regimes where the posteriors do not have closed

forms. A common solution to this has been to use *approximate* sampling techniques to generate samples from *approximations* of the posteriors [36, 66, 104, 166]. Such approaches have been demonstrated to work effectively in practice [160, 186], but it is unclear how to maintain performance over arbitrary time horizons while using approximate sampling. Indeed, to the best of our knowledge the strongest regret guarantees for Thompson sampling with approximate samples, due to Lu and Van Roy [104], require a model whose complexity grows with the time horizon to guarantee optimal performance. Further, it was recently shown theoretically by Phan, Yadkori, and Domke [150] that a naïve usage of approximate sampling algorithms with Thompson sampling can yield a drastic drop in performance.

As such, even in the well studied and relatively simple setting of multi-armed bandits, there is still a lack of understanding on how to design versatile model-based algorithms for sequential decision-making under uncertainty— a key ingredient in developing reliable algorithms for societal-systems.

Contributions In the following chapters we analyze Thompson sampling with approximate sampling methods in a class of multi-armed bandit algorithms where the rewards are unbounded, but their distributions are log-concave. In Chapter 7 we derive novel posterior contraction rates for posteriors when the rewards are generated from such distributions and under general assumptions on the priors. Using these rates, we show in Chapter 8 that Thompson sampling with samples from the true posterior achieves finite-time optimal frequentist regret. Further, the regret guarantee we derive has explicit constants and explicit dependencies on the dimension of the parameter spaces, variance of the reward distributions, and the quality of the prior distributions.

In Chapter 9 we present a simple counterexample demonstrating the relationship between the approximation error to the posterior and the resulting regret of the algorithm. Building on the insight provided by this example, we propose two approximate sampling schemes based on Langevin dynamics to generate samples from approximate posteriors and analyze their impact on the regret of Thompson sampling. We first analyze samples generated from the unadjusted Langevin algorithm (ULA) and specify the runtime, hyperparameters, and initialization required to achieve an approximation error which provably maintains the optimal regret guarantee of exact Thompson sampling over finite-time horizons. Crucially, we initialize the ULA algorithm from the approximate sample generated in the previous round to make use of the posterior concentration property and ensure that only a *constant* number of iterations are required to achieve the optimal regret guarantee. Under slightly stronger assumptions, we then demonstrate that a stochastic gradient variant called *stochastic gradient Langevin dynamics* (SGLD) requires only a *constant* batch size in addition to the constant number of iterations to achieve logarithmic regret. Since the computational complexity of this sampling algorithm does not scale with the time horizon, the proposed method is a true "anytime" algorithm. Finally, we conclude Chapter 9 by validating these theoretical results in numerical simulations where we find that Thompson sampling with our approximate sampling schemes maintain the desirable performance of exact Thompson sampling.

Our results suggest that the tailoring of approximate sampling algorithms to work with Thompson sampling can overcome the phenomenon studied in Phan, Yadkori, and Domke [150], where approximation error in the samples can yield linear regret. Indeed, our results suggest that it is possible for Thompson sampling to achieve order-optimal regret guarantees with an efficiently implementable approximate sampling algorithm.

6.1 Preliminaries

In this work we analyze Thompson sampling strategies for the K-armed stochastic multi-armed bandit (MAB) problem. In such problems, there is a set of K options, or "arms," $\mathcal{A} = \{1, ..., K\}$, from which a player must choose at each round $t = 1, 2, ...$ After choosing an arm $A_t \in \mathcal{A}$ in round t, the player receives a real-valued reward X_{A_t} drawn from a fixed yet unknown distribution associated with the arm, p_{A_t}. The random rewards obtained from playing an arm repeatedly are i.i.d. and independent of the rewards obtained from choosing other arms.

Throughout this paper, we assume that the reward distribution for each arm is a member of a parametric family parametrized by $\theta_a \in \mathbb{R}^{d_a}$, such that the true reward distribution is $p_a(X) = p_a(X; \theta_a^*)$, where θ_a^* is unknown. Moreover, we assume throughout this chapter that the parametric families are log-concave and Lipschitz smooth in θ_a:

Assumption 11 (Assumption on the family $p_a(X|\theta_a)$ around θ_a^*)**.** *Assume that* $\log p_a(x|\theta_a)$ *is* L_a*-smooth and* m_a*-strongly concave around* θ_a^* *for all* $X \in \mathbb{R}$*:*

$$
\begin{aligned}
-\log p_a(x|\theta_a^*) - \nabla_\theta \log p_a(x|\theta_a^*)^\top (\theta_a - \theta_a^*) + \frac{m_a}{2}\|\theta_a - \theta_a^*\|^2 \leq -\log p_a(x|\theta_a) \\
\leq -\log p_a(x|\theta_a^*) - \nabla_\theta \log p_a(x|\theta_a^*)^\top (\theta_a - \theta_a^*) + \frac{L_a}{2}\|\theta_a - \theta_a^*\|^2, \quad \forall \theta_a \in \mathbb{R}^{d_a}, x \in \mathbb{R}.
\end{aligned}
$$

Additionally we make assumptions on the true distribution of the rewards:

Assumption 12 (Assumption on true reward distribution $p_a(X|\theta_a^*)$)**.** *For every* $a \in \mathcal{A}$ *assume that* $p_a(X; \theta_a^*)$ *is strongly log-concave in* X *with some parameter* ν_a*, and that* $\nabla_\theta \log p_a(x|\theta_a^*)$ *is* L_a*-Lipschitz in* X*:*

$$
-\left(\nabla_x \log p_a(x|\theta_a^*) - \nabla_x \log p_a(x'|\theta_a^*)\right)^T (x - x') \geq \nu_a \|x - x'\|_2^2, \quad \forall x, x' \in \mathbb{R}.
$$

$$
\|\nabla_\theta \log p_a(x|\theta_a^*) - \nabla_\theta \log p_a(x'|\theta_a^*)\| \leq L_a \|x - x'\|_2, \quad \forall x, x' \in \mathbb{R}.
$$

Parameters ν_a and L_a provide lower and upper bounds to the sub- and super-Gaussianity of the true reward distributions. We further define $\kappa_a = \max\{L_a/m_a, L_a/\nu_a\}$ to be the condition number of the model class. Finally, we assume that for each arm $a \in \mathcal{A}$ there is a linear map such that for all $\theta_a \in \mathbb{R}^{d_a}$, $\mathbb{E}_{x \sim p_a(x|\theta_a)}[X] = \alpha_a^T \theta_a$, with $\|\alpha_a\| = A_a$.

We now review Thompson sampling, the pseudo-code for which is presented in Algorithm 2. A key advantage of Thompson sampling over frequentist algorithms for multi-armed bandit problems is its flexibility in incorporating prior information. In this paper, we assume that the prior distributions over the parameters of the arms have smooth log-concave densities:

Assumption 13 (Assumptions on the prior distribution)**.** *For every $a \in \mathcal{A}$ assume that $\log \pi_a(\theta_a)$ is concave with L_a-Lipschitz gradients for all $\theta_a \in \mathbb{R}^{d_a}$:*[1]

$$\|\nabla_\theta \pi_a(\theta_a) - \nabla_\theta \pi_a(\theta'_a)\| \leq L_a \|\theta_a - \theta'_a\|, \quad \forall \theta_a, \theta'_a \in \mathbb{R}^{d_a}.$$

We remark that we do *not* assume the prior is strongly log-concave, but that some structure is needed to prove finite-time concentration rates for the posterior. In particular, Assumption 13 allows us to prove the following proposition.

Proposition 16. *If the prior distribution over θ_a satisfies Assumption 13. Then:*

$$\sup_{\mathbb{R}^{d_a}} \nabla \log \pi_a(\theta_a)^T (\theta_a - \theta_a^*) \leq g_a^* - \log \pi_a(\theta_a^*)$$

Where $g_a^* = \max_{\theta \in \mathbb{R}^d} \log \pi_a(\theta_a)$.

Proof. Let $\log \pi_a(\theta_a) = g(\theta_a)$. From the concavity of g, we know that

$$\nabla g(\theta_a)^T (\theta_a - \theta_a^*) \leq g(\theta_a) - g(\theta_a^*)$$

Since this holds for all $\theta \in \mathbb{R}^{d_a}$, we take the supremum of both sides and get that:

$$\sup_{\mathbb{R}^{d_a}} \nabla g(\theta_a)^T (\theta_a - \theta_a^*) \leq g^* - g(\theta_a^*)$$

□

Let $B_a := g_a^* - \log \pi_a(\theta_a^*)$. Note that if the prior is centered on the correct value of θ_a^*, then $B_a = 0$. The parameter B_a is the prior dependence that appears in our concentration and regret bounds throughout the paper.

Given the priors and likelihoods, Thompson sampling proceeds by maintaining a posterior distribution over the parameters of each arm a at each round t. Given the likelihood family, $p(X|\theta_a)$, the prior, $\pi(\theta_a)$, and the n data samples from an arm a, $X_{a,1}, \cdots, X_{a,n}$, let $F_{n,a} : \mathbb{R}^{d_a} \to \mathbb{R}$ be $F_{n,a}(\theta_a) = \frac{1}{n} \sum_{i=1}^n \log p_a(X_{a,i}|\theta_a)$, be the average log-likelihood of the data. Then the posterior distribution over the parameter θ_a at round t, denoted $\mu_a^{(n)}$, satisfies:

$$p_a(\theta_a | X_{a,1}, \cdots, X_{a,n}) \propto \pi_a(\theta_a) \prod_{i=1}^t \left(p_a(X_t|\theta_a)\right)^{\mathbb{1}\{A_t = a\}} = \exp\left(n F_{n,a}(\theta_a) + \log \pi(\theta_a)\right).$$

[1] We remark that the Lipschitz constants are all assumed to be the same to simplify notation.

Algorithm 2 Thompson sampling

Input : Priors π_a for $a \in \mathcal{A}$, posterior scaling parameter γ_a

Set $\mu_{a,t} = \pi_a$ for $a \in \mathcal{A}$ **for** $t = 0, 1, \cdots$ **do**

  Sample $\theta_{a,t} \sim \mu_a^{(T_a(t))}[\gamma_a]$

  Choose action $A_t = \text{argmax}_{a \in \mathcal{A}} \, \alpha_a^T \theta_{a,t}$.

  Receive reward X_{A_t}.

  Update (approximate) posterior distribution for arm A_t: $\mu_a^{(T_a(t+1))}$.

For any $\gamma_a > 0$ we denote the scaled posterior[2] as $\mu_a^{(n)}[\gamma_a]$, whose density is proportional to:

$$\exp\left(\gamma_a(nF_{n,a}(\theta_a) + \log \pi(\theta_a))\right). \tag{6.1}$$

Letting $T_a(t)$ be the number of samples received from arm a after t rounds, a Thompson sampling algorithm, at each round t, first samples the parameters of each arm a from their (scaled) posterior distributions: $\theta_{a,t} \sim \mu_a^{(T_a(t))}[\gamma_a]$ and then chooses the arm for which the sample has the highest value:

$$A_t = \underset{a \in \mathcal{A}}{\text{argmax}} \, \alpha_a^T \theta_{a,t}.$$

A player's objective in MAB problems is to maximize her cumulative reward over any fixed time horizon T. The measure of performance most commonly used in the MAB literature is known as the *expected regret* $R(T)$, which corresponds to the expected difference between the accrued reward and the reward that would have been accrued had the learner selected the action with the highest mean reward during all steps $t = 1, \cdots, T$.[3] Recalling that $\bar{r}_a$ is the mean reward for arm $a \in \mathcal{A}$, the regret is given by:

$$R(T) := \mathbb{E}\left[\sum_{t=1}^{T} \bar{r}_{a^*} - \bar{r}_{A_t}\right],$$

where $\bar{r}_{a^*} = \max_{a \in \mathcal{A}} \bar{r}_a$. Without loss of generality, we assume throughout this chapter that the optimal arm, $a^* = \text{argmax}_{a \in \mathcal{A}} \, \bar{r}_a$, is arm 1. Further, we assume that the optimal arm is unique[4]: $\bar{r}_1 > \bar{r}_a$ for $a > 1$.

Traditional treatment of Thompson sampling algorithms often overlooks one of its most critical aspects: ensuring compatibility between the mechanism that produces samples from the posterior distributions and the algorithm's regret guarantees. This issue is usually addressed by assuming that the prior distributions and the reward distributions are conjugate

[2] In Chapter 8 we explain the use of scaled posteriors is required to obtain optimal regret guarantees for our bandit algorithms.

[3] We remark that the analysis of Thompson sampling has often been focused on a different quantity known as the Bayes regret, which is simply the expectation of $R(T)$ over the priors: $\mathbb{E}_\pi[R(T)]$. However, in an effort to demonstrate that Thompson sampling is an effective alternative to frequentist methods like UCB, we analyze the frequentist regret $R(T)$.

[4] We introduce this assumption merely for the purpose of simplifying our analysis.

pairs. Although this approach is simple and prevalent in the literature [see, e.g., 166], it fails to capture more complex distributional families for which this assumption may not hold. Indeed, it was recently shown in Phan, Yadkori, and Domke [150] that if the samples come from distributions that approximate the posteriors with a constant error, the regret may grow at a linear rate. A more nuanced understanding of the relationship between the quality of the samples and the regret of the algorithms is, however, still lacking.

Notation

Before summarizing the contents of Part II, we first give an overview of all of the notation used throughout this part of the dissertation.

Symbol	Meaning
$\mathcal{A}$	set of arms in bandit environment
K	number of arms in the bandit environment $\lvert\mathcal{A}\rvert$
T	Time horizon
A_t	arm pulled at time t by the algorithm $A_t \in \mathcal{A}$
$T_a(t)$	number of times arm a has been pulled by time t
X_{A_t}	reward from choosing arm A_t at time t
θ_a	parameters of likelihood functions such that, $\theta_a \in \mathbb{R}^{d_a}$
d_a	dimension of parameter space for arm a
$p_a(x\lvert\theta_a)$	parametric family of reward distributions for arm a
$\pi_a(\theta_a)$	prior distribution over the parameters for arm a
$\mu_a^{(n)}$	probability measure associated with the posterior over the parameters of arm a after n samples from arm a
$\mu_a^{(n)}[\gamma_a]$	probability measure associated with the (scaled) posterior over the parameters of arm a after n samples from arm a
$\widehat{\mu}_a^{(n)}$	probability measure resulting from running the Langevin MCMC algorithm described in Algorithm 3 which approximates $\mu_a^{(n)}$
$\bar{\mu}_a^{(n)}[\gamma_a]$	probability measure resulting from an approximate sampling method which approximates $\mu_a^{(n)}[\gamma_a]$
θ_a^*	true parameter value for arm a
$\theta_{a,t}$	sampled parameter for arm a at time t of the Thompson Sampling algorithm: $\theta_{a,t} \sim \mu_a^{(n)}$
$\bar{r}_a$	mean of the reward distribution for arm a: $\bar{r}_a = \mathbb{E}[X_a\lvert\theta_a^*]$
α_a	vector in $\mathbb{R}^{d_a}$ such that $\bar{r}_a = \alpha_a^T\theta_a^*$
$r_{a,t}(T_a(t))$	estimate of mean of arm a at round t: $r_{a,t}(T_a(t)) = \alpha_a^T\theta_{a,t}$
A_a	norm of α_a
m_a	Strong log-concavity parameter of the family $p_a(x;\theta)$ in θ for all x.
ν_a	Strong log-concavity parameter of the true reward distribution $p_a(x;\theta^*)$ in x.
$F_{n,a}(\theta_a)$	Averaged log likelihood over the data points: $F_{n,a}(\theta_a) = \frac{1}{n}\sum_{i=1}^n \log p_a(X_i,\theta_a)$
L_a	Lipschitz constant for the true reward distribution and likelihood families $p_a(x;\theta^*)$ in x.
κ_a	condition number of the likelihood family $\kappa_a = \max\left(\frac{L_a}{m_a}, \frac{L_a}{\nu_a}\right)$.
B_a	reflects the quality of the prior: $B_a = \frac{\max_\theta \pi_a(\theta)}{\pi_a(\theta^*)}$

6.2 Overview of Part II

Part II of this dissertation focuses on the analysis of Thompson sampling with approximate sampling methods in a class of multi-armed bandit algorithms where the rewards are unbounded, but their distributions are log-concave. In Chapter 7 we derive new posterior contraction rates for posteriors when the rewards are generated from such distributions and under general assumptions on the priors. Using these rates, we show in Chapter 8 that Thompson sampling with samples from the true posterior achieves finite-time optimal frequentist regret. Further, the regret guarantee we derive has explicit constants and explicit dependencies on the dimension of the parameter spaces, variance of the reward distributions, and the quality of the prior distributions.

In Chapter 9 we present a simple counter-example demonstrating the relationship between the approximation error of the approximation to the posterior and the resulting regret of the algorithm. Building on the insight provided by this example, we analyze two approximate sampling schemes and their impact on the regret of Thompson sampling. We first analyze the regret of Thompson sampling when the samples are generated from the unadjusted Langevin algorithm (ULA). We specify the runtime and hyperparameters needed to achieve an approximation error which provably maintains the optimal regret guarantee of the exact algorithm over finite-time horizons. Under slightly stronger assumptions, we then show that a stochastic gradient variant of ULA also achieves logarithmic regret. Notably both the iteration complexity and the sample complexity for stochastic gradient estimate do not scale with the time horizon, yielding a true 'anytime' algorithm.

Our results suggest that the tailoring of approximate sampling algorithms to work with Thompson sampling can overcome the phenomenon studied in [150] where approximation error in the samples results in linear regret. Indeed, our results suggest that it is possible for Thompson sampling to achieve order-optimal regret guarantees with approximate sampling.

Chapter 7

Posterior Concentration Results

Core to the analysis of Thompson sampling is understanding the behavior of the posterior distributions over the parameters of the arms' distributions as the algorithm progresses and samples from the arms are collected. In particular, characterizing how the posterior density over the parameters given data,

$$P(\theta|X_1, ..., X_n),$$

concentrates around the true parameter value, is an essential step in developing a deeper understanding of Thompson Sampling.

The literature on understanding how posteriors evolve as data is collected goes back to Doob [47] and his proof of the asymptotic normality of posteriors. More recently, there has been a line of work [see, e.g., 60, 187] that derives rates of convergence of posteriors in various regimes, mostly following the framework first developed in Ghosal, Ghosh, and Vaart [59] for finite- and infinite-dimensional models. Such results, though quite general, do not have explicit constants or forms which make them amenable for use in analyzing bandit algorithms. Indeed, finite-time rates remain an active area of research but have been developed using information-theoretic arguments [172], and more recently through the analysis of stochastic differential equations [130], though in both cases the assumptions, burn-in times, and lack of precise constants make them difficult to integrate with the analysis of Thompson sampling. Due to this, Thompson sampling has, for the most part, been only well understood for conjugate prior/likelihood families like beta/Bernoulli and Gaussian/Gaussian [8], or in more generality in well-behaved families such as one-dimensional exponential families with uninformative priors [86] or finitely supported prior/likelihood pairs [69].

In this Chapter we derive posterior concentration rates for parameters in d-dimensions and for a large class of priors and likelihoods by analyzing the moments of a stochastic differential equation for which the posterior is the limiting distribution. Our results expand upon the recent derivation of novel contraction rates for posterior distributions presented in Mou et al. [130] to hold for a finite number of samples and may be of independent interest. We make use of these concentration results to show that Thompson sampling with such priors and likelihoods results in order-optimal regret guarantees.

7.1 Posterior Concentration in Log-Concave Families

To begin, we note that classic results [139] guarantee that, as $t \to \infty$ the distribution P_t of θ_t which evolves according to:

$$d\theta_t = \frac{1}{2}\nabla_\theta F_{n,a}(\theta_t)dt + \frac{1}{2n}\nabla_\theta \log \pi_a(\theta_t))dt + \frac{1}{\sqrt{n\gamma_a}}dB_t, \tag{7.1}$$

is given by:

$$\lim_{t\to\infty} P_t(\theta|X_1, ..., X_n) \propto \exp(-\gamma_a \left(nF_{n,a}(\theta) + \log\ \pi_a(\theta)\right)),$$

almost surely. Comparing with Eq. (6.1), this limiting distribution is the scaled posterior distribution $\mu_a^{(n)}[\gamma_a]$. Thus, by analyzing the limiting properties of θ_t as it evolves according to the stochastic differential equation, we can derive properties of the scaled posterior distribution.

To do so, we first show that with high probability the gradient of $F_{n,a}(\theta^*)$ concentrates around zero (given the data $X_1, ..., X_n$). More precisely we show using well known results on the concentration of Lipschitz functions of strongly log-concave random variables that $\nabla_\theta F_{a,n}(\theta_a^*)$ has sub-Gaussian tails:

Proposition 17. *The random variable $\|\nabla_\theta F_{a,n}(\theta_a^*)\|$ is $L_a\sqrt{\frac{d_a}{n\nu_a}}$-sub-Gaussian.*

The proof follows from the concentration of Lipschitz functions under strongly log-concave densities and relies on Assumption 12.

Proof. Recall that the true density $p_a(x|\theta_a^*)$ is ν_a-strongly log-concave in x and that $\nabla_\theta \log p_a(x|\theta_a^*)$ is L_a-Lipschitz in x. Notice that $\nabla_\theta F_a(\theta_a^*) = 0$ since θ_a^* is the point maximizing the population likelihood.

Let's consider the random variable $Z = \nabla_\theta \log p_a(x|\theta_a^*)$. Since $\mathbb{E}[Z] = \nabla_\theta F_a(\theta_a^*)$, the random variable Z is centered.

We start by showing Z is a subgaussian random vector. Let $v \in \mathbb{S}_{d_a}$ be an arbitrary point in the d_a-dimensional sphere and define the function $V : \mathbb{R}^{d_a} \to \mathbb{R}$ as $V(x) = \langle \nabla_\theta \log p_a(x|\theta_a^*), v\rangle$. This function is L_a-Lipschitz. Indeed let $x_1, x_2 \in \mathbb{R}^{d_a}$ be two arbitrary points in $\mathbb{R}^{d_a}$:

$$\begin{aligned}
|V(x_1) - V(x_2)| &= |\langle \nabla_\theta \log p_a(x_1|\theta_a^*) - \nabla_\theta \log p_a(x_2|\theta_a^*), v\rangle| \\
&\leq \|\nabla_\theta \log p_a(x_1|\theta_a^*) - \nabla_\theta \log p_a(x_2|\theta_a^*)\|_2 \|v\|_2 \\
&= \|\nabla_\theta \log p_a(x_1|\theta_a^*) - \nabla_\theta \log p_a(x_2|\theta_a^*)\|_2 \\
&\leq L_a\|x_1 - x_2\|
\end{aligned}$$

The first inequality follows by Cauchy-Schwartz, the second inequality by the Lipschitz assumption on the gradients. After a simple application of Proposition 2.18 in Ledoux [96], we conclude that $V(x)$ is subgaussian with parameter $\frac{L_a}{\sqrt{\nu_a}}$.

Since the projection of Z onto an arbitrary direction v of the unit sphere is subgaussian, with a parameter independent of v, we conclude the random vector Z is subgaussian with the same parameter $\frac{L_a}{\sqrt{\nu_a}}$. Consequently, the vector $\nabla_\theta F_{a,n}(\theta_a^*)$, being an average of n i.i.d. subgaussian vectors with parameter $\frac{L_a}{\sqrt{\nu_a}}$ is also subgaussian with parameter $\frac{L_a}{\sqrt{n\nu_a}}$.

Since $\nabla_\theta F_{a,n}(\theta_a^*)$ is a subgaussian vector with parameter $\frac{L_a}{\sqrt{n\nu_a}}$, Lemma 1 of [79] implies it is norm subgaussian with parameter $\frac{L_a\sqrt{d_a}}{\sqrt{n\nu_a}}$.

□

Conditioning on this high-probability event, we then analyze how the potential function:

$$V(\theta_t) = \frac{1}{2}e^{\alpha t}||\theta_t - \theta^*||_2^2,$$

evolves along trajectories of the stochastic differential equation, where $\alpha > 0$. By bounding the supremum of $V(\theta_t)$, we construct bounds on the higher moments of the random variable $||\theta - \theta^*||$. These moment bounds translate directly into the posterior concentration bound of $\theta \sim \mu_a^{(n)}$ around θ^* presented in the following theorem.

Theorem 18. *Suppose that Assumptions 11-13 hold, then for $\delta \in (0, e^{-1/2})$:*

$$\mathbb{P}_{\theta\sim\mu_a^{(n)}[\gamma_a]}\left(\|\theta_a - \theta_a^*\|_2 > \sqrt{\frac{2e}{m_a n}\left(\frac{d_a}{\gamma_a} + \log B_a + \left(\frac{32}{\gamma_a} + 8d_a\kappa_a^2\right)\log\left(1/\delta\right)\right)}\right) < \delta,$$

where $B_a = \max_{\theta\in\mathbb{R}^d}\frac{\pi_a(\theta)}{\pi_a(\theta_a^*)}$.

Theorem 18 guarantees that the scaled posterior distribution over the parameters of the arms concentrate at rate $\frac{1}{\sqrt{n}}$, where n is the number of times the arm has been pulled.

Proof. The proof makes use of the techniques used to prove Theorem 1 in Mou et al. [130]: analyzing how a carefully designed potential function evolves along trajectories of the s.d.e. By a careful accounting of terms and constants, however, we are able to keep explicit constants and derive tighter bounds which hold for any finite number of samples. Throughout the proof we drop the dependence on a and condition on the high-probability event, $G_{a,n}(\delta_1)$, defined in Proposition 17, which guarantees that the norm of the likelihood gradients concentrates with probability at least $1 - \delta_1$.

Consider the s.d.e.:

$$d\theta_t = \frac{1}{2}\nabla_\theta F_n(\theta_t)dt + \frac{1}{2n}\nabla_\theta \log\pi(\theta_t))dt + \frac{1}{\sqrt{n\gamma}}dB_t,$$

and the potential function given by:

$$V(\theta) = \frac{1}{2}e^{\alpha t}\|\theta - \theta^*\|_2^2,$$

for a choice of $\alpha > 0$. The idea is that bounds on the p-th moments of $V(\theta_t)$ can be translated into bounds on the p-th moments of $V(\theta)$ where $\theta \sim \mu^{(n)}$, due to the fact that $\lim_{t\to\infty} \theta_t = \theta \sim \mu^{(n)}$. The square-root growth in p of these moments will imply that $\|\theta - \theta^*\|_2$ has subgaussian tails with a rate that we make explicit.

We begin by using Ito's Lemma on $V(\theta_t)$:

$$V(\theta_t) = T1 + T2 + T3 + T4,$$

where:

$$\begin{aligned}
T1 &= -\frac{1}{2}\int_0^t e^{\alpha s}\langle \theta^* - \theta_s, \nabla_\theta F_n(\theta_s)\rangle ds + \frac{\alpha}{2}\int_0^t e^{\alpha s}\|\theta_s - \theta^*\|_2^2 ds \\
T2 &= \frac{1}{2n}\int_0^t e^{\alpha s}\langle \theta_s - \theta^*, \nabla_\theta \log \pi(\theta_s)\rangle ds \\
T3 &= \frac{d}{2n\gamma}\int_0^t e^{\alpha s} ds \\
T4 &= \frac{1}{\sqrt{n\gamma}}\int_0^t e^{\alpha s}\langle \theta_s - \theta^*, dB_s\rangle.
\end{aligned}$$

Let us first upper bound $T1$:

$$\begin{aligned}
T1 &= -\frac{1}{2}\int_0^t e^{\alpha s}\langle \theta^* - \theta_s, \nabla_\theta F_n(\theta_s)\rangle ds + \frac{\alpha}{2}\int_0^t e^{\alpha s}\|\theta_s - \theta^*\|_2^2 ds \\
&= -\frac{1}{2}\int_0^t e^{\alpha s}\langle \theta^* - \theta_s, \nabla_\theta F_n(\theta_s) - \nabla_\theta F_n(\theta^*)\rangle ds + \frac{\alpha}{2}\int_0^t e^{\alpha s}\|\theta_s - \theta^*\|_2^2 ds \\
&\quad - \frac{1}{2}\int_0^t e^{\alpha s}\langle \theta^* - \theta_s, \nabla_\theta F_n(\theta^*)\rangle ds \\
&\overset{(i)}{\leq} \frac{\alpha - m}{2}\int_0^t e^{\alpha s}\|\theta_s - \theta^*\|_2^2 ds - \frac{1}{2}\int_0^t e^{\alpha s}\langle \theta^* - \theta_s, \nabla_\theta F_n(\theta^*)\rangle ds \\
&\overset{(ii)}{\leq} \frac{\alpha - m}{2}\int_0^t e^{\alpha s}\|\theta_s - \theta^*\|_2^2 ds + \frac{1}{2}\int_0^t e^{\alpha s}\|\theta^* - \theta_s\| \underbrace{\|\nabla_\theta F_n(\theta^*)\|}_{:=\epsilon(n)} ds,
\end{aligned}$$

where in (i) we use the strong-concavity property from Assumption 11, and in (ii) we use Cauchy-Shwartz.

Using Young's inequality for products, where the constant is m, gives:

$$T1 \leq \frac{2\alpha - m}{4}\int_0^t e^{\alpha s}\|\theta_s - \theta^*\|_2^2 ds + \frac{\epsilon(n)^2}{4m}\int_0^t e^{\alpha s} ds.$$

Finally, choosing $\alpha = m/2$ the first term on the RHS vanishes. Evaluating the integral in the second term on the RHS gives:

$$T1 \leq \frac{\epsilon(n)^2}{2m^2}\left(e^{\alpha t} - 1\right) \leq \frac{\epsilon(n)^2}{m^2} e^{\alpha t}.$$

Given our assumption on the prior, our choice of $\alpha = \frac{m}{2}$ and simple algebra, we can upper bound $T2$ and $T3$ as:

$$T2 = \frac{1}{2n}\int_0^t e^{\alpha s}\langle \theta_s - \theta^*, \nabla_\theta \log \pi(\theta_s)\rangle ds \leq \frac{\log B}{2\alpha n}(e^{\alpha t} - 1) \leq \frac{\log B}{nm}e^{\alpha t}$$

$$T3 = \frac{d}{2n\gamma}\int_0^t e^{\alpha s} ds \leq \frac{d}{\gamma n m}e^{\alpha t}.$$

We proceed to bound $T4$. Let's start by defining:

$$M_t = \int_0^t e^{\alpha s}\langle \theta_s - \theta^*, dB_s\rangle,$$

so that:

$$T4 = \frac{M_t}{\sqrt{n\gamma}}.$$

Combining all the upper bounds of $T1, T2, T3$, and $T4$ we have that:

$$V(\theta_t) \leq \left(\frac{\epsilon(n)^2}{m^2} + \frac{d}{\gamma n m} + \frac{\log B}{nm}\right)e^{\alpha t} + \frac{M_t}{\sqrt{\gamma n}}.$$

To find a bound for the $p-$th moments of V, we upper bound the p-th moments of the supremum of M_t where $p \geq 1$:

$$\begin{aligned}
\mathbb{E}\left[\sup_{0\leq t\leq T}|M_t|^p\right] &\overset{(i)}{\leq} (8p)^{\frac{p}{2}}\mathbb{E}\left[\langle M, M\rangle_T^{\frac{p}{2}}\right] \\
&= (8p)^{\frac{p}{2}}\mathbb{E}\left[\left(\int_0^T e^{2\alpha s}\|\theta_s - \theta^*\|_2^2\right)^{\frac{p}{2}} ds\right] \\
&\overset{(ii)}{\leq} (8p)^{\frac{p}{2}}\mathbb{E}\left[\left(\sup_{0\leq t\leq T} e^{\alpha t}\|\theta_t - \theta^*\|_2^2 \int_0^T e^{\alpha s} ds\right)^{\frac{p}{2}}\right] \\
&= (8p)^{\frac{p}{2}}\mathbb{E}\left[\left(\sup_{0\leq t\leq T} e^{\alpha t}\|\theta_t - \theta^*\|_2^2 \frac{(e^{\alpha T} - 1)}{\alpha}\right)^{\frac{p}{2}}\right] \\
&\overset{(iii)}{\leq} \left(\frac{8pe^{\alpha T}}{\alpha}\right)^{\frac{p}{2}}\mathbb{E}\left[\left(\sup_{0\leq t\leq T} e^{\alpha t}\|\theta_t - \theta^*\|_2^2\right)^{\frac{p}{2}}\right].
\end{aligned}$$

Inequality (i) is a direct consequence of the Burkholder-Gundy-Davis inequality [159], (ii) follows by pulling out the supremum out of the integral, (iii) holds because $e^{\alpha T} - 1 \leq e^{\alpha T}$.

Now, let us consider the moments of $V(\theta_t)$:

$$\begin{aligned}
\mathbb{E}\left[\left(\sup_{0\leq t\leq T} V(\theta_t)\right)^p\right]^{\frac{1}{p}} &\leq \mathbb{E}\left[\left(\sup_{0\leq t\leq T}\left(\frac{\epsilon(n)^2}{m^2} + \frac{d}{\gamma n m} + \frac{\log B}{nm}\right)e^{\alpha t} + \frac{|M_t|}{\sqrt{\gamma n}}\right)^p\right]^{\frac{1}{p}} \\
&\leq \mathbb{E}\left[\left(\sup_{0\leq t\leq T}\left(\frac{\epsilon(n)^2}{m^2} + \frac{d}{\gamma n m} + \frac{\log B}{nm}\right)e^{\alpha t} + \sup_{0\leq t\leq T}\frac{|M_t|}{\sqrt{\gamma n}}\right)^p\right]^{\frac{1}{p}}
\end{aligned}$$

Via the Minkowski Inequality, and the fact $\epsilon(n)$ is independent of t, we can expand the above as:

$$\mathbb{E}\left[\left(\sup_{0\leq t\leq T} V(\theta_t)\right)^p\right]^{\frac{1}{p}} \leq \underbrace{\left(\frac{d}{\gamma nm}+\frac{\log B}{nm}\right)e^{\alpha T}}_{:=U_T} + \frac{e^{\alpha T}}{m^2}\mathbb{E}\left[\epsilon(n)^{2p}\right]^{\frac{1}{p}} + \mathbb{E}\left[\left(\sup_{0\leq t\leq T}\frac{|M_t|}{\sqrt{n}}\right)^p\right]^{\frac{1}{p}}$$

Since, from Proposition 17, we know that $\epsilon(n)$ is a $L\sqrt{\frac{d}{n\nu}}$-sub-Gaussian vector, we know that:

$$\mathbb{E}\left[\epsilon(n)^{2p}\right]^{\frac{1}{p}} \leq \left(2L\sqrt{\frac{2dp}{n\nu}}\right)^2$$

Using our upper bound on the supremum of M_t gives:

$$\mathbb{E}\left[\left(\sup_{0\leq t\leq T} V(\theta_t)\right)^p\right]^{\frac{1}{p}} \leq U_T + \frac{e^{\alpha T}8dL^2}{\nu m^2 n}p + \mathbb{E}\left[\left(\frac{8pe^{\alpha T}}{\gamma\alpha n}\right)^{\frac{p}{2}}\left(\sup_{0\leq t\leq T} e^{\alpha t}\|\theta_t-\theta^*\|_2^2\right)^{\frac{p}{2}}\right]^{\frac{1}{p}} \quad (7.2)$$

We proceed by bounding the second term on the RHS of the expression above:

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{8pe^{\alpha T}}{\alpha n}\right)^{\frac{p}{2}}\left(\sup_{0\leq t\leq T} e^{\alpha t}\|\theta_t-\theta^*\|_2^2\right)^{\frac{p}{2}}\right]^{\frac{1}{p}} &\overset{(i)}{\leq} \mathbb{E}\left[\frac{2^{p-1}}{2}\left(\frac{8pe^{\alpha T}}{\alpha\gamma n}\right)^p + \frac{1}{2^p}\left(\sup_{0\leq t\leq T} e^{\alpha t}\|\theta_t-\theta^*\|_2^2\right)^p\right]^{\frac{1}{p}} \\
&\overset{(ii)}{\leq} 2^{\frac{p-2}{p}}\mathbb{E}\left[\left(\frac{8pe^{\alpha T}}{\alpha\gamma n}\right)^p\right]^{\frac{1}{p}} + \frac{1}{2}\mathbb{E}\left[\left(\sup_{0\leq t\leq T} e^{\alpha t}\|\theta_t-\theta^*\|_2^2\right)^p\right]^{\frac{1}{p}} \\
&\overset{(iii)}{\leq} 16\ \mathbb{E}\left[\left(\frac{pe^{\alpha T}}{\alpha\gamma n}\right)^p\right]^{\frac{1}{p}} + \underbrace{\frac{1}{2}\mathbb{E}\left[\left(\sup_{0\leq t\leq T} e^{\alpha t}\|\theta_t-\theta^*\|_2^2\right)^p\right]^{\frac{1}{p}}}_{I}
\end{aligned}$$

Inequality (i) follows from using Young's inequality for products on the term inside the expectation with constant 2^{p-1}, inequality (ii) is a consequence of the Minkowski inequality and (iii) because $2^{\frac{p-2}{p}} \leq 2$. We note now that the second term I on the right hand side above is exactly:

$$\frac{1}{2}\mathbb{E}\left[\left(\sup_{0\leq t\leq T} V(\theta_t)\right)^p\right]^{\frac{1}{p}}$$

Plugging this into Equation 7.2 and rearranging gives:

$$\frac{1}{2}\mathbb{E}\left[\left(\sup_{0\leq t\leq T} V(\theta_t)\right)^p\right]^{\frac{1}{p}} \leq U_T + \frac{16e^{\alpha T}}{\alpha\gamma n}p + \frac{e^{\alpha T}8dL^2}{\nu m^2 n}p,$$

which finally results in:

$$\mathbb{E}\left[\left(\sup_{0\leq t\leq T} V(\theta_t)\right)^p\right]^{\frac{1}{p}} \leq \frac{2}{mn}\left(\frac{d}{\gamma} + \log B + \left(\frac{32}{\gamma} + \frac{8dL^2}{\nu m}\right)p\right)e^{\alpha T}. \tag{7.3}$$

Given this control on the moments of the supremum of $V(\theta_t)$ (recall $V(\theta) = \frac{1}{2}e^{\alpha t}\|\theta-\theta^*\|_2^2$), we finally construct the bound on the moments of $\|\theta_T - \theta^*\|$:

$$\begin{aligned}
\mathbb{E}[\|\theta_T - \theta^*\|^p]^{\frac{1}{p}} &= \mathbb{E}\left[e^{-\frac{p\alpha T}{2}} V(\theta_T)^{\frac{p}{2}}\right]^{\frac{1}{p}} \\
&\overset{(i)}{\leq} \mathbb{E}\left[e^{-\frac{p\alpha T}{2}}\left(\sup_{0\leq t\leq T} V(\theta_t)\right)^{\frac{p}{2}}\right]^{\frac{1}{p}} \\
&= e^{-\frac{\alpha T}{2}}\left(\mathbb{E}\left[\left(\sup_{0\leq t\leq T} V(\theta_t)\right)^{\frac{p}{2}}\right]^{\frac{2}{p}}\right)^{\frac{1}{2}} \\
&\overset{(ii)}{\leq} e^{-\frac{\alpha T}{2}}\left(\frac{2}{mn}\left(\frac{d}{\gamma} + \log B + \left(\frac{16}{\gamma} + \frac{4dL^2}{\nu m}\right)p\right)e^{\alpha T}\right)^{\frac{1}{2}} \\
&= \sqrt{\frac{2}{mn}}\left(\frac{d}{\gamma} + \log B + \left(\frac{16}{\gamma} + \frac{4dL^2}{\nu m}\right)p\right)^{\frac{1}{2}}.
\end{aligned}$$

Inequality (i) follows from taking the supremum of $V(\theta_t)$, inequality (ii) from plugging in the upper bound from Equation 7.3.

Taking the limit as $T \to \infty$ and using Fatou's Lemma, we therefore have that the moments of $\mathbb{E}[\|\theta - \theta^*\|^p]^{\frac{1}{p}}$, with probability at least $1 - \delta_1$, grow at a rate of $\sqrt{p}$:

$$\mathbb{E}[\|\theta - \theta^*\|^p]^{\frac{1}{p}} \leq \liminf_{T\to\infty} \mathbb{E}[\|\theta_T - \theta^*\|^p]^{\frac{1}{p}} \tag{7.4}$$

$$= \sqrt{\frac{2}{mn}}\left(\frac{d}{\gamma} + \log B + \left(\frac{16}{\gamma} + \frac{4dL^2}{\nu m}\right)p\right)^{\frac{1}{2}}. \tag{7.5}$$

To simplify notation, let $D = \left(\frac{d}{\gamma} + \log B\right)$, and $\sigma = \left(\frac{16}{\gamma} + \frac{4dL^2}{\nu m}\right)$. Therefore we have:

$$\mathbb{E}[\|\theta - \theta^*\|^p]^{\frac{1}{p}} \leq \sqrt{\frac{2}{mn}\left(D + \sigma p\right)} \tag{7.6}$$

The result (7.6), guarantees us that the norm of the uncentered random variable $\theta - \theta^*$ has subgaussian tails. We make the parameters explicit via Markov's inequality:

$$\begin{aligned}
\mathbb{P}_{\theta\sim\mu_a^{(n)}}\left(\|\theta - \theta^*\| > \epsilon\right) &\leq \frac{\mathbb{E}[\|\theta - \theta^*\|^p]}{\epsilon^p} \\
&\leq \left(\frac{\sqrt{2\left(D + \sigma p\right)}}{\sqrt{mn}\epsilon}\right)^p.
\end{aligned}$$

Choosing $p = 2 \log 1/\delta$ and letting

$$\epsilon = e^{\frac{1}{2}} \sqrt{\frac{2}{mn} (D + \sigma p)}$$

gives us our desired solution:

$$\mathbb{P}_{\theta \sim \mu_a^{(n)}[\gamma_a]} \left(\|\theta - \theta^*\|_2 > \sqrt{\frac{2e}{mn} \left(\frac{d}{\gamma} + \log B + \left(\frac{32}{\gamma} + \frac{8dL^2}{\nu m} \right) \log (1/\delta) \right)} \right) < \delta,$$

for $\delta \leq e^{-0.5}$. □

7.2 Chapter Summary

We remark that the posterior concentration result we derived in this chapter in Theorem 18 has a number of desirable properties. Through the presence of B_a, it reflects an explicit dependence on the quality of the prior. In particular, $B = 0$ if the prior is properly centered such that its mode is at θ^* or if the prior is uninformative or nearly flat everywhere. We further remark that the concentration result also scales with the variance of θ_a which is on the order of $\frac{d}{mn}$. The bound also has an explicit dependence on the quality of the data received from the arm through its dependence on $1/\delta_1$. Lastly, we remark that this concentration result holds for any $n > 0$ and the constants are explicitly defined in terms of the smoothness and structural assumptions on the priors, likelihoods, and reward distributions. This makes it more amenable for use in constructing regret guarantees, since we do not have to wait for a burn-in period for the result to hold as in [172] and [130]. Moreover, the dependence on the dimension of the parameter space and constants is explicit. These properties allow us to use this result to prove the order-optimal regret of exact Thompson Sampling (in this family of bandit problems) in the next chapter.

Chapter 8

Exact Thompson Sampling

In this chapter we build upon the posterior concentration results from Chapter 7 to give finite-time regret guarantees for exact Thompson sampling in log-concave bandits— a larger family of priors and posteriors than have previously been analyzed in the literature. For clarity of exposition, in the first section we given an overview of the proof of logarithmic regret of Thompson Sampling in log-concave bandit problems, and present the details of the proof in Section 8.2

8.1 Regret Bounds for Exact Thompson Sampling in Log-Concave Bandits

To analyze the regret of exact Thompson sampling we proceed as is common in regret proofs for multi-armed bandits by upper bounding $T_a(T)$, the number of times a sub-optimal arm $a \in \mathcal{A}$ is pulled up to time T. Without loss of generality we assume throughout this section that arm 1 is the optimal arm, and define the filtration associated with a run of the algorithm as $\mathcal{F}_t = \{A_1, X_1, A_2, X_2, ..., A_t, X_t\}$.

We first define the low-probability event that the mean calculated from the value of $\theta_{a,t}$ sampled from the posterior at time $t \leq T$, $r_{a,t}(T_a(t))$, is greater than $\bar{r}_1 - \epsilon$ (recall that $\bar{r}_1$ is the optimal arm's mean): $E_a(t) = \{r_{a,t}(T_a(t)) \geq \bar{r}_1 - \epsilon\}$ for some $\epsilon > 0$. Given these events, we proceed to decompose the expected number of pulls of a sub-optimal arm $a \in \mathcal{A}$ as:

$$\mathbb{E}[T_a(T)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}(A_t = a)\right] = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}(A_t = a, E_a^c(t))\right]}_{\text{I}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}(A_t = a, E_a(t))\right]}_{\text{II}}. \quad (8.1)$$

These two terms satisfy the following standard bounds (see, e.g., [94]):

Lemma 4 (Bounding I and II)**.** *For a sub-optimal arm $a \in \mathcal{A}$, we have that:*

$$\mathrm{I} \leq \mathbb{E}\left[\sum_{s=1}^{T-1} \frac{1}{p_{1,s}} - 1\right]; \tag{8.2}$$

$$\mathrm{II} \leq 1 + \mathbb{E}\left[\sum_{s=1}^{T} \mathbb{I}\left(p_{a,s} > \frac{1}{T}\right)\right], \tag{8.3}$$

where $p_{a,s} = \mathbb{P}(r_{a,t}(s) > \bar{r}_1 - \epsilon | \mathcal{F}_{t-1})$, for some $\epsilon > 0$.

The proof of these results are standard for the regret of Thompson sampling and can be found in Section 8.2, Lemmas 6 and 7, for completeness.

Given Lemma 4, we see that to bound the regret of Thompson Sampling it is sufficient to bound the two terms I and II.

To bound term I, we first show that for all times $t = 1, ..., T$, and number of samples collected from arm 1, the probability $p_{1,n} = \mathbb{P}(r_{1,t}(n) > \bar{r}_1 - \epsilon | \mathcal{F}_{t-1})$ is lower bounded by a constant depending only on the quality of the prior for arm 1. This guarantees the posterior for the optimal arm is approximately optimistic with at least a constant probability, and requires a proper choice of γ_1. We note the unscaled posterior provides the correct concentration with respect to the number of data samples $T_a(t)$, when $T_a(t)$ is large. This is sufficient to upper bound the trailing terms of I, that is, summands in Equation 8.2 for large s. Unfortunately concentration is not enough to bound term I, since the early summands of Equation 8.2 corresponding to small values of s could be extremely large. Intuitively, the random variable $r_{1,t}(s)$ can be thought of as centered around the posterior mean of arm 1. Though this is close to the true value of $\bar{r}_1$ with high probability, when $T_1(t)$ is small, concentration alone does not preclude the possibility that the posterior mean underestimates $\bar{r}_1$ by a value of at least ϵ. In order to ensure $p_{1,s}$ is large enough in these cases, we require $r_{1,t}(s)$ to have sufficient variance to overcome this potential underestimation bias. We show that a scaled posterior $\mu_a^{(T_a(t))}[\gamma_a]$ with $\gamma_a = (8 d_a \kappa_a^3)^{-1}$ in Algorithm 2 ensures $r_{1,t}(s)$ has enough variance.

Lemma 5. *Suppose the likelihood and reward distributions satisfy Assumptions 11-13, then for all $n = 1, ..., T$ and $\gamma_1 = \frac{1}{8 d_1 \kappa_1^3}$:*

$$\mathbb{E}\left[\frac{1}{p_{1,n}}\right] \leq C\sqrt{B_1 \kappa_1},$$

where C is a universal constant independent of the problem-dependent parameters.

We find that a proper choice of γ_1 is required to ensure that that the posterior on the optimal arm has a large enough variance to guarantee a degree of optimism despite the randomness in its mean. Scaling up the posterior was also noted to be necessary in linear bandits (see, e.g., [3, 9]) to ensure optimal regret. In practice, since we do not know a priori which is the optimal arm, we must scale the posterior of each arm by a parameter γ_a.

The quantity $B_1 = \frac{\max_\theta \pi_1(\theta)}{\pi_1(\theta_1^*)}$ captures a worst case dependence on the quality of the prior for the optimal arm, and can be seen as the expected number of samples from the prior until an optimistic sample is observed.

By using this upper bound in combination with the posterior concentration result derived in Theorem 18, we can further bound I and II. We note that in contrast with simple subgaussian concentration bounds, our posterior concentration rates have a bias term decreasing at a rate of $1/\sqrt{\text{number of samples}}$. In our analysis we carefully track and control the effects of this bias term ensuring it does not compromise our log-regret guarantees. Indeed, using the posterior concentration in the bounds from Lemma 4 we show that, for $\gamma_a = \frac{1}{8 d_a \kappa_a^3}$ there are two universal constants $C_1, C_2 > 0$ independent of the problem-dependent parameters such that:

$$\begin{aligned} \text{I} &\leq C_1 \sqrt{\kappa_1 B_1} \left\lceil \frac{A_1^2}{m_1 \Delta_a^2} (D_1 + \sigma_1) \right\rceil + 1; \\ \text{II} &\leq \frac{C_2 A_a^2}{m_a \Delta_a^2} (D_a + \sigma_a \log(T)), \end{aligned}$$

where for $a \in \mathcal{A}$, D_a and σ_a are given by:

$$D_a = \log B_a + d_a^2 \kappa_a^3, \qquad \sigma_a = d_a \kappa_a^3 + d_a \kappa_a^2.$$

Finally, combining all these observations we obtain the following regret guarantee:

Theorem 19 (Regret of Exact Thompson Sampling)**.** *When the likelihood and true reward distributions satisfy Assumptions 11-13 and $\gamma_a = \frac{1}{8 d_a \kappa_a^3}$ we have that the expected regret after $T > 0$ rounds of Thompson sampling with exact sampling satisfies:*

$$\mathbb{E}[R(T)] \leq \sum_{a>1} \frac{C A_a^2}{m_a \Delta_a} \left(\log B_a + d_a^2 \kappa_a^3 + d_a \kappa_a^3 \log(T) \right) + \sqrt{\kappa_1 B_1} \frac{C A_1^2}{m_1 \Delta_a} \left(\log B_1 + d_1^2 \kappa_1^3 \right) + \Delta_a,$$

where C is a universal constant independent of problem-dependent parameters.

The proof of the theorem follows directly from the bounds on term I and II in (8.1). We remark that this regret bound gives an $O\left(\frac{\log(T)}{\Delta}\right)$ asymptotic regret guarantee, but holds for any $T > 0$. This further highlights that Thompson sampling is a competitive alternative to UCB algorithms since it achieves the optimal problem-dependent rate for multi-armed bandit algorithms first presented in Lai and Robbins [91].

Our bound also has explicit dependencies on the dimension of the parameter space of the likelihood distributions for each arm, as well as on the quality of the priors through the presence of B_a and B_1. We note that the dependence on the priors does not distinguish between "good" and "bad" priors. Indeed, the parameter $B_a \geq 1$ is worst case, and does not capture the potential advantages of good priors in Thompson sampling, that we observe in our numerical experiments in Section 9.5. Further, we remark that our bound exhibits a

worse dependence on the prior for the optimal arm ($O(\sqrt{B_1}\log(B_1))$) than for sub-optimal arms ($O(\log(B_a))$). This is also a worst case dependence which captures the expected number of samples from the prior until an approximately optimistic sample is observed, which we believe to be unavoidable.

Finally, we note that our regret bound scales with the variances of the reward and likelihood families since $\frac{1}{m_a}$ and $\frac{1}{\nu_a}$ reflect the variance of the likelihoods in θ and the rewards X_a respectively.

8.2 Detailed Proofs of the Regret of Exact Thompson Sampling

We now present the detailed proof of logarithmic regret of Thompson sampling under our assumptions with samples from the true posterior. We begin with the decomposition of the number of times a sub-optimal arm has been pulled from (8.1).

$$\mathbb{E}[T_a(T)] = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}(A_t = a)\right] = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}(A_t = a, E_a^c(t))\right]}_{I} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}(A_t = a, E_a(t))\right]}_{II}.$$

In Lemma 6 we upper bound (I), and then bound term (II) in Lemmas 7. We note that this proof follows a similar structure to that of the regret bound for Thompson sampling for Bernoulli bandits and bounded rewards in [7]. However, to give the regret guarantees that incorporate the quality of the priors as well as the potential errors and lack of independence resulting from the approximate sampling methods we discuss in Section 9 the proof is more complex.

Lemma 6 (Bounding I)**.** *For a sub-optimal arm $a \in \mathcal{A}$, we have that:*

$$I = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}(A_t = a, E_a^c(t))\right] \leq \mathbb{E}\left[\sum_{s=1}^{T-1}\frac{1}{p_{1,s}} - 1\right].$$

where $p_{a,s} = \mathbb{P}(r_{a,t}(s) > \bar{r}_1 - \epsilon|\mathcal{F}_{t-1})$, for some $\epsilon > 0$.

Proof. To bound term I of (8.1), we first recall A_t is the arm achieving the largest sample reward mean at round t. Further, we define A'_t to be the arm achieving the maximum sample mean value among all the suboptimal arms:

$$A'_t = \underset{a\in\mathcal{A}, a\neq 1}{\operatorname{argmax}}\, r_a(t, T_a(t)).$$

Since $\mathbb{E}\left[\mathbb{I}(A_t = a, E_a^c(t))\right] = \mathbb{P}\left(A_t = a, E_a^c(t)\right)$, we aim to bound $\mathbb{P}(A_t = a, E_a^c(t)|\mathcal{F}_{t-1})$. We note that the following inequality holds:

$$\begin{aligned}\mathbb{P}(A_t = a, E_a^c(t)|\mathcal{F}_{t-1}) &\leq \mathbb{P}(A_t' = a, E_a^c(t)|\mathcal{F}_{t-1})(\mathbb{P}(r_1(t, T_1(t)) \leq \bar{r}_1 - \epsilon|\mathcal{F}_{t-1})) \\ &= \mathbb{P}(A_t' = a, E_a^c(t)|\mathcal{F}_{t-1})(1 - \mathbb{P}(E_1(t)|\mathcal{F}_{t-1})). \end{aligned} \tag{8.4}$$

We also note that the term $\mathbb{P}(A_t' = a, E_a^c(t)|\mathcal{F}_{t-1})$ can be bounded as follows:

$$\begin{aligned}\mathbb{P}(A_t = 1, E_a^c(t)|\mathcal{F}_{t-1}) &\overset{(i)}{\geq} \mathbb{P}(A_t' = a, E_a^c(t), E_1(t)|\mathcal{F}_{t-1}) \\ &= \mathbb{P}(A_t' = a, E_a^c(t)|\mathcal{F}_{t-1})\mathbb{P}(E_1(t).|\mathcal{F}_{t-1}) \end{aligned} \tag{8.5}$$

Inequality (i) holds because $\{A_t' = a, E_a^c(t), E_1(t)\} \subseteq \{A_t = 1, E_a^c(t), E_1(t)\}$. The equality is a consequence of the conditional independence of $E_1(t)$ and $\{A_t' = a, E_a^c(t)\}$ (conditioned on $\mathcal{F}_{t-1}$). [1]

Assuming $\mathbb{P}(E_1(t)|\mathcal{F}_{t-1}) > 0$ and[2] putting inequalities 8.4 and 8.5 together gives the following upper bound for $\mathbb{P}(A_t = a, E_a^c(t)|\mathcal{F}_{t-1})$:

$$\mathbb{P}(A_t = a, E_a^c(t)|\mathcal{F}_{t-1}) \leq \mathbb{P}(A_t = 1, E_a^c(t)|\mathcal{F}_{t-1}) \left(\frac{1 - \mathbb{P}(E_1(t)|\mathcal{F}_{t-1})}{\mathbb{P}(E_1(t)|\mathcal{F}_{t-1})}\right).$$

Letting $P(E_1(t)|\mathcal{F}_{t-1}) := p_{1,T_1(t)}$ and noting that$\{A_t = 1, E_a^c(t)\} \subseteq \{A_t = 1\}$:

$$\mathbb{P}(A_t = a, E_a^c(t)|\mathcal{F}_{t-1}) \leq \mathbb{P}(A_t = 1|\mathcal{F}_{t-1}) \left(\frac{1}{p_{1,T_1(t)}} - 1\right). \tag{8.6}$$

Now, we use this to give an upper bound on the term of interest:

[1] The conditional independence property holds for all of our sampling mechanisms because the sample distributions for the two distinct arms $(a, 1)$ are always conditionally independent on $\mathcal{F}_{t-1}$

[2] In all the cases we consider, including approximate sampling schemes, this property holds. In that case, since the Gaussian noise in the Langevin diffusion ensures all sets of the form (a, b) have nonzero probability mass.

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}(A_t=a,E_a^c(t))\right] &\overset{(i)}{=} \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{I}\left(A_t=a,E_a^c(t)\right)|\mathcal{F}_{t-1}\right]\right] \\
&\overset{(ii)}{=} \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{P}\left(A_t=a,E_a^c(t)|\mathcal{F}_{t-1}\right)\right] \\
&\overset{(iii)}{\leq} \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{P}(A_t=1|\mathcal{F}_{t-1})\left(\frac{1}{p_{1,T_1(t)}}-1\right)\right] \\
&\overset{(iv)}{=} \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{I}(A_t=1)|\mathcal{F}_{t-1}\right]\left(\frac{1}{p_{1,T_1(t)}}-1\right)\right] \\
&\overset{(v)}{=} \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}(A_t=1)\left(\frac{1}{p_{1,T_1(t)}}-1\right)\right] \\
&\overset{(vi)}{\leq} \mathbb{E}\left[\sum_{s=1}^{T-1}\frac{1}{p_{1,s}}-1\right].
\end{aligned}$$

Here the equality (i) is a consequence of the tower property, and equality (ii) by noting that:

$$\mathbb{E}\left[\mathbb{I}\left(A_t=a,E_a^c(t)\right)|\mathcal{F}_{t-1}\right]=\mathbb{P}\left(A_t=a,E_a^c(t)|\mathcal{F}_{t-1}\right).$$

Inequality (iii) follows by from Equation 8.6, and equality (iv) follows by definition. Finally, equality (v) follows by the tower property and the last line each the fact that $T_1(t)=s$ and $A_t=1$ can only happen once for every $s=1,...,T$. This completes the proof. □

Given the bound on (I) from (8.1), we now present a bound on (II).

Lemma 7 (Bounding II - exact posterior)**.** *For a sub-optimal arm $a\in\mathcal{A}$, we have that:*

$$II=\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}(A_t=a,E_a(t))\right]\leq 1+\mathbb{E}\left[\sum_{s=1}^{T}\mathbb{I}\left(p_{a,s}>\frac{1}{T}\right)\right].$$

where $p_{a,s}=\mathbb{P}(r_{a,t}(s)>\bar{r}_1-\epsilon|\mathcal{F}_{t-1})$, for some $\epsilon>0$.

Proof. The upper bound for term II in (8.1) follows the exact same proof as in [7], and we recreate it for completeness below. Let $\mathcal{T}=\{t:p_{a,T_a(t)}>\frac{1}{T}\}$, then:

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}(A_t=a,E_a(t))\right]\leq\underbrace{\mathbb{E}\left[\sum_{t\in\mathcal{T}}\mathbb{I}(A_t=a)\right]}_{I}+\underbrace{\mathbb{E}\left[\sum_{t\notin\mathcal{T}}\mathbb{I}(E_a(t))\right]}_{II}. \tag{8.7}$$

By definition, term I in (8.7) satisfies:

$$\sum_{t\in\mathcal{T}}\mathbb{I}(A_t=a)=\sum_{t\in\mathcal{T}}\mathbb{I}\left(A_t=a,p_{a,T_a(t)}>\frac{1}{T}\right)\leq\sum_{s=1}^{T}\mathbb{I}\left(p_{a,s}>\frac{1}{T}\right).$$

To address term II in (8.7), we note that, by definition: $\mathbb{E}[\mathbb{I}(E_a(t))|\mathcal{F}_{t-1}]=p_{a,T_a(t)}$. Therefore, using the definition of the set of times $\mathcal{T}$, we can construct this simple upper bound:

$$\begin{aligned}\mathbb{E}\left[\sum_{t\notin\mathcal{T}}\mathbb{I}(E_a(t))\right]&=\mathbb{E}\left[\sum_{t\notin\mathcal{T}}\mathbb{E}\left[\mathbb{I}\left(E_a(t)\right)|\mathcal{F}_{t-1}\right]\right]\\&=\mathbb{E}\left[\sum_{t\notin\mathcal{T}}p_{a,t}\right]\\&\leq\sum_{t\notin\mathcal{T}}\frac{1}{T}\\&\leq 1.\end{aligned}$$

Using the two upper bounds for terms I and II in (8.7) gives out desired result:

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}(A_t=a,E_a(t))\right]\leq 1+\mathbb{E}\left[\sum_{s=1}^{T}\mathbb{I}\left(p_{a,s}>\frac{1}{T}\right)\right].$$

□

Regret of Exact Thompson Sampling

We now present two technical lemmas and their proofs which enable us to bound the regret of exact Thompson sampling. The first is Lemma 5, which we restate below, which provides a lower bound on the probability of an arm begin optimistic in terms of the quality of the prior.

Lemma. *Suppose the likelihood and reward distributions satisfy Assumptions 11-13, then for all* $n=1,...,T$ *and* $\gamma_1=\frac{\nu_1 m_1^2}{8d_1L_1^3}$*:*

$$\mathbb{E}\left[\frac{1}{p_{1,n}}\right]\leq 64\sqrt{\frac{L_1}{m_1}B_1}.$$

Proof. Throughout this proof we drop the dependence on the arm to simplify notation (unless necessary). We first analyze $\|\theta^*-\theta_u\|^2$ where θ_u is the mode of the posterior of arm 1 after having received n samples from the arm which satisfies:

$$\frac{1}{n}\nabla\log\pi_1(\theta_u)+\nabla F_{1,n}(\theta_u)=0$$

Given this definition, and letting $\hat{\theta} = \theta_u - \theta^*$ we have that:

$$\hat{\theta}^T \left(\nabla F_n(\theta^*) - \nabla F_n(\theta_u)\right) - \frac{1}{n}\hat{\theta}^T \nabla \log \pi(\theta_u) = \hat{\theta}^T \nabla F_n(\theta^*)$$

$$m\|\hat{\theta}\|^2 \leq \frac{m}{2}\|\hat{\theta}\|^2 + \frac{1}{2m}\|\nabla F_n(\theta^*)\|^2 + \frac{\log B_1}{n}$$

$$\|\hat{\theta}\|^2 \leq \frac{1}{m^2}\|\nabla F_n(\theta^*)\|^2 + \frac{2\log B_1}{mn}.$$

Noting that $|a^T(\theta^* - \theta_u)| \leq \sqrt{A^2\|\hat{\theta}\|^2}$ we find that:

$$p_{1,s} = Pr\left(\alpha^T\left(\theta - \theta_u\right) \geq \alpha^T\left(\theta^* - \theta_u\right) - \epsilon\right)$$

$$\geq Pr\left(\alpha^T(\theta - \theta_u) \geq \underbrace{\sqrt{\frac{2A^2\log B_1}{nm} + \frac{A^2}{m^2}\|\nabla F_n(\theta^*)\|^2}}_{=t}\right),$$

where we note that $\|F_n(\theta^*)\|$ in Proposition 1 is a 1-dimensional $\frac{dL_a}{\sqrt{n\nu}}$ sub-Gaussian random variable.

Now, since we know that the posterior over θ is $\gamma(n+1)L$-smooth and γmn-strongly log concave, with mode θ_u, we know from [169], Theorem 3.8, that the marginal density of $\alpha^T\theta$ is $\frac{\gamma(n+1)L}{A^2}$-smooth and $\frac{\gamma mn}{A^2}$-strongly log-concave.

Thus we have that:

$$Pr\left(\alpha^T\left(\theta - \theta_u\right) \geq t\right) \geq \sqrt{\frac{nm}{(n+1)L}} Pr(Z \geq t)$$

where $Z \sim \mathcal{N}\left(0, \frac{A^2}{\gamma(n+1)L}\right)$.

Now using a lower bound on the cumulative density function of a Gaussian random variable, we find that, for $\sigma^2 = \frac{A^2}{\gamma(n+1)L}$:

$$p_{1,s} \geq \sqrt{\frac{nm}{2\pi(n+1)L}} \begin{cases} \frac{\sigma t}{t^2+\sigma^2} e^{-\frac{t^2}{2\sigma^2}} & : t > \frac{A}{\sqrt{\gamma(n+1)L}} \\ 0.34 & : t \leq \frac{A}{\sqrt{\gamma(n+1)L}} \end{cases}$$

Thus we have that:

$$\frac{1}{p_{1,s}} \leq \sqrt{\frac{2\pi(n+1)L}{nm}} \begin{cases} \frac{t^2+\sigma^2}{\sigma t} e^{\frac{t^2}{2\sigma^2}} & : t > \frac{A}{\sqrt{\gamma(n+1)L}} \\ \frac{1}{0.34} & : t \leq \frac{A}{\sqrt{\gamma(n+1)L}} \end{cases}$$

$$\leq \sqrt{\frac{2\pi(n+1)L}{nm}} \begin{cases} \left(\frac{t}{\sigma} + 1\right) e^{\frac{t^2}{2\sigma^2}} & : t > \frac{A}{\sqrt{\gamma(n+1)L}} \\ 3 & : t \leq \frac{A}{\sqrt{\gamma(n+1)L}} \end{cases}$$

Taking the expectation of both sides with respect to the samples $X_1, ..., X_n$, letting $\kappa = L/m$, and using the fact that $\frac{n+1}{n} \leq 2$ for $n \geq 1$ we find that:

$$\mathbb{E}\left[\frac{1}{p_{1,s}}\right] \leq 6\sqrt{\pi\kappa} + 2\sqrt{\pi\kappa}\mathbb{E}\left[\left(\frac{\sqrt{\frac{2A^2\log B_1}{nm} + \frac{A^2}{m^2}\|\nabla F_n(\theta^*)\|^2}}{\sigma} + 1\right)e^{\frac{t^2}{2\sigma^2}}\right]$$

Noting that $\sqrt{\frac{2A^2\log B_1}{nm} + \frac{A^2}{m^2}\|\nabla F_n(\theta^*)^T\|^2} \leq A\sqrt{\frac{2\log B_1}{nm}} + \frac{A}{m}\|\nabla F_n(\theta^*)\|$, and letting $Y = \|\nabla F_n(\theta^*)\|$ to simplify notation, this further simplifies:

$$\mathbb{E}\left[\frac{1}{p_{1,s}}\right] \leq 6\sqrt{\pi\kappa} + 2\sqrt{\pi\kappa}\mathbb{E}\left[\left(\sqrt{4\gamma\kappa\log B_1} + \frac{A}{m\sigma}Y\right)e^{2\gamma\kappa\log B_1 + \frac{(n+1)\gamma L}{2m^2}Y^2}\right]$$

Using the Cauchy-Schwartz inequality we can further expand this upper bound and find that:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{p_{1,s}}\right] \leq\ & 6\sqrt{\pi\kappa} + 2\sqrt{\pi\kappa}e^{2\gamma\kappa\log B_1} \\ & \cdot\left(\sqrt{4\gamma\kappa\log B_1}\mathbb{E}\left[e^{\frac{(n+1)\gamma L}{2m^2}Y^2}\right] + \frac{A}{m\sigma}\sqrt{\mathbb{E}\left[Y^2\right]}\sqrt{\mathbb{E}\left[e^{\frac{(n+1)\gamma L}{m^2}Y^2}\right]}\right)\end{aligned}$$

Since Y is sub-Gaussian, Y^2 is sub-exponential such that:

$$\mathbb{E}\left[e^{\lambda Y^2}\right] \leq e \quad \text{and} \quad \mathbb{E}\left[Y^2\right] \leq 2\frac{dL^2}{\nu n}$$

for $\lambda < \frac{n\nu}{4dL^2}$. Therefore, if

$$\gamma = \frac{\nu m^2}{8dL^3},$$

simplifying the bound further gives:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{p_{1,s}}\right] &\leq 6\sqrt{\pi\kappa} + 2\sqrt{\pi\kappa}e^{2\gamma\kappa\log B_1}\left(\sqrt{4\gamma\kappa\log B_1}e + 2\sqrt{\frac{e\gamma(n+1)L}{m^2}\frac{dL^2}{\nu n}}\right) \\ &\leq 6\sqrt{\pi\kappa} + 2\sqrt{\pi\kappa}e^{\frac{\log B_1}{4}}(\sqrt{\frac{\log B_1}{2}}e + 2\sqrt{e}),\end{aligned}$$

where we have used the fact that $\kappa, d \geq 1$ and the fact that we can assume without loss of generality that $L/\nu \geq 1$. Thus, this bound simplifies to:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{p_{1,s}}\right] &\leq 6\sqrt{\pi\kappa} + 2\sqrt{\pi\kappa}e^{2\gamma\kappa\log B_1}\left(\sqrt{4\gamma\kappa\log B_1}e + 2\sqrt{\frac{e\gamma(n+1)L}{m^2}\frac{dL^2}{\nu n}}\right) \\
&\leq 2\sqrt{\pi\kappa}\,(B_1)^{\frac{1}{4}}\left(\sqrt{\frac{\log B_1}{2}}e + 7\right) \\
&\leq 4\sqrt{\pi\kappa}\,(B_1)^{\frac{1}{4}}\left(\sqrt{\log B_1} + 4\right) \\
&\leq 64\sqrt{\kappa B_1}
\end{aligned}$$

where we used the fact that $x^{1/4}(\sqrt{\log x}+4) \leq 8\sqrt{x}$ for $x \geq 1$ and $\sqrt{\pi} < 2$ to simplify our bound. □

The last technical lemma upper bounds the two terms defined in Lemma 4.

Lemma 8. *Suppose the likelihood, true reward distributions, and priors satisfy Assumptions 11-13, then for* $\gamma_a = \frac{\nu_a m_a^2}{8 d_a L_a^3}$*:*

$$\sum_{s=1}^{T-1}\mathbb{E}\left[\frac{1}{p_{1,s}} - 1\right] \leq 64\sqrt{\frac{L_1}{m_1}B_1}\left[\frac{8eA_1^2}{m\Delta_a^2}(D_1 + 4\sigma_1\log 2)\right] + 1 \tag{8.8}$$

$$\sum_{s=1}^{T}\mathbb{E}\left[\mathbb{I}\left(p_{a,s} > \frac{1}{T}\right)\right] \leq \frac{8eA_a^2}{m\Delta_a^2}(D_a + 2\sigma_a\log(T)) \tag{8.9}$$

Where for $a \in \mathcal{A}$*,* D_a *is given by:*

$$D_a = \log B_a + \frac{8d_a^2L_a^3}{m_a^2\nu_a} \qquad \sigma_a = \frac{256 d_a L_a^3}{m_a^2\nu_a} + \frac{8d_aL_a^2}{m_a\nu_a}$$

Proof. We begin by showing that (8.8) holds. To do so, we first note that, by definition $p_{1,s}$ satisfies:

$$\begin{aligned}
p_{1,s} &= \mathbb{P}(r_{1,t}(s) > \bar{r}_1 - \epsilon|\mathcal{F}_{t-1}) && (8.10)\\
&= 1 - \mathbb{P}(r_{1,t}(s) - \bar{r}_1 < -\epsilon|\mathcal{F}_{t-1}) && (8.11)\\
&\geq 1 - \mathbb{P}(|r_{1,t}(s) - \bar{r}_1| > \epsilon|\mathcal{F}_{t-1}) && (8.12)\\
&\geq 1 - \mathbb{P}_{\theta\sim\mu_1^{(s)}}\left(\|\theta - \theta^*\| > \frac{\epsilon}{A_1}\right), && (8.13)
\end{aligned}$$

where the last inequality follows from the fact that $r_{1,t}(s)$ and $\bar{r}_1$ are A_a-Lipschitz functions of $\theta \sim \mu_1^{(s)}$ and θ^* respectively.

We then use the fact that the posterior distribution $\mathbb{P}_{\theta\sim\mu_1^{(s)}}$ satisfies the concentration bound from Theorem 18 for $\delta \in (0, e^{-1/2})$. Therefore, we have that:

$$\mathbb{P}_{\theta\sim\mu_1^{(s)}}\left(\|\theta-\theta^*\| > \frac{\epsilon}{A_1}\right) \leq \exp\left(-\frac{1}{2\sigma_1}\left(\frac{mn\epsilon^2}{2eA_1^2} - D_1\right)\right), \tag{8.14}$$

where we use the constant D_1 and σ_1 defined in the proof of Theorem 18 to simplify notation. We remark that this bound is not useful unless:

$$n > \frac{2eA_1^2}{\epsilon^2 m} D_1.$$

Thus, choosing $\epsilon = (\bar{r}_1 - \bar{r}_a)/2 = \Delta_a/2$ and ℓ as:

$$\ell = \left\lceil \frac{8eA_1^2}{m\Delta_a^2}(D_1 + 2\sigma_1 \log 2) \right\rceil.$$

we proceed as follows:

$$\begin{aligned}\sum_{s=\ell}^{T-1} \mathbb{E}\left[\frac{1}{p_{1,s}} - 1\right] &\leq \sum_{s=0}^{T-1} \frac{1}{1-\frac{1}{2}\delta(s)} - 1 \\ &\leq \int_{s=1}^{\infty} \frac{1}{1-\frac{1}{2}\delta(s)} - 1 ds,\end{aligned}$$

where:

$$\frac{1}{2}\delta(s) = \frac{1}{2}\exp\left(-\frac{1}{2\sigma_1}\left(\frac{m\epsilon^2}{2eA_1^2}s\right)\right) \leq e^{-1/2}, \forall s \geq \ell$$

and the first inequality follows from our choice of ℓ and the second by upper bounding the sum by an integral. To finish, we write $\delta(s) = \exp(-c * s)$, and solve the integral to find that:

$$\int_{s=1}^{\infty} \frac{1}{1-\frac{1}{2}\delta(s)} - 1 ds = \frac{\log 2 - \log\left(2e^c - 1\right)}{c} + 1 \leq \frac{\log 2}{c} + 1.$$

Plugging in for c gives:

$$\begin{aligned}\sum_{s=1}^{T-1} \mathbb{E}\left[\frac{1}{p_{1,s}} - 1\right] &\leq \sum_{s=1}^{\ell-1} \mathbb{E}\left[\frac{1}{p_{1,s}} - 1\right] + \frac{8eA_1^2}{m\Delta_a^2} 2\sigma_1 \log 2 + 1 \\ &\leq 64\sqrt{\frac{L_1}{m_1}B_1}\left\lceil \frac{8eA_1^2}{m\Delta_a^2}(D_1 + 4\sigma_1 \log 2)\right\rceil + 1\end{aligned}$$

To show that (8.9) holds, we do a similar derivation as in (8.13):

$$\begin{aligned}
\sum_{s=1}^{T} \mathbb{E}\left[\mathbb{I}\left(p_{a,s} > \frac{1}{T}\right)\right] &= \sum_{s=1}^{T} \mathbb{E}\left[\mathbb{I}\left(\mathbb{P}(r_{a,t}(s) - \bar{r}_a > \Delta_a - \epsilon | \mathcal{F}_{t-1}) > \frac{1}{T}\right)\right] \\
&= \sum_{s=1}^{T} \mathbb{E}\left[\mathbb{I}\left(\mathbb{P}(r_{a,t}(s) - \bar{r}_a > \frac{\Delta_a}{2} | \mathcal{F}_{t-1}) > \frac{1}{T}\right)\right] \\
&\leq \sum_{s=1}^{T} \mathbb{E}\left[\mathbb{I}\left(\mathbb{P}\left(|r_{a,t}(s) - \bar{r}_a| > \frac{\Delta_a}{2} \middle| \mathcal{F}_{t-1}\right) > \frac{1}{T}\right)\right] \\
&\leq \sum_{s=1}^{T} \mathbb{E}\left[\mathbb{I}\left(\mathbb{P}_{\theta \sim \mu_a^{(s)}[\gamma_a]}\left(\|\theta - \theta^*\| > \frac{\Delta_a}{2A_a}\right) > \frac{1}{T}\right)\right].
\end{aligned}$$

Using the posterior concentration result from Theorem 18 we upper bound the number of pulls $\bar{n}$ of arm a such that for all $n \geq \bar{n}$:

$$\mathbb{P}_{\theta \sim \mu_a^{(n)}[\gamma_a]}\left(\|\theta - \theta^*\| > \frac{\Delta_a}{2A_a}\right) \leq \frac{1}{T}.$$

Since the posterior for arm a after n pulls of arm a has the same form as in (8.14), and $1/T \leq e^{-0.5}$ we can choose $\bar{n}$ as:

$$\bar{n} = \frac{8eA_a^2}{m\Delta_a^2}(D_a + 2\sigma_a \log(T)).$$

This completes the proof.

□

Given these lemmas the proof of Theorem 19 is straightforward. For clarity, we restate the theorem below:

Theorem. *When the likelihood and true reward distributions satisfy Assumptions 11-13 and $\gamma_a = \frac{\nu_a m_a^2}{8 d_a L_a^3}$ we have that the expected regret after $T > 0$ rounds of Thompson sampling with exact sampling satisfies:*

$$\mathbb{E}[R(T)] \leq \sum_{a>1} \frac{CA_a^2}{m_a \Delta_a}\left(\log B_a + d_a^2 \kappa_a^3 + d_a \kappa_a^3 \log(T)\right) + \sqrt{\kappa_1 B_1} \frac{CA_1^2}{m_1 \Delta_a}\left(1 + \log B_1 + d_1^2 \kappa_1^3\right) + \Delta_a$$

Where C is a universal constant independent of problem-dependent parameters.

Proof. We invoke Lemmas 6 and 7, to find that:

$$\mathbb{E}\left[T_a(T)\right] \leq \underbrace{\sum_{s=1}^{T-1} \mathbb{E}\left[\frac{1}{p_{1,s}} - 1\right]}_{(I)} + \underbrace{\sum_{s=1}^{T} \mathbb{E}\left[\mathbb{I}\left(1 - p_{a,s} > \frac{1}{T}\right)\right]}_{(II)} \tag{8.15}$$

Now, invoking Lemma 8, we use the upper bounds for terms (I) and (II) in the regret decomposition and expanding D_a and D_1 to give:

$$\begin{aligned}\mathbb{E}[R(T)] &\leq \sum_{a>1} \frac{8eA_a^2}{m_a\Delta_a}\left(\log B_a + 8d_a\kappa_a^3\left(d_a + 66\log(T)\right)\right) \\ &\qquad + \sqrt{\kappa_1 B_1}\frac{512eA_a^2}{m_1\Delta_a^2}\left(1 + \log B_1 + 8d_1\kappa_1^3\left(d_1 + 132\log(2)\right)\right) + \Delta_a \\ &\leq \sum_{a>1} \frac{CA_a^2}{m_a\Delta_a}\left(\log B_a + d_a^2\kappa_a^3 + d_a\kappa_a^3\log(T)\right) \\ &\qquad + \sqrt{\kappa_1 B_1}\frac{CA_1^2}{m_1\Delta_a}\left(\log B_1 + d_1^2\kappa_1^3\right) + \Delta_a.\end{aligned}$$

□

8.3 Chapter Summary

In this chapter, we used our posterior contraction rates to get sharp, finite-time regret bounds for *exact* Thompson sampling multi-dimensional log-concave families with arbitrary log-concave priors. This generalizes the result of [86] to a more general class or priors and higher dimensional parametric families. In the next section we build upon these results to derive the first provably optimal approximate Thompson Sampling algorithm.

Chapter 9

Approximate Thompson Sampling

In this Chapter we present two approximate sampling schemes for generating samples from approximations of the posteriors at each round. For both, we give the values of the hyper-parameters and computation time needed to guarantee an approximation error which does *not* result in a drastic change in the regret of the Thompson sampling algorithm.

Before doing so, however, we first present a simple counterexample to illustrate that in the worst case, Thompson sampling with approximate samples incurs an irreducible regret dependent on the error between the posterior and the approximation to the posterior. In particular, by allowing the approximation error to decrease over time, we extract a relationship between the order of the regret and the level of approximation.

Example 6. *Consider a Gaussian bandit instance of two arms $\mathcal{A} = \{1, 2\}$ having mean rewards $\bar{r}_1$ and $\bar{r}_2$ and known unit variances. Further assume that the unknown parameters are the means of the distributions such that $\theta_a^* = \bar{r}_a$, and consider the case when the learner makes use of a zero-mean, unit-variance Gaussian prior over θ_a for $a = 1, 2$. Under these assumptions, after $X_{a,1}, \cdots, X_{a,n}$ samples, the posterior updates satisfy the following formulae [131]:*

$$P_{a,n}(\theta_a) \propto \mathcal{N}\left(\frac{n}{n+1}, \frac{1}{n+1}\right).$$

Let $\bar{r}_1 = 1$ and $\bar{r}_2 = 0$ such that Arm 1 is optimal. We now show there exists an approximate posterior $\tilde{P}_{a,t}$ of arm 2, satisfying $\mathrm{TV}(\tilde{P}_{2,t}, P_{2,t}) \leq n^{-\alpha}$ and such that if samples from $P_{1,t}$ and $\tilde{P}_{2,t}$ were to be used by a Thompson sampling algorithm, its regret would satisfy: $R(T) = \Omega(T^{1-\alpha})$.

We substantiate this claim by a simple construction. Let $\tilde{P}_{a,t}$ be $(1 - n^{-\alpha})P_{a,t} + n^{-\alpha}\delta_2$, where δ_2 denotes a delta mass centered at 2. $\tilde{P}_{a,t}$ is a mixture distribution between the true posterior and a point mass.

Clearly, for all $t \geq C$ for some universal constant C, with probability at least $n^{-\alpha}$ the posterior sample from arm 2 will be larger than the sample from arm 1. Since $t > n$, $t^{-\alpha} < n^{-\alpha}$ for $\alpha > 0$ and since the suboptimality gap equals 1, we conclude $R(T) = \Omega(\sum_{t=1}^{T} t^{-\alpha})$. Thus, to incur logarithmic regret, one needs $TV(\tilde{P}_{2,t}, P_{2,t}) = \Omega\left(\frac{1}{n}\right)$.

Example 6 extends upon the insights in [150] that constant approximation error can incur linear regret to highlight the fact that to achieve logarithmic regret, the total variation distance between the approximation of the posterior $\hat{\mu}_a^{(n)}$ and the true posterior $\mu_a^{(n)}$ must decrease as samples are collected. In particular it illustrates that the rate at which the approximation error decreases is directly linked to the resulting regret bound.

To generate samples from approximations to posteriors we propose two Langevin Markov Chain Monte Carlo (MCMC) algorithms. These algorithms can be seen as discretizations of the *continuous-time Langevin dynamics*, the stochastic process represented by the following stochastic differential equation :

$$\mathrm{d}\theta_t = -\nabla U(\theta_t)\ \mathrm{d}t + \sqrt{2}\ \mathrm{d}B_t.$$

We first encountered this continuous time Langevin dynamics in Eq. (7.1), where we have set $U(\theta) = -\gamma_a \left(nF_{n,a}(\theta) + \log \pi_a(\theta)\right) = -\gamma_a \sum_{i=1}^n \log p_a\left(x_{a,i}|\theta\right) - \gamma_a \log \pi_a(\theta)$ to prove posterior concentration of $\mu_a^{(n)}[\gamma_a]$.

One important feature of the Langevin dynamics is that its invariant distribution is proportional to $e^{-U(\theta)}$. We can therefore also use it to generate samples distributed according to the unscaled posterior distribution $\mu_a^{(n)}$. Via letting $U(\theta) = -\sum_{i=1}^n \log p_a\left(x_{a,i}|\theta\right) - \log \pi_a(\theta)$, we obtain continuous time dynamics which generates trajectories that converge towards the posterior distribution $\mu_a^{(n)}$ exponentially fast.

To obtain an implementable algorithm, we apply Euler-Maruyama discretization to the Langevin dynamics and arrive at the following ULA update:

$$\theta_{(i+1)h^{(n)}} \sim \mathcal{N}\left(\theta_{ih^{(n)}} - h^{(n)}\nabla U(\theta_{ih^{(n)}}), 2h^{(n)}\mathrm{I}\right).$$

Since $\nabla U(\theta) = -\sum_{i=1}^n \nabla \log p_a\left(x_{a,i}|\theta\right) - \nabla \log \pi_a(\theta)$ in the above update rule, the computation complexity within each iteration of the Langevin algorithm grows with the number of data being collected, n. To cope with the growing number of terms in $\nabla U(\theta)$, we take a stochastic gradient approach and define $\widehat{U}(\theta) = -\frac{n}{|\mathcal{S}|}\sum_{x_k \in \mathcal{S}} \nabla \log p_a(x_k|\theta) - \nabla \log \pi_a(\theta)$, where $\mathcal{S}$ is a subset of the dataset $\{x_{a,1}, \cdots, x_{a,n}\}$. For simplicity, we form $\mathcal{S}$ via subsampling uniformly from $\{x_{a,1}, \cdots, x_{a,n}\}$. Substituting the stochastic gradient $\nabla\widehat{U}$ for the full gradient ∇U in the above update rule results in the SGLD algorithm.

9.1 Convergence Rates for Langevin Algorithms

Given our intuition from Example 6 we first propose an unadjusted Langevin algorithm (ULA) [48] which generates samples from an approximate posterior which monotonically approaches the true posterior as data is collected and provably maintains the regret guarantee of exact Thompson sampling. Important to this effort, we demonstrate that the number of steps inside the ULA procedure does not scale with the time horizon, though the number of gradient evaluations scale with the number of times an arm has been pulled. To address this, we propose the stochastic gradient Langevin dynamics (SGLD) [192] variant

of ULA which has appealing computational benefits: under slightly stronger assumptions, SGLD takes constant number of iterations as well as constant number of data samples in the stochastic gradient estimate while maintaining the order-optimal regret of the exact Thompson sampling algorithm. Once again, for clarity of exposition we defer detailed proofs to the end of the chapter in Section 9.4.

Algorithm 3 (Stochastic Gradient) Langevin Algorithm for Arm a

Input : Data $\{x_{a,1}, \cdots, x_{a,n}\}$;
MCMC sample $\theta_{a,Nh^{(n-1)}}$ from last round

Set $\theta_0 = \theta_{a,t-1}$ for $a \in \mathcal{A}$

for $i = 0, 1, \cdots N$ **do**

Uniformly subsample $\mathcal{S} \subseteq \{x_{a,1}, \cdots, x_{a,n}\}$.

Compute $\nabla \widehat{U}(\theta_{ih^{(n)}}) = -\frac{n}{|\mathcal{S}|} \sum_{x_k \in \mathcal{S}} \nabla \log p_a(x_k|\theta_{ih^{(n)}}) - \nabla \log \pi_a(\theta_{ih^{(n)}})$.

Sample $\theta_{(i+1)h^{(n)}} \sim \mathcal{N}\left(\theta_{ih^{(n)}} - h^{(n)} \nabla \widehat{U}(\theta_{ih^{(n)}}), 2h^{(n)}\mathrm{I}\right)$.

Output: $\theta_{a,Nh^{(n)}} = \theta_{Nh^{(n)}}$ and $\theta_{a,t} \sim \mathcal{N}\left(\theta_{Nh^{(n)}}, \frac{1}{nL_a\gamma_a} I\right)$

As described in Algorithm 3, in each round t of the bandit algorithm we run the (stochastic gradient) Langevin algorithm for N steps to generate a sample of desirable quality for each arm. In particular, we first run a Langevin MCMC algorithm to generate a sample from an approximation to the unscaled posterior. To achieve the scaling with γ_a that we require for the analysis of the regret, we add zero-mean Gaussian noise with variance $\frac{1}{\gamma_a L_a n}$ to this sample. The distribution of the resulting sample has the same characteristics as those from the scaled posterior analyzed in Sec. 8.

Given Assumptions 14 and 13, we prove that running ULA with exact gradients provides appealing convergence properties. In particular, for a number of iterations independent of the number of rounds t or the number of samples from an arm, $n = T_a(t)$, ULA converges to an accuracy in Wasserstein-p distance which maintains the logarithmic regret of the exact algorithm (for more information on such metrics see Villani [189]). We note parenthetically that working with the Wasserstein-p distance provides us with a tighter MCMC convergence analysis (than with the total variation distance used in Example 6) that helps in conjunction with the regret bounds. The proofs of the ULA and SGLD convergence require a uniform strong log-concavity and Lipschitz smoothness condition of the family $p_a(X|\theta_a)$ over the parameter θ_a, a strengthening of Assumption 11.

Assumption 14 (Assumption on the family $p_a(X|\theta_a)$— strengthened for approximate sampling)**.** *Assume that* $\log p_a(x|\theta_a)$ *is* L_a*-smooth and* m_a*-strongly concave over the parameter*

θ_a:

$$-\log p_a(x|\theta'_a) - \nabla_\theta \log p_a(x|\theta'_a)^\top (\theta_a - \theta'_a) + \frac{m_a}{2}\|\theta_a - \theta'_a\|^2 \leq -\log p_a(x|\theta_a)$$
$$\leq -\log p_a(x|\theta'_a) - \nabla_\theta \log p_a(x|\theta'_a)^\top (\theta_a - \theta'_a) + \frac{L_a}{2}\|\theta_a - \theta'_a\|^2, \quad \forall \theta_a, \theta'_a \in \mathbb{R}^{d_a}, x \in \mathbb{R}.$$

This assumption allows us to prove a tight bound on the approximation error of ULA in the following theorem.

Theorem 20 (ULA Convergence)**.** *Suppose that Assumptions 12- 14 hold. We take step size* $h^{(n)} = \frac{1}{32}\frac{m_a}{n\left(L_a+\frac{1}{n}L_a\right)^2} = \mathcal{O}\left(\frac{1}{nL_a\kappa_a}\right)$ *and number of steps* $N = 640\frac{\left(L_a+\frac{1}{n}L_a\right)^2}{m_a^2} = \mathcal{O}\left(\kappa_a^2\right)$ *in Algorithm 3. If the posterior distribution satisfy the concentration inequality that* $\mathbb{E}_{\theta\sim\mu_a^{(n)}}\left[\|\theta-\theta^*\|^p\right]^{\frac{1}{p}} \leq \frac{1}{\sqrt{n}}\widetilde{D}$*, then for any positive even integer* p*, we have convergence of the ULA algorithm in* W_p *distance to the posterior* $\mu_a^{(n)}$*:* $W_p\left(\widehat{\mu}_a^{(n)}, \mu_a^{(n)}\right) \leq \frac{2}{\sqrt{n}}\widetilde{D}$*,* $\forall \widetilde{D} \geq \sqrt{\frac{32}{m_a}d_a p}$*.*

Although the number of iterations required for ULA to converge is constant with respect to the time horizon t, the number of gradient computations over the likelihood function within each iteration is $T_a(t)$. To tackle this issue, we sub-sample the data at each iteration and use a stochastic gradient MCMC method [106]. To be able to get convergence guarantees despite the larger variance this method incurs, we make a slightly stronger Lipschitz smoothness assumption on the parametric family of likelihoods.

Assumption 15 (Joint Lipschitz smoothness of the family $\log p_a(X|\theta_a)$: for SGLD)**.** *Assume a joint Lipschitz smoothness condition, which strengthens Assumptions 14 and 12 to impose the Lipschitz smoothness on the entire bivariate function* $\log p_a(x;\theta)$*:*[1]

$$\|\nabla_\theta \log p_a(x|\theta_a) - \nabla_\theta \log p_a(x'|\theta_a)\| \leq L_a \|\theta_a - \theta'_a\| + L_a^* \|x - x'\|, \quad \forall \theta_a, \theta'_a \in \mathbb{R}^{d_a}, x, x' \in \mathbb{R}.$$

Under this stronger assumption, we prove the fast convergence of the SGLD method in the following Theorem 21. Specifically, we demonstrate that for a suitable choice of stepsize $h^{(n)}$, number of iterations N, and size of the minibatch $k = |\mathcal{S}|$, samples generated by Algorithm 3 are distributed sufficiently close to the true posterior to ensure the optimal regret guarantee. By examining the number of iterations, N, and the size of the minibatch, k, we confirm that the algorithmic and sample complexity of our method do not grow with the number of rounds t, as advertised.

Theorem 21 (SGLD Convergence)**.** *Suppose that Assumptions 12- 14 hold, and further assume that the family* $\log p_a(x;\theta)$*, prior distributions, and that the true reward distributions satisfy Assumption 15. If we take the batch size* $k = \mathcal{O}\left(\kappa_a^2\right)$*, step size* $h^{(n)} = \mathcal{O}\left(\frac{1}{n}\frac{1}{\kappa_a L_a}\right)$ *and*

[1]For simplicity of notation, we let Lipschitz constants $L_a^* = L_a$ in the main paper.

number of steps $N = \mathcal{O}(\kappa_a^2)$ *in the SGLD algorithm, then for* $\delta_1 \in (0,1)$*, with probability at least* $1-\delta_1$ *with respect to* $X_{a,1}, ...X_{a,n}$*, we have convergence of the SGLD algorithm in the Wasserstein-p distance. In particular, between the n-th and the* $(n+1)$*-th pull to arm a, samples* $\theta_{a,t}$ *approximately follow the posterior* $\mu_a^{(n)}$*:*

$$W_p\left(\widehat{\mu}_a^{(n)}, \mu_a^{(n)}\right) \leq \sqrt{\frac{8}{nm_a}} \left(d_a + \log B_a + \left(32 + 8d_a\kappa_a^2\right) p\right)^{\frac{1}{2}},$$

where $\widehat{\mu}_a^{(n)}$ *is the probability measure associated with any of the sample(s)* $\theta_{a,Nh_a^{(n)}}$ *between the n-th and the* $(n+1)$*-th pull of arm a.*

We remark that we are able to keep the number of iterations, N, for both algorithms constant by initializing the current round of the approximate sampling algorithm using the output of the last round of the Langevin MCMC algorithm. If we initialized the algorithm independently from the prior, we would need $O(\log T_a(t))$ iterations to achieve this result, which would in turn yield a Thompson sampling algorithm for which the computational complexity grows with the time horizon. We note that this warm-starting procedure complicates the regret proof for the approximate Thompson sampling algorithms since the samples used by Thompson sampling are no longer independent.

By scrutinizing the stepsize $h^{(n)}$ and the accuracy level of the sample distribution in the Wasserstein distance $W_p\left(\widehat{\mu}_a^{(n)}, \mu_a^{(n)}\right)$, we note that we are taking smaller steps to get increasingly accurate MCMC samples as more data are being collected. This is due to the need of decreasing the error incurred by discretizing the continuous Langevin dynamics and stochastically estimating the gradient of the log posterior. However, the number of iterations and subsampled gradients are not increasing since the concentration of the posterior provides us with stronger contraction of the continuous Langevin dynamics and requires less work because $\mu_a^{(n)}$ and $\mu_a^{(n+1)}$ are closer.

We restate Theorem 21 and give explicit values of the hyper-parameters in Theorem 23 in the appendix, but remark that the proof of this theorem is novel in the MCMC literature. It builds upon and strengthens [49] by taking into account the discretization and stochastic gradient error to achieve strong convergence guarantees in the Wasserstein-p distance up to any finite order p. Other related works on the convergence of ULA can provide upper bounds in the Wassertein distances up to the second order (i.e., for $p \leq 2$) [see, e.g., 38, 42, 107, 188]. This bound in the Wasserstein-p distance for arbitrarily large p is necessary in guaranteeing the following Lemma 9, a similar concentration result as in Theorem 18 for the approximate samples $\theta_{a,t} \sim \bar{\mu}_a^{(n)}[\gamma_a]$.

Lemma 9. *Suppose that Assumptions 12- 14 hold, and further assume that the family* $\log p_a(x;\theta)$*, prior distributions, and that the true reward distributions satisfy Assumption 15, then for* $\delta \in (0, e^{-1/2})$*, the sample* $\theta_{a,t}$ *resulting from running the (stochastic gradient) ULA*

with N steps, a step size of $h^{(n)}$, and a batch size k as defined in Theorem 21 satisfies:

$$\mathbb{P}_{\theta_{a,t}\sim\bar{\mu}_a^{(n)}[\gamma_a]}\left(\|\theta_{a,t}-\theta_a^*\|_2 > \sqrt{\frac{36e}{m_a n}\left(d_a+\log B_a+2\left(\sigma_a+\frac{d_a}{18\kappa_a\gamma_a}\right)\log 1/\delta\right)}\right)<\delta.$$

where $\sigma_a = 16 + 4d_a\kappa_a^2$.

9.2 Regret of Approximate Thompson Sampling with Langevin Algorithms

Given that the concentration results of the samples from ULA and SGLD have the same form as that of exact Thompson sampling, we now show that approximate Thompson sampling achieves the same *finite*-time optimal regret guarantees (up to constant factors) as the exact Thompson sampling algorithm. To show this, we require an analgous result to Lemma 5 on the anti-concentration properties of the approximations to the scaled posteriors:

Lemma 10. *Suppose that Assumptions 12- 14 hold, and further assume that the family $\log p_a(x;\theta)$, prior distributions, and that the true reward distributions satisfy Assumption 15, then, if $\gamma_1 = O\left(\frac{1}{d_1\kappa_1^3}\right)$, for all $n = 1, ..., T$ all samples from the the (stochastic gradient) ULA method with the hyperparameters and runtime as described in Theorem 21 satisfy:*

$$\mathbb{E}\left[\frac{1}{p_{1,n}}\right] \leq C\sqrt{B_1},$$

where C is a universal constant independent of problem-dependent parameters.

The proof of Lemma 10 is similar to that of Lemma 5, but we are able to save a factor of $\sqrt{\kappa_1}$ due to the fact that the last step of the approximate sampling scheme samples $\theta_{a,t}$ from a Gaussian distribution as opposed to a strongly-log concave distribution which we must approximate with a Gaussian.

Given this lemma and our concentration results presented in the previous Chapter, the proof of logarithmic regret is essentially the same as that of the regret for exact Thompson sampling. However, more care has to be taken to deal with the fact that the samples from the approximate posteriors are no longer independent due to the fact that we warm-start our proposed sampling algorithms using previous samples. We cope with this issue by constructing concentration rates (of a similar form as in Lemma 9) on the distributions of the samples given the initial sample is sufficiently well behaved (see Lemmas 17 and 18). We then show that this happens with sufficiently high probability to maintain similar upper bounds on terms I and II from Lemma 4 in Lemma 19, which in turn allows us to prove the following Theorem in Appendix 9.4.

Theorem 22 (Regret of Thompson sampling with a (stochastic gradient) Langevin algorithm)**.** *Suppose that Assumptions 12- 14 hold, and further assume that the family* $\log p_a(x;\theta)$*, prior distributions, and that the true reward distributions satisfy Assumption 15, then the expected regret after* $T > 0$ *rounds of Thompson sampling with the (stochastic gradient) ULA method with the hyper-parameters and runtime as described in Theorem 21 satisfies:*

$$\mathbb{E}[R(T)] \leq \sum_{a>1} \frac{CA_a^2}{m_a \Delta_a} \left(\log B_a + d_a + d_a^2 \kappa_a^2 \log T\right) + \sqrt{B_1} \frac{CA_1^2}{m_1 \Delta_a} \left(1 + \log B_1 + d_1^2 \kappa_1^2 + d_1 \kappa_1^2 \log T\right) + 3\Delta_a,$$

where C *is a universal constant that is independent of problem dependent parameter and the scale parameter* $\gamma_a = O\left(\frac{1}{d_a \kappa_a^3}\right)$*.*

We note that Theorem 21 allows for SGLD to be implemented with a constant number of steps per iteration and a constant batch size with only the step size decreasing linearly with the number of samples. Combining this with our regret guarantee shows that an anytime algorithm for Thompson sampling with approximate samples can indeed achieve logarithmic regret.

Further, we remark that this bound exhibits a *worse* dependence on the quality of the prior on the optimal arm than in the exact sampling regime. In particular, we pay a $d_1^2\sqrt{B_1}\log T$ in this regret bound as opposed to $d_1^2\sqrt{B_1}$. Our regret bound in the approximate sampling regime does exhibit a slightly better dependence on the condition number of the family. This, we believe, is an artifact of our analysis and is due to the fact that a lower bound on the exact posterior was needed to invoke Gaussian anti-concentration results which were not needed in the approximate sampling regime due to the design of the proposed sampling algorithm. We empirically validate these results in numerical experiments at the end of the chapter.

9.3 Detailed Proofs of for the Convergence of Langevin Algorithms

In this section we supply the proofs of concentration for approximate samples from both the ULA and SGLD MCMC methods. In particular we quantify the computation complexity of generating samples which are distributed close enough to the posterior. To do so, we require a slightly stronger assumption on the family of likelihoods for the MCMC sampling methods to converge.

Assumption 16 (Assumption on the family $p_a(X|\theta_a)$: strengthened for approximate sampling)**.** *Assume that* $\log p_a(x|\theta_a)$ *is* L_a*-smooth and* m_a*-strongly concave over the parameter*

θ_a:

$$-\log p_a(x|\theta'_a) - \nabla_\theta \log p_a(x|\theta'_a)^\top (\theta_a - \theta'_a) + \frac{m_a}{2}\|\theta_a - \theta'_a\|^2 \leq -\log p_a(x|\theta_a)$$
$$\leq -\log p_a(x|\theta'_a) - \nabla_\theta \log p_a(x|\theta'_a)^\top (\theta_a - \theta'_a) + \frac{L_a}{2}\|\theta_a - \theta'_a\|^2, \quad \forall \theta_a, \theta'_a \in \mathbb{R}^{d_a}, x \in \mathbb{R}.$$

Before presenting our proofs, we first include a table summarizing the notation we use within Algorithm 3.

Symbol	Meaning
N	number of steps of the approximate sampling algorithm
$h^{(n)}$	step size of the approximate sampling algorithm after n samples from the arm
$\theta_{ih^{(n)}}$	MCMC sample generated within i-th iteration of Algorithm 3
$\mu_{ih^{(n)}}$	measure of $\theta_{ih^{(n)}}$
k	batch-size of the stochastic gradient Langevin algorithm

Convergence of the Unadjusted Langevin Algorithm (ULA)

We first prove tight bounds on the approximation error of ULA. If function $\log p_a(x;\theta)$ satisfies the Lipschitz smoothness condition in Assumption 11, then we can leverage gradient based MCMC algorithms to generate samples with convergence guarantees in the p-Wasserstein distance. As stated in Algorithm 3, we initialize ULA in the n-th round from the last iterate in the $(n-1)$-th round. Given this, we prove Theorem 20, which is restated below.

Theorem (ULA Convergence)**.** *Assume that the likelihood* $\log p_a(x;\theta)$ *and prior* π_a *satisfy Assumption 14 and Assumption 13. We take step size* $h^{(n)} = \frac{1}{32}\frac{m_a}{n\left(L_a+\frac{1}{n}L_a\right)^2} = \mathcal{O}\left(\frac{1}{nL_a\kappa_a}\right)$ *and number of steps* $N = 640\frac{\left(L_a+\frac{1}{n}L_a\right)^2}{m_a^2} = \mathcal{O}\left(\kappa_a^2\right)$ *in Algorithm 3. If the posterior distribution satisfy the concentration inequality that* $\mathbb{E}_{\theta\sim\mu_a^{(n)}}\left[\|\theta-\theta^*\|^p\right]^{\frac{1}{p}} \leq \frac{1}{\sqrt{n}}\widetilde{D}$*, then for any positive even integer* p*, we have convergence of the ULA algorithm in* W_p *distance to the posterior* $\mu_a^{(n)}$*:* $W_p\left(\widehat{\mu}_a^{(n)}, \mu_a^{(n)}\right) \leq \frac{2}{\sqrt{n}}\widetilde{D}$*,* $\forall \widetilde{D} \geq \sqrt{\frac{32}{m_a}d_a p}$*.*

Proof of Theorem 20. We use induction to prove this theorem.

- For $n = 1$, we initialize at θ_0 which is within a $\sqrt{\frac{d_a}{m_a}}$-ball from the maximum of the target distribution, $\theta_p^* = \arg\max p_a(\theta|x_1)$, where $p_a(\theta|x_1) \propto p_a(x_1|\theta)\pi_a(\theta)$ and negative $\log p_a(\theta|x_1)$ is m_a-strongly convex and $(L_a + L_a)$-Lipschitz smooth. Invoking Lemma 16, we obtain that for $\mathrm{d}\mu_a^{(1)} = p_a(\theta|x_1)\mathrm{d}\theta$, Wasserstein-$p$ distance between the target distribution and the point mass at its mode: $W_p\left(\mu_a^{(1)}, \delta\left(\theta_p^*\right)\right) \leq 5\sqrt{\frac{1}{m_a}d_a p}$. Therefore, $W_p\left(\mu_a^{(1)}, \delta\left(\theta_0\right)\right) \leq W_p\left(\mu_a^{(1)}, \delta\left(\theta_p^*\right)\right) + \left\|\theta_0 - \theta_p^*\right\| \leq 6\sqrt{\frac{1}{m_a}d_a p}$. We then

invoke Lemma 12, with initial condition $\mu_0 = \delta\left(\theta_p^*\right)$, to obtain the convergence in the N-th iteration of Algorithm 3 after the first pull to arm a:

$$W_p^p\left(\mu_{Nh^{(1)}}, \mu_a^{(1)}\right) \leq \left(1 - \frac{m_a}{8}h^{(1)}\right)^{p\cdot N} W_p^p\left(\delta\left(\theta_0\right), \mu_a^{(1)}\right) + 2^{5p}\frac{\left(L_a + L_a\right)^p}{m_a^p}\left(d_a p\right)^{p/2}\left(h^{(1)}\right)^{p/2},$$

where we have substituted in the strong convexity m_a for $\widehat{m}$ and the Lipschitz smoothness $(L_a + L_a)$ for $\widehat{L}$. Plugging in the step size,

$$h^{(1)} = \frac{1}{32}\frac{m_a}{\left(L_a + L_a\right)^2} \leq \min\left\{\frac{m_a}{32\left(L_a + L_a\right)^2}, \frac{1}{1024}\frac{m_a^2}{\left(L_a + L_a\right)^2}\frac{\widetilde{D}^2}{d_a p}\right\},$$

and number of steps $N = \frac{20}{m_a}\frac{1}{h^{(1)}} = 640\frac{\left(L_a + L_a\right)^2}{m_a^2}$, $W_p^p\left(\widehat{\mu}_a^{(1)}, \mu_a^{(1)}\right) = W_p^p\left(\mu_{Nh^{(1)}}, \mu_a^{(1)}\right) \leq 2\widetilde{D}^p$.

- Assume that after the $(n-1)$-th pull and before the n-th pull to the arm a, the ULA algorithm guarantees that $W_p\left(\widehat{\mu}_a^{(n-1)}, \mu_a^{(n-1)}\right) \leq \frac{2}{\sqrt{n-1}}\widetilde{D}$. We now prove that after the n-th pull and before the $(n+1)$-th pull, it is guaranteed that $W_p\left(\widehat{\mu}_a^{(n)}, \mu_a^{(n)}\right) \leq \frac{2}{\sqrt{n}}\widetilde{D}$. We first obtain from the assumed posterior concentration inequality:

$$W_p(\mu_a^{(n)}, \delta\left(\theta^*\right)) \leq \mathbb{E}_{\theta\sim\mu_a^{(n)}}\left[\|\theta - \theta^*\|^p\right]^{\frac{1}{p}} \leq \frac{1}{\sqrt{n}}\widetilde{D}. \tag{9.1}$$

Therefore, for $n \geq 2$,

$$W_p\left(\mu_a^{(n)}, \mu_a^{(n-1)}\right) \leq W_p(\mu_a^{(n)}, \delta\left(\theta^*\right)) + W_p(\mu_a^{(n-1)}, \delta\left(\theta^*\right)) \leq \frac{3}{\sqrt{n}}\widetilde{D}.$$

We combine this bound with the induction hypothesis and obtain that

$$W_p\left(\mu_a^{(n)}, \widehat{\mu}_a^{(n-1)}\right) \leq W_p\left(\mu_a^{(n)}, \mu_a^{(n-1)}\right) + W_p\left(\mu_a^{(n-1)}, \widehat{\mu}_a^{(n-1)}\right) \leq \frac{8}{\sqrt{n}}\widetilde{D}.$$

From Lemma 12, we know that for $\widehat{m} = n \cdot m_a$ and $\widehat{L} = n \cdot L_a + L_a$, with initial condition $\mu_0 = \widehat{\mu}_a^{(n-1)}$, with accurate gradient,

$$W_p^p\left(\mu_{ih^{(n)}}, \mu_a^{(n)}\right) \leq \left(1 - \frac{\widehat{m}}{8}h^{(n)}\right)^{p\cdot i} W_p^p\left(\widehat{\mu}_a^{(n-1)}, \mu_a^{(n)}\right) + 2^{5p}\frac{\widehat{L}^p}{\widehat{m}^p}\left(d_a p\right)^{p/2}\left(h^{(n)}\right)^{p/2}.$$

If we take step size $h^{(n)} = \frac{1}{32}\frac{\widehat{m}}{\widehat{L}^2} \leq \min\left\{\frac{\widehat{m}}{32\widehat{L}^2}, \frac{1}{1024}\frac{1}{n}\frac{\widehat{m}^2}{\widehat{L}^2}\frac{\widetilde{D}^2}{d_a p}\right\}$ and number of steps taken in the ULA algorithm from $(n-1)$-th pull until the n-th pull to be: $\widehat{N} \geq \frac{20}{\widehat{m}}\frac{1}{h^{(n)}}$,

$$W_p^p\left(\widehat{\mu}_a^{(n)}, \mu_a^{(n)}\right) = W_p^p\left(\mu_{\widehat{N}h^{(n)}}, \mu_a^{(n)}\right) \tag{9.2}$$

$$\leq \left(1 - \frac{\widehat{m}}{8}h^{(n)}\right)^{p\cdot\widehat{N}}\frac{8^p\widetilde{D}^p}{n^{p/2}} + 2^{5p}\frac{\widehat{L}^p}{\widehat{m}^p}\left(d_a p\right)^{p/2}\left(h^{(n)}\right)^{p/2} \tag{9.3}$$

$$\leq \frac{2\widetilde{D}^p}{n^{p/2}}, \tag{9.4}$$

leading to the result that $W_p\left(\widehat{\mu}_a^{(n)}, \mu_a^{(n)}\right) \leq \frac{2}{\sqrt{n}}\widetilde{D}$.

Since at least one round would have passed from the $(n-1)$-th pull to the n-th pull to arm a, taking number of steps in each round t to be $N = \frac{20}{\widehat{m}}\frac{1}{h^{(n)}} = 640\frac{\left(L_a+\frac{1}{n}L_a\right)^2}{m_a^2}$ suffices.

Therefore, $N = 640\frac{\left(L_a+\frac{1}{n}L_a\right)^2}{m_a^2} = \mathcal{O}\left(\frac{L_a^2}{m_a^2}\right)$. □

Convergence of the stochastic gradient Langevin algorithm (SGLD)

If $\log p_a(x;\theta)$ satisfies a stronger joint Lipschitz smoothness condition in Assumption 15, similar guarantees can be obtained for stochastic gradient MCMC algorithms.

Theorem 23 (SGLD Convergence)**.** *Assume that the family* $\log p_a(x;\theta)$ *and prior* π_a *satisfy Assumption 14, Assumption 13, and Assumption 15. We take number of data samples in the stochastic gradient estimate* $k = 32\frac{(L_a^*)^2}{m_a\nu_a} = 32\kappa_a^2$*, step size* $h^{(n)} = \frac{1}{32}\frac{m_a}{n\left(L_a+\frac{1}{n}L_a\right)^2} = \mathcal{O}\left(\frac{1}{nL_a\kappa_a}\right)$ *and number of steps* $N = 1280\frac{\left(L_a+\frac{1}{n}L_a\right)^2}{m_a^2} = \mathcal{O}\left(\kappa_a^2\right)$ *in Algorithm 3. If the posterior distribution satisfy the concentration inequality that* $\mathbb{E}_{\theta\sim\mu_a^{(n)}}\left[\|\theta-\theta^*\|^p\right]^{\frac{1}{p}} \leq \frac{1}{\sqrt{n}}\widetilde{D}$*, then for any positive even integer* p*, we have convergence of the ULA algorithm in* W_p *distance to the posterior* $\mu_a^{(n)}$*:* $W_p\left(\widehat{\mu}_a^{(n)}, \mu_a^{(n)}\right) \leq \frac{2}{\sqrt{n}}\widetilde{D}$*,* $\forall\widetilde{D} \geq \sqrt{\frac{32}{m_a}d_a p}$.

Proof of Theorem 23. Similar to Theorem 20, we use induction to prove this theorem. After the first pull to arm a, we take the same $640\frac{\left(L_a+\frac{1}{n}L_a\right)^2}{m_a^2}$ number of steps to converge to $W_p^p\left(\widehat{\mu}_a^{(1)}, \mu_a^{(1)}\right) \leq 2\widetilde{D}^p$.

Assume that after the $(n-1)$-th pull and before the n-th pull to the arm a, the SGLD algorithm guarantees that $W_p\left(\widehat{\mu}_a^{(n-1)}, \mu_a^{(n-1)}\right) \leq \frac{2}{\sqrt{n-1}}\widetilde{D}$. We prove that after the n-th pull and before the $(n+1)$-th pull, it is guaranteed that $W_p\left(\widehat{\mu}_a^{(n)}, \mu_a^{(n)}\right) \leq \frac{2}{\sqrt{n}}\widetilde{D}$. Following the proof of Theorem 20, we combine the assumed posterior concentration inequality and the induction hypothesis to obtain:

$$W_p\left(\mu_a^{(n)}, \widehat{\mu}_a^{(n-1)}\right) \leq W_p\left(\mu_a^{(n)}, \mu_a^{(n-1)}\right) + W_p\left(\mu_a^{(n-1)}, \widehat{\mu}_a^{(n-1)}\right) \leq \frac{8}{\sqrt{n}}\widetilde{D}.$$

Denote function U as the negative log-posterior density over parameter θ. From Lemma 12, we know that for $\widehat{m} = n\cdot m_a$ and $\widehat{L} = n\cdot L_a + L_a$, with initial condition that $\mu_0 = \widehat{\mu}_a^{(n-1)}$, if the difference between the stochastic gradient $\nabla\widehat{U}$ and the exact one ∇U is bounded as

$\mathbb{E}\left[\left\|\nabla U(\theta)-\nabla\widehat{U}(\theta)\right\|^p|\theta\right]\leq\Delta_p$, then

$$W_p^p\left(\mu_{ih^{(n)}},\mu_a^{(n)}\right)\leq\left(1-\frac{\widehat{m}}{8}h^{(n)}\right)^{p\cdot i}W_p^p\left(\widehat{\mu}_a^{(n-1)},\mu_a^{(n)}\right)+2^{5p}\frac{\widehat{L}^p}{\widehat{m}^p}\left(d_ap\right)^{p/2}\left(h^{(n)}\right)^{p/2}+2^{2p+3}\frac{\Delta_p}{\widehat{m}^p}.$$

We demonstrate in the following Lemma 11 that

$$\Delta_p\leq 2\frac{n^{p/2}}{k^{p/2}}\left(\frac{\sqrt{d_ap}L_a^*}{\sqrt{\nu_a}}\right)^p.$$

Lemma 11. *Denote $\widehat{U}$ as the stochastic estimator of U. Then for stochastic gradient estimate with k data points,*

$$\mathbb{E}\left[\left\|\nabla\widehat{U}(\theta)-\nabla U(\theta)\right\|^p|\theta\right]\leq 2\frac{n^{p/2}}{k^{p/2}}\left(\frac{\sqrt{d_ap}L_a^*}{\sqrt{\nu_a}}\right)^p.$$

If we take the number of samples in the stochastic gradient estimator $k=32\frac{(L_a^*)^2}{m_a\nu_a}$, then $\Delta_p\leq\frac{2}{32^{p/2}}\left(n\cdot m_a\right)^{p/2}\cdot\left(p\cdot d_a\right)^{p/2}\leq 2^{-2p-5}\frac{\widehat{m}^p\widetilde{D}^p}{n^{p/2}}$ for any $p\geq 2$. Consequently, $2^{2p+3}\frac{\Delta_p}{\widehat{m}^p}\leq\frac{1}{4}\frac{\widetilde{D}^p}{n^{p/2}}$.

If we take step size $h^{(n)}=\frac{1}{32}\frac{\widehat{m}}{\widehat{L}^2}\leq\min\left\{\frac{\widehat{m}}{32\widehat{L}^2},\frac{1}{1024}\frac{1}{n}\frac{\widehat{m}^2}{\widehat{L}^2}\frac{\widetilde{D}^2}{d_ap}\right\}$ and number of steps taken in the SGLD algorithm from $(n-1)$-th pull till n-th pull to be: $\widehat{N}\geq\frac{40}{\widehat{m}}\frac{1}{h^{(n)}}$,

$$\begin{aligned}W_p^p\left(\widehat{\mu}_a^{(n)},\mu_a^{(n)}\right)&=W_p^p\left(\mu_{\widehat{N}h^{(n)}},\mu_a^{(n)}\right)\\&\leq\left(1-\frac{\widehat{m}}{8}h^{(n)}\right)^{p\cdot\widehat{N}}\frac{8^p\widetilde{D}^p}{n^{p/2}}+2^{5p}\frac{\widehat{L}^p}{\widehat{m}^p}\left(d_ap\right)^{p/2}\left(h^{(n)}\right)^{p/2}+2^{2p+3}\frac{\Delta_p}{\widehat{m}^p}\\&\leq\frac{2\widetilde{D}^p}{n^{p/2}},\end{aligned}$$

leading to the result that $W_p\left(\widehat{\mu}_a^{(n)},\mu_a^{(n)}\right)\leq\frac{2}{\sqrt{n}}\widetilde{D}$. Since at least one round would have past from the $(n-1)$-th pull to the n-th pull to arm a, taking number of steps in each round t to be $N=\frac{40}{\widehat{m}}\frac{1}{h^{(n)}}$ suffices.

Therefore, $N=1280\frac{\left(L_a+\frac{1}{n}L_a\right)^2}{m_a^2}=\mathcal{O}\left(\frac{L_a^2}{m_a^2}\right)$. □

Proof of Lemma 11. We first develop the expression:

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla U(\theta)-\nabla\widehat{U}(\theta)\right\|^p\right]&=n^p\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\nabla\log p(x_i|\theta_a)-\frac{1}{k}\sum_{j=1}^k\nabla\log p(x_j|\theta_a)\right\|^p\right]\\&=\frac{n^p}{k^p}\mathbb{E}\left[\left\|\sum_{j=1}^k\left(\frac{1}{n}\sum_{i=1}^n\nabla\log p(x_i|\theta_a)-\nabla\log p(x_j|\theta_a)\right)\right\|^p\right].\end{aligned}$$

We note that

$$\nabla \log p(x_j|\theta_a) - \frac{1}{n}\sum_{i=1}^{n} \nabla \log p(x_i|\theta_a) = \frac{1}{n}\sum_{i\neq j} \left(\nabla \log p(x_j|\theta_a) - \nabla \log p(x_i|\theta_a)\right).$$

By the joint Lipschitz smoothness Assumption 15, we know that $\nabla \log p(x|\theta_a)$ is a Lipschitz function of x:

$$\left\|\nabla \log p(x_j|\theta_a) - \nabla \log p(x_i|\theta_a)\right\| \leq L_a^* \left\|x_j - x_i\right\|.$$

On the other hand, the data x follows the true distribution $p(x;\theta^*)$, which by Assumption 12 is ν_a-strongly log-concave. Applying Theorem 3.16 in [191], we obtain that

$$\left(\nabla \log p(x_j|\theta_a) - \nabla \log p(x_i|\theta_a)\right)$$

is $\frac{2L_a^*}{\sqrt{\nu_a}}$-sub-Gaussian. Leveraging the Azuma-Hoeffding inequality for martingale difference sequences [191], we obtain that sum of the $(n-1)$ sub-Gaussian random variables:

$$\left(\nabla \log p(x_j|\theta_a) - \frac{1}{n}\sum_{i=1}^{n} \nabla \log p(x_i|\theta_a)\right),$$

is $\frac{2\sqrt{n-1}L_a^*}{n\sqrt{\nu_a}}$-sub-Gaussian. In the same vein, $\left(\sum_{j=1}^{k}\left(\frac{1}{n}\sum_{i=1}^{n} \nabla \log p(x_i|\theta_a) - \nabla \log p(x_j|\theta_a)\right)\right)$ is $\frac{2\sqrt{k(n-1)}L_a^*}{n\sqrt{\nu_a}}$-sub-Gaussian. We then invoke the $\frac{2\sqrt{d_a k(n-1)}L_a^*}{n\sqrt{\nu_a}}$-sub-Gaussianity of

$$\left\|\sum_{j=1}^{k}\left(\frac{1}{n}\sum_{i=1}^{n} \nabla \log p(x_i|\theta_a) - \nabla \log p(x_j|\theta_a)\right)\right\|$$

and have

$$\mathbb{E}\left[\left\|\sum_{j=1}^{k}\left(\frac{1}{n}\sum_{i=1}^{n} \nabla \log p(x_i|\theta_a) - \nabla \log p(x_j|\theta_a)\right)\right\|^p\right] \leq 2\left(\frac{2\sqrt{d_a k(n-1)p}L_a^*}{en\sqrt{\nu_a}}\right)^p.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla U(\theta) - \nabla \widehat{U}(\theta)\right\|^p\right] &= \frac{n^p}{k^p}\mathbb{E}\left[\left\|\sum_{j=1}^{k}\left(\frac{1}{n}\sum_{i=1}^{n} \nabla \log p(x_i|\theta_a) - \nabla \log p(x_j|\theta_a)\right)\right\|^p\right] \\ &\leq 2\frac{n^{p/2}}{k^{p/2}}\left(\frac{2\sqrt{d_a p}L_a^*}{e\sqrt{\nu_a}}\right)^p \leq 2\frac{n^{p/2}}{k^{p/2}}\left(\frac{\sqrt{d_a p}L_a^*}{\sqrt{\nu_a}}\right)^p.\end{aligned}$$

□

Concentration of Approximate Samples from the (Stochastic Gradient) Langevin Algorithm

In this section, we examine convergence of the (stochastic gradient) Langevin algorithm to the posterior distribution over a-th arm at the n-th round. Since only the a-th arm and n-th round are considered, we drop these two indices in the notation whenever suitable. We also define some notation that will only be used within this subsection. For example, we focus on the θ parameter and denote the posterior measure $\mathrm{d}\mu_a^{(n)}(x;\theta) = \mathrm{d}\mu^*(\theta) = \exp\left(-U(\theta)\right)\mathrm{d}\theta$ as the target distribution.

Symbol	Meaning
μ^*	posterior distribution, μ_a^n
U	potential (i.e., negative log posterior density)
θ_U^*	minimum of the potential U (or mode of the posterior μ^*)
θ_t	interpolation between $\theta_{ih^{(n)}}$ and $\theta_{(i+1)h^{(n)}}$, for $t \in [ih^{(n)}, (i+1)h^{(n)}]$
μ_t	measure associated with θ_t
θ_t^*	an auxiliary stochastic process with initial distribution μ^* which follows dynamics (9.9)
$\widehat{m}$	strong convexity of the potential U, nm_a
$\widehat{L}$	Lipschitz smoothness of the potential U, $nL_a + L_a$

We also formally define the Wasserstein-p distance used in the main text. Given a pair of distributions μ and ν on $\mathbb{R}^d$, a *coupling* γ is a joint distribution over the product space $\mathbb{R}^d \times \mathbb{R}^d$ that has μ and ν as its marginal distributions. We let $\Gamma(\mu,\nu)$ denote the space of all possible couplings of μ and ν. With this notation, the Wasserstein-p distance is given by

$$W^p(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p \,\mathrm{d}\gamma(x,y). \tag{9.5}$$

We use the following (stochastic gradient) Langevin algorithm to generate approximate samples from the posterior distribution $\mu_a^{(n)}(\theta)$ at n-th round. For $i = 0, \cdots, T$,

$$\theta_{(i+1)h^{(n)}} \sim \mathcal{N}\left(\theta_{ih^{(n)}} - h^{(n)}\nabla\widehat{U}(\theta_{ih^{(n)}}), 2h^{(n)}\mathrm{I}\right), \tag{9.6}$$

where $\nabla\widehat{U}(\theta_{ih^{(n)}})$ is a stochastic estimate of $\nabla U(\theta_{ih^{(n)}})$. We prove in the following Lemma 12 the convergence of this algorithm within n-th round.

Lemma 12. *Assume that the potential U is $\widehat{m}$-strongly convex and $\widehat{L}$-Lipschitz smooth. Further assume that the p-th moment between the true gradient and the stochastic one satisfies:*

$$\mathbb{E}\left[\left\|\nabla U(\theta_{ih^{(n)}}) - \nabla\widehat{U}(\theta_{ih^{(n)}})\right\|^p \middle| \theta_{ih^{(n)}}\right] \le \Delta_p.$$

Then at i-th step, for $\mu_{ih^{(n)}}$ following the (stochastic gradient) Langevin algorithm with $h \leq \frac{\widehat{m}}{32\widehat{L}^2}$,

$$W_p^p\left(\mu_{ih^{(n)}}, \mu^*\right) \leq \left(1 - \frac{\widehat{m}}{8}h^{(n)}\right)^{p\cdot i} W_p^p\left(\mu_0, \mu^*\right) + 2^{5p}\frac{\widehat{L}^p}{\widehat{m}^p}\left(dp\right)^{p/2}\left(h^{(n)}\right)^{p/2} + 2^{2p+3}\frac{\Delta_p}{\widehat{m}^p}. \quad (9.7)$$

Remark 8. *When $\Delta_p = 0$, Lemma 12 provides convergence rate of the unadjusted Langevin algorithm (ULA) with the exact gradient.*

Proof of Lemma 12. We first interpolate a continuous time stochastic process, θ_t, between $\theta_{ih^{(n)}}$ and $\theta_{(i+1)h^{(n)}}$. For $t \in [ih^{(n)}, (i+1)h^{(n)}]$,

$$\mathrm{d}\theta_t = \nabla\widehat{U}(\theta_{ih^{(n)}})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad (9.8)$$

where B_t is standard Brownian motion. This process connects $\theta_{ih^{(n)}}$ and $\theta_{(i+1)h^{(n)}}$ and approximates the following stochastic differential equation which maintains the exact posterior distribution:

$$\mathrm{d}\theta_t^* = \nabla U(\theta_t^*)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t. \quad (9.9)$$

For a θ_t^* initialized from μ^* and following equation (9.9), θ_t^* will always have distribution μ^*.

We therefore design a coupling between the two processes: θ_t and θ_t^*, where θ_t follows equation (9.8) (and thereby interpolates Algorithm 3) and θ_t^* initializes from μ^* and follows equation (9.9) (and thereby preserves μ^*). By studying the difference between the two processes, we will obtain the convergence rate in terms of the Wasserstein-p distance.

For $t = ih^{(n)}$, we let $\theta_{ih^{(n)}}$ to couple optimally with $\theta^*_{ih^{(n)}}$, so that for

$$\left(\theta_{ih^{(n)}}, \theta^*_{ih^{(n)}}\right) \sim \gamma^* \in \Gamma_{opt}\left(\mu_{ih^{(n)}}, \mu^*_{ih^{(n)}}\right),$$

$\mathbb{E}\left[\left\|\theta_{ih^{(n)}} - \theta^*_{ih^{(n)}}\right\|^p\right] = W_p^p\left(\mu_{ih^{(n)}}, \mu^*\right)$. For $t \in [ih^{(n)}, (i+1)h^{(n)}]$, we choose a synchronous coupling $\bar{\gamma}\left(\theta_t, \theta_t^* | \theta_{ih^{(n)}}, \theta^*_{ih^{(n)}}\right) \in \Gamma\left(\mu_t(\theta_t|\theta_{ih^{(n)}}), \mu_t^*(\theta_t^*|\theta_{ih^{(n)}})\right)$ for the laws of θ_t and θ_t^*. (A synchonous coupling simply means that we use the same Brownian motion B_t in defining θ_t

and θ_t^*.) We then obtain that for any pair $(\theta_t, \theta_t^*) \sim \bar{\gamma}$,

$$\begin{aligned}
\frac{\mathrm{d}\|\theta_t - \theta_t^*\|^p}{\mathrm{d}t} &= \|\theta_t - \theta_t^*\|^{p-2} \left\langle \theta_t - \theta_t^*, \frac{\mathrm{d}\theta_t}{\mathrm{d}t} - \frac{\mathrm{d}\theta_t^*}{\mathrm{d}t} \right\rangle \\
&= p\|\theta_t - \theta_t^*\|^{p-2} \left\langle \theta_t - \theta_t^*, -\nabla U(\theta_t) + \nabla U(\theta_t^*) \right\rangle \\
&+ p\|\theta_t - \theta_t^*\|^{p-2} \left\langle \theta_t - \theta_t^*, \nabla U(\theta_t) - \nabla \widehat{U}(\theta_{ih^{(n)}}) \right\rangle \\
&\leq -p\widehat{m}\, \|\theta_t - \theta_t^*\|^p + p\, \|\theta_t - \theta_t^*\|^{p-1} \left\| \nabla U(\theta_t) - \nabla \widehat{U}(\theta_{ih^{(n)}}) \right\| \qquad (9.10) \\
&\leq -p\widehat{m}\, \|\theta_t - \theta_t^*\|^p \qquad (9.11) \\
&\quad + p \left(\frac{p-1}{p} \left(\frac{p\widehat{m}}{2(p-1)} \right) \|\theta_t - \theta_t^*\|^p + \frac{1}{p} \frac{1}{\left(\frac{p\widehat{m}}{2(p-1)} \right)^{p-1}} \left\| \nabla U(\theta_t) - \nabla \widehat{U}(\theta_{ih^{(n)}}) \right\|^p \right) \qquad (9.12) \\
&\leq -\frac{p\widehat{m}}{2} \|\theta_t - \theta_t^*\|^p + \frac{2^{p-1}}{\widehat{m}^{p-1}} \left\| \nabla U(\theta_t) - \nabla \widehat{U}(\theta_{ih^{(n)}}) \right\|^p, \qquad (9.13)
\end{aligned}$$

where equation (9.12) follows from Young's inequality.

Equivalently, we can obtain

$$\frac{\mathrm{d}e^{\frac{p\widehat{m}}{2}t}\|\theta_t - \theta_t^*\|^p}{\mathrm{d}t} \leq e^{\frac{p\widehat{m}}{2}t} \frac{2^{p-1}}{\widehat{m}^{p-1}} \left\| \nabla U(\theta_t) - \nabla \widehat{U}(\theta_{ih^{(n)}}) \right\|^p.$$

By the fundamental theorem of calculus,

$$\|\theta_t - \theta_t^*\|^p \leq e^{-\frac{p\widehat{m}}{2}\left(t - ih^{(n)}\right)} \left\| \theta_{ih^{(n)}} - \theta_{ih^{(n)}}^* \right\|^p + \frac{2^{p-1}}{\widehat{m}^{p-1}} \int_{ih^{(n)}}^t e^{-\frac{p\widehat{m}}{2}(t-s)} \left\| \nabla U(\theta_s) - \nabla \widehat{U}(\theta_{ih^{(n)}}) \right\|^p \mathrm{d}s. \qquad (9.14)$$

Taking expectation on both sides, we obtain that

$$\begin{aligned}
\mathbb{E}\left[\|\theta_t - \theta_t^*\|^p\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\theta_t - \theta_t^*\|^p \mid \theta_{ih^{(n)}}, \theta_{ih^{(n)}}^*\right]\right] \\
&\leq e^{-\frac{p\widehat{m}}{2}\left(t - ih^{(n)}\right)} \mathbb{E}\left[\left\|\theta_{ih^{(n)}} - \theta_{ih^{(n)}}^*\right\|^p\right] \\
&+ \frac{2^{p-1}}{\widehat{m}^{p-1}} \int_{ih^{(n)}}^t e^{-\frac{p\widehat{m}}{2}(t-s)} \mathbb{E}\left[\left\| \nabla U(\theta_s) - \nabla \widehat{U}(\theta_{ih^{(n)}}) \right\|^p\right] \mathrm{d}s. \qquad (9.15)
\end{aligned}$$

In the above expression, the integral and expectation are exchanged using Tonelli's theorem, since

$$\left\| \nabla U(\theta_s) - \nabla \widehat{U}(\theta_{ih^{(n)}}) \right\|^p$$

is positive measurable.

We further expand the expected error $\mathbb{E}\left[\left\|\nabla U(\theta_s) - \nabla \widehat{U}(\theta_{ih^{(n)}})\right\|^p\right]$:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla U(\theta_s) - \nabla \widehat{U}(\theta_{ih^{(n)}})\right\|^p\right] \\
&= \mathbb{E}\left[\left\|\nabla U(\theta_s) - \nabla U(\theta_{ih^{(n)}}) + \nabla U(\theta_{ih^{(n)}}) - \nabla \widehat{U}(\theta_{ih^{(n)}})\right\|^p\right] \\
&\leq \frac{1}{2}\mathbb{E}\left[\|2\left(\nabla U(\theta_s) - \nabla U(\theta_{ih^{(n)}})\right)\|^p\right] + \frac{1}{2}\mathbb{E}\left[\left\|2\left(\nabla U(\theta_{ih^{(n)}}) - \nabla \widehat{U}(\theta_{ih^{(n)}})\right)\right\|^p\right] \\
&= 2^{p-1}\mathbb{E}\left[\|\nabla U(\theta_s) - \nabla U(\theta_{ih^{(n)}})\|^p\right] + 2^{p-1}\mathbb{E}\left[\mathbb{E}\left[\left\|\nabla U(\theta_{ih^{(n)}}) - \nabla \widehat{U}(\theta_{ih^{(n)}})\right\|^p \middle| \theta_{ih^{(n)}}\right]\right] \\
&\leq 2^{p-1}\widehat{L}^p \cdot \mathbb{E}\left[\|\theta_s - \theta_{ih^{(n)}}\|^p\right] + 2^{p-1}\Delta_p. \qquad (9.16)
\end{aligned}$$

Plugging into equation (9.14), we have that

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_t - \theta_t^*\|^p\right] \\
&\leq e^{-\frac{p\widehat{m}}{2}\left(t - ih^{(n)}\right)}\mathbb{E}\left[\left\|\theta_{ih^{(n)}} - \theta^*_{ih^{(n)}}\right\|^p\right] \\
&+ 2^{2p-2}\frac{\widehat{L}^p}{\widehat{m}^{p-1}}\int_{ih^{(n)}}^t e^{-\frac{p\widehat{m}}{2}(t-s)}\mathbb{E}\left[\|\theta_s - \theta_{ih^{(n)}}\|^p\right]\mathrm{d}s + 2^{2p-2}(t - ih^{(n)})\frac{\Delta_p}{\widehat{m}^{p-1}}. \qquad (9.17)
\end{aligned}$$

We provide an upper bound for $\int_{ih^{(n)}}^t e^{-\frac{p\widehat{m}}{2}(t-s)}\mathbb{E}\left[\|\theta_s - \theta_{ih^{(n)}}\|^p\right]\mathrm{d}s$ in the following lemma.

Lemma 13. *For $h^{(n)} \leq \frac{\widehat{m}}{32\widehat{L}^2}$, and for $t \in [ih^{(n)}, (i+1)h^{(n)}]$,*

$$\begin{aligned}
&\int_{ih^{(n)}}^t e^{-\frac{p\widehat{m}}{2}(t-s)}\mathbb{E}\left[\|\theta_s - \theta_{ih^{(n)}}\|^p\,\mathrm{d}s\right] \\
&\leq 2^{3p-3}\widehat{L}^p\left(t - ih^{(n)}\right)^{p+1} W_p^p\left(\mu_{ih^{(n)}}, \mu^*\right) + \frac{8^p}{2}\left(t - ih^{(n)}\right)^{p/2+1}(dp)^{p/2} + 2^{2p-2}(t - ih^{(n)})^{p+1}\cdot\Delta_p. \qquad (9.18)
\end{aligned}$$

Applying this upper bound to equation (9.17), we obtain that for $h^{(n)} \leq \frac{\widehat{m}}{32\widehat{L}^2}$, and for $t \in [ih^{(n)}, (i+1)h^{(n)}]$,

$$\begin{aligned}
\mathbb{E}\left[\|\theta_t - \theta_t^*\|^p\right] &\leq e^{-\frac{p\widehat{m}}{2}\left(t - ih^{(n)}\right)}\mathbb{E}\left[\left\|\theta_{ih^{(n)}} - \theta^*_{ih^{(n)}}\right\|^p\right] + 2^{5p-5}\frac{\widehat{L}^{2p}}{\widehat{m}^{p-1}}\left(t - ih^{(n)}\right)^{p+1} W_p^p\left(\mu_{ih^{(n)}}, \mu^*\right) \\
&\quad + 2^{5p-3}\frac{\widehat{L}^p}{\widehat{m}^{p-1}}\left(t - ih^{(n)}\right)^{p/2+1}(dp)^{p/2} + 2^{4p-4}\frac{\widehat{L}^p}{\widehat{m}^{p-1}}(t - ih^{(n)})^{p+1}\cdot\Delta_p \\
&\quad + 2^{2p-2}(t - ih^{(n)})\frac{\Delta_p}{\widehat{m}^{p-1}} \\
&\leq \left(1 - \frac{\widehat{m}}{4}\left(t - ih^{(n)}\right)\right)^p \mathbb{E}\left[\left\|\theta_{ih^{(n)}} - \theta^*_{ih^{(n)}}\right\|^p\right] + 2^{5p-5}\frac{\widehat{L}^{2p}}{\widehat{m}^{p-1}}\left(t - ih^{(n)}\right)^{p+1} W_p^p\left(\mu_{ih^{(n)}}, \mu^*\right) \\
&\quad + 2^{5p-3}\frac{\widehat{L}^p}{\widehat{m}^{p-1}}\left(t - ih^{(n)}\right)^{p/2+1}(dp)^{p/2} + 2^{2p}(t - ih^{(n)})\frac{\Delta_p}{\widehat{m}^{p-1}}.
\end{aligned}$$

Recognizing that $\widehat{\gamma}\left(\theta_t, \theta_t^*\right) = \mathbb{E}_{\left(\theta_{ih^{(n)}}, \theta^*_{ih^{(n)}}\right)\sim\gamma^*}\left[\bar{\gamma}\left(\theta_t, \theta_t^* | \theta_{ih^{(n)}}, \theta^*_{ih^{(n)}}\right)\right]$ is a coupling, we achieve the upper bound for $W_p^p\left(\mu_t, \mu^*\right)$:

$$
\begin{aligned}
W_p^p\left(\mu_t, \mu^*\right) &\leq \mathbb{E}_{(\theta_t, \theta_t^*)\sim\widehat{\gamma}}\left[\|\theta_t - \theta_t^*\|^p\right] \\
&\leq \left(1 - \frac{\widehat{m}}{4}\left(t - ih^{(n)}\right)\right)^p \mathbb{E}_{\left(\theta_{ih^{(n)}}, \theta^*_{ih^{(n)}}\right)\sim\gamma^*}\left[\left\|\theta_{ih^{(n)}} - \theta^*_{ih^{(n)}}\right\|^p\right] \\
&\quad + 2^{5p-5}\frac{\widehat{L}^{2p}}{\widehat{m}^{p-1}}\left(t - ih^{(n)}\right)^{p+1} W_p^p\left(\mu_{ih^{(n)}}, \mu^*\right) + 2^{5p-3}\frac{\widehat{L}^p}{\widehat{m}^{p-1}}\left(t - ih^{(n)}\right)^{p/2+1}(dp)^{p/2} \\
&\quad + 2^{2p}(t - ih^{(n)})\frac{\Delta_p}{\widehat{m}^{p-1}}. \\
&\leq \left(1 - \frac{\widehat{m}}{8}\left(t - ih^{(n)}\right)\right)^p W_p^p\left(\mu_{ih^{(n)}}, \mu^*\right) + 2^{5p-3}\frac{\widehat{L}^p}{\widehat{m}^{p-1}}\left(t - ih^{(n)}\right)^{p/2+1}(dp)^{p/2} \qquad (9.19) \\
&\quad + 2^{2p}(t - ih^{(n)})\frac{\Delta_p}{\widehat{m}^{p-1}}. \qquad (9.20)
\end{aligned}
$$

Taking $t = (i+1)h^{(n)}$, the recurring bound reads

$$
W_p^p\left(\mu_{(i+1)h^{(n)}}, \mu^*\right) \leq \left(1 - \frac{\widehat{m}}{8}h^{(n)}\right)^p W_p^p\left(\mu_{ih^{(n)}}, \mu^*\right) + 2^{5p-3}\frac{\widehat{L}^p}{\widehat{m}^{p-1}}(dp)^{p/2}\left(h^{(n)}\right)^{p/2+1} + \frac{4^p}{\widehat{m}^{p-1}}h^{(n)}\Delta_p.
$$

We finish the proof by invoking the recursion i times:

$$
\begin{aligned}
W_p^p\left(\mu_{ih^{(n)}}, \mu^*\right) &\leq \left(1 - \frac{\widehat{m}}{8}h^{(n)}\right)^p W_p^p\left(\mu_{(i-1)h^{(n)}}, \mu^*\right) + 2^{5p-3}\frac{\widehat{L}^p}{\widehat{m}^{p-1}}(dp)^{p/2}\left(h^{(n)}\right)^{p/2+1} + \frac{4^p}{\widehat{m}^{p-1}}h^{(n)}\Delta_p \\
&\leq \left(1 - \frac{\widehat{m}}{8}h^{(n)}\right)^{p\cdot i} W_p^p\left(\mu_0, \mu^*\right) \\
&\quad + \sum_{k=0}^{i-1}\left(1 - \frac{\widehat{m}}{8}h^{(n)}\right)^{p\cdot k}\cdot\left(2^{5p-3}\frac{\widehat{L}^p}{\widehat{m}^{p-1}}(dp)^{p/2}\left(h^{(n)}\right)^{p/2+1} + \frac{4^p}{\widehat{m}^{p-1}}h^{(n)}\Delta_p\right) \\
&\leq \left(1 - \frac{\widehat{m}}{8}h^{(n)}\right)^{p\cdot i} W_p^p\left(\mu_0, \mu^*\right) + 2^{5p}\frac{\widehat{L}^p}{\widehat{m}^p}(dp)^{p/2}\left(h^{(n)}\right)^{p/2} + 2^{2p+3}\frac{\Delta_p}{\widehat{m}^p}. \qquad (9.21)
\end{aligned}
$$

□

Supporting proofs for Lemma 12

Proof of Lemma 13. We use the update rule of ULA to develop $\int_{ih^{(n)}}^{t} e^{-\frac{p\widehat{m}}{2}(t-s)}\mathbb{E}\left[\|\theta_s - \theta_{ih^{(n)}}\|^p\right]\mathrm{d}s$:

$$
\begin{aligned}
&\int_{ih^{(n)}}^{t} e^{-\frac{p\widehat{m}}{2}(t-s)}\mathbb{E}\left[\|\theta_s - \theta_{ih^{(n)}}\|^p\,\mathrm{d}s\right] \\
&= \int_{ih^{(n)}}^{t} e^{-\frac{p\widehat{m}}{2}(t-s)}\mathbb{E}\left[\left\|-(s-ih^{(n)})\left(\nabla U(\theta_{ih^{(n)}}) - \left(\nabla U(\theta_{ih^{(n)}}) - \nabla\widehat{U}(\theta_{ih^{(n)}})\right)\right) + \sqrt{2}(B_s - B_{ih^{(n)}})\right\|^p\right]\mathrm{d}s \\
&\leq 2^{2p-2}(t-ih^{(n)})^p \int_{ih^{(n)}}^{t} e^{-\frac{p\widehat{m}}{2}(t-s)}\mathbb{E}\left[\|\nabla U(\theta_{ih^{(n)}})\|^p\right]\mathrm{d}s \\
&+ 2^{3p/2-1}\int_{ih^{(n)}}^{t} e^{-\frac{p\widehat{m}}{2}(t-s)}\mathbb{E}\left[\|B_s - B_{ih^{(n)}}\|^p\right]\mathrm{d}s \\
&+ 2^{2p-2}(t-ih^{(n)})^p \int_{ih^{(n)}}^{t} e^{-\frac{p\widehat{m}}{2}(t-s)}\mathbb{E}\left[\left\|\nabla U(\theta_{ih^{(n)}}) - \nabla\widehat{U}(\theta_{ih^{(n)}})\right\|^p\right]\mathrm{d}s \\
&\leq 2^{2p-2}\widehat{L}^p\left(t-ih^{(n)}\right)^{p+1}\mathbb{E}\left[\|\theta_{ih^{(n)}} - \theta_U^*\|^p\right] + 2^{3p/2-1}\int_{ih^{(n)}}^{t}\mathbb{E}\left[\|B_s - B_{ih^{(n)}}\|^p\right]\mathrm{d}s \\
&\qquad + 2^{2p-2}\left(t-ih^{(n)}\right)^{p+1}\Delta_p. \qquad (9.22)
\end{aligned}
$$

where θ_U^* is the fixed point of U. We then use the following lemma to simplify the above expression.

Lemma 14. *The integrated p-th moment of the Brownian motion can be bounded as:*

$$\int_{ih^{(n)}}^{t}\mathbb{E}\left\|B_s - B_{ih^{(n)}}\right\|^p\mathrm{d}s \leq 2\left(\frac{dp}{e}\right)^{p/2}\left(t-ih^{(n)}\right)^{p/2+1}. \qquad (9.23)$$

We also provide bound for the p-th moment of $\|\theta_{ih^{(n)}} - \theta_U^*\|$.

Lemma 15. *For $\theta_{ih^{(n)}} \sim \mu_{ih^{(n)}}$,*

$$\mathbb{E}\left\|\theta_{ih^{(n)}} - \theta_U^*\right\|^p \leq 2^{p-1}W_p^p\left(\mu_{ih^n}, \mu^*\right) + \frac{10^p}{2}\left(\frac{dp}{\widehat{m}}\right)^{p/2}. \qquad (9.24)$$

Plugging the results into equation (9.22), we obtain that for $h^{(n)} \leq \frac{\widehat{m}}{32\widehat{L}^2}$, and for $t \in$

$[ih^{(n)}, (i+1)h^{(n)}]$,

$$\begin{aligned}
&\int_{ih^{(n)}}^{t} e^{-\frac{p\widehat{m}}{2}(t-s)}\mathbb{E}\left[\|\theta_s - \theta_{ih^{(n)}}\|^p \,\mathrm{d}s\right] \\
&\leq 2^{3p-3}\widehat{L}^p\left(t-ih^{(n)}\right)^{p+1} W_p^p\left(\mu_{ih^n}, \mu^*\right) + \frac{40^p}{8}\widehat{L}^p\left(t-ih^{(n)}\right)^{p+1}\left(\frac{dp}{\widehat{m}}\right)^{p/2} \\
&+ \left(\frac{8}{e}\right)^{p/2}(dp)^{p/2}\left(t-ih^{(n)}\right)^{p/2+1} + 2^{2p-2}(t-ih^{(n)})^{p+1}\cdot\Delta_p \\
&\leq 2^{3p-3}\widehat{L}^p\left(t-ih^{(n)}\right)^{p+1} W_p^p\left(\mu_{ih^n}, \mu^*\right) + \frac{8^p}{2}\left(t-ih^{(n)}\right)^{p/2+1}(dp)^{p/2} + 2^{2p-2}(t-ih^{(n)})^{p+1}\Delta_p.
\end{aligned} \tag{9.25}$$

□

Proof of Lemma 14. The Brownian motion term can be upper bounded by higher moments of a normal random variable:

$$\int_{ih^{(n)}}^{t}\mathbb{E}\left\|B_s - B_{ih^{(n)}}\right\|^p \mathrm{d}s \leq \left(t-ih^{(n)}\right)\mathbb{E}\left\|B_t - B_{ih^{(n)}}\right\|^p = \left(t-ih^{(n)}\right)^{p/2+1}\mathbb{E}\left\|v\right\|^p,$$

where v is a standard d-dimensional normal random variable. We then invoke the $\sqrt{d}$ sub-Gaussianity of $\|v\|$ and have (assuming p to be an even integer):

$$\mathbb{E}\left\|v\right\|^p \leq \frac{p!}{2^{p/2}\,(p/2)!}d^{p/2} \leq \frac{e^{1/12p}\sqrt{2\pi p}(p/e)^p}{2^{p/2}\sqrt{\pi p}(p/2e)^{p/2}}d^{p/2} \leq 2\left(\frac{dp}{e}\right)^{p/2}.$$

□

Proof of Lemma 15. For the $\mathbb{E}\left\|\theta_{ih^{(n)}} - \theta_U^*\right\|^p$ term, we note that any coupling of a distribution with a delta measure is their product measure. Therefore, $\mathbb{E}\left\|\theta_{ih^{(n)}} - \theta_U^*\right\|^p$ relates to the p-Wasserstein distance between $\mu_{ih^{(n)}}$ and the delta measure at the fixed point θ_U^*, $\delta\left(\theta_U^*\right)$:

$$\begin{aligned}
\mathbb{E}\left\|\theta_{ih^{(n)}} - \theta_U^*\right\|^p = W_p^p\left(\mu_{ih^{(n)}}, \delta\left(\theta_U^*\right)\right) &\leq \left(W_p\left(\mu_{ih^{(n)}}, \mu^*\right) + W_p\left(\mu^*, \delta\left(\theta_U^*\right)\right)\right)^p \\
&\leq 2^{p-1}W_p^p\left(\mu_{ih^{(n)}}, \mu^*\right) + 2^{p-1}W_p^p\left(\mu^*, \delta\left(\theta_U^*\right)\right).
\end{aligned}$$

We then bound $W_p^p\left(\mu^*, \delta\left(\theta_U^*\right)\right)$ in the following lemma.

Lemma 16. *Assume the posterior μ^* is $\widehat{m}$-strongly log-concave. Then for $\theta_U^* = \arg\max \mu^*$,*

$$W_p^p\left(\mu^*, \delta\left(\theta_U^*\right)\right) \leq 5^p\left(\frac{dp}{\widehat{m}}\right)^{p/2}. \tag{9.26}$$

Therefore,

$$\mathbb{E}\left\|\theta_{ih^{(n)}}^{(n)} - \theta_n^*\right\|^p \leq 2^{p-1}W_p^p\left(\mu_{ih^{(n)}}, \mu^*\right) + \frac{10^p}{2}\left(\frac{dp}{\widehat{m}}\right)^{p/2}.$$

□

Proof of Lemma 16. We first decompose $W_p\left(\mu^*, \delta\left(\theta_U^*\right)\right)$ into two terms:

$$W_p\left(\mu^*, \delta\left(\theta_U^*\right)\right) \leq W_p\left(\mu^*, \delta\left(\mathbb{E}_{\theta\sim\mu^*}\left[\theta\right]\right)\right) + \left\|\theta_U^* - \mathbb{E}_{\theta\sim\mu^*}\left[\theta\right]\right\|.$$

By the celebrate relation between mean and mode for 1-unimodal distributions [see, e.g., 20, Theorem 7], we can first bound the difference between mean and mode:

$$\left(\theta_U^* - \mathbb{E}_{\theta\sim\mu^*}\left[\theta\right]\right)^T \Sigma^{-1} \left(\theta_U^* - \mathbb{E}_{\theta\sim\mu^*}\left[\theta\right]\right) \leq 3.$$

where Σ is the covariance matrix of μ^*. Therefore,

$$\left\|\theta_U^* - \mathbb{E}_{\theta\sim\mu^*}\left[\theta\right]\right\|^2 \leq \frac{3}{\widehat{m}}. \tag{9.27}$$

We then bound $W_p\left(\mu^*, \delta\left(\mathbb{E}_{\theta\sim\mu^*}\left[\theta\right]\right)\right)$. Since the coupling between μ^* and the delta measure $\delta\left(\mathbb{E}_{\theta\sim\mu^*}\left[\theta\right]\right)$ is their product measure, we can directly obtain that the p-Wasserstein distance is the p-th moments of μ^*:

$$W_p^p\left(\mu^*, \delta\left(\mathbb{E}_{\theta\sim\mu^*}\left[\theta\right]\right)\right) = \int \left\|\theta - \mathbb{E}_{\theta\sim\mu^*}\left[\theta\right]\right\|^p \mathrm{d}\mu^*(\theta).$$

We invoke the Herbst argument [see, e.g., 95] to obtain the p-th moment bound. We first note that for an $\widehat{m}$-strongly log-concave distribution, it has a log Sobolev constant of $\widehat{m}$. Then using the Herbst argument, we know that $x \sim \mu^*$ is a sub-Gaussian random vector with parameter $\sigma^2 = \frac{1}{2\widehat{m}}$:

$$\int e^{\lambda u^T\left(\theta - \mathbb{E}_{\theta\sim\mu^*}[\theta]\right)} \mathrm{d}\mu^*(\theta) \leq e^{\frac{\lambda^2}{4\widehat{m}}}, \quad \forall \left\|u\right\| = 1.$$

Hence θ is $2\sqrt{\frac{d}{\widehat{m}}}$ norm-sub-Gaussian, which implies that

$$\left(\mathbb{E}_{\theta\sim\mu^*}\left[\left\|\theta - \mathbb{E}_{\theta\sim\mu^*}\left[\theta\right]\right\|^p\right]\right)^{1/p} \leq 2e^{1/e}\sqrt{\frac{dp}{\widehat{m}}}. \tag{9.28}$$

Combining equations (9.27) and (9.28), we obtain the final result that

$$\begin{aligned} W_p^p\left(\mu^*, \delta\left(\theta_U^*\right)\right) &\leq \left(2e^{1/e}\sqrt{\frac{dp}{\widehat{m}}} + \sqrt{\frac{3}{\widehat{m}}}\right)^p \\ &\leq 5^p \left(\frac{dp}{\widehat{m}}\right)^{p/2}. \end{aligned}$$

□

To conclude this section we provide the results which guarantee concentration of the approximate samples resulting from ULA and SGLD.

Lemma 17. *Assume that the likelihood* $\log p_a(x;\theta)$*, prior distribution, and true distributions satisfy Assumptions 1-3, and that arm* a *has been chosen* $n = T_a(t)$ *times up to iteration* t *of the Thompson sampling algorithm. Further, assume that we choose the stepsize step size* $h^{(n)} = \frac{1}{32}\frac{m_a}{n\left(L_a+\frac{1}{n}L_a\right)^2} = \mathcal{O}\left(\frac{m_a}{nL_a^2}\right)$ *, and number of steps* $N = 640\frac{\left(L_a+\frac{1}{n}L_a\right)^2}{m_a^2} = \mathcal{O}\left(\frac{L_a^2}{m_a^2}\right)$ *in Algorithm 3 then for* $\delta_2 \in (0, e^{-1/2})$*:*

$$\mathbb{P}_{\theta_{a,t}\sim\bar{\mu}_a^{(n)}[\gamma_a]}\left(\|\theta_{a,t}-\theta_a^*\|_2 > \sqrt{\Gamma}\middle| Z_{n-1}\right) < \delta_2,$$

where,

$$\Gamma = \frac{36e}{m_a n}\left(d_a + \log B_a + 2\sigma\log 1/\delta_1 + 2\left(\sigma_a + \frac{m_a d_a}{18L_a\gamma_a}\right)\log 1/\delta_2\right)$$

and $Z_{t-1} = \{\|\theta_{a,t-1}-\theta_a^*\| \le C(n)\}$ *for:*

$$C(n) = \sqrt{\frac{18e}{nm_a}}\left(d_a + \log B_a + 2\sigma\log 1/\delta_1\right)^{\frac{1}{2}},$$

$\sigma = 16 + \frac{4d_aL_a^2}{\nu_a m_a}$*, and where* $\theta_{a,t-1}$ *is the sample from the previous round of the Thompson sampling algorithm for arm* a*.*

Proof. We begin as in the proof of Theorem 21, except that we now take $\mu_0 = \delta_{\theta_{a,t-1}}$, where $\theta_{a,t-1}$ is the sample from the previous step of the algorithm:

$$W_p^p\left(\mu_{ih^{(n)}}, \mu_a^{(n)}\right) \le \left(1 - \frac{\widehat{m}}{8}h^{(n)}\right)^{p\cdot i} W_p^p\left(\delta(\theta_{a,t-1}), \mu_a^{(n)}\right) + \frac{80^p}{2}\frac{\widehat{L}^p}{\widehat{m}^p}\left(dp\right)^{p/2}\left(h^{(n)}\right)^{p/2}.$$

We first use the triangle inequality on the first term on the RHS:

$$\begin{aligned} W_p\left(\delta(\theta_{a,t-1}), \mu_a^{(n)}\right) &\le W_p\left(\delta(\theta_{a,t-1}), \delta_{\theta_a^*}\right) + W_p\left(\delta(\theta_a^*), \mu_a^{(n)}\right) \\ &= \|\theta_a^* - \theta_{a,t-1}\| + + W_p\left(\delta(\theta_a^*), \mu_a^{(n)}\right) \\ &\le C(n) + \frac{\tilde{D}}{\sqrt{n}}, \end{aligned}$$

where we have used the fact that $\|\theta_a^* - \theta_{a,t-1}\| \le C(n)$ by assumption, and the definition of $\tilde{D}$ from the proof of Theorem 20: $\widetilde{D} = \sqrt{\frac{2}{m_a}}\left(d_a + \log B_a + \sigma p\right)^{\frac{1}{2}}$.

Since:

$$C(n) = \sqrt{\frac{18e}{m_a}}\left(d_a + \log B_a + 2\sigma\log 1/\delta_1\right)^{\frac{1}{2}},$$

we can further expand this upper bound:

$$\begin{aligned} W_p\left(\delta_{\theta_{a,t-1}}, \mu_a^{(n)}\right) &\leq \frac{\tilde{D}}{\sqrt{n}} + C(n) \\ &\leq 8\sqrt{\frac{2}{m_a n}}\left(d_a + \log B_a + 2\sigma \log 1/\delta_1 + \sigma p\right)^{\frac{1}{2}}, \end{aligned}$$

where to derive this result we have used the fact that $\sqrt{2(x+y)} \geq \sqrt{x} + \sqrt{y}$.

Letting $\bar{D} = \sqrt{\frac{2}{m_a n}}\left(d_a + \log B_a + 2\sigma \log 1/\delta_1 + \sigma p\right)^{\frac{1}{2}}$, we see that our final result is:

$$W_p\left(\delta_{\theta_{a,t-1}}, \mu_a^{(n)}\right) \leq \frac{8}{\sqrt{n}}\bar{D},$$

where $\tilde{D} < \bar{D}$. Using the same choice of $h^{(n)}$ and number of steps N as in the proof or Theorem 20 guarantees us that:

$$W_p^p\left(\mu_{ih^{(n)}}, \mu_a^{(n)}\right) \leq 2\left(\frac{\bar{D}}{\sqrt{n}}\right)^p.$$

Further combining this with the triangle inequality, and the fact that $\tilde{D} < \bar{D}$ gives us that:

$$W_p\left(\mu_{ih^{(n)}}, \delta_{\theta^*}\right) \leq \frac{\tilde{D}}{\sqrt{n}} + \frac{\bar{D}}{\sqrt{n}} \leq 3\frac{\bar{D}}{\sqrt{n}}.$$

Now, since the sample returned by the Langevin algorithm is given by:

$$\theta_a = \theta_N + Z, \tag{9.29}$$

where $Z \sim \mathcal{N}\left(0, \frac{1}{nL_a\gamma_a} I\right)$, it remains to bound the distance between the approximate posterior $\hat{\mu}_a^{(n)}$ of θ_a and the distribution of $\theta_{Nh^{(n)}}$. Since $\theta_a - \theta_{Nh^{(n)}} = Z$, for any even integer p, we find that:

$$\begin{aligned} W_p^p\left(\bar{\mu}_a^{(n)}, \bar{\mu}_a^{(n)}[\gamma_a]\right) &= \left(\inf_{\gamma \in \Gamma\left(\bar{\mu}_a^{(n)}, \bar{\mu}_a^{(n)}[\gamma_a]\right)} \int \|\theta_a - \theta_N\|^p \, \mathrm{d}\theta_a \mathrm{d}\theta_N\right)^{1/p} \\ &\leq \mathbb{E}[\|Z\|^p]^{\frac{1}{p}} \\ &\leq \sqrt{\frac{d}{nL_a\gamma_a}}\left(\frac{2^{p/2}\mathbf{\Gamma}(\frac{p+1}{2})}{\sqrt{\pi}}\right)^{1/p} \\ &\leq \sqrt{\frac{d}{nL_a\gamma_a}}\left(2^{p/2}\left(\frac{p}{2}\right)^{p/2}\right)^{1/p} \\ &\leq \sqrt{\frac{dp}{nL_a\gamma_a}}, \end{aligned}$$

where we have used upper bound of the Stirling type for the Gamma function $\mathbf{\Gamma}(\cdot)$ in the second last inequality.

Thus, we have, via the triangle inequality once again, that:

$$W_p\left(\bar{\mu}_a^{(n)[\gamma_a]}, \delta_{\theta^*}\right) \leq 3\frac{\bar{D}}{\sqrt{n}} + \sqrt{\frac{dp}{nL_a\gamma_a}}$$
$$\leq \sqrt{\frac{36}{m_a n}}\left(d_a + \log B_a + 2\sigma_a \log 1/\delta_1 + \left(\sigma_a + \frac{d_a}{18L_a\gamma_a}\right)p\right)^{\frac{1}{2}},$$

which, by the same derivation as in the proof of Theorem 18, gives us the desired result. □

We remark that via an identical argument, the following Lemma holds as well:

Lemma 18. *Assume that the family $\log p_a(x;\theta)$ and the prior π_a satisfy Assumptions 1-15 and that arm a has been chosen $n = T_a(t)$ times up to iteration t of the Thompson sampling algorithm. If we take number of data samples in the stochastic gradient estimate $k = 32\frac{(L_a^*)^2}{m_a\nu_a}$, step size $h^{(n)} = \frac{1}{32}\frac{m_a}{n\left(L_a+\frac{1}{n}L_a\right)^2} = \mathcal{O}\left(\frac{m_a}{nL_a^2}\right)$ and number of steps $N = 1280\frac{\left(L_a+\frac{1}{n}L_a\right)^2}{m_a^2} = \mathcal{O}\left(\frac{L_a^2}{m_a^2}\right)$ in Algorithm 3, then for $\delta_2 \in (0, e^{-1/2})$::*

$$\mathbb{P}_{\theta_{a,t}\sim\bar{\mu}_a^{(n)}[\gamma_a]}\left(\|\theta_{a,t} - \theta_a^*\|_2 > \sqrt{\Gamma}\middle| Z_{n-1}\right) < \delta_2.,$$

where,

$$\Gamma = \frac{36e}{m_a n}\left(d_a + \log B_a + 2\sigma\log 1/\delta_1 + 2\left(\sigma_a + \frac{m_a d_a}{18L_a\gamma_a}\right)\log 1/\delta_2\right)$$

and $Z_{t-1} = \{\|\theta_{a,t-1} - \theta_a^*\| \leq C(n)\}$ for the parameters:

$$C(n) = \sqrt{\frac{18e}{nm_a}}\left(d_a + \log B_a + 2\sigma\log 1/\delta_1\right)^{\frac{1}{2}}, \qquad \sigma = 16 + \frac{4d_a L_a^2}{\nu_a m_a},$$

and $\theta_{a,t-1}$ being the sample from the previous round of the Thompson sampling algorithm over arm a.

9.4 Detailed Proofs of the Regret of Approximate Sampling

For the proof of Theorem 22, we proceed similarly as for the proof of Theorem 19, but require another intermediate lemma to deal with the fact that the samples from the arms are no longer conditionally independent given the filtration (due to the fact that we use the last sample as the initialization of the filtration). To do so, we first define the event:

$$Z_a(T) = \cap_{t=1}^{T-1} Z_{a,t},$$

where:

$$Z_{a,t} = \left\{ \|\theta_{a,t} - \theta_a^*\| < \sqrt{\frac{18e}{nm_a}} \left(d_a + \log B_a + 2\left(16 + \frac{4dL_a^2}{\nu_a m_a}\right) \log 1/\delta_1 \right)^{\frac{1}{2}} \right\},$$

Lemma 19. *Suppose the likelihood and reward distributions satisfy Assumptions 1-4, Then the regret of a Thompson sampling algorithm with approximate sampling can be decomposed as:*

$$\mathbb{E}[R(T)] \leq \sum_{a>1} \Delta_a \mathbb{E}\left[T_a(T) \,\middle|\, Z_a(T) \cap Z_1(T) \right] + 2\Delta_a \tag{9.30}$$

Proof. We begin by conditioning on the event $Z_a(T) \cap Z_1(T)$ for each $a \in \mathcal{A}$, where we note that by construction $p_Z = \mathbb{P}((Z_a(T)^c \cup Z_1(T)^c)) \leq \mathbb{P}(Z_1(T)^c) + \mathbb{P}(Z_a(T)^c) = 2T\delta_1)$ (since via Lemma 9, the probability of each event in $Z_a(T)^c$ and $Z_1(T)^c$ is less than δ_1).

Therefore, we must have that:

$$\begin{aligned}
\mathbb{E}[T_a(T)] &\leq \mathbb{E}\left[T_a(T) \,\middle|\, Z_a(T) \cap Z_1(T) \right] + \mathbb{E}\left[T_a(T) \,\middle|\, (Z_a(T)^c \cup Z_1(T)^c) \right] p_Z \\
&\leq \mathbb{E}\left[T_a(T) \,\middle|\, Z_a(T) \cap Z_1(T) \right] + 2T\delta_3 \mathbb{E}\left[T_a(T) \,\middle|\, (Z_a(T)^c \cup Z_1(T)^c) \right] \\
&\leq \mathbb{E}\left[T_a(T) \,\middle|\, Z_a(T) \cap Z_1(T) \right] + 2\delta_3 T^2,
\end{aligned}$$

where in the first line we use the fact that $1 - p_Z \leq 1$ and in the last line we used the fact that $T_a(T)$ is trivially less than T. Choosing $\delta_1 = 1/T^2 \leq e^{-1/2}$ completes the proof. □

With this decomposition in hand, we can now proceed as in Lemma 5 to provide anti-concentration guarantees for the approximate posteriors.

Lemma 20. *Suppose the likelihood and true reward distributions satisfy Assumptions 1-4: then if $\gamma_1 = \frac{\nu m^2}{32(16L\nu m + 4dL^3)}$, for all $n = 1, ..., T$ all samples from the the (stochastic gradient) ULA method with the hyperparameters and runtime as described in Theorem 21 satisfy:*

$$\mathbb{E}\left[\frac{1}{p_{1,n}} \right] \leq 27\sqrt{B_1}$$

Proof. We begin by using the last step of our Langevin Dynamics and show that it exhibits the desired anti-concentration properties. In particular, we know that $\theta_{1,t} \sim \mathcal{N}(\theta_{1,Nh}, \frac{1}{\gamma} I)$, such that:

$$p_{1,s} = Pr\left(\alpha^T(\theta - \theta_{1,Nh}) \geq \alpha^T(\theta^* - \theta_{1,Nh}) - \epsilon\right)$$
$$\geq Pr\left(Z \geq \underbrace{A\|\theta_{1,Nh} - \theta_*\|}_{:=t}\right)$$

where $Z \sim \mathcal{N}(0, \frac{A^2}{nL\gamma}I)$ by construction.

Now using a lower bound on the cumulative density function of a Gaussian random variable, we find that, for $\sigma^2 = \frac{A^2}{nL\gamma}$:

$$p_{1,s} \geq \sqrt{\frac{1}{2\pi}} \begin{cases} \frac{\sigma t}{t^2+\sigma^2} e^{-\frac{t^2}{2\sigma^2}} & : t > \frac{A}{\sqrt{nL\gamma}} \\ 0.34 & : t \leq \frac{A}{\sqrt{nL\gamma}} \end{cases}$$

Thus we have that:

$$\frac{1}{p_{1,s}} \leq \sqrt{2\pi} \begin{cases} \left(\frac{t}{\sigma} + 1\right) e^{\frac{t^2}{2\sigma^2}} & : t > \frac{A}{\sqrt{nL\gamma}} \\ 3 & : t \leq \frac{A}{\sqrt{nL\gamma}} \end{cases}$$

Taking the expectation of both sides with respect to the samples $X_1, ..., X_n$, we find that:

$$\mathbb{E}\left[\frac{1}{p_{1,s}}\right] \leq 3\sqrt{2\pi} + \sqrt{2\pi}\mathbb{E}\left[\left(\sqrt{nL\gamma}\|\theta_{1,Nh} - \theta_*\| + 1\right) e^{nL\gamma\|\theta_{1,Nh} - \theta_*\|^2}\right]$$
$$\leq 3\sqrt{2\pi} + \sqrt{2\pi nL\gamma}\sqrt{\mathbb{E}\left[\|\theta_{1,Nh} - \theta_*\|^2\right]}\sqrt{\mathbb{E}\left[e^{nL\gamma\|\theta_{1,Nh} - \theta_*\|^2}\right]} + \sqrt{2\pi}\mathbb{E}\left[e^{\frac{nL\gamma}{2}\|\theta_{1,Nh} - \theta_*\|^2}\right]$$

Now, we remark that, from Theorems 20 and 23, we have that for both approximate sampling schemes:

$$\mathbb{E}\left[\|\theta_{1,Nh} - \theta^*\|^2\right] \leq \frac{18}{mn}\left(d + \log B + 32 + \frac{8dL^2}{\nu m}\right)$$

Further, we note that $\|\theta_{1,Nh} - \theta^*\|^2$ is a sub-exponential random variable. To see this, we analyze its moment generating function:

$$\mathbb{E}[e^{nL\gamma\|\theta_{1,Nh} - \theta^*\|^2}] = 1 + \sum_{i=1}^{\infty} \mathbb{E}\left[\frac{(nL\gamma)^i\|\theta_{1,Nh} - \theta_*\|^{2i}}{i!}\right]$$

Borrowing the notation from the proof of Theorem 18, we know that

$$\mathbb{E}\left[\|\theta_{1,Nh} - \theta_*\|^{2p}\right] \leq 3\left(\frac{2D}{mn} + \frac{4\sigma p}{mn}\right)^p$$

where:

$$D = d + \log B \quad \text{and} \quad \sigma = 16 + \frac{4dL^2}{\nu m}$$

Plugging this in above gives:

$$\begin{aligned}
\mathbb{E}[e^{\gamma\|\theta_{1,Nh}-\theta_*\|^2}] &\leq 1 + 3\sum_{i=1}^{\infty} \frac{\left(\frac{2nL\gamma D + 4nL\gamma\sigma i}{mn}\right)^i}{i!} \\
&\leq 1 + \frac{3}{2}\sum_{i=1}^{\infty}\frac{1}{i!}\left(\frac{4nL\gamma D}{mn}\right)^i + \frac{3}{2}\sum_{i=1}^{\infty}\frac{1}{i!}\left(\frac{8nL\gamma\sigma i}{nm}\right)^i \\
&\leq \frac{3}{2}e^{\frac{4nL\gamma D}{mn}} + \frac{3}{2}\sum_{i=1}^{\infty}\left(\frac{8nL\gamma e\sigma}{nm}\right)^i
\end{aligned}$$

where, we have use the identities $(x+y)^i \leq 2^{i-1}(x^i + y^i)$ for $i \geq 1$, and $i! \geq (i/e)^i$ to simplify the bound.

If $\gamma \leq \frac{m}{32L\sigma}$, then we have that:

$$\mathbb{E}[e^{nL\gamma\|\theta_{1,Nh}-\theta_*\|^2}] \leq \frac{3}{2}\left(e^{\frac{4nL\gamma D}{m}} + 2.5\right),$$

which, together with the upper bound on γ gives:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{p_{1,s}}\right] &\leq 3\sqrt{2\pi} + \frac{3}{2}\sqrt{\frac{16\pi nL\gamma}{m}(D+2\sigma)}\left(e^{\frac{2nL\gamma D}{m}} + 2\right) + \frac{3}{2}\sqrt{2\pi}\left(e^{\frac{4nL\gamma D}{m}} + 7.5\right) \\
&\leq 3\sqrt{2\pi} + \frac{3}{2}\left(\sqrt{\frac{\pi(d+\log B)}{2\sigma}} + \sqrt{\pi}\right)\left(e^{\frac{d+\log B}{16\sigma}} + 2\right) + \frac{3}{2}\sqrt{2\pi}\left(e^{\frac{d+\log B}{8\sigma}} + 2.5\right),
\end{aligned}$$

where we used the sub-additivity of $\sqrt{x}$, the fact that $\sqrt{\frac{3}{2}} < \frac{3}{2}$, $\sqrt{2.5} < 2$ and substituted in the values for σ and D to simplify the bound. Finally since $\frac{L^2}{m\nu} > 1$, we find that $\sigma > \max(4d, 1)$, allowing us to simplify the bound further to:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{p_{1,s}}\right] &\leq 3\sqrt{2\pi} + \frac{3}{2}\sqrt{\frac{\pi}{8} + \frac{\log B}{2}}\left(2B^{1/16} + 2\right) + \frac{3}{2}\sqrt{2\pi}\left(2B^{1/8} + 2.5\right) \\
&\leq 18 + \frac{3}{\sqrt{2}}\underbrace{\left(B^{1/16} + B^{1/16}\sqrt{\log B} + \log B + 2B^{1/8}\right)}_{I} \\
&\leq 18 + 12/\sqrt{2}\sqrt{B} \leq 27\sqrt{B},
\end{aligned}$$

where to simplify the bound we used the fact that $\sqrt{\pi} < 2$ and $I \leq 4\sqrt{B}$ and that $18 + 12/\sqrt{2}x \leq 27x$ for $x \geq 1$.

□

With this lemma in hand, we can now proceed as in Lemma 8 to finalize the proof of Theorem 22.

Lemma 21. *Suppose the likelihood, true reward distributions, and priors satisfy Assumptions 1-4, the samples are generated from the sampling schemes described in Theorem 23 and Theorem 20, and* $\gamma_a = \frac{m_a}{32 L_a \sigma_a}$ *then:*

$$\sum_{s=1}^{T-1} \mathbb{E}\left[\frac{1}{\widehat{p}_{1,s}} - 1 \middle| Z_1(T)\right] \leq 27\sqrt{B_1}\left\lceil \frac{144 e A_1^2}{m\Delta_a^2}(d_1 + \log B_1 + 4\sigma_1 \log T + 12 d_1 \sigma_1 \log 2)\right\rceil + 1 \tag{9.31}$$

$$\sum_{s=1}^{T} \mathbb{E}\left[\mathbb{I}\left(\widehat{p}_{a,s} > \frac{1}{T}\right) \middle| Z_a(T)\right] \leq \frac{144 e A_a^2}{m\Delta_a^2}(d_a + \log B_a + 10 d_a \sigma_a \log(T)), \tag{9.32}$$

where $\widehat{p}_{a,s}$ *is the distribution of a sample from the approximate posterior* $\widehat{\mu}_a$ *after* s *samples have been collected, and for* $a \in \mathcal{A}$, σ_a *is given by:*

$$\sigma_a = 16 + \frac{4 d_a L_a^2}{m_a \nu_a}.$$

Proof. We begin by showing that (9.31) holds. To do so, we proceed identically as in the proof of Lemma 8 to note that, by definition $\widehat{p}_{1,s}$ satisfies:

$$\begin{aligned}
\widehat{p}_{1,s} &= \mathbb{P}(r_{1,t}(s) > \bar{r}_1 - \epsilon | \mathcal{F}_{t-1}) && (9.33)\\
&= 1 - \mathbb{P}(r_{1,t}(s) - \bar{r}_1 < -\epsilon | \mathcal{F}_{t-1}) && (9.34)\\
&\geq 1 - \mathbb{P}(|r_{1,t}(s) - \bar{r}_1| > \epsilon | \mathcal{F}_{t-1}) && (9.35)\\
&\geq 1 - \mathbb{P}_{\theta \sim \widehat{\mu}_1^{(s)}}\left(\|\theta - \theta^*\| > \frac{\epsilon}{A_1}\right), && (9.36)
\end{aligned}$$

where the last inequality follows from the fact that $r_{1,t}(s)$ and $\bar{r}_1$ are A_a-Lipschitz functions of $\theta \sim \mu_1^{(s)}$ and θ^* respectively.

We then use the fact that conditioned on $Z_1(T)$, the approximate posterior distribution $\mathbb{P}_{\theta \sim \widehat{\mu}_1^{(s)}}$ satisfies the identical concentration bounds from Lemmas 18 and Lemma 17. Substituting in the assumed value of γ_1, and simplifying, we have that the distribution of the samples conditioned on $Z_1(T)$ satisfy:

$$\mathbb{P}_{\theta_{1,t} \sim \bar{\mu}_1^{(s)}[\gamma_1]}\left(\|\theta_{1,t} - \theta_1^*\|_2 > \sqrt{\frac{36e}{m_1 n}(d_1 + \log B_1 + 4\sigma_1 \log T + 6 d_1 \sigma_1 \log 1/\delta_2)} \middle| Z_{n-1}\right) < \delta_2.$$

Equivalently, we have that:

$$\mathbb{P}_{\theta\sim\bar{\mu}_1^{(s)}[\gamma_1]}\left(\|\theta-\theta^*\|>\frac{\epsilon}{A_1}\right)\leq exp\left(-\frac{1}{6d_1\sigma_1}\left(\frac{m_1 n\epsilon^2}{36eA_1^2}-\bar{D}_1\right)\right), \tag{9.37}$$

where we define $\bar{D}_1 = d_1 + \log B_1 + 4\sigma\log T$, to simplify notation. We remark that this bound is not useful unless:

$$n > \frac{16eA_1^2}{\epsilon^2 m_1}\bar{D}_1.$$

Thus, choosing $\epsilon = (\bar{r}_1 - \bar{r}_a)/2 = \Delta_a/2$, we can choose ℓ as:

$$\ell = \left\lceil \frac{144eA_1^2}{m\Delta_a^2}(\bar{D}_1 + 6d_1\sigma_1\log 2)\right\rceil.$$

With this choice of ℓ, we proceed exactly as in the proof of Lemma 8 to find that :

$$\begin{aligned}\sum_{s=1}^{T-1}\mathbb{E}\left[\frac{1}{\widehat{p}_{1,s}}-1\middle|Z_1(T)\right] &\leq 27\sqrt{B_1}\ell + \sum_{s=\ell}^{T-1}\mathbb{E}\left[\frac{1}{p_{1,s}}-1\middle|Z_1(T)\right]\\ &\leq 27\sqrt{B_1}\left\lceil\frac{144eA_1^2}{m\Delta_a^2}(\bar{D}_1+12d_1\sigma_1\log 2)\right\rceil+1,\end{aligned}$$

where we used the upper bound from Lemma 20 to bound the first ℓ terms in the first inequality.

To show that (9.32) holds, we use a similar derivation as in (9.36):

$$\sum_{s=1}^{T}\mathbb{E}\left[\mathbb{I}\left(p_{a,s}>\frac{1}{T}\right)\middle|Z_a(T)\right]\leq\sum_{s=1}^{T}\mathbb{E}\left[\mathbb{I}\left(\mathbb{P}_{\theta\sim\bar{\mu}_a^{(s)}[\gamma_a]}\left(\|\theta-\theta^*\|>\frac{\Delta_a}{2A_a}\right)>\frac{1}{T}\right)\middle|Z_a(T)\right]$$

Since on the event $Z_a(T)$, the posterior concentration result from Lemmas 18 and Lemma 17 holds, it remains to upper bound the number of pulls $\bar{n}$ of arm a such that for all $n \geq \bar{n}$:

$$\mathbb{P}_{\theta\sim\bar{\mu}_a^{(n)}[\gamma_a]}\left(\|\theta-\theta^*\|>\frac{\Delta_a}{2A_a}\right)\leq\frac{1}{T}.$$

Since the posterior for arm a after n pulls of arm a has the same form as in (8.14), we can choose $\bar{n}$ as:

$$\bar{n} = \frac{144eA_a^2}{m\Delta_a^2}(\bar{D}_a + 6d_a\sigma_a\log(T)).$$

Using the fact that $d_a >\geq 1$ to simplify the bound completes the proof. □

Putting together the results of Lemmas 19 and 21 gives us our final theorem:

Theorem (Regret of Thompson sampling with (stochastic gradient) Langevin algorithm)**.** *When the likelihood and true reward distributions satisfy Assumptions 1-4: we have that the expected regret after $T > 0$ rounds of Thompson sampling with the (stochastic gradient) ULA method with the hyper-parameters and runtime as described in Lemmas 17 (and 18 respectively), and $\gamma_a = \frac{\nu_a m_a^2}{32(16L_a\nu_a m_a + 4d_a L_a^3)} = O\left(\frac{1}{d_a\kappa_a^3}\right)$ satisfies:*

$$\mathbb{E}[R(T)] \leq \sum_{a>1} \frac{CA_a^2}{m_a\Delta_a}\left(d_a + \log B_a + d_a^2\kappa_a^2 \log T\right) + \frac{C\sqrt{B_1}A_1^2}{m_1\Delta_a}\left(1 + \log B_1 + d_1\kappa_1^2 \log T + d_1^2\kappa_1^2\right) + 3\Delta_a,$$

where C is a universal constant that is independent of problem-dependent parameters and $\kappa_a = L_a/m_a$.

Proof. To begin, we invoke Lemma 19, which shows that we only need to bound the number of times a suboptimal arm $a \in \mathcal{A}$ is chosen on the 'nice' event $Z_1(T) \cap Z_a(T)$ where the gradient of the log likelihood has concentrated and the approximate samples have been in high probability regions of the posteriors. We then invoke Lemmas 6 and 7, to find that:

$$\mathbb{E}\left[T_a(T) \,\middle|\, Z_1(T) \cap Z_a(T)\right] \leq 1 + \ell \tag{9.38}$$

$$+ \underbrace{\sum_{s=\ell}^{T-1} \mathbb{E}\left[\frac{1}{p_{1,s}} - 1 \middle| Z_1(T)\right]}_{(I)} + \underbrace{\sum_{s=1}^{T} \mathbb{E}\left[\mathbb{I}\left(1 - p_{a,s} > \frac{1}{T}\right) \middle| Z_a(T)\right]}_{(II)} \tag{9.39}$$

Now, invoking Lemma 8, we use the upper bounds for terms (I) and (II) in the regret decomposition, use our choice of both δ_1 and $\delta_3 = 1/T^2$, expanding D_a and D_1, and use the fact that $\lceil x \rceil \leq x + 1$ to give that:

$$\begin{aligned}
\mathbb{E}[R(T)] &\leq \sum_{a>1} \frac{144eA_a^2}{m_a\Delta_a}\left(d_a + \log B_a + 10d_a\left(16 + \frac{4d_aL_a^2}{\nu_a m_a}\right)\log(T)\right) \\
&\quad + 27\sqrt{B_1}\frac{144eA_1^2}{m_1\Delta_a}\left(1 + d_1 + \log B_1 + 4\left(16 + \frac{4d_1L_a^2}{\nu_1 m_1}\right)(\log T + 3d_1\log 2)\right) + 3\Delta_a. \\
&\leq \sum_{a>1} \frac{CA_a^2}{m_a\Delta_a}\left(d_a + \log B_a + d_a^2\kappa_a^2\log T\right) \\
&\quad + \frac{C\sqrt{B_1}A_1^2}{m_1\Delta_a}\left(1 + \log B_1 + d_1\kappa_1^2\log T + d_1^2\kappa_1^2\right) + 3\Delta_a.
\end{aligned}$$

Using the fact that $\kappa_a \geq 1$ and that $d_1 \geq 1$ allows us to simplify to get our desired result. □

9.5 Numerical Experiments

In this section we empirically validate the effectiveness of approximate Thompson sampling in log-concave multi-armed bandit instances. We benchmark against both UCB and exact Thompson Sampling across three different Gaussian multi-armed bandit instances with 10 arms. We remark that the use of Gaussian bandit instances is due to the fact that the closed form for the posteriors allows for us to properly benchmark against exact Thompson Sampling and UCB, though our theory applies to a broader family of prior/likelihood pairs.

Experimental Setup

In all three instances we keep the reward distributions for each arm fixed such that their means are evenly spaced from 0 to 10 ($\bar{r}_1 = 1$, $\bar{r}_2 = 2$, and so on), and their variances are all 1. In each instance we use different priors over the means of the arms to analyze whether the approximate Thompson Sampling algorithms preserve the performance of exact Thompson Sampling.

In the first instance, the priors reflect the correct orderings of the means. We use Gaussian priors with variance 4, and means evenly spaced between 5 and 10 such that $\mathbb{E}_{\pi_1}[X] = 5$, and $\mathbb{E}_{\pi_{10}}[X] = 10$. In the second instance, the prior for each arm is a Gaussian with mean 7.5 and variance 4. Finally, the third instance is 'adversarial' in the sense that the priors reflects the complete opposite ordering of the means. In particular, the priors are still Gaussians such that their means are evenly spaced between 5 and 10 with variance 4, but this time $\mathbb{E}_{\pi_1}[X] = 10$, and $\mathbb{E}_{\pi_{10}}[X] = 5$.

As suggested in our theoretical analysis in Section 9, we use a constant number of steps for both ULA and SGLD to generate samples from the approximate posteriors. In particular, for ULA, we take $N = 100$ and double that number for SGLD $N = 200$. We also choose the stepsize for both algorithms to be $\frac{1}{32T_a(t)}$. For SGLD, we use a batch size of $\min(T_a(t), 32)$. Further, since $d_a = \kappa_a = 1$ since this is a Gaussian family, we take the scaling to be $\gamma_a = 1$.

The regret is calculated as $\sum_{t=1}^{T} \bar{r}_{10} - \bar{r}_{A_t}$ for the three algorithms and is averaged across 100 runs. Finally, for the implementation of UCB, we used the time-horizon tuned UCB [94] and the known variance, σ^2 of the arms in the upper confidence bounds (to maintain a level playing field between algorithms):

$$UCB_a(t) = \frac{1}{T_a(t)} \sum_{i=1}^{t-1} X_{A_i} \mathbb{1}\{A_i = a\} + \sqrt{\frac{4\sigma^2 \log 2T}{T_a(t)}}.$$

Empirical Results

We observe significant performance gains from the (approximate) Thompson sampling approach over the deterministic UCB algorithm when the priors are suggestive or even non-informative of the appealing arms. When the priors are adversarial to the algorithm, the

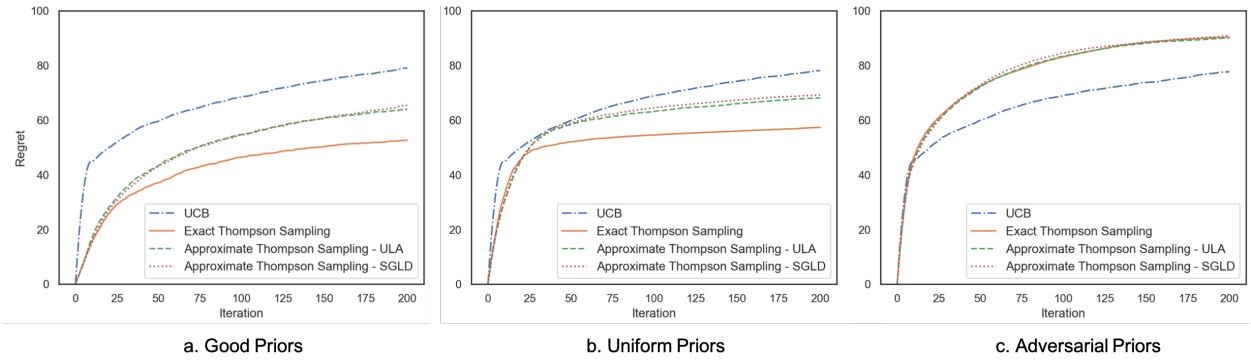


Figure 9.1: Performance of exact and approximate Thompson Sampling vs UCB on Gaussian bandits with a. 'good priors' (priors reflecting the correct ordering of the arms' means), b. the same priors on all the arms' means, and c. 'bad' priors (priors reflecting the exact opposite ordering of the arms' means). Each line is the regret averaged across 100 runs of the algorithm.

UCB algorithm outperforms the Thompson sampling approach as expected. (This case corresponds to the constant B_a in the Theorems 19 and 22 being large). Also as the theory predicts, we observe little difference between the exact and the approximate Thompson sampling methods in terms of the regret. If we zoom in and scrutinize further, we can see that exact Thompson Sampling slightly outperforms the Thompson sampling with SGLD in the 'googd' prior case. This might be due to the added stochasticity from the approximate sampling techniques, which adds unnecessary exploration.

9.6 Chapter Summary

Although Thompson sampling has been used successfully for decades and has been shown to have appealing theoretical properties there remains a lack of understanding of how approximate sampling affects its regret guarantees.

In this chapter we first derived new posterior contraction rates for log-concave likelihood families with arbitrary log-concave priors which captured key dependencies between the posterior distributions and various problem-dependent parameters like the prior quality and the parameter dimension. We then used these rates to show that exact Thompson sampling in MAB problems where the reward distributions are log-concave achieves the optimal finite-time regret guarantee for MAB bandit problems shown in [91]. As a direction for future work, we note that although our regret bound demonstrates a dependence on the quality of the prior, it still is unable to capture the potential advantages of good priors.

We then demonstrated that Thompson sampling using samples generated from ULA, and under slightly stronger assumptions, SGLD, could still achieve the optimal regret guarantee with constant algorithmic as well as sample complexity in the stochastic gradient estimate. Thus, by designing approximate sampling algorithms specifically for use with Thompson

sampling, we were able to construct a computationally tractable anytime Thompson sampling algorithm with approximate samples with end-to-end guarantees of logarithmic regret.

Part III

Learning Models of Human Decision-Making

Chapter 10

Models of Human Decision-Making

The modeling and learning of human decision-making behavior is increasingly becoming important as critical systems begin to rely more on automation and artificial intelligence. Yet, in this task we face a number of challenges, not least of which is the fact that humans are known to behave in ways that are not completely rational. For example, there is mounting evidence to support the fact that humans often use *reference points*—e.g., the *status quo* or former experiences or recent expectations about the future that are otherwise perceived to be related to the decision the human is making [87, 183]. It has also been observed that their decisions are impacted by their perception of the external world (exogenous factors) and their present state of mind (endogenous factors) as well as how the decision is *framed* or presented [184].

The success of *descriptive* behavioral models in capturing human behavior has long been touted by the psychology community and, more recently, by the economics community. In the engineering context, humans have largely been modeled, under rationality assumptions, from the so-called *normative* point of view where things are modeled *as they ought to be*, which is counter to a descriptive *as is* point of view.

However, risk-sensitivity in the context of learning to control stochastic dynamic systems (see, *e.g.*, [27, 58, 111]) has been fairly extensively explored in engineering and computer science. Many of these approaches are targeted at mitigating risks due to uncertainties in controlling a system such as a plant or robot. Much of this work simply handles *risk-aversion* by leveraging techniques such as exponential utility functions or minimizing mean-variance type criteria.

Complex risk-sensitive behavior arising from human interaction with automation is only recently coming into focus. Human decision makers can be at once risk-averse and risk-seeking depending their frame of reference. The adoption of diverse behavioral models in engineering—in particular, in learning and control—is growing due to the fact that humans are increasingly playing an integral role in automation both at the individual and societal scale. Learning accurate models of human decision-making is important for both *prediction* and *description*. For example, control/incentive schemes need to predict human behavior as a function of external stimuli including not only potential disturbances but also the con-

trol/incentive mechanism itself. On the other hand, policy makers and regulatory agencies, e.g., are interested in interpreting human reactions to implemented regulations and policies.

Approaches for integrating the risk-sensitivity in the control and reinforcement learning problems via behavioral models have recently emerged [90, 109, 126, 133, 174]. These approaches largely assume a risk-sensitive Markov decision process (MDP) formulated based on a model that captures behavioral aspects of the human's decision-making process. We refer the problem of learning the optimal policy in this setting as the *forward* problem. We are interested in solving the so-called *inverse* problem which seeks to estimate the decision-making process given a set of demonstrations. In order to do so, a well formulated forward problem with convergence guarantees is required.

Inverse reinforcement learning in the context of recovering policies directly (or indirectly via first learning a representation for the reward) has long been studied in the context expected utility maximization and MDPs [1, 137, 158]. We may care about, e.g., producing the value and reward functions (or at least, characterize the space of these functions) that produce behaviors matching that which is observed. On the other hand, we may want to extract the optimal policy from a set of demonstrations so that we can reproduce the behavior in support of, e.g., designing incentives or control policies. In this paper, our focus is on the combination of these two tasks.

We model human decision-makers as *risk-sensitive Q-learning agents* where we exploit very rich behavioral models from behavioral psychology and economics that capture a whole spectrum of risk-sensitive behaviors and loss aversion. We first derive a reinforcement learning algorithm that leverages coherent risk metrics and behavioral value functions such as those deriving from prospect theory. We provide convergence guarantees via a contraction mapping argument. In comparison to previous work in this area [173], we show that the behavioral value functions we introduce satisfy the assumptions of our theorems.

Given the forward risk-sensitive reinforcement learning algorithm, we propose a gradient-based learning algorithm for inferring the decision-making model parameters from demonstrations—that is, we propose a framework for solving the *inverse risk-sensitive reinforcement learning* problem with theoretical guarantees. We show that the gradient of the loss function with respect to the model parameters is well-defined and computable via a contraction map argument. We demonstrate the efficacy of the learning scheme on the canonical Grid World example and a passenger's view of ride-sharing modeled as an MDP with parameters estimated from real-world data.

10.1 Related Work

Before presenting our results, we first comment on related work. The primary motivation for most other works in this domain is to get a prescriptive model or algorithm for humans amidst autonomy so that the human can be controlled and accounted for. For example, in [174]—one of the motivating previous works for the forward MDP model we use in this paper—their approach to learning the decision-making model is to parameterize unknown quantities of

interest, sample the parameter space, and use a model selection criteria (specifically, the Bayesian information criteria) to select parameters that best fit the observed behavior. We, on the other hand, derive a well-formulated gradient-based procedure for finding the value function and policy best matching the observed behavior. Moreover, in contrast to [174], we introduce new value functions that satisfy our theorems for the forward and inverse problems and retain the salient features of the empirically observed behavioral psychology and economics models.

In [109], the authors take a similar approach to ours in leveraging risk metrics to capture risk sensitivity. However, they focus their efforts on estimating the risk metric used by the human decision maker by leveraging the well-known representation theorem for coherent risk metrics [56]. They couple the resulting optimization problem with classical inverse reinforcement learning procedures for learning the reward (that is, they parameterize the reward function over a set of basis functions), yet their approach does not differentiate between the reward and the decision-making model.

Our approach, in comparison, focuses on estimating the value function and the agent's behavior which also induces the risk metric via the acceptance level set. Specifically, we consider a broad class of risk metrics generated by value functions, formulate the MDP model based on this, and learn the parameters of the value function that generates the risk metric and results in a policy that best matches the agent's observed behavior. The parameters of the value function, which ultimately drive the decision making model and specify the risk measure, are highly interpretable in terms of the degree of risk sensitivity and loss aversion. Thus, our technique supports prescriptive and descriptive analysis, both of which are important for the design of incentives and policies that takes into consideration the nuances of human decision-making behavior.

10.2 Overview of Part III

The remainder of this part of the dissertation is organized as follows. In Chapter 11, we overview the model we assume for risk-sensitive agents and show that it is amenable to integration of behavioral models from prospect theory. We then present our risk-sensitive Q-learning convergence results. In Chapter 12, we formulate the inverse reinforcement learning problem and propose a gradient–based algorithm to solve it. Examples that demonstrate the ability of the proposed scheme to capture a wide breadth of risk-sensitive behaviors are provided in Section 12.2. Finally, we conclude with some discussion and comments on future work in Section 12.3.

Chapter 11

Risk-Sensitive Reinforcement Learning

In order to learn a decision-making model for an agent who faces sequential decisions in an uncertain environment, we leverage a risk-sensitive Q-learning model that integrates coherent risk metrics with behavioral models. In particular, the model we use is based on a model first introduced in [70] and later refined in [126, 174].

The primary difference between the work presented in this chapter and previous work is that we (i) introduce a new prospect theory based value function and (ii) provide a convergence theorem whose assumptions are satisfied for the behavioral models we use. Under the assumption that the agent is making decisions according to this model, in the sequel we formulate a gradient–based method for learning the policy as well as parameters of the agent's value function.

11.1 Markov Decision Processes

Throughout this part of the dissertation we consider a class of finite MDPs consisting of a state space X, an admissible action space $A(x) \subset A$ for each $x \in X$, a transition kernel $P(x'|x,a)$ that denotes the probability of moving from state x to x' given action a, and a reward function[1] $r : X \times A \times W \to \mathbb{R}$ where W is the space of bounded disturbances and has distribution $P_r(\cdot|x,a)$. Including disturbances allows us to model random rewards; we use the notation $R(x',a)$ to denote the random reward having distribution $P_r(\cdot|x,a)$.

In the classical expected utility maximization framework, the agent seeks to maximize the expected discounted rewards by selecting a Markov policy π—that is, for an infinite horizon MDP, the optimal policy is obtained by maximizing

$$J(x_0) = \max_\pi \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t R(x_t, a_t)\right] \tag{11.1}$$

[1]We note that it is possible to consider the more general reward structure $r : X \times A \times X \times W \to \mathbb{R}$, however we exclude this case in order to not further bog down the notation.

where x_0 is the initial state and $\gamma \in (0, 1)$ is the discount factor.

The risk-sensitive reinforcement learning problem transforms the above problem to account for a salient features of the human decision-making process such as loss aversion, reference point dependence, and risk-sensitivity. Specifically, we introduce two key components, *value functions* and *valuation functions*, that allow for our model to capture these features. The former captures risk-sensitivity, loss-aversion, and reference point dependence in its transformation of outcome values to their value as perceived by the agent and the latter generalizes the expectation operator to more general measures of risk—specifically, *coherent risk measures*.

11.2 Value Functions

Given the environmental and reward uncertainties, we model the outcome of each action as a real-valued random variable $Y(i) \in \mathbb{R}$, $i \in I$ where I denotes a finite event space and Y is the outcome of i–th event with probability $\mu(i)$ where $\mu \in \Delta(I)$, the space of probability distributions on I. Much like the standard expected utility framework, an agent makes choices based on the value of their outcome as defined by their *value function* $v : \mathbb{R} \to \mathbb{R}$.

There are a number of existing approaches to defining value functions that capture risk-sensitivity and loss aversion. These approaches derive from a variety of fields including behavioral psychology/economics, mathematical finance, and even neuroscience.

One of the principal features of human decision-making is that losses are perceived more significant that a gain of equal true value. The models with the greatest efficacy in capturing this affect are convex and concave in different regions of the outcome space. Prospect theory, developed by Kahneman and Tvsersky [81, 182], is built on one such model. The form of the value function introduced in prospect theory is given by

$$v(y) = \begin{cases} k_+(y - y_o)^{\zeta_+}, & y > y_o \\ -k_-(y_o - y)^{\zeta_-}, & y \leq y_o \end{cases} \tag{11.2}$$

where y_o is the *reference point* that the decision-maker compares outcomes against in determining if the decision is a loss or gain.

The parameters $(k_+, k_-, \zeta_+, \zeta_-)$ control the degree of loss-aversion and risk-sensitivity. For example, the following are risk preferences for different parameter values:

(i) $0 < \zeta_+, \zeta_- < 1$: risk-averse preferences on gains and risk-seeking preferences on losses (concave in gains, convex in losses);

(ii) $\zeta_+ = \zeta_- = 1$: risk-neutral preferences;

(iii) $\zeta_+, \zeta_- > 1$: risk-averse preferences on losses and risk-seeking preferences on gains (convex in gains, concave in losses).

Experimental results for a series of one-off decisions have indicated that typically both ζ_+ and ζ_- are less than one thereby indicating that humans are risk-averse on gains and risk-seeking on losses—that is, v is concave for $y > y_o$ and convex otherwise).

In addition to the non-linear transformation of outcome values, in prospect theory the effect of under/over-weighting the likelihood of events that has been commonly observed in human behavior is modeled via *warping* of event probabilities [67, 198]. Other concepts such as framing effects, reference dependence, and loss aversion—captured, *e.g.*, in the (k_+, k_-) parameters in (11.2)—have also been widely observed in experimental studies on human decision-making (see, *e.g.*, [33, 167, 185]).

Outside of the prospect theory value, other mappings have been proposed to capture risk-sensitivity. For example, the mapping proposed in [126] where the authors develop a risk-sensitive reinforcement learning procedure, is the linear mapping

$$v(y) = \begin{cases} (1-\kappa)y, & y > y_o \\ (1+\kappa)y, & y \leq y_o \end{cases} \tag{11.3}$$

with $\kappa \in (-1, 1)$. This value function can be viewed as a special case of the prospect value function introduced above.

Another example is the entropic map which is given by

$$v(y) = \exp(\lambda y) \tag{11.4}$$

where λ controls the degree of risk-sensitivity. The entropic map, however, is either convex or concave on the entire outcome space.

Motivated by the empirical evidence supporting the prospect theoretic value function and numerical considerations of our algorithm, which are discussed in greater detail in subsequent sections, we introduce a new value function which retains the shape of the prospect theory value function—i.e. its convex–concave structure—while improving the performance (in terms of convergence speed) of the gradient-based inverse reinforcement learning algorithm we propose in Section 12. In particular, we define the locally Lipschitz-prospect (ℓ-prospect) value function given by

$$v(y) = \begin{cases} k_+(y - y_o + \epsilon)^{\zeta_+} - k_+\epsilon^{\zeta_+}, & y > y_o \\ -k_-(y_o - y + \epsilon)^{\zeta_-} + k_-\epsilon^{\zeta_-}, & y \leq y_o \end{cases} \tag{11.5}$$

with $k_+, k_-, \zeta_+, \zeta_- > 0$ and $\epsilon > 0$, a small constant. This value function is Lipschitz continuous on a bounded domain. Moreover, the derivative of the ℓ-prospect function is bounded away from zero at the reference point. Hence, in practice it has better numerical properties.

We remark that, for given parameters $(k_+, k_-, \zeta_+, \zeta_-)$, the ℓ-prospect function has the same risk-sensitivity as the prospect value function with those same parameters. Moreover, as $\epsilon \to 0$ the ℓ-prospect value function approaches the prospect value function and thus, qualitatively speaking, the degree of Lipschitzness decreases as $\epsilon \to 0$.

The fact that each of these value functions are defined by a small number of parameters that are highly interpretable in terms of risk-sensitivity and loss-aversion is one of the motivating factors for integrating them into a reinforcement learning framework. It is our aim

to design learning algorithms that will ultimately provide the theoretical underpinnings for designing incentives and control policies taking into consideration salient features of human decision-making behavior.

11.3 Valuation Functions via Coherent Risk Metrics

To further capture risk-sensitivity, *valuation functions* generalize the expectation operator, which considers *average* or *expected* outcomes,[2] to measures of risk.

Definition 16 (Monetary Risk Measure [56])**.** *A functional $\rho : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ on the space $\mathcal{X}$ of measurable functions defined on a probability space $(\Omega, \mathcal{F}, P)$ is said to be a monetary risk measure if $\rho(0)$ is finite and if, for all $X, X' \in \mathcal{X}$, ρ satisfies the following:*

1. *(monotone) $X \leq X' \implies \rho(X) \leq \rho(X')$*
2. *(translation invariant) $m \in \mathbb{R} \implies \rho(X + m) = \rho(X) + m$*

If a monetary risk measure ρ satisfies

$$\rho(\lambda X + (1 - \lambda)X') \leq \lambda\rho(X) + (1 - \lambda)\rho(X'), \tag{11.6}$$

for $\lambda \in [0, 1]$, then it is a *convex risk measure*. If, additionally, ρ is *positive homogeneous*, i.e. it satisfies the condition that

$$\lambda \geq 0 \implies \rho(\lambda X) = \lambda\rho(X), \tag{11.7}$$

then we call ρ a *coherent risk measure*.

We will primarily focus on convex measures of risk that are generated by a set of *acceptable* positions. Let $\mathcal{M}_1(\Omega, \mathcal{F})$ be the space of probability measures on $(\Omega, \mathcal{F})$.

Definition 17 (Acceptable Positions)**.** *Consider a value function v, a probability measure $P \in \mathcal{M}_1(\Omega, \mathcal{F})$, and $v_0 = v(y_0)$ with y_0 in the domain of v. The set*

$$\mathcal{A} = \{X \in \mathcal{X} |\ \mathbb{E}_P[v(X)] \geq v_0\} \tag{11.8}$$

is the set of acceptable positions where v_0 is the acceptance level*.*

The above definition can be extending to the entire class of probability measures on $(\Omega, \mathcal{F})$ as follows:

$$\mathcal{A} = \cap_{P \in \mathcal{M}_1(\Omega, \mathcal{F})}\{X \in \mathcal{X} |\ \mathbb{E}_P[v(X)] \geq v(y_P)\} \tag{11.9}$$

with constants y_P such that $\sup_{P \in \mathcal{M}_1(\Omega, \mathcal{F})} y_P < \infty$.

[2] In the case of two events, the valuation function can also capture warping of probabilities. Alternative approaches to reinforcement learning based on cumulative prospect theory for the more general case have been examined [90].

Proposition 18 ([56, proposition 4.7])**.** *Suppose the class of* acceptable positions $\mathcal{A}$ *is a non-empty subset of* $\mathcal{X}$ *satisfying*

1. $\inf\{m \in \mathbb{R} | X + m \in \mathcal{A}\} > -\infty$, $\forall\ X \in \mathcal{X}$, *and*
2. *given* $X \in \mathcal{A}, Y \in \mathcal{X}$, $Y \geq X \implies Y \in \mathcal{A}$.

Then $\mathcal{A}$ *induces a monetary measure of risk* $\rho_{\mathcal{A}}$. *If* $\mathcal{A}$ *is convex, then* $\rho_{\mathcal{A}}$ *is a convex measure of risk. Furthermore, if* $\mathcal{A}$ *is a cone, then* $\rho_{\mathcal{A}}$ *is a coherent risk metric.*

Note that monetary measure of risk induced by a set of acceptable positions $\mathcal{A} \subset \mathcal{X}$ is

$$\rho_{\mathcal{A}}(X) = \inf\{z \in \mathbb{R} |\ z + X \in \mathcal{A}\}. \tag{11.10}$$

The following proposition is key for extending the expectation operator to more general measures of risk.

Proposition 19 ([56, proposition 4.104])**.** *Consider the acceptance level set*

$$\mathcal{A} = \{X \in \mathcal{X} |\ \mathbb{E}_P[v(X)] \geq v_0\} \tag{11.11}$$

for a continuous value function v, *acceptance level* $v_0 = v(y_0)$ *for some* y_0 *in the domain of* v, *and probability measure* P. *Suppose that* v *is strictly increasing on* $(y_0 - \varepsilon, \infty)$ *for some* $\varepsilon > 0$. *Then the corresponding* $\rho_{\mathcal{A}}$ *is a convex measure of risk which is continuous from below. Moreover,* $\rho_{\mathcal{A}}(X)$ *is the unique solution to*

$$\mathbb{E}_P[v(X - m)] = v_0. \tag{11.12}$$

Proposition 19 also implies that for each value function, we can define an acceptance set which in turn induces a convex risk metric ρ. Let us consider an example.

Example 7 (Entropic Risk Metric [56])**.** *Consider the entropic value function* $v(y) = \exp(\lambda y)$. *It has been used extensively in the field of risk measures [56], in neuroscience to capture risk sensitivity in motor control [133] and even more so in control of MDPs (see,* e.g., *[34, 41, 111]).*

The entropic value function with an acceptance level v_0 *can be used to define the acceptance set*

$$\mathcal{A} = \{m \in \mathbb{R} |\ \mathbb{E}[\exp(-\lambda(m + Y))] \leq v_0\}. \tag{11.13}$$

The risk metric in this case is given by

$$\begin{aligned} \rho(X) &= \inf\{m \in \mathbb{R} |\ \mathbb{E}[\exp(-\lambda(m + Y))] \leq v_0\} && (11.14)\\ &= \tfrac{1}{\lambda} \log \mathbb{E}[\exp(-\lambda Y)] - \tfrac{1}{\lambda} \log(v_0). && (11.15) \end{aligned}$$

The parameter $\lambda \in \mathbb{R}$ controls the risk preference; indeed, this can be seen by considering the Taylor expansion [56, Example 4.105]. As a further comment, this particular risk metric is equivalent (up to an additive constant) to the so called entropic risk measure *which is given by*

$$\rho(Y) = \sup_{P' \in \mathcal{M}_1(P)} \left(\mathbb{E}_{P'}[-Y] - \frac{1}{\lambda} H(P'|P) \right) \tag{11.16}$$

where $\mathcal{M}_1(P)$ is the set of all measures on $(\Omega, \mathcal{F})$ that are absolutely continuous with respect to P and where $H(\cdot|\cdot)$ is the relative entropy function. ■

Let us recall the concept of a *valuation function* introduced and used in [11, 56, 174].

Definition 18 (Valuation Function)**.** *A mapping $\mathcal{V} : \mathbb{R}^{|I|} \times \Delta(I) \to \mathbb{R}$ is called a* valuation function *if for each $\mu \in \Delta(I)$, (i) $\mathcal{V}(Y, \mu) \leq \mathcal{V}(Z, \mu)$ whenever $Y \leq Z$ (monotonic) and (ii) $\mathcal{V}(Y + y\mathbf{1}, \mu) = \mathcal{V}(Y, \mu) + y$ for any $y \in \mathbb{R}$ (translation invariant).*

Such a map is used to characterize an agent's preferences—that is, one prefers (Y, μ) to (Z, ν) whenever $\mathcal{V}(Z, \nu) \leq \mathcal{V}(Y, \mu)$.

We will consider valuation functions that are convex risk metrics induced by a value function v and a probability measure μ. To simplify notation, from here on out we will suppress the dependence on the probability measure μ.

For each state–action pair, we define $\mathcal{V}(Y|x, a) : \mathbb{R}^{|I|} \times X \times A \to \mathbb{R}$ a *valuation map* such that $\mathcal{V}_{x,a} \equiv \mathcal{V}(\cdot|x, a)$ is a valuation function induced by an acceptance set with respect to value function v and acceptance level v_0.

If we let $\mathcal{V}^{\pi}_x(Y) = \sum_{a \in A(x)} \pi(a|x) \mathcal{V}_{x,a}(Y)$, the optimization problem in (11.1) generalizes to

$$\tilde{J}_T(\pi, x_0) = \mathcal{V}^{\pi_0}_{x_0} \Big[R[x_0, a_0] + \gamma \mathcal{V}^{\pi_1}_{x_1} \big[R[x_1, a_1] + \cdots + \gamma \mathcal{V}^{\pi_T}_{x_T} [R(x_T, a_T)] \cdots \big] \Big] \tag{11.17}$$

where we define $\max_{\pi} \tilde{J}(\pi, x_0) = \lim_{T \to \infty} \tilde{J}_T(\pi, x_0)$.

11.4 Risk-Sensitive Q-Learning

In the classical reinforcement learning framework, the Bellman equation is used to derive a Q-learning procedure. Generalizations of the Bellman equation for risk-sensitive reinforcement learning—derived, *e.g.*, in [126, 173]—have been used to formulate an action–value function or Q-learning procedure for the risk-sensitive reinforcement learning problem. In particular, as shown in [173], if V^* satisfies

$$V^*(x_0) = \max_{a \in A(x)} \mathcal{V}_{x,a}(R(x, a) + \gamma V^*), \tag{11.18}$$

then $V^* = \max_{\pi} \tilde{J}(\pi, x_0)$ holds for all $x_0 \in X$; moreover, a deterministic policy is optimal if $\pi^*(x) = \arg\max_{a \in A(x)} \mathcal{V}_{x,a}(R + \gamma V^*)$ [173, Thm. 5.5]. The action–value function $Q^*(x, a) =$

$\mathcal{V}_{x,a}(R + \gamma V^*)$ is defined such that (11.18) becomes

$$Q^*(x,a) = \mathcal{V}_{x,a}\left(R + \gamma \max_{a \in A(x')} Q^*(x',a)\right), \tag{11.19}$$

for all $(x,a) \in X \times A$.

Given a value function v and acceptance level v_0, we use the coherent risk metric induced state-action valuation function given by

$$\mathcal{V}_{x,a}(Y) = \sup\{z \in \mathbb{R} |\ \mathbb{E}[v(Y-z)] \geq v_0\} \tag{11.20}$$

where the expectation is taken with respect to $\mu = P(x'|x,a)P_r(w|x,a)$. Hence, by a direct application of proposition 19, if v is continuous and strictly increasing, then $\mathcal{V}_{x,a}(Y) = z^*(x,a)$ is the unique solution to $\mathbb{E}[v(Y - z^*(x,a))] = v_0$.

As shown in [174, proposition 3.1], by letting $Y = R + \gamma V^*$, we have that $z^*(x,a)$ corresponds to $Q^*(x,a)$ and, in particular,

$$\mathbb{E}\left[v\left(r(x,a,w) + \gamma \max_{a' \in A(x')} Q^*(x',a') - Q^*(x,a)\right)\right] = v_0 \tag{11.21}$$

where, again, the expectation is taken with respect to $\mu = P(x'|x,a)P_r(w|x,a)$.

The above leads naturally to a Q-learning procedure,

$$Q(x_t,a_t) \leftarrow Q(x_t,a_t) + \alpha_t(x_t,a_t)\big[v(y_t) - v_0\big], \tag{11.22}$$

where the non-linear transformation v is applied to the temporal difference $y_t = r_t + \gamma \max_a Q(x_{t+1},a) - Q(x_t,a_t)$ instead of simply the reward r_t. Transformation of the temporal differences avoids certain pitfalls of the reward transformation approach such as poor convergence performance. This procedure has convergence guarantees even in this more general setting under some assumptions on the value function v.

Theorem 24 (Q-learning Convergence [174, Theorem 3.2])**.** *Suppose that $v : Y \to \mathbb{R}$ is in $C(Y,\mathbb{R})$, is strictly increasing in y and there exists constants $\varepsilon, L > 0$ such that $\varepsilon \leq \frac{v(y)-v(y')}{y-y'} \leq L$ for all $y \neq y'$. Moreover, suppose that there exists a $\bar{y}$ such that $v(\bar{y}) = v_0$. If the non-negative learning rates $\alpha_t(x,a)$ are such that $\sum_{t=0}^{\infty} \alpha_t(x,a) = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2(x,a) < \infty$, $\forall (x,a) \in X \times A$, then the procedure in* (11.22) *converges to $Q^*(x,a)$ for all $(x,a) \in X \times A$ with probability one.*

The assumptions on α_t are fairly standard and the core of the convergence proof is based on the Robbins–Siegmund Theorem appearing in the seminal work [163].

We note that the assumptions on the value function v of Theorem 24 are fairly restrictive, excluding many of the value functions presented in Section 11.2. For example, value functions of the form e^x and x^ζ do not satisfy the global Lipschitz condition.

We generalize the convergence result in Theorem 24 by modifying the assumptions on the value function v to ensure that we have convergence of the Q-learning procedure for the ℓ-prospect and entropic value functions.

Assumption 17. *The value function $v \in C^1(Y, \mathbb{R})$ satisfies the following:*

(i) it is strictly increasing in y and there exists a $\bar{y}$ such that $v(\bar{y}) = v_0$;

(ii) it is locally Lipschitz on any ball of finite radius centered at the origin;

Note that in comparison to the assumptions of Theorem 24, we have removed the assumption that the derivative of v is bounded away from zero, and relaxed the global Lipschitz assumption on v. We remark that the ℓ-prospect and entropic value functions satisfy these assumptions for all parameters and MDPs.

Let $\mathcal{X}$ be a complete metric space endowed with the L_∞ norm and let $\mathcal{Q} \subset \mathcal{X}$ be the space of maps $Q : X \times A \to \mathbb{R}$. Further, define $\tilde{v} \equiv v - v_0$. We then re-write the Q–update equation in the form

$$Q_{t+1}(x,a) = \Big(1 - \frac{\alpha_t}{\alpha}\Big) Q_t(x,a) + \frac{\alpha_t}{\alpha}\big(\alpha(v(y_t) - v_0) + Q_t(x,a)\big) \tag{11.23}$$

where $\alpha \in (0, \min\{L^{-1}, 1\}]$ and we have suppressed the dependence of α_t on (x,a). This is a standard update equation form in, *e.g.*, the stochastic approximation algorithm literature [89, 162, 181]. In addition, we define the map given by

$$(TQ)(x,a) = \alpha \mathbb{E}_{x',w}\big[\tilde{v}\big(r(x,a,w) + \gamma \max_{a' \in A} Q(x',a') - Q(x,a)\big)\big] + Q(x,a) \tag{11.24}$$

which we will prove is a contraction.

Theorem 25. *Suppose that v satisfies Assumption 17 and that for each $(x,a) \in X \times A$ the reward $r(x,a,w)$ is bounded almost surely—that is, there exists $0 < M < \infty$ such that $|r| < M$ almost surely. Moreover, let $\alpha \in (0, \min\{1, L^{-1}\}]$, for L, the Lipschitz constant of v on $B_K(0)$.*

a. Let $B_K(0) \subset \mathcal{Q}$ be a closed ball of radius $K > 0$ centered at zero. Then, $T : \mathcal{Q} \to \mathcal{X}$ is a contraction.

b. Suppose K is chosen such that

$$\frac{\max\{|\tilde{v}(M)|, |\tilde{v}(-M)|\}}{(1-\gamma)} < K \min_{y \in I_K} D\tilde{v}(y) \tag{11.25}$$

where $I_K = [-M - K, M + K]$. Then, T has a unique fixed point in $B_K(0)$.

The proof of Theorem 25 relies on the following fixed point theorem.

Theorem 26 (Fixed Point Theorem [93, Theorem 2.2])**.** *Let (X,d) be a complete metric space and let $B_k(y) = \{x \in X |\ d(x,y) < k\}$ be a ball of radius r, where $k > 0$, centered at $y \in X$. Let $f : B_k(y) \to X$ be a contraction map with contraction constant $h < 1$. Further, assume that $d(y, f(y)) < k(1-h)$. Then, f has a unique fixed point in $B_k(y)$.*

Proof of Theorem 25.a. We claim that T is a contraction with constant $\bar{\alpha} = (1-\alpha(1-\gamma)\varepsilon_K)$ where $\varepsilon_K = \min\{D\tilde{v}(y)|\ y \in I_K\}$. Indeed, let $y(Q(x,a)) = r(x,a,w) + \gamma \max_{a'} Q(x',a') - Q(x,a)$ be the temporal difference and define $g(x',a') = \max_{a'} Q(x',a')$. For any $Q \in B_K(0)$ we note that the temporal differences are bounded—in fact, $y(Q(x,a)) \in I_K = [-M - K, M + K]$. Due to the monotonicity assumption on v, we have that for any $y', y \in I_K$, $\tilde{v}(y) - \tilde{v}(y') = \xi(y - y')$ for some $\xi \in [\varepsilon_K, L]$. Recall the contraction map defined in (11.24):

$$(TQ)(x,a) = \alpha \mathbb{E}_{x',w}\big[\tilde{v}\big(y(Q(x,a))\big)\big] + Q(x,a) \tag{11.26}$$

Then, for any Q_1 and Q_2, we have that

$$\begin{aligned}
(TQ_1 - TQ_2)(x,a) &= \alpha \mathbb{E}_{x',w}[\tilde{v}(y(Q_1(x,a))) - \tilde{v}(y(Q_2(x,a)))] + Q_1(x,a) - Q_2(x,a) \\
&\leq \alpha \mathbb{E}_{x',w}[\xi_{x',w}(\gamma(g_1(x',a') + g_2(x',a')) - Q_1(x,a) + Q_2(x,a))] \\
&\qquad + Q_1(x,a) - Q_2(x,a) \\
&\leq \alpha\gamma \mathbb{E}_{x',w}[\xi_{x',w}(g_1(x',a') + g_2(x',a'))] \\
&\qquad + (1 - \alpha \mathbb{E}_{x',w}[\xi_{x',w}])(Q_1(x,a) - Q_2(x,a)).
\end{aligned}$$

Hence,

$$\begin{aligned}
|(TQ_1 - TQ_2)(x,a)| &\leq (1 - \alpha(1-\gamma)\mathbb{E}_{x',w}[\xi_{x',w}])\|Q_1 - Q_2\|_\infty \\
&\leq (1 - \alpha(1-\gamma)\varepsilon_K)\|Q_1 - Q_2\|_\infty.
\end{aligned}$$

We claim that the constant $\bar{\alpha}_K = 1-\alpha(1-\gamma)\varepsilon_K < 1$. Indeed, recall that $0 < \alpha \leq \min\{1, L^{-1}\}$ so that if $\alpha = L^{-1}$, then $\bar{\alpha}_K < 1$ since $L = \max_{y \in I_K} D\tilde{v}(y)$ and $\varepsilon_K = \min_{y \in I_K} D\tilde{v}(y)$. On the other hand, if $\alpha = 1$, then $1 \leq L^{-1} \leq (\varepsilon_K)^{-1}$ so that $\varepsilon_K \leq 1$ which, in turn, implies that $\bar{\alpha}_K < 1$. If $0 < \alpha < \min\{1, L^{-1}\}$, then $\bar{\alpha}_K < 1$ follows trivially from the implications in the above two cases. Thus, T is a contraction on $B_K(0)$ with the constant $\bar{\alpha}_K = (1 - \alpha(1-\gamma)\varepsilon_K) < 1$. □

Proof of Theorem 25.b. Suppose K is chosen such that

$$\frac{\max\{|\tilde{v}(M)|, |\tilde{v}(-M)|\}}{1-\gamma} < K \min_{y \in I_K} D\tilde{v}(y). \tag{11.27}$$

Now, we argue that T applied to the zero map, $0 \in B_K(0)$, is strictly less than $K(1 - \bar{\alpha}_K)$. Indeed, for any $\alpha \in (0, \min\{1, L^{-1}\}]$,

$$\|T(0)\| \leq \alpha \max\{|v(M)|, |v(-M)|\} < (1-\gamma)K\varepsilon_K\alpha = K(1 - \bar{\alpha}_K)$$

Combinging the above fact with the fact that T is a contraction, the assumptions of Theorem 26 hold and, hence there is a unique fixed point $Q^*(x,a) \in B_K(0)$ for each $(x,a) \in X \times A$. □

The following proposition shows that the ℓ-prospect and entropic value functions satisfy the assumption in (11.25). Moreover, it shows that the value functions which satisfy Assumption 17 also satisfy (11.25).

Proposition 20. *Consider a MDP with reward $r : X \times A \times W \to \mathbb{R}$ bounded almost surely by M and $\gamma \in (0,1)$ and consider the condition*

$$\frac{\max\{|\tilde{v}(M)|, |\tilde{v}(-M)|\}}{(1-\gamma)} < K \min_{y \in I_K} D\tilde{v}(y). \tag{11.28}$$

a. Suppose v satisfies Assumption 17 and that for some $\varepsilon > 0$, $\varepsilon < \frac{v(y)-v(y')}{y-y'}$ for all $y \neq y'$. Then (11.28) holds.

b. Suppose v is an ℓ-prospect value function with arbitrary parameters $(k_-, k_+, \zeta_-, \zeta_+)$ satisfying Assumption 17. Then there exists a K such that the ℓ-prospect value function satisfies (11.28).

c. Suppose that v is an entropic value function. Then there exists a $C > 0$ such that for any $|\lambda| \in (0, C)$ where v satisfies Assumption 17, (11.28) holds with $K = (\lambda)^{-1}$.

Proof of Proposition 20.a. Suppose v satisfies Assumption 17 and that for some $\varepsilon > 0$, $\varepsilon < \frac{v(y)-v(y')}{y-y'}$ for all $y \neq y'$. Then there exists a value of K, say $\bar{K}$, such that (11.28) holds for all $K > \bar{K}$. Indeed since $\min_{K>0} \varepsilon_K > \varepsilon$, for all K satisfying

$$\frac{\max\{|\tilde{v}(M)|, |\tilde{v}(-M)|\}}{\varepsilon(1-\gamma)} < K,$$

(11.28) must hold. □

Proof of Proposition 20.b. We now show that for the ℓ-prospect value function, (11.28) holds for any choice of parameters $(k_-, k_+, \zeta_-, \zeta_+)$. Indeed, for $\zeta_+, \zeta_- \geq 1$ and any choice of k_-, k_+,

$$\min_{K>0} \varepsilon_K > \varepsilon > 0$$

where $\varepsilon = \min\{\lim_{y \uparrow 0} D\tilde{v}(y), \lim_{y \downarrow 0} D\tilde{v}(y)\}$. Therefore, with $\zeta_+, \zeta_- \geq 1$, for any K such that

$$\frac{\max\{|\tilde{v}(M)|, |\tilde{v}(-M)|\}}{\varepsilon(1-\gamma)} < K,$$

(11.28) must hold. For the case when either $\zeta_+ < 1$ or $\zeta_- < 1$ or both, we note that

$$\min_{y \in I_K} D\tilde{v}(y) = \min\left\{ \min_{y \in \{M+K, -M-K\}} D\tilde{v}(y), \varepsilon \right\}.$$

so that we need only show that for $\zeta_+ < 1$ and $\zeta_- < 1$, there exists a K such that

$$\frac{\max\{|\tilde{v}(M)|, |\tilde{v}(-M)|\}}{1-\gamma} < K D\tilde{v}(K+M) \tag{11.29}$$

and

$$\frac{\max\{|\tilde{v}(M)|, |\tilde{v}(-M)|\}}{1-\gamma} < KD\tilde{v}(-K-M), \tag{11.30}$$

respectively. Note that

$$D\tilde{v}(y) = \begin{cases} k_+\zeta_+(y-y_0+\epsilon)^{\zeta_+-1}, & y \geq y_0 \\ k_-\zeta_-(y_0-y+\epsilon)^{\zeta_--1}, & y < y_0 \end{cases}$$

Without loss of generality, we show (11.29) must hold for $\zeta_+ < 1$ and reference point $y_0 = 0$ (the proof for $\zeta_- < 1$ follows an exactly analogous argument). Plugging $D\tilde{v}(K+M)$ in and rearranging, we get that we need to find a K such that

$$\frac{\max\{|\tilde{v}(M)|, |\tilde{v}(-M)|\}}{(1-\gamma)\xi_+k_+} < K(K+M+\epsilon)^{\xi_+-1}$$

Since the right-hand side above is a function of K that is zero at $K = 0$ and approaches infinity as $K \to \infty$, and the left-hand side is a finite constant, there is some $\bar{K}$ such that for all $K > \bar{K}$, the above holds. Thus, for the ℓ-prospect value function, our assumptions are satisfied and there always exists a value of K to choose in Theorem 25.b. □

Proof of Proposition 20.c. Suppose v is an entropic map. We note that, for the entropic map, $\min_{y\in I_K} D\tilde{v}(y)$ must occur at either $K+M$ or $-K-M$ if $\lambda < 0$ or $\lambda > 0$, respectively. Without loss of generality, let $\lambda > 0$. First, consider that the derivative of $\tilde{v}$,

$$D\tilde{v}(y) = \frac{1}{\lambda}e^{\lambda y},$$

is minimized on I_K at $-M-K$ for any M and K. Moreover, $|\tilde{v}(M)| > |\tilde{v}(-M)|$. Hence, with $K = \lambda^{-1}$, we can derive conditions on λ for which (11.28) holds. In other words, with the specified K, we use (11.28) to dertermine which values of λ are admissible. Indeed, from (11.28), we have

$$\frac{e^{\lambda M}}{1-\gamma} < \frac{1}{\lambda}e^{-\lambda M-1}$$

which reduces to

$$\lambda e^{2\lambda M} < (1-\gamma)e^{-1}.$$

Let $x = 2\lambda M$, so that

$$xe^x < 2M(1-\gamma)e^{-1}.$$

Now, we can apply the Lambert W function which satisfies $W(xe^x) = x$ for $x \geq 0$, to get that

$$x < W(2M(1-\gamma)e^{-1}),$$

so that

$$\lambda < \frac{W(2M(1-\gamma)e^{-1})}{2M}.$$

Thus, if $|\lambda| < \frac{W(2M(1-\gamma)e^{-1})}{2M}$, then for the choice $K = \frac{1}{\lambda}$, Theorem 25.b holds. □

With Theorem 25 and Proposition 20, we now prove convergence of Q-learning for risk-sensitive reinforcement learning.

Theorem 27 (Q-learning Convergence on $B_K(0)$)**.** *Suppose that v satisfies Assumption 17 and that for each $(x, a) \in X \times A$ the reward $r(x, a, w)$ is bounded almost surely—that is, there exists $0 < M < \infty$ such that $|r| < M$ almost surely. Moreover, suppose the ball $B_K(0)$ is chosen such that* (11.25) *holds. If the non-negative learning rates $\alpha_t(x, a)$ are such that $\sum_{t=0}^{\infty} \alpha_t(x, a) = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2(x, a) < \infty$, $\forall (x, a) \in X \times A$, then the procedure in* (11.22) *converges to $Q^* \in B_K(0)$ with probability one.*

Given Theorem 25, the proof of Theorem 27 follows directly the same proof as provided in [173]. The key aspect of the proof is combining the fixed point result of Theorem 25 with Robbins–Siegmund Theorem [163].

Theorem 25, proposition 20, and Theorem 27 extend the results for risk-sensitive reinforcement learning presented in [173] by relaxing the assumptions on the value functions for which the Q-learning procedure converges.

11.5 Chapter Summary

In this chapter we introduced the *forward* model under which we assume agents make decisions in uncertain dynamic environments. We introduced formally the concept of an MDP and built up the necessary concepts to formally define the risk-sensitive reinforcement learning problems. Crucially, this risk-sensitive reinforcement learning framework is amenable to the integration of expressive behavioral models from behavioral economics and mathematical finance. Building on previous work on risk-sensitive reinforcement learning we then derived a procedure for Q-learning under more relaxed conditions on agents' value functions that in previous work.

In the next chapter, we investigate the inverse problem of learning an agent's value function or utility from data.

Chapter 12

Inverse Risk-Sensitive Reinforcement Learning

In this chapter, we formulate the inverse risk-sensitive reinforcement learning problem. To begin, we select a parametric class of policies, $\{\pi_\theta\}_\theta$, $\pi_\theta \in \Pi$ and parametric value function $\{v_\theta\}_\theta$, $v_\theta \in \mathcal{F}$ where $\mathcal{F}$ is a family of value functions and $\theta \in \Theta \subset \mathbb{R}^d$.

We use value functions such as those described in Section 11.2; e.g., if v is the prospect theory value function defined in (11.2), then the parameter vector is $\theta = (k_-, k_+, \zeta_-, \zeta_+, \gamma, \beta)$. For mappings v and Q, we now indicate their dependence on θ—that is, we will write $Q(x, a, \theta)$ and $v_\theta(y) = v(y, \theta)$ where $v : Y \times \Theta \to \mathbb{R}$. Note that since y is the temporal difference it also depends on θ and we will indicate this dependence where it is not directly obvious by writing $y(\theta)$.

In the inverse reinforcement learning literature, it is common to assume that agents' policies (the way they choose actions in states) are stochastic and derived from a smooth map G that operates on the action-value function space. This defines a parametric policy space. A common form of G is the space of Boltzmann policies of the form

$$G_\theta(Q)(a|x) = \frac{\exp(\beta Q(x, a, \theta))}{\sum_{a' \in A} \exp(\beta Q(x, a', \theta))} \tag{12.1}$$

where $\beta > 0$ controls how close $G_\theta(Q)$ is to a *greedy policy* which we define to be any policy π such that $\sum_{a \in A} \pi(a|x) Q(x, a, \theta) = \max_{a \in A} Q(x, a, \theta)$ at all states $x \in X$. We will utilize policies of this form. Note that, as is pointed out in [136], the benefit of selecting strictly stochastic policies is that if the true agent's policy is deterministic, uniqueness of the solution is forced.

We aim to *tune* the parameters so as to minimize some loss $\ell(\pi_\theta)$ which is a function of the parameterized policy π_θ. By an abuse of notation, we introduce the shorthand $\ell(\theta) = \ell(\pi_\theta)$.

12.1 An Optimization approach to Inverse Risk-Sensitive Reinforcement Learning

The optimization problem is specified by

$$\min_{\theta\in\Theta}\{\ell(\theta)|\ \pi_\theta = G_\theta(Q^*), v_\theta \in \mathcal{F}\} \tag{12.2}$$

Given a set of *demonstrations* $\mathcal{D} = \{(x_k, a_k)\}_{k=1}^N$, it is our goal to recover the policy and estimate the value function.

There are several possible loss functions that may be employed. For example, suppose we elect to minimize the negative weighted log-likelihood of the demonstrated behavior which is given by

$$\ell(\theta) = \textstyle\sum_{(x,a)\in\mathcal{D}} w(x,a)\log(\pi_\theta(x,a)) \tag{12.3}$$

where $w(x,a)$ may, e.g., be the normalized empirical frequency of observing (x,a) pairs in $\mathcal{D}$, i.e. $n(x,a)/N$ where $n(x,a)$ is the frequency of (x,a).

Related to maximizing the log-likelihood, an alternative loss function is the relative entropy or Kullback-Leibler (KL) divergence between the empirical distribution of the state-action trajectories and their distribution under the learned policy—that is,

$$\ell(\theta) = \textstyle\sum_{x\in\mathcal{D}_x} D_{\mathrm{KL}}(\hat{\pi}(\cdot|x)||\pi_\theta(\cdot|x)) \tag{12.4}$$

where

$$D_{\mathrm{KL}}(P||Q) = \textstyle\sum_i P(i)\log\left(P(i)/Q(i)\right) \tag{12.5}$$

is the KL divergence, $\mathcal{D}_x \subset \mathcal{D}$ is the sequence of observed states, and $\hat{\pi}$ is the empirical distribution on the trajectories of $\mathcal{D}$.

Computing Gradients in Inverse Risk-Sensitive Reinforcement Learning

We propose to solve the problem of estimating the parameters of the agent's value function via gradient methods. This requires computing the derivative of $Q^*(x,a,\theta)$ with respect to θ. Since $Q^*(x,a,\theta)$ is only defined as the fixed point of a contraction map. obtaining its gradients is highly non-trivial, and in fact it is not clear if its gradients even exist.

Hence, given the form of the Q-learning procedure where the temporal differences are transformed as in (11.22), in this section we show that it is in fact differentiable, and derive a procedure for obtaining the derivative.

Using some basic calculus, given the form of smoothing map G_θ in (12.1), we can compute the derivative of the policy π_θ with respect to θ_k for an element of $\theta \in \Theta$:

$$\begin{aligned} D_{\theta_k}\pi_\theta(a|x) &= \pi_\theta(a|x) D_{\theta_k}\ln(\pi_\theta(a|x)) \\ &= \pi_\theta(a|x)\beta\big(D_{\theta_k}Q^*(x,a,\theta) - \textstyle\sum_{a'\in A}\pi_\theta(a'|x)D_{\theta_k}Q^*(x,a',\theta)\big). \end{aligned} \tag{12.6}$$

We show that $D_{\theta_k} Q_\theta^*$ can be calculated almost everywhere on Θ by solving fixed-point equations similar to the Bellman-optimality equations.

To do this, we require some assumptions on the value function v.

Assumption 18. *The value function $v \in C^1(Y \times \Theta, \mathbb{R})$ satisfies the following conditions:*

(i) v is strictly increasing in y and for each $\theta \in \Theta$, there exists a $\bar{y}$ such that $v(\bar{y}, \theta) = v_0$;

(ii) for each $\theta \in \Theta$, on any ball centered around the origin of finite radius, v is locally Lipschitz in y with constant $L_y(\theta)$ and locally Lipschitz on Θ with constant L_θ;

(iii) there exists $\varepsilon > 0$ such that $\varepsilon \leq \frac{v(y,\theta)-v(y',\theta)}{y-y'}$ for all $y \neq y'$.

Define $L_y = \max_\theta L_y(\theta)$ and $L = \max_\theta \{L_y(\theta), L_\theta\}$. As before, let $\tilde{v} \equiv v - v_0$. We re-write the Q–update equation as

$$Q_{t+1}(x,a,\theta) = \left(1 - \tfrac{\alpha_t}{\alpha}\right) Q_t(x,a,\theta) + \tfrac{\alpha_t}{\alpha}\left(\alpha(v(y_t(\theta),\theta) - v_0) + Q_t(x,a,\theta)\right) \tag{12.7}$$

where $y_t(\theta) = r_t + \gamma \max_a Q_t(x_{t+1}, a, \theta) - Q_t(x_t, a_t, \theta)$ is the temporal difference, $\alpha \in (0, \min\{L^{-1}, 1\}]$ and we have suppressed the dependence of α_t on (x, a). In addition, define the map T such that

$$(TQ)(x,a,\theta) = \alpha \mathbb{E}_{x',w} \tilde{v}(y(\theta),\theta) + Q(x,a,\theta) \tag{12.8}$$

where $y(\theta) = r(x,a,w) + \gamma \max_{a' \in A} Q(x',a',\theta) - Q(x,a,\theta)$. This map is a contraction for each θ. Indeed, fixing θ, when v satisfies Assumption 18, then for cases where $v_0 = 0$, T was shown to be a contraction in [126] and in the more general setting (i.e. $v_0 \neq 0$), in [174].

Our first main result on inverse risk-sensitive reinforcement learning, which is the theoretical underpinning of our gradient-based algorithm, gives us a mechanism to compute the derivative of Q_θ^* with respect to θ as a solution to a fixed-point equation via a contraction mapping argument.

Let $D_i \tilde{v}(\cdot,\cdot)$ be the derivative of $\tilde{v}$ with respect to the i–th argument where $i = 1, 2$.

Theorem 28. *Assume that $v \in C^1(Y \times \Theta, \mathbb{R})$ satisfies Assumption 18. Then the following statements hold:*

a. Q_θ^* is locally Lipschitz continuous as a function of θ—that is, for any $(x,a) \in X \times A$, $\theta, \theta' \in \Theta$, $|Q^*(x,a,\theta) - Q^*(x,a,\theta')| \leq C\|\theta - \theta'\|$ for some $C > 0$;

b. except on a set of measure zero, the gradient $D_\theta Q_\theta^*$ is given by the solution of the fixed–point equation

$$\begin{aligned}\phi_\theta(x,a) =& \alpha \mathbb{E}_{x',w}\big[D_2 \tilde{v}(y(\theta),\theta) + D_1 \tilde{v}(y(\theta),\theta) \\ & \cdot (\gamma \phi_\theta(x', a^*_{x'}) - \phi_\theta(x,a))\big] + \phi_\theta(x,a)\end{aligned} \tag{12.9}$$

where $\phi_\theta : X \times A \to \mathbb{R}^d$ and $a^*_{x'}$ is the action that maximizes $\sum_{a' \in A} \pi(a|x') Q(x',a,\theta)$ where π is any policy that is greedy with respect to Q_θ.

To give a high-level outline of this proof, we use an induction argument combined with a contraction mapping argument on the map

$$\begin{aligned}(S\phi_\theta)(x,a) =&\alpha\mathbb{E}_{x',w}\big[D_2\tilde{v}(y(\theta),\theta) + D_1\tilde{v}(y(\theta),\theta) \\ &\cdot(\gamma\phi_\theta(x',a^*_{x'}) - \phi_\theta(x,a))\big] + \phi_\theta(x,a). \end{aligned} \tag{12.10}$$

The almost everywhere differentiability follows from Rademacher's Theorem (see, *e.g.*, [71, Thm. 3.1]).

The gradient algorithm and Theorem 28 are consistent with the gradient descent framework which uses the *contravariant* gradient for learning as introduced in [10] for Riemannian parameter spaces Θ. Of course, when Θ is Euclidean and the coordinate system is orthonormal, the gradient we normally use (*covariant* derivative) coincides with the contravariant gradient. However, using the covariant derivative does not generalize to admissible parameter spaces with more structure.

Moreover, as is pointed out in [136], the trajectories that result from the solution to the gradient algorithm are equivalent up to reparameterization through a smooth invertible mapping with a smooth inverse. Contravariant gradient methods have been shown to be asymptotically efficient in a probabilistic sense and thus, they tend to avoid *plateaus* [10, 149].

Given these comments, we now proceed to prove Theorem 28. To do so, we formally define the concept of a subdifferential.

Definition 19 (Fréchet Subdifferentials)**.** *Let U be a Banach space and U^* its dual. The Fréchet subdifferential of $f : U \to \mathbb{R}$ at $u \in U$, denoted by $\partial f(u)$ is the set of $u^* \in U^*$ such that*

$$\liminf_{h\to 0\ h\neq 0} \|h\|^{-1}\left(f(u+h) - f(u) - \langle u^*, u\rangle\right) \geq 0. \tag{12.11}$$

Given this definition we recall a useful property of subdifferentials.

Proposition 21 ([88, 136])**.** *For a finite family $(f_i)_{i\in I}$ of real-valued functions (where I is a finite index set) defined on U, let $f(u) = \max_{i\in I} f_i(u)$. If $u^* \in \partial f_i(u)$ and $f_i(u) = f(u)$, then $u^* \in \partial f_i(u)$. If $f_1, f_2 : U \to \mathbb{R}$, $\alpha_1, \alpha_2 \geq 0$, then $\alpha_1\partial f_1 + \alpha_2\partial f_2 \subset \partial(\alpha_1 f_1 + \alpha_1 f_2)$.*

Finally, we present a useful proposition on the convergence of subdifferentials of a sequence of functions which is a crucial to our proof.

Proposition 22 ([136, 148])**.** *Suppose that $(f_n)_{n\in\mathbb{N}}$ is a sequence of real-valued functions on U which converge pointwise to f. Let $u \in U$, $u^*_n \in \partial f_n(u) \subset U^*$ and suppose that (u^*_n) is weak*–convergent to u^* and is bounded. Moreover, suppose that at u, for any $\varepsilon > 0$, there exists an $N > 0$ and $\delta > 0$ such that for any $n \geq N$, $h \in B_U(0,\delta)$, a δ–ball around 0, $f_n(u+h) \geq f_n(u) + \langle u^*_n, h\rangle - \varepsilon\|h\|$. Then $u^* \in \partial f(u)$.*

Given these definitions and results now provide the proof for parts (a) and (b) of Theorem 28.

Proof of Theorem 28.a. Let $Q_0(x,a,\theta) \equiv 0$. Then it is trivial that $Q_0(x,a,\theta)$ is locally Lipschitz in θ on Θ. Supposing that $Q_t(x,a,\theta)$ is L_t–locally Lipschitz in θ, then we need to show that $TQ_t(x,a,\theta)$ is locally Lipschitz which we recall is defined by

$$(TQ)(x,a,\theta) = \alpha \mathbb{E}_{x',w} \tilde{v}(y(\theta),\theta) + Q(x,a,\theta) \tag{12.12}$$

where $y(\theta) = r(x,a,w) + \gamma \max_{a' \in A} Q(x',a',\theta) - Q(x,a,\theta)$.

Since $\tilde{v} \equiv v - v_0$, it also satisfies Assumption 18. Let $L_y = \max\{L_y(\theta) | \theta \in \Theta\}$ and define $g_t(x,\theta) = \max_{a'} Q_t(x,a',\theta)$. Note that since Q_t is assumed Lipschitz with constant L_t, so is g_t. Supressing the dependent of TQ on (x,a), we have that

$$\begin{aligned} TQ_t(\theta) - TQ_t(\theta') &= \alpha \mathbb{E}_{x',w}[\tilde{v}(y(\theta),\theta) - \tilde{v}(y(\theta'),\theta')] + Q_t(x,a,\theta) - Q_t(x,a,\theta') \\ &= \alpha \mathbb{E}_{x',w}[\tilde{v}(y(\theta),\theta) - \tilde{v}(y(\theta'),\theta) + \tilde{v}(y(\theta'),\theta) - \tilde{v}(y(\theta'),\theta')] + Q_t(\theta) - Q_t(\theta'). \end{aligned}$$

Due to the monotonicity of $\tilde{v}$ in y, we know that for all y_1, y_2 there exists $\xi \in [\varepsilon, L_y]$ such that

$$\tilde{v}(y_1,\theta) - \tilde{v}(y_2,\theta) = \xi(y_1 - y_2).$$

Hence,

$$\begin{aligned} &\mathbb{E}_{x',w}[\tilde{v}(y(\theta),\theta) - \tilde{v}(y(\theta'),\theta) + \tilde{v}(y(\theta'),\theta) - \tilde{v}(y(\theta'),\theta')] \\ &\quad = \mathbb{E}_{x',w}[\xi_{x',w}(y(\theta) - y(\theta')) + \tilde{v}(y(\theta'),\theta) - \tilde{v}(y(\theta'),\theta')] \end{aligned}$$

where we simply denote the dependence of ξ on x' and w, the components subject to randomness. Then,

$$\begin{aligned} TQ_t(\theta) - TQ_t(\theta') &= \alpha \mathbb{E}_{x',w}\left[\xi_{x',w}(y(\theta) - y(\theta')) + \tilde{v}(y(\theta'),\theta) - \tilde{v}(y(\theta'),\theta')\right] + Q_t(\theta) - Q_t(\theta') \\ &= \alpha\gamma \mathbb{E}_{x',w}[\xi_{x',w}(g_t(x',\theta) - g_t(x',\theta'))] - \alpha \mathbb{E}_{x',w}[\xi_{x',w}(Q_t(\theta) - Q_t(\theta'))] \\ &\qquad + \alpha \mathbb{E}_{x',w}[\tilde{v}(y(\theta'),\theta) - \tilde{v}(y(\theta'),\theta')] + Q_t(\theta) - Q_t(\theta') \\ &= \alpha\gamma \mathbb{E}_{x',w}[\xi_{x',w}(g_t(x',\theta) - g_t(x',\theta'))] - \alpha \mathbb{E}_{x',w}[\xi_{x',w}](Q_t(\theta) - Q_t(\theta')) \\ &\qquad + \alpha \mathbb{E}_{x',w}[\tilde{v}(y(\theta'),\theta) - \tilde{v}(y(\theta'),\theta')] + Q_t(\theta) - Q_t(\theta') \\ &= (1 - \alpha \mathbb{E}_{x',w}[\xi_{x',w}])(Q_t(\theta) - Q_t(\theta')) + \alpha\gamma \mathbb{E}_{x',w}[\xi_{x',w}(g_t(x',\theta) - g_t(x',\theta'))] \\ &\qquad + \alpha \mathbb{E}_{x',w}[\tilde{v}(y(\theta'),\theta) - \tilde{v}(y(\theta'),\theta')] \end{aligned}$$

so that

$$\|TQ_t(\theta) - TQ_t(\theta')\| \leq ((1 - \alpha(1-\gamma)\varepsilon) + \alpha L_\theta) L_t \|\theta - \theta'\|.$$

Hence, letting $\bar{\alpha} = 1 - \alpha(1-\gamma)\varepsilon$, we have that $TQ_t(\cdot,\cdot,\theta)$ is L_{t+1}–locally Lipschitz with $L_{t+1} = \bar{\alpha} L_t + \alpha L_\theta$. With $L_0 = 0$, by iterating, we get that

$$L_{t+1} = (\bar{\alpha}^t + \cdots + \bar{\alpha} + 1)\alpha L_\theta.$$

As stated in Section 12.1, T is a contraction so that $T^n Q_0 \to Q^*_\theta = Q^*(\cdot,\cdot,\theta)$ as $n \to \infty$. Hence, by the above argument, Q^*_θ is $\alpha L_\theta / (1 - \bar{\alpha})$–Lipschitz continuous. □

Proof of Theorem 28.b. Consider a fixed vector $\theta \in \mathbb{R}^d$. We now show that the operator S acting on the space of functions $\phi_\theta : X \times A \to \mathbb{R}^d$ and defined by

$$(S\phi_\theta)(x,a) = \alpha \mathbb{E}_{x',w}\big[D_2\tilde{v}(y(\theta),\theta) + D_1\tilde{v}(y(\theta),\theta)(\gamma\phi_\theta(x',a^*_{x'}) - \phi_\theta(x,a))\big] + \phi_\theta(x,a) \tag{12.13}$$

is a contraction where $a^*_{x'}$ is the action that maximizes $\sum_{a'\in A}\pi(a|x)Q(x,a,\theta)$ for any greedy policy π with respect to Q_θ. Indeed,

$$\begin{aligned}(S\phi_\theta - S\phi'_\theta)(x,a) = \alpha\mathbb{E}_{x',w}[D_1\tilde{v}(y(\theta),\theta)\big(\gamma(\phi_\theta(x',a^*_{x'}) - \phi'_\theta(x',a^*_{x'})) - (\phi_\theta(x,a) - \phi'_\theta(x,a))\big)] \\ + \phi_\theta(x,a) - \phi'_\theta(x,a) \\ \leq (1 - \alpha(1-\gamma)\mathbb{E}_{x',w}[D_1\tilde{v}(y(\theta),\theta)])\|\phi_\theta - \phi'_\theta\|_\infty\end{aligned}$$

so that, by Assumption 18,

$$\|(S\phi_\theta - S\phi'_\theta)(x,a)\| \leq (1 - \alpha(1-\gamma)\varepsilon)\|\phi_\theta - \phi'_\theta\|_\infty.$$

Thus, $\bar{\alpha}$ is the required constant for ensuring S is a contraction. We remark that S operates on each of the d components of θ separately and hence, it is a contraction when restricted to each individual component.

Let π denote a greedy policy with respect to Q^*_θ and let π_n be a sequence of policies that are greedy with respect to $Q_n = T^n Q_0$ where ties are broken so that $\sum_{(x,a)\in X\times A}|\pi(a|x) - \pi_n(a|x)|$ is minimized. Then for large enough n, $\pi_n = \pi$. Denote by S_{π_n} the map S defined in (12.10) where π_n is the implemented policy. Consider the sequence $\phi_{\theta,n}$ such that $\phi_{\theta,0} = 0$ and $\phi_{\theta,n+1} = S_{\pi_n}\phi_{\theta,n}$. For large enough n, $\phi_{\theta,n+1} = S_\pi\phi_{\theta,n}$. Applying the (local) contraction mapping theorem (see, *e.g.*, [168, Theorem 3.18]) we get that $\lim_{n\to\infty} S^n\phi_0$ converges to a unique fixed point.

Moreover, by induction and Proposition 21, $\phi_{\theta,n}(x,a) \in \partial_\theta Q_n(x,a,\theta)$. Hence, by Proposition 22, the limit is a subdifferential of Q^*_θ since $\tilde{v}$ is Lipschitz on Y and Θ and the derivatives of $\tilde{v}$ are uniformly bounded. Since by part (a), Q^*_θ is locally Lipschitz in θ, Rademacher's Theorem (see, *e.g.*, [71, Thm. 3.1]) tells us it is differentiable almost everywhere (except a set of Lebesgue measure zero). Since Q^*_θ is differentiable, its subdifferential is its derivative. □

Theorem 28 gives us a procedure—namely, a fixed–point equation which is a contraction—to compute the derivative $D_{\theta_k}Q^*$ so that we can compute the derivative of our loss function $\ell(\theta)$. Hence the gradient method provided in Algorithm 4 for solving the inverse risk-sensitive reinforcement learning problem is well formulated.

Algorithm 4 Gradient-Based Risk-Sensitive IRL

Input : Observed data $\mathcal{D}$, Number of iterations N

Initialize: $\theta \leftarrow \theta_0$ **while** $k < N$ *&* $\|\ell(\theta) - \ell(\theta_-)\| \geq \delta$ **do**

  $\eta_k \leftarrow$ **LineSearch**$(\ell(\theta_-)), D_\theta\ell(\theta)$

  $\theta \leftarrow \theta - \eta_k D_\theta \ell(\theta)$

  $k \leftarrow k + 1$

return θ

Remark 9. *The prospect theory value function v given in* (11.2) *is not globally Lipschitz in y—in particular, it is not Lipschitz near the reference point y_o—for values of ζ_+ and ζ_- less than one. Moreover, for certain parameter combinations, it may not even be differentiable. The ℓ-prospect function, on the other hand, is locally Lipschitz and its derivative near the reference point is bounded away from zero. This makes it a more viable candidate for numerical implementation. Its derivative, however, is not bounded away from zero as $y \to \infty$.*

This being said, we note that if the procedure for computing Q^* follows an algorithm which implements repeated applications of the map T is initialized with $Q_0(x,a)$ being finite for all (x,a) and r is bounded for all possible (x,a,w) pairs, then the derivative of $\tilde{v}$ will always be bounded away from zero for all realized values of y in the procedure. An analogous statement can be made regarding the computation of $D_\theta Q^*$. Hence, the procedures for computing Q^* and $D_\theta Q^*$ for all the value functions we consider (excluding the classical prospect value function) are guaranteed to converge (except on a set of measure zero).

Let us translate this remark into a formal result. Consider a modified version of Assumption 18:

Assumption 19. *The value function $v \in C^1(Y \times \Theta, \mathbb{R})$ satisfies the following:*

(i) it is strictly increasing in y and for each $\theta \in \Theta$, there exists a $\bar{y}$ such that $v(\bar{y}, \theta) = v_0$;

(ii) for each $\theta \in \Theta$, it is Lipschitz in y with constant $L_y(\theta)$ and locally Lipschitz on Θ with constant L_θ.

Simply speaking, analogous to Assumption 17, we have removed the uniform lower bound on the derivative of v. Moreover, Theorem 25 gives us that T, as defined in (12.8), is a contraction on a ball of finite radius for each θ under Assumption 17.

Theorem 29. *Assume that $v \in C^1(Y \times \Theta, \mathbb{R})$ satisfies Assumption 19 and that the reward $r : X \times A \times W \to \mathbb{R}$ is bounded almost surely by $M > 0$. Then the following statements hold.*

1. *For any ball $B_K(0)$, Q^*_θ is locally Lipschitz-continuous on $B_K(0)$ as a function of θ—that is, for any $(x,a) \in X \times A$, $\theta, \theta' \in \Theta$, $|Q^*(x,a,\theta) - Q^*(x,a,\theta')| \leq C\|\theta - \theta'\|$ for some $C > 0$.*

2. *For each θ, let $B_K(0)$ be the ball with radius K satisfying*

$$\frac{\max\{|\tilde{v}(M,\theta)|, |\tilde{v}(-M,\theta)|\}}{1-\gamma} < K \min_{y \in I_K} .D\tilde{v}(y,\theta) \tag{12.14}$$

Except on a set of measure zero, the gradient $D_\theta Q^*_\theta(x,a) \in B_K(0)$ is given by the solution of the fixed–point equation

$$\phi_\theta(x,a) = \alpha \mathbb{E}_{x',w}\big[D_2\tilde{v}(y(\theta),\theta) + D_1\tilde{v}(y(\theta),\theta)(\gamma\phi_\theta(x', a^*_{x'}) - \phi_\theta(x,a))\big] + \phi_\theta(x,a)$$

where $\phi_\theta : X \times A \to \mathbb{R}^d$ and $a^*_{x'}$ is the action that maximizes $\sum_{a' \in A} \pi(a|x')Q(x',a,\theta)$ with π being any policy that is greedy with respect to Q_θ.

The proof the above theorem follows the same techniques as in Theorem 25 and Theorem 28. Below, we provide an outline of the proof, and for brevity, direct the reader to the analogous sections of the two proceeding proofs as required.

Proof of Theorem 29.a. For each θ, the proof that $TQ(x, a, \theta)$ is a contraction, and thus has a fixed point $Q^*_\theta \in B_K(0)$, follows directly that of Theorem 25 where instead of Q_1 and Q_2 we have $Q(\theta)$ and $Q(\theta')$. Given that T is a contraction, the proof that $Q^*_\theta \in B_K(0)$ is Lipschitz with constant $\alpha L_\theta/(1 - \bar{\alpha}_K)$ follows a similar argument to Theorem 28. □

Proof of Theorem 29.b. The proof that S is a contraction on $B_K(0)$ follows a similar argument to that of Theorem 28, part (b). Indeed,

$$\begin{aligned}(S\phi_\theta - S\phi'_\theta)(x, a) &= \alpha\mathbb{E}_{x',w}[D_1\tilde{v}(y(\theta), \theta)\big(\gamma(\phi_\theta(x', a^*_{x'}) - \phi'_\theta(x', a^*_{x'})) - (\phi_\theta(x, a) - \phi'_\theta(x, a))\big)] \\ &\quad + \phi_\theta(x, a) - \phi'_\theta(x, a) \\ &\leq (1 - \alpha(1 - \gamma)\mathbb{E}_{x',w}[D_1\tilde{v}(y(\theta), \theta)])\|\phi_\theta - \phi'_\theta\|_\infty\end{aligned}$$

so that, by Assumption 17,

$$\|(S\phi_\theta - S\phi'_\theta)(x, a)\| \leq (1 - \alpha(1 - \gamma)\varepsilon_K)\|\phi_\theta - \phi'_\theta\|_\infty$$

where $\varepsilon_K = \min\{D_1 v(y, \theta)|\ y \in I_K\}$. Note that $\bar{\alpha}_K = 1 - \alpha(1 - \gamma)\varepsilon_K < 1$ for the same reasons as given in the proof of Theorem 25 since $\alpha \in (0, \min\{1, L^{-1}\}]$.

For each $\theta \in \Theta$, let $B_K(0)$ be the ball with radius K satisfying

$$\frac{\max\{|\tilde{v}(M, \theta)|, |\tilde{v}(-M, \theta)|\}}{1 - \gamma} < K \min_{y \in I_K} D\tilde{v}(y, \theta).$$

Then, for each θ, S satisfies Theorem 26 so that it has a unique fixed point in $B_K(0)$.

Following the same argument as in the proof of Theorem 28, part (b), by induction and Proposition 21, $\phi_{\theta,n}(x, a) \in \partial_\theta Q_n(x, a, \theta)$. Hence, by Proposition 22, the limit is a subdifferential of Q^*_θ. By part (a), Q^*_θ is locally Lipschitz in θ so that Rademacher's Theorem (see, *e.g.*, [71, Thm. 3.1]) implies it is differentiable almost everywhere (except a set of Lebesgue measure zero). Since Q^*_θ is differentiable, its subdifferential is its derivative.

□

Note that for each fixed θ, condition (12.14) is the same as condition (11.25). Moreover, proposition 20 shows that for the ℓ-prospect and entropic value functions, such a K must exist for any choice of parameters.

Complexity

Small dataset size is often a challenge in modeling sequential human decision-making owing in large part to the frequency and time scale on which decisions are made in many applications. To properly understand how our gradient-based approach performs for different amounts

of data, we analyze the case when the loss function, $\ell(\theta)$, is either the negative of the log-likelihood of the data—see (12.3) above—or the sum over states of the KL divergence between the policy under our learned value function and the the empirical policy of the agent—see (12.4) above. These are two of the more common loss functions used in the literature.

We first note that maximizing the log-likelihood is equivalent to minimizing a weighted sum over states of the KL divergence between the empirical policy of the *true* agent, $\hat{\pi}_n$, and the policy under the learned value function, π_θ. In particular, through some algebraic manipulation the weighted log-likelihood can be re-written as

$$\ell(\theta) = \textstyle\sum_{x\in\mathcal{D}_x} w(x) D_{KL}(\hat{\pi}_n(\cdot|x)||\pi_\theta(\cdot|x)) \tag{12.15}$$

where $w(x)$ is the frequency of state x normalized by $|\mathcal{D}| = N$. This approach has the added benefit that it is independent of θ and therefore will not be affected by scaling of the value functions [136].

Both cost functions are natural metrics for performance in that they minimize a measure of the divergence between the optimal policy under the learned agent and empirical policy of the true agent. While the KL-divergence is not suitable for our analysis, since it is not a metric on the space of probability distributions, it does provide an upper bound on the total variation (TV) distance via Pinsker's inequality:

$$\delta(\hat{\pi}_n(\cdot|x), \pi_\theta(\cdot|x)) \leq \sqrt{2 D_{KL}(\hat{\pi}_n(\cdot|x)||\pi_\theta(\cdot|x))} \tag{12.16}$$

where $\delta(\pi(\cdot|x), \pi_\theta(\cdot|x))$ is the TV distance between $\hat{\pi}_n(\cdot|x)$ and $\pi_\theta(\cdot|x)$, defined as

$$\delta(\hat{\pi}_n(\cdot|x), \pi_\theta(\cdot|x)) = \tfrac{1}{2}\|\pi_\theta(\cdot|x) - \hat{\pi}_n(\cdot|x)\|_1. \tag{12.17}$$

The TV distance between distributions is a proper metric. Furthermore, use of the two cost functions described above will also translate to minimizing the TV distance as it is upper bounded by the KL divergence.

We first note that, for each state x, we would ideally like to get a bound on $\delta(\pi(\cdot|x), \pi_\theta(\cdot|x))$, the TV distance between the agent's true policy $\pi(\cdot|x)$ and the estimated policy $\pi_\theta(\cdot|x)$. However, we only have access to the empirical policy $\hat{\pi}_n$. We therefore use the triangle inequality to get an upper bound on $\delta(\pi(\cdot|x), \pi_\theta(\cdot|x))$, in terms of values for which we can calculate explicitly or construct bounds. In particular, we derive the following bound:

$$\delta(\pi_\theta(\cdot|x), \pi(\cdot|x)) \leq \delta(\hat{\pi}_n(\cdot|x), \pi_\theta(\cdot|x)) + \delta(\hat{\pi}_n(\cdot|x), \pi(\cdot|x)). \tag{12.18}$$

Note that $\delta(\hat{\pi}_n(\cdot|x), \pi_\theta(\cdot|x))$ is tantamount to a training error as metricized by the TV distance, and is upper bounded by a function of the KL divergence (which appears in the loss function) via (12.16).

The first term in (12.18), $\delta(\hat{\pi}_n(\cdot|x), \pi(\cdot|x))$, is the distance between the empirical policy and the true policy in state x. Using the Dvoretzky Kiefer-Wolfowitz inequality (see, *e.g.*, [112, 127]), this term can be bounded above with high probability. Indeed,

$$\Pr(\|\pi(\cdot|x) - \hat{\pi}_n(\cdot|x)\|_1 > \epsilon) \leq 2|A| e^{-2n\epsilon^2/|A|^2},\ \epsilon > 0 \tag{12.19}$$

where n is the number of samples from the distribution $\pi(\cdot|x)$ and $|A|$ is the cardinality of the action set. Combining this bound with (12.18), we get that, with probability $1-\nu$,

$$\delta(\pi_\theta(\cdot|x), \pi(\cdot|x)) \leq |A| \left(\frac{2}{n} \log \frac{2|A|}{\nu}\right)^{1/2} + \delta(\hat{\pi}_n(\cdot|x), \pi_\theta(\cdot|x)). \tag{12.20}$$

Supposing Algorithm 1 achieves a sufficiently small training error $\varepsilon > 0$, the second term above can be bounded above by a calculable small amount which we define notationally to be $\bar{\varepsilon} > 0$. Supposing $\bar{\varepsilon}$ is also sufficiently small, the dominating term in the distance between π and π_θ is the first term on the right-hand side in (12.20). This gives us a $O(n^{-1/2})$ convergence rate on the per state level. This rate is seen qualitatively in our experiments on sample complexity outlined in Section 12.2.

We note that this bound is for each individual state x. Thus, for states that are visited more frequently by the agent, we have better guarantees on how well the policy under the learned value function approximates the true policy. Moreover, it suggests ways of designing data collection schemes to better understand the agent's actions in less explored regions of the state space.

12.2 Examples

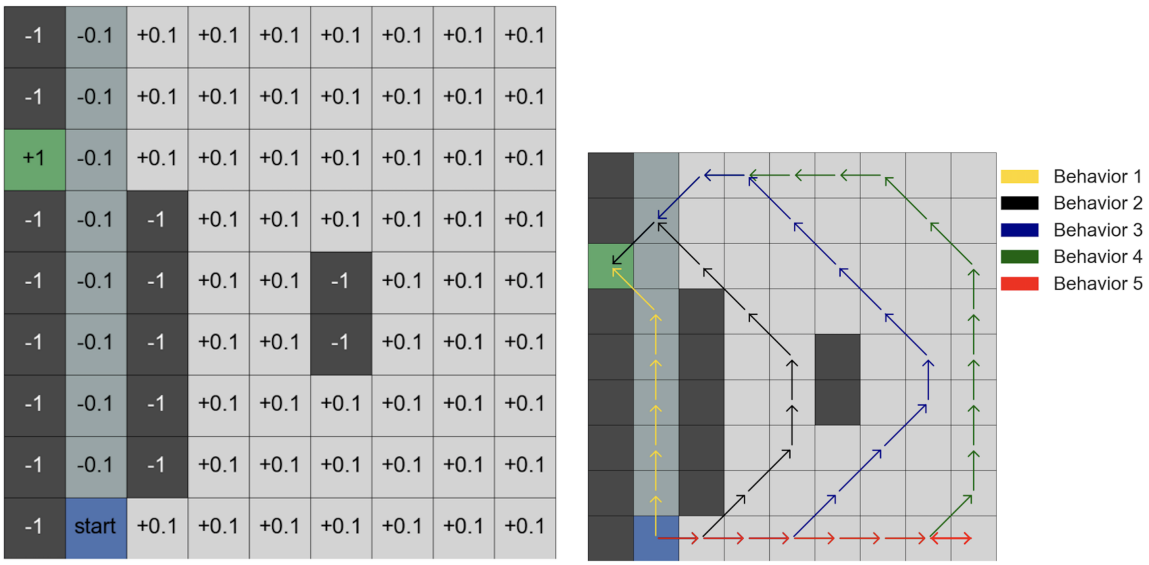
Let us now demonstrate the performance of the proposed method on two examples. While we are able to formulate the inverse risk-sensitive reinforcement learning problem for parameter vectors θ that include γ and β, in the following examples we use $\gamma = 0.95$ and $\beta = 4$. The purpose of doing this is to explore the effects of changing the value function parameters on the resulting policy.

In all experiments, our optimization objective is the negative log-likelihood of the data, defined in (12.3) and the valuation function we use is induced by an acceptance level set defined for a value function that we specify and acceptance level of zero. Furthermore, for the prospect and ℓ-prospect value functions, we use a reference point of zero[1]. These choices are aimed at further deconflating our observations of the behavior—in terms of risk-sensitivity and loss-aversion—that results from different choices of the value function parameters from other characteristics of the MDP or learning algorithm.

Grid World

In our first test of the proposed gradient-based inverse risk-sensitive reinforcement learning approach, we utilize data from agents operating on the canonical Grid World MDP. In the remainder, we describe the setup of the MDP, the three types of experiments we conduct, and qualitative results on sample complexity. The three experiments are described as follows:

[1]Individually, the acceptance level and the reference point can be recentered around zero without loss of generality.



(a) (b)

Figure 12.1: (a) Grid World layout showing the reward structure. (b) The five behavior profiles of risk-sensitive policies through the Grid World. These five paths correspond to the maximum likelihood paths of agents with various parameter combinations for their prospect, ℓ-prospect and entropic map value functions. To generate each behavior with the prospect and ℓ-prospect value functions, the following parameter combinations $(\{k_-, k_+, \zeta_-, \zeta_+\})$ were used: *Behavior 1*: $\{0.1, 1.0, 0.5, 1.5\}$; *Behavior 2*: $\{1.0, 1.0, 1.0, 1.0\}$; *Behavior 3*: $\{1.0, 1.0, 1.1, 0.9\}$; *Behavior 4*: $\{5.0, 1.0, 1.1, 0.8\}$; *Behavior 5*: $\{5.0, 1.0, 1.5, 0.7\}$. To generate the behaviors with the entropic map value function, we varied λ from 1 to -1.

1. Learning the value function of an agent with the correct model for the value function (e.g., learning a prospect value function when the agent also has a prospect value function);

2. Learning the value function of an agent with the wrong model for the value function (e.g., learning an entropic map value function when the agent has a prospect value function);

3. Exploring the dependence of our training error on the number of sample trajectories collected from the agent.

We measure the performance of the gradient-based approach via the TV norm, defined in (12.17), of the difference between the policy in state x of the true agent and the policy in state x under the learned value function.

Value Function	**Prospect**		**ℓ-prospect**		**Entropic**	
Behavior	**Mean**	**Variance**	**Mean**	**Variance**	**Mean**	**Variance**
Behavior 1	1.9e-2	6.3e-4	1.3e-2	2.3e-4	1.6e-3	5.1e-6
Behavior 2	1.5e-2	2.0e-4	1.0e-2	9.6e-5	2.6e-4	1.4e-7
Behavior 3	2.0e-2	3.6e-4	1.1e-2	1.3e-4	2.2e-3	1.5e-5
Behavior 4	1.6e-2	2.0e-4	1.2e-2	1.4e-4	4.6e-4	1.8e-7
Behavior 5	4.7e-2	3.0e-3	1.0e-2	3.4e-4	6.6e-4	2.2e-7

(a) Learning with the Correct Model

Value Function	**Mean**	**Variance**
Prospect	1.5e-2	1.6e-4
ℓ-prospect	1.5e-2	1.6e-4
Entropic Map	5.4e-2	1.4e-2

(b) Learning with an Incorrect Model

Table 12.1: (a) In this experiment, the learning is done with the same type of value function as that of the agent from which the data was collected. We report the mean and variance across all states in the grid of the TV distance between the true policy and the policy under the learned value function. We note that these are the results of the best of five randomly sampled initial sets of parameters. (b) In this experiment, 10,000 trajectories were sampled from the policy of an agent with the prospect value function with $\{k_-, k_+, \zeta_-, \zeta_+\} = \{2.0, 1.0, 0.9, 0.7\}$. We then used this data to learn prospect, ℓ-prospect, and entropic map value functions. We report the mean and variance across all states in the grid of the TV distance between the true policy and the policy under the learned value function. Again, we note that we present the best of five randomly sampled initial sets of parameters.

Setup

Our instantiation of Grid World is shown in Fig. 12.1a. An agent operating in this MDP starts in the blue box and aims to maximize their value function over an infinite time horizon. Every square in the grid represents a state, and the action space allows movement in to any of the eight neighboring states $A = \{N, NE, E, SE, S, SW, W, NW\}$. Each action corresponds to a movement in the specified direction (where we have used the usual abbreviations for directions). The black and green states are absorbing, meaning that once an agent enters that state they can never leave no matter their action. In all the other states, the agent moves in their desired direction with probability 0.93 and they move in any of the other seven directions with probability 0.01. To make the grid finite, any action taking the agent out of the grid has probability zero, and the other actions are re-weighted accordingly. The reward structure of our instantiation of the Grid World is shown in Fig. 12.1a as well. The agent gets a reward of -1 and $+1$ for being in the black and green states respectively. In the darker gray states, the agent gets a reward of -0.1. In all other states the agent is given a reward of $+0.1$.

Learning with the correct model of the value function

This experiment is intended to validate our approach on a simple example. We trained agents with various parameter combinations of the four value functions described in Section 11. The resulting policies of these agents are classified into five behavior profiles via their maximum likelihood path through the MDP. These behaviors are outlined in Fig. 12.1b. Each behavior corresponds to the maximum likelihood path resulting from a different risk profile: *Behavior 1* corresponds to a profile that is risk-seeking on gains, *Behavior 2* corresponds to a profile that is risk neutral on gains and losses (this is also the behavior corresponding to the non-risk-sensitive reinforcement learning approach), and *Behaviors 3-5* correspond to behaviors that are increasingly risk averse on losses and increasingly weigh losses more than gains.

We sampled 1,000 trajectories from the policies of these agents and used the data to learn the value function of the agent using our gradient-based approach. In this experiment, the learned value function is of the same type as that of the agent. For example, the data sampled from the policy of an agent having a prospect value function and exhibiting *Behavior 1* is used to learn the parameters of a prospect value function. We note that due to the non-convexity of the problem, we use five randomly generated initial parameter choices.

The results we report are associated with the value function that achieves the minimum value of the objective. In Table 12.1a, we report the mean TV distance between the two policies across all states, as well as the variance in the TV distance across states. In all the cases considered in Table 12.1a, the learned value functions produce policies that correctly match the maximum likelihood path of the true agent.

We remark that the performance for learning a prospect value function was consistently worse than learning an ℓ-prospect function. This is most likely due to the fact that the prospect value function is not Lipschitz around the reference point. Thus, we have no guarantees of differentiability of Q^* with respect to θ for the prospect value function. This translates to numerical issues in calculating the gradient which, in turn, results in worse performance.

The entropic value function performs best of the four value functions, primarily due to the fact that there is only one parameter to learn, and the rewards and losses are all relatively small. In fact, in all the cases the learned entropic map value function coincided with the true value function of the agent, thereby indicating that the objective function was relatively convex around the parameter values we tried.

Learning with an incorrect model of the value function

The second experiment consists of learning different types of value functions from the same dataset. This is a more realistic experiment since the value function of human subjects will very likely be different than any model we could choose. The motivation for this experiment is to ensure that the results and risk-profiles learned were consistent across our choice of model.

The experiment uses 10,000 samples from an agent with a prospect value function and learned prospect, ℓ-prospect, and entropic map value functions. The mean TV distance between the policy of the true agent and the policies under the learned value functions are shown in Table 12.1b. The true agent's value function has parameters $\{k_-, k_+, \zeta_-, \zeta_+\} = \{2.0, 1.0, 0.9, 0.7\}$—that is, it is risk-seeking in losses, risk-averse in gains, and loss averse.

Again, the learned value functions all have policies that replicated the maximum likelihood behavior of the true agent. We note that the ℓ-prospect and prospect functions perform as well as each other on this data, but the ℓ-prospect function showed none of the numerical issues that we encountered with the prospect function (see Section 12.2 for further detail on numerical considerations). Further, learning with the ℓ-prospect function is markedly faster than with the prospect function. Again, this is most likely due to the fact that the prospect function is not locally Lipschitz continuous around the reference point. Thus, the values of α required to make the various contraction maps converge to their fixed points are vanishingly small. This results in slow convergence.

The fact that the entropic value functions does not perform as well is most likely due to the fact that it cannot accurately match the shape of the prospect function at these values; e.g., the entropic map is always either convex or concave.

Qualitative results on sample complexity

One of the challenges in modeling human decision-making is the lack of access to large datasets, particularly when it comes to sequential decisions that are made over longer periods of time. This is counter to the usual learning scenarios addressed in the much of the learning literature. For instance, if the focus is learning to control a robot, then it may be possible to generate a large number of demonstrations very quickly. This motivates our third experiment with the Grid World MDP—i.e. an experiment that allows us to better understand how the performance of our approach varies with the size of the dataset.

In this experiment, we first train an agent with an entropic map value function and then create sets of sample trajectories from the agent's policy varying between zero and 10,000 in size. Next, using each of these sample sets, we learn the value function via our approach and plot the mean TV distance across all states between the true policy of the agent and the policy under the learned value function. This is shown in Fig. 12.2.

First, we note that more data does translate to consistently better results. This matches our intuition that the better our data matches the policy of the true agent, the better we can learn a value function that would be associated with that policy. Of particular interest, though, is the rate at which the average TV across all the states decreases with the number of trajectories sampled. The rate, which is on the order of $x^{-0.54}$, is very close to the asymptotic rate, derived in Section 7, of $O(x^{-1/2})$. This suggests that the dominating factor in the performance of our algorithm is how well our data matches the underlying policy, and not the non-convexity of the objective function. In fact, this provides empirical evidence that the second term in (12.20)—i.e. $\delta(\hat{\pi}_n(\cdot|x), \pi_\theta(\cdot|x))$—must also be $O(x^{-1/2})$.

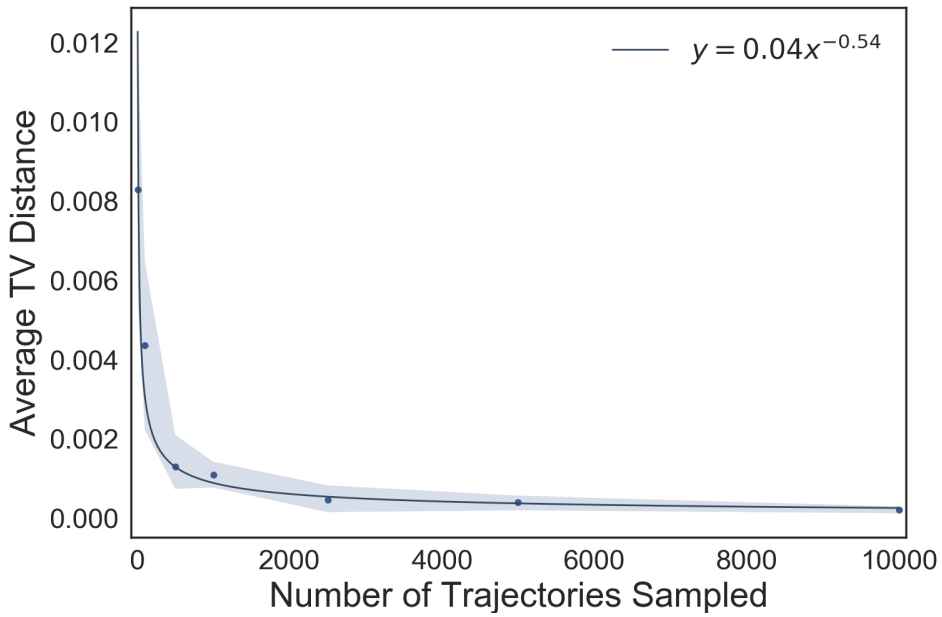


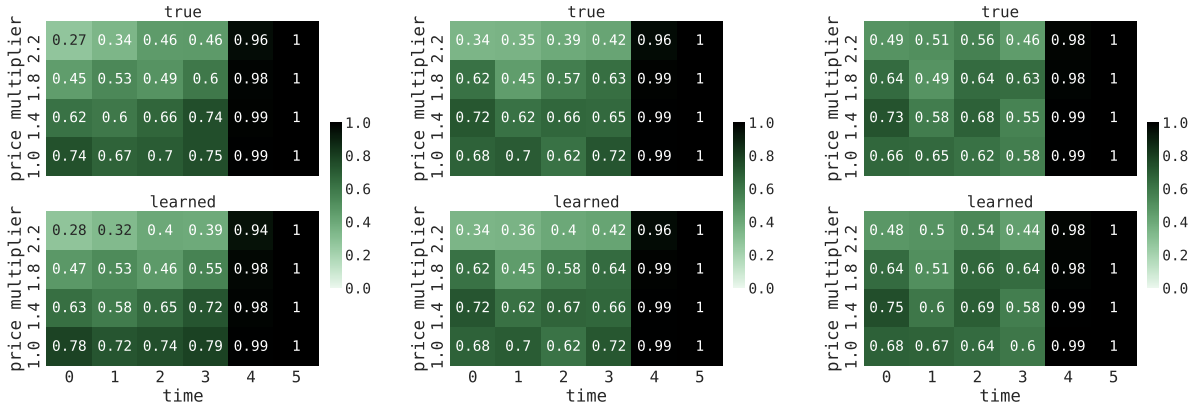
Figure 12.2: The mean TV distance across all states between the policy of the agent and the policy under the learned value function, as a function of the number of trajectories in the dataset. To construct each data point, we sample five different datasets of the same number of trajectories from the agent's policy. We then try five random initial parameter values per dataset and take the value function that achieves the minimum value of the objective. Finally, we calculate the mean TV distance between the policy of the agent and the policy under the learned value function for each dataset and then average these values. The bars show the 95% confidence interval around the mean of the five datasets of the given size. Finally, we note that the trendline $y = 0.04x^{-0.54}$ is the best fit of the form $y = ax^b$ to the data points, for constant terms a, b.

A Passenger's View of Ride-Sharing

In addition to the Grid World example, we explore a ride-sharing example for which the MDP is created from real-world data and we simulate agents with different risk preference and loss aversion profiles[2].

Reference dependence models are increasingly being used to model travel choices and activity scheduling [100]. More broadly, behavioral modeling has been used quite extensively

[2]We adopt the *surge pricing* model here due to the availability of data even though ride-sharing services such as Uber are moving towards personalized pricing schemes that combine data on exogenous factors such as demand and driver supply with data on riders' individual choices to deterime a price that reflects what the rider is willing to pay. This kind of pricing model motivates even more strongly the need for techniques that are considerate of how humans actually make decisions.



(a) risk-seeking in gains/risk-averse in losses (convex/concave), $\zeta_+ = \zeta_- = 1.5$

(b) risk-neutral, $\zeta_+ = \zeta_- = 1.0$

(c) risk-averse in gains/risk-seeking in losses (concave/convex), $\zeta_+ = \zeta_- = 0.5$

Figure 12.3: Plots showing the probabilities of taking a ride in each state under the true and learned optimal policies for true and learned agents with prospect value functions. The true agent has prospect gain parameters of $k_+ = 0.5$ and $k_- = 1.0$ for all three plots. The value function used for the right most graphic (Fig. 12.3c) is most representative of human decision-making since humans tend to be risk-averse in gains, risk-seeking in losses, are loss averse. In these plots, the trend we see is that the more risk-averse, the less likely they policy suggests taking the ride.

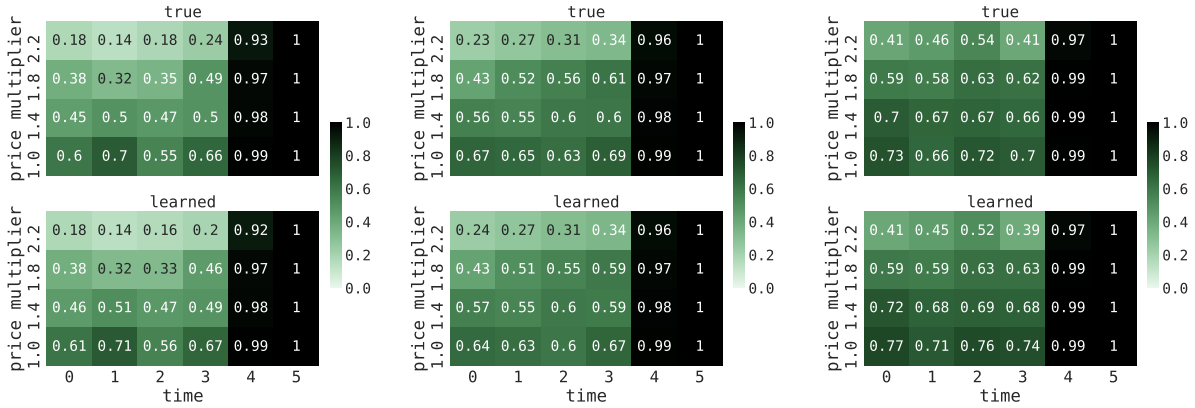
in transportation to model travel choices (see, e.g., [13, 31, 85, 200]).

Ride-sharing is a disruptive technology that offers commuters an alternative mode of transport. Many ride-sharing companies set prices based on both supply of drivers and demand of passengers. As a result, the price can fluctuate over time and space and passengers react differently to price changes depending on their risk preferences.

From the passenger's view point, we model the ride-sharing MDP as follows. The action space is $A = \{0, 1\}$ where 0 corresponds to 'wait' and 1 corresponds to 'ride.' The state space $X = \mathcal{X} \times \mathcal{T} \cup \{x_\mathrm{f}\}$ where $\mathcal{X}$ is a finite set of surge price multipliers, $\mathcal{T} = \{0, \ldots, T_\mathrm{f}\}$ is the part of the state corresponding to the time index, and x_f is a terminal state representing the completed ride that occurs when a ride is taken. At time t, the state is notationally given by (x_t, t). The reward r_t is modeled as a random variable that depends on the current price as well as a random variable $Z(t)$ for travel time. In particular, for any time $t < T_\mathrm{f}$ the reward is given by

$$R(x_t, a_t) = \begin{cases} \bar{r}, & a_t = 0 \text{ ('wait')} \\ \tilde{r}_t, & a_t = 1 \text{ ('ride')} \end{cases} \tag{12.21}$$

with $\bar{r} < 0$ a constant and $\tilde{r}_t = S_t - x_t(p_\mathrm{base} + p_\mathrm{mile} D + p_\mathrm{min} Z(t))$ where D is the distance in miles, S_t is a time dependent satisfaction (we selected it to linearly decrease in time from some initial satisfaction level), and p_base, p_mile, and p_min are the base, per mile, and per min



(a) less loss-averse $k_+ = 1.5, k_- = 0.5$

(b) no loss-aversion, $k_+ = k_- = 1.0$

(c) more loss-averse, $k_+ = 0.5, k_- = 1.5$

Figure 12.4: Plots showing the probabilities of taking a ride in each state under the true and learned optimal policies for true and learned agents with prospect value functions. The true agent has prospect parameters of $\zeta_- = \zeta_+ = 1.0$ for all three plots, while we vary (k_+, k_-) to capture different degrees of loss-aversion. In these plots, the trend we see is that the more loss-averse the agent (under both the learned and true value functions), the more likely they are to take the ride.

prices, respectively.

At the final time T_{final}, the agent is forced to take the ride if they have not selected to take a ride at a prior time. This reflects the fact that the agent presumably needs to get from their origin to their destination and the reward structure reflects the dissatisfaction the agent feels as a result of having to ultimately take the ride despite the potential desire to wait.

Using the Uber Movement[3] platform for travel time statistics, base pricing data[4] and surge pricing data[5] for Washington D.C., we examined several locations and hours which have different characteristics in terms of travel time and price statistics. We generate the distribution for $Z(t)$ from these data sets as well as the surge price intervals and transition probability matrix. Since the core risk-sensitive behaviors we observe are similar across the different locations, we report only on one.

Specifically, we report on a ride-sharing MDP generated for origin GPS= $(-77.027046, 38.926749)$ and destination GPS= $(-76.935773, 38.885964)$ [6] in Washington D.C. at 5AM.

[3]Uber Movement: `https://movement.uber.com/cities`

[4]The base, per min, and per mile prices can be found here: `http://uberestimate.com/prices/Washington-DC/`

[5]The surge pricing data we used was originally collected by and has been made publicly available here: `https://github.com/comp-journalism/2016-03-wapo-uber`. The data we use was collected over three minute intervals in period between November 14 to November 28, 2016.

[6]Note that these correspond to Uber Movement id's 197 and 113, respectively.

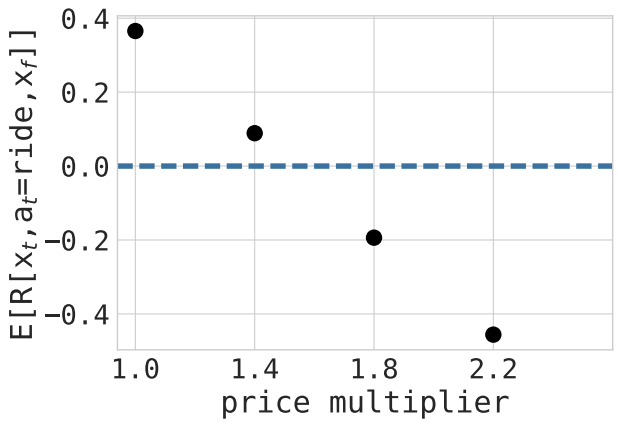


Figure 12.5: Expected rewards for each time step in the ride-sharing example. Notice that the rewards can either be gains (positive values) or losses (negative values) given that we take the reference point to be $r_0 = 0$.

Value Function	**Prospect**		**Entropic**		**ℓ-prospect**	
Preferences	**Mean**	**Variance**	**Mean**	**Variance**	**Mean**	**Variance**
Risk-Averse Gains /Risk-Seeking Losses[7]	1.3e-2	3.5e-4	0.9e-3	1.0e-6	1.0e-2	1.4e-4
Risk-Neutral	0.6e-2	4.9e-5	1.4e-3	2.0e-6	6.6e-3	1.0e-4
Risk-Seeking Gains /Risk-Averse Losses[8]	1.1e-2	1.7e-4	1.1e-3	1.5e-6	1.1e-2	1.1e-4

Table 12.2: Averaged TV error and variance over 10 different initializations of the algorithm for different risk-preference profiles. The last column shows the error when using an ℓ-prospect agent with $\epsilon =$ 1e-2 to learn a prospect agent.

The transition probability kernel $P : X \times A \times X \to [0, 1]$ is estimated from the ride-sharing data. The travel-time data is available on an hourly basis and the price change data is available on a three minute basis. Hence, we use the three minute price change data for each hour to derive a static transition matrix by empirically estimating the transition probabilities where we bin prices in the following way. For prices in $[1.0, 1.2)$, $x = 1.0$; for prices in $[1.2, 1.6)$, $x = 1.4$; for prices in $[1.6, 2.0)$, $x = 1.8$; otherwise $x = 2.2$. Hence, $\mathcal{X} = \{1.0, 1.4, 1.8, 2.2\}$. In the time periods we examine, the max price multiplier was 2.2. We set the reference point y_0 and acceptance level v_0 to be zero.

With this model, the transition matrix for the price multipliers is given by

$$P = \begin{bmatrix} 0.876 & 0.099 & 0.017 & 0.008 \\ 0.347 & 0.412 & 0.167 & 0.074 \\ 0.106 & 0.353 & 0.259 & 0.282 \\ 0.086 & 0.219 & 0.143 & 0.552 \end{bmatrix} \tag{12.22}$$

for each time. The travel time distribution is a standard normal distribution truncated to

the upper and lower bounds specified by the Uber Movement data. Measured in seconds, we use location parameter 2371, scale parameter 100, and 1554 and 3619 as the upper and lower bound, respectively.

The graphics in Fig. 12.3 and Fig. 12.4 show the state space as a grid with the probability of taking a ride under the true and learned optimal policies overlaid on each state. For the examples depicted in these figures, we consider the true and learned agents to have prospect value functions.

In Fig. 12.3, we fix the true agent's parameters to show a range of behaviors from risk-seeking in losses/risk-averse in gains to risk-averse in losses/risk-seeking in gains. There is empirical evidence supporting the fact that human's are more like the former. Moreover, in these examples we use $(k_+, k_-) = (0.5, 1)$ to capture that humans tend to be loss-averse—that is, for losses and gains of equal value, the loss is perceived as more significant.

On the other hand, in Fig. 12.4 we fix the true agent's parameters to show a range of behaviors depending on the degree of loss-aversion. In particular, we fix $\zeta_- = \zeta_+ = 1$ and vary the ratio of k_- to k_+, where a higher ratio corresponds to more loss averse preferences.

In each of the graphics in Fig. 12.3 and Fig. 12.4, we see that the learned policy is very close to the true policy. In addition, in Fig. 12.3, we observe that the more risk-averse the agent is (in gains or losses), the more likely they are to take the ride. This trend can be seen by noting the sign of the expected rewards—in Fig. 12.5, we see that the reward is positive for $x_t \in \{1.0, 1.4\}$ and is negative for $x_t \in \{1.8, 2.2\}$—and examining the corresponding rows in Fig. 12.3 for negative and positive rewards. In Fig. 12.4, observe that the more loss averse the agent, the more likely they are to take the ride uniformly. This is reasonable as the satisfaction level is linear decreasing in time.

In Table 12.2, we show the mean and variance of the total variation error for the ride-sharing example where we varied the risk preference profiles, holding $(k_-, k_+) = (1, 1)$, using agents with prospect and entropic value functions. In addition, we show the error for different risk profiles when we learn a true prospect agent with an ℓ-prospect agent. Recall that the prospect value function does not meet the requirements of our theorem where as the ℓ-prospect value function does as it is Lipschitz.

Numerical Considerations

We end the experimental results section with some observations on the convergence speed and the implementation of Algorithm 1.

First, we note that the two contraction mappings (11.23) and (12.6) are sensitive to the learning rate α. A very small choice of α results in convergence of the sequence of Q-functions to the fixed point being too slow to be practically useful. On the other hand, a large choice of α makes the sequence diverge. Thus, choosing α has a large effect on the runtime of the overall algorithm as the computation of Q^* and $D_\theta Q^*$ both depend on the choice of α.

We further remark that numerical observations suggest that the condition $\alpha \in (0, \min\{L^{-1}, 1\}]$ is fairly restrictive and that larger values of α give faster convergence. Hence, our implementation of Algorithm 1 includes an adaptive scheme to find the largest possible α. In

particular, if two consecutive iteration elements in the sequence are observed to diverge in the L_∞ norm, we decrease α by a fixed constant. As long successive elements in the sequence converge, we periodically increase α by another constant. This allows us to noticeably speed up the implementation of our algorithm. Adaptively choosing the step-size α also allows us to train the prospect function agents more accurately, since these were particularly susceptible to changes in the value of α due to the fact that the value function is non-Lipschitz around the reference point.

To speed up the gradient-descent algorithm, we also implement a back-tracking line search. We do this to address the computationally intensive gradient calculation. Specifically, the line-search allows us to exploit each gradient calculation fully. The backtracking line search also leads to a noticeable speed up in the implementation of our algorithm, which allows us to tackle larger MDPs.

12.3 Chapter Summary

We present a new gradient based technique for learning risk-sensitive decision-making models of agents operating in uncertain environments. We find that while there are a number of technical issues related to learning prospect theory based agents—namely, their value functions are not Lipschitz for parameter combinations that best capture human decision-making (i.e. when $0 < \zeta_-, \zeta_+ < 1$)—we are still able to numerically learn the policies of these agents. Moreover, we introduce a Lipschitz variation of the prospect value function, which retains the convex-concave structure of the prospect theory value function while satisfying the assumptions of our theorems on a bounded domain and possessing better numerical properties. We demonstrate the algorithm's performance for agents based on several types of behavioral models and do so on two examples: the canonical Grid World problem and a passenger's view ride-sharing where the parameters of the ride-sharing MDP are learned from real-world data.

Looking forward, we remark that we assumed the reference point was known and fixed. We are examining techniques that have formal performance guarantees (e.g., on convergence) that allow us to simultaneously estimate reference points. We are also examining the use of other risk-metrics beyond convex risk measures derived from acceptance sets that will allow us to leverage the benefits of cumulative prospect theory which has been shown to be more flexible in approximating human decision-making and has recently been adopted in the reinforcement learning literature [90].

Part IV

Future Directions: Algorithms in Societal-Systems

Chapter 13

Future Directions

This dissertation represents is a small part of what is really an emerging research agenda on developing algorithms for societal systems. Moving forward, there is a large scope for leveraging tools in dynamical systems, machine learning, stochastic analysis, and blending them with ideas game theory and behavioral economics to develop a fundamental understanding of the challenges posed by the deployment of learning algorithms in societal-scale systems.

An overarching goal at the heart of this research agenda is developing a *unified design methodology for learning algorithms* that not only have provable guarantees of performance for individual agents, but also achieve societal goals— i.e., system-wide objectives like fairness. This entails moving beyond analyzing the interactions between algorithms and developing a fundamental understanding of the consequences of game theoretic interactions between heterogeneous agents.

To achieve this goal, we see three broad themes which expand upon the foundations laid in this dissertation: (i) understanding the impacts of algorithmic decision making in economic settings, (ii) learning-based mechanism design, and (iii) high-confidence decision-making in dynamic environments.

Understanding the impacts of algorithmic decision making in economic settings. In this dissertation, we developed an understanding of the equilibria and dynamics of learning algorithms in competitive settings. Moving forwards, it will be crucial to understand more broadly the impacts of learning algorithms on societal objectives in societal-scale systems. Of particular importance is developing a game theoretic understanding of the *feedback loops* between learning algorithms and the data sources on which they rely. Indeed, there is a growing recognition of the fact that people are not merely sources of static, i.i.d data, but are in fact *strategic data sources* who can *dynamically* alter their data to achieve their goals. Examples of this abound in real-world settings, like the gaming of surge pricing algorithms by drivers[1] and understanding these issues can have a large impact on the deployment of algorithms in real-world settings [205].

[1] Uber drivers reportedly triggering higher fares through Surge Club, Digital Trends, 2019.

Learning-based mechanism design. Algorithms are deployed in real-world settings under the premise that they have beneficial impacts on societal systems, but there is growing evidence that suggests that poorly designed algorithms can have disastrous impacts on societal welfare (e.g. using click through rates as proxies for user engagement has led to the widespread proliferation of conspiracy theories on social networks). Core to this research agenda on algorithms in societal-systems is the development of learning-based mechanism design which entails designing algorithms for societal systems that can adapt to the game theoretic structure of the problems to achieve societal goals. This includes designing machine learning algorithms that are robust to— or can adapt to— the gamification of the data, and designing algorithms for online incentive design to drive systems towards better outcomes[108, 202].

High-confidence decision-making in dynamic environments. In this dissertation we focused on analyzing model-based learning in the simplest dynamic problem – multi-armed bandits– and extended the ideas to an algorithm for robotics that had impressive empirical successes. Despite this, understanding the fundamental limits of how to learn models and adapt them to uncertainty *online* is still an open question in complex settings like reinforcement learning and robotics, though there is no shortage of evidence that model-based methods are vastly more efficient than model-free methods at most tasks. As such, an interesting avenue of future work is to continue to develop a systematic understanding of how to learn and use models in uncertain, dynamic environments— a line of work that belongs in the emerging and exciting research area known as Assured Autonomy. The ultimate goal of this line of work is to develop model-based algorithms that are flexible enough to overcome issues of model mis-specification while having the safety and robustness guarantees required to be used in real world settings.

13.1 Concluding Remarks

The massive scale at which autonomous systems are being deployed has opened the doors for new opportunities for innovation in society. This new reality poses new challenges that can only be addressed through an interdisciplinary research agenda at the intersection of economics and engineering and computer science. This dissertation drew on tools from dynamical systems theory, statistics, machine learning, and economics to begin to develop a fundamental understanding of the challenges posed by the deployment of learning algorithms in societal-scale systems. Moving forwards, this research agenda has a tremendous potential to have a beneficial impact in application domains like matching markets, transport systems, and online platforms.

Bibliography

[1] P. Abbeel and A. Y. Ng. "Apprenticeship Learning via Inverse Reinforcement Learning". In: *Proceedings of the Twenty-first International Conference on Machine Learning*. 2004. DOI: 10.1145/1015330.1015430.

[2] S. Abdallah and V. Lesser. "A Multiagent Reinforcement Learning Algorithm with Non-linear Dynamics". In: *Journal of Artificial Intelligence Research* (2008).

[3] M. Abeille and A. Lazaric. "Linear Thompson sampling revisited". In: *Electron. J. Statist.* 11.2 (2017), pp. 5165–5197.

[4] J. Abernethy, K. A. Lai, and A. Wibisono. "Last-Iterate Convergence Rates for Min-Max Optimization: Convergence of Hamiltonian Gradient Descent and Consensus Optimization". In: *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*. Vol. 132. Proceedings of Machine Learning Research. PMLR, 2021, pp. 3–47.

[5] R. Abraham, J. E. Marsden, and T. Ratiu. *Manifolds, Tensor Analysis, and Applications*. 2nd. Springer, 1988.

[6] L. Adolphs et al. "Local saddle point optimization: A curvature exploitation approach". In: *Proceedings of the 22th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2019.

[7] S. Agrawal and N. Goyal. "Analysis of Thompson sampling for the multi-armed bandit problem". In: *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*. 2012, pp. 39.1–39.26.

[8] S. Agrawal and N. Goyal. "Further Optimal Regret Bounds for Thompson Sampling". In: *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2013, pp. 99–107.

[9] S. Agrawal and N. Goyal. "Thompson Sampling for Contextual Bandits with Linear Payoffs". In: *Proceedings of the 30th International Conference on Machine Learning (ICML)*. 2013, pp. 127–135.

[10] S. Amari. "Natural Gradient Works Efficiently in Learning". In: *Neural Computation* 10.2 (1998), pp. 251–276. DOI: 10.1162/089976698300017746.

[11] P. Artzner et al. "Coherent Measures of Risk". In: *Mathematical Finance* 9.3 (1999), pp. 203–228. DOI: 10.1111/1467-9965.00068.

[12] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47 (2002), pp. 235–256.

[13] E. Avineri and P. Bovy. "Identification of Parameters for a Prospect Theory Model for Travel Choice Analysis". In: *Transportation Research Record* 2082 (2008). DOI: 10.3141/2082-17.

[14] W. Azizian et al. "A Tight and Unified Analysis of Gradient-Based Methods for a Whole Spectrum of Differentiable Games". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 2863–2873.

[15] D. Balduzzi et al. "The Mechanics of n-Player Differentiable Games". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 354–363.

[16] B. Banerjee and J. Peng. "Adaptive Policy Gradient in Multiagent Learning". In: *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*. ACM, 2003, pp. 686–692. DOI: 10.1145/860575.860686.

[17] T. Bansal et al. "Emergent Complexity via Multi-Agent Competition". In: *International Conference on Learning Representations (ICLR)*. 2018.

[18] T. Basar. "On the uniqueness of the Nash solution in Linear-Quadratic differential Games". In: *International Journal of Game Theory* 5 (1976), pp. 65–90. DOI: 10.1007/BF01753310.

[19] T. Basar and G. Olsder. *Dynamic Noncooperative Game Theory*. Society for Industrial and Applied Mathematics, 1998.

[20] S. Basu and A. DasGupta. "The mean, median, and mode of unimodal distributions: a characterization". In: *Teor. Veroyatnost. i Primenen.* 41 (2 1996), pp. 336–352.

[21] M. Benaim and M. W. Hirsch. "Learning Processes, Mixed Equilibria and Dynamical Systems Arising from Repeated Games". In: *Games and Economic Behavior* (1997).

[22] M. Benaïm. "A Dynamical System Approach to Stochastic Approximations". In: *SIAM Journal on Control and Optimization* (1996).

[23] M. Benaïm. "Dynamics of stochastic approximation algorithms". In: *Séminaire de Probabilités XXXIII*. Springer Berlin Heidelberg, 1999, pp. 1–68.

[24] M. Benaïm and M. W. Hirsch. "Dynamics of Morse-Smale urn processes". In: *Ergodic Theory and Dynamical Systems* 15.6 (Dec. 1995).

[25] M. Benaïm and M. W. Hirsch. "Mixed equilibria and dynamical systems arising from fictitious play in perturbed games". In: *Games and Economic Behavior* 29 (1999), pp. 36–72.

[26] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[27] V. S. Borkar and S. P. Meyn. "Risk-Sensitive Optimal Control for Markov Decision Processes with Monotone Cost". In: *Mathematics of Operations Research* 27.1 (2002), pp. 192–209. DOI: 10.1287/moor.27.1.192.334.

[28] L. Bottou. "Large-Scale Machine Learning with Stochastic Gradient Descent". In: *Proceedings in Computational Statistics* (2010).

[29] M. Bravo, D. Leslie, and P. Mertikopoulos. "Bandit learning in concave N-person games". In: *Advances in Neural Information Processing Systems*. 2018.

[30] H. W. Broer and F. Takens. "Chapter 1–Preliminaries of Dynamical Systems Theory". In: *Handbook of Dynamical Systems*. Ed. by F Takens Henk Broer and B Hasselblatt. Vol. 3. Handbook of Dynamical Systems. Elsevier Science, 2010, pp. 1–42.

[31] P. Burnett. "Disaggregate behavioral models of travel decisions other than mode choice: a review and contribution to spatial choice theory". In: *Transportation Research Board Special Report* 149 (1974).

[32] D. Calderone et al. "Understanding the impact of parking on urban mobility via routing games on queue-flow networks". In: *2016 IEEE 55th Conference on Decision and Control (CDC)*. 2016, pp. 7605–7610.

[33] C. F. Camerer. "An experimental test of several generalized utility theories". In: *J. Risk and Uncertainty* 2.1 (1989), pp. 61–104. DOI: 10.1007/BF00055711.

[34] R. Cavazos-Cadena. "Optimality equations and inequalities in a class of risk-sensitive average cost Markov decision chains". In: *Mathematical Methods of Operations Research* 71.1 (2010), pp. 47–84. DOI: 10.1007/s00186-009-0285-6.

[35] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[36] O. Chapelle and L. Li. "An empirical evaluation of Thompson sampling". In: *Advances in Neural Information Processing Systems 24 (NeurIPS)*. 2011, pp. 2249–2257.

[37] B. Chasnov et al. "Convergence Analysis of Gradient-Based Learning in Continuous Games". In: ed. by Ryan P. Adams and Vibhav Gogate. Vol. 115. Proceedings of Machine Learning Research. Tel Aviv, Israel: PMLR, 2020, pp. 935–944.

[38] X. Cheng and P. L. Bartlett. "Convergence of Langevin MCMC in KL-divergence". In: *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*. 2018, pp. 186–211.

[39] A. S. Chivukula and W. Liu. "Adversarial learning games with deep learning models". In: *International Joint Conference on Neural Networks* (2017).

[40] C. Conley. "Isolated Invariant Sets and the Morse Index". In: *CBMS Regional Conference Series in Mathematics*. 1978.

[41] S. P. Coraluppi and S. I. Marcus. "Mixed risk-neutral/minimax control of discrete-time, finite-state Markov decision processes". In: *IEEE Trans. Autom. Control* 45.3 (2000), pp. 528–532. DOI: 10.1109/9.847737.

[42] A. S. Dalalyan and A. Karagulyan. "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient". In: *Stoch. Process. and their Appl.* 129.12 (2019), pp. 5278–5311.

[43] C. Daskalakis, P. Goldberg, and Papadimitriou C. "The complexity of computing a Nash Eequilibrium". In: *SIAM Journal on Computing* 39 (Feb. 2009), pp. 195–259.

[44] C. Daskalakis and I. Panageas. "The Limit Points of (Optimistic) Gradient Descent in Min-max Optimization". In: *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*. 2018.

[45] C. Daskalakis et al. "Traning GANs with Optimism". In: (2017). arXiv preprint.

[46] S. Dean et al. "On the Sample Complexity of the Linear Quadratic Regulator". In: *Foundations of Computational Mathematics* (2019). DOI: 10.1007/s10208-019-09426-y.

[47] J. L. Doob. "Application of the theory of martingales". In: *Le Calcul des Probabilites et ses Applications* (1949), pp. 23–27.

[48] A. Durmus and E. Moulines. "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm". In: *Ann. Appl. Probab.* 27.3 (June 2017), pp. 1551–1587.

[49] A. Durmus and E. Moulines. "Sampling from strongly log-concave distributions with the Unadjusted Langevin Algorithm". arXiv preprint. 2016.

[50] J. Engwerda. "On Scalar Feedback Nash Equilibria in the Infinite Horizon LQ-Game". In: *IFAC Proceedings Volumes* (1998).

[51] M. Fazel et al. "Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator". In: *International Conference of Machine Learning*. 2018.

[52] T. Fiez, B. Chasnov, and L. Ratliff. "Implicit Learning Dynamics in Stackelberg Games: Equilibria Characterization, Convergence Analysis, and Empirical Study". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 3133–3144.

[53] S. D. Flåm and G. H. Greco. "Non-Cooperative Games; Methods of Subgradient Projection and Proximal Point". In: *Advances in Optimization*. Springer Berlin Heidelberg, 1992, pp. 406–419.

[54] A. Flaxman, A. Kalai, and B. McMahan. "Online Convex Optimization in the Bandit Setting: Gradient Descent Without a Gradient". In: *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*. 2005.

[55] J. Foerster et al. "Learning with Opponent-Learning Awareness". In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[56] H. Föllmer and A. Schied. "Convex measures of risk and trading constraints". In: *Finance and Stochastics* 6.4 (2002), pp. 429–447. DOI: 10.1007/s007800200072.

[57] D. Fudenberg and D. K. Levine. *The theory of learning in games*. Vol. 2. MIT press, 1998.

[58] P. Geibel and F. Wysotzki. "Risk-Sensitive Reinforcement Learning Applied to Control under Constraints". In: *Journal of Artificial Intelligence Research* 24 (2005), pp. 81–108.

[59] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. "Convergence rates of posterior distributions". In: *Ann. Statist.* 28.2 (Apr. 2000), pp. 500–531.

[60] S. Ghosal and A. W. van der Vaart. "Convergence Rates of Posterior Distributions for Noniid Observations". In: *Ann. Statist.* 35.1 (2007), pp. 192–223.

[61] C. G. Gibson et al. "Topological Stability of Smooth Mappings". In: *Lecture Notes in Mathematics*. Vol. 552. Springer-Verlag, 1976.

[62] G. Gidel, T. Jebara, and S. Lacoste-Julien. "Frank-Wolfe Algorithms for Saddle Point Problems". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2017.

[63] G. Gidel et al. "A Variational Inequality Perspective on Generative Adversarial Networks". In: *International Conference on Learning Representations*. 2019.

[64] R. Giordano, T. Broderick, and M. I. Jordan. "Covariances, robustness, and variational Bayes". In: *Journal of Machine Learning Research* (2018).

[65] M. Golubitsky and V. Guillemin. *Stable Mappings and Their Singularities*. Springer-Verlag, 1973.

[66] C. A. Gómez-Uribe. "Online algorithms for parameter mean and variance estimation in dynamic regression". arXiv preprint. 2016.

[67] R. Gonzalez and G. Wu. "On the Shape of the Probability Weighting Function". In: *Cognitive Psychology* 38.1 (1999), pp. 129–166. DOI: 10.1006/cogp.1998.0710.

[68] I. J. Goodfellow et al. "Generative Adversarial Networks". In: *Advances in Neural Information Processing Systems 27*. 2014.

[69] A. Gopalan, S. Mannor, and Y. Mansour. "Thompson Sampling for Complex Online Problems". In: *Proceedings of the 31st International Conference on Machine Learning (ICML)*. Proceedings of Machine Learning Research. 2014, pp. 100–108.

[70] M. Heger. "Consideration of Risk in Reinforcement Learning". In: *Proc. Eleventh Inter. Conf. Machine Learning*. 1994, pp. 105–111.

[71] J. Heinonen. "Lectures on Lipschitz Analysis". In: *Lectures at the 14th Jyväskylä Summer School* (2004).

[72] A. Héliou, J. Cohen, and P. Mertikopoulos. "Learning with Bandit Feedback in Potential Games". In: *NIPS*. 2017.

[73] M. Heusel et al. "GANs trained by a two time-scale update rule converge to a local Nash equilibrium". In: *Advances in Neural Information Processing Systems 30*. Dec. 2017.

[74] M. W. Hirsch. *Differential topology*. Springer–Verlag, 1976.

[75] C. H. Hommes and M. Ochea. *Multiple Steady States, Limit Cycles and Chaotic Attractors in Evolutionary Games with Logit Dynamics*. CeNDEF Working Papers 10-04. Universiteit van Amsterdam, Center for Nonlinear Dynamics in Economics and Finance, 2010. URL: `https://ideas.repec.org/p/ams/ndfwpp/10-04.html`.

[76] C. H. Hommes and M. I. Ochea. "Multiple equilibria and limit cycles in evolutionary games with Logit Dynamics". In: *Games and Economic Behavior* (2012).

[77] M. Jaderberg et al. "Human-level performance in 3D multiplayer games with population-based reinforcement learning". In: *Science* (2019).

[78] C. Jin, P. Netrapalli, and M. Jordan. "What is Local Optimality in Nonconvex-Nonconcave Minimax Optimization?" In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. PMLR, 2020, pp. 4880–4889.

[79] C. Jin et al. "A short note on concentration inequalities for random vectors with subGaussian norm". arXiv preprint. 2019.

[80] M. I. Jordan. "Artificial Intelligence: The Revolution Hasn't Happened Yet". In: *Medium* (2018).

[81] D. Kahneman and A. Tversky. "Prospect Theory: An Analysis of Decision under Risk". In: *Econometrica* 47.2 (1979), pp. 263–291.

[82] R. E. Kalman. "Contributions to the theory of optimal control". In: *Boletin de la Sociedad Matematica Mexicana* 5 (1960).

[83] E. Kaufmann, N. Korda, and R. Munos. "Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis". In: *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*. 2012.

[84] J. Kelley. *General Topology*. Van Nostrand Reinhold Company, 1955.

[85] F. S. Koppelman. "Non-linear utility functions in models of travel choice behavior". In: *Transportation* 10.2 (1981), pp. 127–146. DOI: `10.1007/BF00165262`.

[86] N. Korda, E. Kaufmann, and R. Munos. "Thompson Sampling for 1-Dimensional Exponential Family Bandits". In: *Advances in Neural Information Processing Systems 26 (NeurRIPS)*. 2013, pp. 1448–1456.

[87] B. Koszegi and M. Rabin. "A Model of Reference-Dependent Preferences". In: *The Quarterly Journal of Economics* 121.4 (2006), pp. 1133–1165.

[88] A. Ya. Kruger. "On Fréchet Subdifferentials". In: *J. Mathematical Sciences* 116.3 (2003).

[89] H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

[90] P. L.A. et al. "Cumulative Prospect Theory Meets Reinforcement Learning: Prediction and Control". In: *Proc. 33rd Intern. Conf. on Machine Learning*. Vol. 48. 2016.

[91] T. L. Lai and H. Robbins. "Asymptotically Efficient Adaptive Allocation Rules". In: *Adv. Appl. Math.* 6 (1 1985), pp. 4–22.

[92] M. Lanctot et al. "A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning". In: *Advances in Neural Information Processing Systems 30*. 2017.

[93] A. Latif. "Banach Contraction Principle and Its Generalizations". In: ed. by Saleh Almezel, Qamrul Hasan Ansari, and Mohamed Amine Khamsi. Springer International Publishing, 2014, pp. 33–64. DOI: 10.1007/978-3-319-01586-6_2.

[94] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[95] M. Ledoux. "Concentration of measure and logarithmic Sobolev inequalities". In: *Seminaire de probabilites XXXIII*. Springer, 1999, pp. 120–216.

[96] M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical surveys and monographs. American Mathematical Society, 2001.

[97] J. Lee. *Introduction to smooth manifolds*. Springer, 2012.

[98] J. D. Lee et al. "Gradient Descent Only Converges to Minimizers". In: *29th Annual Conference on Learning Theory*. 2016.

[99] D. S. Leslie and E. J. Collins. "Individual Q-Learning in Normal Form Games". In: *SIAM J. Control and Optimization* (2005).

[100] Q. Li et al. "A reference-dependent user equilibrium model for activity-travel scheduling". In: *Transportation* 43.6 (2016), pp. 1061–1077. DOI: 10.1007/s11116-016-9725-3.

[101] S. Li et al. "Robust Multi-Agent Reinforcement Learning via Minimax Deep Deterministic Policy Gradient". In: *Proceedings of the AAAI Conference*. 2019.

[102] T-Y. Li and Z. Gajic. "Lyapunov Iterations for Solving Coupled Algebraic Riccati Equations of Nash Differential Games and Algebraic Riccati Equations of Zero-Sum Games". In: *New Trends in Dynamic Games and Applications*. 1995.

[103] R. Lowe et al. "Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments". In: *Advances in Neural Information Processing Systems 30*. 2017.

[104] X. Lu and B. Van Roy. "Ensemble Sampling". In: *Advances in Neural Information Processing Systems 32 (NeurIPS)*. 2017, pp. 3260–3268.

[105] D. L. Lukes and D. L. Russell. "A global theory for linear-quadratic differential games". In: *Journal of Mathematical Analysis and Applications* (1971).

[106] Y.-A Ma, T. Chen, and E. Fox. "A Complete Recipe for Stochastic Gradient MCMC". In: *Advances in Neural Information Processing Systems 28 (NeurIPS)*. 2015, pp. 2917–2925.

[107] Y.-A. Ma et al. "Sampling Can be Faster than Optimization". In: *Proc. Natl. Acad. Sci. U.S.A.* 116.42 (2019), pp. 20881–20885.

[108] C. Maheshwari et al. "Zeroth-Order Methods for Convex-Concave Minmax Problems: Applications to Decision-Dependent Risk Minimization". In: (2021). arXiv preprint.

[109] A. Majumdar et al. "Risk-sensitive Inverse Reinforcement Learning via Coherent Risk Models". In: *Robotics: Science and Systems*. 2017.

[110] D. Malik et al. "Derivative-Free Methods for Policy Optimization: Guarantees for Linear Quadratic Systems". In: *Proceedings of Machine Learning Research*. 2019.

[111] S. I. Marcus et al. "Risk Sensitive Markov Decision Processes". In: *Systems and Control in the Twenty-First Century*. Ed. by Christopher I. Byrnes et al. Birkhäuser Boston, 1997, pp. 263–279. DOI: 10.1007/978-1-4612-4120-1_14.

[112] P. Massart. "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality". In: *The Annals of Probability* 18 (1990), pp. 1269–1283.

[113] E. Mazumdar, M. Jordan, and S. Sastry. "On Finding Local Nash Equilibria (and only Local Nash equilibria) in Zero-sum Continuous Games". In: (2019). arXiv preprint.

[114] E. Mazumdar and L. Ratliff. "Local Nash Equilibria are Isolated, Strict Local Nash Equilibria in 'Almost All' Zero-Sum Continuous Games". In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. 2019, pp. 6899–6904.

[115] E. Mazumdar, L. Ratliff, and S. Sastry. "Inverse Risk-Sensitive Reinforcement Learning via Gradient Methods". In: *IEEE Conference on Decision and Control (CDC)* (2017).

[116] E. Mazumdar, L. Ratliff, and S. Sastry. "On Gradient-Based Learning in Continuous Games". In: *SIAM Journal on Mathematics of Data Science* 2 (Jan. 2020), pp. 103–131.

[117] E. Mazumdar et al. "High Confidence Sets for Trajectories of Stochastic Time-Varying Nonlinear Systems". In: *2020 IEEE 59th Conference on Decision and Control (CDC)* (2020).

[118] E. Mazumdar et al. "On Approximate Thompson Sampling with Langevin Algorithms". In: *International Conference on Machine Learning (ICML)* (2020).

[119] E. Mazumdar et al. "Policy Gradient Algorithms Have No Guarantees of Convergence in Linear Quadratic Games." In: *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (2020).

[120] P. Mertikopoulos. "Online optimization and learning in games: Theory and applications". In: 2019.

[121] P. Mertikopoulos, C. H. Papadimitriou, and G. Piliouras. "Cycles in adversarial regularized learning". In: *Proceedings of the 29th annual ACM-SIAM symposium on discrete algorithms*. 2018.

[122] P. Mertikopoulos and M. Staudigl. "On the Convergence of Gradient-Like Flows with Noisy Gradient Input". In: *SIAM Journal on Optimization* (2018).

[123] P. Mertikopoulos and Z. Zhou. "Learning in games with continuous action sets and unknown payoff functions". In: *Mathematical Programming* (2019).

[124] P. Mertikopoulos et al. "Mirror descent in saddle-point problems: Going the extra (gradient) mile". In: *CoRR* abs/1807.02629 (2018).

[125] L. M. Mescheder, S. Nowozin, and A. Geiger. "The numerics of GANs". In: *Advances in Neural Information Processing Systems 30*. 2017.

[126] O. Mihatsch and R. Neuneier. "Risk-Sensitive Reinforcement Learning". In: *Machine Learning* 49.2 (2002), pp. 267–290. DOI: 10.1023/A:1017940631555.

[127] P. W. Millar. "Asymptotic minimax theorems for the sample distribution function". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 48.3 (1979), pp. 233–252. DOI: 10.1007/BF00537522.

[128] A. Mokhtari, A. Ozdaglar, and S. Pattathil. "A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1497–1507.

[129] D. Monderer and L. S. Shapley. "Potential Games". In: *Games and Economic Behavior* 14 (1996).

[130] W. Mou et al. "A Diffusion Process Perspective on Posterior Contraction Rates for Parameters". arXiv preprint. 2019.

[131] K. Murphy. "Conjugate Bayesian analysis of the Gaussian distribution". 2007.

[132] V. Nagarajan and Z. Kolter. "Gradient descent GAN optimization is locally stable". In: *Advances in Neural Information Processing Systems 30*. 2017.

[133] A. J. Nagengast, D. A. Braun, and D. M. Wolpert. "Risk-Sensitive Optimal Feedback Control Accounts for Sensorimotor Behavior under Uncertainty". In: *PLOS Computational Biology* 6.7 (2010), pp. 1–15. DOI: 10.1371/journal.pcbi.1000857.

[134] J. Nash. "Non-cooperative games". In: *Annals of Mathematics* 54 (1951), pp. 289–295.

[135] A. Nemirovski et al. "Robust Stochastic Approximation Approach to Stochastic Programming". In: *SIAM Journal on Optimization* (2009).

[136] G. Neu and C. Szepesvári. "Apprenticeship Learning Using Inverse Reinforcement Learning and Gradient Methods". In: *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*. 2007, pp. 295–302.

[137] A. Y. Ng and S. Russell. "Algorithms for Inverse Reinforcement Learning". In: *Proceedings of the 17th International Conference on Machine Learning*. 2000, pp. 663–670.

[138] N. Nisan et al. *Algorithmic Game Theory*. Cambridge, UK: Cambridge University Press, 2007.

[139] B. Øksendal. *Stochastic Differential Equations*. 6th. Springer, Berlin, 2003.

[140] S. Omidshafiei et al. "Deep Decentralized Multi-task Multi-Agent Reinforcement Learning under Partial Observability". In: (2017). arXiv preprint.

[141] M. Osborne. *A Course in Game Theory*. MIT Press, 1994.

[142] R. S. Palais. "Morse theory on Hilbert manifolds". In: *Topology* 2.4 (1963), pp. 299–340.

[143] J. Palis and S. Smale. "Structural Stability Theorems". In: *Proceedings of the Symposium on Pure Mathematics* (1970).

[144] I. Panageas and G. Piliouras. "Gradient Descent Only Converges to Minimizers: Non-Isolated Critical Points and Invariant Regions". In: *Innovations in Theoretical Computer Science*. 2016.

[145] C. Papadimitriou and G. Piliouras. "Game Dynamics As the Meaning of a Game". In: *ACM SIGecom Exchanges* (2018).

[146] R. Pemantle. "A survey of random processes with reinforcement". In: *Probability Surveys* 4.1–79 (2007).

[147] R. Pemantle. "Nonconvergence to Unstable Points in Urn Models and Stochastic Approximations". In: *Annuals Probability* (1990).

[148] J. P. Penot. "On the Interchange of Subdifferentiation and Epi-Convergence". In: *J. Mathematical Analysis and Applications* 196.2 (1995), pp. 676–698. DOI: 10.1006/jmaa.1995.1434.

[149] J. Peters, S. Vijayakumar, and S. Schaal. "Natural Actor-critic". In: *Proceedings of the 16th European Conference on Machine Learning*. 2005, pp. 280–291. DOI: 10.1007/11564096_29.

[150] M. Phan, Y. A. Yadkori, and J. Domke. "Thompson Sampling and Approximate Inference". In: *Advances in Neural Information Processing Systems 32 (NeurIPS)*. 2019, pp. 8804–8813.

[151] L. Pint et al. "Robust Adversarial Reinforcement Learning". In: *Proceedings of the International Conference on Machine Learning*. 2017.

[152] C. Possieri and M. Sassano. "An algebraic geometry approach for the computation of all linear feedback Nash equilibria in LQ differential games". In: *2015 54th IEEE Conference on Decision and Control (CDC)*. 2015.

[153] L. Ratliff and E. Mazumdar. "Inverse Risk-Sensitive Reinforcement Learning". In: *IEEE Transactions on Automatic Control* 65.3 (2020), pp. 1256–1263.

[154] L. Ratliff et al. "To observe or not to observe: Queuing game framework for urban parking". In: *2016 IEEE 55th Conference on Decision and Control (CDC)*. 2016, pp. 5286–5291.

[155] L. J. Ratliff, S. A. Burden, and S. S Sastry. "On the Characterization of Local Nash Equilibria in Continuous Games". In: *IEEE Transactions on Automatic Control* (2016).

[156] L. J. Ratliff, S. A. Burden, and S. S. Sastry. "Characterization and computation of local Nash equilibria in continuous games". In: *Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing*. 2013.

[157] L. J. Ratliff, S. A. Burden, and S. S. Sastry. "Generictiy and Structural Stability of Non–Degenerate Differential Nash Equilibria". In: *Proceedings of the American Control Conference*. 2014.

[158] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. "Maximum Margin Planning". In: *Proceedings of the 23rd International Conference on Machine Learning*. 2006, pp. 729–736. DOI: 10.1145/1143844.1143936.

[159] Y. Ren. "On the Burkholder-Davis-Gundy inequalities for continuous martingales". In: *Stat. Probabil. Lett.* 78.17 (2008), pp. 3034–3039.

[160] C. Riquelme, Tucker G., and J. Snoek. "Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling". arXiv preprint. 2018.

[161] J. W. Robbin. "A Structural Stability Theorem". In: *Annals of Mathematics* (1971).

[162] H. Robbins and S. Monro. "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

[163] H. Robbins and D. Siegmund. "A Convergence Theorem for Non Negative Almost Supermartingales and Some Applications". In: *Herbert Robbins Selected Papers*. Springer New York, 1985.

[164] J. B. Rosen. "Existence and Uniqueness of Equilibrium Points for Concave N-Person Games". In: *Econometrica* (1965).

[165] D. Russo and B. Van Roy. "An Information-Theoretic Analysis of Thompson Sampling". In: *J. Mach. Learn. Res.* 17 (1 2016), pp. 1–30.

[166] D. Russo et al. "A Tutorial on Thompson Sampling". arXiv preprint. 2017.

[167] Herbert S. "Bounded rationality in social science: Today and tomorrow". In: *Mind & Society* 1.1 (Mar. 2000), pp. 25–39. DOI: 10.1007/bf02512227.

[168] S. Sastry. *Nonlinear Systems*. Springer New York, 1999.

[169] A. Saumard and J. A. Wellner. "Log-concavity and strong log-concavity: A review". In: *Statist. Surv.* 8 (2014), pp. 45–114.

[170] D. Scieur et al. "Integration Methods and Optimization Algorithms". In: *Advances in Neural Information Processing Systems 30*.

[171] S. L. Scott. "A modern Bayesian look at the multi-armed bandit". In: *Applied Stochastic Models in Business and Industry* 26.6 (2010), pp. 639–658.

[172] X. Shen and L. Wasserman. "Rates of convergence of posterior distributions". In: *Ann. Statist.* 29.3 (June 2001), pp. 687–714.

[173] Y. Shen, W. Stannat, and K. Obermayer. "Risk-Sensitive Markov Control Processes". In: *SIAM J. Control Optimization* 51.5 (2013), pp. 3652–3672.

[174] Y. Shen, M. J. Tobia, and K. Obermayer. "Risk-Sensitive Reinforcement Learning". In: *Neural Computation* 26 (2014), pp. 1298–1328.

[175] M. Shub. *Global Stability of Dynamical Systems*. Springer-Verlag, 1978.

[176] S. P. Singh, M. J. Kearns, and Y. Mansour. "Nash Convergence of Gradient Dynamics in General-Sum Games". In: *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*. 2000.

[177] S. Smale. "Differentiable Dynamical Systems". In: *Bulletin of the American Mathematical Society* (1967).

[178] S. Srinivasan et al. "Actor-Critic Policy Optimization in Partially Observable Multiagent Environments". In: *Advances in Neural Information Processing Systems 31*. 2018.

[179] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT press, 2017.

[180] W. Thompson. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". In: *Biometrika* 25.3/4 (1933), pp. 285–294.

[181] J. N. Tsitsiklis. "Asynchronous Stochastic Approximation and Q-Learning". In: *Machine Learning* 16.3 (1994), pp. 185–202. DOI: 10.1023/A:1022689125041.

[182] A. Tversky and D. Kahneman. "Advances in prospect theory: Cumulative representation of uncertainty". In: *Journal of Risk Uncertainty* 5.4 (1992), pp. 297–323. DOI: 10.1007/bf00122574.

[183] A. Tversky and D. Kahneman. "Loss Aversion in Riskless Choice: A Reference-Dependent Model". In: *The Quarterly Journal of Economics* 106.4 (1991), pp. 1039–1061. DOI: 10.2307/2937956.

[184] A. Tversky and D. Kahneman. "Rational Choice and the Framing of Decisions". In: *The Journal of Business* 59.4 (1986), pp. S251–S278.

[185] A. Tversky and D. Kahneman. "The framing of decisions and the psychology of choice". In: *Science* 211.4481 (Jan. 1981), pp. 453–458. DOI: 10.1126/science.7455683.

[186] I. Urteaga and C. Wiggins. "Variational inference for the multi-armed contextual bandit". In: *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2018, pp. 698–706.

[187] A. W. van der Vaart and J. H. van Zanten. "Rates of contraction of posterior distributions based on Gaussian process priors". In: *Ann. Statist.* 36.3 (June 2008), pp. 1435–1463.

[188] S. Vempala and A. Wibisono. "Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices". In: *Advances in Neural Information Processing Systems 32 (NeurIPS)*. 2019, pp. 8094–8106.

[189] C. Villani. *Optimal Transport: Old and New*. Wissenschaften. Springer, Berlin, 2009.

[190] O. Vinyals et al. "Grandmaster level in StarCraft II using multi-agent reinforcement learning". In: *Nature* (2019).

[191] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[192] M. Welling and Y. W. Teh. "Bayesian learning via stochastic gradient Langevin dynamics". In: *Proceedings of the 28th international conference on machine learning (ICML)*. 2011, pp. 681–688.

[193] E. Wesson and R. Rand. "Hopf Bifurcations in Delayed Rock–Paper–Scissors Replicator Dynamics". In: *Dynamic Games and Applications* (2016).

[194] T. Westenbroek et al. "Adaptive Control for Linearizable Systems Using On-Policy Reinforcement Learning". In: *2020 IEEE 59th Conference on Decision and Control (CDC)* (2020).

[195] T. Westenbroek et al. "Feedback Linearization for Unknown Systems via Reinforcement Learning". In: *International Conference on Robotics and Automation (ICRA)* (2020).

[196] T. Westenbroek et al. *Technical Report: Adaptive Control for Linearizable Systems Using On-Policy Reinforcement Learning*. arXiv preprint. 2020.

[197] Ashia C. Wilson, Benjamin Recht, and Michael I. Jordan. "A Lyapunov Analysis of Momentum Methods in Optimization". In: (2016). arXiv preprint.

[198] G. Wu and R. Gonzalez. "Curvature of the Probability Weighting Function". In: *Management Science* 42.12 (1996), pp. 1676–1690. DOI: 10.1287/mnsc.42.12.1676.

[199] H. Xu, C. Caramanis, and S. Mannor. "Robustness and regularization of support vector machines". In: *Journal of Machine Learning Research* 10 (Dec. 2009), pp. 1485–1510. ISSN: 1532-4435.

[200] H. Xu, J. Zhou, and W. Xu. "A decision-making rule for modeling travelers' route choice behavior based on cumulative prospect theory". In: *Transportation Research Part C: Emerging Technologies* 19.2 (2011), pp. 218–228. DOI: 10.1016/j.trc.2010.05.009.

[201] L. Yang. "Active learning with a drifting distribution". In: *Advances in Neural Information Processing Systems*. 2011.

[202] Y. Yu et al. "Fast Distributionally Robust Learning with Variance Reduced Min-Max Optimization". In: *arXiv preprint* (2021).

[203] C. Zhang and V. Lesser. "Multi-Agent Learning with Policy Prediction". In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. 2010.

[204] K. Zhang, Z. Yang, and T. Basar. "Policy Optimization Provably Converges to Nash Equilibria in Zero-Sum Linear Quadratic Games". In: *Advances in Neural Information Processing Systems 32*. 2019.

[205] T. Zrnic et al. *Who Leads and Who Follows in Strategic Classification?* arXiv preprint. 2021.