Title
Learning in Mean-Field Games and Continuous-Time Stochastic Control Problems

Author
Hu, Anran

Publication Date
2022

Peer reviewed|Thesis/dissertation

Learning in Mean-Field Games and Continuous-Time Stochastic Control Problems

by

Anran Hu

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Xin Guo, Chair
Professor Anil Aswani
Professor Paul Grigas
Professor Jiantao Jiao

Summer 2022

Learning in Mean-Field Games and Continuous-Time Stochastic Control Problems

Abstract

Learning in Mean-Field Games and Continuous-Time Stochastic Control Problems

by

Anran Hu

Doctor of Philosophy in Engineering – Industrial Engineering and Operations Research

University of California, Berkeley

Professor Xin Guo, Chair

In recent years, there has been an ever-increasing demand for building reliable and versatile agents in applications arising from numerous fields including autonomous driving, supply chain, manufacturing, e-commerce and finance. To meet these challenging demands, researches in decision making systems have drawn upon a wide range of tools from applied probability, reinforcement learning (RL), stochastic control and game theory. This dissertation focuses on developing new methodologies and efficient algorithms with provable performance guarantees to deal with complex environments such as large population competitions and continuous-time systems.

The first part of this dissertation focuses on designing and analyzing RL algorithms for large population games. Large population games have appeared in many real-world problems. Examples include massive multiplayer online role-playing games, high frequency trading, and the sharing economy. However, in general, it becomes increasingly difficult to solve such problems as the number of players in the game grows. Mean field game (MFG) provides an ingenious and tractable aggregation approach to approximate the otherwise challenging N-player stochastic games. In Chapter 1, we present a general mean-field game (GMFG) framework for simultaneous learning and decision-making in stochastic games with a large population. It first establishes the existence of a unique Nash Equilibrium to this GMFG, and demonstrates that naively combining reinforcement learning with the fixed-point approach in classical MFGs yields unstable algorithms. It then proposes value-based and policy-based reinforcement learning algorithms (GMF-V and GMF-P, respectively) with smoothed policies, with analysis of their convergence properties and computational complexities. Experiments on an equilibrium product pricing problem demonstrate that GMF-V-Q and GMF-P-TRPO, two specific instantiations of GMF-V and GMF-P, respectively, with Q-learning and TRPO, are both efficient and robust in the GMFG setting. Moreover, their performance is superior in convergence speed, accuracy, and stability when compared with existing algorithms for multi-agent reinforcement learning in the N-player setting.

The second part of this dissertation focuses on designing and analyzing RL algorithms for continuous-time stochastic dynamical systems. As most physical systems in science and engineering evolve continuously in time, many real-world control tasks, such as those in aerospace, automotive industry and robotics, are naturally formulated in terms of continuous-time dynamical systems. Nevertheless, the mainstream RL algorithms have been designed for discrete-time systems, despite that they are widely applied to physical tasks in continuous-time systems. Continuous-time RL algorithms have also been developed in the past decades. But the theoretical guarantees of these works are limited to the asymptotic convergence and the non-asymptotic guarantees remain unknown. In Chapter 2, we take the first step towards designing algorithms with non-asymptotic guarantees for solving finite-time-horizon continuous-time linear quadratic (LQ) RL problems in an episodic setting, where both the state and control coefficients are unknown to the controller. We first propose a least-squares algorithm based on continuous-time observations and controls, and establish a logarithmic regret bound of magnitude $\mathcal{O}((\ln M)(\ln\ln M))$, with M being the number of learning episodes. The analysis consists of two components: perturbation analysis, which exploits the regularity and robustness of the associated Riccati differential equation; and parameter estimation error, which relies on sub-exponential properties of continuous-time least-squares estimators. We further propose a practically implementable least-squares algorithm based on discrete-time observations and piecewise constant controls, which achieves similar logarithmic regret with an additional term depending explicitly on the time stepsizes used in the algorithm. In Chapter 3, we extend the results beyond linear-quadratic problems, where the unknown linear jump-diffusion process is controlled subject to non-smooth convex costs. We show that the associated linear-convex (LC) control problems admit Lipchitz continuous optimal feedback controls and further prove the Lipschitz stability of the feedback controls. The analysis relies on a stability analysis of the associated forward-backward stochastic differential equation. We then propose a least-squares algorithm which achieves a regret of the order $\mathcal{O}(\sqrt{N\ln N})$ on linear-convex learning problems with jumps, where N is the number of learning episodes; the analysis leverages the Lipschitz stability of feedback controls and concentration properties of sub-Weibull random variables. Numerical experiment confirms the convergence and the robustness of the proposed algorithm.

To Bolin and Xiuzhen

Contents

List of Figures

List of Tables

Acknowledgments

First, I am really fortunate to be advised by Professor Xin Guo, who not only has provided me with great guidance on my researches during my PhD studies, but also has been caring and supporting in my daily life. As early as in my first year of PhD, I enjoyed and learned a lot in her course on applied stochastic processes. When I was at the beginning of PhD years and knew little about how to do researches, she encouraged me a lot and gave me faith in becoming a good researcher. Working with Xin has always been a pleasant and fruitful journey. She is sharp and can always provide insightful suggestions and comments in our every discussion. In addition, she has always been a great role model to me as a successful female researcher. She taught me to be strong, confident, persistent and open-minded. I would never become who I am now without her instruction and guidance.

Next, I would like to thank my thesis and oral defense committee members, Professor Anil Aswani, Professor Paul Grigas and Professor Jiantao Jiao. I really appreciate Anil for the discussions we had on research topics in my early years. I am also grateful for Paul sharing his teaching materials on IEOR 242, which was super helpful to me when I taught this course as an instructor for the first time in my life. I also sincerely thank Jiantao for agreeing to be my committee in the last-minute.

I would also like to thank my manager Xinyang Shen and my mentor Jingchen Wu during my internship at Amazon in the summer of 2019. They taught me how to apply what I knew to solve real world problems, and I learned a lot from them on how to deal with massive data and build simple models to tackle difficult problems.

In addition, I am very lucky to be in a research group with a caring and encouraging atmosphere. I am greatly thankful to Renyuan Xu, who is like a sister to me and was extremely helpful in my earlier stage of researches. She provided me with so much support to encourage me to walk through the hard time. I would also like to thank Yufei Zhang, who is aways enthusiastic and encouraging. I am deeply impressed by how smart yet modest he is and I really enjoy the collaborations with him. I am also grateful to Jiacheng Zhang who is always nice and supportive. The discussions with him have always been pleasant and insightful. I would also like to thank Matteo Basei, who has been a great collaborator and I have learned a lot about presentation skills from him. Besides the wonderful collaborators, I also sincerely thank my talented research group members Haoyang Cao, Xiaoli Wei, Nan Yang, Wenpin Tang, Yusuke Kikuchi, Haotian Gu, Xinyu Li, Mahan Tajrobehkar, Alberto Gennaro, Othmane Mounjid and Fangyuan Zhao. I enjoyed a lot the group lunches and discussions together.

Moreover, I would like to express my sincere gratitude to my friends and cohorts: Jin Xie, Yongyi Guo, Qitian Chen, Jingjing Bai, Zhimei Ren, Shuqi Cen, Shi Dong, Ya Wen, Zeyu Zheng, Junyu Cao, Ying Cao, Mo Zhou, Jiaying Shi, Xu Rao, Han Feng, Meng Qi, Tianyi Lin, Heyuan Liu, Hansheng Jiang, Yuhao Ding, Quico Spaen, Erik Bertelli, SangWoo Park, Shunan Jiang, Mo Liu, Mengxin Wang, Alfonso Lobos and Cristobal Pais.

Finally, I'm deeply indebted to my parents, Bolin Hu and Xiuzhen Gui, who have always been there to support me. Their love has always been my source of courage to move forward.

Last but not least, I would also like to sincerely thank my husband, Junzi Zhang, who has always been by my side.

Chapter 1

A General Framework for Learning Mean-Field Games

1.1 Introduction

Motivating example. This paper is motivated by the following Ad auction problem for an advertiser. An Ad auction is a stochastic game on an Ad exchange platform among a large number of players, the advertisers. In between the time a web user requests a page and the time the page is displayed, usually within a millisecond, a Vickrey-type of second-best-price auction is run to incentivize interested advertisers to bid for an Ad slot to display advertisement. Each advertiser has limited information before each bid: first, her own valuation for a slot depends on some random conversion of clicks for the item; secondly, she, should she win the bid, only knows the reward after the users activities on the website are finished. In addition, she has a budget constraint in this repeated auction.

The question is, how should she bid in this online sequential repeated game when there is a large population of bidders competing on the Ad platform, with random conversions of clicks and rewards?

Besides Ad auctions, there are many other real-world problems involving a large number of players and uncertain systems. Examples include massive multi-player online role-playing games [91], high frequency tradings [106], and the sharing economy [79].

Our work. Motivated by these problems, we consider a general framework of simultaneous learning and decision-making in stochastic games with a large population. We formulate a general mean-field-game (GMFG) with incorporation of action distributions and (randomized) relaxed policies. This general framework can also be viewed as a generalized version of MFGs of extended McKean-Vlasov type [3], which is a different paradigm from the classical MFG. It is also beyond the scope of the existing reinforcement learning (RL) framework for Markov decision processes (MDP), as MDP is technically equivalent to a single player stochastic game.

On the theory front, this general framework differs from the existing MFGs. We establish under appropriate technical conditions the existence and uniqueness of the Nash equilibrium (NE) to this GMFG. On the computational front, we show that naively combining reinforcement learning with the three-step fixed-point approach in classical MFGs yields unstable algorithms. We then propose both value based and policy based reinforcement learning algorithms with smoothed policies (GMF-V and GMF-P, respectively), establish the convergence property and analyze the computational complexity (see Section 1.7 for all proof details). Finally, we apply GMF-V-Q and GMF-P-TRPO, which are two specific instantiations of GMF-V and GMF-P, respectively, with Q-learning and TRPO, to an equilibrium product pricing problem[1]. Both algorithms have demonstrated to be efficient and robust in the GMFG setting. Their performance is superior in terms of convergence speed, accuracy and stability, when compared with existing algorithms for multi-agent reinforcement learning in the N-player setting. Note that an earlier and preliminary version [76] has been published in NeurIPS. Nevertheless, the conference version focuses only on GMF-V-Q, whereas this paper provides a new meta framework for learning mean-field-game which combines (1) the three-step fixed point approach, (2) the smoothing techniques, and (3) the single-agent algorithms with sample complexity guarantees in the sub-routine. This general framework incorporates both value-based algorithms and policy-based algorithms. In addition, the policy-based RL algorithm (GMF-P-TRPO) in this paper is the first globally convergent policy-based algorithm for solving mean-field-games. Numerical results show that it achieves similar performance as the Q-learning based algorithm (GMF-V-Q) in [76].

Related works. On learning large population games with mean-field approximations, [166] focuses on inverse reinforcement learning for MFGs without decision making, with its extension in [36] for agent-level inference; [167] studies an MARL problem with a first-order mean-field approximation term modeling the interaction between one player and all the other finite players, which has been generalized to the setting with partially observable states in [146]; and [95] and [168] consider model-based adaptive learning for MFGs in specific models (*e.g.*, linear-quadratic and oscillator games). More recently, [115] studies the local convergence of actor-critic algorithms on finite time horizon MFGs, and [144] proposes a policy-gradient based algorithm and analyzes the so-called local NE for reinforcement learning in infinite time horizon MFGs. For learning large population games without mean-field approximation, see [82, 93] and the references therein. In the specific topic of learning auctions with a large number of advertisers, [28] and [92] explore reinforcement learning techniques to search for social optimal solutions with real-word data, and [90] uses MFGs to model the auction system with unknown conversion of clicks within a Bayesian framework.

However, none of these works consider the problem of simultaneous learning and decision-making in a general MFG framework. Neither do they establish the existence and uniqueness of the (global) NE, nor do they present model-free learning algorithms with complexity

[1]The numerical experiments on the application of GMF-V-Q to the motivating Ad auction problem can be found in the conference version of our paper [76].

analysis and convergence to the NE. Note that in principle, global results are harder to obtain compared to local results.

Following the conference version [76] of the current paper, various efforts have been made to extend our reinforcement learning work in [76] to more general MFG settings. These include linear-quadratic MFGs in both discrete-time setting [60, 153, 154] and in continuous-time setting [77, 163, 47], MFGs with general continuous state and/or action spaces [7], entropy regularized MFGs in discrete time [7, 164, 165, 41] and in continuous time [77], and non-stationary MFGs [117]. In particular, [41] interprets the softmax smoothing technique proposed in [76] from a smoothed equilibrium perspective. In addition, different frameworks based on monotonicity assumptions (instead of the contractivity assumption in [76]) have also been proposed, and fictitious play algorithms with policy and mean-field averaging [52, 131] and online mirror descent algorithms [129] have been proposed to solve MFGs under such assumptions. There are also some recent extensions to reinforcement learning of MFGs with strategic complementarity [105] and multiple agent types [66, 145]. These algorithms for reinforcement learning of MFGs have also been applied in economics [10], in finance [44], in animal behavior simulation [130], and in concave utility reinforcement learning [64]. In the meantime, the idea of simultaneous learning and decision making with mean-field interaction has been used for analyzing collaborative games with social optimal solution [31, 32, 73, 110, 162, 127, 61, 42, 9].

Notations. Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space and $\mathcal{X}$ is equipped with the Borel σ-field $\mathcal{B}(\mathcal{X})$, meaning the σ-field generated by the open sets of $\mathcal{X}$. Denote $\mathcal{P}(\mathcal{X})$ for the set of (Borel) probability measures on $\mathcal{X}$. $\mathcal{W}_p$ denotes the Wasserstein distance of order p such that

$$W_p(\mu,\mu') = \inf\left\{\left(\int_{\mathcal{X}\times\mathcal{X}} d^p_{\mathcal{X}}(x,x')\nu(dx,dx')\right)\colon \nu\in\mathcal{P}(\mathcal{X}\times\mathcal{X}) \text{ with marginals } \mu,\mu'\in\mathcal{P}(\mathcal{X})\right\}.$$

$\mathcal{P}(\mathcal{X})$ is always equipped with $W_1(\mu,\mu')$. The Borel σ-field of $\mathcal{P}(\mathcal{X})$ is the σ-field induced by the evaluation $\mathcal{P}(\mathcal{X})\ni\mu\mapsto\mu(C)$ for any Borel set $C\subset\mathcal{X}$. Note that the Borel σ-field of $\mathcal{P}(\mathcal{X})$ is generated by W_1. (See e.g. [157] and [100]).

Given two measurable spaces $(\mathcal{Y},\mathcal{B}(\mathcal{Y}))$ and $(\mathcal{X},\mathcal{B}(\mathcal{X}))$, we say a measure-valued function $f:\mathcal{Y}\to\mathcal{P}(\mathcal{X})$ is measurable if $\Lambda_C\circ f:\mathcal{Y}\to[0,1]$ is measurable for any $C\in\mathcal{B}(\mathcal{X})$, where $\Lambda_C:\mathcal{P}(\mathcal{X})\ni\mu\mapsto\mu(C)\in[0,1]$.

1.2 Framework of General MFG (GMFG)

1.2.1 Background: classical N-player Markovian game and MFG

Let us first recall the classical N-player game. There are N players in a game. At each step t, the state of player i $(=1,2,\cdots,N)$ is $s^i_t\in\mathcal{S}\subseteq\mathbb{R}^d$ and she takes an action $a^i_t\in\mathcal{A}\subseteq\mathbb{R}^p$. Here d,p are positive integers. The state space $(\mathcal{S},d_{\mathcal{S}})$ and the action space $(\mathcal{A},d_{\mathcal{A}})$ are two compact metric spaces, including the case of $\mathcal{S}$ and $\mathcal{A}$ being finite. Given

the current state profile of N-players $\boldsymbol{s}_t = (s_t^1, \dots, s_t^N) \in \mathcal{S}^N$ and the action a_t^i, player i will receive a reward $r^i(\boldsymbol{s}_t, a_t^i)$ sampled from a distribution $R^i(\boldsymbol{s}_t, a_t^i)$ and her state will change to s_{t+1}^i according to a transition probability function $P^i(\boldsymbol{s}_t, a_t^i)$. In particular, the probability transition $P^i : \mathcal{S}^N \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ and the distribution of the reward function $R^i : \mathcal{S}^N \times \mathcal{A} \to \mathcal{P}([0, \mathrm{R}_{\max}])$ are both measurable functions with some constant $\mathrm{R}_{\max} > 0$.

A Markovian game further restricts the admissible policy/control for player i to be of the form $a_t^i \sim \pi_t^i(\boldsymbol{s}_t)$ with π_t^i measurable. That is, $\pi_t^i : \mathcal{S}^N \to \mathcal{P}(\mathcal{A})$ maps each state profile $\boldsymbol{s} \in \mathcal{S}^N$ to a randomized action.The accumulated reward (a.k.a. the value function) for player i, given the initial state profile $\boldsymbol{s}$ and the policy profile sequence $\boldsymbol{\pi} := \{\boldsymbol{\pi}_t\}_{t=0}^{\infty}$ with $\boldsymbol{\pi}_t = (\pi_t^1, \dots, \pi_t^N)$, is then defined as

$$V^i(\boldsymbol{s}, \boldsymbol{\pi}) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^i(\boldsymbol{s}_t, a_t^i) \Big| \boldsymbol{s}_0 = \boldsymbol{s}\right], \tag{1.2.1}$$

where $\gamma \in (0, 1)$ is the discount factor, $a_t^i \sim \pi_t^i(\boldsymbol{s}^t)$, and $s_{t+1}^i \sim P^i(\boldsymbol{s}_t, a_t^i)$. The goal of each player is to maximize her value function over all admissible policy sequences such that (1.2.1) is finite.

In general, this type of stochastic N-player game is notoriously hard to analyze, especially when N is large [126]. Mean field game (MFG), pioneered by [87] and [104] in the continuous settings and later developed in [20, 68, 86, 109, 138] for discrete settings, provides an ingenious and tractable aggregation approach to approximate the otherwise challenging N-player stochastic games. The basic idea for an MFG goes as follows. Assume all players are identical, indistinguishable and interchangeable, when $N \to \infty$, one can view the limit of other players' states $\boldsymbol{s}_t^{-i} = (s_t^1, \dots, s_t^{i-1}, s_t^{i+1}, \dots, s_t^N)$ as a population state distribution μ_t with $\mu_t(s) := \lim_{N\to\infty} \frac{\sum_{j=1, j\neq i}^{N} \mathbf{I}_{s_t^j = s}}{N}$.[2] Due to the homogeneity of the players, one can then focus on a single (representative) player. At time t, after the representative player chooses her action a_t according to some policy π_t, she will receive reward $r(s_t, a_t, \mu_t)$ and her state will evolve under a *controlled stochastic dynamics* of a mean-field type $P(\cdot|s_t, a_t, \mu_t)$. Here the policy π_t depends on both the current state s_t and the current population state distribution μ_t such that $\pi_t : \mathcal{S} \times \mathcal{P}(\mathcal{S}) \to \mathcal{P}(\mathcal{A})$. Then, in mean-field limit, one may consider instead the following optimization problem,

$$\begin{aligned} \text{maximize}_{\boldsymbol{\pi}} \quad & V(s, \boldsymbol{\pi}, \boldsymbol{\mu}) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, \mu_t) | s_0 = s\right] \\ \text{subject to} \quad & s_{t+1} \sim P(s_t, a_t, \mu_t), \quad a_t \sim \pi_t(s_t, \mu_t), \end{aligned}$$

where $\boldsymbol{\pi} := \{\pi_t\}_{t=0}^{\infty}$ denotes the policy sequence and $\boldsymbol{\mu} := \{\mu_t\}_{t=0}^{\infty}$ the distribution flow.

1.2.2 General MFG (GMFG)

In the classical MFG setting, the reward and the dynamic for each player are known. They depend only on the state of the player s_t, the action of this particular player a_t, and

[2] Here the indicator function $\mathbf{I}_{s_t^j = s} = 1$ if $s_t^j = s$ and 0 otherwise.

the population state distribution μ_t. In contrast, in the motivating auction example, the reward and the dynamic are unknown; they rely on the actions of *all* players, as well as on s_t and μ_t.

We therefore define the following general MFG (GMFG) framework. At time t, after the representative player chooses her action a_t according to some measurable policy $\pi : \mathcal{S} \times \mathcal{P}(\mathcal{S}) \to \mathcal{P}(\mathcal{A})$, she will receive a (possibly random) reward $r(s_t, a_t, \mathcal{L}_t)$ sampled from distribution $R(s_t, a_t, \mathcal{L}_t)$ and her state will evolve according to $P(\cdot|s_t, a_t, \mathcal{L}_t)$, with $\mathcal{L}_t = \mathbb{P}_{s_t,a_t} \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ the joint distribution of the state and the action, *i.e.*, the population state-action pair. This joint distribution $\mathcal{L}_t$ has marginal distributions α_t for the population action and μ_t for the population state. Note the inclusion of α_t allows the reward and the dynamic to depend on all players' actions. Here $P : \mathcal{S} \times \mathcal{A} \times \mathcal{P}(\mathcal{S} \times \mathcal{A}) \to \mathcal{P}(\mathcal{S})$ and $R : \mathcal{S} \times \mathcal{A} \times \mathcal{P}(\mathcal{S} \times \mathcal{A}) \to \mathcal{P}([0, \mathrm{R}_{\max}])$ are measurable functions with some constant $\mathrm{R}_{\max} > 0$. The objective of the player is to solve the following control problem:

$$\begin{aligned} &\text{maximize}_{\boldsymbol{\pi}} \quad V(s, \boldsymbol{\pi}, \boldsymbol{\mathcal{L}}) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, \mathcal{L}_t) | s_0 = s\right] \\ &\text{subject to} \quad s_{t+1} \sim P(s_t, a_t, \mathcal{L}_t), \quad a_t \sim \pi_t(s_t, \mu_t). \end{aligned} \tag{GMFG}$$

Here the expectation in the objective function is always taken for all randomness in the system. In addition, $\boldsymbol{\mathcal{L}} := \{\mathcal{L}_t\}_{t=0}^{\infty}$ which may be time dependent. That is, an infinite-time horizon MFG may have time-dependent NE solutions due to the mean information process in the MFG. This is fundamentally different from the theory of MDP where the optimal control, if exists uniquely, would be time independent in an infinite time horizon setting.

In this paper, we will analyze the existence of NE to GMFG. For ease of exposition, we will first focus on stationry NEs. Accordingly, for notational brevity, we abbreviate $\boldsymbol{\pi} = \{\pi\}_{t=0}^{\infty}$ and $\boldsymbol{\mathcal{L}} = \{\mathcal{L}\}_{t=0}^{\infty}$ as π and $\mathcal{L}$, respectively. We will show in the end how this stationary constraint can be relaxed (*cf.* Section 1.9).

Definition 1.2.1 (Stationary NE for GMFGs)**.** In (GMFG), a player-population profile ($\pi^\star$, $\mathcal{L}^\star$) is called a stationary NE if

1. (Single player side) For any policy π and any initial state $s \in \mathcal{S}$,

$$V\left(s, \pi^\star, \mathcal{L}^\star\right) \geq V\left(s, \pi, \mathcal{L}^\star\right). \tag{1.2.2}$$

2. (Population side) $\mathbb{P}_{s_t,a_t} = \mathcal{L}^\star$ for all $t \geq 0$, where $\{s_t, a_t\}_{t=0}^{\infty}$ is the dynamics under the policy $\pi^\star$ starting from $s_0 \sim \mu^\star$, with $a_t \sim \pi^\star(s_t, \mu^\star)$, $s_{t+1} \sim P(\cdot|s_t, a_t, \mathcal{L}^\star)$, and $\mu^\star$ being the population state marginal of $\mathcal{L}^\star$.

The single player side condition captures the optimality of $\pi^\star$, when the population side is fixed. The population side condition ensures the "consistency" of the solution: it guarantees that the state and action distribution flow of the single player does match the population state and action sequence $\boldsymbol{\mathcal{L}}^\star := \{\mathcal{L}^\star\}_{t=0}^{\infty}$.

1.2.3 Examples of GMFG

Here we provide three examples under the framework of GMFG.

A toy example. Take a two-state dynamic system with two choices of controls. The state space $\mathcal{S} = \{0, 1\}$, the action space $\mathcal{A} = \{L, R\}$. Here the action L means to move left and R means to move right. The dynamic of the representative agent in the mean-field system $\{s_t\}_{t\geq 1}$ goes as follows: if the agent is in state s_t and she takes action $a_t = L$ at time t, then $s_{t+1} = 0$; if she takes action $a_t = R$, then $s_{t+1} = 1$. At the end of each round, the agent will receive a reward $-W_2(\mu_t, B) - W_2(\beta_t(s_t, \cdot), B)$, which depends on all agents, where W_2 is the ℓ_2-Wasserstein distance. Here $\mu_t(\cdot)$ denotes the state distribution of the mean-field population at time t, $\beta_t(s, \cdot) := \mathcal{L}_t(s, \cdot)/\mu_t(s)$ denotes the action distribution of the population in state s ($s = 0, 1$) at time t (set $\beta_t(s, \cdot) := (0.5, 0.5)$ when $\mu_t(s) = 0$), and B is a given Bernoulli distribution with parameter p ($0 < p < 1$).

As a demonstrating example, here we provide the calculation for one stationary NE solution. Note that $-W_2(\mu, B) \leq 0$ for any distribution μ over $\mathcal{S}$. Similarly, $-W_2(\alpha, B) \leq 0$ for any distribution α over $\mathcal{A}$. Hence for each policy $\boldsymbol{\pi}$, given population distribution flow $\boldsymbol{\mathcal{L}} = \{\mathcal{L}_t\}_{t=1}^{\infty}$,

$$V(0, \boldsymbol{\pi}, \boldsymbol{\mathcal{L}}) \;=\; -\sum_{t=1}^{\infty} \gamma^t \mathbb{E}[W_2(\mu_t, B) + W_2(\beta_t(s_t, \cdot), B) | s_0 = 0] \;\leq\; 0, \tag{1.2.3}$$

and

$$V(1, \boldsymbol{\pi}, \boldsymbol{\mathcal{L}}) \;=\; -\sum_{t=1}^{\infty} \gamma^t \mathbb{E}[W_2(\mu_t, B) + W_2(\beta_t(s_t, \cdot), B) | s_0 = 1] \;\leq\; 0. \tag{1.2.4}$$

It is easy to check that $\mu^\star = (p, 1-p)$ and $\pi^\star(s, \mu^\star) = (p, 1-p)$ ($s = 0, 1$). is a pair of stationary mean-field solution. And $\mathcal{L}^\star$ is defined with $\mathcal{L}^\star(s, a) = \mu^\star(s)\pi^\star(a|s, \mu^\star)$ for any $s \in \mathcal{S}$, $a \in \mathcal{A}$, accordingly, where $\pi(a|s, \mu)$ is defined as the probability of taking action a following the action distribution $\pi(s, \mu)$. In this case, the corresponding optimal value function is defined as

$$V(0, \pi^\star, \mathcal{L}^\star) = V(1, \pi^\star, \mathcal{L}^\star) = 0,$$

which reaches the upper bound in (1.2.3) and (1.2.4).

Repeated auction. Take a representative advertiser in the auction aforementioned in the motivating example in Section 1.1. Denote $s_t \in \{0, 1, 2, \cdots, s_{\max}\}$ as the budget of this player at time t, where $s_{\max} \in \mathbb{N}^+$ is the maximum budget allowed on the Ad exchange with a unit bidding price. Denote $a_t \in \{0, 1, 2, \cdots, a_{\max}\}$ as the bid price submitted by this player, where $a_{\max}$ is the maximum bid set by the bidder, and α_t as the bidding/(action) distribution of the population. At time t, all advertisers are randomly divided into different groups and each group of advertisers competes for one slot to display their ads. Assuming

that there are M advertisers in each group, then the representative advertiser competes with $M-1$ other representative players whose bidding prices are independently sampled from α_t. Let w_t^M denote whether the representative player wins the bid. Then if she takes action a_t, the probability she will win the bid is $\mathbb{P}(w_t^M = 1) = F_{\alpha_t}(a_t)^{M-1}$, where F_{α_t} is the cumulative distribution function of a random variable $X \sim \alpha_t$.

If this advertiser does not win the bid, her reward $r_t = 0$. If she wins, there are several components in her reward: a_t^M, the second best bid in a Vickrey auction, paid by the winning advertiser; v_t, the conversion of clicks of the slot; and ρ, the rate of penalty for overshooting if the payment a_t^M exceeds her budget s_t. Therefore, at each time t, her reward with bid a_t and budget s_t is

$$r_t = \mathbf{I}_{\{w_t^M=1\}} \left[(v_t - a_t^M) - (1+\rho)\mathbf{I}_{\{s_t<a_t^M\}}(a_t^M - s_t) \right], \tag{1.2.5}$$

where the first term is the profit of wining the auction and the second term is the penalty of overshooting. And the budget dynamics s_t follows,

$$s_{t+1} = \begin{cases} s_t, & w_t^M \neq 1, \\ s_t - a_t^M, & w_t^M = 1 \text{ and } a_t^M \leq s_t, \\ 0, & w_t^M = 1 \text{ and } a_t^M > s_t. \end{cases} \tag{1.2.6}$$

That is, if this player does not win the bid, the budget remains the same; if she wins and has sufficient money to pay, her budget will decrease from s_t to $s_t - a_t^M$; however, if she wins but does not have enough money to pay, her budget will be 0 after the payment and there will be a penalty in the reward function.

Notice that both distributions of w_t^M and a_t^M depend on the population distribution $\mathcal{L}_t$ (or more specifically α_t). In fact, the reward function $r(s_t, a_t) = r_t$ and the transition probability $s_{t+1} \sim P(\cdot|s_t, a_t, \mathcal{L}_t)$ specified by (1.2.5) and (1.2.6) are fully characterized by the probabilities $\mathbb{P}(w_t^M = 1, a_t^M \leq \cdot|s_t, a_t, \mathcal{L}_t)$ and $\mathbb{P}(w_t^M = 0)$ (since $r_t = 0$ and $s_{t+1} = s_t$ whenever $w_t^M = 0$), with

$$\mathbb{P}(w_t^M = 1, a_t^M \leq x|s_t, a_t, \mathcal{L}_t) = F_{\alpha_t}(\min\{x, a_t\})^{M-1}, \quad \mathbb{P}(w_t^M = 0) = 1 - F_{\alpha_t}(a_t)^{M-1}.$$

Clearly the above model fits into the framework of (GMFG), with the following transition probability.

$$\mathbb{P}(s'|s, a, \mathcal{L}) = \begin{cases} F_\alpha(a)^{M-1} - F_\alpha(\min\{s, a\})^{M-1}, & s' = 0, \\ 1 - F_\alpha(a)^{M-1}, & s' = s, \\ F_\alpha(\min\{s - s', a\})^{M-1} - F_\alpha(\min\{s - s' - 1, a\})^{M-1}, & 0 < s' < s, \end{cases} \tag{1.2.7}$$

where α is the action marginal of $\mathcal{L}$. The reward model can be explicitly written similarly.

In practice, one may modify the dynamics of s_{t+1} with a non-negative random budget fulfillment $\Delta(s_{t+1})$ after the auction clearing such that $\hat{s}_{t+1} = s_{t+1} + \Delta(s_{t+1})$ [8, 75].

Experiments of this repeated auction problem can be found in the conference version [76] of this paper, and will not be repeated here.

Equilibrium price. Another example, adapted from [74, Section 3] is to consider a large number (continuum) of homogeneous firms producing the same product under perfect competition, and the price of the product is determined endogenously by the supply-demand equilibrium [22]. Each firm, meanwhile, maintains a certain inventory level of the raw materials for production.

Given the homogeneity of the firms, it is sufficient to focus on a representative firm paired with the population distribution. In each period t, the representative firm decides a quantity q_t to consume the raw materials for production and a quantity h_t to replenish the inventory of raw materials. For simplicity, we assume each unit of the raw material is used to produce one unit of the product. Both the new products and ordered raw materials will be available at the end of this given period t. The representative agent makes decision based on her current inventory level of the raw material, denoted as s_t, which evolves according to

$$s_{t+1} = s_t - \min\{q_t, s_t\} + h_t. \tag{1.2.8}$$

Note that if the firm overproduces and exceeds her current inventory capacity (i.e., $q_t > s_t$), then the firm will pay a cost for an emergency order of the raw material. Finally, the reward during this period t is given by

$$r_t = (p_t - c_0)\, q_t - c_1\, q_t^2 - c_2\, h_t - (c_2 + c_3)\, \max\{q_t - s_t, 0\} - c_4\, s_t. \tag{1.2.9}$$

Here p_t is the selling price of the product of all firms; $c_0 > 0$ is the manufacturing cost and labor cost for making one unit of the product; $c_1 > 0$ is the quadratic cost which can be viewed as the transient price impact associated with the production level q_t; $c_2 > 0$ is the cost of regular orders of the raw materials; $c_3 > 0$ is the additional cost for the emergency order of the raw materials; and finally, $c_4 > 0$ is the inventory cost.

The price p_t is determined according to the supply-demand equilibrium on the market at each moment. On one hand, the normalized demand (per producer) on the market $D(p_t)$ follows ([74])

$$D(p_t) := d p_t^{-\sigma}, \tag{1.2.10}$$

where d denotes some benchmark demand level and σ is the elasticity of demand that can be interpreted as the elasticity of substitution between the given product and any other good. On the other hand, the (average) supply in this market is given by the average production of all firms which follows $\mathbb{E}_{q_t \sim \pi_t}[q_t]$ under some policy π_t. If all firms are restricted to stationary policies (denoted as π), then this leads to a stationary equilibrium price q which satisfies the supply-demand equilibrium:

$$\mathbb{E}_{q \sim \pi}[q] = d\, p^{-\sigma}. \tag{1.2.11}$$

To fit into the theoretical framework proposed in Section 1.2, we set $\mathcal{S} = \{0, 1, \cdots, S\}$ and $\mathcal{A} = \{(q, h) \,|\, q \in \{0, 1, \cdots, Q\}$ and $h \in \{0, 1, \cdots, H\}\}$ for some positive integers S, Q and H.

1.3 Solution for GMFGs

We now establish the existence and uniqueness of the stationary NE to (GMFG), by generalizing the classical fixed-point approach for MFGs to this GMFG setting. (See [87] and [104] for the classical case.) It consists of three steps.

Step A. Fix $\mathcal{L}$, (GMFG) becomes the classical single-player optimization problem. Indeed, with $\mathcal{L}$ fixed, the population state distribution μ is also fixed, and hence the space of admissible policies is reduced to the single-player case. Solving (GMFG) is now reduced to finding a policy $\pi^\star_{\mathcal{L}} \in \Pi := \{\pi \,|\, \pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})\}$ to maximize

$$\begin{aligned} V(s, \pi_{\mathcal{L}}, \mathcal{L}) := \quad & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, \mathcal{L}) | s_0 = s\right], \\ \text{subject to} \quad & s_{t+1} \sim P(s_t, a_t, \mathcal{L}), \quad a_t \sim \pi_{\mathcal{L}}(s_t). \end{aligned}$$

Notice that with $\mathcal{L}$ fixed, one can safely suppress the dependency on μ in the admissible policies.

Now given this fixed $\mathcal{L}$ and the solution $\pi^\star_{\mathcal{L}}$ to the above optimization problem, one can define a mapping from the fixed population distribution $\mathcal{L}$ to a chosen optimal randomized policy sequence. That is,

$$\Gamma_1 : \mathcal{P}(\mathcal{S} \times \mathcal{A}) \to \Pi,$$

such that $\pi^\star_{\mathcal{L}} = \Gamma_1(\mathcal{L})$. Note that the optimal policy of an MDP in general may not be unique. To ensure that Γ_1 is a single-valued instead of set-valued mapping, here Γ_1 includes a policy selection component to select a single optimal policy from the set of optimal policies for a given $\mathcal{L}$, which is guaranteed to exist by Zermelo's Axiom of Choice. For example, when the action space is finite, one can utilize the **argmax-e** operator and set the "maximizing" actions with equal probabilities (see Section 1.4.1 for the detailed definition). In addition, for non-degenerate linear-quadratic MFGs [60] and general MFGs where the Bellman mappings are strongly concave in actions [7] and the action space is convex in the Euclidean space, the optimal policy $\pi^\star_{\mathcal{L}}$ for a given $\mathcal{L}$ is unique under appropriate assumptions. Hence no policy selection is needed in such cases.

Note that this $\pi^\star_{\mathcal{L}}$ satisfies the single player side condition in Definition 1.2.1 for the population state-action pair $\mathcal{L}$,

$$V\left(s, \pi^\star_{\mathcal{L}}, \mathcal{L}\right) \geq V\left(s, \pi, \mathcal{L}\right), \tag{1.3.1}$$

for any policy π and any initial state $s \in \mathcal{S}$.

As in the MFG literature [87], a feedback regularity condition is needed for analyzing Step A.

Assumption 1. *There exists a constant $d_1 \geq 0$, such that for any $\mathcal{L}, \mathcal{L}' \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$,*

$$D(\Gamma_1(\mathcal{L}), \Gamma_1(\mathcal{L}')) \leq d_1 W_1(\mathcal{L}, \mathcal{L}'), \tag{1.3.2}$$

where

$$D(\pi, \pi') := \sup_{s \in \mathcal{S}} W_1(\pi(s), \pi'(s)), \tag{1.3.3}$$

and W_1 is the ℓ_1-Wasserstein distance (a.k.a. earth mover distance) between probability measures [67, 132, 157].

Step B. Given $\pi^\star_{\mathcal{L}}$ obtained from Step A, update the initial $\mathcal{L}$ to $\mathcal{L}'$ following the controlled dynamics $P(\cdot|s_t, a_t, \mathcal{L})$.

Accordingly, for any admissible policy $\pi \in \Pi$ and a joint population state-action pair $\mathcal{L} \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$, define a mapping $\Gamma_2 : \Pi \times \mathcal{P}(\mathcal{S} \times \mathcal{A}) \to \mathcal{P}(\mathcal{S} \times \mathcal{A})$ as follows:

$$\Gamma_2(\pi, \mathcal{L}) := \hat{\mathcal{L}} = \mathbb{P}_{s_1, a_1}, \tag{1.3.4}$$

where $a_1 \sim \pi(s_1)$, $s_1 \sim \mu P(\cdot|\cdot, a_0, \mathcal{L})$, $a_0 \sim \pi(s_0)$, $s_0 \sim \mu$, and μ is the population state marginal of $\mathcal{L}$.

One needs a standard assumption in this step.

Assumption 2. *There exist constants d_2, $d_3 \geq 0$, such that for any admissible policies π, π_1, π_2 and joint distributions $\mathcal{L}, \mathcal{L}_1, \mathcal{L}_2$,*

$$W_1(\Gamma_2(\pi_1, \mathcal{L}), \Gamma_2(\pi_2, \mathcal{L})) \leq d_2 D(\pi_1, \pi_2), \tag{1.3.5}$$

$$W_1(\Gamma_2(\pi, \mathcal{L}_1), \Gamma_2(\pi, \mathcal{L}_2)) \leq d_3 W_1(\mathcal{L}_1, \mathcal{L}_2). \tag{1.3.6}$$

Step C. Repeat Step A and Step B until $\mathcal{L}'$ matches $\mathcal{L}$.

This step is to ensure the population side condition. To ensure the convergence of the combined step one and step two, it suffices if $\Gamma : \mathcal{P}(\mathcal{S} \times \mathcal{A}) \to \mathcal{P}(\mathcal{S} \times \mathcal{A})$ with $\Gamma(\mathcal{L}) := \Gamma_2(\Gamma_1(\mathcal{L}), \mathcal{L})$ is a contractive mapping under the W_1 distance. Then by the Banach fixed point theorem and the completeness of the related metric spaces (*cf.* Section 1.10.1), there exists a unique stationary NE of the GMFG. That is,

Theorem 1.3.1 (Existence and Uniqueness of stationary GMFG solution)**.** *Given Assumptions 1 and 2, and assume $d_1 d_2 + d_3 < 1$. Then there exists a unique stationary NE to* (GMFG).

Proof. [Proof of Theorem 1.3.1] First by Definition 1.2.1 and the definitions of Γ_i $(i = 1, 2)$, $(\pi, \mathcal{L})$ is a stationary NE iff $\mathcal{L} = \Gamma(\mathcal{L}) = \Gamma_2(\Gamma_1(\mathcal{L}), \mathcal{L})$ and $\pi = \Gamma_1(\mathcal{L})$, where $\Gamma(\mathcal{L}) = \Gamma_2(\Gamma_1(\mathcal{L}), \mathcal{L})$. This indicates that for any $\mathcal{L}_1, \mathcal{L}_2 \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$,

$$\begin{aligned} &W_1(\Gamma(\mathcal{L}_1), \Gamma(\mathcal{L}_2)) = W_1(\Gamma_2(\Gamma_1(\mathcal{L}_1), \mathcal{L}_1), \Gamma_2(\Gamma_1(\mathcal{L}_2), \mathcal{L}_2)) \\ &\leq W_1(\Gamma_2(\Gamma_1(\mathcal{L}_1), \mathcal{L}_1), \Gamma_2(\Gamma_1(\mathcal{L}_2), \mathcal{L}_1)) + W_1(\Gamma_2(\Gamma_1(\mathcal{L}_2), \mathcal{L}_1), \Gamma_2(\Gamma_1(\mathcal{L}_2), \mathcal{L}_2)) \\ &\leq (d_1 d_2 + d_3) W_1(\mathcal{L}_1, \mathcal{L}_2). \end{aligned} \tag{1.3.7}$$

And since $d_1 d_2 + d_3 \in [0, 1)$, by the Banach fixed-point theorem, we conclude that there exists a unique fixed-point of Γ, or equivalently, a unique stationary MFG solution to (GMFG). □

Remark 1.3.1 (Existence and Uniqueness of the GMFG solution)**.** *(1) In general, there may multiple optimal policies in Step A under a fixed mean-field information $\mathcal{L}$. In this case, the candidate fixed point(s) are the fixed point(s) of a set-valued map as described in [100]. To simplify the analysis, we specify a rule in Step A to select one optimal policy to ensure that Γ is an injection.*

(2) In the MFG literature, the uniqueness of the MFG solution can be verified under the small parameter condition [29] or the monotonicity condition [104]. Our condition of $d_1 d_2 + d_3 < 1$ extends the small parameter condition in [29] for strict controls to relaxed controls.

(3) Finally, Theorem 1.3.1 can be extended to a non-stationary setting, as will be shown in Section 1.9.

Remark 1.3.2. *Assumptions 1 and 2 can be more explicit in specific problem settings.*

For instance, when the action space is the Euclidean space or its convex subset, explicit conditions on P and r have been described for the linear-quadratic MFG (LQ-MFG) [60] and later generalized in [7].

When the action space is finite, the following lemma explicitly characterizes Assumption 2.

Lemma 1.3.2. *Suppose that $\max_{s,a,\mathcal{L},s'} P(s'|s,a,\mathcal{L}) \leq c_1$, and that $P(s'|s,a,\cdot)$ is c_2-Lipschitz in W_1, i.e.,*

$$|P(s'|s,a,\mathcal{L}_1) - P(s'|s,a,\mathcal{L}_2)| \leq c_2 W_1(\mathcal{L}_1, \mathcal{L}_2). \tag{1.3.8}$$

Then in Assumption 2, d_2 and d_3 can be chosen as

$$d_2 = \frac{2 diam(\mathcal{S}) diam(\mathcal{A}) |\mathcal{S}| c_1}{d_{\min}(\mathcal{A})} \tag{1.3.9}$$

and $d_3 = \frac{diam(\mathcal{S}) diam(\mathcal{A}) c_2}{2}$, respectively. Here $d_{\min}(\mathcal{A}) = \min_{a \neq a' \in \mathcal{A}} \|a - a'\|_2$, which is guaranteed to be positive when $\mathcal{A}$ is finite.

When entropy regularization is introduced into the system (see e.g., [7, 164]), Assumption 1 can be reduced to boundedness and Lipschitz continuity conditions on P and r as in Lemma 1.3.2. Moreover, Theorem 1.3.1 and all subsequent theoretical results hold whenever the composed mapping Γ is contractive (in W_1), independent of Assumptions 1 or 2. In Section 1.8.2, we numerically verify that the Γ mapping is contractive for various choices of the model parameters in our tested problems.

1.4 Naive algorithm and stabilization techniques

In this section, we design algorithms for the GMFG. Since the reward and transition distributions are unknown, this is simultaneously learning the system and finding the NE of

the game. We will focus on the case with finite state and action spaces, *i.e.*, $|\mathcal{S}|, |\mathcal{A}| < \infty$. We will look for stationary (time independent) NEs. This stationarity property enables developing appropriate stationary reinforcement learning algorithms, suitable for an infinite time horizon game. Instead of knowing the transition probability P and the reward r explicitly, the algorithms we propose only assume access to a simulator oracle, which is described below. This is not restrictive in practice. For instance, in the ad auction example, one may adopt the bid recommendation perspective of the publisher, say Google, Facebook or Amazon, who acts as the auctioneer and owns the Ad slot inventory on its own Ad exchange platform. In this case, a high quality auction simulator is typically built and maintained by a team of the publisher. See also [144] for more examples.

Simulator oracle. For any policy $\pi \in \Pi$, given the current state $s \in \mathcal{S}$, for any population distribution $\mathcal{L}$, one can obtain a *sample* of the next state $s' \sim P(\cdot|s, \pi(s), \mathcal{L})$, a reward $r = r(s, \pi(s), \mathcal{L})$, and the next population distribution $\mathcal{L}' = \mathbb{P}_{s',\pi(s')}$. For brevity, we denote the simulator as $(s', r, \mathcal{L}') = \mathcal{G}(s, \pi, \mathcal{L})$. This simulator oracle can be weakened to fit the N-player setting, see Section 1.6.

In the following, we begin with a naive algorithm that simply combines the three-step fixed point approach with general RL algorithms, and demonstrate that this algorithm can be unstable (Section 1.4.1). We then propose some smoothing and projection techniques to resolve the issue (Section 1.4.2). In Section 1.5.1 and Section 1.5.2, we design general value-based and policy-based RL algorithms, and establish the corresponding convergence and complexity results. These two algorithms include most of the RL algorithms in the literature. We then illustrate by two concrete examples based on Q-learning and trust-region policy optimization algorithms.

1.4.1 Naive algorithm and its issue

We follow the three-step fixed-point approach described in Section 1.3. Notice the fact that with $\mathcal{L}$ fixed, Step A in Section 1.3 becomes a standard learning problem for an infinite horizon discounted MDP. More specifically, the MDP to be solved is $\mathcal{M}_{\mathcal{L}} = (\mathcal{S}, \mathcal{A}, P_{\mathcal{L}}, r_{\mathcal{L}}, \gamma)$, where $P_{\mathcal{L}}(s'|s, a) = P(s'|s, a, \mathcal{L})$ and $r_{\mathcal{L}}(s, a) = r(s, a, \mathcal{L})$. In general, for an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, for any policy π one can define its value functions $V^{\pi}_{\mathcal{M}}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s\right]$ and its Q-functions $Q^{\pi}_{\mathcal{M}}(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s, a_0 = a\right]$, where s_t, a_t is the trajectory under policy π. One can also define the optimal Q-function as the unique solution of the Bellman equation:

$$Q^{\star}_{\mathcal{M}}(s, a) = \mathbb{E}[r(s, a)] + \gamma \max_{a'} \sum_{s' \in \mathcal{S}} P(s'|s, a) Q^{\star}_{\mathcal{M}}(s', a')$$

for all s, a and its optimal value function $V^{\star}_{\mathcal{M}}(s) = \max_a Q^{\star}_{\mathcal{M}}(s, a)$ for all s. We also use the shorthand $V^{\star}_{\mathcal{L}} = V^{\star}_{\mathcal{M}_{\mathcal{L}}}$ and $Q^{\star}_{\mathcal{L}} = Q^{\star}_{\mathcal{M}_{\mathcal{L}}}$ for notational brevity. Whenever the context is clear, we may omit $\mathcal{M}$, $\mathcal{L}$ and $\mathcal{M}_{\mathcal{L}}$ for notational convenience.

Given the optimal Q-function $Q^\star_{\mathcal{L}}$, one can obtain an optimal policy $\pi^\star_{\mathcal{L}}$ with $\pi^\star_{\mathcal{L}}(s) = \textbf{argmax-e}(Q^\star_{\mathcal{L}}(s,\cdot))$. Here the **argmax-e** operator is defined so that actions with equal maximum Q-values would have equal probabilities to be selected. Hereafter, we specify Γ_1 as a mapping to the aforementioned choice of the optimal policy, *i.e.*, the s-component $\Gamma_1(\mathcal{L})_s = \textbf{argmax-e}(Q^\star_{\mathcal{L}}(s,\cdot))$ for any $s \in \mathcal{S}$.

The population update in Step B can then be directly obtained from the simulator $\mathcal{G}$ following policy $\pi^\star_{\mathcal{L}}$. Combining these two steps leads to the following naive algorithm (Algorithm 1).

Algorithm 1 Naive Reinforcement Learning for GMFGs

1: **Input**: Initial population state-action pair $\mathcal{L}_0$
2: **for** $k = 0, 1, \cdots$ **do**
3: Obtain the optimal Q-function $Q_k(s,a) = Q^\star_{\mathcal{L}_k}(s,a)$ of an MDP with dynamics $P_{\mathcal{L}_k}(s'|s,a)$ and reward distributions $R_{\mathcal{L}_k}(s,a)$.
4: Compute $\pi_k \in \Pi$ with $\pi_k(s) = \textbf{argmax-e}\,(Q_k(s,\cdot))$.
5: Sample $s \sim \mu_k$, where μ_k is the population state marginal of $\mathcal{L}_k$, and obtain $\mathcal{L}_{k+1}$ from $\mathcal{G}(s, \pi_k, \mathcal{L}_k)$.
6: **end for**

Unfortunately, in practice, one cannot obtain the exact optimal Q-function Q_k. In fact, invoking any commonly used RL algorithm with the simulator $\mathcal{G}$ leads to an approximation $\hat{Q}_k$ of the actual Q_k. This approximation error is then magnified by the discontinuous and sensitive **argmax-e**, which eventually leads to an unstable algorithm (see Figure 1.4 for an example of divergence). To see why **argmax-e** is not continuous, consider the following simple example. Let $x = (1,1)$, then $\textbf{argmax-e}(x) = (1/2, 1/2)$. For any $\epsilon > 0$, let $y_\epsilon = (1, 1-\epsilon)$, then $\textbf{argmax-e}(y_\epsilon) = (1,0)$. Hence $\lim_{\epsilon\to 0} y_\epsilon = x$ but

$$\lim_{\epsilon\to 0} \textbf{argmax-e}(y_\epsilon) \neq \textbf{argmax-e}(x).$$

This instability issue will be addressed by introducing smoothing and projection techniques.

1.4.2 Restoring stability

Smoothing techniques. To address the instability caused, we replace **argmax-e** with a smooth function that is a good approximation to **argmax-e** while being Lipschitz continuous. One such candidate is the softmax operator $\textbf{softmax}_c : \mathbb{R}^n \to \mathbb{R}^n$, with

$$\textbf{softmax}_c(x)_i = \frac{\exp(cx_i)}{\sum_{j=1}^n \exp(cx_j)}, \quad i = 1, \ldots, n,$$

for some positive constant c. The resulting policies are sometimes called Boltzmann policies, and are widely used in the literature of reinforcement learning [11, 78].

The softmax operator can be generalized to a wide class of operators. In fact, for positive constants $c,\ c' > 0$, one can consider a parametrized family $\mathcal{F}_{c,c'} \subseteq \{f_{c,c'} : \mathbb{R}^n \to \mathbb{R}^n\}$ of all "smoothed" **argmax-e**'s, *i.e.*, all $f_{c,c'} : \mathbb{R}^n \to \mathbb{R}^n$ that satisfies the following two conditions:

- Condition 1: $f_{c,c'}$ is c-Lipschitz, *i.e.*, $\|f_{c,c'}(x) - f_{c,c'}(y)\|_2 \leq c\|x - y\|_2$.
- Condition 2: $f_{c,c'}$ is a good approximation of **argmax-e**, *i.e.*,

$$\|f_{c,c'}(x) - \textbf{argmax-e}(x)\|_2 \leq 2n \exp(-c'\delta),$$

  where $\delta = x_{\max} - \max_{x_j < x_{\max}} x_j$, $x_{\max} = \max_{i=1,\ldots,n} x_i$, and $\delta := \infty$ when all x_j are equal.

Notice that $\mathcal{F}_{c,c'}$ is closed under convex combinations, *i.e.*, if $f_{c,c'},\ g_{c,c'} \in \mathcal{F}_{c,c'}$, then for any $\lambda \in [0,1]$, $\lambda f_{c,c'} + (1-\lambda) g_{c,c}$ also satisfies the two conditions. Hence $\mathcal{F}_{c,c'}$ is convex.

To have a better idea of what $\mathcal{F}_{c,c'}$ looks like, we describe a subset $\mathcal{B}_{c,c'}$ of $\mathcal{F}_{c,c'}$ consisting of the generalized softmax operator $\textbf{softmax}_h : \mathbb{R}^n \to \mathbb{R}^n$, defined as

$$\textbf{softmax}_h(x)_i = \frac{\exp(h(x_i))}{\sum_{j=1}^n \exp(h(x_j))}, \quad i = 1, \ldots, n, \tag{1.4.1}$$

where $h : \mathbb{R} \to \mathbb{R}$ satisfies $c'(x-y) \leq h(x) - h(y) \leq c(x-y)$ for any $x \geq y$. When h is continuously differentiable, a sufficient condition is that $c' \leq h'(x) \leq c$. In particular, if $h(x) \equiv cx$ for some constant $c > 0$, the operator reduces to the classical softmax operator, in which case we overload the notation to write $\textbf{softmax}_h$ as $\textbf{softmax}_c$.

This operator is Lipschitz continuous and close to the **argmax-e** (see Lemmas 1.7.2 and 1.7.3 in Section 3.6), and in particular one can show that $\mathcal{B}_{c,c'} \subseteq \mathcal{F}_{c,c'}$. As a result, even though smoothed (*e.g.*, Boltzmann) policies are not optimal, the difference between the smoothed and the optimal one can always be controlled by choosing a function h with appropriate parameters c, c'. Note that other smoothing operators (*e.g.*, Mellowmax [11], which is a softmax operator with time-varying and problem dependent temperatures) may also be considered.

Error control in updating $\mathcal{L}$. Given the sub-optimality of the smoothed policy, one needs to characterize the difference between the optimal policy and the non-optimal ones. In particular, one can define the action gap between the best and the second best actions in terms of the Q-value as

$$\delta^s(\mathcal{L}) := \max_{a' \in \mathcal{A}} Q^\star_{\mathcal{L}}(s, a') - \max_{a \notin \text{argmax}_{a \in \mathcal{A}} Q^\star_{\mathcal{L}}(s,a)} Q^\star_{\mathcal{L}}(s, a) > 0.$$

Action gap is important for approximation algorithms [19], and is closely related to the problem-dependent bounds for regret analysis in reinforcement learning and multi-armed bandits, and advantage learning algorithms including A3C [116].

The problem is: in order for the learning algorithm to converge in terms of $\mathcal{L}$ (Theorems 1.5.1 and 1.5.5), one needs to ensure a definite differentiation between the optimal policy and the sub-optimal ones. This is problematic as the infimum of $\delta^s(\mathcal{L})$ over an infinite number of $\mathcal{L}$ can be 0. To address this, the population distribution at step k, say $\mathcal{L}_k$, needs to be projected to a finite grid, called ϵ-net. The relation between the ϵ-net and action gaps is as follows:

For any $\epsilon > 0$, there exist a positive function $\phi(\epsilon)$ and an ϵ-net $S_\epsilon := \{\mathcal{L}^{(1)}, \dots, \mathcal{L}^{(N_\epsilon)}\} \subseteq \mathcal{P}(\mathcal{S} \times \mathcal{A})$, with the properties that $\min_{i=1,\dots,N_\epsilon} d_{TV}(\mathcal{L}, \mathcal{L}^{(i)}) \leq \epsilon$ for any $\mathcal{L} \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$, and that $\max_{a' \in \mathcal{A}} Q^\star_{\mathcal{L}^{(i)}}(s, a') - Q^\star_{\mathcal{L}^{(i)}}(s, a) \geq \phi(\epsilon)$ for any $i = 1, \dots, N_\epsilon$, $s \in \mathcal{S}$, and any $a \notin argmax_{a \in \mathcal{A}} Q^\star_{\mathcal{L}^{(i)}}(s, a)$.

Here the existence of ϵ-nets is trivial due to the compactness of the probability simplex $\mathcal{P}(\mathcal{S} \times \mathcal{A})$, and the existence of $\phi(\epsilon)$ comes from the finiteness of the action set $\mathcal{A}$.

In practice, $\phi(\epsilon)$ often takes the form of $D\epsilon^\alpha$ with $D > 0$ and the exponent $\alpha > 0$ characterizing the decay rate of the action gaps. In general, experiments are robust with respect to the choice of ϵ-net.

In the next section, we propose value based and policy based algorithms for learning GMFG.

1.5 RL Algorithms for (stationary) GMFGs

1.5.1 Value-based algorithms

We start by introducing the following definition.

Definition 1.5.1 (Value-based Guarantee)**.** For an arbitrary MDP $\mathcal{M}$, we say that an algorithm has a *value-based guarantee* with parameters $\{C^{(i)}_{\mathcal{M}}, \alpha^{(i)}_1, \alpha^{(i)}_2, \alpha^{(i)}_3, \alpha^{(i)}_4\}_{i=1}^m$, if for any $\epsilon,\ \delta > 0$, after obtaining

$$T_{\mathcal{M}}(\epsilon, \delta) = \sum_{i=1}^{m} C^{(i)}_{\mathcal{M}} \left(\frac{1}{\epsilon}\right)^{\alpha^{(i)}_1} \left(\log \frac{1}{\epsilon}\right)^{\alpha^{(i)}_2} \left(\frac{1}{\delta}\right)^{\alpha^{(i)}_3} \left(\log \frac{1}{\delta}\right)^{\alpha^{(i)}_4} \tag{1.5.1}$$

samples from the simulator oracle $\mathcal{G}$, with probability at least $1 - 2\delta$, it outputs an approximate Q-function $\hat{Q}^{T_{\mathcal{M}}(\epsilon,\delta)}$ which satisfies $\|\hat{Q}^{T_{\mathcal{M}}(\epsilon,\delta)} - Q^\star\|_\infty \leq \epsilon$. Here the norm $\|\cdot\|_\infty$ is understood element-wisely.

1.5.1.1 GMF-V

We now state the first main algorithm (Algorithm 2). It applies to any algorithm *Alg* with a value-based guarantee.

Here $\mathbf{Proj}_{S_\epsilon}(\mathcal{L}) = \text{argmin}_{\mathcal{L}^{(1)},\dots,\mathcal{L}^{(N_\epsilon)}} d_{TV}(\mathcal{L}^{(i)}, \mathcal{L})$. For computational tractability, it is sufficient to choose S_ϵ as a truncation grid so that projection of $\tilde{\mathcal{L}}_k$ onto the ϵ-net reduces to

Algorithm 2 GMF-V(*Alg*, $f_{c,c'}$)

1: **Input**: Initial $\mathcal{L}_0$, ϵ-net S_ϵ, temperatures $c,\ c' > 0$, tolerances $\epsilon_k,\ \delta_k > 0$, $k = 0, 1, \ldots$.
2: **for** $k = 0, 1, \cdots$ **do**
3: Apply *Alg* to find the approximate Q-function $\hat{Q}_k^\star = \hat{Q}^{T_k}$ of the MDP $\mathcal{M}_{\mathcal{L}_k}$, where $T_k = T_{\mathcal{M}_{\mathcal{L}_k}}(\epsilon_k, \delta_k)$.
4: Compute $\pi_k(s) = f_{c,c'}(\hat{Q}_k^\star(s, \cdot))$.
5: Sample $s \sim \mu_k$ (μ_k is the population state marginal of $\mathcal{L}_k$), obtain $\tilde{\mathcal{L}}_{k+1}$ from $\mathcal{G}(s, \pi_k, \mathcal{L}_k)$.
6: Find $\mathcal{L}_{k+1} = \mathbf{Proj}_{S_\epsilon}(\tilde{\mathcal{L}}_{k+1})$
7: **end for**

truncating $\tilde{\mathcal{L}}_k$ to a certain number of digits. For instance, in experiments in Section 1.8, the number of digits is chosen to be 4. Appropriate choices of the hyper-parameters $c,\ c',\ \epsilon$ and tolerances $\epsilon_k,\ \delta_k$ ($k \geq 0$) are given in Theorems 1.5.1. Our experiment shows the algorithm is robust with respect to these hyper-parameters.

We next establish the convergence of the above GMF-V algorithm to an approximate Nash equilibrium of (GMFG), with complexity analysis.

Theorem 1.5.1 (Convergence and complexity of GMF-V)**.** *Assume the same assumptions as Theorem 1.3.1, and $f_{c,c'} \subseteq \mathcal{F}_{c,c'}$. Suppose that Alg has a value-based guarantee with parameters*

$$\{C_{\mathcal{M}}^{(i)}, \alpha_1^{(i)}, \alpha_2^{(i)}, \alpha_3^{(i)}, \alpha_4^{(i)}\}_{i=1}^m.$$

For any $\epsilon,\ \delta > 0$, set $\delta_k = \delta/K_{\epsilon,\eta}$, $\epsilon_k = (k+1)^{-(1+\eta)}$ for some $\eta \in (0, 1]$ ($k = 0, \ldots, K_{\epsilon,\eta} - 1$), and $c \geq c' \geq \frac{\log(1/\epsilon)}{\phi(\epsilon)}$. Then with probability at least $1 - 2\delta$,

$$W_1(\mathcal{L}_{K_{\epsilon,\eta}}, \mathcal{L}^\star) \leq C_0 \epsilon.$$

Here $K_{\epsilon,\eta} := \left\lceil 2 \max\left\{(\eta\epsilon/c)^{-1/\eta}, \log_d(\epsilon/\max\{diam(\mathcal{S})diam(\mathcal{A}), c\}) + 1\right\}\right\rceil$ is the number of outer iterations, and the constant C_0 is independent of δ, ϵ and η.
Moreover, the total number of samples $T = \sum_{k=0}^{K_{\epsilon,\eta}-1} T_{\mathcal{M}_{\mathcal{L}_k}}(\delta_k, \epsilon_k)$ is bounded by

$$T \leq \sum_{i=1}^m \frac{2^{\alpha_2^{(i)}}}{2\alpha_1^{(i)} + 1} C_{\mathcal{M}}^{(i)} K_{\epsilon,\eta}^{2\alpha_1^{(i)}+1} (K_{\epsilon,\eta}/\delta)^{\alpha_3^{(i)}} \left(\log(K_{\epsilon,\eta}/\delta)\right)^{\alpha_2^{(i)}+\alpha_4^{(i)}}. \tag{1.5.2}$$

The proof of Theorem 1.5.1 (in Section 1.7.4) depends on the Lipschitz continuity of the smoothing operator $f_{c,c'}$, the closeness between $f_{c,c'}$ and the **argmax-e** (Lemma 1.7.3 in Section 1.7.3), and the complexity of *Alg* provided by the value-based guarantee.

1.5.1.2 GMF-V-Q: GMF-V with Q-learning

As an example of the GMF-V algorithm, we describe algorithm GMF-V-Q, a Q-learning based GMF-V algorithm. For an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, the synchronous Q-learning

algorithm approximates the value iteration by stochastic approximation. At each step l, with state s and action a, the system reaches state s' according to the controlled dynamics, and the Q-function approximation Q_l is updated by

$$\hat{Q}^{l+1}(s,a) = (1-\beta_l)\hat{Q}^l(s,a) + \beta_l \left[r(s,a) + \gamma \max_{\bar{a}} \hat{Q}^l(s',\bar{a}) \right], \quad \forall s \in \mathcal{S},\, a \in \mathcal{A}, \tag{1.5.3}$$

where $\hat{Q}^0(s,a) = C$ for some constant $C \in \mathbb{R}$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, and the step size β_l can be chosen as ([53])

$$\beta_l = |l+1|^{-h}, \tag{1.5.4}$$

with $h \in (1/2, 1)$.

The corresponding synchronous Q-learning based algorithm with the standard **softmax** operator is GMF-V-Q (Algorithm 3), and will be used in the experiment (Section 1.8).

Algorithm 3 Q-learning for GMFGs (GMF-V-Q)

1: **Input**: Initial $\mathcal{L}_0$, ϵ-net S_ϵ, tolerances $\epsilon_k,\ \delta_k > 0$, $k = 0, 1, \ldots$.
2: **for** $k = 0, 1, \cdots$ **do**
3: Perform synchronous Q-learning with stepsizes (1.5.4) for $T_k = T_{\mathcal{M}_{\mathcal{L}_k}}(\epsilon_k, \delta_k)$ iterations to find the approximate Q-function $\hat{Q}^\star_k = \hat{Q}^{T_k}$ of the MDP $\mathcal{M}_{\mathcal{L}_k}$.
4: Compute $\pi_k \in \Pi$ with $\pi_k(s) = \mathbf{softmax}_c(\hat{Q}^\star_k(s,\cdot))$.
5: Sample $s \sim \mu_k$ (μ_k is the population state marginal of $\mathcal{L}_k$), obtain $\tilde{\mathcal{L}}_{k+1}$ from $\mathcal{G}(s, \pi_k, \mathcal{L}_k)$.
6: Find $\mathcal{L}_{k+1} = \mathbf{Proj}_{S_\epsilon}(\tilde{\mathcal{L}}_{k+1})$
7: **end for**

Let us first recall the following sample complexity result for synchronous Q-learning method.

Lemma 1.5.2 ([53]: sample complexity of synchronous Q-learning)**.** *For an MDP, say $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, suppose that the Q-learning algorithm takes step-sizes (1.5.4). Then $\|\hat{Q}^{T_{\mathcal{M}}(\delta,\epsilon)} - Q^\star_{\mathcal{M}}\|_\infty \le \epsilon$ with probability at least $1 - 2\delta$. Here $\hat{Q}^T$ is the T-th update in the Q-learning updates (1.5.3), $Q^\star_{\mathcal{M}}$ is the (optimal) Q-function, and*

$$T_{\mathcal{M}}(\epsilon,\delta) = \Omega\left(\left(\frac{V_{\max}^2 \log\left(\frac{|\mathcal{S}||\mathcal{A}|V_{\max}}{\delta\beta\epsilon} \right)}{\beta^2\epsilon^2} \right)^{\frac{1}{h}} + \left(\frac{1}{\beta} \log \frac{V_{\max}}{\epsilon} \right)^{\frac{1}{1-h}} \right),$$

where $\beta = (1-\gamma)/2$, $V_{\max} = R_{\max}/(1-\gamma)$, and $R_{\max}$ is such that a.s. $0 \le r(s,a) \le R_{\max}$.

This lemma implies immediately the value-based guarantee (as in Definition 1.5.1) and the convergence for GMF-V-Q. Similar results can be established for asynchronous Q-learning method, as shown in Section 1.10.2.

Corollary 1.5.3. *The synchronous Q-learning algorithm with appropriate choices of step-sizes (cf. (1.5.4)) satisfies the value-based guarantee with parameters* $\{\tilde{C}_{\mathcal{M}}^{(i)}, \alpha_1^{(i)}, \alpha_2^{(i)}, \alpha_3^{(i)}, \alpha_4^{(i)}\}_{i=1}^3$, *where* $C_{\mathcal{M}}^{(i)}(i = 1, 2, 3)$ *are constants depending on* $|\mathcal{S}|, |\mathcal{A}|, V_{\max}, \beta$ *and* h, *and*

$$\begin{aligned}
&\alpha_1^{(1)} = 2/h,\ \alpha_2^{(1)} = 1/h,\ \alpha_3^{(1)} = \alpha_4^{(1)} = 0;\\
&\alpha_1^{(2)} = 2/h,\ \alpha_2^{(2)} = \alpha_3^{(2)} = 0 \text{ and } \alpha_4^{(2)} = 1/h;\\
&\alpha_1^{(3)} = 0,\ \alpha_2^{(3)} = 1/(1-h),\ \alpha_3^{(3)} = 0 \text{ and } \alpha_4^{(3)} = 0.
\end{aligned}$$

In addition, assume the same assumptions as Theorem 1.3.1, then for Algorithm 3 with synchronous Q-learning method, with probability at least $1-2\delta$, $W_1(\mathcal{L}_{K_{\epsilon,\eta}}, \mathcal{L}^\star) \le C_0\epsilon$, *where* $K_{\epsilon,\eta}$ *is defined as in Theorem 1.5.1. And the total number of samples* $T = \sum_{k=0}^{K_{\epsilon,\eta}-1} T_{\mathcal{M}_{\mathcal{L}_k}}(\epsilon_k, \delta_k)$ *is bounded by*

$$T \le O\left(K_{\epsilon,\eta}^{\frac{4}{h}+1}\left(\log\frac{K_{\epsilon,\eta}}{\delta}\right)^{\frac{1}{h}} + \left(\log\frac{K_{\epsilon,\eta}}{\delta}\right)^{\frac{1}{1-h}}\right).$$

1.5.2 Policy-based algorithms

In addition to algorithms with value-based guarantees (*cf.* Definition 1.5.1), there are also numerous algorithms with *policy-based guarantees*.

Definition 1.5.2 (Policy-based Guarantee)**.** For an arbitrary MDP $\mathcal{M}$, we say that an algorithm has a *policy-based guarantee* with parameters $\{C_{\mathcal{M}}^{(i)}, \alpha_1^{(i)}, \alpha_2^{(i)}, \alpha_3^{(i)}, \alpha_4^{(i)}\}_{i=1}^m$, if for any $\epsilon,\ \delta > 0$, after obtaining

$$T_{\mathcal{M}}(\epsilon, \delta) = \sum_{i=1}^{m} C_{\mathcal{M}}^{(i)}\left(\frac{1}{\epsilon}\right)^{\alpha_1^{(i)}}\left(\log\frac{1}{\epsilon}\right)^{\alpha_2^{(i)}}\left(\frac{1}{\delta}\right)^{\alpha_3^{(i)}}\left(\log\frac{1}{\delta}\right)^{\alpha_4^{(i)}} \tag{1.5.5}$$

samples from the simulator oracle $\mathcal{G}$, with probability at least $1-2\delta$, it outputs an approximate policy $\pi_{T_{\mathcal{M}}(\epsilon,\delta)}$, which satisfies $V_{\mathcal{M}}^\star(s) - V_{\mathcal{M}}^{\pi_{T_{\mathcal{M}}(\epsilon,\delta)}}(s) \le \epsilon,\ \forall\, s \in \mathcal{S}$.

1.5.2.1 GMF-P

Before we present policy-based RL algorithms, let us first establish a connection between policy-based and value-based guarantees.

To start, take any policy $\pi \in \Pi$, consider the following synchronous temporal difference (TD) iterations:

$$\tilde{Q}_\pi^{l+1}(s,a) = (1-\beta_l)\tilde{Q}_\pi^l(s,a) + \beta_l\left[r(s,a) + \gamma\tilde{Q}_\pi^l(s',a')\right], \quad \forall s \in \mathcal{S},\ a \in \mathcal{A}, \tag{1.5.6}$$

where $a' \sim \pi(s')$, $\tilde{Q}_\pi^0(s,a) = C$ for some constant $C \in \mathbb{R}$ and any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, and the step size $\beta_l = (l+1)^{-h}$ for some $h \in (1/2, 1)$.

Then we have

Lemma 1.5.4. *Suppose that the algorithm Alg satisfies a policy-based guarantee with parameters* $\{C_{\mathcal{M}}^{(i)}, \alpha_1^{(i)}, \alpha_2^{(i)}, \alpha_3^{(i)}, \alpha_4^{(i)}\}_{i=1}^m$. *Let* $\tilde{Q}_\pi^l$ *be defined by (1.5.6). Then for any* $\delta \in (0,1)$ *and* $\epsilon > 0$, *with probability at least* $1-2\delta$, $\left\|\tilde{Q}^l_{\pi_{T_{\mathcal{M}}(\epsilon,\delta/2)}} - Q^\star_{\mathcal{M}}\right\|_\infty \le \epsilon$ *if*

$$l = \Omega\left(\left(\frac{V_{\max}\log\left(\frac{|\mathcal{S}||\mathcal{A}|V_{\max}}{\delta\beta^2\epsilon}\right)}{\beta^4\epsilon^2}\right)^{1/h} + \left(\frac{1}{\beta}\log\frac{V_{\max}}{\beta\epsilon}\right)^{1/(1-h)}\right), \tag{1.5.7}$$

where $V_{\max} = R_{\max}/(1-\gamma)$ *and* $\beta = (1-\gamma)/2$.

Consequently, the algorithm Alg (combined with TD updates (1.5.6)) also has a value-based guarantee with parameters $\{\tilde{C}_{\mathcal{M}}^{(i)}, \alpha_1^{(i)}, \alpha_2^{(i)}, \alpha_3^{(i)}, \alpha_4^{(i)}\}_{i=1}^{m+3}$, *where* $\tilde{C}_{\mathcal{M}}^{(i)}$ *is some constant multiple of* $C_{\mathcal{M}}^{(i)}$ *(*$i = 1,\dots,m$*),* $\tilde{C}_{\mathcal{M}}^{(m+i)}$ *(*$i = 1,2,3$*) are constants depending on* $V_{\max}$, $|\mathcal{S}|$, $|\mathcal{A}|$, β *and* h, *and we have*

$$\begin{aligned}
&\alpha_1^{(m+1)} = 2/h,\ \alpha_2^{(m+1)} = 1/h,\ \alpha_3^{(m+1)} = \alpha_4^{(m+1)} = 0;\\
&\alpha_1^{(m+2)} = 2/h,\ \alpha_2^{(m+2)} = \alpha_3^{(m+2)} = 0 \text{ and } \alpha_4^{(m+2)} = 1/h;\\
&\alpha_1^{(m+3)} = 0,\ \alpha_2^{(m+3)} = 1/(1-h),\ \alpha_3^{(m+3)} = 0 \text{ and } \alpha_4^{(m+3)} = 0.
\end{aligned} \tag{1.5.8}$$

The above lemma indicates that any algorithm with a policy-based guarantee also satisfies a value-based guarantee with similar parameters (when combined with the TD updates). The policy-based algorithm GMF-P (Algorithm 4) makes use of Lemma 1.5.4 to select the hyper-parameter l so that the resulting $\tilde{Q}^l_{\pi_{T_{\mathcal{M}}(\epsilon,\delta/2)}}$ forms a good value-based certificate.

Algorithm 4 GMF-P(*Alg*, $f_{c,c'}$)

1: **Input**: Initial $\mathcal{L}_0$, ϵ-net S_ϵ, temperatures c, $c' > 0$, tolerances ϵ_k, $\delta_k > 0$, $k = 0, 1, \dots$.
2: **for** $k = 0, 1, \cdots$ **do**
3: Apply *Alg* to find the approximate policy $\hat{\pi}_k = \pi_{T_k}$ of the MDP $\mathcal{M}_k := \mathcal{M}_{\mathcal{L}_k}$, where $T_k = T_{\mathcal{M}_k}(\epsilon_k, \delta_k/2)$.
4: Compute $\tilde{Q}^{l_k}_{\hat{\pi}_k}$ using TD updates (1.5.6) for MDP $\mathcal{M}_k$, with l_k satisfying (1.5.7) (with ϵ and δ replaced by ϵ_k and $\delta_k/2$, respectively).
5: Compute $\pi_k(s) = f_{c,c'}(\tilde{Q}^{l_k}_{\hat{\pi}_k}(s,\cdot))$.
6: Sample $s \sim \mu_k$ (μ_k is the population state marginal of $\mathcal{L}_k$), obtain $\tilde{\mathcal{L}}_{k+1}$ from $\mathcal{G}(s, \pi_k, \mathcal{L}_k)$.
7: Find $\mathcal{L}_{k+1} = \mathbf{Proj}_{S_\epsilon}(\tilde{\mathcal{L}}_{k+1})$
8: **end for**

We next present the convergence property for the GMF-P algorithm by combining the proofs of Lemma 1.5.4 and Theorem 1.5.1.

Theorem 1.5.5 (Convergence and complexity of GMF-P)**.** *Assume the same assumptions as in Theorem 1.3.1, and in addition that $f_{c,c'} \subseteq \mathcal{F}_{c,c'}$. Suppose that Alg has a policy-based guarantee with parameters*

$$\{C_{\mathcal{M}}^{(i)}, \alpha_1^{(i)}, \alpha_2^{(i)}, \alpha_3^{(i)}, \alpha_4^{(i)}\}_{i=1}^m.$$

Then for any ϵ, $\delta > 0$, set $\delta_k = \delta/K_{\epsilon,\eta}$, $\epsilon_k = (k+1)^{-(1+\eta)}$ for some $\eta \in (0,1]$ ($k = 0, \ldots, K_{\epsilon,\eta} - 1$), and $c \geq c' \geq \frac{\log(1/\epsilon)}{\phi(\epsilon)}$, with probability at least $1 - 2\delta$,

$$W_1(\mathcal{L}_{K_{\epsilon,\eta}}, \mathcal{L}^\star) \leq C_0 \epsilon.$$

Here $K_{\epsilon,\eta} := \left\lceil 2\max\left\{(\eta\epsilon/c)^{-1/\eta}, \log_d(\epsilon/\max\{diam(\mathcal{S})diam(\mathcal{A}), c\}) + 1\right\}\right\rceil$ is the number of outer iterations, and the constant C_0 is independent of δ, ϵ and η.
Moreover, the total number of samples $T = \sum_{k=0}^{K_{\epsilon,\eta}-1} T_{\mathcal{M}_{\mathcal{L}_k}}(\delta_k, \epsilon_k)$ is bounded by

$$T \leq \sum_{i=1}^{m+3} \frac{2^{\alpha_2^{(i)}}}{2\alpha_1^{(i)} + 1} \tilde{C}_{\mathcal{M}}^{(i)} K_{\epsilon,\eta}^{2\alpha_1^{(i)}+1} (K_{\epsilon,\eta}/\delta)^{\alpha_3^{(i)}} \left(\log(K_{\epsilon,\eta}/\delta)\right)^{\alpha_2^{(i)}+\alpha_4^{(i)}}, \tag{1.5.9}$$

where the parameters $\{\tilde{C}_{\mathcal{M}}^{(i)}, \alpha_1^{(i)}, \alpha_2^{(i)}, \alpha_3^{(i)}, \alpha_4^{(i)}\}_{i=1}^{m+3}$ are defined in Lemma 1.5.4.

1.5.2.2 GMF-P-TRPO: GMF-P with TRPO

A special form of the GMF-P algorithm utilizes the trust region policy optimization (TRPO) algorithm [140, 141]. We call it GMF-P-TRPO.

Sample-based TRPO [141] assumes access to a ν-restart model. That is, it can only access sampled trajectories and restarts according to the distribution ν. Here we pick ν such that $C^{\pi^\star} := \left\| \frac{d^{\pi^\star}_{\text{Unif}_{\mathcal{S}}}}{\nu} \right\|_\infty = \max_{s\in\mathcal{S}} \left| \frac{d^{\pi^\star}_{\text{Unif}_{\mathcal{S}}}(s)}{\nu(s)} \right| < \infty$, where $d_\rho^\pi = (1-\gamma)\rho(I - \gamma P^\pi)^{-1}$ and $\text{Unif}_{\mathcal{S}}$ is the uniform distribution on set $\mathcal{S}$. Sample-based TRPO samples M_0 trajectories per episode. The initial state s_0 at the beginning of each episode is sampled from ν. In every trajectory m ($m = 1, 2, \cdots, M_0$) of the l-th episode, it first samples $s_m \sim d_\nu^{\pi_l}$ and takes an action $a_m \sim \text{Unif}_{\mathcal{A}}$ where $\text{Unif}_{\mathcal{A}}$ is the uniform distribution on the set $\mathcal{A}$. Then, by following the current π_l, it estimates $Q^{\pi_l}(s_m, a_m)$ using a rollout. Denote this estimate as $\hat{Q}^{\pi_l}(s_m, a_m, m)$ and observe that it is (nearly) an unbiased estimator of $Q^{\pi_l}(s_m, a_m)$. We assume that each rollout runs sufficiently long so that the bias is sufficiently small. Sample-Based TRPO updates the policy at the end of the l-th episode, by the following proximal problem

$$\pi_{l+1} \in \arg\max_{\pi \in \Delta_{\mathcal{A}}^{|\mathcal{S}|}} \left\{ \frac{1}{M_0} \sum_{m=1}^{M_0} \frac{1}{t_l(1-\gamma)} B_w(s_m; \pi, \pi_l) + \langle \hat{\nabla} V^{\pi_l}[m], \pi(s_m) - \pi_l(s_m) \rangle \right\},$$

where the estimation of the gradient is

$$\hat{\nabla} V^{\pi_l}[m] := \frac{1}{1-\gamma} |\mathcal{A}| \hat{Q}^{\pi_l}(s_m, \cdot, m) \circ \mathbf{I}_{\{\cdot = a_m\}}.$$

Given two policies π_1 and π_2, we denote their Bregman distance associated with a strongly convex function w as $B_w(s;\pi_1,\pi_2) = B_w(\pi_1(s),\pi_2(s))$, where $B_w(x,y) := w(x) - w(y) - \langle \nabla w(y), x-y\rangle$ and $\pi_i(s) \in P(\mathcal{A})$ $(i=1,2)$. Denote $B_w(\pi_1,\pi_2) \in \mathbb{R}^{|\mathcal{S}|}$ as the corresponding state-wise vector. Here we consider two common cases for w: when $w(x) = \frac{1}{2}\|x\|_2^2$ is the Euclidean distance, $B_w(x,y) = \frac{1}{2}\|x-y\|_2^2$; when $w(x) = H(x)$ is the negative entropy, $B_w(x,y) = d_{\mathrm{KL}}(x||y)$. We refer to [141, Section 6.2] for more detailed discussion on Sample-based TRPO.

The above guarantee follows from the sample complexity result below by specifying $\mu := \mathrm{Unif}_{\mathcal{S}}$. Notice that here for any $\mu \in \mathcal{P}(\mathcal{S})$, we define $V^\star(\mu) := \sum_{s\in\mathcal{S}} \mu(s) V^\star(s)$, and similarly $V^{\pi_k}(\mu) := \sum_{s\in\mathcal{S}} \mu(s) V^{\pi_k}(s)$.

The sample complexity of TRPO algorithm can be characterized as below.

Lemma 1.5.6 (Theorem 5 in [141]: sample complexity of TRPO)**.** *Let $\{\pi_l\}_{l\geq 0}$ be the sequence generated by Sample-Based TRPO, using*

$$M_0 \geq \Omega\left(\frac{|\mathcal{A}|^2 C^2 (|\mathcal{S}|\log|\mathcal{A}| + \log 1/\delta)}{(1-\gamma)^2\epsilon^2}\right)$$

samples in each episode, with $t_l = \frac{(1-\gamma)}{C_{\omega,1}C\sqrt{l+1}}$. Let $\{V^N_{best}\}_{N\geq 0}$ be the sequence of best achieved values, $V^N_{best}(\mu) := \max_{l=0,1,\cdots,N} V^{\pi_l}(\mu)$, where $\mu \in \mathcal{P}(\mathcal{S})$. Then with probability greater than $1-\delta$ for every $\epsilon > 0$, the following holds for all $N \geq 1$:

$$V^\star(\mu) - V^N_{best}(\mu) \leq O\left(\frac{C_{\omega,1}C}{(1-\gamma)^2\sqrt{N}} + \frac{C^{\pi^\star}\epsilon}{(1-\gamma)^2}\right).$$

Here $C > 0$ is the upper bound on the reward function r, $C_{w,1} = \sqrt{|\mathcal{A}|}$ in the euclidean case and $C_{w,1} = 1$ in the non-euclidean case, $C_{w,2} = 1$ for the euclidean case and $C_{w,2} = |\mathcal{A}|^2$ for the non-euclidean case. Note that unlike the case of Q-learning, here we are only guaranteed to have *certain* iterate among iterations $0,\ldots,N$ that satisfy the desired sub-optimality bound. Note that this is a common pattern of the theoretical results for policy optimization algorithms in the RL literature [4, 161], unless the (oracle) access to exact policy gradients is assumed [114]. For simplicity, hereafter we assume an oracle access to such an iterate after running TRPO. In practice, with additional (polynomial number of) samples, one can explicitly identify a single policy satisfying the desired bound with high probability; see *e.g.*, the two-phase technique in [65].

Note that [141, Theorem 5] has both regularized version and unregularized version of TRPO. Here we only adopt the unregularized version which fits the framework of Algorithm 4. For more materials on regularized MDPs and reinforcement learning, we refer the readers to [122, 63, 48].

Based on the sample complexity in Lemma 1.5.6, the following policy-based guarantee for TRPO algorithm and the convergence result for GMF-P-TRPO can be obtained.

Corollary 1.5.7. *Let* $t_l = \frac{(1-\gamma)}{C_{\omega,1}C\sqrt{l+1}}$, *then TRPO algorithm satisfies the policy-based guarantee with parameters* $\{\tilde{C}^{(i)}_{\mathcal{M}}, \alpha^{(i)}_1, \alpha^{(i)}_2, \alpha^{(i)}_3, \alpha^{(i)}_4\}^2_{i=1}$, *where* $C^{(i)}_{\mathcal{M}}(i = 1, 2)$ *are constants depending on* $|\mathcal{S}|, |\mathcal{A}|, V_{\max}, \beta$ *and* h, *and we have:*

$$\alpha^{(1)}_1 = 5/2,\ \alpha^{(1)}_j = 0 \text{ for } j = 2, 3, 4,$$
$$\alpha^{(2)}_1 = 5/2,\ \alpha^{(2)}_4 = 1, \alpha^{(2)}_2 = \alpha^{(2)}_3 = 0.$$

In addition, under same assumptions as Theorem 1.3.1, then for Algorithm 4 using TRPO method, with probability at least $1 - 2\delta$, $W_1(\mathcal{L}_{K_{\epsilon,\eta}}, \mathcal{L}^\star) \le C_0\epsilon$, *where* $K_{\epsilon,\eta}$ *is defined as in Theorem 1.5.1. And the total number of samples* $T = \sum_{k=0}^{K_{\epsilon,\eta}-1} T_{\mathcal{M}_{\mathcal{L}_k}}(\delta_k, \epsilon_k)$ *is bounded by*

$$T \le O\left(K^6_{\epsilon,\eta}\left(\log\frac{K_{\epsilon,\eta}}{\delta}\right) + K^{\frac{4}{h}+1}_{\epsilon,\eta}\left(\log\frac{K_{\epsilon,\eta}}{\delta}\right)^{\frac{1}{h}} + \left(\log\frac{K_{\epsilon,\eta}}{\delta}\right)^{\frac{1}{1-h}}\right).$$

1.6 Applications to N-player Games

In this section, we discuss a potential application of our modeling and approach to N-player settings. To this end, we consider extensions of Algorithms 2 and 4 with weaker assumptions on the simulator access. In particular, we weaken the simulator oracle assumption in Section 1.4 as follows.

Weak simulator oracle. For each player i, given any policy $\pi \in \Pi$, the current state $s_i \in \mathcal{S}$, for any empirical population state-action distribution $\mathcal{L}_N$, one can obtain a *sample* of the next state $s'_i \sim P_{\mathcal{L}_N}(\cdot|s_i, \pi(s_i)) = P(\cdot|s_i, \pi(s_i), \mathcal{L}_N)$ and a reward $r = r_{\mathcal{L}_N}(s_i, \pi(s_i)) = r(s_i, \pi(s_i), \mathcal{L}_N)$. For brevity, we denote the simulator as $(s'_i, r) = \mathcal{G}_W(s_i, \pi, \mathcal{L}_N)$.

We say that $\mathcal{L}_N$ is an empirical population state-action distribution of N-players if for each $s \in \mathcal{S},\ a \in \mathcal{A}$, $\mathcal{L}_N(s, a) = \frac{1}{N}\sum_{i=1}^N \mathbf{I}_{s_i=s, a_i=a}$ for some state-action profile of $\{s_i, a_i\}^N_{i=1}$. Equivalently, this holds if $N\mathcal{L}_N(s, a)$ is a non-negative integer for each $s \in \mathcal{S},\ a \in \mathcal{A}$, and $\sum_{s,a}\mathcal{L}_N(s, a) = 1$. We denote the set of empirical population state-action distributions as $\mathbf{Emp}_N$.

RL algorithms with access only to $\mathcal{G}_W$. Compared to the original simulator oracle $\mathcal{G}$, the weak simulator $\mathcal{G}_W$ only accepts empirical population state-action distributions as inputs, and does not directly output the next (empirical) population state-action distribution.

To make use of the simulator $\mathcal{G}_W$, we modify Algorithm 2 and Algorithm 4 to algorithms (Algorithms 5 and 6). In particular, see Step 6 in Algorithm 5 and Step 7 in Algorithm 6 for generating empirical distributions from simulator $\mathcal{G}_W$.

One can observe that $\mathbf{Emp}_N$ already serves as an $1/N$-net. So one can directly use it without additional projections. The definition of $\mathcal{L}_k$ also makes sure that $\mathcal{L}_k \in \mathbf{Emp}_N$ as required for the input of the weaker simulator.

Algorithm 5 GMF-VW(*Alg*, $f_{c,c'}$): weak simulator

Input: Initial $\mathcal{L}_0$, temperatures $c,\ c' > 0$, tolerances $\epsilon_k,\ \delta_k > 0$, $k = 0, 1, \ldots$.
for $k = 0, 1, \cdots$ **do**
  Apply Alg to find the approximate Q-function $\hat{Q}_k^\star = \hat{Q}^{T_k}$ of the MDP $\mathcal{M}_{\mathcal{L}_k}$, where $T_k = T_{\mathcal{M}_{\mathcal{L}_k}}(\epsilon_k, \delta_k)$.
  Compute $\pi_k(s) = f_{c,c'}(\hat{Q}_k^\star(s, \cdot))$.
  for $i = 1, 2, \cdots, N$ **do**
    Sample $s_i \overset{\text{i.i.d.}}{\sim} \mu_k$, then obtain s_i' i.i.d. from $\mathcal{G}_W(s_i, \pi_k, \mathcal{L}_k)$ and $a_i' \overset{\text{i.i.d.}}{\sim} \pi_k(s_i')$.
  end for
  Compute $\mathcal{L}_{k+1}$ with $\mathcal{L}_{k+1}(s, a) = \frac{1}{N}\sum_{i=1}^N \mathbf{I}_{s_i'=s, a_i'=a}$.
end for

Algorithm 6 GMF-PW(*Alg*, $f_{c,c'}$): weak simulator

Input: Initial $\mathcal{L}_0$, temperatures $c,\ c' > 0$, tolerances $\epsilon_k,\ \delta_k > 0$, $k = 0, 1, \ldots$.
for $k = 0, 1, \cdots$ **do**
  Apply Alg to find the approximate policy $\hat{\pi}_k = \pi_{T_k}$ of the MDP $\mathcal{M}_k := \mathcal{M}_{\mathcal{L}_k}$, where $T_k = T_{\mathcal{M}_{\mathcal{L}_k}}(\epsilon_k, \delta_k/2)$.
  Compute $\tilde{Q}_{\hat{\pi}_k}^{l_k}$ using TD updates (1.5.6) for MDP $\mathcal{M}_k$, with l_k satisfying (1.5.7) (with ϵ and δ replaced by ϵ_k and $\delta_k/2$, respectively).
  Compute $\pi_k(s) = f_{c,c'}(\hat{Q}^{\hat{\pi}_k}_{\mathcal{M}_k, M_k, l_k}(s, \cdot))$.
  for $i = 1, 2, \cdots, N$ **do**
    Sample $s_i \overset{\text{i.i.d.}}{\sim} \mu_k$, then obtain s_i' i.i.d. from $\mathcal{G}_W(s_i, \pi_k, \mathcal{L}_k)$ and $a_i' \overset{\text{i.i.d.}}{\sim} \pi_k(s_i')$.
  end for
  Compute $\mathcal{L}_{k+1}$ with $\mathcal{L}_{k+1}(s, a) = \frac{1}{N}\sum_{i=1}^N \mathbf{I}_{s_i'=s, a_i'=a}$.
end for

Convergence results similar to Theorems 1.5.1 and 1.5.5 can be obtained for Algorithms 5 and 6, respectively. (See Section 1.10.3.) Here the major difference is an additional $O(1/\sqrt{N})$ term in the finite step error bound. It is worth mentioning that $O(1/\sqrt{N})$ is consistent with the literature on MFG approximation errors of finite N-player games [87].

1.7 Proof of the main results

1.7.1 Proof of Lemma 1.3.2

In this section, we provide the proof of Lemma 1.3.2.

Proof. [Proof of Lemma 1.3.2] We begin by noticing that $\mathcal{L}' = \Gamma_2(\pi, \mathcal{L})$ can be expanded and computed as follows:

$$\mu'(s') = \sum\nolimits_{s\in\mathcal{S},a\in\mathcal{A}} \mu(s)P(s'|s,a,\mathcal{L})\pi(a|s), \quad \mathcal{L}'(s',a') = \mu'(s')\pi(a'|s'), \tag{1.7.1}$$

where μ is the state marginal distribution of $\mathcal{L}$.

Now by the inequalities (1.10.3), we have

$$\begin{aligned}
&W_1(\Gamma_2(\pi_1,\mathcal{L}),\Gamma_2(\pi_2,\mathcal{L})) \le \mathrm{diam}(\mathcal{S}\times\mathcal{A}) d_{TV}(\Gamma_2(\pi_1,\mathcal{L}),\Gamma_2(\pi_2,\mathcal{L}))\\
=&\frac{\mathrm{diam}(\mathcal{S}\times\mathcal{A})}{2} \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} \left| \sum_{s\in\mathcal{S},a\in\mathcal{A}} \mu(s)P(s'|s,a,\mathcal{L}) \left(\pi_1(a|s)\pi_1(a'|s') - \pi_2(a|s)\pi_2(a'|s')\right)\right|\\
\le&\frac{\mathrm{diam}(\mathcal{S}\times\mathcal{A})}{2} \max_{s,a,\mathcal{L},s'} P(s'|s,a,\mathcal{L}) \sum_{s,a,s',a'} \mu(s)(\pi_1(a|s)+\pi_2(a|s))|\pi_1(a'|s') - \pi_2(a'|s')|\\
\le&\frac{\mathrm{diam}(\mathcal{S}\times\mathcal{A})}{2} \max_{s,a,\mathcal{L},s'} P(s'|s,a,\mathcal{L}) \sum_{s',a'} |\pi_1(a'|s') - \pi_2(a'|s')| \cdot (1+1)\\
=&2\mathrm{diam}(\mathcal{S}\times\mathcal{A}) \max_{s,a,\mathcal{L},s'} P(s'|s,a,\mathcal{L}) \sum_{s'} d_{TV}(\pi_1(s'),\pi_2(s'))\\
\le&\frac{2\mathrm{diam}(\mathcal{S}\times\mathcal{A})\max_{s,a,\mathcal{L},s'} P(s'|s,a,\mathcal{L})|\mathcal{S}|}{d_{\min}(\mathcal{A})} D(\pi_1,\pi_2) = \frac{2\mathrm{diam}(\mathcal{S})\mathrm{diam}(\mathcal{A})|\mathcal{S}|c_1}{d_{\min}(\mathcal{A})} D(\pi_1,\pi_2).
\end{aligned} \tag{1.7.2}$$

Similarly, we have

$$\begin{aligned}
&W_1(\Gamma_2(\pi,\mathcal{L}_1),\Gamma_2(\pi,\mathcal{L}_2)) \le \mathrm{diam}(\mathcal{S}\times\mathcal{A}) d_{TV}(\Gamma_2(\pi,\mathcal{L}_1),\Gamma_2(\pi,\mathcal{L}_2))\\
=&\frac{\mathrm{diam}(\mathcal{S}\times\mathcal{A})}{2} \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} \left| \sum_{s\in\mathcal{S},a\in\mathcal{A}} \mu(s)\pi(a|s)\pi(a'|s') \left(P(s'|s,a,\mathcal{L}_1) - P(s'|s,a,\mathcal{L}_2)\right)\right|\\
\le&\frac{\mathrm{diam}(\mathcal{S}\times\mathcal{A})}{2} \sum_{s,a,s',a'} \mu(s)\pi(a|s)\pi(a'|s') \left|P(s'|s,a,\mathcal{L}_1) - P(s'|s,a,\mathcal{L}_2)\right|\\
\le&\frac{\mathrm{diam}(\mathcal{S})\mathrm{diam}(\mathcal{A})c_2}{2}.
\end{aligned} \tag{1.7.3}$$

This completes the proof. □

1.7.2 Proof of Lemma 1.5.4

For notation simplicity, in the following analysis we fix the MDP and omit the notation $\mathcal{M}$.

We begin by establishing the convergence rate of the synchronous TD updates (1.5.6).

Lemma 1.7.1. *Take $\tilde{Q}^l_\pi$ from (1.5.6). Then for any $\delta \in (0,1)$ and $\epsilon > 0$, with probability at least $1-\delta$, $\|\tilde{Q}^l_{\pi_{T(\epsilon,\delta)}} - Q^{\pi_{T(\epsilon,\delta)}}\|_\infty \leq \epsilon$ if*

$$l = \Omega\left(\left(\frac{V_{\max}\log\left(\frac{|\mathcal{S}||\mathcal{A}|V_{\max}}{\delta\beta\epsilon}\right)}{\beta^2\epsilon^2}\right)^{1/h} + \left(\frac{1}{\beta}\log\frac{V_{\max}}{\epsilon}\right)^{1/(1-h)}\right), \tag{1.7.4}$$

where $V_{\max} = R_{\max}/(1-\gamma)$ and $\beta = (1-\gamma)/2$.

The proof is adapted from that of [53, Theorem 2], with the max term in the Bellman operator modified to actions sampled from the current policy π. The details are omitted.

Proof. [Proof of Lemma 1.5.4] First, if $V^\star(s') - V^{\pi_{T(\epsilon,\delta/2)}}(s') \leq \epsilon$, then

$$\begin{aligned} |Q^{\pi_{T(\epsilon,\delta/2)}}(s,a) - Q^\star(s,a)| &= \gamma\left|\sum_{s'\in\mathcal{S}} P(s'|s,a)V^{\pi_{T(\epsilon,\delta/2)}}(s') - \sum_{s'\in\mathcal{S}} P(s'|s,a)V^\star(s')\right| \\ &\leq \gamma\sum_{s'\in\mathcal{S}} P(s'|s,a)\,|V^{\pi_{T(\epsilon,\delta/2)}}(s') - V^\star(s')| \leq \gamma\epsilon < \epsilon. \end{aligned} \tag{1.7.5}$$

for any $s \in \mathcal{S}$, $a \in \mathcal{A}$. Since *Alg* is assumed to satisfying the policy-based guarantee, (1.7.5) holds with probability at least $1-\delta$.

In addition, by Lemma 1.7.1, whenever l satisfies (1.5.7), with probability at least $1 - \delta/2 \geq 1-\delta$,

$$\|\tilde{Q}^l_{\pi_{T(\epsilon,\delta/2)}} - Q^{\pi_{T(\epsilon,\delta/2)}}\|_\infty \leq (1-\gamma)\epsilon. \tag{1.7.6}$$

Combining (1.7.5) and (1.7.6), then for any l satisfying (1.5.7), with probability at least $1-2\delta$, we have

$$\|\tilde{Q}^l_{\pi_{T(\epsilon,\delta/2)}} - Q^\star\|_\infty \leq \gamma\epsilon + (1-\gamma)\epsilon = \epsilon.$$

The above result shows that for any $\delta \in (0,1)$ and $\epsilon > 0$, after obtaining $T(\epsilon,\delta/2) + |\mathcal{S}||\mathcal{A}|l$ samples (with l satisfying the lower bound (1.5.7)) from the simulator, with probability at least $1-2\delta$, it outputs an approximate Q-function $\tilde{Q}^l_{\pi_{T(\epsilon,\delta/2)}}$ which satisfies $\|\tilde{Q}^l_{\pi_{T(\epsilon,\delta/2)}} - Q^\star\|_\infty \leq \epsilon$. Thus *Alg* also has a value-based guarantee with parameters

$$\{\tilde{C}^{(i)}_{\mathcal{M}}, \alpha^{(i)}_1, \alpha^{(i)}_2, \alpha^{(i)}_3, \alpha^{(i)}_4\}_{i=1}^{m+3}, \tag{1.7.7}$$

specified in (1.5.8). Here the first m groups of parameters come from $T(\epsilon,\delta/2)$ while the last three groups of parameters come from $|\mathcal{S}||\mathcal{A}|l$ (with the lower bound (1.5.7) of l plugged in here). □

1.7.3 Proof of $\mathcal{B}_{c,c'} \subseteq \mathcal{F}_{c,c'}$

Lemma 1.7.2. *Suppose that $h : \mathbb{R} \to \mathbb{R}$ satisfies $h(a) - h(b) \le c(a-b)$ for any $a \ge b \in \mathbb{R}$. Then the softmax function $\mathbf{softmax}_h$ is c-Lipschitz, i.e., $\|\mathbf{softmax}_h(x) - \mathbf{softmax}_h(y)\|_2 \le c\|x-y\|_2$ for any $x,\ y \in \mathbb{R}^n$.*

Proof. [Proof of Lemma 1.7.2] Notice that $\mathbf{softmax}_h(x) = \mathbf{softmax}(\tilde{h}(x))$, where

$$\mathbf{softmax}(x)_i = \frac{\exp(x_i)}{\sum_{j=1}^n \exp(x_j)} (i = 1, \dots, n)$$

is the standard softmax function and $\tilde{h}(x)_i = h(x_i)$ for $i = 1, \dots, n$. Now since **softmax** is 1-Lipschitz continuous (*cf.* [62, Proposition 4]), and $\tilde{h}$ is c-Lipschitz continuous, we conclude that the composition $\mathbf{softmax} \circ \tilde{h}$ is c-Lipschitz continuous. □

Notice that for a finite set $\mathcal{X} \subseteq \mathbb{R}^k$ and any two (discrete) distributions $\nu,\ \nu'$ over $\mathcal{X}$, we have

$$W_1(\nu, \nu') \le \mathrm{diam}(\mathcal{X}) d_{TV}(\nu, \nu') = \frac{\mathrm{diam}(\mathcal{X})}{2}\|\nu - \nu'\|_1 \le \frac{\mathrm{diam}(\mathcal{X})\sqrt{|\mathcal{X}|}}{2}\|\nu - \nu'\|_2, \quad (1.7.8)$$

where in computing the ℓ_1-norm, $\nu,\ \nu'$ are viewed as vectors of length $|\mathcal{X}|$.

Lemma 1.7.2 implies that for any $x,\ y \in \mathbb{R}^{|\mathcal{X}|}$, when $\mathbf{softmax}_c(x)$ and $\mathbf{softmax}_c(y)$ are viewed as probability distributions over $\mathcal{X}$, we have

$$W_1(\mathbf{softmax}_c(x), \mathbf{softmax}_c(y)) \le \frac{\mathrm{diam}(\mathcal{X})\sqrt{|\mathcal{X}|}c}{2}\|x-y\|_2 \le \frac{\mathrm{diam}(\mathcal{X})|\mathcal{X}|c}{2}\|x-y\|_\infty.$$

Lemma 1.7.3. *Suppose that $h : \mathbb{R} \to \mathbb{R}$ satisfies $c'(a-b) \le h(a) - h(b)$ for any $a \le b \in \mathbb{R}$. Then for any $x \in \mathbb{R}^n$, the distance between the $\mathbf{softmax}_h$ and the **argmax-e** mapping is bounded by*

$$\|\mathbf{softmax}_h(x) - \mathbf{argmax\text{-}e}(x)\|_2 \le 2n\exp(-c'\delta),$$

where $\delta = x_{\max} - \max_{x_j < x_{\max}} x_j$, $x_{\max} = \max_{i=1,\dots,n} x_i$, and $\delta := \infty$ when all x_j are equal.

Similar to Lemma 1.7.2, Lemma 1.7.3 implies that for any $x \in \mathbb{R}^{|\mathcal{X}|}$, viewing $\mathbf{softmax}_h(x)$ as probability distributions over $\mathcal{X}$ leads to

$$W_1(\mathbf{softmax}_h(x), \mathbf{argmax\text{-}e}(x)) \le \mathrm{diam}(\mathcal{X})|\mathcal{X}|\exp(-c\delta).$$

Proof. [Proof of Lemma 1.7.3] Without loss of generality, assume that $x_1 = x_2 = \cdots = x_m = \max_{i=1,\ldots,n} x_i = x^\star > x_j$ for all $m < j \leq n$. Then

$$\textbf{argmax-e}(x)_i = \begin{cases} \frac{1}{m}, & i \leq m, \\ 0, & otherwise. \end{cases}$$

$$\textbf{softmax}_h(x)_i = \begin{cases} \frac{e^{h(x^\star)}}{me^{h(x^\star)} + \sum_{j=m+1}^n e^{h(x_j)}}, & i \leq m, \\ \frac{e^{h(x_i)}}{me^{h(x^\star)} + \sum_{j=m+1}^n e^{h(x_j)}}, & otherwise. \end{cases}$$

Therefore

$$\begin{aligned} &\|\textbf{softmax}_h(x) - \textbf{argmax-e}(x)\|_2 \leq \|\textbf{softmax}_h(x) - \textbf{argmax-e}(x)\|_1 \\ =&m\left(\frac{1}{m} - \frac{e^{h(x^\star)}}{me^{h(x^\star)} + \sum_{j=m+1}^n e^{h(x_j)}}\right) + \frac{\sum_{i=m+1}^n e^{h(x_i)}}{me^{h(x^\star)} + \sum_{j=m+1}^n e^{h(x_j)}} \\ =&\frac{2\sum_{i=m+1}^n e^{h(x_i)}}{me^{h(x^\star)} + \sum_{i=m+1}^n e^{h(x_i)}} = \frac{2\sum_{i=m+1}^n e^{-c'\delta_i}}{m + \sum_{i=m+1}^n e^{-c\delta_i}} \\ \leq&\frac{2}{m}\sum_{i=m+1}^n e^{-c'\delta_i} \leq \frac{2(n-m)}{m}e^{-c'\delta} \leq 2ne^{-c'\delta}, \end{aligned}$$

with $\delta_i = x^\star - x^i$. □

We are now ready to present the proofs of Theorems 1.5.1 and 1.5.5.

1.7.4 Proof of Theorems 1.5.1 and 1.5.5

Proof. [Proof of Theorem 1.5.1] Here we prove the case when we are using GMF-V and *Alg* has a value-based guarantee. Define $\hat{\Gamma}_1^k(\mathcal{L}_k) := f_{c,c'}\left(\hat{Q}_k^\star\right)$. In the following, $\pi = f_{c,c'}(Q_\mathcal{L})$ is understood as the policy π with $\pi(s) = f_{c,c'}(Q_\mathcal{L}(s,\cdot))$. Let $\mathcal{L}^\star$ be the population state-action pair in a stationary NE of (GMFG). Then $\pi_k = \hat{\Gamma}_1^k(\mathcal{L}_k)$. Denoting $d := d_1d_2 + d_3$, we see

$$\begin{aligned} W_1(\tilde{\mathcal{L}}_{k+1}, \mathcal{L}^\star) &= W_1(\Gamma_2(\pi_k, \mathcal{L}_k), \Gamma_2(\Gamma_1(\mathcal{L}^\star), \mathcal{L}^\star)) \\ &\leq W_1(\Gamma_2(\Gamma_1(\mathcal{L}_k), \mathcal{L}_k), \Gamma_2(\Gamma_1(\mathcal{L}^\star), \mathcal{L}^\star)) + W_1(\Gamma_2(\Gamma_1(\mathcal{L}_k), \mathcal{L}_k), \Gamma_2(\hat{\Gamma}_1^k(\mathcal{L}_k), \mathcal{L}_k)) \\ &\leq W_1(\Gamma(\mathcal{L}_k), \Gamma(\mathcal{L}^\star)) + d_2 D(\Gamma_1(\mathcal{L}_k), \hat{\Gamma}_1^k(\mathcal{L}_k)) \\ &\leq (d_1d_2 + d_3)W_1(\mathcal{L}_k, \mathcal{L}^\star) + d_2 D(\textbf{argmax-e}(Q^\star_{\mathcal{L}_k}), f_{c,c'}(\hat{Q}_k^\star)) \\ &\leq dW_1(\mathcal{L}_k, \mathcal{L}^\star) + d_2 D(f_{c,c'}(\hat{Q}_k^\star), f_{c,c'}(Q^\star_{\mathcal{L}_k})) \\ &\quad + d_2 D(\textbf{argmax-e}(Q^\star_{\mathcal{L}_k}), f_{c,c'}(Q^\star_{\mathcal{L}_k})) \\ &\leq dW_1(\mathcal{L}_k, \mathcal{L}^\star) + \frac{cd_2\text{diam}(\mathcal{A})|\mathcal{A}|}{2}\|\hat{Q}_k^\star - Q^\star_{\mathcal{L}_k}\|_\infty + d_2 D(\textbf{argmax-e}(Q^\star_{\mathcal{L}_k}), f_{c,c'}(Q^\star_{\mathcal{L}_k})). \end{aligned}$$

Since $\mathcal{L}_k \in S_\epsilon$ by the projection step, by Lemma 1.7.3 and the algorithm Alg has a policy-based guarantee, with the choice of $T_k = T_{\mathcal{M}_{\mathcal{L}_k}}(\delta_k, \epsilon_k))$, we have, with probability at least $1 - 2\delta_k$,

$$W_1(\tilde{\mathcal{L}}_{k+1}, \mathcal{L}^\star) \leq dW_1(\mathcal{L}_k, \mathcal{L}^\star) + \frac{cd_2\text{diam}(\mathcal{A})|\mathcal{A}|}{2}\epsilon_k + d_2\text{diam}(\mathcal{A})|\mathcal{A}|e^{-c'\phi(\epsilon)}. \tag{1.7.9}$$

Finally, with probability at least $1 - 2\delta_k$,

$$\begin{aligned} W_1(\mathcal{L}_{k+1}, \mathcal{L}^\star) &\leq W_1(\tilde{\mathcal{L}}_{k+1}, \mathcal{L}^\star) + W_1(\tilde{\mathcal{L}}_{k+1}, \mathbf{Proj}_{S_\epsilon}(\tilde{\mathcal{L}}_{k+1})) \\ &\leq dW_1(\mathcal{L}_k, \mathcal{L}^\star) + \frac{cd_2\text{diam}(\mathcal{A})|\mathcal{A}|}{2}\epsilon_k + d_2\text{diam}(\mathcal{A})|\mathcal{A}|e^{-c'\phi(\epsilon)} + \epsilon. \end{aligned}$$

This implies that with probability at least $1 - 2\sum_{k=0}^{K-1}\delta_k$,

$$\begin{aligned} W_1(\mathcal{L}_K, \mathcal{L}^\star) \leq & d^K W_1(\mathcal{L}_0, \mathcal{L}^\star) + \frac{cd_2\text{diam}(\mathcal{A})|\mathcal{A}|}{2}\sum_{k=0}^{K-1} d^{K-k}\epsilon_k \\ & + \frac{(d_2\text{diam}(\mathcal{A})|\mathcal{A}|e^{-c'\phi(\epsilon)} + \epsilon)(1 - d^K)}{1 - d}. \end{aligned} \tag{1.7.10}$$

Since ϵ_k is summable, we have $\sup_{k\geq 0}\epsilon_k < \infty$,

$$\sum_{k=0}^{K-1} d^{K-k}\epsilon_k \leq \frac{\sup_{k\geq 0}\epsilon_k}{1-d} d^{\lfloor (K-1)/2 \rfloor} + \sum_{k=\lceil (K-1)/2 \rceil}^{\infty} \epsilon_k.$$

Now plugging in $K = K_{\epsilon,\eta}$, with the choice of δ_k and $c = \frac{\log(1/\epsilon)}{\phi(\epsilon)}$, and noticing that $d \in [0, 1)$, we have with probability at least $1 - 2\delta$,

$$\begin{aligned} W_1(\mathcal{L}_{K_{\epsilon,\eta}}, \mathcal{L}^\star) \leq & d^{K_{\epsilon,\eta}} W_1(\mathcal{L}_0, \mathcal{L}^\star) \\ & + \frac{cd_2\text{diam}(\mathcal{A})|\mathcal{A}|}{2}\left(\frac{\sup_{k\geq 0}\epsilon_k}{1-d} d^{\lfloor (K_{\epsilon,\eta}-1)/2 \rfloor} + \sum_{k=\lceil (K_{\epsilon,\eta}-1)/2 \rceil}^{\infty} \epsilon_k\right) \\ & + \frac{(d_2\text{diam}(\mathcal{A})|\mathcal{A}| + 1)\epsilon}{1-d}. \end{aligned} \tag{1.7.11}$$

Setting $\epsilon_k = (k+1)^{-(1+\eta)}$, then when $K_{\epsilon,\eta} \geq 2(\log_d(\epsilon/c) + 1)$,

$$\frac{\sup_{k\geq 0}\epsilon_k}{1-d} d^{\lfloor (K_{\epsilon,\eta}-1)/2 \rfloor} \leq \frac{\epsilon/c}{1-d}.$$

Similarly, when $K_{\epsilon,\eta} \geq 2(\eta\epsilon/c)^{-1/\eta}$,

$$\sum_{k=\left\lceil \frac{K_{\epsilon,\eta}-1}{2} \right\rceil}^{\infty} \epsilon_k \leq \epsilon/c.$$

Finally, when $K_{\epsilon,\eta} \geq \log_d(\epsilon/(\text{diam}(\mathcal{S})\text{diam}(\mathcal{A})))$, $d^{K_{\epsilon,\eta}} W_1(\mathcal{L}_0, \mathcal{L}^\star) \leq \epsilon$, since $W_1(\mathcal{L}_0, \mathcal{L}^\star) \leq \text{diam}(\mathcal{S} \times \mathcal{A}) = \text{diam}(\mathcal{S})\text{diam}(\mathcal{A})$.

In summary, if $K_{\epsilon,\eta} = \lceil 2 \max\{(\eta\epsilon/c)^{-1/\eta}, \log_d(\epsilon/\max\{\text{diam}(\mathcal{S})\text{diam}(\mathcal{A}), c\}) + 1\}\rceil$, then with probability at least $1 - 2\delta$,

$$W_1(\mathcal{L}_{K_{\epsilon,\eta}}, \mathcal{L}^\star) \leq \left(1 + \frac{d_2 \text{diam}(\mathcal{A})|\mathcal{A}|(2-d)}{2(1-d)} + \frac{(d_2 \text{diam}(\mathcal{A})|\mathcal{A}| + 1)}{1-d}\right)\epsilon = O(\epsilon). \quad (1.7.12)$$

Finally, if we are using GMF-V and have assumed that *Alg* satisfies a value-based guarantee with parameters $\{C_{\mathcal{M}}^{(i)}, \alpha_1^{(i)}, \alpha_2^{(i)}, \alpha_3^{(i)}, \alpha_4^{(i)}\}_{i=1}^m$, plugging in ϵ_k and δ_k into $T_{\mathcal{M}_\mathcal{L}}(\delta_k, \epsilon_k)$, and noticing that $k \leq K_{\epsilon,\eta}$ and $\sum_{k=0}^{K_{\epsilon,\eta}-1}(k+1)^\alpha \leq \frac{K_{\epsilon,\eta}^{\alpha+1}}{\alpha+1}$, we have

$$\begin{aligned}
T &= \sum_{k=0}^{K_{\epsilon,\eta}} \sum_{i=1}^{m} C_{\mathcal{M}}^{(i)} \left(\frac{1}{\epsilon_k}\right)^{\alpha_1^{(i)}} \left(\log \frac{1}{\epsilon_k}\right)^{\alpha_2^{(i)}} \left(\frac{1}{\delta_k}\right)^{\alpha_3^{(i)}} \left(\log \frac{1}{\delta_k}\right)^{\alpha_4^{(i)}} \\
&= \sum_{k=0}^{K_{\epsilon,\eta}} \sum_{i=1}^{m} (1+\eta)^{\alpha_2^{(i)}} C_{\mathcal{M}}^{(i)} (k+1)^{\alpha_1^{(i)}(1+\eta)} (\log(k+1))^{\alpha_2^{(i)}} (K_{\epsilon,\eta}/\delta)^{\alpha_3^{(i)}} (\log(K_{\epsilon,\eta}/\delta))^{\alpha_4^{(i)}} \\
&\leq \sum_{i=1}^{m} \frac{(1+\eta)^{\alpha_2^{(i)}}}{\alpha_1^{(i)}(1+\eta)+1} C_{\mathcal{M}}^{(i)} K_{\epsilon,\eta}^{\alpha_1^{(i)}(1+\eta)+1} (\log(K_{\epsilon,\eta}+1))^{\alpha_2^{(i)}} (K_{\epsilon,\eta}/\delta)^{\alpha_3^{(i)}} (\log(K_{\epsilon,\eta}/\delta))^{\alpha_4^{(i)}} \\
&\leq \sum_{i=1}^{m} \frac{2^{\alpha_2^{(i)}}}{2\alpha_1^{(i)}+1} C_{\mathcal{M}}^{(i)} K_{\epsilon,\eta}^{2\alpha_1^{(i)}+1} (K_{\epsilon,\eta}/\delta)^{\alpha_3^{(i)}} (\log(K_{\epsilon,\eta}/\delta))^{\alpha_2^{(i)}+\alpha_4^{(i)}},
\end{aligned} \quad (1.7.13)$$

which completes the proof of the value-based case. □

Proof. [Proof of Theorem 1.5.5] If we use GMF-P and assume that *Alg* has the policy-based guarantee, then by Lemma 1.5.4,

$$\mathbb{P}\left(\left\|\tilde{Q}_{\hat{\pi}_k}^{l_k} - Q_{\mathcal{L}_k}^\star\right\|_\infty > \epsilon\right) \leq 2\delta. \quad (1.7.14)$$

Hence one can simply replace $\hat{Q}_k^\star$ by $\tilde{Q}_{\hat{\pi}_k}^{l_k}$ in the proof of Theorem 1.5.1, and obtain the same bound on $W_1(\mathcal{L}_{K_{\epsilon,\eta}}, \mathcal{L}^\star)$ (*cf.* (1.7.12)). The only difference is that in each iteration, the required number of samples $T_{\mathcal{M}_\mathcal{L}}$ now has parameters $\{\tilde{C}_{\mathcal{M}}^{(i)}, \alpha_1^{(i)}, \alpha_2^{(i)}, \alpha_3^{(i)}, \alpha_4^{(i)}\}_{i=1}^{m+3}$ as defined in Lemma 1.5.4. Hence repeating the proof of (1.7.13) leads to (1.5.9). □

1.8 Experiments

In this section, we report the performance of the proposed GMF-V-Q Algorithm and GMF-P-TRPO Algorithm with an equilibrium pricing model (see Section 1.2.3). The objectives of the experiments include 1) testing the convergence and stability of both GMF-V-Q

and GMF-P-TRPO in the GMFG setting, 2) empirically verifying the contractive property of mapping Γ, and 3) comparing GMF-V-Q and GMF-P-TRPO with existing multi-agent reinforcement learning algorithms, including the Independent Learner (IL) algorithm [150, 85] and the MF-Q[3] algorithm [167]. Another set of experiments for the repeated auction model (see Section 1.2.3) is demonstrated in the short version [76].

1.8.1 Set-up and parameter configuration

We introduce two testing environments in our numerical experiments, one is the GMFG environment with a continuum of agents (i.e., infinite number of agents) descried in Section 1.2.3 and the other one is an N-player environment with a weak simulator.

Equilibrium price as an N-player game. We also consider an N-player game version of the equilibrium price model, which is the GMFG version described above with an N-player weak simulator oracle as described in Section 1.6. In particular, Take N companies. At each time t, company i decides a quantity q_t^i for production and a quantity h_t^i to replenish the inventory. Let s_t^i denote the current inventory level of company i at time t. Then similar to Section 1.2.3, the inventory level evolves according to

$$s_{t+1}^i = s_t^i - \min\{q_t^i, s_t^i\} + h_t$$

and the reward of company i at time t is given by

$$r_t^i = (p_t - c_0)q_t^i - c_1(q_t^i)^2 - c_2 h_t^i - (c_2 + c_3)\max\{q_t^i - s_t^i, 0\} - c_4 s_t^i.$$

Here p_t, the price of the product at time t, is determined according to the supply-demand equilibrium on the market. The total supply is $\sum_{i=1}^N q_t^i$, while the total demand is assumed to be $d_N p_t^{-\sigma}$, where $d_N = dN$ is supposed to be linearly growing as N grows, *i.e.*, the number of customers grows proportionally to the number of producers in the market. Then by equating supply and demand, we obtain that

$$\frac{1}{N}\sum_{i=1}^N q_t^i = d p_t^{-\sigma},$$

and by taking the limit $N \to \infty$, we obtain the mean-field counterpart (1.2.11).

In this setting, accordingly, we test the performance of GMF-VW-Q, which is GMF-VW (Algorithm 5) with synchronous Q-learning and the standard **softmax** operator (*cf.* Algorithm 7) and GMF-PW-TRPO, which is GMF-PW (Algorithm 6) with TRPO and the standard **softmax** operator.[4]

[3]Note that MF-Q is designed for global states and coupled local actions, while in our equilibrium price example we have coupled local (private) states and decoupled local actions. To suit this setting, we adapt MF-Q by replacing the mean-field action term with the mean-field state term.

[4]For the sake of brevity, we omit the algorithm frame for GMF-PW-TRPO.

Algorithm 7 Q-learning for GMFGs (GMF-VW-Q): weak simulator

Input: Initial $\mathcal{L}_0$, ϵ-net S_ϵ, tolerances $\epsilon_k,\ \delta_k > 0,\ k = 0, 1, \ldots$.
for $k = 0, 1, \cdots$ **do**
  Perform Q-learning with hyper-parameters in Lemma 1.5.2 for $T_k = T_{\mathcal{M}_{\mathcal{L}_k}}(\epsilon_k, \delta_k)$ iterations to find the approximate Q-function $\hat{Q}^\star_k = \hat{Q}^{T_k}$ of the MDP $\mathcal{M}_{\mathcal{L}_k}$.
  Compute $\pi_k \in \Pi$ with $\pi_k(s) = \mathbf{softmax}_c(\hat{Q}^\star_k(s, \cdot))$.
  for $i = 1, 2, \cdots, N$ **do**
    Sample $s_i \overset{\text{i.i.d.}}{\sim} \mu_k$, then obtain s'_i i.i.d. from $\mathcal{G}_W(s_i, \pi_k, \mathcal{L}_k)$ and $a'_i \overset{\text{i.i.d.}}{\sim} \pi_k(s'_i)$.
  end for
  Compute $\mathcal{L}_{k+1}$ with $\mathcal{L}_{k+1}(s, a) = \frac{1}{N}\sum_{i=1}^N \mathbf{I}_{s'_i=s, a'_i=a}$.
end for

Parameters. The model parameters are (unless otherwise specified): $\gamma = 0.2$, $d = 50$ and $\sigma = 2$. $S = Q = H = 10$ and hence $|\mathcal{S}| = 10$ and $|\mathcal{A}| = 100$. $c_0 = 0.5$, $c_1 = 0.1$, $c_2 = 0.5$, $c_3 = 0.2$ and $c_4 = 0.2$.

The algorithm parameters are (unless otherwise specified): the temperature parameter is set as $c = 4.0$ and the learning rate is set as $\eta = 0.01$ [5]. For simplicity, we set the inner iteration T_k to be $100 \times |\mathcal{S}| \times |\mathcal{A}|$. The 90%-confidence intervals are calculated with 20 sample paths.

1.8.2 Performance evaluation in the GMFG setting.

Our experiments show that GMF-V-Q and GMF-P-TRPO Algorithms are efficient and robust.

Performance metric. We adopt the following metric to measure the difference between a given policy π and an NE (here $\epsilon_0 > 0$ is a safeguard, and is taken as 0.1 in the experiments):

$$C_{MF}(\pi) = \frac{\max_{\pi'} \mathbb{E}_{s\sim\mu}[V(s, \pi', \mathcal{L})] - V(s, \pi, \mathcal{L})}{|\max_{\pi'} \mathbb{E}_{s\sim\mu}[V(s, \pi', \mathcal{L})]| + \epsilon_0}.$$

Here μ is the invariant distribution of the transition matrix P^π, where

$$P^\pi(s, s') = \sum_{a\in\mathcal{A}} P(s'|s, a)\pi(a|s)$$

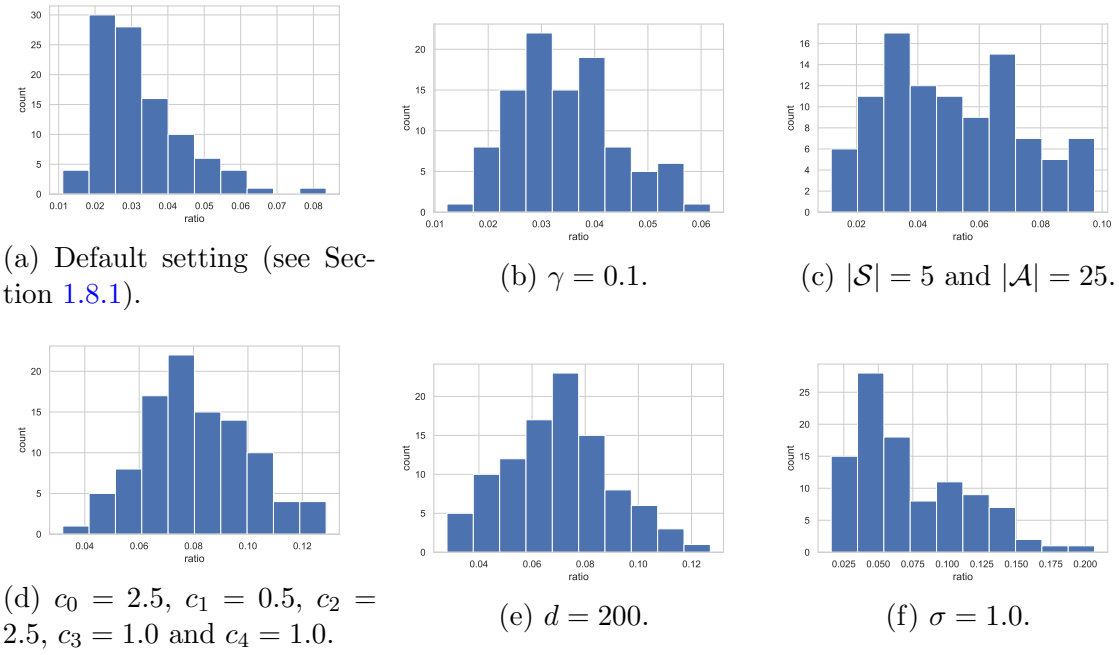
for $s, s' \in \mathcal{S}$, and $\mathcal{L}(s, a) = \mu(s)\pi(a|s)$ for $s, a \in \mathcal{S} \times \mathcal{A}$. Note that in the equilibrium product pricing model we are considering here, the transition model P is independent of the

[5] Lemma 1.5.2 indicates that the learning rate should be inversely proportional to the current visitation number of a given state-action pair, we observe that constant learning rate works well in practice which is easier to implement.

mean-field term $\mathcal{L}$, and hence we write $P(s'|s,a) = P(s'|s,a,\mathcal{L})$. In general, an additional mean-field matching error term needs to be added into the definition of $C_{MF}(\pi)$. Clearly $C_{MF}(\pi) \geq 0$, and $C_{MF}(\pi^\star) = 0$ if and only if $(\pi^\star, \mathcal{L}^\star)$ is an NE where $\mathcal{L}^\star$ is the invariant distribution of $P^{\pi^\star}$. A similar metric without normalization has been adopted in [41].

Contractiveness of mapping Γ. As explained in Remark 1.3.1 from Section 1.3, the contractiveness property of Γ is the key for establishing the uniqueness of MFG solution and hence the convergence of the GMFG algorithm. To empirically verify whether this property holds for the equilibrium price example, we plot the value of $\frac{\|\Gamma(\mathcal{L}_1)-\Gamma(\mathcal{L}_2)\|_1}{\|\mathcal{L}_1-\mathcal{L}_2\|_1}$ for randomly generated state-action distributions $\mathcal{L}_1$ and $\mathcal{L}_2$. Technically speaking, Γ is contractive and there exists a unique MFG solution if the value of $\frac{\|\Gamma(\mathcal{L}_1)-\Gamma(\mathcal{L}_2)\|_1}{\|\mathcal{L}_1-\mathcal{L}_2\|_1}$ is smaller than one for all choices of $\mathcal{L}_1$ and $\mathcal{L}_2$.

W observe from Figure 1.1 that, with various of choices of different model parameters, the quantity $\frac{\|\Gamma(\mathcal{L}_1)-\Gamma(\mathcal{L}_2)\|_1}{\|\mathcal{L}_1-\mathcal{L}_2\|_1}$ is always smaller than 0.3 indicating that Γ is contractive.



(a) Default setting (see Section 1.8.1).

(b) $\gamma = 0.1$.

(c) $|\mathcal{S}| = 5$ and $|\mathcal{A}| = 25$.

(d) $c_0 = 2.5$, $c_1 = 0.5$, $c_2 = 2.5$, $c_3 = 1.0$ and $c_4 = 1.0$.

(e) $d = 200$.

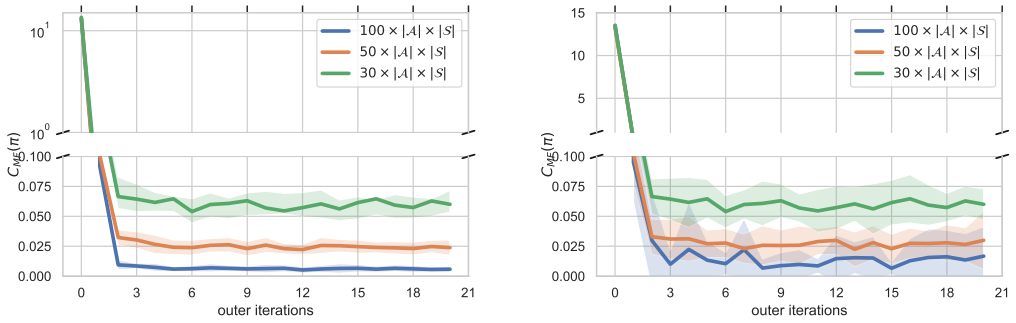
(f) $\sigma = 1.0$.

Figure 1.1: Histogram of $\frac{\|\Gamma(\mathcal{L}_1)-\Gamma(\mathcal{L}_2)\|_1}{\|\mathcal{L}_1-\mathcal{L}_2\|_1}$ under various settings ($\mathcal{L}_1$ and $\mathcal{L}_2$ are randomly sampled according to the uniform distribution).

Convergence and stability. Both GMF-V-Q and GMF-P-TRPO are efficient and robust. First, both GMF-V-Q and GMF-P-TRPO converge within about 5 outer iterations; secondly, as the number of inner iterations increases, the error decreases (Figure 1.2); and finally, the convergence is robust with respect to both the change of number of states and

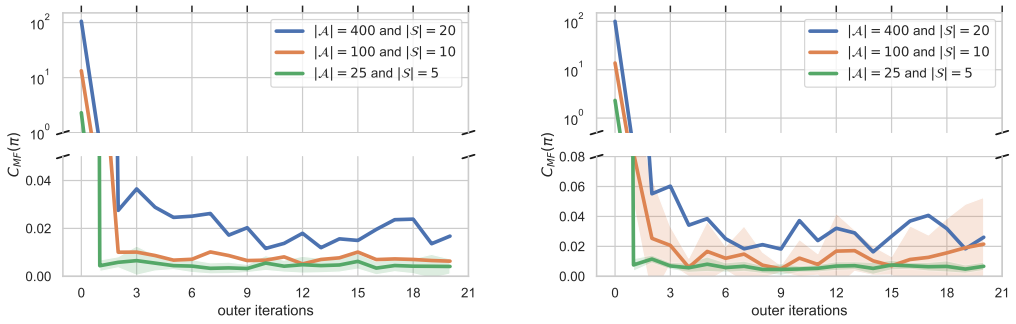
actions (Figure 1.3). The performance of GMF-V-Q is (slightly) more stable than GMF-P-TRPO with a smaller variance across 20 repeated experiments (see Figure 1.2a versus Figure 1.2b or Figure 1.3a versus Figure 1.3b). This is due to the fact that GMF-P-TRPO uses asynchronous updates, which leads to slightly less stable performance compared to GMF-V-Q, which uses synchronous updates.

In contrast, the Naive algorithms, i.e., GMF-V-Q without smoothing (denoted as GMF-V-Q-nonsmoothing) and GMF-P-TRPO without smoothing (denoted as GMF-P-TRPO-nonsmoothing), do not converge even with 50 outer iterations and $200 \times |\mathcal{S}| \times |\mathcal{A}|$ inner iterations within each outer iteration. In particular, GMF-V-Q-nonsmoothing and GMF-P-TRPO-nonsmoothing present different unstable behaviors (see Figure 1.4). The joint distribution $\mathcal{L}_t$ from GMF-V-Q-nonsmoothing keeps fluctuating (Figure 1.4a) whereas the joint distribution $\mathcal{L}_t$ from GMF-P-TRPO (without smoothing) is trapped around the initialization which is far away from the true equilibrium distribution (Figure 1.4b).



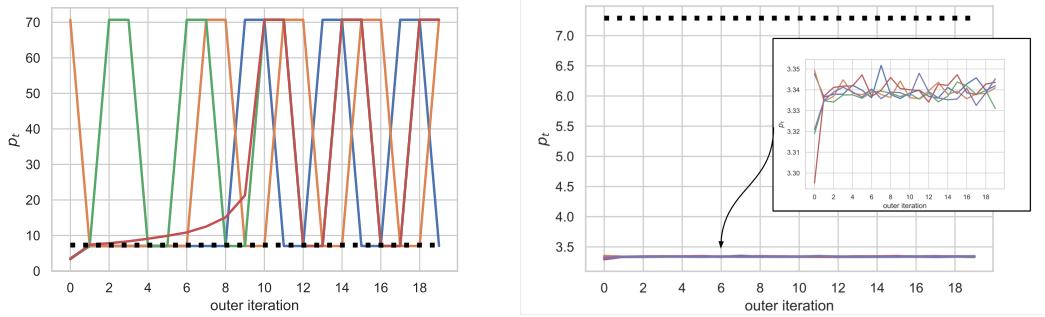
(a) GMF-V-Q. (b) GMF-P-TRPO.

Figure 1.2: Convergence with different number of inner iterations ($|\mathcal{A}| = 100$ and $|\mathcal{S}| = 10$).



(a) GMF-V-Q. (b) GMF-P-TRPO.

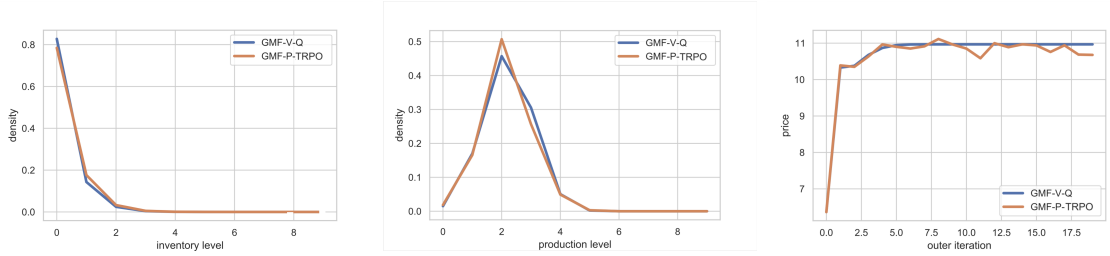
Figure 1.3: Convergence with different size of state space and action space.



(a) Q-learning (5 sample paths). (b) TRPO (5 sample paths).

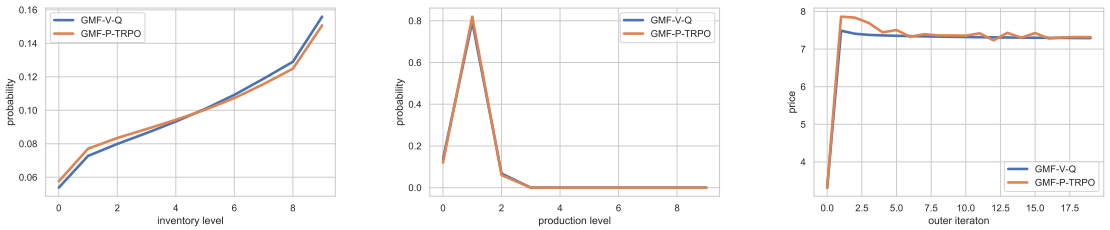
Figure 1.4: Fluctuations of algorithms without smoothing (Dotted black line: theoretical value of the equilibrium price).

Model verification and interpretation of equilibrium scenario. In Figures 1.5 and 1.6, we run both algorithms for 20 outer iterations with the same number of inner iterations (100,000 $= 100 \times |\mathcal{A}| \times |\mathcal{S}|$) within each outer iteration. The final equilibrium inventory distribution and production distribution from both algorithms are close to each other.



(a) Equilibrium inventory distribution. (b) Equilibrium production distribution. (c) Equilibrium price at the end of each outer iteration.

Figure 1.5: GMF-V-Q versus GMF-P-TRPO ($\sigma = 1.3$ and one trajectory).



(a) Equilibrium inventory distribution. (b) Equilibrium production distribution. (c) Equilibrium price at the end of each outer iteration.

Figure 1.6: GMF-V-Q versus GMF-P-TRPO ($\sigma = 2.0$ and one trajectory).

Note that the demand elasticity σ captures how sensitive demand for a product is compared to the changes in other economic factors, such as price or income. When σ is increased from 1.3 to 2.0 indicating that the demand is more sensitive to price rise, the equilibrium price decreases from 10.9 to 7.3 (see Figures 1.5c and 1.6c) and the distribution of the equilibrium production level is centered towards smaller values (see Figures 1.5b and 1.6b). The equilibrium inventory level has a huge mass at 0 when $\sigma = 1.3$. This implies that producers do not keep large inventories and pay the inventory cost in the equilibrium. On the other hand, the equilibrium inventory is more uniformly distributed when $\sigma = 2$.

1.8.3 Performance evaluation in the N-player setting

Performance metric. Similar to the performance metric introduced in Section 1.8.2 for the GMFG setting, we adopt the following metric to measure the difference between a given policy π and an NE under THE N-player setting (here $\epsilon_0 > 0$ is a safeguard, and is taken as 0.1 in the experiments):

$$C(\boldsymbol{\pi}) = \frac{1}{N|\mathcal{S}|^N} \sum_{i=1}^{N} \sum_{\boldsymbol{s} \in \mathcal{S}^N} \frac{\max_{\pi^i} V^i(\boldsymbol{s}, (\boldsymbol{\pi}^{-i}, \pi^i)) - V^i(\boldsymbol{s}, \boldsymbol{\pi})}{|\max_{\pi^i} V^i(\boldsymbol{s}, (\boldsymbol{\pi}^{-i}, \pi^i))| + \epsilon_0}.$$

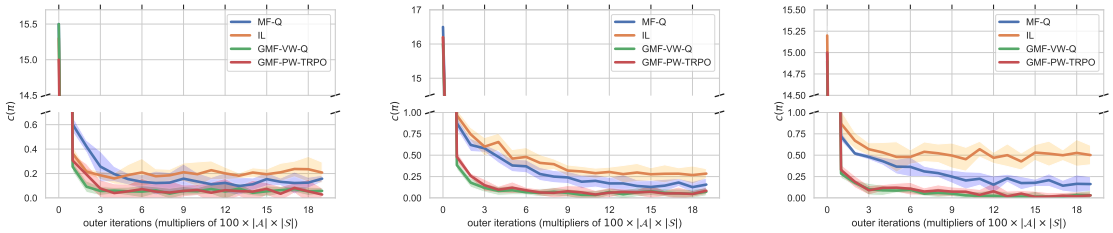
Clearly $C(\boldsymbol{\pi}) \geq 0$, and $C(\boldsymbol{\pi}^\star) = 0$ if and only if $\boldsymbol{\pi}^\star$ is an NE. Policy $\arg\max_{\pi_i} V_i(\boldsymbol{s}, (\boldsymbol{\pi}^{-i}, \pi_i))$ is called the best response to $\boldsymbol{\pi}^{-i}$. A similar metric without normalization has been adopted in [128].

Existing algorithms for N-player games. To test the effectiveness of GMF-VW-Q for approximating N-player games, we next compare GMF-VW-Q with the IL algorithm and the MF-Q algorithm. The IL algorithm [150] considers N independent players and each player solves a decentralized reinforcement learning problem ignoring other players in the system. The MF-Q algorithm [167] extends the NASH-Q Learning algorithm for the N-player game introduced in [84], adds the aggregate actions ($\bar{\boldsymbol{a}}_{-i} = \frac{\sum_{j \neq i} a_j}{N-1}$) from the opponents, and works for the class of games where the interactions are only through the average actions of N players.

Results and analysis. Our experiment (Figure 1.7) shows that GMF-VW-Q and GMF-PW-TRPO achieve similar performance, and both of them are superior in terms of convergence rate, accuracy, and stability for approximating an N-player game. In general, both algorithms converge faster than IL and MF-Q and achieve the smallest errors.

For instance, when $N = 20$, IL Algorithm converges with the largest error 0.220. The error from MF-Q is 0.101, smaller than IL but still bigger than the error from GMF-VW-Q. The GMF-VW-Q and GMF-PW-TRPO converge with the lowest error 0.065. Moreover, as N increases, the error of GMF-VW-Q and GMF-PW-TRPO decease while the errors of both MF-Q and IL increase significantly. As $|\mathcal{S}|$ and $|\mathcal{A}|$ increase, GMF-VW-Q and GMF-PW-TRPO are robust with respect to this increase of dimensionality, while both MF-Q and

IL clearly suffer from the increase of the dimensionality with decreased convergence rate and accuracy. Therefore, GMF-VW-Q and GMF-PW-TRPO are more scalable than IL and MF-Q, when the system is complex and the number of players N is large.



(a) $|\mathcal{S}| = 10, |\mathcal{A}| = 100, N = 20$. (b) $|\mathcal{S}| = 15, |\mathcal{A}| = 225, N = 20$. (c) $|\mathcal{S}| = 10, |\mathcal{A}| = 100, N = 60$.

Figure 1.7: Learning accuracy based on $C(\boldsymbol{\pi})$.

1.9 Extension: Existence and uniqueness for non-stationary NE of GMFGs

In this section, we describe the setting of non-stationary NE for GMFGs and establish the corresponding results of existence and uniqueness.

Definition 1.9.1 (NE for GMFGs)**.** In (GMFG), a player-population profile $(\boldsymbol{\pi}^\star, \boldsymbol{\mathcal{L}}^\star) := (\{\pi_t^\star\}_{t=0}^\infty, \{\mathcal{L}_t^\star\}_{t=0}^\infty)$ is called an NE if

1. (Single player side) Fix $\boldsymbol{\mathcal{L}}^\star$, for any policy sequence $\boldsymbol{\pi} := \{\pi_t\}_{t=0}^\infty$ and initial state $s \in \mathcal{S}$,
$$V\left(s, \boldsymbol{\pi}^\star, \boldsymbol{\mathcal{L}}^\star\right) \geq V\left(s, \boldsymbol{\pi}, \boldsymbol{\mathcal{L}}^\star\right). \tag{1.9.1}$$
2. (Population side) $\mathbb{P}_{s_t, a_t} = \mathcal{L}_t^\star$ for all $t \geq 0$, where $\{s_t, a_t\}_{t=0}^\infty$ is the dynamics under the policy sequence $\boldsymbol{\pi}^\star$ starting from $s_0 \sim \mu_0^\star$, with $a_t \sim \pi_t^\star(s_t, \mu_t^\star)$, $s_{t+1} \sim P(\cdot|s_t, a_t, \mathcal{L}_t^\star)$, and $\mu_t^\star$ being the population state marginal of $\mathcal{L}_t^\star$.

Step A. Fix $\boldsymbol{\mathcal{L}} := \{\mathcal{L}_t\}_{t=0}^\infty$, (GMFG) becomes the classical optimization problem. Indeed, with $\boldsymbol{\mathcal{L}}$ fixed, the population state distribution sequence $\boldsymbol{\mu} := \{\mu_t\}_{t=0}^\infty$ is also fixed, hence the space of admissible policies is reduced to the single-player case. Solving (GMFG) is now reduced to finding a policy sequence $\pi_{t,\boldsymbol{\mathcal{L}}}^\star \in \Pi := \{\pi \,|\, \pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})\}$ over all admissible $\boldsymbol{\pi}_{\boldsymbol{\mathcal{L}}} = \{\pi_{t,\boldsymbol{\mathcal{L}}}\}_{t=0}^\infty$, to maximize

$$\begin{aligned} V(s, \boldsymbol{\pi}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{\mathcal{L}}) := \quad & \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t, \mathcal{L}_t) | s_0 = s\right], \\ \text{subject to} \quad & s_{t+1} \sim P(s_t, a_t, \mathcal{L}_t), \quad a_t \sim \pi_{t,\boldsymbol{\mathcal{L}}}(s_t). \end{aligned}$$

Notice that with $\boldsymbol{\mathcal{L}}$ fixed, one can safely suppress the dependency on μ_t in the admissible policies. Moreover, given this fixed $\boldsymbol{\mathcal{L}}$ sequence and the solution $\boldsymbol{\pi}^\star_{\boldsymbol{\mathcal{L}}} := \{\pi^\star_{t,\boldsymbol{\mathcal{L}}}\}_{t=0}^\infty$, one can define a mapping from the fixed population distribution sequence $\boldsymbol{\mathcal{L}}$ to an optimal randomized policy sequence. That is,

$$\Gamma_1 : \{\mathcal{P}(\mathcal{S}\times\mathcal{A})\}_{t=0}^\infty \to \{\Pi\}_{t=0}^\infty,$$

such that $\boldsymbol{\pi}^\star_{\boldsymbol{\mathcal{L}}} = \Gamma_1(\boldsymbol{\mathcal{L}})$. Note that this $\boldsymbol{\pi}^\star_{\boldsymbol{\mathcal{L}}}$ sequence satisfies the single player side condition in Definition 1.9.1 for the population state-action pair sequence $\boldsymbol{\mathcal{L}}$. That is, $V\left(s,\boldsymbol{\pi}^\star_{\boldsymbol{\mathcal{L}}},\boldsymbol{\mathcal{L}}\right) \geq V\left(s,\boldsymbol{\pi},\boldsymbol{\mathcal{L}}\right)$, for any policy sequence $\boldsymbol{\pi} = \{\pi_t\}_{t=0}^\infty$ and any initial state $s \in \mathcal{S}$.

Accordingly, a similar feedback regularity condition is needed in this step.

Assumption 3. *There exists a constant $d_1 \geq 0$, such that for any $\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{L}}' \in \{\mathcal{P}(\mathcal{S}\times\mathcal{A})\}_{t=0}^\infty$,*

$$D(\Gamma_1(\boldsymbol{\mathcal{L}}),\Gamma_1(\boldsymbol{\mathcal{L}}')) \leq d_1\mathcal{W}_1(\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{L}}'), \tag{1.9.2}$$

where

$$\begin{aligned} D(\boldsymbol{\pi},\boldsymbol{\pi}') &:= \sup_{s\in\mathcal{S}} \mathcal{W}_1(\boldsymbol{\pi}(s),\boldsymbol{\pi}'(s)) = \sup_{s\in\mathcal{S}}\sup_{t\in\mathbb{N}} W_1(\pi_t(s),\pi'_t(s)),\\ \mathcal{W}_1(\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{L}}') &:= \sup_{t\in\mathbb{N}} W_1(\mathcal{L}_t,\mathcal{L}'_t), \end{aligned} \tag{1.9.3}$$

and W_1 is the ℓ_1-Wasserstein distance between probability measures.

Step B. Based on the analysis in Step A and $\boldsymbol{\pi}^\star_{\boldsymbol{\mathcal{L}}} = \{\pi^\star_{t,\boldsymbol{\mathcal{L}}}\}_{t=0}^\infty$, update the initial sequence $\boldsymbol{\mathcal{L}}$ to $\boldsymbol{\mathcal{L}}'$ following the controlled dynamics $P(\cdot|s_t,a_t,\mathcal{L}_t)$.

Accordingly, for any admissible policy sequence $\boldsymbol{\pi} \in \{\Pi\}_{t=0}^\infty$ and a joint population state-action pair sequence $\boldsymbol{\mathcal{L}} \in \{\mathcal{P}(\mathcal{S}\times\mathcal{A})\}_{t=0}^\infty$, define a mapping $\Gamma_2 : \{\Pi\}_{t=0}^\infty \times \{\mathcal{P}(\mathcal{S}\times\mathcal{A})\}_{t=0}^\infty \to \{\mathcal{P}(\mathcal{S}\times\mathcal{A})\}_{t=0}^\infty$ as follows:

$$\Gamma_2(\boldsymbol{\pi},\boldsymbol{\mathcal{L}}) := \hat{\boldsymbol{\mathcal{L}}} = \{\mathbb{P}_{s_t,a_t}\}_{t=0}^\infty, \tag{1.9.4}$$

where $s_{t+1} \sim \mu_t P(\cdot|\cdot,a_t,\mathcal{L}_t)$, $a_t \sim \pi_t(s_t)$, $s_0 \sim \mu_0$, and μ_t is the population state marginal of $\mathcal{L}_t$.

One also needs a similar assumption in this step.

Assumption 4. *There exist constants $d_2,\ d_3 \geq 0$, such that for any admissible policy sequences $\boldsymbol{\pi},\boldsymbol{\pi}^1,\boldsymbol{\pi}^2$ and joint distribution sequences $\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{L}}^1,\boldsymbol{\mathcal{L}}^2$,*

$$\mathcal{W}_1(\Gamma_2(\boldsymbol{\pi}^1,\boldsymbol{\mathcal{L}}),\Gamma_2(\boldsymbol{\pi}^2,\boldsymbol{\mathcal{L}})) \leq d_2 D(\boldsymbol{\pi}^1,\boldsymbol{\pi}^2), \tag{1.9.5}$$

$$\mathcal{W}_1(\Gamma_2(\boldsymbol{\pi},\boldsymbol{\mathcal{L}}^1),\Gamma_2(\boldsymbol{\pi},\boldsymbol{\mathcal{L}}^2)) \leq d_3 \mathcal{W}_1(\boldsymbol{\mathcal{L}}^1,\boldsymbol{\mathcal{L}}^2). \tag{1.9.6}$$

Similarly, Assumption 4 can be reduced to Lipschitz continuity and boundedness of the transition dynamics P under certain conditions.

Step C. Repeat Step A and Step B until $\boldsymbol{\mathcal{L}}'$ matches $\boldsymbol{\mathcal{L}}$.

This step is to take care of the population side condition. To ensure the convergence of the combined step A and step B, it suffices if $\Gamma : \{\mathcal{P}(\mathcal{S} \times \mathcal{A})\}_{t=0}^{\infty} \to \{\mathcal{P}(\mathcal{S} \times \mathcal{A})\}_{t=0}^{\infty}$ is a contractive mapping under the $\mathcal{W}_1$ distance, with $\Gamma(\boldsymbol{\mathcal{L}}) := \Gamma_2(\Gamma_1(\boldsymbol{\mathcal{L}}), \boldsymbol{\mathcal{L}})$. Then by the Banach fixed point theorem and the completeness of the related metric spaces, there exists a unique NE to the GMFG.

In summary, we have

Theorem 1.9.1 (Existence and Uniqueness of GMFG solution)**.** *Given Assumptions 3 and 4, and assuming that $d_1 d_2 + d_3 < 1$, there exists a unique NE to* (GMFG)*.*

The proof of Theorem 1.9.1 can be established by modifying appropriately the fixed-point approach for the stationary GMFG in Theorem 1.3.1.

1.10 Appendix

1.10.1 Distance metrics and completeness

This section reviews some basic properties of the Wasserstein distance. It then proves that the metrics defined in the main text are indeed distance functions and define complete metric spaces.

ℓ_1-Wasserstein distance and dual representation. The ℓ_1 Wasserstein distance over $\mathcal{P}(\mathcal{X})$ for $\mathcal{X} \subseteq \mathbb{R}^k$ is defined as

$$W_1(\nu, \nu') := \inf_{M \in \mathcal{M}(\nu, \nu')} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|_2 \mathrm{d}M(x, y). \tag{1.10.1}$$

where $\mathcal{M}(\nu, \nu')$ is the set of all measures (couplings) on $\mathcal{X} \times \mathcal{X}$, with marginals ν and ν' on the two components, respectively.

The Kantorovich duality theorem enables the following equivalent dual representation of W_1:

$$W_1(\nu, \nu') = \sup_{\|f\|_L \leq 1} \left| \int_{\mathcal{X}} f d\nu - \int_{\mathcal{X}} f d\nu' \right|, \tag{1.10.2}$$

where the supremum is taken over all 1-Lipschitz functions f, *i.e.*, f satisfying $|f(x) - f(y)| \leq \|x - y\|_2$ for all $x, y \in \mathcal{X}$.

The Wasserstein distance W_1 can also be related to the total variation distance via the following inequalities [67]:

$$d_{\min}(\mathcal{X}) d_{TV}(\nu, \nu') \leq W_1(\nu, \nu') \leq \mathrm{diam}(\mathcal{X}) d_{TV}(\nu, \nu'), \tag{1.10.3}$$

where $d_{\min}(\mathcal{X}) = \min_{x \neq y \in \mathcal{X}} \|x - y\|_2$, which is guaranteed to be positive when $\mathcal{X}$ is finite.

When $\mathcal{S}$ and $\mathcal{A}$ are compact, for any compact subset $\mathcal{X} \subseteq \mathbb{R}^k$, and for any $\nu, \nu' \in \mathcal{P}(\mathcal{X})$, $W_1(\nu, \nu') \leq \text{diam}(\mathcal{X}) d_{TV}(\nu, \nu') \leq \text{diam}(\mathcal{X}) < \infty$, where $\text{diam}(\mathcal{X}) = \sup_{x,y \in \mathcal{X}} \|x - y\|_2$ and d_{TV} is the total variation distance. Moreover, one can verify

Lemma 1.10.1. *Both D and $\mathcal{W}_1$ are distance functions, and they are finite for any input distribution pairs. In addition, both $(\{\Pi\}_{t=0}^{\infty}, D)$ and $(\{\mathcal{P}(\mathcal{S} \times \mathcal{A})\}_{t=0}^{\infty}, \mathcal{W}_1)$ are complete metric spaces.*

These facts enable the usage of Banach fixed-point mapping theorem for the proof of existence and uniqueness (Theorems 1.9.1 and 1.3.1).

Proof. [Proof of Lemma 1.10.1] It is known that for any compact set $\mathcal{X} \subseteq \mathbb{R}^k$, $(\mathcal{P}(\mathcal{X}), W_1)$ defines a complete metric space [25]. Since $W_1(\nu, \nu') \leq \text{diam}(\mathcal{X})$ is uniformly bounded for any $\nu,\ \nu' \in \mathcal{P}(\mathcal{X})$, we know that $\mathcal{W}_1(\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{L}}') \leq \text{diam}(\mathcal{X})$ and $D(\boldsymbol{\pi}, \boldsymbol{\pi}') \leq \text{diam}(\mathcal{X})$ as well, so they are both finite for any input distribution pairs. It is clear that they are distance functions based on the fact that W_1 is a distance function.

Finally, we show the completeness of the two metric spaces $(\{\Pi\}_{t=0}^{\infty}, D)$ and $(\{\mathcal{P}(\mathcal{S} \times \mathcal{A})\}_{t=0}^{\infty}, \mathcal{W}_1)$. Take $(\{\Pi\}_{t=0}^{\infty}, D)$ for example. Suppose that $\boldsymbol{\pi}^k$ is a Cauchy sequence in $(\{\Pi\}_{t=0}^{\infty}, D)$. Then for any $\epsilon > 0$, there exists a positive integer N, such that for any $m,\ n \geq N$,

$$D(\boldsymbol{\pi}^n, \boldsymbol{\pi}^m) \leq \epsilon \Longrightarrow W_1(\pi_t^n(s), \pi_t^m(s)) \leq \epsilon \text{ for any } s \in \mathcal{S},\ t \in \mathbb{N}, \tag{1.10.4}$$

which implies that $\pi_t^k(s)$ forms a Cauchy sequence in $(\mathcal{P}(\mathcal{A}), W_1)$, and hence by the completeness of $(\mathcal{P}(\mathcal{A}), W_1)$, $\pi_t^k(s)$ converges to some $\pi_t(s) \in \mathcal{P}(\mathcal{A})$. As a result, $\boldsymbol{\pi}^n \to \boldsymbol{\pi} \in \{\Pi\}_{t=0}^{\infty}$ under metric D, which shows that $(\{\Pi\}_{t=0}^{\infty}, D)$ is complete.

The completeness of $(\{\mathcal{P}(\mathcal{S} \times \mathcal{A})\}_{t=0}^{\infty}, \mathcal{W}_1)$ can be proved similarly. □

The same argument for Lemma 1.10.1 shows that both D and W_1 are distance functions and are finite for any input distribution pairs, with both (Π, D) and $(\mathcal{P}(\mathcal{S} \times \mathcal{A}), W_1)$ again complete metric spaces.

1.10.2 Bounds for GMF-V-Q using asynchronous Q-learning

In the main text, we have shown the results by using synchronous Q-learning algorithm. Here for the completeness, we also show the corresponding results for asynchronous Q-learning algorithm.

For asynchronous Q-learning algorithm, at each step l with the state s and an action a, the system reaches state s' according to the controlled dynamics and the Q-function approximation Q_l is updated according to

$$\hat{Q}^{l+1}(s, a) = (1 - \beta_l(s, a))\hat{Q}^l(s, a) + \beta_l(s, a) \left[r(s, a) + \gamma \max_{\bar{a}} \hat{Q}^l(s', \bar{a}) \right], \tag{1.10.5}$$

where $\hat{Q}^0(s,a) = C$ for some constant $C \in \mathbb{R}$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, and the step size $\beta_l(s,a)$ can be chosen as ([53])

$$\beta_l(s,a) = \begin{cases} |\#(s,a,l)+1|^{-h}, & (s,a) = (s_l, a_l), \\ 0, & \text{otherwise.} \end{cases} \tag{1.10.6}$$

with $h \in (1/2, 1)$. Here $\#(s,a,l)$ is the number of times up to time l that one visits the state-action pair (s,a). The algorithm then proceeds to choose action a' based on $\hat{Q}^{l+1}$ with appropriate exploration strategies, including the ϵ-greedy strategy.

Lemma 1.10.2 ([53]: sample complexity of asynchronous Q-learning)**.** *For an MDP, say $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, suppose that the Q-learning algorithm takes step-sizes (1.5.4). Also suppose that the covering time of the state-action pairs is bounded by L with probability at least $1-p$ for some $p \in (0,1)$. Then $\|\hat{Q}^{T_{\mathcal{M}}(\delta,\epsilon)} - Q^\star_{\mathcal{M}}\|_\infty \le \epsilon$ with probability at least $1-2\delta$. Here $\hat{Q}^T$ is the T-th update in the Q-learning updates (1.5.3), $Q^\star_{\mathcal{M}}$ is the (optimal) Q-function, and*

$$T_{\mathcal{M}}(\delta,\epsilon) = \Omega\left(\left(\frac{L\log_p(\delta)}{\beta}\log\frac{V_{\max}}{\epsilon}\right)^{\frac{1}{1-h}} + \left(\frac{\left(L\log_p(\delta)\right)^{1+3h} V_{\max}^2 \log\left(\frac{|\mathcal{S}||\mathcal{A}|V_{\max}}{\delta\beta\epsilon}\right)}{\beta^2\epsilon^2}\right)^{\frac{1}{h}}\right),$$

where $\beta = (1-\gamma)/2$, $V_{\max} = R_{\max}/(1-\gamma)$, and $R_{\max}$ is such that a.s. $0 \le r(s,a) \le R_{\max}$.

Here the covering time L of a state-action pair sequence is defined to be the number of steps needed to visit all state-action pairs starting from any arbitrary state-action pair. Also notice that the l_∞ norm above is defined in an element-wise sense, *i.e.*, for $M \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, we have $\|M\|_\infty = \max_{s\in\mathcal{S}, a\in\mathcal{A}} |M(s,a)|$.

Corollary 1.10.3 (Value-based guarantee of asynchronous Q-learning algorithm)**.** *The asynchronous Q-learning algorithm with appropriate choices of step-sizes (cf. (1.5.4)) satisfies the following value-based guarantee, where $C^{(i)}_{\mathcal{M}} (i = 1,2,3)$ are constants depending on $|\mathcal{S}|, |\mathcal{A}|, V_{\max}, \beta$ and h, and we have:*

$$\alpha_2^{(1)} = \alpha_4^{(1)} = \frac{1}{1-h},\ \alpha_1^{(1)} = \alpha_3^{(2)} = 0,$$
$$\alpha_1^{(2)} = \frac{2}{h},\ \alpha_4^{(2)} = \frac{2+3h}{h},\ \alpha_j^{(2)} = 0 \text{ for } j = 2,3,$$
$$\alpha_1^{(3)} = \frac{2}{h},\ \alpha_2^{(3)} = \frac{1}{h},\ \alpha_4^{(3)} = \frac{1+3h}{h},\ \alpha_3^{(3)} = 0.$$

In addition, assume the same assumptions as Theorem 1.3.1, then for Algorithm 3 with asynchronous Q-learning method, with probability at least $1-2\delta$, $W_1(\mathcal{L}_{K_{\epsilon,\eta}}, \mathcal{L}^\star) \le C_0\epsilon$, where $K_{\epsilon,\eta}$

is defined as in Theorem 1.5.1. And the total number of samples $T = \sum_{k=0}^{K_{\epsilon,\eta}-1} T_{\mathcal{M}_{\mathcal{L}_k}}(\delta_k, \epsilon_k)$ *is bounded by*

$$T \leq O\left(K_{\epsilon,\eta}^{\frac{4}{h}+1}\left(\log\frac{K_{\epsilon,\eta}}{\delta}\right)^{3+\frac{2}{h}} + \left(\log\frac{K_{\epsilon,\eta}}{\delta}\right)^{\frac{2}{1-h}}\right).$$

1.10.3 Weak simulator

In this section, we state the counterpart of Theorems 1.5.1 and 1.5.5 for Algorithms 5 and 6, respectively. Notice that here the major difference is the additional $O(1/\sqrt{N})$ term.

We first (re)state the relation between $\mathbf{Emp}_N$ (which serves as a $1/N$-net) and action gaps:

For any positive integer N*, there exist a positive constant* $\phi_N > 0$*, with the property that* $\max_{a' \in \mathcal{A}} Q^\star_{\mathcal{L}}(s, a') - Q^\star_{\mathcal{L}}(s, a) \geq \phi_N$ *for any* $\mathcal{L} \in \boldsymbol{Emp}_N$*,* $s \in \mathcal{S}$*, and any* $a \notin argmax_{a \in \mathcal{A}} Q^\star_{\mathcal{L}}(s, a)$*.*

Now we are ready to state the convergence results.

Theorem 1.10.4 (Convergence and complexity of GMF-VW)**.** *Assume the same assumptions as Theorem 1.3.1. Suppose that Alg has a value-based guarantee with parameters*

$$\{C_{\mathcal{M}}^{(i)}, \alpha_1^{(i)}, \alpha_2^{(i)}, \alpha_3^{(i)}, \alpha_4^{(i)}\}_{i=1}^m.$$

For any $\epsilon,\ \delta > 0$*, set* $\delta_k = \delta/K_{\epsilon,\eta}$*,* $\epsilon_k = (k+1)^{-(1+\eta)}$ *for some* $\eta \in (0, 1]$ $(k = 0, \ldots, K_{\epsilon,\eta} - 1)$*, and* $c' = c = \frac{\log(1/\epsilon)}{\phi_N}$*.*[6] *Then with probability at least* $1 - 4\delta$*,*

$$W_1(\mathcal{L}_{K_{\epsilon,\eta}}, \mathcal{L}^\star) \leq C\epsilon + \frac{diam(\mathcal{S})diam(\mathcal{A})|\mathcal{S}||\mathcal{A}|}{2(1-d)}\sqrt{\frac{1}{2N}\log(|\mathcal{S}||\mathcal{A}|K_{\epsilon,\eta}/\delta)}.$$

Here $K_{\epsilon,\eta} := \left\lceil 2\max\left\{(\eta\epsilon/c)^{-1/\eta}, \log_d(\epsilon/\max\{diam(\mathcal{S})diam(\mathcal{A}), c\}) + 1\right\}\right\rceil$ *is the number of outer iterations, and the constant* C *is independent of* δ*,* ϵ *and* η*.*

Moreover, the total number of samples $T = \sum_{k=0}^{K_{\epsilon,\eta}-1} T_{\mathcal{M}_{\mathcal{L}_k}}(\delta_k, \epsilon_k)$ *is bounded by*

$$T \leq \sum_{i=1}^m \frac{2^{\alpha_2^{(i)}}}{2\alpha_1^{(i)} + 1} C_{\mathcal{M}}^{(i)} K_{\epsilon,\eta}^{2\alpha_1^{(i)}+1} (K_{\epsilon,\eta}/\delta)^{\alpha_3^{(i)}} \left(\log(K_{\epsilon,\eta}/\delta)\right)^{\alpha_2^{(i)}+\alpha_4^{(i)}}. \tag{1.10.7}$$

Theorem 1.10.5 (Convergence and complexity of GMF-PW)**.** *Assume the same assumptions as in Theorem 1.3.1. Suppose that Alg has a policy-based guarantee with parameters*

$$\{C_{\mathcal{M}}^{(i)}, \alpha_1^{(i)}, \alpha_2^{(i)}, \alpha_3^{(i)}, \alpha_4^{(i)}\}_{i=1}^m.$$

[6] Here we actually only need $c' = \Omega(\frac{\log(1/\epsilon)}{\phi_N})$ and $c = O(\frac{\log(1/\epsilon)}{\phi_N})$, and the corresponding result will differ only in some absolute constants.

Then for any $\epsilon,\ \delta > 0$, *set* $\delta_k = \delta/K_{\epsilon,\eta}$, $\epsilon_k = (k+1)^{-(1+\eta)}$ *for some* $\eta \in (0,1]$ $(k = 0,\dots,K_{\epsilon,\eta}-1)$, *and* $c' = c = \frac{\log(1/\epsilon)}{\phi_N}$,[7] *with probability at least* $1-4\delta$,

$$W_1(\mathcal{L}_{K_{\epsilon,\eta}},\mathcal{L}^\star) \le C\epsilon + \frac{diam(\mathcal{S})\,diam(\mathcal{A})|\mathcal{S}||\mathcal{A}|}{2(1-d)}\sqrt{\frac{1}{2N}\log(|\mathcal{S}||\mathcal{A}|K_{\epsilon,\eta}/\delta)}.$$

Here $K_{\epsilon,\eta} := \left\lceil 2\max\left\{(\eta\epsilon/c)^{-1/\eta}, \log_d(\epsilon/\max\{diam(\mathcal{S})\,diam(\mathcal{A}), c\}) + 1\right\}\right\rceil$ *is the number of outer iterations, and the constant* C *is independent of* δ, ϵ *and* η.

Moreover, the total number of samples $T = \sum_{k=0}^{K_{\epsilon,\eta}-1} T_{\mathcal{M}_{\mathcal{L}_k}}(\delta_k,\epsilon_k)$ *is bounded by*

$$T \le \sum_{i=1}^{m+1} \frac{2^{\alpha_2^{(i)}}}{2\alpha_1^{(i)}+1}\tilde{C}_{\mathcal{M}}^{(i)} K_{\epsilon,\eta}^{2\alpha_1^{(i)}+1}(K_{\epsilon,\eta}/\delta)^{\alpha_3^{(i)}}\left(\log(K_{\epsilon,\eta}/\delta)\right)^{\alpha_2^{(i)}+\alpha_4^{(i)}}, \tag{1.10.8}$$

where the parameters $\{\tilde{C}_{\mathcal{M}}^{(i)}, \alpha_1^{(i)}, \alpha_2^{(i)}, \alpha_3^{(i)}, \alpha_4^{(i)}\}_{i=1}^{m+1}$ *are defined in Lemma 1.5.4.*

The key to the proof of Theorems 1.10.4 and 1.10.5 is the following lemma, which follows from the Hoeffding inequality.

Lemma 1.10.6. *The expectation*

$$\mathbb{E}\left[\mathcal{L}_{k+1}(s',a')\right] = \pi_k(s',a')\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mu_k(s)P(s'|s,a,\mathcal{L}_k)\pi_k(s,a) = \Gamma_2(\pi_k,\mathcal{L}_k).$$

In addition, we have that for any $t > 0$, $s' \in \mathcal{S}$ *and* $a' \in \mathcal{A}$,

$$\mathbb{P}\left(|\mathcal{L}_{k+1}(s',a') - \mathbb{E}\left[\mathcal{L}_{k+1}(s',a')\right]| \ge t\right) \le 2\exp\left(-2Nt^2\right). \tag{1.10.9}$$

The above lemma essentially states that the iterates $\mathcal{L}_{k+1}$ of Algorithms 5 and 6 are very close to the "$\tilde{\mathcal{L}}_{k+1}$" obtained from the (strong) simulator $\mathcal{G}(s,\pi_k,\mathcal{L}_k)$ with $s \sim \mu_k$ following line 5 in Algorithm 2 and line 6 in Algorithm 4. This bridges the gap between the weak and the strong simulators. In particular, by noticing that

$$\begin{aligned} W_1(\mathcal{L}_{k+1},\mathcal{L}^\star) &\le W_1(\mathcal{L}_{k+1},\Gamma_2(\pi_k,\mathcal{L}_k)) + W_1(\Gamma_2(\pi_k,\mathcal{L}_k),\Gamma_2(\Gamma_1(\mathcal{L}^\star),\mathcal{L}^\star)) \\ &= W_1(\mathcal{L}_{k+1},\mathbb{E}\left[\mathcal{L}_{k+1}\right]) + W_1(\Gamma_2(\pi_k,\mathcal{L}_k),\Gamma_2(\Gamma_1(\mathcal{L}^\star),\mathcal{L}^\star)) \\ &\le \frac{\mathrm{diam}(\mathcal{S})\mathrm{diam}(\mathcal{A})|\mathcal{S}||\mathcal{A}|}{2}\left\|\mathcal{L}_{k+1} - \mathbb{E}\left[\mathcal{L}_{k+1}\right]\right\|_\infty + W_1(\Gamma_2(\pi_k,\mathcal{L}_k),\Gamma_2(\Gamma_1(\mathcal{L}^\star),\mathcal{L}^\star)), \end{aligned}$$

one can bound the first term with high probability via (1.10.9). The second term is then bounded in exactly the same way as the proofs for Theorems 1.5.1 and 1.5.5, and hence we omit the details.

[7] Here again we actually only need $c' = \Omega(\frac{\log(1/\epsilon)}{\phi_N})$ and $c = O(\frac{\log(1/\epsilon)}{\phi_N})$, and the corresponding result will differ only in some absolute constants.

1.10.4 Adaption of MF-Q

Note that MF-Q is designed for global states and coupled local actions, while in our equilibrium price example we have coupled local (private) states and decoupled local actions. To suit this setting, we adapt MF-Q by replacing the mean-field action term with the mean-field state term. In addition, our comparison is within the tabular setting, which is the setting theoretically analyzed in [167]. In this case, the Q functions in MF-Q are also functions of only states and actions but not the mean-field terms, which are iteratively changing together with the Q-functions and claimed to be converging to a unique point. This is not the case for the function approximation setting (where Q-functions in MF-Q are indeed dependent on the mean-field terms and function approximation enables learning and generalizing such dependencies via experience replay). We leave it a future work to extend and compare our algorithms with MF-Q in the deep approximation setting.

Chapter 2

Logarithmic Regret for Episodic Continuous-Time Linear-Quadratic Reinforcement Learning over a Finite-Time Horizon

2.1 Introduction

Reinforcement learning (RL) for linear quadratic (LQ) control problems has been one of the most active areas for both the control and the reinforcement learning communities. Over the last few decades, significant progresses have been made in the discrete-time setting.

2.1.1 Discrete-time RL

In the area of adaptive control with unknown dynamics parameters, the goal is to find optimal stationary policy that stabilizes the unknown dynamics and minimizes the long term average cost ([88, 103]). For an infinite-time horizon LQ system, it has been shown that persistent excitation conditions [72] are critical to the parameter identification. Meanwhile, algorithms with asymptotic convergence in both the parameter estimation and the optimal control have been developed in [69], [99] and [30]: the first one assumes that costs only depend on state variables and the other two consider both state and control costs and use a cost-biased least-squared estimation method. See [55, 57] and references therein for recent developments of (randomised) adaptive control algorithms for LQ systems.

Following the seminal works of [14, 13] and [124], *non-asymptotic* regret bound analysis for RL algorithms has been one of the main topics, and has been developed for tabular Markov decision problems.

The non-asymptotic analysis of adaptive LQ problem by [1] utilizes the Optimism in the Face of Uncertainty principle to construct a sequence of improving confidence regions for

the unknown model parameters, and solves a non-convex constrained optimization problem for each confidence region; their algorithm achieves an $\mathcal{O}(\sqrt{T})$ regret bound, with T being the number of time steps. To reduce the computational complexity and to avoid the non-convexity issue, [2] and [125] propose Thompson-sampling-based algorithms and derive $\mathcal{O}(\sqrt{T})$ regret bounds in the Bayesian setting; [45] proposes a robust adaptive control algorithm to solve a convex sub-problem in each step and achieves an $\mathcal{O}(T^{2/3})$ regret bound. The gap between these regret bounds is removed by [112] and [40] via two different approaches for the same $\mathcal{O}(\sqrt{T})$ frequentist regret bound. Later, [142] establishes a lower bound on the regret of order $\mathcal{O}(\sqrt{d_u^2 d_x T})$, where d_u and d_x are the dimensions of the actions and the states, and shows that a simple variant of certainty equivalent control matches the lower bound in both T and the dimensions. Similar regret bounds have also been established under different settings and assumptions, such as [35] in the adversarial setting and [101] without a stabilizing controller at the early stages of agent-environment interaction.

All the analyses are in discrete-time with an infinite time horizon. In all these problems, adaptive control algorithms are shown to achieve logarithmic regret bounds when additional information regarding the parameters of the system (often referred to as identifiability conditions) is available. Indeed, [56, 58] prove that certainty equivalent adaptive regulator achieves logarithmic regret bounds if the system parameter satisfies certain sparsity or low-rankness conditions. [34] establishes logarithmic regret bounds when either the state transition matrix is unknown, or when the state-action transition matrix is unknown and the optimal policy is non-degenerate. In partially observable linear dynamical systems, which takes linear-quadratic Gaussian problem as a special case, [102] proposes an algorithm with a logarithmic regret bound, under the assumption that one has access to a set in which all controllers persistently excite the system to approximate the optimal control. Logarithmic regret bounds in the adversarial setting with known dynamics parameters have been established in [5, 59].

2.1.2 Continuous-time RL.

Most real-world control systems, such as those in aerospace, automotive industry and robotics, are naturally continuous-time dynamical systems.

So are their related physical tasks, such as inverted pendulum problems, cart-pole balancing problems, and legged robot problems. Continuous-time finite-time horizon LQ control problems can be found in portfolio optimization [160], algorithmic trading [33], production management of exhaustible resources [71], and biological movement systems [16].

Analysis for continuous-time LQ-RL and general RL problems, however, is fairly limited. The primary approach is to develop learning algorithms after discretizing both the time and the space spaces, and establish the convergence as discretization parameters tend to zero. For instance, [120] proposes a policy gradient algorithm and shows the convergence of the policy gradient estimate to the true gradient. [121, 119] design learning algorithms by discretizing Bellman equations of the underlying control problems and prove the asymptotic convergence of their algorithms. For the LQ system, attentions have been mostly on algorithms designs,

including the integral reinforcement learning algorithm in [118], and the policy iteration algorithm in [136]. Yet, very little is known regarding the convergence rate or the regret bound of all these algorithms. Indeed, despite the natural analytical connection between LQ control and RL, the best known theoretical work for continuous-time LQ-RL is still due to [50], where an asymptotically sublinear regret for an ergodic model has been derived via a weighted least-squares-based estimation approach. Nevertheless, the exact order of the regret bound has not been studied.

Issues and challenges from non-asymptotic analysis. It is insufficient and improper to rely solely on the analysis and algorithms for the discrete-time RL to solve the continuous-time problems. There is a mismatch between the algorithms timescale for the former and the underlying systems timescale for the latter. When designing algorithms that make observations and take actions at discrete time points, it is important to take the model mismatch into consideration. For instance, the empirical studies in [149] suggest that vanilla Q-learning methods exhibit degraded performance as the time stepsize decreases, while a proper scaling of learning rates with stepsize leads to more robust performance.

The questions are therefore: A) How to quantify the precise impacts of the observation stepsize and action stepsize on algorithm performance? B) How to derive non-asymptotic regret analysis for learning algorithms in continuous-time LQ-RL (or general RL) system, analogous to the discrete-time LQ-RL counterpart?

There are technical reasons behind the limited theoretical progress in the continuous-time domain for RL, including LQ-RL. In addition to the known difficulty for analyzing stochastic control problems, the learning component compounds the problem complexity and poses new challenges.

For instance, the counterpart in the continuous-time problem to the algebraic equations in [112] for the discrete-time version is the regularity and stability of the continuous-time Riccati equation and the regularity of feedback controls. While Riccati equation and its robustness and existence and uniqueness of optimal controls have been well studied in the control literature, regularity of feedback controls with respect to underlying models is completely new for control theory and crucial for algorithm design and its robustness analysis. Moreover, deriving the *exact* order of the regret bound requires developing new and different techniques than those used for the *asymptotic* regret analysis in [50].

Our work and contributions. This paper studies finite-time horizon continuous-time LQ-RL problems in an episodic setting.

- It first proposes a greedy least-squares algorithm based on continuous-time observations and controls. At each iteration, the algorithm estimates the unknown parameters by a regularized least-squares estimator based on observed trajectories, then designs linear feedback controls via the Riccati differential equation for the estimated model. It identifies conditions under which the unknown state transition matrix and state-action transition matrix are uniquely identifiable under the optimal policies. (Remark 2.2.1 and Proposition

2.2.1). By exploiting the identifiability of coefficients, this continuous-time least-squares algorithm is shown to have a logarithmic regret of the magnitude $\mathcal{O}((\ln M)(\ln\ln M))$, with M being the number of learning episodes (Theorem 2.2.2). To the best of our knowledge, this is the first non-asymptotic logarithmic regret bound for continuous-time LQ-RL problems with unknown state and control coefficients.

- It then proposes a practically implementable least-squares algorithm based on discrete-time observations and controls. At each iteration, the algorithm estimates the unknown parameters by observing continuous-time trajectories at discrete time points, then designs a piecewise constant linear feedback control via Riccati difference equations for an associated discrete-time LQ-RL problem. It shows that the regret of the discrete-time least-squares algorithm is of the magnitude $\mathcal{O}\big((\ln M)(\ln\ln M) + \sum_{\ell=0}^{\ln M} 2^\ell \tau_\ell^2\big)$, where τ_ℓ is the time stepsize used in the $(\ell+1)$-th update of model parameters (Theorem 2.2.3).

  Our analysis shows that scaling the regularization parameter of the discrete-time least-squares estimator with respect to time stepsize is critical for a robust performance of the algorithm in different timescales (Remark 2.2.3). To the best of our knowledge, this is the first discrete-time algorithm with rigorous regret bound for continuous-time LQ-RL problems.

Different from the least-squares algorithms for the ergodic LQ problems (see e.g., [50, 112]), our continuous-time least-squares algorithm constructs feedback controls via Riccati differential equations instead of the algebraic equations in [112]. Here, the regularity and stability of the continuous-time Riccati equation is analyzed in order to establish the robustness of feedback controls.

Moreover, our analysis for the estimation error exploits extensively the sub-exponential tail behavior of the least-squares estimators. This probabilistic approach differs from the asymptotic sublinear regret analysis in [50]; it establishes the exact order of the logarithmic regret bound by the concentration inequality for the error bound.

In addition, our analysis also exploits an important self-exploration property of finite-time horizon continuous-time LQ-RL problems, for which the time-dependent optimal feedback matrices ensure that the optimal state and control processes span the entire parameter space. This property allows us to design exploration-free learning algorithms with logarithmic regret bounds. Furthermore, we provide explicit conditions on models that guarantees the successful identification of the unknown parameters with optimal feedback policies. This is in contrast to the identification conditions for logarithmic regret bounds in discrete-time infinite-time-horizon LQ problems. Our conditions apply to arbitrary finite time-horizon problems, without imposing sparsity or low-rankness conditions on system parameters as in [56, 58] or requiring these parameters to be partially known to the controller as in [34, 59].

Finally, our analysis provides the precise parameter estimation error in terms of the sample size and time stepsize, and quantifies the performance gap between applying a piecewise-constant policy from an incorrect model and applying the optimal policy. The misspecification error scales linearly with respect to the stepsize, and the performance gap depends

quadratically with respect to the time stepsize and the magnitude of parameter perturbations. Our analysis is based on the first-order convergence of Riccati difference equations and a uniform sub-exponential tail bound of discrete-time least-squares estimators.

Notation. For each $n \in \mathbb{N}$, we denote by $I = I_n$ the $n \times n$ identity matrix, and by $\mathbb{S}_0^n$ (resp. $\mathbb{S}_+^n$) the space of symmetric positive semidefinite (resp. definite) matrices. We denote by $|\cdot|$ the Euclidean norm of a given Euclidean space, by $\|\cdot\|_2$ the matrix norm induced by Euclidean norms, and by $A^\top$ and $\mathrm{tr}(A)$ the transpose and trace of a matrix A, respectively. For each $T > 0$, filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} = \{\mathcal{F}_t\}_{t\in[0,T]}, \mathbb{P})$ satisfying the usual condition and Euclidean space $(E, |\cdot|)$, we introduce the following spaces:

- $C([0,T];E)$ is the space of continuous functions $\phi : [0,T] \to E$ satisfying $\|\phi\|_{C([0,T];E)} = \sup_{t\in[0,T]} |\phi_t| < \infty$;
- $C^1([0,T];E)$ is the space of continuously differentiable functions $\phi : [0,T] \to E$ satisfying $\|\phi\|_{C^1([0,T];E)} = \sup_{t\in[0,T]}(|\phi_t| + |\phi'_t|) < \infty$;
- $\mathcal{S}^2(E)$ is the space of E-valued $\mathbb{F}$-progressively measurable càdlàg processes $X : \Omega \times [0,T] \to E$ satisfying $\|X\|_{\mathcal{S}^2(E)} = \mathbb{E}[\sup_{t\in[0,T]} |X_t|^2]^{1/2} < \infty$;
- $\mathcal{H}^2(E)$ is the space of E-valued $\mathbb{F}$-progressively measurable processes $X : \Omega \times [0,T] \to E$ satisfying $\|X\|_{\mathcal{H}^2(E)} = \mathbb{E}[\int_0^T |X_t|^2 \,\mathrm{d}t]^{1/2} < \infty$.

For notation simplicity, we denote by $C \in [0,\infty)$ a generic constant, which depends only on the constants appearing in the assumptions and may take a different value at each occurrence.

2.2 Problem formulation and main results

2.2.1 Linear-quadratic reinforcement learning problem

In this section, we consider the linear-quadratic reinforcement learning (LQ-RL) problem, where the drift coefficient of the state dynamics is unknown to the controller.

More precisely, let $T \in (0,\infty)$ be a given terminal time, W be an n-dimensional standard Brownian motion defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$ be the filtration generated by W augmented by the $\mathbb{P}$-null sets. Let $x_0 \in \mathbb{R}^n$ be a given initial state and $(A^\star, B^\star) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times d}$ be fixed but unknown matrices, consider the following problem:

$$\inf_{U\in\mathcal{H}^2(\mathbb{R}^d)} J^{\theta^\star}(U), \quad \text{with} \quad J^{\theta^\star}(U) = \mathbb{E}\left[\int_0^T \left((X_t^{\theta^\star,U})^\top Q X_t^{\theta^\star,U} + (U_t)^\top R U_t\right) \mathrm{d}t\right], \tag{2.2.1}$$

where for each $U \in \mathcal{H}^2(\mathbb{R}^d)$, the process $X^{\theta^\star,U} \in \mathcal{S}^2(\mathbb{R}^n)$ satisfies the following controlled dynamics associated with the parameter $\theta^\star = (A^\star, B^\star)^\top$:

$$\mathrm{d}X_t = (A^\star X_t + B^\star U_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad t \in [0,T]; \quad X_0 = x_0, \tag{2.2.2}$$

with given matrices $Q \in \mathbb{S}_0^n$ and $R \in \mathbb{S}_+^d$. Note that we assume the loss functional (2.2.1) only involves a time homogeneous running cost to allow a direct comparison with infinite-time horizon RL problems (see e.g., [51]), but similar analysis can be performed if the cost

functions are time inhomogeneous, a terminal cost is included, or the Brownian motion W in (2.2.2) is scaled by an known nonsingular diffusion matrix.

If the parameter $\theta^\star = (A^\star, B^\star)^\top$ are known to the controller, then (2.2.1)-(2.2.2) reduces to the classical LQ control problems. In this case, it is well known that (see e.g., [169] and the references therein), the optimal control $U^{\theta^\star}$ of (2.2.1)-(2.2.2) is given in a feedback form by

$$U_t^{\theta^\star} = \psi^{\theta^\star}(t, X_t^{\theta^\star}), \quad \text{with } \psi^{\theta^\star}(t,x) = K_t^{\theta^\star} x, \ \forall (t,x) \in [0,T] \times \mathbb{R}^n, \tag{2.2.3}$$

where $K_t^{\theta^\star} = -R^{-1}(B^\star)^\top P_t^{\theta^\star}$ for all $t \in [0,T]$, $(P_t^{\theta^\star})_{t\in[0,T]}$ solves the Riccati equation

$$\tfrac{\mathrm{d}}{\mathrm{d}t} P_t + (A^\star)^\top P_t + P_t A^\star - P_t (B^\star R^{-1} (B^\star)^\top) P_t + Q = 0, \quad t \in [0,T]; \quad P_T = 0, \tag{2.2.4}$$

and $X^{\theta^\star}$ is the state process governed by the following dynamics:

$$\mathrm{d}X_t = (A^\star X_t + B^\star K_t^{\theta^\star} X_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad t \in [0,T]; \quad X_0 = x_0. \tag{2.2.5}$$

To solve the LQ-RL problem (2.2.1)-(2.2.2) with unknown $\theta^\star$, the controller searches for the optimal control while simultaneously learning the system, i.e., the matrices $A^\star, B^\star$. In an episodic (also known as reset or restart) learning framework, the controller improves her knowledge of the underlying dynamics X_t through successive learning episodes, in order to find a control that is close to the optimal one.

Mathematically, it goes as follows. Let $M \in \mathbb{N}$ be the total number of learning episodes. In the i-th learning episode, $i = 1, \ldots, M$, a feedback control ψ^i is exercised, and the state process X^{ψ^i} evolves according to the dynamics (2.2.2) controlled by the policy ψ^i:

$$\mathrm{d}X_t = (A^\star X_t + B^\star \psi^i(t, X_t))\mathrm{d}t + \mathrm{d}W_t^i, \quad t \in [0,T]; \quad X_0 = x_0. \tag{2.2.6}$$

Here $W^i, i = 1, 2, \ldots, M$ are independent n-dimensional Brownian motions defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The (expected) cost of learning in the i-th episode is then given by

$$J^{\theta^\star}(U^{\psi^i}) = \mathbb{E}\left[\int_0^T \left((X_t^{\psi^i})^\top Q X_t^{\psi^i} + (U_t^{\psi^i})^\top R U_t^{\psi^i}\right) \mathrm{d}t\right], \text{ with } U_t^{\psi^i} := \psi^i(t, X_t^{\psi^i}),\ t \in [0,T], \tag{2.2.7}$$

and the (expected) regret of learning up to $M \in \mathbb{N}$ episodes (with the sequence of controls $(U^{\psi^i})_{i=1}^M$) is defined as follows:

$$R(M) = \sum_{i=1}^M \Big(J^{\theta^\star}(U^{\psi^i}) - J^{\theta^\star}(U^{\theta^\star})\Big), \tag{2.2.8}$$

where $J^{\theta^\star}(U^{\theta^\star})$ is the optimal cost of (2.2.1)-(2.2.2) when $A^\star, B^\star$ are known. Intuitively, the regret characterizes the cumulative loss from taking sub-optimal policies in all episodes.

In the following, we shall propose several least-squares-based learning algorithms to solve (2.2.1)-(2.2.2), and prove that they achieve logarithmic regrets if $\theta^\star$ is identifiable (see Remark 2.2.1 for details).

2.2.2 Continuous-time least-squares algorithm and its regret bound

In this section, we consider a continuous-time least-squares algorithm, which chooses the optimal feedback control based on the current estimation of the parameter, and updates the parameter estimation based on the whole trajectories of the state dynamics.

More precisely, let $\theta = (A, B)^\top \in \mathbb{R}^{(n+d)\times n}$ be the current estimate of the unknown parameter $\theta^\star$, then the controller would exercise the optimal feedback control ψ^θ for (2.2.1)-(2.2.2) with $\theta^\star$ replaced by θ, i.e.,

$$\psi^\theta(t,x) = K_t^\theta x, \quad K_t^\theta := -R^{-1}B^\top P_t^\theta, \quad \forall (t,x) \in [0,T]\times\mathbb{R}^n, \tag{2.2.9}$$

where P^θ satisfies the Riccati equation (3.5.1) with $\theta^\star$ replaced by θ:

$$\tfrac{\mathrm{d}}{\mathrm{d}t}P_t + A^\top P_t + P_t A - P_t(BR^{-1}B^\top)P_t + Q = 0, \quad t\in[0,T]; \quad P_T = 0. \tag{2.2.10}$$

This leads to the state process X^{ψ^θ} satisfying (cf. (2.2.6)):

$$\mathrm{d}X_t = (A^\star X_t + B^\star \psi^\theta(t,X_t))\,\mathrm{d}t + \mathrm{d}W_t, \quad t\in[0,T]; \quad X_0 = x_0. \tag{2.2.11}$$

We proceed to derive an ℓ_2-regularized least-squares estimation for $\theta^\star$ based on sampled trajectories of X^{ψ^θ}. Observing from (2.2.11) that

$$Z_t^{\psi^\theta}(\mathrm{d}X_t^{\psi^\theta})^\top = Z_t^{\psi^\theta}(Z_t^{\psi^\theta})^\top\theta^\star\mathrm{d}t + Z_t^{\psi^\theta}(\mathrm{d}W_t)^\top, \quad \text{with } Z_t^{\psi^\theta} = \begin{pmatrix} X^{\psi^\theta} \\ \psi^\theta(t,X_t^{\psi^\theta}) \end{pmatrix} \text{ for all } t\in[0,T].$$

Hence the martingale property of the Itô integral implies that

$$\theta^\star = \left(\mathbb{E}\left[\int_0^T Z_t^{\psi^\theta}(Z_t^{\psi^\theta})^\top\,\mathrm{d}t\right]\right)^{-1}\mathbb{E}\left[\int_0^T Z_t^{\psi^\theta}(\mathrm{d}X_t^{\psi^\theta})^\top\right], \tag{2.2.12}$$

provided that $\mathbb{E}\big[\int_0^T Z_t^{\psi^\theta}(Z_t^{\psi^\theta})^\top\,\mathrm{d}t\big]$ is invertible. This suggests a practical rule to improve one's estimate θ for the true parameter $\theta^\star$, by replacing the expectations in (2.2.12) with empirical averages over independent realizations. More precisely, let $m\in\mathbb{N}$ and

$$(X_t^{\psi^\theta,i}, \psi^\theta(t,X_t^{\psi^\theta,i}))_{t\in[0,T]}, \quad i = 1,\dots,m,$$

be trajectories of m independent realizations of the state and control processes, we shall update the estimate θ by the following rule, inspired by (2.2.12):

$$\theta \longleftarrow \left(\frac{1}{m}\sum_{i=1}^m\int_0^T Z_t^{\psi^\theta,i}(Z_t^{\psi^\theta,i})^\intercal\,\mathrm{d}t + \frac{1}{m}I\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^m\int_0^T Z_t^{\psi^\theta,i}(\mathrm{d}X_t^{\psi^\theta,i})^\intercal\right), \tag{2.2.13}$$

where $Z_t^{\psi^\theta,i} := \begin{pmatrix} X_t^{\psi^\theta,i} \\ \psi^\theta(t,X_t^{\psi^\theta,i}) \end{pmatrix}$ for all $t\in[0,T]$ and $i=1,\dots,m$, and I is the $(n+d)\times(n+d)$ identity matrix.

The regularization term $\frac{1}{m}I$ in (2.2.13) guarantees the required matrix inverse and vanishes as $m \to \infty$. The estimator (2.2.13) can be equivalently expressed as an ℓ_2-regularized least-squares estimator, as pointed out in [51] for the ergodic LQ-RL problem.

We summarize the continuous-time least-squares algorithm as follows.

Algorithm 8 Continuous-time least-squares algorithm

1: **Input**: Choose an initial estimation θ_0 of $\theta^\star$ and numbers of learning episodes $\{m_\ell\}_{\ell\in\mathbb{N}\cup\{0\}}$.
2: **for** $\ell = 0, 1, \cdots$ **do**
3: Obtain the feedback control ψ^{θ_ℓ} as (2.2.9) with $\theta = \theta_\ell$.
4: Execute the feedback control ψ^{θ_ℓ} for m_ℓ independent episodes, and collect the trajectory data $(X_t^{\psi^{\theta_\ell},i}, \psi^{\theta_\ell}(t, X_t^{\psi^{\theta_\ell},i}))_{t\in[0,T]}$, $i = 1, \dots, m_\ell$.
5: Obtain an updated estimation $\theta_{\ell+1}$ by using (2.2.13) and the m_ℓ trajectories collected above.
6: **end for**

Note that Algorithm 8 operates in cycles, with m_ℓ the number of episodes in the ℓ-th cycle. Hence, the regret of learning up to M episodes (cf. (2.2.8)) can be upper bounded by the accumulated regret at the end of the L-th cycle, where L is the smallest integer such that $\sum_{\ell=0}^{L} m_\ell \geq M$.

In this section, we analyze the regret of Algorithm 8 based on the following assumptions of the learning problem (2.2.1)-(2.2.2).

H.1. *(1)* $T \in (0,\infty)$, $n, d \in \mathbb{N}$, $x_0 \in \mathbb{R}^n$, $A^\star \in \mathbb{R}^{n\times n}$, $B^\star \in \mathbb{R}^{n\times d}$, $Q \in \mathbb{S}_0^n$ *and* $R \in \mathbb{S}_+^d$.

(2) $\{v \in \mathbb{R}^d \mid (K_t^{\theta^\star})^\top v = 0,\ \forall t \in [0,T]\} = \{0\}$, *with* $K^{\theta^\star}$ *defined in* (2.2.3).

Before discussing the regret of Algorithm 8, we make the following remark of (H.1).

Remark 2.2.1 (**Self-exploration of finite-time horizon RL problems**). *(H.1(1)) is the standard assumption for finite-time horizon LQ-RL problems (see e.g., [80]), except that H.1(1) allows Q to be positive semidefinite, which is important for costs depending on partial states. (H.15) corresponds to the identifiability of the true parameter $\theta^\star$ by executing the optimal policy $K^{\theta^\star}$. In fact, as shown in Proposition 2.3.10, under (H.1(1)), (H.15) is equivalent to the following statement:*

(2') if $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^d$ satisfy $u^\top X^{\theta^\star} + v^\top U^{\theta^\star} = 0$ for $\mathrm{d}\mathbb{P} \otimes \mathrm{d}t$-almost everywhere in $\Omega \times [0,T]$, then $u = 0$ and $v = 0$, where $X^{\theta^\star}$ and $U^{\theta^\star}$ are the optimal state and control processes of (2.2.1)-(2.2.2) defined by (2.2.5) and (2.2.3), respectively,

Item (2') indicates an important self-exploration property of finite-time horizon continuous-time RL problems. In particular, the time-dependent optimal feedback matrix $K^{\theta^\star}$ and the non-degenerate noises guarantee the non-degeneracy of the space spanned by $X^{\theta^\star}$ and $U^{\theta^\star}$,

enabling learning the parameters sufficiently well. This self-exploration property is critical for our design of exploration-free learning algorithms for (2.2.1)-(2.2.2) with a logarithmic regret (see Theorems 2.2.2 and 2.2.3).

One can easily show that (H.15) holds if the optimal policy $(K^{\theta^\star})_{t\in[0,T]}$ is nondegenerate, i.e., $\sup_{t\in[0,T]} \lambda_{\min}\left((K_t^{\theta^\star})(K_t^{\theta^\star})^\top\right) > 0$. Similar nondegeneracy condition has been imposed in [34] for discrete-time ergodic LQ-RL problems. In particular, by assuming that the optimal stationary policy satisfies $\lambda_{\min}(K^\star(K^\star)^\top) > 0$ (along with other controllablity conditions), they propose learning algorithms with a logarithmic regret, under the assumption that only the control coefficient $B^\star$ is unknown. In contrast, we allow both the state coefficient $A^\star$ and the control coefficient $B^\star$ to be unknown.

Moreover, the following proposition provides sufficient conditions of (H.15), which are special cases of Proposition 2.3.11.

Proposition 2.2.1. *Let $n, d \in \mathbb{N}$, $Q \in \mathbb{S}_0^n$ and $R \in \mathbb{S}_+^d$.*

(1) If $(B^\star)^\top Q B^\star \in \mathbb{S}_+^d$, then (H.15) holds for all $T > 0$.

(2) Assume that the algebraic Riccati equation $(A^\star)^\top P + PA^\star - P(B^\star R^{-1}(B^\star)^\top)P + Q = 0$ admits a unique maximal solution $P_\infty^\star \in \mathbb{S}_+^n$. Let $K_\infty^\star = -R^{-1}(B^\star)^\top P_\infty^\star$, and for each $T > 0$, let $P^{\star,(T)} \in C([0,T];\mathbb{S}_0^n)$ be defined in (3.5.1). Assume that $\lim_{T\to\infty} P_0^{\star,(T)} = P_\infty^\star$ and $K_\infty^\star(K_\infty^\star)^\top \in \mathbb{S}_+^d$. Then there exists $T_0 > 0$, such that (H.15) holds for all $T \geq T_0$.

Proposition 2.2.1 provides two sets of conditions for (H.15) under two different scenarios: Item (1) applies to an arbitrary finite $T > 0$, and Item (2) only applies to sufficiently large T. Item (2) assumes the asymptotic behavior of solutions to Riccati differential equations, which can be ensured by the stabilizability of the pair $(A^\star, B^\star)$ and detectability of the pair $(A^\star, Q^{1/2})$ (see [24, Theorems 10.9 and 10.10]). Note that our subsequent analysis is based on (H.1), and does not require stabilizability assumptions.

Remark 2.2.2 (Stabilizability of $(A^\star, B^\star)$ and dependence on T). *Since the LQ-RL problem (2.2.1)-(2.2.2) is over the time horizon $[0,T]$ with a fixed $T < \infty$, in general one does not need additional conditions on $(A^\star, B^\star)$ for the well-definedness of (2.2.1)-(2.2.2). If $T = \infty$, then some stabilizability/controllability conditions of $(A^\star, B^\star)$ may be required for (2.2.1)-(2.2.2) to ensure a well-defined solution (see e.g., [46]). Under these conditions, different algorithms have been shown to achieve sub-linear regret with respect to the number of decision steps (see e.g., [112, 40]), and even logarithmic regrets provided that further identifiability assumptions are satisfied (see e.g., [56, 58, 34, 102]); see Section 2.1.1 for more details. For $T < \infty$, the regrets of learning algorithms for (2.2.1)-(2.2.2) in general depend exponentially on the time horizon T (e.g., the constants C_0, C' in Theorem 2.2.2), as the moments of the optimal state process $X^{\theta^\star}$ and control process $U^{\theta^\star}$ may grow exponentially with respect to T. It would be interesting to quantify the precise dependence of the regret bounds on T. This would entail deriving precise a priori bounds of solutions to (2.2.10) and*

estimating the norm $\|(\mathbb{E}[\int_0^T Z_t^{\theta^\star}(Z_t^{\theta^\star})^\top \mathrm{d}t])^{-1}\|_2$ *in terms of* $(A^\star, B^\star, Q, T)$*, and is left for future research.*

We are now ready to state the main result of this section, which shows that the regret of Algorithm 8 grows logarithmically with respect to the number of episodes.

Theorem 2.2.2. *Suppose (H.1) holds and let* $\theta_0 = (A_0, B_0)^\top \in \mathbb{R}^{(n+d)\times d}$ *such that (H.15) holds with* θ_0*. Then there exists a constant* $C_0 > 0$ *such that for all* $C \geq C_0$*, and* $\delta \in (0, \frac{3}{\pi^2})$*, if one sets* $m_0 = C(-\ln\delta)$ *and* $m_\ell = 2^\ell m_0$ *for all* $\ell \in \mathbb{N}$*, then with probability at least* $1 - \frac{\pi^2\delta}{3}$*, the regret of Algorithm 8 given by (2.2.8) satisfies*

$$R(M) \leq C'\big((\ln M)(\ln\ln M) + (-\ln\delta)(\ln M)\big), \quad \forall M \in \mathbb{N},$$

where C' *is a constant independent of* M *and* δ*.*

To simplify the presentation, we analyze the performance of Algorithm 8 by assuming the number of learning episodes $\{m_\ell\}_\ell$ is doubled between two successive updates of the estimation of $\theta^\star$. Similar regret results can be established for Algorithm 8 with different choices of $\{m_\ell\}_\ell$. Under this specific choice of $\{m_\ell\}_\ell$, for any $M \in \mathbb{N}$, Algorithm 8 splits M episodes into $L = \lceil \log_2(\frac{M}{m_0}+1)\rceil - 1$ cycles, where the ℓ-th cycle, $\ell = 0, 1, \ldots, L-1$, contains m_ℓ episodes, and the remaining $M - \sum_{\ell=0}^{L-1} m_\ell$ episodes are in the last cycle.

Sketched proof of Theorem 2.2.2. We outline the key steps of the proof of Theorem 2.2.2, and present the detailed arguments to Section 2.3.3. By exploiting the regularity and robustness of solutions to (2.2.10), we prove that the performance gap $J^{\theta^\star}(U^{\psi^\theta}) - J^{\theta^\star}(U^{\theta^\star})$ is of the magnitude $\mathcal{O}(|\theta - \theta^\star|^2)$, for all *a-priori* bounded θ (Proposition 2.3.8). We then establish a uniform sub-exponential property for the (deterministic and stochastic) integrals in (2.2.12), which along with (H.15) and Bernstein's inequality leads to the following estimate of the parameter estimation error: for all $\delta \in (0, 1/2)$, all sufficiently large $m \in \mathbb{N}$, and all θ sufficiently close to $\theta^\star$,

$$|\hat\theta - \theta^\star| \leq \mathcal{O}\Big(\sqrt{\frac{-\ln\delta}{m}} + \frac{-\ln\delta}{m} + \frac{(-\ln\delta)^2}{m^2}\Big), \quad \text{with probability } 1 - 2\delta, \tag{2.2.14}$$

where $\hat\theta$ is generated by (2.2.13) with ψ^θ (Proposition 3.3.12). Then for each $\delta > 0$, applying (2.2.14) with $\delta_\ell = \delta/(\ell+1)^2$ for all $\ell \in \mathbb{N}\cup\{0\}$ shows that with probability $1 - 2\sum_{\ell=0}^\infty \delta_\ell = 1 - \frac{\pi^2\delta}{3}$,

$$|\hat\theta_{\ell+1} - \theta^\star|^2 \lesssim \frac{-\ln\delta_\ell}{m_\ell} + \frac{(-\ln\delta_\ell)^2}{m_\ell^2} + \frac{(-\ln\delta_\ell)^4}{m_\ell^4}, \quad \forall \ell \in \mathbb{N}, \tag{2.2.15}$$

where $\lesssim$ means the inequality holds with a multiplicative constant independent of δ and ℓ. By the quadratic performance gap and the choice of $\{m_\ell\}_\ell$, the regret of Algorithm

8 up to the M-th episode can be bounded by the regret at the end of L-th cycle with $L = \lceil \log_2(\frac{M}{m_0} + 1) \rceil - 1$:

$$R(M) \lesssim \sum_{\ell=0}^{L} m_\ell |\theta_\ell - \theta^\star|^2 \lesssim \sum_{\ell=0}^{L} (-\ln \delta_\ell) \left(1 + \frac{-\ln \delta_\ell}{m_\ell} + \frac{(-\ln \delta_\ell)^3}{m_\ell^3} \right). \tag{2.2.16}$$

Observe that the choices of $\{\delta_\ell\}_\ell$ and $\{m_\ell\}_\ell$ ensure that $\sup_{\delta \in (0, \frac{3}{\pi^2}), \ell \in \mathbb{N}} \frac{-\ln \delta_\ell}{m_\ell} < \infty$. Hence, the right-hand side of (2.2.16) is of the magnitude $\mathcal{O}\left(\sum_{\ell=0}^{L} (-\ln \delta_\ell) \right)$, which along with the choices of δ_ℓ and L leads to the desired regret bound; see the end of Section 2.3.3 for more details.

2.2.3 Discrete-time least-squares algorithm and its regret bound

Note that Algorithm 8 in Section 2.2.2 requires executing feedback controls and observing corresponding state trajectories continuously. A common practice to solve continuous-time RL problems is by assuming that at each learning episode the dynamics only evolves in discrete time, and then estimate parameters according to discrete-time RL algorithms (see e.g., [121, 119, 120, 149]). As the true dynamics evolves continuously, it is necessary to quantify the impact of reaction stepsize on the algorithm performance.

In this section, we analyze the performance of the above procedure for solving (2.2.1)-(2.2.2). We adapt regularized least-squares algorithms for discrete-time LQ problems to the present setting, and establish their regret bounds in terms of the discretization stepsize. Our analysis shows that a proper scaling of the regularization term in the least-squares estimation in terms of stepsize is critical for a robust performance with respect to different timescales.

More precisely, for a given cycle (i.e., the index ℓ in Algorithm 8), let $\theta = (A, B)^\top \in \mathbb{R}^{(n+d)\times n}$ be the current estimate of $\theta^\star$ in (2.2.2), and let $\{t_i\}_{i=0}^N$, $N \in \mathbb{N}$, be a uniform partition of $[0, T]$ with stepsize $\tau = T/N$. We then assume that (2.2.1)-(2.2.2) is piecewise constant between any two grid points $\{t_i\}_{i=0}^N$, choose actions and make observations every τ, and update the estimated parameter based on these observations. To this end, we consider the following discrete-time LQ control problem with parameter θ:

$$\inf_{U \in \mathcal{H}_N^2(\mathbb{R}^d)} J_N(U), \quad \text{with } J_N(U) = \mathbb{E}\left[\sum_{i=0}^{N-1} \left((X_{t_i}^{U,\tau})^\top Q X_{t_i}^{U,\tau} + U_{t_i}^\top R U_{t_i} \right) \tau \right], \tag{2.2.17}$$

where $\mathcal{H}_N^2(\mathbb{R}^d) = \{U \in \mathcal{H}^2(\mathbb{R}^d) \mid U_t = U_{t_i}, t \in [t_i, t_{i+1}), i = 0, \dots, N-1\}$, and $(X_{t_i}^{U,\tau})_{i=0}^{N-1}$ are defined by

$$X_{t_{i+1}}^{U,\tau} - X_{t_i}^{U,\tau} = (A X_{t_i}^{U,\tau} + B U_{t_i})\tau + W_{t_{i+1}} - W_{t_i}, \quad i = 0, \dots N-1; \quad X_0^{U,\tau} = x_0. \tag{2.2.18}$$

Note that for simplicity, our strategy is constructed by assuming a discrete-time dynamics arising from an Euler discretization of (2.2.2) (with the estimated parameter θ); similar analysis can be performed with a high-order approximation of (2.2.1)-(2.2.2).

It is well-known that (see e.g., [24]), the optimal control of (2.2.17)-(2.2.18) is given by the following feedback form:

$$U_t = \psi^{\theta,\tau}(t, X_t^{U,\tau}), \quad \text{with } \psi^{\theta,\tau}(t,x) = K_t^{\theta,\tau} x, \quad \forall (t,x) \in [0,T) \times \mathbb{R}^n, \tag{2.2.19}$$

where $K^{\theta,\tau} : [0,T) \to \mathbb{R}^{d\times n}$ is the piecewise constant function (with stepsize $\tau = T/N$) defined by

$$\begin{aligned} P_{t_{i-1}}^{\theta,\tau} &= \tau Q + (I+\tau A)^\top P_{t_i}^{\theta,\tau}(I+\tau A) - (I+\tau A)^\top P_{t_i}^{\theta,\tau}\tau B(R+\tau B^\top P_{t_i}^{\theta,\tau} B)^{-1} B^\top P_{t_i}^{\theta,\tau}(I+\tau A), \\ &\qquad \forall i = 0,\dots,N-1; \quad P_T^{\theta,\tau} = 0, \\ K_t^{\theta,\tau} &= -(R+\tau B^\top P_{t_{i+1}}^{\theta,\tau} B)^{-1} B^\top P_{t_{i+1}}^{\theta,\tau}(I+\tau A), \quad t \in [t_i, t_{i+1}),\ i = 0,\dots,N-1. \end{aligned} \tag{2.2.20}$$

We then implement the piecewise constant strategy $\psi^{\theta,\tau}$ defined in (2.2.19) on the original system (2.2.2) for m episodes, and update the estimated parameter θ by observing (2.2.2) with stepsize $\tau = T/N$. More precisely, let $X^{\psi^{\theta,\tau}} \in \mathcal{S}^2(\mathbb{R}^n)$ be the state process associated with $\psi^{\theta,\tau}$:

$$\mathrm{d}X_t = (A^\star X_t + B^\star K_{t_i}^{\theta,\tau} X_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad t \in [t_i, t_{i+1}],\ i = 0,\dots,N-1; \quad X_0 = x_0, \tag{2.2.21}$$

and $(X_t^{\psi^{\theta,\tau},j})_{t\in[0,T]}$, $j = 1,\dots,m$, $m \in \mathbb{N}$, be m independent trajectories of $X^{\psi^{\theta,\tau}} \in \mathcal{S}^2(\mathbb{R}^n)$, we update the parameter θ according to the following discrete-time least-squares estimator:

$$\theta \leftarrow \operatorname*{arg\,min}_{\theta\in\mathbb{R}^{(n+d)\times n}} \sum_{j=1}^{m}\sum_{i=0}^{N-1} \|X_{t_{i+1}}^{\psi^{\theta,\tau},j} - X_{t_i}^{\psi^{\theta,\tau},j} - \tau\theta^\top Z_{t_i}^{\psi^{\theta,\tau},j}\|_2^2 + \tau \mathrm{tr}(\theta^\top\theta), \tag{2.2.22}$$

with $Z_{t_i}^{\psi^{\theta,\tau},j} := \begin{pmatrix} X_{t_i}^{\psi^{\theta,\tau},j} \\ K_{t_i}^{\theta,\tau} X_{t_i}^{\psi^{\theta,\tau},j} \end{pmatrix}$ for all i,j. The update (2.2.22) is consistent with the agent's assumption that the state evolves according to (2.2.18) between two grid points. Setting the derivative (with respect to θ) of the right-hand side of (2.2.22) to zero leads to

$$-\sum_{j=1}^{m}\sum_{i=0}^{N-1} \tau Z_{t_i}^{\psi^{\theta,\tau},j}\left(\left(X_{t_{i+1}}^{\psi^{\theta,\tau},j} - X_{t_i}^{\psi^{\theta,\tau},j}\right)^\top - \tau (Z_{t_i}^{\psi^{\theta,\tau},j})^\top\theta\right) + \tau\theta = 0.$$

Dividing both sides by τ/m and rearranging the terms give the following equivalent expression of the discrete-time least squares estimator (2.2.22):

$$\theta \longleftarrow \left(\frac{1}{m}\sum_{j=1}^{m}\sum_{i=0}^{N-1} Z_{t_i}^{\psi^{\theta,\tau},j}(Z_{t_i}^{\psi^{\theta,\tau},j})^\top\tau + \frac{1}{m}I\right)^{-1}\left(\frac{1}{m}\sum_{j=1}^{m}\sum_{i=0}^{N-1} Z_{t_i}^{\psi^{\theta,\tau},j}\left(X_{t_{i+1}}^{\psi^{\theta,\tau},j} - X_{t_i}^{\psi^{\theta,\tau},j}\right)^\top\right). \tag{2.2.23}$$

Remark 2.2.3 (**Scaling hyper-parameters with timescales**). *In principle, when applying discrete-time RL algorithms in a continuous environment, it is critical to adopt a proper scaling of the hyper-parameters for a robust performance with respect to different timescales. Indeed, scaling the regularization term* $\mathrm{tr}(\theta^\top\theta)$ *in* (2.2.22) *with respect to the stepsize* τ *is essential for the robustness of* (2.2.23) *for all small stepsizes* τ*. If one updates* θ *by minimizing the following* ℓ_2*-regularized loss function with a given hyper-parameter* $\alpha < 1$ *such that*

$$\arg\min_{\theta\in\mathbb{R}^{(n+d)\times n}} \sum_{j=1}^{m}\sum_{i=0}^{N-1} \|X_{t_{i+1}}^{\psi^{\theta,\tau},j} - X_{t_i}^{\psi^{\theta,\tau},j} - \tau\theta^\top Z_{t_i}^{\psi^{\theta,\tau},j}\|_2^2 + \tau^\alpha \mathrm{tr}(\theta^\top\theta), \tag{2.2.24}$$

then the corresponding discrete-time estimator is given by

$$\theta^\tau := \left(\frac{1}{m}\sum_{j=1}^{m}\sum_{i=0}^{N-1} Z_{t_i}^{\psi^{\theta,\tau},j}(Z_{t_i}^{\psi^{\theta,\tau},j})^\top\tau + \frac{1}{\tau^{1-\alpha}m}I\right)^{-1}\left(\frac{1}{m}\sum_{j=1}^{m}\sum_{i=0}^{N-1} Z_{t_i}^{\psi^{\theta,\tau},j}\left(X_{t_{i+1}}^{\psi^{\theta,\tau},j} - X_{t_i}^{\psi^{\theta,\tau},j}\right)^\top\right).$$

Observe that for any given $m \in \mathbb{N}$*, the estimator* θ^τ *degenerates to zero as the stepsize* τ *tends to zero. Hence, to ensure the viability of* θ^τ *across different timescales, the number of episodes* m *has to increase appropriately when* τ *tends to zero. In contrast, by choosing* $\alpha = 1$ *in* (2.2.24), (2.2.23) *admits a continuous-time limit* (2.2.13) *as* $\tau \to 0$*, and leads to a learning algorithm in which the episode numbers and the time stepsize can be chosen independently (see Theorem 2.2.3).*

We now summarize the discrete-time least-squares algorithm as follows.

Algorithm 9 Discrete-time least-squares algorithm

1: **Input**: Choose an initial estimation θ_0 of $\theta^\star$, numbers of learning episodes $\{m_\ell\}_{\ell\in\mathbb{N}\cup\{0\}}$ and numbers of intervention points $\{N_\ell\}_{\ell\in\mathbb{N}\cup\{0\}}$.
2: **for** $\ell = 0, 1, \cdots$ **do**
3: Obtain the piecewise constant control $\psi^{\theta_\ell,\tau_\ell}$ as (2.2.19) with $\tau = T/N_\ell$ and $\theta = \theta_\ell$.
4: Execute the control $\psi^{\theta_\ell,\tau_\ell}$ for m_ℓ independent episodes, and collect the data $X_{t_i}^{\psi^{\theta_\ell,\tau_\ell},j}$, $i = 0, \ldots, N_\ell$, $j = 1, \ldots, m_\ell$.
5: Obtain an updated estimation $\theta_{\ell+1}$ by using (2.2.23) and the data $(X_{t_i}^{\psi^{\theta_\ell,\tau_\ell},j})_{i=0,\ldots,N_\ell,j=1,\ldots,m_\ell}$.
6: **end for**

Again, as the ℓ-th cycle of Algorithm 9 contains m_ℓ episodes, for each $M \in \mathbb{N}$, the regret of learning up to M episodes (cf. (2.2.8)) can be upper bounded by the accumulated regret at the end of the L-th cycle, where L is the smallest integer such that $\sum_{\ell=0}^{L} m_\ell \geq M$. The following theorem is an analogue of Theorem 2.2.2 for Algorithm 9.

Theorem 2.2.3. *Suppose (H.1) holds and let* $\theta_0 = (A_0, B_0)^\top \in \mathbb{R}^{(n+d)\times d}$ *such that (H.15) holds with* θ_0*. Then there exists* $C_0 > 0$ *and* $n_0 \in \mathbb{N}$ *such that for all* $C \geq C_0$*, and* $\delta \in (0, \frac{3}{\pi^2})$*,*

if one sets $m_0 = C(-\ln\delta)$, $m_\ell = 2^\ell m_0$ *and* $N_\ell \geq n_0$ *for all* $\ell \in \mathbb{N}\cup\{0\}$, *then with probability at least* $1 - \frac{\pi^2\delta}{3}$, *the regret of Algorithm 9 given by (2.2.8) satisfies*

$$R(M) \leq C'\left((\ln M)(\ln\ln M) + (-\ln\delta)(\ln M) + (-\ln\delta)\sum_{\ell=0}^{\ln M} 2^\ell N_\ell^{-2}\right), \quad \forall M \in \mathbb{N}, \tag{2.2.25}$$

where C' *is a constant independent of* M, δ *and* $(N_\ell)_{\ell\in\mathbb{N}\cup\{0\}}$.

Remark 2.2.4. *Theorem 2.2.3 provides a general regret bound of Algorithm 9 with any time discretization steps* $\{N_\ell\}_{\ell\geq 0}$, *where* N_ℓ *is the number of intervention points in the* ℓ*-th cycle. Compared with Algorithm 8, the regret of Algorithm 9 has an additional term* $(-\ln\delta)\sum_{\ell=0}^{\ln M} 2^\ell N_\ell^{-2}$*: for each learning episode, one achieves a sub-optimal loss by adjusting her policy in the discrete time and also suffers from model misspecification error in parameter estimation from discrete-time observations. Specifically,*

- *if the time discretization step is fixed for all cycles, i.e.,* $N_\ell = T/\tau$ *for all* ℓ, *then the last term of (2.2.25) is of the magnitude:*

$$\mathcal{O}\left((-\ln\delta)\sum_{\ell=0}^{\ln M} 2^\ell N_\ell^{-2}\right) = \mathcal{O}\left((-\ln\delta)\tau^2\sum_{\ell=0}^{\ln M} 2^\ell\right) = \mathcal{O}((-\ln\delta)\tau^2 M),$$

  and consequently Algorithm 9 achieves a sub-optimal linear regret;

- *if the time discretization step of the* ℓ*-th cycle increases exponentially in terms of* ℓ, *e.g.,* $N_\ell = \sqrt{2}^\ell N_0$ *for* $\ell = 1, \ldots, \ln M$, *then the last term of (2.2.25) is of the magnitude:*

$$\mathcal{O}\left((-\ln\delta)\sum_{\ell=0}^{\ln M} 2^\ell N_\ell^{-2}\right) = \mathcal{O}\left((-\ln\delta)\sum_{\ell=0}^{\ln M} N_0^{-2}\right) = \mathcal{O}((-\ln\delta)\ln M),$$

  which guarantees that the regret of Algorithm 9 is still logarithmic in M.

Sketched proof of Theorem 2.2.3. We point out the main differences between the proofs of Theorems 2.2.2-2.2.3, and give the detailed proof of Theorem 2.2.3 in Section 2.3.4. Compared with Theorem 2.2.2, the essential challenges in proving Theorem 2.2.3 are to quantify the precise dependence of the performance gap and the parameter estimation error on the stepsize. To this end, we first prove a first-order convergence of (2.2.20) to (2.2.9) as the stepsize tends to zero. Then by exploiting the affine structure of (2.2.21), we establish the following quadratic performance gap for a piecewise constant policy $\psi^{\theta,\tau}$ (Proposition 2.3.12):

$$J^{\theta^\star}(U^{\psi^{\theta,\tau}}) - J^{\theta^\star}(U^{\theta^\star}) \leq C(|\theta - \theta^\star|^2 + \tau^2). \tag{2.2.26}$$

The analysis of the parameter estimation error is somewhat involved, as the state trajectories are merely α-Hölder continuous in time with $\alpha < 1/2$. By leveraging the analytic expression of $X^{\psi^{\theta,\tau}}$, we first show the first-order convergence of $\hat{\theta}^\tau$ to $\theta^\star$ with

$$\hat{\theta}^\tau := \left(\mathbb{E}\left[\sum_{i=0}^{N-1} Z_{t_i}^{\psi^{\theta,\tau}} (Z_{t_i}^{\psi^{\theta,\tau}})^\top \tau\right]\right)^{-1} \left(\mathbb{E}\left[\sum_{i=0}^{N-1} Z_{t_i}^{\psi^{\theta,\tau}} \left(X_{t_{i+1}}^{\psi^{\theta,\tau}} - X_{t_i}^{\psi^{\theta,\tau}}\right)^\top\right]\right). \tag{2.2.27}$$

We then prove that (2.2.23) enjoys a uniform sub-exponential tail bound for all θ close to $\theta^\star$ and small τ. Comparing (2.2.23) with (2.2.27) and applying the above results allow for bounding the estimation error of (2.2.23) by (2.2.14) with an additional $\mathcal{O}(\tau)$ term (Proposition 2.3.14).

2.3 Proofs of Theorems 2.2.2 and 2.2.3

To simplify the notation, for any given $N, m \in \mathbb{N}$ and control $\psi : [0,T] \times \mathbb{R}^n \to \mathbb{R}^d$ that is affine in the spatial variable, we introduce the following random variables associated with continuous-time observations:

$$\begin{aligned} V^\psi &= \int_0^T Z_t^\psi (Z_t^\psi)^\top \,\mathrm{d}t, & Y^\psi &= \int_0^T Z_t^\psi (\mathrm{d}X_t^\psi)^\top, \\ V^{\psi,m} &= \frac{1}{m}\sum_{j=1}^m \int_0^T Z_t^{\psi,j} (Z_t^{\psi,j})^\top \,\mathrm{d}t, & Y^{\psi,m} &= \frac{1}{m}\sum_{j=1}^m \int_0^T Z_t^{\psi,j} (\mathrm{d}X_t^{\psi,j})^\top, \end{aligned} \tag{2.3.1}$$

and the random variables associated with discrete-time observations with stepsize $\tau = T/N$:

$$\begin{aligned} V^{\psi,\tau} &= \sum_{i=0}^{N-1} Z_{t_i}^\psi (Z_{t_i}^\psi)^\top \tau, & Y^{\psi,\tau} &= \sum_{i=0}^{N-1} Z_{t_i}^\psi (X_{t_{i+1}}^\psi - X_{t_i}^\psi)^\top, \\ V^{\psi,\tau,m} &= \frac{1}{m}\sum_{j=1}^m \sum_{i=0}^{N-1} Z_{t_i}^{\psi,j} (Z_{t_i}^{\psi,j})^\top \tau, & Y^{\psi,\tau,m} &= \frac{1}{m}\sum_{j=1}^m \sum_{i=0}^{N-1} Z_{t_i}^{\psi,j} (X_{t_{i+1}}^{\psi,j} - X_{t_i}^{\psi,j})^\top, \end{aligned} \tag{2.3.2}$$

where X^ψ is the state process associated with the parameter $\theta^\star$ and the control ψ (cf. (2.2.6)), $Z_t^\psi = \begin{pmatrix} X^\psi \\ \psi(t, X_t^\psi) \end{pmatrix}$ for all $t \in [0,T]$, and $(X^{\psi,j}, Z^{\psi,j})_{j=1}^m$ are independent copies of (X^ψ, Z^ψ).

2.3.1 Convergence and stability of Riccati equations and feedback controls

Lemma 2.3.1. *Suppose (H.1(1)) holds. Then for all $\theta = (A,B)^\top \in \mathbb{R}^{(n+d)\times n}$, the Riccati equation*

$$\tfrac{\mathrm{d}}{\mathrm{d}t}P_t + A^\top P_t + P_t A - P_t B R^{-1} B^\top P_t + Q = 0, \quad t \in [0,T]; \quad P_T = 0. \tag{2.3.3}$$

admits a unique solution $P^\theta \in C([0,T];\mathbb{R}^{n\times n})$. Moreover, the map $\mathbb{R}^{(n+d)\times n} \ni \theta \mapsto P^\theta \in C^1([0,T];\mathbb{R}^{n\times n})$ is continuously differentiable.

Proof. It has been shown in [169, Corollary 2.10 on p. 297] that under (H.1(1)), for all $\theta = (A, B)^\top \in \mathbb{R}^{(n+d)\times n}$, (2.3.3) admits a unique solution $P^\theta \in C([0,T];\mathbb{R}^{n\times n})$ such that $P_t^\theta \in \mathbb{S}_0^n$ for all $t \in [0,T]$. It remains to to prove the continuous differentiability of $\theta \mapsto P^\theta$.

To this end, consider the Banach spaces $\mathbb{X} = \mathbb{R}^{(n+d)\times n} \times C^1([0,T];\mathbb{R}^{n\times n})$ and $\mathbb{Y} = C([0,T];\mathbb{R}^{n\times n}) \times \mathbb{R}^{n\times n}$, and the operator $\Phi : \mathbb{X} \to \mathbb{Y}$ defined by

$$\mathbb{X} \ni (\theta, P) \mapsto \Phi(\theta, P) \coloneqq (F(\theta, P), P_T) \in \mathbb{Y},$$

where $F(\theta,P)_t = \frac{\mathrm{d}}{\mathrm{d}t}P_t + A^\top P_t + P_t A - P_t B R^{-1} B^\top P_t + Q$ for all $t \in [0,T]$. Observe that for all $\theta \in \mathbb{R}^{(n+d)\times n}$, $\Phi(P^\theta, \theta) = 0$. Moreover, one can easily show that for any $(P,\theta) \in \mathbb{X}$, Φ is continuously Fréchet differentiable at (P,θ), and the partial derivative $\frac{\partial}{\partial P}\Phi(\theta,P) : C^1([0,T];\mathbb{R}^{n\times n}) \to \mathbb{Y}$ is a bounded linear operator such that for all $\tilde{P} \in C^1([0,T];\mathbb{R}^{n\times n})$,

$$\frac{\partial}{\partial P}\Phi(\theta,P)(\tilde{P}) = \begin{pmatrix} \left(\frac{\mathrm{d}}{\mathrm{d}t}\tilde{P}_t + A^\top \tilde{P}_t + \tilde{P}_t A - \tilde{P}_t B R^{-1} B^\top P_t - P_t B R^{-1} B^\top \tilde{P}_t\right)_{t\in[0,T]} \\ \tilde{P}_T \end{pmatrix} \in \mathbb{Y}.$$

Classical well-posedness results of linear differential equations and the boundedness of P imply that $\frac{\partial}{\partial P}\Phi(P,\theta) : C^1([0,T];\mathbb{R}^{n\times n}) \to \mathbb{Y}$ has a bounded inverse (and hence a bijection). Thus, applying the implicit function theorem (see [38, Theorem 7.13-1]) to Φ proves that $\mathbb{R}^{(n+d)\times n} \ni \theta \mapsto P^\theta \in C^1([0,T];\mathbb{R}^{n\times n})$ is continuously differentiable. □

The following lemma establishes the stability of the Riccati difference operator, which is crucial for the subsequent convergence analysis.

Lemma 2.3.2. *Suppose (H.1(1)) holds. For each $\theta = (A,B)^\top \in \mathbb{R}^{(n+d)\times n}$ and $N \in \mathbb{N}$, let $\tau = T/N$ and the function $\Gamma_\tau^\theta : \mathbb{S}_0^n \to \mathbb{S}_0^n$ such that for all $P \in \mathbb{S}_0^n$,*

$$\Gamma_\tau^\theta(P) \coloneqq \tau Q + (I+\tau A)^\top P(I+\tau A) - (I+\tau A)^\top P \tau B(R+\tau B^\top P B)^{-1} B^\top P(I+\tau A). \quad (2.3.4)$$

Then for all $P, P' \in \mathbb{S}_0^n$,

(1) $\|\Gamma_\tau^\theta(P)\|_2 \le \tau\|Q\|_2 + (1+\tau\|A\|_2)^2\|P\|_2$,

(2) $\|\Gamma_\tau^\theta(P) - \Gamma_\tau^\theta(P')\|_2 \le \left(1+\tau\|R^{-1}\|_2\|B\|_2^2 \max\{\|P\|_2, \|P'\|_2\}\right)^2 (1+\tau\|A\|_2)^2\|P-P'\|_2$.

Proof. Item (1) follows directly from the definition of Γ_τ^θ and the identity that $\|\Gamma_\tau^\theta(P)\|_2 = \sup\{x^\top \Gamma_\tau^\theta(P)x \mid x \in \mathbb{R}^n, |x| = 1\}$. We now prove Item (2). Let $\delta P = P - P'$ and $\delta\Gamma(P) = \Gamma_\tau^\theta(P) - \Gamma_\tau^\theta(P')$, by [24, Lemma 10.1],

$$\delta\Gamma(P) = F^\top \delta P F - F^\top \delta P \tau B(\tau B^\top P \tau B + \tau R)^{-1}\tau B^\top \delta P F,$$

with $F = (I - \tau B(\tau B^\top P' \tau B + \tau R)^{-1}\tau B^\top P')(I+\tau A)$. Thus for all $x \in \mathbb{R}^n$, $x^\top \delta\Gamma(P)x \le \|\delta P\|_2\|F\|_2^2|x|^2$, which along with $\|(\tau B^\top P' B + R)^{-1}\|_2 \le \|R^{-1}\|_2$ implies

$$x^\top \delta\Gamma(P)x \le \|\delta P\|_2(1+\tau\|R^{-1}\|_2\|B\|_2^2\|P'\|_2)^2(1+\tau\|A\|_2)^2|x|^2, \quad x \in \mathbb{R}^n.$$

Hence, interchanging the roles of P and P' in the above inequality and taking the supremum over $x \in \mathbb{R}^n$ lead to the desired estimate. □

The following proposition establishes the first-order convergence of the Riccati difference equation and the associated feedback controls, as the stepsize tends to zero.

Proposition 2.3.3. *Suppose (H.1(1)) holds and let Θ be a bounded subset of $\mathbb{R}^{(n+d)\times n}$. For each $\theta = (A, B)^\top \in \Theta$ and $N \in \mathbb{N}$, let $(P_i^{\theta,\tau})_{i=0}^N$ such that $P_N^{\theta,\tau} = 0$ and $P_i^{\theta,\tau} = \Gamma_\tau^\theta(P_{i+1}^{\theta,\tau})$ for all $i = 0, \dots, N-1$, with Γ_τ^θ defined in (2.3.4) with $\tau = T/N$. Then there exists a constant $C \geq 0$ such that for all $\theta \in \Theta, N \in \mathbb{N}$,*

$$\sup_{i=0,\dots,N-1} \sup_{t\in[i\tau,(i+1)\tau)} \left(\|P_t^\theta - P_i^{\theta,\tau}\|_2 + \|K_t^\theta - K_i^{\theta,\tau}\|_2\right) \leq C\tau,$$

where $P^\theta \in C^1([0,T];\mathbb{R}^{n\times n})$ satisfies (2.3.3), $K_t^\theta = -R^{-1}B^\top P_t^\theta$ for all $t \in [0,T]$ and $K_i^{\theta,\tau} = -(R+\tau B^\top P_{i+1}^{\theta,\tau}B)^{-1}B^\top P_{i+1}^{\theta,\tau}(I+\tau A)$ for all $i = 0, \dots, N-1$.

Proof. Throughout this proof, we shall fix $\theta \in \Theta, N \in \mathbb{N}$, let $t_i = i\tau$ for all $i = 0, \dots, N$, and denote by C a generic constant independent of N and θ. By the continuity of the map $\theta \mapsto P^\theta$ (Lemma 2.3.1) and the boundedness of Θ, there exists a constant C such that $\|P^\theta\|_{C^1([0,T];\mathbb{R}^{n\times n})} \leq C$ for all $\theta \in \Theta$, which implies $\|P_t^\theta - P_s^\theta\|_2 \leq C|t-s|$ for all $t, s \in [0,T]$. Consequently, it suffices to prove $\|P_{t_i}^\theta - P_i^{\theta,\tau}\|_2 + \|K_{t_{i+1}}^\theta - K_i^{\theta,\tau}\|_2 \leq C\tau$ for all $i = 0, \dots, N-1$.

We start by making two important observations. By Lemma 2.3.2 Item (1), $\|P_i^{\theta,\tau}\|_2 \leq \tau C + (1+C\tau)\|P_{i+1}^{\theta,\tau}\|_2$ for all $i = 0, \dots, N-1$, which along with Gronwall's inequality gives $\|P_i^{\theta,\tau}\|_2 \leq C$ for all $i = 0, \dots, N$. Moreover, by (2.3.4), for all $P \in \mathbb{S}_0^n$,

$$\begin{aligned}\Gamma_\tau^\theta(P) = {} & \tau Q + P + \tau(A^\top P + PA) + \tau^2 A^\top PA \\ & - \tau\Big(PB(R+\tau B^\top PB)^{-1}B^\top P + \tau(A^\top H + HA^\top) + \tau^2 A^\top HA\Big),\end{aligned}$$

with $H \coloneqq PB(R+\tau B^\top PB)^{-1}B^\top P$. Hence for any given $i = 0, \dots, N-1$, we see from (2.3.3) that

$$\begin{aligned}& P_{t_i}^\theta - \Gamma_\tau^\theta(P_{t_{i+1}}^\theta) \\ & = \int_{t_i}^{t_{i+1}} \left(A^\top(P_t^\theta - P_{t_{i+1}}^\theta) + (P_t^\theta - P_{t_{i+1}}^\theta)A\right) \mathrm{d}t - \int_{t_i}^{t_{i+1}} (P_t^\theta BR^{-1}B^\top P_t^\theta - P_{t_{i+1}}^\theta BR^{-1}B^\top P_{t_{i+1}}^\theta)\,\mathrm{d}t \\ & \quad - \int_{t_i}^{t_{i+1}} (P_{t_{i+1}}^\theta BR^{-1}B^\top P_{t_{i+1}}^\theta - P_{t_{i+1}}^\theta B(R+\tau B^\top P_{t_{i+1}}^\theta B)^{-1}B^\top P_{t_{i+1}}^\theta)\,\mathrm{d}t \\ & \quad + \tau^2(-A^\top P_{t_{i+1}}^\theta A + A^\top H_{i+1}^\theta + H_{i+1}^\theta A^\top + \tau A^\top H_{i+1}^\theta A),\end{aligned}$$

with $H_{i+1}^\theta = P_{t_{i+1}}^\theta B(R+\tau B^\top P_{t_{i+1}}^\theta B)^{-1}B^\top P_{t_{i+1}}^\theta$. Since $\|P^\theta\|_{C^1([0,T];\mathbb{R}^{n\times n})} \leq C$ and $R \in \mathbb{S}_+^d$, we have $\|P_{t_i}^\theta - \Gamma_\tau^\theta(P_{t_{i+1}}^\theta)\|_2 \leq C\tau^2$ for all $i = 0, \dots, N-1$.

We are ready to show $\max_{i=0,\dots,N-1}(\|P_{t_i}^\theta - P_i^{\theta,\tau}\|_2 + \|K_{t_{i+1}}^\theta - K_i^{\theta,\tau}\|_2) \leq C\tau$. For any given $i = 0, \dots, N-1$, by Lemma 2.3.2 Item (2) and the uniform boundedness of $(P_{t_i}^\theta)_{i=0}^N$ and

$(P_i^{\theta,\tau})_{i=0}^N$,

$$\begin{aligned}\|P_{t_i}^\theta - P_i^{\theta,\tau}\|_2 &\le \|P_{t_i}^\theta - \Gamma_\tau^\theta(P_{t_{i+1}}^\theta)\|_2 + \|\Gamma_\tau^\theta(P_{t_{i+1}}^\theta) - \Gamma_\tau^\theta(P_{i+1}^{\theta,\tau})\|_2 \\ &\le C\tau^2 + \big(1+\tau C\max\{\|P_{t_{i+1}}^\theta\|_2, \|P_{i+1}^{\theta,\tau}\|_2\}\big)^2(1+\tau)\|P_{t_{i+1}}^\theta - P_{i+1}^{\theta,\tau}\|_2 \\ &\le C\tau^2 + \big(1+\tau C\big)\|P_{t_{i+1}}^\theta - P_{i+1}^{\theta,\tau}\|_2,\end{aligned}$$

which along with Gronwall's inequality and $P_T^\theta = P_N^{\theta,\tau} = 0$ shows the desired convergence rate of $(P_i^{\theta,\tau})_{i=1}^N$. Furthermore, for all $i = 0, \dots, N-1$,

$$\begin{aligned}\|K_{t_{i+1}}^\theta - K_i^{\theta,\tau}\|_2 \le{}& \|(R^{-1} - (R+\tau B^\top P_{i+1}^{\theta,\tau}B)^{-1})B^\top P_{t_{i+1}}^\theta\|_2 \\ &+ \|(R+\tau B^\top P_{i+1}^{\theta,\tau}B)^{-1}B^\top(P_{t_{i+1}}^\theta - P_{i+1}^{\theta,\tau}(I+\tau A))\|_2 \le C\tau,\end{aligned}$$

from the facts that $\|P_{t_i}^\theta\|_2 \le C$, $\|P_i^{\theta,\tau}\|_2 \le C$ and $\|P_{t_i}^\theta - P_i^{\theta,\tau}\|_2 \le C\tau$ for all i. □

2.3.2 Concentration inequalities for least-squares estimators

In this section, we analyze the concentration behavior of the least-squares estimators (2.2.13) and (2.2.23). We first recall the definition of sub-exponential random variables (see e.g., [158]).

Definition 2.3.1. *A random variable X with mean $\mu = \mathbb{E}[X]$ is (ν, b)-sub-exponential for $\nu, b \in [0,\infty)$ if $\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{\nu^2\lambda^2/2}$ for all $|\lambda| < 1/b$.*

Note that a $(\nu, 0)$-sub-exponential random variable is usually called a sub-Gaussian random variable. It is well-known that products of sub-Gaussian random variables are sub-exponential, and the class of sub-exponential random variables forms a vector space. Moreover, sub-exponential random variables enjoy the following concentration inequality (also known as Bernstein's inequality; see e.g., [158, Equation 2.18 p. 29]).

Lemma 2.3.4. *Let $m \in \mathbb{N}$, $\nu, b \in [0,\infty)$ and $(X_i)_{i=1}^m$ be independent (ν, b)-sub-exponential random variables with $\mu = \mathbb{E}[X_i]$ for all $i = 1, \dots, m$. Then for all $\epsilon \ge 0$,*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m X_i - \mu\right| \ge \epsilon\right) \le 2\exp\left(-\min\left\{\frac{m\epsilon^2}{2\nu^2}, \frac{m\epsilon}{2b}\right\}\right).$$

The following lemma shows double iterated Itô integrals are sub-exponential random variables.

Lemma 2.3.5. *Let $L \ge 0$ and $g, h : [0,T]\times[0,T] \to \mathbb{R}^{n\times n}$ be measurable functions such that $|g(t,s)| \le L$ and $|h(t,s)| \le L$ for all $t, s \in [0,T]$. Then there exist $\nu, b \in [0,\infty)$, depending polynomially on L, n, T, such that*

(1) $\int_0^T \big(\int_0^t g(t,s)\,\mathrm{d}W_s\big)^\top \mathrm{d}W_t$,

(2) $\int_0^T \left(\int_0^t g(t,s)\,\mathrm{d}W_s\right)^\top \left(\int_0^t h(t,s)\,\mathrm{d}W_s\right)\mathrm{d}t$

are (ν, b)*-sub-exponential,*

Proof. We first prove Item (1) by assuming without loss of generality that $\|g(t,s)\|_2 \le L$ for all $t, s \in [0,T]$, and by defining $V^q := \int_0^T \left(\int_0^t q(t,s)\,\mathrm{d}W_s\right)^\top \mathrm{d}W_t$ for any bounded measurable function $q : [0,T]\times[0,T] \to \mathbb{R}^{n\times n}$. By similar arguments as [37, Lemma 3.2], we have for all $t \in [0,T]$ and $0 \le \lambda < \frac{1}{2T}$,

$$\mathbb{E}[\exp(2\lambda V^{\frac{g}{L}})] \le \mathbb{E}[\exp(2\lambda V^{I_n})] = \left(\frac{1}{\sqrt{1-2\lambda T}}\exp(-\lambda T)\right)^n.$$

As $\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \le e^{2\lambda^2}$ for all $|\lambda| \le 1/4$, we see $\mathbb{E}[\exp(2\lambda V^{g/L})] \le \exp(2n\lambda^2 T^2)$ for all $0 \le \lambda < \frac{1}{4T}$. Consequently, for all $0 \le \lambda < \frac{1}{2LT}$,

$$\mathbb{E}[\exp(\lambda V^g)] = \mathbb{E}\left[\exp\left(2\frac{\lambda L}{2}V^{\frac{g}{L}}\right)\right] \le \exp\left(\frac{nL^2T^2\lambda^2}{2}\right).$$

Replacing g by $-g$ shows the above estimate holds for $|\lambda| < \frac{1}{2LT}$, which implies the desired sub-exponential property of V^g.

For Item (2), observe that for each $t \in [0,T]$, the Itô formula allows one to express the product $\left(\int_0^t g(t,s)\,\mathrm{d}W_s\right)^\top \left(\int_0^t h(t,s)\,\mathrm{d}W_s\right)$ as a linear combination of double iterated Itô integrals and deterministic integrals. Then the desired sub-exponential property follows from the stochastic Fubini theorem (see e.g., [155]) and Item (1). □

The following theorem establishes the concentration properties of the random variables involved in the least-squares estimators.

Theorem 2.3.6. *Suppose (H.1(1)) holds and let* Θ *be a bounded subset of* $\mathbb{R}^{(n+d)\times n}$*. For each* $\theta \in \Theta$ *and* $N \in \mathbb{N}$*, let* ψ^θ *be defined in* (2.2.9)*, and* $\psi^{\theta,\tau}$ *be defined in* (2.2.20) *with stepsize* $\tau = T/N$*. Then there exist constants* $C, \nu, b > 0$ *such that for all* $\theta \in \Theta$*,* $N, m \in \mathbb{N}$ *and* $\epsilon > 0$*,*

$$\max\Big\{\mathbb{P}(|V^{\psi^\theta,m} - \mathbb{E}[V^{\psi^\theta}]| \ge \epsilon), \mathbb{P}(|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]| \ge \epsilon),$$
$$\mathbb{P}(|V^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[V^{\psi^{\theta,\tau},\tau}]| \ge \epsilon), \mathbb{P}(|Y^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[Y^{\psi^{\theta,\tau},\tau}]| \ge \epsilon)\Big\} \le C\exp\left(-\frac{1}{C}\min\left\{\frac{m\epsilon^2}{\nu^2}, \frac{m\epsilon}{b}\right\}\right),$$

where V^{ψ^θ}*,* Y^{ψ^θ}*,* $V^{\psi^\theta,m}$*,* $Y^{\psi^\theta,m}$ *are defined in* (2.3.1)*, and* $V^{\psi^{\theta,\tau},\tau}$*,* $Y^{\psi^{\theta,\tau},\tau}$*,* $V^{\psi^{\theta,\tau},\tau,m}$*,* $Y^{\psi^{\theta,\tau},\tau,m}$ *are defined in* (2.3.2)*.*

Proof. We first show there exist $\nu, b > 0$ such that all entries of V^{ψ^θ}, Y^{ψ^θ} $Y^{\psi^{\theta,\tau},\tau}$, $V^{\psi^{\theta,\tau},\tau}$ are (ν, b)-sub-exponential for all $\theta \in \Theta$ and $N \in \mathbb{N}$. By (2.3.1), we have

$$V^{\psi^\theta} = \int_0^T \begin{pmatrix} X_t^{\psi^\theta} \\ K_t^\theta X_t^{\psi^\theta}\end{pmatrix}\begin{pmatrix}(X_t^{\psi^\theta})^\top & (K_t^\theta X_t^{\psi^\theta})^\top\end{pmatrix}\mathrm{d}t, \quad Y^{\psi^\theta} = V^{\psi^\theta}\theta^\star + \int_0^T \begin{pmatrix} X_t^{\psi^\theta} \\ K_t^\theta X_t^{\psi^\theta}\end{pmatrix}(\mathrm{d}W_t)^\top.$$

Moreover, applying the variation-of-constants formula (see e.g., [113, Theorem 3.1 p. 96]) to (2.2.11) shows that $X_t^{\psi^\theta} = \Phi_t^\theta\big(x_0+\int_0^t(\Phi_s^\theta)^{-1}\,\mathrm{d}W_s\big)$ for all $t\in[0,T]$, where $\Phi^\theta\in C([0,T];\mathbb{R}^{n\times n})$ is the fundamental solution of $\mathrm{d}\Phi_t^\theta = (A^\star + B^\star K_t^\theta)\Phi_t^\theta\,\mathrm{d}t$. The continuity of $\mathbb{R}^{(n+d)\times n}\ni\theta\mapsto K^\theta\in C([0,T];\mathbb{R}^{d\times n})$ (cf. Proposition 2.3.1) and the boundedness of Θ implies that $K^\theta,\Phi^\theta,(\Phi^\theta)^{-1}$ are uniformly bounded for all $\theta\in\Theta$. Consequently, from Lemma 2.3.5, there exist $\nu,b>0$ such that all entries of V^{ψ^θ} and Y^{ψ^θ} are (ν,b)-sub-exponential.

Similarly, by (2.2.21) and (2.3.2),

$$
\begin{aligned}
V^{\psi^{\theta,\tau},\tau} &= \int_0^T \sum_{i=0}^{N-1} \mathbf{1}_{[t_i,t_{i+1})}(t) \begin{pmatrix} X_{t_i}^{\psi^{\theta,\tau}} \\ K_{t_i}^{\theta,\tau} X_{t_i}^{\psi^{\theta,\tau}} \end{pmatrix} \begin{pmatrix} (X_{t_i}^{\psi^{\theta,\tau}})^\top & (K_{t_i}^{\theta,\tau} X_{t_i}^{\psi^{\theta,\tau}})^\top \end{pmatrix} \mathrm{d}t, \\
Y^{\psi^{\theta,\tau},\tau} &= \int_0^T \sum_{i=0}^{N-1} \mathbf{1}_{[t_i,t_{i+1})}(t) \begin{pmatrix} X_{t_i}^{\psi^{\theta,\tau}} \\ K_{t_i}^{\theta,\tau} X_{t_i}^{\psi^{\theta,\tau}} \end{pmatrix} \begin{pmatrix} (X_t^{\psi^{\theta,\tau}})^\top & (K_t^{\theta,\tau} X_t^{\psi^{\theta,\tau}})^\top \end{pmatrix} (\theta^\star)^\top \mathrm{d}t \\
&\quad + \int_0^T \sum_{i=0}^{N-1} \mathbf{1}_{[t_i,t_{i+1})}(t) \begin{pmatrix} X_{t_i}^{\psi^{\theta,\tau}} \\ K_{t_i}^{\theta,\tau} X_{t_i}^{\psi^{\theta,\tau}} \end{pmatrix} (\mathrm{d}W_t)^\top,
\end{aligned}
$$

where $X_t^{\psi^{\theta,\tau}} = \Phi_t^{\theta,\tau}\big(x_0+\int_0^t(\Phi_s^{\theta,\tau})^{-1}\,\mathrm{d}W_s\big)$ for all $t\in[0,T]$, and $\Phi^{\theta,\tau}\in C([0,T];\mathbb{R}^{n\times n})$ is the fundamental solution of $\mathrm{d}\Phi_t^{\theta,\tau} = (A^\star+B^\star K_t^{\theta,\tau})\Phi_t^{\theta,\tau}\,\mathrm{d}t$. By Proposition 2.3.3, $K^{\theta,\tau},\Phi^{\theta,\tau},(\Phi^{\theta,\tau})^{-1}$ are uniformly bounded for all $\theta\in\Theta$ and $N\in\mathbb{N}$, which along with Lemma 2.3.5 leads to the desired sub-exponential properties of $Y^{\psi^{\theta,\tau},\tau}$ and $V^{\psi^{\theta,\tau},\tau}$.

Finally, since $\mathbb{P}(|\sum_{i=1}^\ell X_i|\ge\epsilon)\le\sum_{i=1}^\ell\mathbb{P}(|X_i|\ge\epsilon/\ell)$ for all $\ell\in\mathbb{N}$ and random variables $(X_i)_{i=1}^\ell$, we can apply Lemma 2.3.4 to each component of V^{ψ^θ}, Y^{ψ^θ} $Y^{\psi^{\theta,\tau},\tau}$ and $V^{\psi^{\theta,\tau},\tau}$, and conclude the desired concentration inequality with a constant C depending polynomially on n,d. □

2.3.3 Regret analysis of continuous-time least-squares algorithm

This section is devoted to the proof of Theorem 2.2.2, which consists of three steps: (1) We first quantify the performance gap between applying feedback controls for an incorrect model and that for the true model; our proof exploits the stability of Riccati equations established in Lemma 2.3.1; (2) We then estimate the parameter estimation error in terms of the number of learning episodes based on the sub-exponential tail behavior of the least-squares estimator (2.2.13); (3) Finally, we estimate the regret for the feedback controls $(\psi^{\theta_\ell})_{\ell\in\mathbb{N}}$ in Algorithm 8, thus establishing Theorem 2.2.2.

Step 1: Analysis of the performance gap. We start by establishing a quadratic expansion of the cost function at any open-loop control.

Proposition 2.3.7. *Suppose (H.1(1)) holds. Let $\psi^{\theta^\star}$ be defined in (2.2.3), $X^{\theta^\star}$ be the state process associated with $\psi^{\theta^\star}$ (cf. (2.2.5)), and $U^{\theta^\star}\in\mathcal{H}^2(\mathbb{R}^d)$ be such that for all $t\in[0,T]$,*

$U_t^{\theta^\star} = \psi^{\theta^\star}(t, X_t^{\theta^\star})$. *Then for all* $U \in \mathcal{H}^2(\mathbb{R}^d)$,

$$J^{\theta^\star}(U) - J^{\theta^\star}(U^{\theta^\star}) \le \|Q\|_2 \|X^{\theta^\star, U} - X^{\theta^\star}\|^2_{\mathcal{H}^2(\mathbb{R}^n)} + \|R\|_2 \|U - U^{\theta^\star}\|^2_{\mathcal{H}^2(\mathbb{R}^d)}, \tag{2.3.5}$$

where $X^{\theta^\star, U}$ *is the state process controlled by* U *(cf.* (2.2.2)*), and* $J^{\theta^\star} : \mathcal{H}^2(\mathbb{R}^d) \to \mathbb{R}$ *is defined in* (2.2.1).

Proof. For notational simplicity, for all $U \in \mathcal{H}^2(\mathbb{R}^d)$ and $\epsilon > 0$, we write $U^\epsilon = U^{\theta^\star} + \epsilon(U - U^{\theta^\star})$, denote by $X^\epsilon = X^{\theta^\star, U^\epsilon}$ the associated state process defined by (2.2.2), and by $X^U = X^0 = X^{\theta^\star, U}$. The affineness of (2.2.2) implies that $X^\epsilon = (1-\epsilon)X^{\theta^\star} + \epsilon X^U$ for all $\epsilon > 0$. Hence, for all $U \in \mathcal{H}^2(\mathbb{R}^d)$,

$$\begin{aligned}
&\lim_{\epsilon \searrow 0} \frac{1}{\epsilon} \left(J^{\theta^\star}(U^\epsilon) - J^{\theta^\star}(U^{\theta^\star}) \right) \\
&= \lim_{\epsilon \searrow 0} \frac{1}{\epsilon} \mathbb{E}\bigg[\int_0^T \bigg(\Big((1-\epsilon)X_t^{\theta^\star} + \epsilon X_t^U \Big)^\top Q \Big((1-\epsilon)X_t^{\theta^\star} + \epsilon X_t^U \Big) - (X_t^{\theta^\star})^\top Q X_t^{\theta^\star} \\
&\quad + \left((1-\epsilon)U_t^{\theta^\star} + \epsilon U_t \right)^\top R \left((1-\epsilon)U_t^{\theta^\star} + \epsilon U_t \right) - (U_t^{\theta^\star})^\top R U_t^{\theta^\star} \bigg) \, \mathrm{d}t \bigg] \\
&= \lim_{\epsilon \searrow 0} \epsilon \, \mathbb{E}\left[\int_0^T \Big((X_t^U - X_t^{\theta^\star})^\top Q (X_t^U - X_t^{\theta^\star}) + (U_t - U_t^{\theta^\star})^\top R (U_t - U_t^{\theta^\star}) \Big) \, \mathrm{d}t \right] \\
&\quad + 2\mathbb{E}\left[\int_0^T \left((X_t^U - X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + (U_t - U_t^{\theta^\star})^\top R U_t^{\theta^\star} \right) \mathrm{d}t \right] \\
&= 2\mathbb{E}\left[\int_0^T \left((X_t^U - X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + (U_t - U_t^{\theta^\star})^\top R U_t^{\theta^\star} \right) \mathrm{d}t \right],
\end{aligned}$$

which is based on the fact that $X^U - X^{\theta^\star} \in \mathcal{H}^2(\mathbb{R}^n)$ and $U - U^\star \in \mathcal{H}^2(\mathbb{R}^d)$. As U^θ is the optimal control of $J^{\theta^\star}$, $J^{\theta^\star}(U) \ge J^{\theta^\star}(U^{\theta^\star})$ for all $U \in \mathcal{H}^2(\mathbb{R}^d)$. Hence for all $U \in \mathcal{H}^2(\mathbb{R}^d)$,

$$\mathbb{E}\left[\int_0^T \left((X_t^U - X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + (U_t - U_t^{\theta^\star})^\top R U_t^{\theta^\star} \right) \mathrm{d}t \right] = \lim_{\epsilon \searrow 0} \frac{1}{2\epsilon} (J^{\theta^\star}(U^\epsilon) - J^{\theta^\star}(U^{\theta^\star})) \ge 0. \tag{2.3.6}$$

We now prove that the above quantity is in fact zero for all $U \in \mathcal{H}^2(\mathbb{R}^d)$. To this end, let $U \in \mathcal{H}^2(\mathbb{R}^d)$ be a given (open-loop) control, and consider $\tilde{U} = U^{\theta^\star} - (U - U^{\theta^\star})$. Then by the affineness of (2.2.2), $X^{\tilde{U}} - X^{\theta^\star}$ satisfies the following controlled dynamics:

$$\mathrm{d}X_t = (A^\star X_t - B^\star (U - U^{\theta^\star})_t) \, \mathrm{d}t, \quad t \in [0, T]; \quad X_0 = 0. \tag{2.3.7}$$

Moreover, one can verify by the affineness of (2.2.2) that $-(X^U - X^{\theta^\star})$ also satisfies the dynamics (2.3.7), which along with the uniqueness of solutions to (2.3.7) shows that $X^{\tilde{U}} - X^{\theta^\star} = -(X^U - X^{\theta^\star})$. Therefore, applying (2.3.6) with $U = \tilde{U}$ implies that

$$\begin{aligned}
0 &\le \mathbb{E}\left[\int_0^T \left((X_t^{\tilde{U}} - X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + (\tilde{U}_t - U_t^{\theta^\star})^\top R U_t^{\theta^\star} \right) \mathrm{d}t \right] \\
&= -\mathbb{E}\left[\int_0^T \left((X_t^U - X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + (U_t - U_t^{\theta^\star})^\top R U_t^{\theta^\star} \right) \mathrm{d}t \right] \le 0.
\end{aligned}$$

Hence for all $U \in \mathcal{H}^2(\mathbb{R}^d)$,

$$\mathbb{E}\left[\int_0^T \left((X_t^U - X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + (U_t - U_t^{\theta^\star})^\top R U_t^{\theta^\star}\right) \mathrm{d}t\right] = 0,$$

which leads to the desired result (2.3.5) due to the following identify:

$$\begin{aligned}
J^{\theta^\star}(U) - J^{\theta^\star}(U^{\theta^\star}) &= \mathbb{E}\left[\int_0^T ((X_t^U)^\top Q X_t^U - (X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + U_t^\top R U_t - (U_t^{\theta^\star})^\top R U_t^{\theta^\star})\,\mathrm{d}t\right] \\
&= \mathbb{E}\left[\int_0^T \left((X_t^U)^\top Q X_t^U - (X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + U_t^\top R U_t - (U_t^{\theta^\star})^\top R U_t^{\theta^\star}\right) \mathrm{d}t\right] \\
&\quad - 2\mathbb{E}\left[\int_0^T \left((X_t^U - X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + (U_t - U_t^{\theta^\star})^\top R U_t^{\theta^\star}\right) \mathrm{d}t\right] \\
&= \mathbb{E}\left[\int_0^T \left((X_t^U - X_t^{\theta^\star})^\top Q (X_t^U - X_t^{\theta^\star}) + (U_t - U_t^{\theta^\star})^\top R (U_t - U_t^{\theta^\star})\,\mathrm{d}t\right)\right].
\end{aligned}$$

□

Armed with Proposition 2.3.7, the following proposition quantifies the quadratic performance gap of a greedy policy ψ^θ.

Proposition 2.3.8. *Suppose (H.1(1)) holds and let Θ be a bounded subset of $\mathbb{R}^{(n+d)\times n}$. For each $\theta \in \Theta$, let ψ^θ be defined in (2.2.9), let X^{ψ^θ} be the state process associated with ψ^θ (cf. (2.2.11)), let $\psi^{\theta^\star}$ be defined in (2.2.3), and let $X^{\theta^\star}$ be the state process associated with $\psi^{\theta^\star}$ (cf. (2.2.5)). Then there exists a constant C such that*

$$|J^{\theta^\star}(U^{\psi^\theta}) - J^{\theta^\star}(U^{\theta^\star})| \le C|\theta - \theta^\star|^2, \quad \forall \theta \in \Theta,$$

where $U_t^{\psi^\theta} = \psi^\theta(t, X_t^{\psi^\theta})$ and $U_t^{\theta^\star} = \psi^{\theta^\star}(t, X_t^{\theta^\star})$ for all $t \in [0,T]$, and $J^{\theta^\star}$ is defined in (2.2.1).

Proof. For all $\theta \in \Theta$, applying Proposition 2.3.7 with $U = U^{\psi^\theta}$ gives

$$\begin{aligned}
&J^{\theta^\star}(U^{\psi^\theta}) - J^{\theta^\star}(U^{\theta^\star}) \\
&\le \|Q\|_2 \|X^{\theta^\star, U^{\psi^\theta}} - X^{\theta^\star}\|^2_{\mathcal{H}^2(\mathbb{R}^n)} + \|R\|_2 \|U^{\psi^\theta} - U^{\theta^\star}\|^2_{\mathcal{H}^2(\mathbb{R}^d)}, \qquad (2.3.8) \\
&\le \|Q\|_2 \|X^{\psi^\theta} - X^{\psi^{\theta^\star}}\|^2_{\mathcal{H}^2(\mathbb{R}^n)} + \|R\|_2 \|\psi^\theta(\cdot, X_\cdot^{\psi^\theta}) - \psi^{\theta^\star}(\cdot, X_\cdot^{\psi^{\theta^\star}})\|^2_{\mathcal{H}^2(\mathbb{R}^d)},
\end{aligned}$$

where the last inequality used the fact that $X^{\theta^\star, U^{\psi^\theta}} = X^{\psi^\theta}$ (see (2.2.11)), and the definitions of U^{ψ^θ} and $U^{\theta^\star}$. It remains to prove

$$\|X^{\psi^\theta} - X^{\psi^{\theta^\star}}\|_{\mathcal{H}^2(\mathbb{R}^n)} + \|\psi^\theta(\cdot, X_\cdot^{\psi^\theta}) - \psi^{\theta^\star}(\cdot, X_\cdot^{\psi^{\theta^\star}})\|_{\mathcal{H}^2(\mathbb{R}^d)} \le C|\theta - \theta^\star|,$$

for a constant C independent of θ.

Observe that by (2.2.9), for all $(t,x)\in[0,T]\times\mathbb{R}^n$, $\psi^\theta(t,x)=K_t^\theta x$ with $K_t^\theta=-R^{-1}B^\top P_t^\theta$. Now by Lemma 2.3.1 and the boundedness of Θ, there exists a constant $C\geq 0$ such that $\|P^\theta\|_{C([0,T;\mathbb{R}^{n\times n})}\leq C$ and $\|P^\theta-P^{\theta^\star}\|_{C([0,T;\mathbb{R}^{n\times n})}\leq C|\theta-\theta^\star|$ for all $\theta\in\Theta\cup\{\theta^\star\}$, which along with $K_t^\theta=-R^{-1}B^\top P_t^\theta$ implies that $\|K^\theta\|_{C([0,T;\mathbb{R}^{d\times n})}\leq C$ and $\|K^\theta-K^{\theta^\star}\|_{C([0,T;\mathbb{R}^{d\times n})}\leq C|\theta-\theta^\star|$. Moreover, observe from (2.2.5) and (2.2.11) that $X_0^{\theta^\star}=X_0^{\psi^\theta}$ and for all $t\in[0,T]$,

$$\mathrm{d}(X^{\psi^{\theta^\star}}-X^{\psi^\theta})_t=\Big((A^\star+B^\star K_t^{\theta^\star})(X^{\psi^{\theta^\star}}-X^{\psi^\theta})_t+B^\star(K_t^{\theta^\star}-K_t^\theta)X_t^{\psi^\theta}\Big)\,\mathrm{d}t,$$

which combined with the boundedness of $K^{\theta^\star}$ and Gronwall's inequality leads to

$$\begin{aligned}\|X^{\psi^{\theta^\star}}-X^{\psi^\theta}\|_{\mathcal{H}^2(\mathbb{R}^n)}&\leq C\|X^{\psi^{\theta^\star}}-X^{\psi^\theta}\|_{\mathcal{S}^2(\mathbb{R}^n)}\\&\leq C\|(K^{\theta^\star}-K^\theta)X^{\psi^\theta}\|_{\mathcal{H}^2(\mathbb{R}^d)}\leq C\|K^{\theta^\star}-K^\theta\|_{C([0,T;\mathbb{R}^{d\times n})}\|X^{\psi^\theta}\|_{\mathcal{H}^2(\mathbb{R}^n)}\\&\leq C|\theta-\theta^\star|,\quad\forall\theta\in\Theta,\end{aligned}$$

where the last inequality follows from $\|X^{\psi^\theta}\|_{\mathcal{H}^2(\mathbb{R}^n)}\leq C$, as K^θ is uniformly bounded. The above inequality further implies

$$\begin{aligned}&\|\psi^\theta(\cdot,X_\cdot^{\psi^\theta})-\psi^{\theta^\star}(\cdot,X_\cdot^{\psi^{\theta^\star}})\|_{\mathcal{H}^2(\mathbb{R}^d)}=\|K_\cdot^\theta X_\cdot^{\psi^\theta}-K_\cdot^{\theta^\star}X_\cdot^{\psi^{\theta^\star}}\|_{\mathcal{H}^2(\mathbb{R}^d)}\\&\leq\|(K_\cdot^\theta-K_\cdot^{\theta^\star})X_\cdot^{\psi^\theta}\|_{\mathcal{H}^2(\mathbb{R}^d)}+\|K_\cdot^{\theta^\star}(X_\cdot^{\psi^\theta}-X_\cdot^{\psi^{\theta^\star}})\|_{\mathcal{H}^2(\mathbb{R}^d)}\\&\leq\|K^{\theta^\star}-K^\theta\|_{C([0,T;\mathbb{R}^{d\times n})}\|X^{\psi^\theta}\|_{\mathcal{H}^2(\mathbb{R}^n)}+\|K^{\theta^\star}\|_{C([0,T;\mathbb{R}^{d\times n})}\|X^{\psi^\theta}-X^{\psi^{\theta^\star}}\|_{\mathcal{H}^2(\mathbb{R}^n)}\\&\leq C|\theta-\theta^\star|,\quad\forall\theta\in\Theta,\end{aligned}$$

which along with (2.3.8) finishes the desired estimate. □

Step 2: Error bound for parameter estimation.

Proposition 2.3.9. *Suppose (H.1(1)) holds and let $\Theta\subset\mathbb{R}^{(n+d)\times n}$ such that there exists $C_1>0$ satisfying $\|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2\leq C_1$ and $|\theta|\leq C_1$ for all $\theta\in\Theta$, with V^{ψ^θ} defined in (2.3.1). Then there exist constants $\bar{C}_1,\bar{C}_2\geq 0$, such that for all $\theta\in\Theta$ and $\delta\in(0,1/2)$, if $m\geq\bar{C}_1(-\ln\delta)$, then with probability at least $1-2\delta$,*

$$|\hat{\theta}-\theta^\star|\leq\bar{C}_2\bigg(\sqrt{\frac{-\ln\delta}{m}}+\frac{-\ln\delta}{m}+\frac{(-\ln\delta)^2}{m^2}\bigg),\tag{2.3.9}$$

where $\hat{\theta}$ denotes the right-hand side of (2.2.13) with the control ψ^θ.

Proof. Let us fix $\delta\in(0,1/2)$ and $\theta\in\Theta$. By (2.2.12) and (2.2.13), we obtain

$$\begin{aligned}\|\hat{\theta}-\theta^\star\|_2&=\|(V^{\psi^\theta,m}+\tfrac{1}{m}I)^{-1}Y^{\psi^\theta,m}-(\mathbb{E}[V^{\psi^\theta}])^{-1}\mathbb{E}[Y^{\psi^\theta}]\|_2\\&\leq\|(V^{\psi^\theta,m}+\tfrac{1}{m}I)^{-1}-(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2\|Y^{\psi^\theta,m}\|_2+\|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2\|Y^{\psi^\theta,m}-\mathbb{E}[Y^{\psi^\theta}]\|_2.\end{aligned}$$

As $E^{-1} - F^{-1} = F^{-1}(F - E)E^{-1}$ for all nonsingular matrices E and F, we have

$$\begin{aligned}
&\|\hat{\theta} - \theta^\star\|_2 \\
&\le \|(V^{\psi^\theta,m} + \tfrac{1}{m}I)^{-1}\|_2 \|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2 \|Y^{\psi^\theta,m}\|_2 \|V^{\psi^\theta,m} - \mathbb{E}[V^{\psi^\theta}] + \tfrac{1}{m}I\|_2 \\
&\quad + \|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2 \|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2 \\
&\le C_1\big(\|(V^{\psi^\theta,m} + \tfrac{1}{m}I)^{-1}\|_2 \|Y^{\psi^\theta,m}\|_2 \|V^{\psi^\theta,m} - \mathbb{E}[V^{\psi^\theta}] + \tfrac{1}{m}I\|_2 + \|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2\big),
\end{aligned} \tag{2.3.10}$$

where the last inequality follows from the assumption $\|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2 \le C_1$.

We now estimate each term in the right-hand side of (2.3.10), and denote by C a generic constant independent of $\theta \in \Theta, \delta \in (0, 1/2), m \in \mathbb{N}$. By Theorem 2.3.6, with probability at least $1 - 2\delta$, $\|V^{\psi^\theta,m} - \mathbb{E}[V^{\psi^\theta}]\|_2 \le \delta_m$ and $\|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2 \le \delta_m$, with the constant δ_m given by

$$\delta_m := C \max\left\{\left(\frac{-\ln \delta}{m}\right)^{\frac{1}{2}}, \frac{-\ln \delta}{m}\right\}. \tag{2.3.11}$$

Let m be a sufficiently large constant satisfying $\delta_m + 1/m \le 1/(2C_1)$, where C_1 is the constant such that $\|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2 \le C_1$ for all $\theta \in \Theta$. Then with probability at least $1 - 2\delta$, $\|V^{\psi^\theta,m} - \mathbb{E}[V^{\psi^\theta}] + \frac{1}{m}I\|_2 \le \delta_m + \frac{1}{m} \le \frac{1}{2C_1}$, which in turn yields

$$\lambda_{\min}(V^{\psi^\theta,m} + \tfrac{1}{m}) \ge \lambda_{\min}(\mathbb{E}[V^{\psi^\theta}]) - \|V^{\psi^\theta,m} - \mathbb{E}[V^{\psi^\theta}] + \tfrac{1}{m}I\|_2 \ge \tfrac{1}{2C_1},$$

or equivalently $\|(V^{\psi^\theta,m} + \frac{1}{m})^{-1}\|_2 \le 2C_1$. Moreover, the continuity of $\mathbb{R}^{(n+d)\times n} \ni \theta \mapsto \mathbb{E}[Y^{\psi^\theta}] \in \mathbb{R}$ implies $\|Y^{\psi^\theta,m}\|_2 \le \|\mathbb{E}[Y^{\psi^\theta}]\|_2 + \|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2 \le C + \|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2$. Hence, by (2.3.10),

$$\begin{aligned}
&|\hat{\theta} - \theta^\star| \\
&\le C\big((1 + \|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2)\|V^{\psi^\theta,m} - \mathbb{E}[V^{\psi^\theta}] + \tfrac{1}{m}I\|_2 + \|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2\big) \\
&\le C\left((\delta_m + \tfrac{1}{m})(1 + \delta_m) + \delta_m\right) \le C\left(\delta_m + \delta_m^2 + \tfrac{1}{m}\right).
\end{aligned}$$

Substituting (3.3.15) into the above estimate yields the desired estimate (3.3.13). As $\delta \in (0, 1/2)$, it is clear that $\delta_m + 1/m \le 1/(2C_1)$ is satisfied for all $m \ge \bar{C}_1(-\ln \delta)$, with a sufficiently large C_1. □

Step 3: Proof of Theorem 2.2.2. The following proposition shows that for any given $\theta = (A, B)^\top \in \mathbb{R}^{(n+d)\times n}$, the full row rank of K^θ is equivalent to the well-definedness of (2.2.12) for all θ' sufficiently close to θ.

Proposition 2.3.10. *Suppose (H.1(1)) holds. For each $\theta \in \mathbb{R}^{(n+d)\times n}$, let V^{ψ^θ} be defined in (2.3.1). Then for any $\theta = (A, B)^\top \in \mathbb{R}^{(n+d)\times n}$, the following properties are equivalent:*

(1) $\{v \in \mathbb{R}^d \mid (K_t^\theta)^\top v = 0,\ \forall t \in [0,T]\} = \{0\}$, *with* K^θ *defined in* (2.2.9)*;*

(2) $\mathbb{E}[V^{\psi^\theta}] \in \mathbb{S}_+^{n+d}$*;*

(3) there exist $\lambda_0, \varepsilon > 0$ *such that* $\lambda_{\min}(\mathbb{E}[V^{\psi^{\theta'}}]) \geq \lambda_0$ *for all* $\theta' \in \Phi_\varepsilon := \{\theta' \in \mathbb{R}^{(n+d)\times n} \mid |\theta' - \theta| \leq \varepsilon\}$*, where* $\lambda_{\min}(Z)$ *is the minimum eigenvalue of* $Z \in \mathbb{S}_0^{n+d}$*.*

Proof. For (1) $\Longrightarrow$ (2): By (2.3.1), $\mathbb{E}[V^{\psi^\theta}] \in \mathbb{S}_+^{n+d}$ if and only if there exists no nonzero $v \in \mathbb{R}^{n+d}$ such that

$$\mathbb{E}\left[\int_0^T v^\top Z_t^{\psi^\theta}(Z_t^{\psi^\theta})^\top v\,\mathrm{d}t\right] = \int_0^T v^\top \begin{pmatrix} I \\ K_t^\theta \end{pmatrix} \mathbb{E}\left[X_t^{\psi^\theta}(X_t^{\psi^\theta})^\top\right] \begin{pmatrix} I & (K_t^\theta)^\top \end{pmatrix} v\,\mathrm{d}t = 0, \quad (2.3.12)$$

where we applied Fubini's theorem for the first identity. By (2.2.5), $X_t^{\psi^\theta} = \Phi_t^\theta\big(x_0 + \int_0^t (\Phi_s^\theta)^{-1}\,\mathrm{d}W_s\big)$ for all $t \in [0,T]$, where $\Phi^\theta \in C([0,T];\mathbb{R}^{n\times n})$ is the fundamental solution of $\mathrm{d}\Phi_t^\theta = (A^\star + B^\star K_t^\theta)\Phi_t^\theta\,\mathrm{d}t$, $K_t^\theta = -R^{-1}B^\top P_t^\theta$ for all $t \in [0,T]$, and P^θ satisfies (2.2.10). Hence,

$$\mathbb{E}\left[X_t^{\psi^\theta}(X_t^{\psi^\theta})^\top\right] = \Phi_t^\theta\left(x_0x_0^\top + \int_0^t (\Phi_s^\theta)^{-1}((\Phi_s^\theta)^{-1})^\top\,\mathrm{d}s\right)(\Phi_t^\theta)^\top \in \mathbb{S}_0^n, \quad \forall t \in [0,T].$$

Then by (2.3.12) and the continuity of $t \mapsto \mathbb{E}\left[X_t^{\psi^\theta}(X_t^{\psi^\theta})^\top\right]$ and $t \mapsto K_t^\theta$, $\mathbb{E}[V^{\psi^\theta}] \in \mathbb{S}_+^{n+d}$ if and only if there exists no nonzero $v \in \mathbb{R}^{n+d}$ such that

$$v^\top \begin{pmatrix} I \\ K_t^\theta \end{pmatrix} \Phi_t^\theta\left(x_0x_0^\top + \int_0^t (\Phi_s^\theta)^{-1}((\Phi_s^\theta)^{-1})^\top\,\mathrm{d}s\right)(\Phi_t^\theta)^\top \begin{pmatrix} I & (K_t^\theta)^\top \end{pmatrix} v = 0, \quad \forall t \in [0,T],$$

where I is the $n \times n$ identity matrix. One can easily deduce by the invertibility of $(\Phi_t^\theta)^{-1}$ for all $t \in [0,T]$ that $\int_0^t (\Phi_s^\theta)^{-1}((\Phi_s^\theta)^{-1})^\top\,\mathrm{d}s \in \mathbb{S}_+^n$ for all $t > 0$, which subsequently shows that $\mathbb{E}[V^{\psi^\theta}] \in \mathbb{S}_+^{n+d}$ if and only if there exists no nonzero $\tilde{v} \in \mathbb{R}^{n+d}$ such that $\begin{pmatrix} I & (K_t^\theta)^\top \end{pmatrix}\tilde{v} = 0$ for all $t \in [0,T]$. Now let us denote without loss of generality that $\tilde{v} = \left(\begin{smallmatrix} u \\ v \end{smallmatrix}\right)$ for some $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^d$. Then the above derivation shows that $\mathbb{E}[V^{\psi^\theta}] \in \mathbb{S}_+^{n+d}$ is equivalent to the following statement:

$$\text{if } u \in \mathbb{R}^n \text{ and } v \in \mathbb{R}^d \text{ satisfy } u + (K_t^\theta)^\top v = 0 \text{ for all } t \in [0,T], \text{ then } u = 0 \text{ and } v = 0. \quad (2.3.13)$$

By (2.2.9), $(K_t^\theta)^\top = -P_t^\theta BR^{-1}$ for all $t \in [0,T]$ and $P_T^\theta = 0$, implying that $K_T^\theta = 0$. Then (2.3.13) can be rewritten as:

$$\text{if } v \in \mathbb{R}^d \text{ satisfies } (K_t^\theta)^\top v = 0 \text{ for all } t \in [0,T], \text{ then } v = 0.$$

For (2) $\iff$ (3): Item (3) clearly implies Item (2). On the other hand, for any given $\theta, \theta' \in \mathbb{R}^{(n+d)\times n}$,

$$\mathrm{d}(X^{\psi^\theta} - X^{\psi^{\theta'}})_t = \Big((A^\star + B^\star K_t^\theta)(X^{\psi^\theta} - X^{\psi^{\theta'}})_t + B^\star(K_t^\theta - K_t^{\theta'})X_t^{\psi^{\theta'}}\Big)\,\mathrm{d}t.$$

Then, we can easily deduce from the continuity of $t \mapsto K^\theta$ (see Lemma 2.3.1) that $\mathbb{R}^{(n+d)\times n} \ni \theta \mapsto Z^{\psi^\theta} \in \mathcal{H}^2(\mathbb{R}^{(n+d)\times n})$ is continuous, which implies the continuity of $\mathbb{R}^{(n+d)\times n} \ni \theta \mapsto V^{\psi^\theta} = \mathbb{E}\big[\int_0^T Z_t^{\psi^\theta}(Z_t^{\psi^\theta})^\top \mathrm{d}t\big] \in \mathbb{S}_0^{n+d}$. Hence, by the continuity of the minimum eigenvalue function, we can conclude Item (2) from Item (3). □

The following proposition provides sufficient conditions for the nondegeneracy of K^θ.

Proposition 2.3.11. *Let $n, d \in \mathbb{N}$, $\theta = (A, B)^\top \in \mathbb{R}^{(n+d)\times n}$, $Q \in \mathbb{S}_0^n$ and $R \in \mathbb{S}_+^d$.*

(1) For all $T > 0$, if $B^\top Q B \in \mathbb{S}_+^d$, then $\{v \in \mathbb{R}^d \mid (K_t^\theta)^\top v = 0,\ \forall t \in [0,T]\} = \{0\}$.

(2) Assume that the algebraic Riccati equation $A^\top P + PA - P(BR^{-1}B^\top)P + Q = 0$ admits a unique maximal solution $P_\infty \in \mathbb{S}_+^n$. Let $K_\infty = -R^{-1}B^\top P_\infty$, and for each $T > 0$, let $P^{(T)} \in C([0,T];\mathbb{S}_0^n)$ be defined in (2.2.10). Assume that $\lim_{T\to\infty} P_0^{(T)} = P_\infty$ and $K_\infty(K_\infty)^\top \in \mathbb{S}_+^d$. Then there exists $T_0 > 0$, such that for all $T \geq T_0$, $\{v \in \mathbb{R}^d \mid (K_t^\theta)^\top v = 0,\ \forall t \in [0,T]\} = \{0\}$.

Proof. To prove Item (1), suppose that $B^\top Q B \in \mathbb{S}_+^n$ and $v \in \mathbb{R}^d$ such that $(K_t^\theta)^\top v = -P_t^\theta B R^{-1} v = 0$ for all $t \in [0,T]$, with P^θ defined in (2.2.10). Setting $u = R^{-1}v$, right multiplying (2.2.10) by Bu, and left multiplying (2.2.10) by $u^\top B^\top$ shows

$$u^\top B^\top(\tfrac{\mathrm{d}}{\mathrm{d}t}P_t^\theta)Bu + A^\top P_t^\theta Bu + u^\top B^\top P_t^\theta ABu - u^\top B^\top P_t^\theta B R^{-1} B^\top P_t^\theta Bu + u^\top B^\top Q Bu = 0.$$

As $P_t^\theta Bu = 0$ for all $t \in (0,T)$, $u^\top B^\top(\frac{\mathrm{d}}{\mathrm{d}t}P_t^\theta)Bu = u^\top B^\top P_t^\theta = 0$ for all $t \in (0,T)$, and hence $u^\top B^\top Q B u = 0$. The assumption of $B^\top Q B \in \mathbb{S}_+^n$ then gives $u = R^{-1}v = 0$, which along with the invertibility of R^{-1} shows that $v = 0$.

To prove Item (2), observe that $\lim_{T\to\infty}(-R^{-1}B^\top P_0^{(T)}) = K_\infty$. As $\lambda_{\min}(K_\infty(K_\infty)^\top) > 0$, there exists $T_0 > 0$ such that for all $T \geq T_0$, $\lambda_{\min}\Big((-R^{-1}B^\top P_0^{(T)})(-R^{-1}B^\top P_0^{(T)})^\top\Big) > 0$. Fix $T \geq T_0$ and consider $v \in \mathbb{R}^d$ such that $(K_t^\theta)^\top v = 0$ for all $t \in [0,T]$. Then the definitions of K^θ and $P^{(T)}$ imply the invertibility of $K_0^\theta(K_0^\theta)^\top$, which yields $v = (K_0^\theta(K_0^\theta)^\top)^{-1}K_0^\theta(K_0^\theta)^\top v = 0$. □

Now we are ready for the proof of Theorem 2.2.2.

Proof of Theorem 2.2.2. As (H.15) holds with $\theta^\star$ and θ_0, we can obtain from Proposition 2.3.10 that, there exist $C_1, \varepsilon > 0$ such that for all $\theta \in \Phi_\varepsilon := \{\theta \mid \mathbb{R}^{(n+d)\times n} \mid |\theta - \theta^\star| \leq \varepsilon\} \cup \{\theta_0\}$, we have $\|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2 \leq C_1$. Then by Proposition 3.3.12, there exist constants $\bar{C}_1, \bar{C}_2 \geq 1$,

such that for all $\theta \in \Theta_\varepsilon$ and $\delta' \in (0, 1/2)$, if $m \geq \bar{C}_1(-\ln \delta')$, then with probability at least $1 - 2\delta'$,

$$|\hat{\theta} - \theta^\star| \leq \bar{C}_2 \left(\sqrt{\frac{-\ln \delta'}{m}} + \frac{-\ln \delta'}{m} + \frac{(-\ln \delta')^2}{m^2} \right), \tag{2.3.14}$$

where $\hat{\theta}$ denotes the right-hand side of (2.2.13) with the control ψ^θ. In the following, we fix $\delta \in (0, 3/\pi^2)$ and $C \geq C_0$, with the constant $C_0 \in (0, \infty)$ satisfying

$$C_0 \geq \bar{C}_1 \left(\sup_{\ell \in \mathbb{N} \cup \{0\}, \delta \in (0, 3/\pi^2)} \frac{-\ln(\delta/(\ell+1)^2)}{2^\ell (-\ln \delta)} \right) \Big/ \min \left\{ \left(\frac{\varepsilon}{3\bar{C}_2} \right)^2, 1 \right\},$$

let $m_0 = C(-\ln \delta)$, and for each $\ell \in \mathbb{N} \cup \{0\}$, let $\delta_\ell = \delta/(\ell+1)^2$, $m_\ell = 2^\ell m_0$, and let $\theta_{\ell+1}$ be generated by (2.2.13) with $m = m_\ell$ and $\theta = \theta_\ell$. Note that the choices of C_0, m_ℓ, δ_ℓ ensure that $m_\ell \geq \bar{C}_1(-\ln \delta_\ell)$, and

$$\bar{C}_2 \left(\sqrt{\frac{-\ln \delta_\ell}{m_\ell}} + \frac{(-\ln \delta_\ell)}{m_\ell} + \frac{(-\ln \delta_\ell)^2}{m_\ell^2} \right) \leq 3\bar{C}_2 \sqrt{\frac{-\ln \delta_\ell}{m_\ell}} \leq \varepsilon, \quad \forall \ell \in \mathbb{N} \cup \{0\}. \tag{2.3.15}$$

We now prove with probability at least $1 - 2\sum_{\ell=0}^{\infty} \delta_\ell = 1 - \frac{\pi^2}{3}\delta$,

$$|\theta_{\ell+1} - \theta^\star| \leq \bar{C}_2 \left(\sqrt{\frac{-\ln \delta_\ell}{m_\ell}} + \frac{(-\ln \delta_\ell)}{m_\ell} + \frac{(-\ln \delta_\ell)^2}{m_\ell^2} \right), \quad \forall \ell \in \mathbb{N} \cup \{0\}. \tag{2.3.16}$$

Let us consider the induction statement for each $k \in \mathbb{N} \cup \{0\}$: with probability at least $1 - 2\sum_{\ell=0}^{k} \delta_\ell$, (2.3.16) holds for all $\ell = 0, \ldots, k$. The fact that $\theta_0 \in \Theta_\varepsilon$ and (2.3.14) yields the induction statement for $k = 0$. Now suppose that the induction statement holds for some $k \in \mathbb{N} \cup \{0\}$. Then the induction hypothesis and (3.3.16) ensure that $|\theta_\ell - \theta^\star| \leq \varepsilon$ for all $\ell = 1, \ldots, k+1$ (and hence $\theta_{k+1} \in \Theta_\varepsilon$) with probability at least $1 - 2\sum_{\ell=0}^{k} \delta_\ell$. Conditioning on this event, we can apply (2.3.14) with $\theta = \theta_{k+1}$, $\delta' = \delta_{k+1} < 1/2$ and $m = m_{k+1} \geq \bar{C}_1(-\ln \delta_{k+1})$, and deduce with probability at least $1 - 2\delta_{k+1}$ that (2.3.16) holds for the index $\ell = k+1$. Combining this with the induction hypothesis yields (2.3.16) holds for the indices $\ell = 0, \ldots, k+1$, with probability at least $1 - 2\sum_{\ell=0}^{k+1} \delta_\ell$.

Observe that for all $i \in \mathbb{N}$, Algorithm 8 generates the i-th trajectory with control ψ^{θ_ℓ} if $i \in (\sum_{j=0}^{\ell-1} m_j, \sum_{j=0}^{\ell} m_j] = (m_0(2^\ell - 1), m_0(2^{\ell+1} - 1)]$ with some $\ell \in \mathbb{N} \cup \{0\}$. Then

conditioning on the event (2.3.16), we can obtain from Proposition 2.3.8 that, for all $M \in \mathbb{N}$,

$$\begin{aligned} R(M) &\leq \sum_{\ell=0}^{\lceil \log_2(\frac{M}{m_0}+1)\rceil-1} m_\ell \Big(J^{\theta^\star}(U^{\psi^{\theta_\ell}}) - J^{\theta^\star}(U^{\theta^\star}) \Big) \leq C' \sum_{\ell=0}^{\lceil \log_2(\frac{M}{m_0}+1)\rceil-1} m_\ell |\theta_\ell - \theta^\star|^2 \\ &\leq C' m_0 + C' \sum_{\ell=0}^{\lceil \log_2(\frac{M}{m_0}+1)\rceil-1} (-\ln \delta_\ell) \Big(1 + \frac{-\ln \delta_\ell}{m_\ell} + \frac{(-\ln \delta_\ell)^3}{m_\ell^3} \Big) \\ &\leq C'(-\ln \delta) + C' \sum_{\ell=1}^{\lceil \log_2 M \rceil} \Big(2 \ln \ell - \ln \delta \Big) \leq C' \left((\ln M)(\ln \ln M) + (\ln M)(-\ln \delta) \right), \end{aligned} \tag{2.3.17}$$

with a constant C' independent of M and δ, where we have used $\sum_{\ell=1}^{n} \ln \ell = \ln(n!) \leq C' n \ln n$ due to Stirling's formula. □

2.3.4 Regret analysis of discrete-time least-squares algorithm

This section is devoted to the proof of Theorem 2.2.3. The main step is similar to the proof of Theorem 2.2.2 in Section 2.3.3. However, one needs to quantity the precise impact of the piecewise constant policies and the discrete-time observations on the performance gap and the parameter estimation error.

Step 1: Analysis of the performance gap. The following proposition shows the performance gap between applying a piecewise constant feedback control for an incorrect model and a continuous-time feedback control for the true model scales quadratically with respect to the stepsize and the parameter errors.

Proposition 2.3.12. *Suppose (H.1(1)) holds and let Θ be a bounded subset of $\mathbb{R}^{(n+d)\times n}$. For each $\theta \in \Theta$ and $N \in \mathbb{N}$, let $\psi^{\theta,\tau}$ be defined in (2.2.19) with stepsize $\tau = T/N$, let $X^{\psi^{\theta,\tau}}$ be the state process associated with $\psi^{\theta,\tau}$ (cf. (2.2.21)), let $\psi^{\theta^\star}$ be defined in (2.2.3), and let $X^{\theta^\star}$ be the state process associated with $\psi^{\theta^\star}$ (cf. (2.2.5)). Then there exists $C > 0$ such that*

$$|J^{\theta^\star}(U^{\psi^{\theta,\tau}}) - J^{\theta^\star}(U^{\theta^\star})| \leq C(N^{-2} + |\theta - \theta^\star|^2), \quad \forall \theta \in \Theta, N \in \mathbb{N}, \tag{2.3.18}$$

where $U_t^{\psi^{\theta,\tau}} = \psi^{\theta,\tau}(t, X_t^{\psi^{\theta,\tau}})$ and $U_t^{\theta^\star} = \psi^{\theta^\star}(t, X_t^{\theta^\star})$ for all $t \in [0,T]$, and $J^{\theta^\star}$ is defined in (2.2.1).

Proof. Let us fix $\theta \in \Theta$ and $N \in \mathbb{N}$. By applying Proposition 2.3.7 with $U = U^{\psi^{\theta,\tau}}$,

$$\begin{aligned} &J^{\theta^\star}(U^{\psi^{\theta,\tau}}) - J^{\theta^\star}(U^{\theta^\star}) \\ &\leq \|Q\|_2 \|X^{\theta^\star, U^{\psi^{\theta,\tau}}} - X^{\theta^\star}\|^2_{\mathcal{H}^2(\mathbb{R}^n)} + \|R\|_2 \|U^{\psi^{\theta,\tau}} - U^{\theta^\star}\|^2_{\mathcal{H}^2(\mathbb{R}^d)}, \\ &\leq \|Q\|_2 \|X^{\psi^{\theta,\tau}} - X^{\psi^{\theta^\star}}\|^2_{\mathcal{H}^2(\mathbb{R}^n)} + \|R\|_2 \|\psi^{\theta,\tau}(\cdot, X_\cdot^{\psi^{\theta,\tau}}) - \psi^{\theta^\star}(\cdot, X_\cdot^{\psi^{\theta^\star}})\|^2_{\mathcal{H}^2(\mathbb{R}^d)}, \end{aligned} \tag{2.3.19}$$

where the last inequality used the fact that $X^{\theta^\star, U^{\psi^{\theta,\tau}}} = X^{\psi^{\theta,\tau}}$ (see (2.2.11)), and the definitions of $U^{\psi^{\theta,\tau}}$ and $U^{\theta^\star}$.

We then prove that there exists a constant C, independent of θ, N, such that

$$\|X^{\psi^{\theta,\tau}} - X^{\psi^{\theta^\star}}\|_{\mathcal{H}^2(\mathbb{R}^n)} + \|\psi^{\theta,\tau}(\cdot, X^{\psi^{\theta,\tau}}_\cdot) - \psi^{\theta^\star}(\cdot, X^{\psi^{\theta^\star}}_\cdot)\|_{\mathcal{H}^2(\mathbb{R}^d)} \le C(N^{-1} + |\theta - \theta^\star|).$$

By setting $\delta X = X^{\theta^\star} - X^{\psi^{\theta,\tau}}$, we obtain from (2.2.5) and (2.2.21) that

$$\mathrm{d}\delta X_t = (A^\star \delta X_t + B^\star K_t^{\theta^\star} \delta X_t + (K_t^{\theta^\star} - K_t^{\theta,\tau}) X_t^{\psi^{\theta,\tau}})\,\mathrm{d}t, \quad t \in [0,T]; \quad \delta X_0 = 0. \tag{2.3.20}$$

Since $\|P^{\theta^\star}\|_{C([0,T];\mathbb{R}^{n\times n})} \le C$ and $K_t^{\theta^\star} = -R^{-1}B^\top P_t^{\theta^\star}$ for all $t \in [0,T]$, $\|K^{\theta^\star}\|_{C([0,T];\mathbb{R}^{d\times n})} \le C$. Moreover, by $\|P_{t_i}^{\theta,\tau}\|_2 \le C$ for all $i = 0, \dots, N$ (see Proposition 2.3.3) and (2.2.20), we have $\|K_t^{\theta,\tau}\|_2 \le C$ for all $t \in [0,T]$, which along with a moment estimate of (2.2.21) yields $\|X^{\psi^{\theta,\tau}}\|_{\mathcal{S}^2(\mathbb{R}^n)} \le C$. Thus, by applying Gronwall's inequality to (2.3.20), Lemma 2.3.1 and Proposition 2.3.3, for all $\theta \in \Theta$ and $N \in \mathbb{N}$,

$$\begin{aligned}
&\|X^{\theta^\star} - X^{\psi^{\theta,\tau}}\|_{\mathcal{H}^2(\mathbb{R}^n)} \le C\|X^{\theta^\star} - X^{\psi^{\theta,\tau}}\|_{\mathcal{S}^2(\mathbb{R}^n)} \\
&\le C\|(K_t^{\theta^\star} - K_t^{\theta,\tau})X^{\psi^{\theta,\tau}}\|_{\mathcal{H}^2(\mathbb{R}^d)} \le C \max_{t\in[0,T]} \|K_t^{\theta^\star} - K_t^{\theta,\tau}\|_2 \\
&\le C \max_{t\in[0,T]} (\|K_t^{\theta} - K_t^{\theta,\tau}\|_2 + \|K_t^{\theta^\star} - K_t^{\theta}\|_2) \le C(N^{-1} + |\theta - \theta^\star|).
\end{aligned} \tag{2.3.21}$$

The above inequality further implies

$$\begin{aligned}
&\|\psi^{\theta,\tau}(\cdot, X^{\psi^{\theta,\tau}}_\cdot) - \psi^{\theta^\star}(\cdot, X^{\psi^{\theta^\star}}_\cdot)\|_{\mathcal{H}^2(\mathbb{R}^d)} = \|K^{\theta,\tau}_\cdot X^{\psi^{\theta,\tau}}_\cdot - K^{\theta^\star}_\cdot X^{\psi^{\theta^\star}}_\cdot\|_{\mathcal{H}^2(\mathbb{R}^d)} \\
&\le \|(K^{\theta,\tau}_\cdot - K^{\theta^\star}_\cdot)X^{\psi^{\theta,\tau}}_\cdot\|_{\mathcal{H}^2(\mathbb{R}^d)} + \|K^{\theta^\star}_\cdot (X^{\psi^{\theta,\tau}}_\cdot - X^{\psi^{\theta^\star}}_\cdot)\|_{\mathcal{H}^2(\mathbb{R}^d)} \\
&\le \|K^{\theta^\star} - K^{\theta,\tau}\|_{C([0,T;\mathbb{R}^{d\times n})} \|X^{\psi^{\theta,\tau}}\|_{\mathcal{H}^2(\mathbb{R}^n)} + \|K^{\theta^\star}\|_{C([0,T;\mathbb{R}^{d\times n})} \|X^{\psi^{\theta,\tau}} - X^{\psi^{\theta^\star}}\|_{\mathcal{H}^2(\mathbb{R}^n)} \\
&\le C(N^{-1} + |\theta - \theta^\star|), \quad \forall \theta \in \Theta, N \in \mathbb{N},
\end{aligned}$$

which along with (2.3.19) finishes the desired estimate. □

Step 2: Error bound for parameter estimation. The following lemma shows that the difference between the expectations of $(V^{\psi^{\theta,\tau},\tau}, Y^{\psi^{\theta,\tau},\tau})$ and of $(V^{\psi^{\theta,\tau}}, Y^{\psi^{\theta,\tau}})$ scales linearly with respect to the stepsize.

Lemma 2.3.13. *Suppose (H.1(1)) holds and let Θ be a bounded subset of $\mathbb{R}^{(n+d)\times n}$. For each $\theta \in \Theta$ and $N \in \mathbb{N}$, let $\tau = T/N$, let $\psi^{\theta,\tau}$ be defined in (2.2.19), let $V^{\psi^{\theta,\tau}}, Y^{\psi^{\theta,\tau}}$ be defined in (2.3.1), and let $V^{\psi^{\theta,\tau},\tau}, Y^{\psi^{\theta,\tau},\tau}$ be defined in (2.3.2). Then there exists a constant C such that*

$$|\mathbb{E}[V^{\psi^{\theta,\tau},\tau} - V^{\psi^{\theta,\tau}}]| + |\mathbb{E}[Y^{\psi^{\theta,\tau},\tau} - Y^{\psi^{\theta,\tau}}]| \le CN^{-1}, \quad \forall \theta \in \Theta, N \in \mathbb{N}.$$

Proof. By (2.2.21), we have for all $i = 0, \dots, N-1$, $X^{\psi^{\theta,\tau}}_{t_{i+1}} - X^{\psi^{\theta,\tau}}_{t_i} = \int_{t_i}^{t_{i+1}} (\theta^\star)^\top Z^{\psi^{\theta,\tau}}_t \,\mathrm{d}t + W_{t_{i+1}} - W_{t_i}$, which implies

$$\mathbb{E}[V^{\psi^{\theta,\tau}} - V^{\psi^{\theta,\tau},\tau}] = \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \mathbb{E}[Z^{\psi^{\theta,\tau}}_t (Z^{\psi^{\theta,\tau}}_t)^\top - Z^{\psi^{\theta,\tau}}_{t_i} (Z^{\psi^{\theta,\tau}}_{t_i})^\top] \,\mathrm{d}t,$$
$$\mathbb{E}[Y^{\psi^{\theta,\tau}} - Y^{\psi^{\theta,\tau},\tau}] = \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \mathbb{E}[(Z^{\psi^{\theta,\tau}}_t - Z^{\psi^{\theta,\tau}}_{t_i})(Z^{\psi^{\theta,\tau}}_t)^\top \theta^\star] \,\mathrm{d}t.$$

Hence it suffices to prove that $|\mathbb{E}[Z^{\psi^{\theta,\tau}}_t (Z^{\psi^{\theta,\tau}}_t)^\top - Z^{\psi^{\theta,\tau}}_{t_i} (Z^{\psi^{\theta,\tau}}_{t_i})^\top]| \le CN^{-1}$ and $|\mathbb{E}[(Z^{\psi^{\theta,\tau}}_t - Z^{\psi^{\theta,\tau}}_{t_i})(Z^{\psi^{\theta,\tau}}_t)^\top]| \le CN^{-1}$ for all $t \in [t_i, t_{i+1}]$ and $i = 0, \dots, N-1$.

Let us fix $i = 0, \dots, N-1$ and $t \in [t_i, t_{i+1}]$. In the following, we shall omit the superscripts of $X^{\psi^{\theta,\tau}}$ and $Z^{\psi^{\theta,\tau}}$ if no confusion occurs. As $t \in [t_i, t_{i+1}]$, by (2.2.21), we have $X_t = e^{Lt} X_{t_i} + \int_{t_i}^t e^{L(t-s)} \,\mathrm{d}W_s$ with $L \coloneqq A^\star + B^\star K^{\theta,\tau}_{t_i}$. Thus,

$$\begin{aligned}
&X_t X_t^\top - X_{t_i} X_{t_i}^\top \\
&= (X_t - X_{t_i} + X_{t_i})(X_t - X_{t_i} + X_{t_i})^\top - X_{t_i} X_{t_i}^\top \\
&= (X_t - X_{t_i})(X_t - X_{t_i})^\top + X_{t_i}(X_t - X_{t_i})^\top + (X_t - X_{t_i}) X_{t_i}^\top \\
&= \left((e^{Lt} - I) X_{t_i} + \int_{t_i}^t e^{L(t-s)} \,\mathrm{d}W_s \right) \left((e^{Lt} - I) X_{t_i} + \int_{t_i}^t e^{L(t-s)} \,\mathrm{d}W_s \right)^\top \\
&\quad + X_{t_i} \left((e^{Lt} - I) X_{t_i} + \int_{t_i}^t e^{L(t-s)} \,\mathrm{d}W_s \right)^\top + \left((e^{Lt} - I) X_{t_i} + \int_{t_i}^t e^{L(t-s)} \,\mathrm{d}W_s \right) X_{t_i}^\top.
\end{aligned}$$

By taking expectations of both sides of the above identity, the martingale property of the Itô integral, and the Itô isometry,

$$\begin{aligned}
\mathbb{E}[X_t X_t^\top - X_{t_i} X_{t_i}^\top] &= (e^{Lt} - I) \mathbb{E}[X_{t_i} X_{t_i}^\top](e^{L^\top t} - I) + \int_{t_i}^t e^{L(t-s)} e^{L^\top (t-s)} \,\mathrm{d}s \\
&\quad + \mathbb{E}[X_{t_i} X_{t_i}^\top](e^{L^\top t} - I) + (e^{Lt} - I) \mathbb{E}[X_{t_i} X_{t_i}^\top] \le C(t - t_i),
\end{aligned}$$

where the last inequality follows from $\|X^{\psi^{\theta,\tau}}\|_{\mathcal{S}^2(\mathbb{R}^n)} \le C$. Since $\psi^{\theta,\tau}(t, X^{\psi^{\theta,\tau}}_t) = K^{\theta,\tau}_{t_i} X^{\psi^{\theta,\tau}}_t$ and $\|K^{\theta,\tau}_{t_i}\|_2 \le C$, one can easily show that $|\mathbb{E}[Z^{\psi^{\theta,\tau}}_t (Z^{\psi^{\theta,\tau}}_t)^\top - Z^{\psi^{\theta,\tau}}_{t_i} (Z^{\psi^{\theta,\tau}}_{t_i})^\top]| \le CN^{-1}$. Furthermore, by $X^{\psi^{\theta,\tau}}_t = e^{Lt} X^{\psi^{\theta,\tau}}_{t_i} + \int_{t_i}^t e^{L(t-s)} \,\mathrm{d}W_s$ and the identity

$$Z^{\psi^{\theta,\tau}}_t (Z^{\psi^{\theta,\tau}}_t)^\top - Z^{\psi^{\theta,\tau}}_{t_i} (Z^{\psi^{\theta,\tau}}_{t_i})^\top = (Z^{\psi^{\theta,\tau}}_t - Z^{\psi^{\theta,\tau}}_{t_i})(Z^{\psi^{\theta,\tau}}_t)^\top + Z^{\psi^{\theta,\tau}}_{t_i} (Z^{\psi^{\theta,\tau}}_t - Z^{\psi^{\theta,\tau}}_{t_i})^\top,$$

we can show that

$$\begin{aligned}
&|\mathbb{E}[(Z^{\psi^{\theta,\tau}}_t - Z^{\psi^{\theta,\tau}}_{t_i})(Z^{\psi^{\theta,\tau}}_t)^\top]| \\
&\le |\mathbb{E}[Z^{\psi^{\theta,\tau}}_t (Z^{\psi^{\theta,\tau}}_t)^\top - Z^{\psi^{\theta,\tau}}_{t_i} (Z^{\psi^{\theta,\tau}}_{t_i})^\top]| + |\mathbb{E}[Z^{\psi^{\theta,\tau}}_{t_i} (Z^{\psi^{\theta,\tau}}_t - Z^{\psi^{\theta,\tau}}_{t_i})^\top]| \\
&\le C \left(N^{-1} + \left| \mathbb{E}\left[Z^{\psi^{\theta,\tau}}_{t_i} (X^{\psi^{\theta,\tau}}_{t_i})^\top (e^{L^\top t} - I) \begin{pmatrix} I & (K^{\theta,\tau}_{t_i})^\top \end{pmatrix} \right] \right| \right) \le CN^{-1},
\end{aligned}$$

by the uniform boundedness of $\|X^{\psi^{\theta,\tau}}\|_{\mathcal{S}^2(\mathbb{R}^n)}$ and $K^{\theta,\tau}$. □

Proposition 2.3.14. *Suppose (H.1(1)) holds, and let $\Theta \subset \mathbb{R}^{(n+d)\times n}$ such that there exists $C_1 > 0$ satisfying $\|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2 \le C_1$ and $|\theta| \le C_1$ for all $\theta \in \Theta$, with V^{ψ^θ} defined in (2.3.1). Then there exist constants $\bar{C}_1, \bar{C}_2 \ge 0$ and $n_0 \in \mathbb{N}$, such that for all $\theta \in \Theta$, $N \in \mathbb{N} \cap [n_0, \infty)$ and $\delta \in (0, 1/2)$, if $m \ge \bar{C}_1(-\ln\delta)$, then with probability at least $1 - 2\delta$,*

$$|\hat{\theta} - \theta^\star| \le \bar{C}_2\left(\sqrt{\frac{-\ln\delta}{m}} + \frac{-\ln\delta}{m} + \frac{(-\ln\delta)^2}{m^2} + \frac{1}{N}\right), \tag{2.3.22}$$

where $\hat{\theta}$ denotes the right-hand side of (2.2.23) with the control $\psi^{\theta,\tau}$ and stepsize $\tau = T/N$.

Proof. We first prove that there exists $n_0 \in \mathbb{N}$ such that for all $N \in \mathbb{N} \cap [n_0, \infty)$ and $\theta \in \Theta$, $\|(\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\|_2 \le C$ for a constant $C > 0$ independent of θ and N. By (2.2.11) and (2.2.21), we have for all $\theta \in \Theta$ and $N \in \mathbb{N}$, $X_0^{\psi^\theta} = X_0^{\psi^\theta,\tau}$ and

$$\mathrm{d}(X^{\psi^\theta} - X^{\psi^\theta,\tau})_t = \Big((A^\star + B^\star K_t^\theta)(X^{\psi^\theta} - X^{\psi^\theta,\tau})_t + B^\star(K_t^\theta - K_t^{\theta,\tau})X_t^{\psi^\theta,\tau}\Big)\,\mathrm{d}t, \quad t \in [0,T].$$

Proposition 2.3.3 shows $\|K_t^\theta - K_t^{\theta,\tau}\|_2 \le CN^{-1}$ for all $t \in [0,T]$, which along with Gronwall's inequality yields $\|X^{\psi^\theta} - X^{\psi^\theta,\tau}\|_{\mathcal{S}^2(\mathbb{R}^n)} \le CN^{-1}$ for all $\theta \in \Theta$ and $N \in \mathbb{N}$. One can further prove that $\|U^{\psi^\theta} - U^{\psi^\theta,\tau}\|_{\mathcal{S}^2(\mathbb{R}^d)} \le CN^{-1}$ with $U_t^{\psi^\theta} = \psi^\theta(t, X_t^{\psi^\theta})$ and $U_t^{\psi^{\theta,\tau}} = \psi^{\theta,\tau}(t, X_t^{\psi^{\theta,\tau}})$ for all $t \in [0,T]$. Thus, we have $|\mathbb{E}[V^{\psi^\theta}] - \mathbb{E}[V^{\psi^{\theta,\tau}}]| \le CN^{-1}$, which along with $\|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2 \le C_1$ implies a uniform bound of $\|(\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\|_2$ for all sufficiently large N.

Let us fix $N \in \mathbb{N} \cap [n_0, \infty)$ and $\theta \in \Theta$ for the subsequent analysis. The invertibility of $\mathbb{E}[V^{\psi^{\theta,\tau}}]$ implies that $\theta^\star = (\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\mathbb{E}[Y^{\psi^{\theta,\tau}}]$(cf. (2.2.12)). Then by (2.2.23), we can derive the following analogues of (2.3.10):

$$\begin{aligned}
&\|\hat{\theta} - \theta^\star\|_2 \\
&= \|(V^{\psi^{\theta,\tau},\tau,m} + \tfrac{1}{m}I)^{-1}Y^{\psi^{\theta,\tau},\tau,m} - (\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\mathbb{E}[Y^{\psi^{\theta,\tau}}]\|_2 \\
&\le \|(V^{\psi^{\theta,\tau},\tau,m} + \tfrac{1}{m}I)^{-1} - (\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\|_2\|Y^{\psi^{\theta,\tau},m,\tau}\|_2 + \|(\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\|_2\|Y^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[Y^{\psi^{\theta,\tau}}]\|_2 \\
&\le \|(V^{\psi^{\theta,\tau},\tau,m} + \tfrac{1}{m}I)^{-1}\|\|(\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\|_2\|Y^{\psi^{\theta,\tau},m,\tau}\|_2\|V^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[V^{\psi^{\theta,\tau}}] + \tfrac{1}{m}I\|_2 \\
&\quad + \|(\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\|_2\|Y^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[Y^{\psi^{\theta,\tau}}]\|_2,
\end{aligned}$$

where $V^{\psi^{\theta,\tau},\tau,m}$ and $Y^{\psi^{\theta,\tau},\tau,m}$ are defined in (2.3.2). Note that

$$\begin{aligned}
\|V^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[V^{\psi^{\theta,\tau}}] + \tfrac{1}{m}I\|_2 &\le \|V^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[V^{\psi^{\theta,\tau},\tau}]\|_2 + \|\mathbb{E}[V^{\psi^{\theta,\tau},\tau}] - \mathbb{E}[V^{\psi^{\theta,\tau}}]\|_2 + \tfrac{1}{m}, \\
\|Y^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[Y^{\psi^{\theta,\tau}}]\|_2 &\le \|Y^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[Y^{\psi^{\theta,\tau},\tau}]\|_2 + \|\mathbb{E}[Y^{\psi^{\theta,\tau},\tau}] - \mathbb{E}[Y^{\psi^{\theta,\tau}}]\|_2,
\end{aligned}$$

where for both inequalities, the first term on the right-hand side can be estimated by Theorem 2.3.6 (uniformly in N), and the second term is of the magnitude $\mathcal{O}(N^{-1})$ due to Lemma 2.3.13. Hence, proceeding along the lines of the proof of Proposition 3.3.12 leads to the desired result. □

Step 3: Proof of Theorem 2.2.3. The proof follows from similar arguments as that of Theorem 2.2.2, and we only present the main steps here. As (H.15) holds with θ_0 and $\theta^\star$, we can obtain from Propositions 2.3.10 and 2.3.14 that, there exists a bounded set $\Phi_\varepsilon \subset \mathbb{R}^{(n+d)\times n}$ and constants $\bar{C}_1, \bar{C}_2 \geq 1$, $n_0 \in \mathbb{N}$ that for all $\theta \in \Phi_\varepsilon$, $N \in \mathbb{N} \cap [n_0, \infty)$ and $\delta' \in (0, 1/2)$, if $m \geq \bar{C}_1(-\ln \delta)$, then with probability at least $1 - 2\delta'$,

$$|\hat{\theta} - \theta^\star| \leq \bar{C}_2 \left(\sqrt{\frac{-\ln \delta'}{m}} + \frac{-\ln \delta'}{m} + \frac{(-\ln \delta')^2}{m^2} + \frac{1}{N} \right), \tag{2.3.23}$$

where $\hat{\theta}$ denotes the right-hand side of (2.2.23) with the control $\psi^{\theta,\tau}$ and stepsize $\tau = T/N$. Then by proceeding along the lines of the proof of Theorem 2.2.2, there exists $C_0 > 0$ and $n_0 \in \mathbb{N}$, such that for any given $\delta \in (0, \frac{3}{\pi^2})$, if $m_0 = C(-\ln \delta)$ with $C \geq C_0$ and $N_\ell \geq n_0$ for all $\ell \in \mathbb{N} \cup \{0\}$, then with probability at least $1 - \frac{\pi^2}{3}\delta$,

$$|\theta_{\ell+1} - \theta^\star| \leq \bar{C}_2 \left(\sqrt{\frac{-\ln \delta_\ell}{m_\ell}} + \frac{(-\ln \delta_\ell)}{m_\ell} + \frac{(-\ln \delta_\ell)^2}{m_\ell^2} + \frac{1}{N_\ell} \right), \quad \forall \ell \in \mathbb{N} \cup \{0\}, \tag{2.3.24}$$

where $\delta_\ell = \delta/(\ell+1)^2$ and $m_\ell = 2^\ell m_0$ for all ℓ. Consequently, we can conclude the desired regret bound from Proposition 2.3.12 (cf. (2.3.17)), with an additional term $\sum_{\ell=0}^{\ln M} m_\ell N_\ell^{-2}$ due to the time discretization errors in (2.3.18) and (2.3.24).

Chapter 3

Reinforcement learning for linear-convex models with jumps via stability analysis of feedback controls

3.1 Introduction

Reinforcement learning (RL) seeks optimal strategies to control an unknown dynamical system by interacting with the random environment through exploration and exploitation [147]. This paper studies a reinforcement learning problem for controlled linear-convex models with unknown drift parameters. The controlled dynamics are with possible jumps, the objectives are extended real-valued nonsmooth convex functions, and the learning is in an episodic setting for a finite-time horizon.

Regret analysis of RL algorithm and stability of controls. RL algorithms are in general characterized by iterations of exploitation and exploration (see e.g. [1, 112, 17]). In the model-based approach, for instance, the agent interacts with the environment via policies based on the present estimation of the unknown model parameters, and then incorporates the responses of these interactions to improve their knowledge of the system. One of the main performance criteria for RL algorithm, called *regret*, is to measure its deviation from the optimality over the learning process.

One key component in regret analysis is the Lipschitz stability of feedback controls which quantifies the mismatch between the assumed and actual models, or the stability of controls with respect to model perturbations. It is to analyze the precise derivation of a pre-computed feedback control from the optimal one, and is also known as the robustness of control policies in the learning community [112, 17, 94]).

Despite the long history of stability of controls in the control literature, its main focus in classical control theory has been restricted to the continuity of value functions and optimal open-loop controls (see e.g. [6, 169, 15, 18, 94]). Studies of high-order stability of controls

such as the Lipschitz stability, has only attracted attention very recently, largely due to its crucial importance in characterizing the precise *regret order* of learning algorithms (see [112, 17, 135]). Analyzing Lipschitz stability of feedback control is technically more challenging. It requires analyzing the derivatives of the value function in a suitable function space, as optimal feedback controls are usually characterized via the derivatives of the value function.

Due to this technical difficulty, most existing works on regret analysis of RL algorithms concentrate on the linear-quadratic (LQ) control framework. In this special setting, the optimal feedback control is an affine function of state variables, whose coefficients satisfy an associated algebraic or ordinary Riccati equation. Consequently, the Lipschitz stability of feedback controls is simplified by analyzing the robustness of the Riccati equation (see e.g. [1, 112, 17]). Unfortunately, these techniques developed specifically for Riccati equations in LQ-RL problems are clearly not applicable for general RL problems (see e.g. [21, 39, 54, 107]). In particular, optimal policies are typically nonlinear in the state variable, especially with the inclusion of entropy regularization for the exploration strategy in the optimization objective (see e.g. [159, 77, 143, 135]).

Our work. This paper consists of three parts.

- It first establishes the Lipschitz stability for finite-time horizon linear-convex control problems, whose dynamics are linear jump-diffusion processes with controlled drifts and possibly degenerate additive noises, and objectives are extended real-valued lower semicontinuous convex functions. Such control problems include as special cases LQ problems with convex control constraints, sparse and switching control of linear systems, and entropy-regularized relaxed control problems (see Examples 3.2.1 and 3.2.2). It shows that these control problems admit Lipchitz continuous optimal feedback controls with linear growth in the spatial variables (Theorem 3.2.5). It further proves that the performance gap between applying feedback controls for an incorrect model and for the true model depends Lipschitz-continuously on the magnitude of perturbations in the model coefficients, even with lower semicontinuous cost functions (Theorem 3.2.7). The Lipschitz stability of feedback controls is extended to entropy-regularized control problems with controlled diffusion in Proposition 3.4.1.

- It then proposes a greedy least-squares (GLS) algorithm for a class of continuous-time linear-convex RL problems in an episodic setting. At each iteration, the GLS algorithm estimates the unknown drift parameters by a regularized least-squares estimator based on observed trajectories, and then designs a feedback control for the estimated model. It establishes that the regret of this GLS algorithm is sublinear, i.e., of the magnitude $\mathcal{O}(\sqrt{N \ln N})$ with N being the number of learning episodes, provided that the least-squares estimator satisfies a general concentration inequality (Theorem 3.3.2). It further characterizes the explicit concentration behaviour of the least-squares estimator (and hence the precise regret bound of the GLS algorithm), depending on tail behaviours of the random

jumps in the state dynamics (Theorem 3.3.3). In the pure diffusion case, a sharper regret bound has been obtained (Theorem 3.3.4).

- It finally verifies the theoretical properties of the proposed GLS algorithm through numerical experiment on a three-dimensional LQ RL problem. It shows the convergence of the least-squares estimations to the true parameters as the number of episodes increases, as well as a sublinear regret as indicated in theoretical results. It also demonstrates the GLS algorithm is robust with respect to initializations.

Our approaches and related works. Optimal control of stochastic systems with parametric uncertainty has been studied in the classical adaptive control literature (see [50, 139, 89, 12]), where stationary policy is constructed to minimize the long term average cost and where the *asymptotic* stability and convergence of an adaptive control law is analyzed when the time horizon goes to infinity. However, research on rate of convergence is virtually non-existent. The problem studied here is different. The main objective is to construct optimal (and time-dependent) policies for *finite-horizon* problems, with the *finite-sample regret* analysis for the learning algorithm. Compared with the classical adaptive control literature, the regret analysis in this work, also known as the non-asymptotic performance analysis, requires novel techniques, consisting of a precise performance estimate of a greedy policy (namely the Lipschitz stability of feedback controls) and a finite-sample analysis of the parameter estimation scheme.

Analyzing the Lipschitz stability of feedback controls in a continuous-time setting requires quantifying the impact of parameter uncertainty on the derivatives of the value functions. [135] studies the so-called exit time problem and the Lipschitz stability of regularized relaxed controls of diffusion processes via a partial differential equation (PDE) approach, which assumes that the diffusion coefficients are non-degenerate and the state process takes values in a compact set. In contrast, we consider (see Section 3.2) unconstrained jump-diffusion process with unbounded drift and (uncontrolled) degenerate noise, and the cost functions are nonsmooth and unbounded. Consequently, the PDE approach requires to deal with a *degenerate nonlocal* PDE with non-Lipschitz nonlinearity, whose solution (i.e., the value function) is unbounded and may be nonsmooth due to the lack of regularization from the Laplacian operator. Here the Lipschitz stability of feedback controls is established by analyzing the stability of the associated coupled forward-backward stochastic differential equations (FBSDEs). This is possible by a) first exploiting the linear-convex structure of the control problem, which enables constructing a Lipschitz continuous feedback control via solutions of coupled FBSDEs, and then b) by extending the stochastic maximum principle in [151] to feedback controls with nonsmooth costs. To the best of our knowledge, this is the first time FBSDE has been used to study stability of *feedback* controls.

Analyzing the (finite-sample) accuracy of the least-squares estimator for jump-diffusion models involves integrations of the state and control processes with respect to Brownian motions and Poisson random measures. Now, the nonlinearity of feedback controls renders it impossible to analyze the tail behaviour of these stochastic integrals as [77] does for

LQ problems with analytical solutions; Additionally, the presence of random jumps implies that the state process is no longer sub-Gaussian, and hence the stochastic integrals in the least-squares estimator no longer sub-exponential. To overcome these difficulties, a convex concentration inequality is employed for SDEs with jumps [111], along with Burkholder's inequality and the Girsanov theorem to characterize precisely the sub-Weibull behaviour of the required stochastic integrals in terms of their Orlicz norms (Lemmas 3.3.6 and 3.3.7). Leveraging recent developments in the theory of sub-Weibull random variables, the precise parameter estimation error of the least-squares estimator is quantified in terms of the sample size.

It is worth pointing out that the stability analysis of feedback controls can be extended (see Section 3.4) to entropy-regularized control problems with controlled diffusion and without the linear-convex structure. Instead of the maximum principle for the linear-convex setting, regularity analysis of the associated fully-nonlinear parabolic PDEs may be needed for nondegenerate noise with regular (such as bounded and high-order differentiable) coefficients. (See the discussion after Proposition 3.4.1 for more details).

Notation. For each $T > 0$, filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} = \{\mathcal{F}_t\}_{t\in[0,T]}, \mathbb{P})$ satisfying the usual condition and Euclidean space $(E, |\cdot|)$, we introduce the following spaces:

- $L^p(0,T;E)$, $p \in [2,\infty]$, is the space of (Borel) measurable functions $\phi : [0,T] \to E$ satisfying $\|\phi\|_{L^p} = (\int_0^T |\phi_t|^p \,\mathrm{d}t)^{1/p} < \infty$ if $p \in [2,\infty)$ and $\|\phi\|_{L^\infty} = \operatorname{ess\,sup}_{t\in[0,T]} |\phi_t| < \infty$ if $p = \infty$;
- $L^2(\Omega; E)$ is the space of E-valued $\mathcal{F}$-measurable random variables X satisfying $\|X\|_{L^2} = \mathbb{E}[|X|^2]^{1/2} < \infty$;
- $\mathcal{S}^2(t,T;E)$, $t \in [0,T]$, is the space of E-valued $\mathbb{F}$-progressively measurable càdlàg processes $Y : \Omega \times [t,T] \to E$ satisfying $\|Y\|_{\mathcal{S}^2} = \mathbb{E}[\sup_{s\in[t,T]} |Y_s|^2]^{1/2} < \infty$;
- $\mathcal{H}^2(t,T;E)$, $t \in [0,T]$, is the space of E-valued $\mathbb{F}$-progressively measurable processes $Z : \Omega \times [t,T] \to E$ satisfying $\|Z\|_{\mathcal{H}^2} = \mathbb{E}[\int_t^T |Z_s|^2 \,\mathrm{d}s]^{1/2} < \infty$;
- $\mathcal{H}^2_\nu(t,T;E)$, $t \in [0,T]$, is the space of E-valued $\mathbb{F}$-progressively measurable processes $M : \Omega \times [t,T] \times \mathbb{R}^p_0 \to E$ satisfying $\|M\|_{\mathcal{H}^2_\nu} = \mathbb{E}[\int_t^T \int_{\mathbb{R}^p_0} |M_s(u)|^2 \nu(\mathrm{d}u)\,\mathrm{d}s]^{1/2} < \infty$, where $\mathbb{R}^p_0 := \mathbb{R}^p \setminus \{0\}$ and ν is a σ-finite measure on $\mathbb{R}^p_0$.

For notational simplicity, we denote $\mathcal{S}^2(E) = \mathcal{S}^2(0,T;E)$, $\mathcal{H}^2(E) = \mathcal{H}^2(0,T;E)$ and $\mathcal{H}^2_\nu(E) = \mathcal{H}^2_\nu(0,T;E)$. We shall also denote by $\langle\cdot,\cdot\rangle$ the usual inner product in a given Euclidean space, by $|\cdot|$ the norm induced by $\langle\cdot,\cdot\rangle$, by $A^\intercal$ the transpose of a matrix A, and by $C \in [0,\infty)$ a generic constant, which depends only on the constants appearing in the assumptions and may take a different value at each occurrence.

3.2 Lipschitz stability of linear-convex control problems

3.2.1 Problem formulation with nonsmooth costs

In this section, we introduce the linear-convex control problems with nonsmooth costs.

Let $T > 0$ be a given terminal time and $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, in which two mutually independent processes, a d-dimensional Brownian motion W and a Poisson random measure $N(\mathrm{d}t, \mathrm{d}u)$ with compensator $\nu(\mathrm{d}u)\mathrm{d}t$, are defined. We assume that ν is a σ-finite measure on $\mathbb{R}_0^p$ equipped with its Borel field $\mathcal{B}(\mathbb{R}_0^p)$ and satisfies $\int_{\mathbb{R}_0^p} \min(1, |u|^2)\, \nu(\mathrm{d}u) < \infty$. We denote by $\tilde{N}(\mathrm{d}t, \mathrm{d}u) = N(\mathrm{d}t, \mathrm{d}u) - \nu(\mathrm{d}u)\mathrm{d}t$ the compensated process of N and by $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$ the filtration generated by W and N and augmented by the $\mathbb{P}$-null sets.

For any given initial state $x_0 \in \mathbb{R}^n$, we consider the following minimization problem

$$V(x_0) = \inf_{\alpha\in\mathcal{H}^2(\mathbb{R}^k)} J(\alpha; x_0), \quad \text{with} \quad J(\alpha; x_0) = \mathbb{E}\left[\int_0^T f(t, X_t^{x_0,\alpha}, \alpha_t)\, \mathrm{d}t + g(X_T^{x_0,\alpha})\right], \tag{3.2.1}$$

where for each $\alpha \in \mathcal{H}^2(\mathbb{R}^k)$, the process $X^{x_0,\alpha}$ satisfies the following controlled dynamics:

$$\mathrm{d}X_t = b(t, X_t, \alpha_t)\, \mathrm{d}t + \sigma(t)\, \mathrm{d}W_t + \int_{\mathbb{R}_0^p} \gamma(t, u)\, \tilde{N}(\mathrm{d}t, \mathrm{d}u), \quad t \in [0, T], \quad X_0 = x_0, \tag{3.2.2}$$

where b, σ, γ, f and g are given functions satisfying the following conditions:

H.2. *$b : [0,T]\times\mathbb{R}^n\times\mathbb{R}^k \to \mathbb{R}^n$, $\sigma : [0,T] \to \mathbb{R}^{n\times d}$, $\gamma : [0,T]\times\mathbb{R}_0^p \to \mathbb{R}^n$, $f : [0,T]\times\mathbb{R}^n\times\mathbb{R}^k \to \mathbb{R}\cup\{\infty\}$, $g : \mathbb{R}^n \to \mathbb{R}$ are measurable functions such that for some $L \geq 0$ and $\lambda > 0$,*

(1) there exist measurable functions $(b_0, b_1, b_2) : [0,T] \to \mathbb{R}^n \times \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times k}$ such that $b(t,x,a) = b_0(t) + b_1(t)x + b_2(t)a$ for all $(t,x,a) \in [0,T]\times\mathbb{R}^n\times\mathbb{R}^k$, with $\|b_0\|_{L^2} + \|b_1\|_{L^\infty} + \|b_2\|_{L^\infty} + \|\sigma\|_{L^2} + \big(\int_0^T \int_{\mathbb{R}_0^p} |\gamma(t,u)|^2\, \nu(\mathrm{d}u)\mathrm{d}t\big)^{1/2} \leq L$.

(2) g is convex and differentiable with an L-Lipschitz derivative such that $|\nabla g(0)| \leq L$.

(3) there exist functions $f_0 : [0,T]\times\mathbb{R}^n\times\mathbb{R}^k \to \mathbb{R}$ and $\mathcal{R} : \mathbb{R}^k \to \mathbb{R}\cup\{\infty\}$ such that

$$f(t,x,a) = f_0(t,x,a) + \mathcal{R}(a), \quad \forall (t,x,a) \in [0,T]\times\mathbb{R}^n\times\mathbb{R}^k.$$

For all $(t,x) \in [0,T]\times\mathbb{R}^n$, $f_0(t,x,\cdot)$ is convex in $\mathbb{R}^k$, $f_0(t,\cdot,\cdot)$ is differentiable in $\mathbb{R}^n\times\mathbb{R}^k$ with an L-Lipschitz derivative, and $|f_0(t,0,0)| + |\partial_{(x,a)} f_0(t,0,0)| \leq L$. Moreover, $\mathcal{R}$ is proper, lower semicontinuous, and convex. [1]

[1] We say a function $\mathcal{R} : \mathbb{R}^k \to \mathbb{R}\cup\{\infty\}$ is proper if it has a nonempty effective domain $\operatorname{dom}\mathcal{R} := \{a \in \mathbb{R}^k \mid \mathcal{R}(a) < \infty\}$.

(4) for all $t \in [0,T]$, $(x,a),(x',a') \in \mathbb{R}^n \times \mathbb{R}^k$, *and* $\eta \in [0,1]$,

$$\eta f(t,x,a) + (1-\eta) f(t,x',a') \geq f(t, \eta x + (1-\eta)x', \eta a + (1-\eta)a') + \eta(1-\eta)\tfrac{\lambda}{2}|a-a'|^2. \tag{3.2.3}$$

Remark 3.2.1. *Throughout this paper, let* $\operatorname{dom}\mathcal{R} = \{a \in \mathbb{R}^k \mid \mathcal{R}(a) < \infty\}$ *be the effective domain of* $\mathcal{R}$ *(or equivalently the effective domain of* f*). Under (H.2), we can show that both the function* f *and its conjugate function*

$$[0,T]\times\mathbb{R}^n\times\mathbb{R}^k \ni (t,x,z) \mapsto f^*(t,x,z) := \sup\{\langle a,z\rangle - f(t,x,a) \mid a \in \mathbb{R}^k\} \in \mathbb{R}\cup\{\infty\} \tag{3.2.4}$$

are normal convex integrands in the sense of [137, Section 14] and hence measurable, which are crucial for the well-definedness of the control problem (3.2.1) and the characterization of optimal controls. Furthermore, the strong convexity condition (H.2(4)) enables us to establish the Lipschitz stability of feedback controls to (3.2.1), which is essential for the analysis of learning algorithms.

Our analysis and results can be extended to control problems with time-space dependent nonsmooth cost function $\mathcal{R} : [0,T]\times\mathbb{R}^n\times\mathbb{R}^k \to \mathbb{R}\cup\{\infty\}$ *by assuming* $\mathcal{R}$ *is a normal convex integrand and satisfies suitable subdifferentability conditions. For notational simplicity and clarity, we choose to refrain from further generalization.*

Note that (H.2) allows the diffusion coefficient σ to be degenerate, hence the stability results in Section 3.2.3 apply to deterministic control problems. Moreover, (H.2) requires neither the effective domain $\operatorname{dom}\mathcal{R}$ to be closed nor the function $\mathcal{R}$ to be bounded or continuous on $\operatorname{dom}\mathcal{R}$, which is important for problems in engineering and machine learning, as shown in the following examples.

Example 3.2.1 (Sparse and switching controls)**.** *Let* $\boldsymbol{A} \subset \mathbb{R}^k$ *be a nonempty closed convex set,* $\delta_{\boldsymbol{A}}$ *be the indicator of* $\boldsymbol{A}$ *satisfying* $\delta_{\boldsymbol{A}}(x) = 0$ *for* $x \in \boldsymbol{A}$ *and* $\delta_{\boldsymbol{A}}(x) = \infty$ *for* $x \in \mathbb{R}^k \setminus \boldsymbol{A}$*, and* $\ell : \mathbb{R}^k \to \mathbb{R}$ *be a lower semicontinuous and convex function. Then* $\mathcal{R} := \ell + \delta_{\boldsymbol{A}}$ *satisfies (H.2(3)). In particular, by setting* $\ell \equiv 0$*, we can consider the linear-convex control problems with smooth running costs and control constraints (see e.g. [21] and [169, Theorem 5.2 on p. 137]), which include the most commonly used linear-quadratic models as special cases.*

More importantly, it is well-known in optimal control literature (see e.g. [39, 54, 107] and references therein) that, one can employ a nonsmooth function ℓ *involving* L^1*-norm of controls to enhance the sparsity and switching property of optimal controls, which are practically important for mimimum fuel problems and optimal device placement problems. Here by sparsity we refer to the situation where the whole vector* α_t *is zero, while by switching control we refer to the phenomena where at most one coordinate of* α_t *is non-zero at each* t*.*

Example 3.2.2 (Regularized relaxed controls)**.** *Consider a regularized control problem arising from reinforcement learning (see e.g. [159, 77, 143, 135]), whose cost function* f *is of the following form:*

$$f(t,x,a) = f_0(t,x) + \langle f_1(t,x), a\rangle + \rho D_{\mathfrak{f}}(a||\mu) \quad \forall (t,x,a) \in [0,T]\times\mathbb{R}^n\times\mathbb{R}^k, \tag{3.2.5}$$

where $f_0 : [0,T]\times\mathbb{R}^n \to \mathbb{R}$, $f_1 : [0,T]\times\mathbb{R}^n \to \mathbb{R}^k$ *are given functions,* $\rho > 0$ *is a regularization parameter, and* $D_{\mathfrak{f}}(\cdot||\mu) : \mathbb{R}^k \to \mathbb{R}\cup\{\infty\}$ *is an* $\mathfrak{f}$*-divergence defined as follows. Let* $\Delta_k := \{a \in [0,1]^k \mid \sum_{i=1}^k a_i = 1\}$, $\mu = (\mu_i)_{i=1}^k \in \Delta_k \cap (0,1)^k$, *and* $\mathfrak{f} : [0,\infty) \to \mathbb{R}\cup\{\infty\}$ *be a lower semicontinuous function which satisfies* $\mathfrak{f}(0) = \lim_{x\to 0}\mathfrak{f}(x)$, $\mathfrak{f}(1) = 0$ *and* $\mathfrak{f}$ *is* κ_μ*-strongly convex on* $[0, \frac{1}{\min_i \mu_i}]$ *with a constant* $\kappa_\mu > 0$. *Then, the* $\mathfrak{f}$*-divergence* $D_{\mathfrak{f}}(\cdot||\mu) : \mathbb{R}^k \to \mathbb{R}\cup\{\infty\}$ *satisfies* $D_{\mathfrak{f}}(a||\mu) = \infty$ *for* $a \notin \Delta_k$ *and*

$$D_{\mathfrak{f}}(a||\mu) := \sum_{i=1}^k \mu_i \mathfrak{f}\big(\tfrac{a_i}{\mu_i}\big) \in \mathbb{R}\cup\{\infty\} \quad \forall a \in \Delta_k.$$

One can easily see from $\mathfrak{f}(1) = 0$ *and the lower semicontinuity of* $\mathfrak{f}$ *that* $D_{\mathfrak{f}}(\cdot||\mu)$ *is a proper, lower semicontinuous function with effective domain* $\operatorname{dom} D_{\mathfrak{f}}(\cdot||\mu) \subset \Delta_k$. *Moreover, by the* κ_μ*-strong convexity of* $\mathfrak{f}$, *we have for all* $a, \tilde{a} \in \Delta_k$, $\eta \in [0,1]$ *that*

$$\begin{aligned}
&\eta D_{\mathfrak{f}}(a||\mu) + (1-\eta) D_{\mathfrak{f}}(\tilde{a}||\mu) \\
&= \sum_{i=1}^k \mu_i \Big(\eta \mathfrak{f}\big(\tfrac{a_i}{\mu_i}\big) + (1-\eta)\mathfrak{f}\big(\tfrac{\tilde{a}_i}{\mu_i}\big)\Big) \geq \sum_{i=1}^k \mu_i \Big(\mathfrak{f}\big(\tfrac{\eta a_i + (1-\eta)\tilde{a}_i}{\mu_i}\big) + \eta(1-\eta)\frac{\kappa_\mu}{2}\big|\tfrac{a_i - \tilde{a}_i}{\mu_i}\big|^2\Big) \\
&\geq D_{\mathfrak{f}}(\eta a + (1-\eta)\tilde{a}||\mu) + \eta(1-\eta)\frac{\kappa_\mu}{2\max_i \mu_i}|a-\tilde{a}|^2,
\end{aligned}$$

which implies the $\frac{\kappa_\mu}{\max_i \mu_i}$*-strong convexity of* $D_{\mathfrak{f}}(\cdot||\mu)$ *in* $\mathbb{R}^k$. *It is clear that for suitable choices of* f_0, f_1, *the function* f *in* (3.2.5) *satisfies (H.2(3)).*

It is important to notice that an $\mathfrak{f}$*-divergence* $D_{\mathfrak{f}}(\cdot||\mu)$ *is in general non-differentiable and unbounded on its effective domain. For example, one may consider the relative entropy (with* $\mathfrak{f}(s) = s\log s$*) and the squared Hellinger divergence (with* $\mathfrak{f}(s) = 2(1-\sqrt{s})$*), which are not subdifferentiable at the boundary of* Δ_k. *Moreover, the reverse relative entropy (with* $\mathfrak{f}(s) = -\log s$*) and the Neyman's* χ^2 *divergence (with* $\mathfrak{f}(s) = \frac{1}{s} - 1$*) are unbounded near the boundary of* Δ_k.

3.2.2 Construction of optimal feedback controls

In this section, we apply the maximum principle to (3.2.1) and explicitly construct optimal feedback controls of (3.2.1) based on the associated coupled FBSDE.

The following proposition shows that under (H.2), the control problem (3.2.1) admits a unique optimal open-loop control.

Proposition 3.2.1. *Suppose (H.2) holds and let* $x_0 \in \mathbb{R}^n$. *Then the cost functional* $J(\alpha; x_0) : \mathcal{H}^2(\mathbb{R}^k) \to \mathbb{R}\cup\{\infty\}$ *is proper, lower semicontinuous, and* λ*-strongly convex. Consequently,* $J(\cdot; x_0)$ *admits a unique minimizer* α^{x_0} *in* $\mathcal{H}^2(\mathbb{R}^k)$.

Proof. The desired properties of J follow directly from the corresponding properties of f, g in (H.2) and the fact that (3.2.2) has affine coefficients. The well-posedness of minimizers

then follows from the standard theory of strongly convex minimization problems on Hilbert spaces (see e.g. [26, Lemma 2.33 (ii)]). □

We then proceed to study optimal feedback controls of (3.2.1). The classical control theory shows that under suitable coercivity and convexity conditions, the optimal open-loop control of (3.2.1) can be expressed in a feedback form, i.e., there exists a measurable function $\psi : [0,T] \times \mathbb{R}^n \to \mathbb{R}^k$ such that $\alpha^{x_0} = \psi(t, X_t^{x_0,\alpha^{x_0}})$ for $\mathrm{d}\mathbb{P} \otimes \mathrm{d}t$ a.e. (see [123] for the case with controlled jump-diffusions and smooth costs and [81] for the case with controlled diffusions and nonsmooth costs). However, since these non-constructive proofs are based on a measurable selection theorem, the resulting feedback policy ψ may not be unique, and may be unstable with respect to perturbations of the state dynamics.

In the subsequent analysis, we give a constructive proof of the existence of Lipschitz continuous feedback controls by exploiting the linear-convex structure of the control problem (3.2.1)-(3.2.2). Such a feedback control can be explicitly represented as solutions of a suitable FBSDE, and hence is Lipschitz stable with respect to perturbations of underlying models (see Theorem 3.2.6).

We first present the precise definitions of feedback controls and the associated state processes.

Definition 3.2.1. *Let $\mathcal{V}$ be the following space of feedback controls:*

$$\mathcal{V} \coloneqq \left\{ \psi : [0,T] \times \mathbb{R}^n \to \mathbb{R}^k \;\middle|\; \begin{array}{l} \psi \text{ is measurable and there exists } C \geq 0 \text{ such that} \\ \text{for all } (t,x,y) \in [0,T] \times \mathbb{R}^n \times \mathbb{R}^n,\ |\psi(t,0)| \leq C \\ \text{and } |\psi(t,x) - \psi(t,y)| \leq C|x-y|. \end{array} \right\} \tag{3.2.6}$$

For any given $x_0 \in \mathbb{R}^n$ and $\psi \in \mathcal{V}$, we say $X^{x_0,\psi} \in \mathcal{S}^2(\mathbb{R}^n)$ is the state process associated with ψ if it satisfies the following dynamics:

$$\mathrm{d}X_t = b(t, X_t, \psi(t, X_t))\,\mathrm{d}t + \sigma(t)\,\mathrm{d}W_t + \int_{\mathbb{R}_0^p} \gamma(t,u)\,\tilde{N}(\mathrm{d}t, \mathrm{d}u),, \quad t \in [0,T], \quad X_0 = x_0. \tag{3.2.7}$$

We say $\psi \in \mathcal{V}$ is an optimal feedback control of (3.2.1) if it holds for $\mathrm{d}\mathbb{P} \otimes \mathrm{d}t$ a.e. that $\alpha_t^{x_0} = \psi(t, X_t^{x_0,\psi})$, where $\alpha^{x_0} \in \mathcal{H}^2(\mathbb{R}^n)$ is the optimal control of (3.2.1).

We then proceed to establish a maximum principle for feedback controls of the control problem (3.2.1) with non-smooth costs. Let $H : [0,T] \times \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $\phi : [0,T] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^k$ such that for all $(t,x,a,y) \in [0,T] \times \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^n$,

$$H(t,x,a,y) \coloneqq \langle b(t,x,a), y \rangle + f(t,x,a), \quad \phi(t,x,y) \coloneqq \operatorname*{arg\,min}_{a \in \mathbb{R}^k} H(t,x,a,y) \in \operatorname{dom} \mathcal{R}. \tag{3.2.8}$$

The following lemma shows that the function ϕ is well-defined and measurable.

Lemma 3.2.2. *Suppose (H.2) holds. Then the function $\phi : [0,T] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^k$ defined in (3.2.8) is measurable and satisfies for all $t \in [0,T]$, $x, y \in \mathbb{R}^n$ that*

$$\phi(t,x,y) = \partial_z f^*(t,x,-b_2(t)^\mathsf{T} y), \tag{3.2.9}$$

where the function f^* is defined in (3.2.4).

Proof. Let $f^* : [0,T] \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R} \cup \{\infty\}$ be the function defined in (3.2.4). Recall that for each $(t,x) \in [0,T] \times \mathbb{R}^n$, $f(t,x,\cdot)$ is λ-strongly convex and lower semicontinuous. Hence by [137, Theorems 11.3 and 11.8], $f^*(t,x,\cdot)$ is finite and differentiable on $\mathbb{R}^k$, and $\partial_z f^*(t,x,z) = \arg\max_{a \in \operatorname{dom} \mathcal{R}} \big(\langle a, z\rangle - f(t,x,a)\big)$ for all $z \in \mathbb{R}^k$. Moreover, by [83, Theorem E4.2.1], $\mathbb{R}^k \ni z \mapsto \partial_z f^*(t,x,z) \in \mathbb{R}^k$ is $1/\lambda$-Lipschitz continuous. Hence, from the definition of ϕ and (H.2(1)), for all $t \in [0,T]$, $x, y \in \mathbb{R}^n$,

$$\begin{aligned}\phi(t,x,y) &= \arg\min_{a \in \mathbb{R}^k} \Big(\langle b(t,x,a), y\rangle + f(t,x,a)\Big) = \arg\max_{a \in \mathbb{R}^k} \Big(\langle a, -b_2(t)^\mathsf{T} y\rangle - f(t,x,a)\Big) \\ &= \partial_z f^*(t,x,-b_2(t)^\mathsf{T} y).\end{aligned} \tag{3.2.10}$$

Note that the measurability of f^* (see Remark 3.2.1) implies that the derivative $\partial_z f^*$ is measurable, which along with the continuity of $z \mapsto \partial_z f^*(t,x,z)$ leads to the measurability of ϕ. □

With the measurable function ϕ in hand, for each $(t,x) \in [0,T] \times \mathbb{R}^n$, let us consider the following coupled FBSDE on $[t,T]$: for all $s \in [t,T]$,

$$\mathrm{d}X_s = b(s, X_s, \phi(s, X_s, Y_s))\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s + \int_{\mathbb{R}_0^p} \gamma(s,u)\,\tilde{N}(\mathrm{d}s, \mathrm{d}u), \qquad X_t = x, \tag{3.2.11a}$$

$$\mathrm{d}Y_s = -\partial_x H(s, X_s, \phi(s, X_s, Y_s), Y_s)\,\mathrm{d}s + Z_s\,\mathrm{d}W_s + \int_{\mathbb{R}_0^p} M_s\,\tilde{N}(\mathrm{d}s, \mathrm{d}u), \quad Y_T = \nabla g(X_T). \tag{3.2.11b}$$

We say a tuple of processes $(X^{t,x}, Y^{t,x}, Z^{t,x}, M^{t,x}) \in \mathbb{S}(t,T) := \mathcal{S}^2(t,T;\mathbb{R}^n) \times \mathcal{S}^2(t,T;\mathbb{R}^n) \times \mathcal{H}^2(t,T;\mathbb{R}^{n\times d}) \times \mathcal{H}^2_\nu(t,T;\mathbb{R}^n)$ is a solution to (3.2.11) (on $[t,T]$ with initial condition $X_t^{t,x} = x$) if it satisfies (3.2.11) $\mathbb{P}$-almost surely.

The next lemma presents several important properties of the Hamiltonian H and the function ϕ defined in (3.2.8), which are essential for the well-posedness and stability of (3.2.11).

Lemma 3.2.3. *Suppose (H.2) holds. Let $\phi : [0,T] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^k$ be the function defined in (3.2.8). Then there exists a constant C such that for all $t \in [0,T]$ and $(x,y), (x',y') \in \mathbb{R}^n \times \mathbb{R}^n$, $|\phi(t,0,0)| \le C$, $|\phi(t,x,y) - \phi(t,x',y')| \le C(|x-x'| + |y-y'|)$ and*

$$\begin{aligned}&\langle b(t,x,\phi(t,x,y)) - b(t,x',\phi(t,x',y')), y - y'\rangle \\ &\quad + \langle -\partial_x H(t,x,\phi(t,x,y),y) + \partial_x H(t,x',\phi(t,x',y'),y'), x - x'\rangle \\ &\le -\lambda |\phi(t,x,y) - \phi(t,x',y')|^2,\end{aligned} \tag{3.2.12}$$

with the constant λ in (H.2).

Proof. We start by showing the boundedness of $\phi(\cdot,0,0)$ by considering $a(t) \coloneqq (\partial_z f^*)(t,0,0)$ for each $t \in [0,T]$, where $f^* : [0,T] \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ is defined as in (3.2.4). The fact that $f(t,0,\cdot)$ is proper, lower semicontinuous and convex implies that $0 \in \widehat{\partial}_a f(t,0,a(t))$ for all $t \in [0,T]$, where $\widehat{\partial}_a f(t,0,a(t))$ is the subdifferential of $f(t,0,\cdot)$ at $a(t)$. Note that $f_0(t,0,\cdot)$ and $\mathcal{R}$ are proper, lower semicontinuous, and convex, and $\operatorname{dom}\mathcal{R} \subset \operatorname{dom} f_0(t,0,\cdot) = \mathbb{R}^k$. Hence by [137, Corollary 10.9], $\widehat{\partial}_a f(t,0,a) = \partial_a f_0(t,0,a) + \widehat{\partial}\mathcal{R}(a)$ for all $(t,a) \in [0,T] \times \mathbb{R}^k$, where $\widehat{\partial}\mathcal{R}(a)$ is the subdifferential of $\mathcal{R}$ at a. Now fix an arbitrary $t_0 \in [0,T]$ and set $a_0 = a(t_0)$. The fact that $0 \in \widehat{\partial}_a f(t_0,0,a_0)$ implies that $-\partial_a f_0(t_0,0,a_0) \in \widehat{\partial}\mathcal{R}(a_0)$ and hence $\partial_a f_0(t,0,a_0) - \partial_a f_0(t_0,0,a_0) \in \widehat{\partial}_a f(t,0,a_0)$ for all $t \in [0,T]$. By the strong convexity condition (3.2.3), for all $t \in [0,T]$, $\xi_1 \in \widehat{\partial}_a f(t,0,a_0)$ and $\xi_2 \in \widehat{\partial}_a f(t,0,a(t))$,

$$\lambda|a_0 - a(t)|^2 \le \langle \xi_1 - \xi_2, a_0 - a(t)\rangle \le |\xi_1 - \xi_2||a_0 - a(t)|.$$

Taking $\xi_1 = \partial_a f_0(t,0,a_0) - \partial_a f_0(t_0,0,a_0)$ and $\xi_2 = 0$ in the above inequality yields

$$|a_0 - a(t)| \le |\partial_a f_0(t,0,a_0) - \partial_a f_0(t_0,0,a_0)|/\lambda \le C,$$

by the linear growth of $\partial_a f_0(t,0,\cdot)$. This implies that $|(\partial_z f^*)(t,0,0)| \le C$ for all $t \in [0,T]$, which along with (3.2.10) leads to the desired uniform boundedness of $\phi(\cdot,0,0)$.

We proceed to establish the Lipschitz continuity of ϕ with respect to (x,y). The $1/\lambda$-Lipschitz continuity of $\partial_z f^*(t,x,\cdot)$ and the boundedness of b_2 imply that ϕ is Lipschitz continuous in y, uniformly with respect to (t,x). It remains to show the Lipschitz continuity of $\partial_z f^*$ with respect to x, which along with (3.2.10) leads to the desired Lipchitz continuity of ϕ. For any given $(t,z) \in [0,T] \times \mathbb{R}^k$ and $x, x' \in \mathbb{R}^n$, let $a = \partial_z f^*(t,x,z)$ and $a' = \partial_z f^*(t,x',z)$. Then we have $z \in \widehat{\partial}_a f(t,x,a)$ and $z \in \widehat{\partial}_a f(t,x',a')$. Moreover, by the convexity of $f(t,x,\cdot)$ for all $(t,x) \in [0,T] \times \mathbb{R}^n$ and similar arguments as above, we can show that $z - \partial_a f_0(t,x',a') + \partial_a f_0(t,x,a') \in \widehat{\partial}_a f(t,x,a')$, which together with the convexity condition (3.2.3) and $z \in \widehat{\partial}_a f(t,x,a)$ leads to

$$\lambda|a' - a| \le |z - \partial_a f_0(t,x',a') + \partial_a f_0(t,x,a') - z| \le L|x - x'|,$$

where we have used the L-Lipchitz continuity of $\partial_a f_0(t,\cdot,\cdot)$. This finishes the proof of the Lipschitz continuity of $\partial_z f^*(t,\cdot,\cdot)$ and $\phi(t,\cdot,\cdot)$.

Finally, we establish the monotonicity condition (3.2.12). By (H.2(4)), for all $t \in [0,T]$, $x, x' \in \mathbb{R}^n, a, a' \in \mathbb{R}^k, y \in \mathbb{R}^n$, the function $\mathbb{R}^n \times \mathbb{R}^k \ni (x,a) \mapsto H(t,x,a,y) \in \mathbb{R}^n \cup \{\infty\}$ satisfies the same convexity condition (3.2.3) as the function f, and hence

$$H(t,x',a',y) - H(t,x,a,y) \ge \langle \xi, x' - x, a' - a\rangle + \tfrac{\lambda}{2}|a' - a|^2 \quad \forall \xi \in \widehat{\partial}_{(x,a)} H(t,x,a,y), \tag{3.2.13}$$

where $\widehat{\partial}_{(x,a)} H(t,x,a,y)$ denotes the subdifferential of the function $H(t,\cdot,y)$ at (x,a). Moreover, for any given $t \in [0,T]$ and $x, y \in \mathbb{R}^n$, the definition of ϕ in (3.2.8) implies that

$0 \in \widehat{\partial}_a H(t,x,\phi(t,x,y),y)$, where $\widehat{\partial}_a H(t,x,\phi(t,x,y),y)$ denotes the subdifferential of the function $H(t,x,\cdot,y)$ at $\phi(t,x,y)$. Now recall that for any Euclidean space E, convex function $F : E \to \mathbb{R} \cup \{\infty\}$ and $x \in \operatorname{dom} F$, $v \in \widehat{\partial} F(x)$ if and only if $\liminf_{\tau \to 0, \tilde{w} \to w} \frac{F(x+\tau\tilde{w})-F(x)}{\tau} \geq \langle v, w \rangle$ for all $w \in E$ (see e.g., Exercise 8.4 and Proposition 8.12 in [137]). Thus, for any $t \in [0,T]$ and $x, y \in \mathbb{R}^n$, $0 \in \widehat{\partial}_a H(t,x,\phi(t,x,y),y)$ yields for all $z \in \mathbb{R}^k$,

$$\liminf_{\tau \to 0, \tilde{z} \to z} \frac{H(t,x,\phi(t,x,y)+\tau\tilde{z},y) - H(t,x,\phi(t,x,y),y)}{\tau} \geq \langle 0, z \rangle = 0. \tag{3.2.14}$$

Moreover, by the convexity of H and the continuity of $\partial_x H$ in (x,a), for any $t \in [0,T]$ and $x, y, w \in \mathbb{R}^n$ and $z \in \mathbb{R}^k$,

$$\begin{aligned} &\liminf_{\tau \to 0, (\tilde{w},\tilde{z}) \to (w,z)} \frac{H(t,x+\tau\tilde{w},\phi(t,x,y)+\tau\tilde{z},y) - H(t,x,\phi(t,x,y)+\tau\tilde{z},y)}{\tau} \\ &\geq \liminf_{\tau \to 0, (\tilde{w},\tilde{z}) \to (w,z)} \frac{\langle \partial_x H(t,x,\phi(t,x,y)+\tau\tilde{z},y), \tau\tilde{w} \rangle}{\tau} \geq \langle \partial_x H(t,x,\phi(t,x,y),y), w \rangle, \end{aligned} \tag{3.2.15}$$

provided that $\phi(t,x,y) + \tau\tilde{z} \in \operatorname{dom} \mathcal{R}$ (cf. (3.2.8)). Then for any $t \in [0,T]$ and $x, y \in \mathbb{R}^n$, adding up (3.2.14) and (3.2.15) and using the fact that $\phi(t,x,y) \in \operatorname{dom} \mathcal{R}$ give for all $(w,z) \in \mathbb{R}^n \times \mathbb{R}^k$,

$$\begin{aligned} &\liminf_{\tau \to 0, (\tilde{w},\tilde{z}) \to (w,z)} \frac{H(t,x+\tau\tilde{w},\phi(t,x,y)+\tau\tilde{z},y) - H(t,x,\phi(t,x,y),y)}{\tau} \\ &\geq \langle \partial_x H(t,x,\phi(t,x,y),y), w \rangle + \langle 0, z \rangle, \end{aligned}$$

which implies

$$(\partial_x H(t,x,\phi(t,x,y),y), 0) \subset \widehat{\partial}_{(x,a)} H(t,x,\phi(t,x,y),y). \tag{3.2.16}$$

Hence for all $t \in [0,T]$, $(x_1,y_1),(x_2,y_2) \in \mathbb{R}^n \times \mathbb{R}^n$, we can define $a_1 = \phi(t,x_1,y_1), a_2 = \phi(t,x_2,y_2)$ and deduce that

$$\begin{aligned} &\langle b(t,x_1,a_1) - b(t,x_2,a_2), y_1 - y_2 \rangle + \langle -\partial_x H(t,x_1,a_1,y_1) + \partial_x H(t,x_2,a_2,y_2), x_1 - x_2 \rangle \\ &= H(t,x_1,a_1,y_1) - H(t,x_2,a_2,y_1) - \langle \partial_x H(t,x_1,a_1,y_1), x_1 - x_2 \rangle \\ &\quad - \big(H(t,x_1,a_1,y_2) - H(t,x_2,a_2,y_2) - \langle \partial_x H(t,x_2,a_2,y_2), x_1 - x_2 \rangle \big) \\ &\leq -\lambda |a_1 - a_2|^2, \end{aligned}$$

which finishes the proof of the desired monotonicity condition. □

The following proposition shows that (3.2.11) admits a unique solution, which is Lipschitz continuous with respect to the initial state. The proof is based on the stability of (3.2.11) under the generalized monotonicity condition (3.2.12) (see Lemma 3.6.1), and follows [134, Corollary 2.4] for the case without jumps.

Proposition 3.2.4. *Suppose (H.2) holds. Then for any given $(t,x) \in [0,T] \times \mathbb{R}^n$, the FBSDE (3.2.11) admits a unique solution $(X^{t,x}, Y^{t,x}, Z^{t,x}, M^{t,x}) \in \mathbb{S}(t,T)$. Moreover, there exists a constant C such that for all $t \in [0,T]$ and $x, x' \in \mathbb{R}^n$, $\|(X^{t,x}, Y^{t,x}, Z^{t,x}, M^{t,x})\|_{\mathbb{S}(t,T)} \leq C(1+|x|)$ and $\|(X^{t,x} - X^{t,x'}, Y^{t,x} - Y^{t,x'}, Z^{t,x} - Z^{t,x'}, M^{t,x} - M^{t,x'})\|_{\mathbb{S}(t,T)} \leq C|x - x'|$.*

Now we are ready to present the main result of this section, which constructs an optimal feedback control of (3.2.1) based on the Hamiltonian (3.2.8) and the solutions to the FBSDE (3.2.11).

Theorem 3.2.5. *Suppose (H.2) holds. Let $\psi : [0,T] \times \mathbb{R}^n \to \mathbb{R}^k$ be the function defined as*

$$\psi(t,x) \coloneqq \phi(t,x,Y_t^{t,x}), \quad (t,x) \in [0,T] \times \mathbb{R}^n, \tag{3.2.17}$$

where the function ϕ is defined in (3.2.8). Then there exists a constant C such that $|\psi(t,0)| \leq C$ and $|\psi(t,x) - \psi(t,x')| \leq C|x - x'|$ for all $t \in [0,T]$, $x, x' \in \mathbb{R}^n$. Moreover, for all $x_0 \in \mathbb{R}^n$, ψ is an optimal feedback control of (3.2.1).

Proof. We first analyze the mapping $[0,T] \times \mathbb{R}^n \ni (t,x) \mapsto v(t,x) \coloneqq Y_t^{t,x} \in \mathbb{R}^n$. Note by Proposition 3.2.4, for any given $(t,x) \in [0,T] \times \mathbb{R}^n$, the solution to (3.2.11) (with initial time t and initial state x) is pathwise unique and Lipschitz continuous with respect to the initial state $x \in \mathbb{R}^n$. Hence, it is well-known that (see e.g., Theorem 3.1 and Remarks 3.2-3.3 in [108]) that the map v can be identified with a deterministic function in the space $\mathcal{V}$ and it holds for all $(t,x) \in [0,T] \times \mathbb{R}^n$ that $\mathbb{P}(\forall s \in [t,T], Y_s^{t,x} = v(s, X_s^{t,x})) = 1$. Thus, from the regularity of ϕ and v, $|\psi(t,0)| \leq C$ and $|\psi(t,x) - \psi(t,x')| \leq C|x - x'|$ for all $x, x' \in \mathbb{R}^n$, i.e., ψ is in the space $\mathcal{V}$.

Now let $x_0 \in \mathbb{R}^n$ be a given initial state and $\tilde{\alpha} \in \mathcal{A}$ satisfy for $\mathrm{d}\mathbb{P} \otimes \mathrm{d}t$ a.e. that $\tilde{\alpha}_t = \phi(t, X_t^{0,x_0}, Y_t^{0,x_0})$. Then for $\mathrm{d}\mathbb{P} \otimes \mathrm{d}t$ a.e., $\tilde{\alpha}_t = \phi(t, X_t^{0,x_0}, v(t, X_t^{0,x_0})) = \psi(t, X_t^{0,x_0})$, and X^{0,x_0} is the solution to (3.2.2) controlled by $\tilde{\alpha}$, because (X^{0,x_0}, Y^{0,x_0}) satisfy (3.2.11a). Since the control problem (3.2.1) admits an unique optimal control in $\mathcal{H}^2(\mathbb{R}^k)$, it suffices to show that $\tilde{\alpha}$ is optimal. By (3.2.16), for $\mathrm{d}\mathbb{P} \otimes \mathrm{d}t$ a.e.,

$$(\partial_x H(t, X_t^{0,x_0}, \phi(t, X_t^{0,x_0}, Y_t^{0,x_0}), Y_t^{0,x_0}), 0) \subset \widehat{\partial}_{(x,a)} H(t, X_t^{0,x_0}, \phi(t, X_t^{0,x_0}, Y_t^{0,x_0}), Y_t^{0,x_0}).$$

Then for any given $\alpha \in \mathcal{H}^2(\mathbb{R}^n)$ with the state process $X^{x_0,\alpha}$ satisfying the controlled dynamics (3.2.2), by the definition of H in (3.2.8), (H.2(2)) and (3.2.13),

$$\begin{aligned}
&J(\alpha; x_0) - J(\tilde{\alpha}; x_0) \\
&= \mathbb{E}\left[g(X_T^{x_0,\alpha}) - g(X_T^{0,x_0}) + \int_0^T (H(t, X_t^{x_0,\alpha}, \alpha_t, Y_t^{0,x_0}) - H(t, X_t^{0,x_0}, \tilde{\alpha}_t, Y_t^{0,x_0}))\, \mathrm{d}t \right] \\
&\quad - \int_0^T \langle b(t, X_t^{x_0,\alpha}, \alpha_t) - b(t, X_t^{0,x_0}, \tilde{\alpha}_t), Y_t^{0,x_0} \rangle\, \mathrm{d}t \Big] \\
&\geq \mathbb{E}\Big[\langle \nabla g(X_T^{0,x_0}), X_T^{x_0,\alpha} - X_T^{0,x_0} \rangle + \int_0^T \langle \partial_x H(t, X_t^{0,x_0}, \phi(t, X_t^{0,x_0}, Y_t^{0,x_0}), Y_t^{0,x_0}), X_t^{x_0,\alpha} - X_t^{0,x_0} \rangle\, \mathrm{d}t \\
&\quad - \int_0^T \langle b(t, X_t^{x_0,\alpha}, \alpha_t) - b(t, X_t^{0,x_0}, \tilde{\alpha}_t), Y_t^{0,x_0} \rangle\, \mathrm{d}t \Big] = 0,
\end{aligned}$$

where the last equality is by applying Itô's formula to the process $(\langle X_t^{x_0,\alpha} - X_t^{0,x_0}, Y_t^{0,x_0}\rangle)_{t\geq 0}$ and by the FBSDE (3.2.11). That is, $\psi \in \mathcal{V}$ is an optimal feedback control of (3.2.1). □

3.2.3 Lipschitz stability of optimal feedback controls and associated costs

In this section, we establish the Lipschitz stability of the optimal feedback controls constructed in Theorem 3.2.5 and their associated costs: that is, they are are Lipschitz continuous with respect to the perturbation in the coefficients of (3.2.2). Such a Lipschitz stability property is crucial for the subsequent analysis of learning algorithms.

More precisely, for any given $x_0 \in \mathbb{R}^n$, we consider a perturbed control problem where the cost functions f, g are the same as those in (3.2.1), and for each $\alpha \in \mathcal{H}^2(\mathbb{R}^n)$, the corresponding state dynamics satisfies the following perturbed dynamics:

$$\mathrm{d}X_t = \tilde{b}(t, X_t, \alpha_t)\,\mathrm{d}t + \tilde{\sigma}(t)\,\mathrm{d}W_t + \int_{\mathbb{R}_0^p} \tilde{\gamma}(t,u)\,\tilde{N}(\mathrm{d}t, \mathrm{d}u), \quad t \in [0,T], \quad X_0 = x_0, \tag{3.2.18}$$

whose coefficients satisfy the following assumption:

H. 3. *$\tilde{b} : [0,T] \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$, $\tilde{\sigma} : [0,T] \to \mathbb{R}^{n\times d}$ and $\tilde{\gamma} : [0,T] \times \mathbb{R}_0^p \to \mathbb{R}^n$ satisfy (H.2(1)) with the same constant L, i.e., there exist measurable functions $(\tilde{b}_0, \tilde{b}_1, \tilde{b}_2) : [0,T] \to \mathbb{R}^n \times \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times k}$ such that $\tilde{b}(t,x,a) = \tilde{b}_0(t) + \tilde{b}_1(t)x + \tilde{b}_2(t)a$ for all $(t,x,a) \in [0,T]\times\mathbb{R}^n\times\mathbb{R}^k$ and $\|\tilde{b}_0\|_{L^2} + \|\tilde{b}_1\|_{L^\infty} + \|\tilde{b}_2\|_{L^\infty} + \|\tilde{\sigma}\|_{L^2} + \big(\int_0^T \int_{\mathbb{R}_0^p} |\tilde{\gamma}(t,u)|^2\,\nu(\mathrm{d}u)\mathrm{d}t\big)^{1/2} \leq L$.*

Under (H.2) and (H.3), Theorem 3.2.5 ensures that an optimal feedback control of the perturbed control problem can be obtained by

$$[0,T] \times \mathbb{R}^n \ni (t,x) \mapsto \tilde{\psi}(t,x) \coloneqq \tilde{\phi}(t,x,\tilde{Y}_t^{t,x}) \in \mathbb{R}^k, \tag{3.2.19}$$

where $\tilde{\phi} : [0,T] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^k$ satisfies for all $(t,x,a,y) \in [0,T] \times \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^n$ that

$$\tilde{\phi}(t,x,y) \coloneqq \operatorname*{arg\,min}_{a\in\mathbb{R}^k} \tilde{H}(t,x,a,y), \quad \tilde{H}(t,x,a,y) \coloneqq \langle \tilde{b}(t,x,a), y\rangle + f(t,x,a), \tag{3.2.20}$$

and for each $(t,x) \in [0,T] \times \mathbb{R}^n$, $(\tilde{X}^{t,x}, \tilde{Y}^{t,x}, \tilde{Z}^{t,x}, \tilde{M}^{t,x}) \in \mathbb{S}(t,T)$ is the solution to the following perturbed FBSDE: for all $s \in [t,T]$,

$$\begin{aligned}
\mathrm{d}X_s &= \tilde{b}(s, X_s, \tilde{\phi}(s, X_s, Y_s))\,\mathrm{d}s + \tilde{\sigma}(s)\,\mathrm{d}W_s + \int_{\mathbb{R}_0^p} \tilde{\gamma}(s,u)\,\tilde{N}(\mathrm{d}s,\mathrm{d}u), \qquad X_t = x,\\
\mathrm{d}Y_s &= -\partial_x \tilde{H}(s, X_s, \tilde{\phi}(s, X_s, Y_s), Y_s)\,\mathrm{d}s + Z_s\,\mathrm{d}W_s + \int_{\mathbb{R}_0^p} M_s\,\tilde{N}(\mathrm{d}s,\mathrm{d}u), \quad Y_T = \nabla g(X_T).
\end{aligned} \tag{3.2.21}$$

The following theorem quantifies the difference of optimal feedback controls in terms of the magnitude of perturbations in the coefficients.

Theorem 3.2.6. *Suppose (H.2) and (H.3) hold. Let $\psi, \tilde{\psi} : [0,T] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^k$ be the functions defined in (3.2.17) and (3.2.19), respectively. Then there exists a constant C such that $|\psi(t,x) - \tilde{\psi}(t,x)| \le C(1+|x|)\mathcal{E}_{per}$ for all $(t,x) \in [0,T] \times \mathbb{R}^n$, with the constant $\mathcal{E}_{per}$ defined by*

$$\begin{aligned}\mathcal{E}_{per} \coloneqq \|b_0 - \tilde{b}_0\|_{L^2} + \|b_1 - \tilde{b}_1\|_{L^\infty} + \|b_2 - \tilde{b}_2\|_{L^\infty} + \|\sigma - \tilde{\sigma}\|_{L^2} \\ + \left(\int_0^T \int_{\mathbb{R}_0^p} |\gamma(t,u) - \tilde{\gamma}(t,u)|^2 \, \nu(\mathrm{d}u)\mathrm{d}t \right)^{1/2}.\end{aligned} \tag{3.2.22}$$

Proof. Throughout this proof, for each $(t,x) \in [0,T] \times \mathbb{R}^n$, let $(X^{t,x}, Y^{t,x}, Z^{t,x}, M^{t,x}) \in \mathbb{S}(t,T)$ and $(\tilde{X}^{t,x}, \tilde{Y}^{t,x}, \tilde{Z}^{t,x}, \tilde{M}^{t,x}) \in \mathbb{S}(t,T)$ be the solutions to (3.2.11) and (3.2.21), respectively, and let C be a generic constant which is independent of $(t,x) \in [0,T] \times \mathbb{R}^n$. Then by Proposition 3.2.4, there exists $C \ge 0$ such that for all $(t,x) \in [0,T] \times \mathbb{R}^n$, $\|(X^{t,x}, Y^{t,x}, Z^{t,x}, M^{t,x})\|_{\mathbb{S}(t,T)} \le C(1+|x|)$ and $\|(\tilde{X}^{t,x}, \tilde{Y}^{t,x}, \tilde{Z}^{t,x}, \tilde{M}^{t,x})\|_{\mathbb{S}(t,T)} \le C(1+|x|)$.

We first estimate the difference between $(X^{t,x}, Y^{t,x}, Z^{t,x}, M^{t,x})$ and $(\tilde{X}^{t,x}, \tilde{Y}^{t,x}, \tilde{Z}^{t,x}, \tilde{M}^{t,x})$ for a given $(t,x) \in [0,T] \times \mathbb{R}^n$. By Lemmas 3.2.3 and 3.6.1,

$$\begin{aligned}&\|(X^{t,x} - \tilde{X}^{t,x}, Y^{t,x} - \tilde{Y}^{t,x}, Z^{t,x} - \tilde{Z}^{t,x}, M^{t,x} - \tilde{M}^{t,x})\|_{\mathbb{S}(t,T)} \\ &\le C\Big\{ \|b(\cdot, \tilde{X}^{t,x}, \phi(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x})) - \tilde{b}(\cdot, \tilde{X}^{t,x}, \tilde{\phi}(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x}))\|_{\mathcal{H}^2} \\ &\quad + \|\partial_x H(\cdot, \tilde{X}^{t,x}, \phi(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x}), \tilde{Y}^{t,x}) - \partial_x \tilde{H}(\cdot, \tilde{X}^{t,x}, \tilde{\phi}(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x}), \tilde{Y}^{t,x})\|_{\mathcal{H}^2} \\ &\quad + \|\sigma - \tilde{\sigma}\|_{L^2} + \left(\int_0^T \int_{\mathbb{R}_0^p} |\gamma(t,u) - \tilde{\gamma}(t,u)|^2 \, \nu(\mathrm{d}u)\mathrm{d}t \right)^{1/2} \Big\}.\end{aligned}$$

It remains to estimate the first two terms on the right-hand side of the above inequality. By (H.2(1)),

$$\begin{aligned}&\|b(\cdot, \tilde{X}^{t,x}, \phi(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x})) - \tilde{b}(\cdot, \tilde{X}^{t,x}, \tilde{\phi}(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x}))\|_{\mathcal{H}^2} \\ &\le \|b(\cdot, \tilde{X}^{t,x}, \phi(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x})) - b(\cdot, \tilde{X}^{t,x}, \tilde{\phi}(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x}))\|_{\mathcal{H}^2} \\ &\quad + \|b(\cdot, \tilde{X}^{t,x}, \tilde{\phi}(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x})) - \tilde{b}(\cdot, \tilde{X}^{t,x}, \tilde{\phi}(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x}))\|_{\mathcal{H}^2} \\ &\le \|b_2\|_{L^\infty} \|\phi(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x}) - \tilde{\phi}(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x})\|_{\mathcal{H}^2} + \|b_0 - \tilde{b}_0\|_{L^2} + \|b_1 - \tilde{b}_1\|_{L^\infty} \|\tilde{X}^{t,x}\|_{\mathcal{H}^2} \\ &\quad + \|b_2 - \tilde{b}_2\|_{L^\infty} \|\tilde{\phi}(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x})\|_{\mathcal{H}^2}.\end{aligned}$$

Note that by (3.2.10), for all $t \in [0,T]$ and $x, y \in \mathbb{R}^n$, $\phi(t,x,y) = (\partial_z f^*)(t,x,-b_2(t)^\mathsf{T} y)$ and $\tilde{\phi}(t,x,y) = (\partial_z f^*)(t,x,-\tilde{b}_2(t)^\mathsf{T} y)$, where f^* is the function defined in (3.2.4). Hence, from the $1/\lambda$-Lipschitz continuity of $\partial_z f^*(t,x,\cdot)$ (see the proof of Lemma 3.2.2),

$$\|\phi(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x}) - \tilde{\phi}(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x})\|_{\mathcal{H}^2} \le C\|b_2 - \tilde{b}_2\|_{L^\infty} \|\tilde{Y}^{t,x}\|_{\mathcal{H}^2} \le C(1+|x|)\mathcal{E}_{\mathrm{per}}, \tag{3.2.23}$$

where the last inequality follows from the moment estimate of $\tilde{Y}^{t,x}$. Moreover, the regularity of $\tilde{\phi}$ (see Lemma 3.2.3) and the moment estimate of $(\tilde{X}^{t,x}, \tilde{Y}^{t,x})$ imply that $\|\tilde{\phi}(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x})\|_{\mathcal{H}^2} \le$

$C(1+|x|)$, which shows that $\|b(\cdot, \tilde{X}^{t,x}, \phi(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x})) - b(\cdot, \tilde{X}^{t,x}, \tilde{\phi}(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x}))\|_{\mathcal{H}^2} \leq C(1+|x|)\mathcal{E}_{\rm per}$. By the definitions of H and $\tilde{H}$, the Lipschitz continuity of $\partial_x f_0$ in (H.2(3)) and (3.2.23),

$$\begin{aligned}
&\|\partial_x H(\cdot, \tilde{X}^{t,x}, \phi(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x}), \tilde{Y}^{t,x}) - \partial_x \tilde{H}(\cdot, \tilde{X}^{t,x}, \tilde{\phi}(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x}), \tilde{Y}^{t,x})\|_{\mathcal{H}^2} \\
&\leq \|(b_1 - \tilde{b}_1)^\intercal \tilde{Y}^{t,x}\|_{\mathcal{H}^2} + \|\partial_x f_0(\cdot, \tilde{X}^{t,x}, \phi(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x})) - \partial_x f_0(\cdot, \tilde{X}^{t,x}, \tilde{\phi}(\cdot, \tilde{X}^{t,x}, \tilde{Y}^{t,x}))\|_{\mathcal{H}^2} \\
&\leq C(1+|x|)\mathcal{E}_{\rm per}.
\end{aligned}$$

Thus, we have proved the stability estimate that $\|(X^{t,x} - \tilde{X}^{t,x}, Y^{t,x} - \tilde{Y}^{t,x}, Z^{t,x} - \tilde{Z}^{t,x}, M^{t,x} - \tilde{M}^{t,x})\|_{\mathbb{S}(t,T)} \leq C(1+|x|)\mathcal{E}_{\rm per}$.

We now establish the stability of feedback controls. By (3.2.10) and the $1/\lambda$-Lipschitz continuity of $\partial_z f^*(t,x,\cdot)$, for all $(t,x) \in [0,T] \times \mathbb{R}^n$,

$$\begin{aligned}
&|\psi(t,x) - \tilde{\psi}(t,x)| = |(\partial_z f^*)(t,x,-b_2(t)^\intercal Y_t^{t,x}) - (\partial_z f^*)(t,x,-\tilde{b}_2(t)^\intercal \tilde{Y}_t^{t,x}))| \\
&\leq |b_2(t)^\intercal Y_t^{t,x} - \tilde{b}_2(t)^\intercal \tilde{Y}_t^{t,x}|/\lambda \leq C(\|b_2 - \tilde{b}_2\|_{L^\infty}|Y_t^{t,x}| + |Y_t^{t,x} - \tilde{Y}_t^{t,x}|) \\
&\leq C(\|b_2 - \tilde{b}_2\|_{L^\infty}\|Y^{t,x}\|_{\mathcal{S}^2} + \|Y_t^{t,x} - \tilde{Y}_t^{t,x}\|_{\mathcal{S}^2}) \leq C(1+|x|)\mathcal{E}_{\rm per}. \qquad \square
\end{aligned}$$

An important application of the Lipschitz stability of feedback controls (Theorem 3.2.6) is the analysis of model misspecification error of a given learning algorithm. One essential component is to examine the performance of the feedback control $\tilde{\psi}$, computed based on the control problem (3.2.1) with the perturbed coefficients $(\tilde{b}, \tilde{\sigma}, \tilde{\gamma}, f, g)$, on the true model with coefficients $(b, \sigma, \gamma, f, g)$. For any given $x_0 \in \mathbb{R}^n$, implementing the feedback control $\tilde{\psi}$ on the original system (3.2.2) will lead to the sub-optimal cost:

$$J(\tilde{\psi}; x_0) := \mathbb{E}\left[\int_0^T f(t, X_t^{x_0,\tilde{\psi}}, \tilde{\psi}(t, X_t^{x_0,\tilde{\psi}}))\,\mathrm{d}t + g(X_T^{x_0,\tilde{\psi}})\right], \tag{3.2.24}$$

where $X^{x_0,\tilde{\psi}} \in \mathcal{S}^2(\mathbb{R}^n)$ is the state process (with coefficients b, σ and γ) associated with $\tilde{\psi}$ (see Definition 3.2.1). The following theorem shows that the difference between this suboptimal cost $J(\tilde{\psi}; x_0)$ and the optimal cost V in (3.2.1) depends Lipschitz-continuously on the magnitude of perturbations in the coefficients.

Theorem 3.2.7. *Suppose (H.2) and (H.3) hold. Let $\psi \in \mathcal{V}$ (resp. $\tilde{\psi} \in \mathcal{V}$) be defined in (3.2.17) (resp. (3.2.19)), and for each $x_0 \in \mathbb{R}^n$, let $X^{x_0,\psi} \in \mathcal{S}^2(\mathbb{R}^n)$ (resp. $X^{x_0,\tilde{\psi}} \in \mathcal{S}^2(\mathbb{R}^n)$) be the state process (3.2.2) associated with ψ (resp. $\tilde{\psi}$), and let $V(x_0)$ (resp. $J(\tilde{\psi}; x_0)$) be defined in (3.2.1) (resp. (3.2.24)). Then there exists a constant C such that for all $x_0 \in \mathbb{R}^n$, $\|X^{x_0,\psi} - X^{x_0,\tilde{\psi}}\|_{\mathcal{S}^2} \leq C(1+|x_0|)\mathcal{E}_{per}$ and $|V(x_0) - J(\tilde{\psi}; x_0)| \leq C(1+|x_0|^2)\mathcal{E}_{per}$, with the constant $\mathcal{E}_{per}$ defined in (3.2.22).*

To prove Theorem 3.2.7, we first establish that the composition of f and the optimal feedback control is Lipschitz continuous, even though the cost function f is merely lower

semicontinuous in the control variable (cf. (H.2(3))). The proof is based on the Fenchel-Young identity:

$$f(t,x,\partial_z f^*(t,x,z)) = \langle z, \partial_z f^*(t,x,z)\rangle - f^*(t,x,z) \in \mathbb{R}, \quad \forall (t,x,z) \in [0,T]\times\mathbb{R}^n\times\mathbb{R}^k,$$

the regularity of f^* and Theorem 3.2.6, and has been given in Section 3.6.2.

Lemma 3.2.8. *Suppose (H.2) and (H.3) hold. Let $\psi, \tilde{\psi} : [0,T]\times\mathbb{R}^n \to \mathbb{R}^k$ be the functions defined in (3.2.17) and (3.2.19), respectively. Then there exists a constant C such that for all $t \in [0,T]$, $x, x' \in \mathbb{R}^n$, $|f(t,x,\psi(t,x)) - f(t,x',\tilde{\psi}(t,x'))| \le C\Big((1+|x|+|x'|)|x-x'| + (1+|x|^2+|x'|^2)\mathcal{E}_{per}\Big)$, where the constant $\mathcal{E}_{per}$ is defined in (3.2.22).*

Proof of Theorem 3.2.7. According to Theorems 3.2.5 and 3.2.6, there exists a constant C such that for all $t \in [0,T]$, $x, x' \in \mathbb{R}^n$, $|\psi(t,0)| + |\tilde{\psi}(t,0)| \le C$, $|\psi(t,x) - \psi(t,x')| + |\tilde{\psi}(t,x) - \tilde{\psi}(t,x')| \le C|x-x'|$, and $|\psi(t,x) - \tilde{\psi}(t,x)| \le C(1+|x|)\mathcal{E}_{\rm per}$. Then, for any given $x_0 \in \mathbb{R}^n$, standard moment and stability estimates of (3.2.7) yield $\|X^{x_0,\psi}\|_{\mathcal{S}^2} + \|X^{x_0,\tilde{\psi}}\|_{\mathcal{S}^2} \le C(1+|x_0|)$ and

$$\begin{aligned}\|X^{x_0,\psi} - X^{x_0,\tilde{\psi}}\|_{\mathcal{S}^2} &\le C\|b(\cdot, X^{x_0,\tilde{\psi}}, \psi(\cdot, X^{x_0,\tilde{\psi}})) - b(\cdot, X^{x_0,\tilde{\psi}}, \tilde{\psi}(\cdot, X^{x_0,\tilde{\psi}}))\|_{\mathcal{H}^2}\\ &\le C\|\psi(\cdot, X^{x_0,\tilde{\psi}}) - \tilde{\psi}(\cdot, X^{x_0,\tilde{\psi}})\|_{\mathcal{H}^2} \le C(1+\|X^{x_0,\tilde{\psi}}\|_{\mathcal{H}^2})\mathcal{E}_{\rm per} \le C(1+|x_0|)\mathcal{E}_{\rm per}.\end{aligned}$$

We now proceed to estimate $|V(x_0) - \tilde{V}(x_0)|$ for any given $x_0 \in \mathbb{R}^n$. By the mean value theorem, (H.2(2)) and the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}[|g(X_T^{x_0,\psi}) - g(X_T^{x_0,\tilde{\psi}})|] &\le C\mathbb{E}[|(1+|X_T^{x_0,\psi}| + |X_T^{x_0,\tilde{\psi}}|)|X_T^{x_0,\psi} - X_T^{x_0,\tilde{\psi}}|]\\ &\le C(1+\|X_T^{x_0,\psi}\|_{L^2} + \|X_T^{x_0,\tilde{\psi}}\|_{L^2})\|X_T^{x_0,\psi} - X_T^{x_0,\tilde{\psi}}\|_{L^2}\\ &\le C(1+|x_0|^2)\mathcal{E}_{\rm per}.\end{aligned}$$

Moreover, from Lemma 3.2.8 and the Cauchy-Schwarz inequality,

$$\begin{aligned}&\mathbb{E}\left[\int_0^T |f(t, X_t^{x_0,\psi}, \psi(t, X_t^{x_0,\psi})) - f(t, X_t^{x_0,\tilde{\psi}}, \tilde{\psi}(t, X_t^{x_0,\tilde{\psi}}))|\,\mathrm{d}t\right]\\ &\le C\mathbb{E}\left[\int_0^T \left((1+|X_t^{x_0,\psi}| + |X_t^{x_0,\tilde{\psi}}|)|X_t^{x_0,\psi} - X_t^{x_0,\tilde{\psi}}| + (1+|X_t^{x_0,\psi}|^2 + |X_t^{x_0,\tilde{\psi}}|^2)\mathcal{E}_{\rm per}\right)\mathrm{d}t\right]\\ &\le C\Big((1+\|X^{x_0,\psi}\|_{\mathcal{H}^2} + \|X^{x_0,\tilde{\psi}}\|_{\mathcal{H}^2})\|X^{x_0,\psi} - X^{x_0,\tilde{\psi}}\|_{\mathcal{H}^2} + (1+\|X^{x_0,\psi}\|_{\mathcal{H}^2}^2 + \|X^{x_0,\tilde{\psi}}\|_{\mathcal{H}^2}^2)\mathcal{E}_{\rm per}\Big)\\ &\le C(1+|x_0|^2)\mathcal{E}_{\rm per}.\end{aligned}$$

Since ψ is an optimal feedback control of (3.2.1) with the initial state $x_0 \in \mathbb{R}^n$, the desired estimate $|V(x_0) - J(\tilde{\psi}; x_0)| \le C(1+|x_0|^2)\mathcal{E}_{\rm per}$ follows. □

3.3 Regret analysis for linear-convex reinforcement learning

The focus of this section is the linear-convex reinforcement learning (RL) problem, where the drift coefficient of the state dynamics (3.2.2) is unknown to the controller, and the objective is to control the system optimally while simultaneously learning the dynamics. We shall propose a greedy least-squares algorithm to solve such problems, and show that the algorithm provides a sublinear regret with high probability guarantees. The analysis of the regret bounds for the algorithm relies on the Lipschitz stability of feedback controls established in Section 3.2.1.

3.3.1 Reinforcement learning problem and least-squares algorithm

The RL problem goes as follows. Let $x_0 \in \mathbb{R}^n$ be a given initial state and $\theta^\star = (A^\star, B^\star) \in \mathbb{R}^{n\times(n+k)}$ be fixed but unknown constants, consider the following problem:

$$V(x_0;\theta^\star) = \inf_{\alpha\in\mathcal{H}^2(\mathbb{R}^k)} J^{\theta^\star}(\alpha;x_0), \quad \text{with} \quad J^{\theta^\star}(\alpha;x_0) = \mathbb{E}\left[\int_0^T f(t, X_t^{x_0,\theta^\star,\alpha}, \alpha_t)\,\mathrm{d}t + g(X_T^{x_0,\theta^\star,\alpha})\right], \tag{3.3.1}$$

where for each $\alpha \in \mathcal{H}^2(\mathbb{R}^k)$, the process $X^{x_0,\theta^\star,\alpha}$ satisfies the following controlled dynamics associated with the parameter $\theta^\star$:

$$\mathrm{d}X_t = (A^\star X_t + B^\star \alpha_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t + \int_{\mathbb{R}_0^p} \gamma(u)\,\tilde{N}(\mathrm{d}t,\mathrm{d}u), \quad t\in[0,T], \quad X_0 = x_0, \tag{3.3.2}$$

with a given constant $\sigma \in \mathbb{R}^{n\times d}$ and given functions $\gamma : \mathbb{R}_0^p \to \mathbb{R}^n$, $f : [0,T]\times\mathbb{R}^n\times\mathbb{R}^k \to \mathbb{R}\cup\{\infty\}$ and $g : \mathbb{R}^n \to \mathbb{R}$. If $\theta^\star = (A^\star, B^\star)$ were known, then (3.3.1) is a control problem.

It is clear that (3.3.1)-(3.3.2) is a special case of (3.2.1)-(3.2.2) with $b(t,x,a) = A^\star x + B^\star a$, $\sigma(t) = \sigma$ and $\gamma(t,u) = \gamma(u)$ for all $(t,x,a,u) \in [0,T]\times\mathbb{R}^n\times\mathbb{R}^k\times\mathbb{R}_0^p$. Hence, if f and g satisfy (H.2) with for some $L \geq 0$ and $\lambda > 0$, then (3.3.1)-(3.3.2) admits an optimal feedback control $\psi^{\theta^\star} \in \mathcal{V}$ as shown in Theorem 3.2.5. Note that to simplify the presentation, we assume that (3.3.2) has time homogenous coefficients as in [1, 112, 17], but similar analysis can be performed if the drift is a linear combination of given time-and-space-dependent basis functions with unknown weights or the diffusion/jump coefficients are also unknown.

To solve (3.3.1)-(3.3.2) with unknown $\theta^\star$, in an episodic reinforcement learning framework, the controller improves their knowledge of the parameter $\theta^\star$ through successive learning episodes. In particular, for each episode $i \in \mathbb{N}$, based on her observations in the past episodes, the controller executes a suitable control policy in $\psi_i \in \mathcal{V}$, whose associated state dynamics (3.3.2) leads to an expected cost $J^{\theta^\star}(\psi_i;x_0)$. To measure the performance of an learning algorithm in this setting, one widely adopted criteria is the (expected) regret of the

algorithm defined as follows (see e.g. [43, 17]):

$$R(N) = \sum_{i=1}^{N} \Big(J^{\theta^\star}(\psi_i; x_0) - V(x_0; \theta^\star) \Big), \quad \forall N \in \mathbb{N}, \tag{3.3.3}$$

where N denotes the total number of learning episodes. Intuitively, this regret characterizes the cumulative loss from taking sub-optimal policies in all episodes.

To start, let us consider a greedy algorithm, which chooses the optimal feedback control based on the current estimation of the parameter, and provides a sublinear regret with respect to the number of episodes N. More precisely, let $\theta = (A, B) \in \mathbb{R}^{n\times(n+k)}$ be the current estimate of $\theta^\star$, then the controller would exercise the optimal feedback control $\psi^\theta \in \mathcal{V}$ defined in Theorem 3.2.5 for the control problem (3.3.1)-(3.3.2) with $\theta^\star$ replaced by θ, which leads to the state process $X^{x_0,\theta} \in \mathcal{S}^2(\mathbb{R}^n)$ satisfying:

$$\mathrm{d}X_t = (A^\star X_t + B^\star \psi^\theta(t, X_t))\,\mathrm{d}t + \sigma\,\mathrm{d}W_t + \int_{\mathbb{R}_0^p} \gamma(u)\,\tilde{N}(\mathrm{d}t, \mathrm{d}u), \quad t \in [0,T], \quad X_0 = x_0. \tag{3.3.4}$$

By the martingale properties of stochastic integrals, we can then estimate $\theta^\star$ based on the process $Z_t^{x_0,\theta} := \begin{pmatrix} X_t^{x_0,\theta} \\ \psi^\theta(t, X_t^{x_0,\theta}) \end{pmatrix}$, $t \in [0,T]$, as follows:

$$(\theta^\star)^\intercal = \left(\mathbb{E}\Big[\int_0^T Z_t^{x_0,\theta} (Z_t^{x_0,\theta})^\intercal \,\mathrm{d}t \Big] \right)^{-1} \mathbb{E}\Big[\int_0^T Z_t^{x_0,\theta} (\mathrm{d}X_t^{x_0,\theta})^\intercal \Big], \tag{3.3.5}$$

provided that $\mathbb{E}\big[\int_0^T Z_t^{x_0,\theta} (Z_t^{x_0,\theta})^\intercal \,\mathrm{d}t \big] \in \mathbb{R}^{(n+k)\times(n+k)}$ is invertible. This motivates us to introduce an iterative procedure to estimate $\theta^\star$, where the expectations in (3.3.5) are replaced by empirical averages over independent realizations. More precisely, let $m \in \mathbb{N}$ and $(X_t^{x_0,\theta,i}, \psi^\theta(t, X_t^{x_0,\theta,i}))_{t\in[0,T]}$, $i = 1, \ldots, m$, be trajectories of m independent realizations of the state and control processes, we shall update the estimate θ, denoted by $\hat{\theta}$, according to (3.3.5):

$$\hat{\theta}^\intercal := \left(\frac{1}{m} \sum_{i=1}^{m} \int_0^T Z_t^{x_0,\theta,i} (Z_t^{x_0,\theta,i})^\intercal \,\mathrm{d}t + \frac{1}{m} \mathbb{I}_{n+k} \right)^{-1} \left(\frac{1}{m} \sum_{i=1}^{m} \int_0^T Z_t^{x_0,\theta,i} (\mathrm{d}X_t^{x_0,\theta,i})^\intercal \right), \tag{3.3.6}$$

where $Z_t^{x_0,\theta,i} := \begin{pmatrix} X_t^{x_0,\theta,i} \\ \psi^\theta(t, X_t^{x_0,\theta,i}) \end{pmatrix}$ for all $t \in [0,T]$ and $i = 1, \ldots, m$, and $\mathbb{I}$ is the $(n+k)\times(n+k)$ identity matrix used to ensure the existence of the required matrix inverse. This leads to the following greedy least-squares (GLS) algorithm:

Algorithm 10 Greedy least-squares (GLS) algorithm

1: **Input**: Choose an initial estimation θ_0 of $\theta^\star$ and numbers of learning episodes $\{m_\ell\}_{\ell\in\mathbb{N}\cup\{0\}}$.
2: **for** $\ell = 0, 1, \cdots$ **do**
3: Obtain the optimal feedback control ψ^{θ_ℓ} for (3.3.1)-(3.3.2) with $\theta^\star = \theta_\ell$ as in Theorem 3.2.5.
4: Execute the feedback control ψ^{θ_ℓ} for m_ℓ indepenent episodes, and collect the trajectory data $(X_t^{x_0,\theta_\ell,i}, \psi^{\theta_\ell}(t, X_t^{x_0,\theta_\ell,i}))_{t\in[0,T]}$, $i = 1, \ldots, m_\ell$.
5: Obtain an updated estimation $\theta_{\ell+1}$ by using (3.3.6) and the m_ℓ trajectories collected above.
6: **end for**

3.3.2 Structural assumptions for learning problems

In this section, we analyze the regret of Algorithm 10 based on the following assumptions of the learning problem (3.3.1)-(3.3.2).

H.4. *(1) Let* $x_0 \in \mathbb{R}^n$, $\theta^\star = (A^\star, B^\star) \in \mathbb{R}^{n\times(n+k)}$, $\sigma \in \mathbb{R}^{n\times d}$, $\gamma : \mathbb{R}_0^p \to \mathbb{R}^n$, $f : [0,T]\times\mathbb{R}^n\times\mathbb{R}^k \to \mathbb{R}\cup\{\infty\}$ *and* $g : \mathbb{R}^n \to \mathbb{R}$ *satisfy (H.2) with some constants* $L \geq 0$ *and* $\lambda > 0$.

(2) There exist $\gamma_{\max} \geq 0$ *and* $\vartheta \in [0,1]$ *such that* $\sup_{q\geq 2} q^{-\vartheta}\big(\int_{\mathbb{R}_0^p} |\gamma(u)|^q\, \nu(\mathrm{d}u)\big)^{1/q} \leq \gamma_{\max}$.

Remark 3.3.1. *Condition (H.4(1)) implies that for each* $\theta = (A, B)$, *the control problem of (3.3.1)-(3.3.2) with* $\theta^\star$ *replaced by* θ *is a nonsmooth linear-convex control problem studied in Section 3.2.*

Condition (H.4(2)) describes the large jumps of the pure jump process

$$L_t := \int_0^t \int_{\mathbb{R}_0^p} \gamma(u)\, \tilde{N}(\mathrm{d}s, \mathrm{d}u), \quad t \in [0,T],$$

which enables estimating the tail behaviour of the state process X^θ, *and subsequently quantifying the parameter estimation error of the least-squares estimator (3.3.6) (see Section 3.3.4.2). If the jump coefficient* γ *is bounded, then one can easily see from* $\int_{\mathbb{R}_0^p} |\gamma(u)|^2\, \nu(\mathrm{d}u) < \infty$ *that (H.4(2)) holds with* $\vartheta = 0$. *Another important case is when* $\gamma(u) = u$ *for all* $u \in \mathbb{R}_0^p$, *under which the process* $(L_t)_{t\in[0,T]}$ *is a Lévy process of pure jumps with Lévy measure* $\nu(\mathrm{d}u)$. *In this case, (H.4(2)) holds with* $\vartheta \in (0,1]$ *if and only if* $\big(\int_{\mathbb{R}_0^p} |u|^q\, \nu(\mathrm{d}u)\big)^{1/q} \leq \mathcal{O}(q^\theta)$ *as* $q \to \infty$.

H.5. $\theta^\star$ *is identifiable, i.e., the optimal control* $\alpha^{x_0,\star} \in \mathcal{H}^2(\mathbb{R}^k)$ *and the optimal state process* $X^{x_0,\theta^\star,\alpha^\star} \in \mathcal{S}^2(\mathbb{R}^n)$ *of (3.3.1)-(3.3.2) (with initial state* x_0 *and parameter* $\theta^\star$*) satisfy the following linear independence condition: if* $u_1 \in \mathbb{R}^n$ *and* $u_2 \in \mathbb{R}^k$ *satisfy* $u_1^\intercal X_t^{x_0,\theta^\star,\alpha^\star} + u_2^\intercal \alpha_t^{x_0,\star} = 0$ *for* $\mathrm{d}\mathbb{P}\otimes\mathrm{d}t$ *a.e., then* u_1 *and* u_2 *are zero vectors.*

Condition (H.5) implies that the true parameter $\theta^\star$ can be uniquely identified if we observe sufficiently many trajectories of the optimal state and control processes of (3.3.1)-(3.3.2). Such a self-exploration property allows us to design *exploration-free* learning algorithms for (3.3.1)-(3.3.2).

The following proposition shows that if the laws of the state processes are supported on the whole space, then (H.5) is equivalent to a self-exploration property of the optimal feedback control. The proof essentially follows the argument of [148, Lemma 6.1], and hence is omitted.

Proposition 3.3.1. *Assume (H.4(1)). Let $\psi \in \mathcal{V}$. Assume that for all $t \in (0,T]$, and any open set $O \subset \mathbb{R}^n$ with positive Lebesgue measure, the state process $X^{\theta^\star,\psi}$ (defined by (3.3.4) with $\psi^\theta = \psi$) satisfies that $\mathbb{P}(\{\omega \in \Omega \mid X_t^{\theta^\star,\psi}(\omega) \in O\}) > 0$. Then the following two statements are equivalent:*

(a) if $u_1 \in \mathbb{R}^n$ and $u_2 \in \mathbb{R}^k$ satisfy $u_1^\mathsf{T} X_t^{\theta^\star,\psi} + u_2^\mathsf{T}\psi(t, X_t^{\theta^\star,\psi}) = 0$ for $\mathrm{d}\mathbb{P} \otimes \mathrm{d}t$ a.e., then u_1 and u_2 are zero vectors;

(b) if $u_1 \in \mathbb{R}^n$ and $u_2 \in \mathbb{R}^k$ satisfy $u_1^\mathsf{T} x + u_2^\mathsf{T}\psi(t,x) = 0$ for almost every $(t,x) \in [0,T]\times\mathbb{R}^n$, then u_1 and u_2 are zero vectors.

Consequently, suppose that (H.4(1)) holds and $\sigma\sigma^\mathsf{T}$ is positive definite, then (H.5) holds if and only if the optimal feedback control $\psi^{\theta^\star}$ of (3.3.1) satisfies Item (b).

Proposition 3.3.1 allows for more explicit expressions of (H.5). For instance, as shown in [17, Proposition 3.9], for quadratic cost functions $g = 0$ and $f(t,x,a) = x^\mathsf{T}Qx + a^\mathsf{T}Ra$ with positive definite matrices Q and R, (H.5) holds if and only if $B^\star$ in (3.3.2) is full column rank. Alternatively, by [148, Proposition 3.3], if (3.3.1)-(3.3.2) has a bounded action set, i.e., $\mathcal{R}$ in (H.2(3)) has a bounded domain $\operatorname{dom}\mathcal{R}$ (cf. Example 3.2.1), then (H.5) holds if and only if the range of $\psi^{\theta^\star}$ contains k linearly independent vectors.

We remark that for general linear-convex learning problems without (H.5), an explicit exploration is necessary for learning [148] . Instead of merely employing greedy polices as in Algorithm 10, they dedicate certain episodes to actively explore the environment with some exploration policy ψ^e satisfying Proposition 3.3.1 Item (b). The numbers of exploration and exploitation episodes are then balanced based on the performance gap in Theorem 3.2.7 and the finite-sample accuracy of the parameter estimator. Note, however, this explicit exploration may yield larger regrets for algorithm in [148] than that in Theorem 3.3.2.

3.3.3 Main results on sublinear regret bounds

We now state the main result which shows that the regret of Algorithm 10 grows at most sublinearly with respect to the number of episodes, provided that the hyper-parameters θ_0 and $\{m_j\}_{j\in\mathbb{N}\cup\{0\}}$ are chosen properly. In particular, we shall choose an initial guess θ_0 of $\theta^\star$ which satisfies the identifiability condition in (H.5) and we shall also double the number of learning episodes between two successive updates of the estimation of $\theta^\star$, which is a commonly

used strategy (the so-called doubling trick) in the design of online learning algorithms (see e.g. [17]). The proof of this theorem is given in Section 3.3.4.3.

To simplify the notation, we introduce the following quantifies for each $x_0 \in \mathbb{R}^n$, $\theta = (A, B) \in \mathbb{R}^{n\times(n+k)}$ and $m \in \mathbb{N}$:

$$\begin{aligned} \overline{U}^{x_0,\theta} &:= \mathbb{E}\bigg[\int_0^T Z_t^{x_0,\theta}(Z_t^{x_0,\theta})^\intercal \,\mathrm{d}t\bigg], & \overline{V}^{x_0,\theta} &:= \mathbb{E}\bigg[\int_0^T Z_t^{x_0,\theta}(\mathrm{d}X_t^{x_0,\theta})^\intercal\bigg], \\ U^{x_0,\theta,m} &:= \frac{1}{m}\sum_{i=1}^m \int_0^T Z_t^{x_0,\theta,i}(Z_t^{x_0,\theta,i})^\intercal \,\mathrm{d}t, & V^{x_0,\theta,m} &:= \frac{1}{m}\sum_{i=1}^m \int_0^T Z_t^{x_0,\theta,i}(\mathrm{d}X_t^{x_0,\theta,i})^\intercal, \end{aligned} \tag{3.3.7}$$

where $X^{x_0,\theta} \in \mathcal{S}^2(\mathbb{R}^n)$ is the solution of (3.3.4), $(X^{x_0,\theta,i})_{i=1}^m$ are independent copies of $X^{x_0,\theta}$, and $Z^{x_0,\theta}$ and $(Z^{x_0,\theta,i})_{i=1}^m$ are defined as in (3.3.5) and (3.3.6), respectively. For any given symmetric matrix A, we denote by $\lambda_{\min}(A)$ the smallest eigenvalue of A.

Theorem 3.3.2. *Suppose (H.4(1)) and (H.5) hold. Assume further that $\lambda_{\min}(\overline{U}^{x_0,\theta_0}) > 0$, and for any given bounded set $\mathcal{K} \subset \mathbb{R}^{n\times(n+k)}$, there exist constants $C_1, C_2 > 0$ and $\beta \geq 1$, such that the following concentration inequality holds for all $\varepsilon \geq 0$, $m \in \mathbb{N}$ and $\theta \in \mathcal{K}$,*

$$\begin{aligned} &\max\Big\{\mathbb{P}\big(\big|U^{x_0,\theta,m} - \overline{U}^{x_0,\theta}\big| \geq \varepsilon\big), \mathbb{P}\big(\big|V^{x_0,\theta,m} - \overline{V}^{x_0,\theta}\big| \geq \varepsilon\big)\Big\} \\ &\quad \leq C_2 \exp\Big(-C_1 \min\Big\{\frac{m\varepsilon^2}{C_2^2}, \Big(\frac{m\varepsilon}{C_2}\Big)^{\frac{1}{\beta}}\Big\}\Big). \end{aligned} \tag{3.3.8}$$

Then there exists a constant $C_0 > 0$, such that for all $C \geq C_0$ and $\delta \in (0, 1/4)$, if we set $m_0 = C(-\ln\delta)^\beta$ and $m_\ell = 2^\ell m_0$ for all $\ell \in \mathbb{N}$, then the regret of Algorithm 10 (cf. (3.3.3)) satisfies the following properties:

(1) It holds with probability at least $1 - 4\delta$ that $R(N) \leq C'\big(\sqrt{N}\sqrt{\ln N} + \sqrt{-\ln\delta}\sqrt{N} + (-\ln\delta)^\beta \ln N\big)$ for all $N \in \mathbb{N}$, where C' is a constant independent of N and δ.

(2) It holds with probability 1 that $R(N) = \mathcal{O}(\sqrt{N \ln N})$ as $N \to \infty$.

The following theorem presents a precise sublinear regret bound of Algorithm 10 for the jump-diffusion model (3.3.2), depending on the jump sizes of the Poisson random measure. The proof follows from Theorem 3.3.2 and Proposition 3.3.9.

Theorem 3.3.3. *Suppose (H.4) and (H.5) hold, and $\lambda_{\min}(\overline{U}^{x_0,\theta_0}) > 0$. Then there exists a constant $C_0 > 0$, such that for all $C \geq C_0$ and $\delta \in (0, 1/4)$, if we set $m_0 = C(-\ln\delta)^{3+\vartheta}$ and $m_\ell = 2^\ell m_0$ for all $\ell \in \mathbb{N}$, then the regret of Algorithm 10 (cf. (3.3.3)) satisfies the following properties:*

(1) It holds with probability at least $1 - 4\delta$ that $R(N) \leq C'\big(\sqrt{N}\sqrt{\ln N} + \sqrt{-\ln\delta}\sqrt{N} + (-\ln\delta)^{3+\vartheta} \ln N\big)$ for all $N \in \mathbb{N}$, where ϑ is the constant in (H.4(2)) and C' is a constant independent of ϑ, N and δ.

(2) It holds with probability 1 that $R(N) = \mathcal{O}(\sqrt{N \ln N})$ *as* $N \to \infty$.

In the case where (3.3.2) is only driven by the Brownian motion, we can exploit the sub-Gaussianity of the state process and obtain a shaper regret bound based on Theorem 3.3.2 and Proposition 3.3.10.

Theorem 3.3.4. *Suppose (H.4) and (H.5) hold with* $\gamma_{\max} = 0$, *and* $\lambda_{\min}(\overline{U}^{x_0,\theta_0}) > 0$. *Then there exists a constant* $C_0 > 0$, *such that for all* $C \geq C_0$ *and* $\delta \in (0, 1/4)$, *if we set* $m_0 = C(-\ln \delta)$ *and* $m_\ell = 2^\ell m_0$ *for all* $\ell \in \mathbb{N}$, *then the regret of Algorithm 10 (cf. (3.3.3)) satisfies the following properties:*

(1) It holds with probability at least $1 - 4\delta$ *that* $R(N) \leq C'\big(\sqrt{N}\sqrt{\ln N} + \sqrt{-\ln \delta}\sqrt{N} + (-\ln \delta)\ln N\big)$ *for all* $N \in \mathbb{N}$, *where* C' *is a constant independent of* N *and* δ.

(2) It holds with probability 1 that $R(N) = \mathcal{O}(\sqrt{N \ln N})$ *as* $N \to \infty$.

Remark 3.3.2. *The condition* $\lambda_{\min}(\overline{U}^{x_0,\theta_0}) > 0$ *in Theorems 3.3.3 and 3.3.4 ensures that the greedy policy* ψ^{θ_0} *explores the parameter space and improves the accuracy of parameter estimation. By Proposition 3.3.1, if (3.3.4) has nondegenerate Brownian noises, then it suffices to choose* θ_0 *such that the corresponding greedy policy* ψ^{θ_0} *enjoys the exploration property stated in Item (b).*

The choice of $m_0 = C_0(-\ln \delta)^\beta$ *along with (H.5) ensures that* $(\theta_\ell)_{\ell \in \mathbb{N}}$ *stays close to* $\theta^\star$ *so that (3.3.8) is applicable. Here* δ *is an arbitrarily small constant indicating the agent's confidence of the regret bound, and* C_0 *is a constant depending on the exploration strength of* $\psi^{\theta^\star}$, *namely the constant* $\lambda_{\min}(\overline{U}^{x_0,\theta^\star}) > 0$ *(see Section 3.3.4.3). Note that to analyze algorithm regrets, it is common to assume some a-priori information on the true parameter and the algorithm being initialized with sufficiently many learning episodes (see e.g., [45]). Obtaining an explicit dependence of* C_0 *on model parameters, however, could be challenging. A practical strategy for validating (H.5) and for choosing the initial episode* m_0 *is to ensure that the obtained estimations* $(\theta_\ell)_{\ell \in \mathbb{N}}$ *remain bounded and that the resulting greedy policies* $(\psi^{\theta_\ell})_{\ell \in \mathbb{N}}$ *satisfy Proposition 3.3.1 Item (b). Our numerical experiments in Section 3.5 demonstrate that the performance of Algorithm 10 is stable with respect to* m_0, *and that a small* m_0 *in general suffices to guarantee a sublinear regret.*

3.3.4 Proofs of sublinear regret bounds

This section is devoted to the proofs of Theorem 3.3.2, 3.3.3 and 3.3.4.

As we have seen in Theorems 3.3.3-3.3.4, an essential step for estimating the regret of Algorithm 10 is to establish the concentration inequality (3.3.8) for the least-squares estimator (3.3.6). Compared to the classical learning problems with Brownian-motion-driven state dynamics (see e.g. [17]), the presence of jumps in the state dynamics creates a crucial

difficulty in quantifying the precise value of β in (3.3.8), since the state variable X^θ is in general not sub-Gaussian, and hence (3.3.8) does not hold with $\beta = 1$.

In the subsequent analysis, we overcome the above difficulty by introducing a notation of sub-Weibull random variables as in [98] and establishing that both deterministic and stochastic integrals preserve sub-Weibull random variables in Section 3.3.4.1. We then show in Section 3.3.4.2 that (3.3.6) behaves like sub-Weibull random variables and (3.3.8) holds with some $\beta \geq 1$, provided that the jumps of the state dynamics are sub-exponential. Finally, we prove the general regret result Theorem 3.3.2 for Algorithm 10 in Section 3.3.4.3.

3.3.4.1 Step 1: Analysis of sub-Weibull random variables

The first step is to analyze integrals of sub-Weibull random variables. We start by recalling the precise definition of sub-Weibull random variables in terms of their Orlicz norms (see [98]).

Definition 3.3.1. *For every $\alpha > 0$, let $\Psi_\alpha : [0, \infty) \to \mathbb{R}$ such that $\Psi_\alpha(x) = e^{x^\alpha} - 1$ for all $x \geq 0$, and let $\|\cdot\|_{\Psi_\alpha}$ be the corresponding Ψ_α-Orlicz (quasi-)norm such that for any given random variable X,*

$$\|X\|_{\Psi_\alpha} \coloneqq \inf\left\{t > 0 \mid \mathbb{E}\left[\Psi_\alpha\left(\tfrac{|X|}{t}\right)\right] \leq 1\right\}.$$

Then a random variable X is said to be sub-Weibull of order $\alpha > 0$, denoted by $X \in \mathrm{subW}(\alpha)$, if $\|X\|_{\Psi_\alpha} < \infty$.

Note that $\|\cdot\|_{\Psi_\alpha}$ is a norm if and only if $\alpha \geq 1$, as otherwise the triangle inequality does not hold. Examples of sub-Weibull random variables include sub-Gaussian and sub-exponential random variables, which correspond to $\mathrm{subW}(2)$ and $\mathrm{subW}(1)$, respectively. We point out that the class of sub-Weibull random variables is closed under multiplication and addition, and for all $\alpha > 0$, there exists a constant C_α, depending only on α, such that

$$C_\alpha^{-1} \sup_{q \geq 1} q^{-1/\alpha} \|X\|_{L^q} \leq \|X\|_{\Psi_\alpha} \leq C_\alpha \sup_{q \geq 1} q^{-1/\alpha} \|X\|_{L^q} \tag{3.3.9}$$

for all random variables X (see [70, Appendix A] for a proof of these properties).

We now present several important lemmas regarding the behavior of integrals of sub-Weibull random variables. The first lemma shows that deterministic integral of a product of sub-Weibull random variables is still sub-Weibull. The proof is based on Definition 3.3.1 and Hölder's inequality, and is given in Section 3.6.2.

Lemma 3.3.5. *For all $\alpha > 0$ and every stochastic process $X, Y : \Omega \times [0, T] \to \mathbb{R}$,*

$$\left\| \int_0^T XY \, \mathrm{d}t \right\|_{\Psi_{\alpha/2}} \leq \left\| \left(\int_0^T |X|^2 \, \mathrm{d}t \right)^{\frac{1}{2}} \right\|_{\Psi_\alpha} \left\| \left(\int_0^T |Y|^2 \, \mathrm{d}t \right)^{\frac{1}{2}} \right\|_{\Psi_\alpha}.$$

The second lemma shows that stochastic integrals preserve the property of being sub-Weibull random variables. The proof is based on the equivalent characterization (3.3.9) of sub-Weibull random variables and Burkholder's inequality, whose details are given in Section 3.6.2.

Lemma 3.3.6. *There exists $C \geq 0$ such that for all $\sigma \in \mathbb{R}^d$, $X \in \mathcal{S}^2(\mathbb{R})$ and every measurable function $\gamma : \mathbb{R}_0^p \to \mathbb{R}$ satisfying (H.4(2)), $\| \int_0^T X_t \sigma^\intercal \, \mathrm{d}W_t \|_{\Psi_{1/2}} \leq C|\sigma| \| (\int_0^T |X|^2 \, \mathrm{d}t)^{\frac{1}{2}} \|_{\Psi_1}$ and*

$$\left\| \int_0^T \int_{\mathbb{R}_0^p} X_t \gamma(u) \, \tilde{N}(\mathrm{d}t, \mathrm{d}u) \right\|_{\Psi_{1/(3+\vartheta)}} \leq C \gamma_{\max} \left(\sup_{p \geq 2} \left\| \left(\int_0^T |X_t|^q \mathrm{d}t \right)^{\frac{1}{q}} \right\|_{\Psi_1} \right),$$

with the constants $\gamma_{\max}$ and ϑ in (H.4(2)).

Lemma 3.3.6 focuses on the case where $(\int_0^T |X|^2 \, \mathrm{d}t)^{1/2} \in \mathrm{subW}(1) \setminus \mathrm{subW}(2)$, which is important for control problems whose state dynamics is driven by a Poisson random measure. Hence we establish the sub-Weibull properties of the stochastic integrals by applying the Burkholder's inequality to estimate the growth of their L^q-norms, precise order of which depends on the constants C_q and $\tilde{C}_q$ in the inequalities (3.6.7) and (3.6.8).

In the case where $(\int_0^T |X|^2 \, \mathrm{d}t)^{1/2} \in \mathrm{subW}(2)$, we can establish the optimal sub-Weibull order $\int_0^T X_t \sigma^\intercal \, \mathrm{d}W_t \in \mathrm{subW}(1)$. Such a characeterization is essential for obtaining a sharper regret bound of Algorithm 10 when the state dynamics is only driven by the Brownian motion. The proof is based on the Girsanov theorem and is given in Section 3.6.2.

Lemma 3.3.7. *There exists $C \geq 0$ such that for all $\sigma \in \mathbb{R}^d$ and $X \in \mathcal{S}^2(\mathbb{R})$,*

$$\| \int_0^T X_t \sigma^\intercal \, \mathrm{d}W_t \|_{\Psi_1} \leq C|\sigma| \| (\int_0^T |X|^2 \, \mathrm{d}t)^{\frac{1}{2}} \|_{\Psi_2}.$$

3.3.4.2 Step 2: Concentration inequalities for the least-squares estimator

Based on the fact that sub-Weibull properties are preserved under algebraic and integral operations as shown in Section 3.3.4.1, we now quantify the precise tail behavior of the least-squares estimator (3.3.6), namely the constant β in (3.3.8), for the jump-diffusion model (3.3.2).

We start by establishing the sub-exponential properties of Lipschitz functionals of the state process X^θ driven by both Brownian motions and Poisson random measures as in (3.3.4). The proof follows as a special case of [111] and is given in Section 3.6.2.

Lemma 3.3.8. *Suppose (H.4) holds. Let $K \in \mathbb{R}$ and $\theta = (A, B) \in \mathbb{R}^{n \times (n+k)}$ satisfy $|\theta| \leq K$. Then there exists $C \geq 0$, depending only on K, T and the constants in (H.4), such that for all $x_0 \in \mathbb{R}^n$ and for every Lipschitz continuous function $\mathfrak{f} : (\mathbb{D}([0,T]; \mathbb{R}^n), d_\infty) \to \mathbb{R}$, the solution $X^{x_0,\theta}$ of (3.3.4) satisfies $\|\mathfrak{f}(X^{x_0,\theta})\|_{\Psi_1} \leq C(\|\mathfrak{f}\|_{\mathrm{Lip}} + |\mathbb{E}[\mathfrak{f}(X^{x_0,\theta})]|)$, where $\mathbb{D}([0,T]; \mathbb{R}^n)$ is the space of $\mathbb{R}^n$-valued càdlàg functions on $[0,T]$ endowed with the uniform metric d_∞, and $\|\mathfrak{f}\|_{\mathrm{Lip}}$ is the Lipschitz constant of $\mathfrak{f}$. (cf. Lemma 3.6.3).*

We now characterize the parameter β in the concentration inequality (3.3.8) based on Lemmas 3.3.5, 3.3.6 and 3.3.8.

Proposition 3.3.9. *Suppose (H.4) holds and let $\mathcal{K} \subset \mathbb{R}^{n\times(n+k)}$ be a bounded set. Then there exist constants $C_1, C_2 \geq 0$ such that (3.3.8) holds for all $\varepsilon \geq 0$, $m \in \mathbb{N}$ and $\theta \in \mathcal{K}$ with $\beta = 3 + \vartheta$, where ϑ is the constant in (H.4(2)).*

Proof. Throughout this proof, let θ be a given constant satisfying $|\theta| \leq K$ for some $K \geq 0$. For notational simplicity, we shall omit the dependence on (x_0, θ) in the subscripts of all random variables, and denote by C_2 a generic constant, which is independent of m and the precise value of θ, and depends possibly on K, x_0, the constants in (H.4) and the dimensions.

Note that for each $i = 1, \dots, m$, the entries of $\int_0^T Z_t^i (Z_t^i)^\mathsf{T} \, \mathrm{d}t$ are one of the three cases:

$$\int_0^T X_{\ell,t}^i X_{j,t}^i \, \mathrm{d}t, \quad \int_0^T X_{\ell,t}^i \psi^\theta(t, X_t^i)_j \, \mathrm{d}t, \quad \int_0^T \psi^\theta(t, X_t^i)_\ell \psi^\theta(t, X_t^i)_j \, \mathrm{d}t \tag{3.3.10}$$

where $X_{\ell,t}^i$ and $\psi^\theta(t, X_t^i)_\ell$ are the ℓth-entry of $X_{\ell,t}^i$ and $\psi^\theta(t, X_t^i)_\ell$, respectively. Similarly, the entries of $\int_0^T Z_t^i (\mathrm{d}X_t^i)^\mathsf{T}$ are one of the two cases:

$$\begin{aligned}
&\int_0^T X_{\ell,t}^i (A^\star X_t^i)_j \, \mathrm{d}t + \int_0^T X_{\ell,t}^i (B^\star \psi^\theta(t, X_t^i))_j \, \mathrm{d}t + \int_0^T X_{\ell,t}^i \sigma_j \, \mathrm{d}W_t^i + \int_0^T \int_{\mathbb{R}_0^p} X_{\ell,t}^i \gamma(u)_j \, \tilde{N}^i(\mathrm{d}t, \mathrm{d}u), \\
&\int_0^T \psi^\theta(t, X_t^i)_\ell (A^\star X_t^i)_j \, \mathrm{d}t + \int_0^T \psi^\theta(t, X_t^i)_\ell (B^\star \psi^\theta(t, X_t^i))_j \, \mathrm{d}t + \int_0^T \psi^\theta(t, X_t^i)_\ell \sigma_j \, \mathrm{d}W_t^i \\
&\quad + \int_0^T \int_{\mathbb{R}_0^p} \psi^\theta(t, X_t^i)_\ell \gamma(u)_j \, \tilde{N}^i(\mathrm{d}t, \mathrm{d}u),
\end{aligned} \tag{3.3.11}$$

where σ_j is the j-th row of $\sigma \in \mathbb{R}^{n\times d}$, γ_j is the j-th entry of the function $\gamma : \mathbb{R}_0^p \to \mathbb{R}^n$, $(W^i)_{i=1}^m$ are m-independent d-dimensional Brownian motion, and $(\tilde{N}^i)_{i=1}^m$ are m-independent compensated Poisson random measures. By the definitions of $U^{x_0,\theta,m}, V^{x_0,\theta,m}$ in (3.3.7), and the inequality that $\mathbb{P}(|\sum_{i=1}^\ell X_i| \geq \varepsilon) \leq \sum_{i=1}^\ell \mathbb{P}(|X_i| \geq \varepsilon/\ell)$ for all $\ell \in \mathbb{N}$ and random variables $(X_i)_{i=1}^\ell$, it suffices to obtain a concentration inequality for each term in (3.3.10) and (3.3.11).

Since $|\theta| \leq K$, by Theorem 3.2.5, there exists $C_2 \geq 0$ such that $|\psi^\theta(t, 0)| \leq C_2$ and $|\psi^\theta(t, x) - \psi^\theta(t, x')| \leq C_2|x - x'|$ for all $t \in [0, T]$, $x, x' \in \mathbb{R}^n$. Then standard moment estimates of (3.3.4) (with the initial condition x_0) shows that $\|X^i\|_{\mathcal{S}^2(\mathbb{R}^n)} \leq C_2$ for all $i = 1, \dots, m$, with a constant C_2 depending on x_0. Then, for each $q \geq 2$, $\ell = 1, \dots, n$ and $j = 1, \dots k$, we consider the functions $\mathfrak{f}_\ell^{(q)}, \bar{\mathfrak{f}}_j^{(q)} : (\mathbb{D}([0, T]; \mathbb{R}^n), d_\infty) \to \mathbb{R}$ satisfying for all $\rho \in \mathbb{D}([0, T]; \mathbb{R}^n)$ that $\mathfrak{f}_\ell^{(q)}(\rho) = \big(\int_0^T |\rho_{\ell,t}|^q \mathrm{d}t\big)^{\frac{1}{q}}$ and $\bar{\mathfrak{f}}_j^{(q)}(\rho) = \big(\int_0^T |\psi^\theta(t, \rho_t)_j|^q \mathrm{d}t\big)^{\frac{1}{q}}$, where $\rho_{\ell,t}$ is the ℓth component of ρ_t and $\psi^\theta(t, \rho_t)_j$ is the jth component of $\psi^\theta(t, \rho_t)$. One can easily show that $\mathfrak{f}_\ell^{(q)}(0) = 0$ and $|\bar{\mathfrak{f}}_j^{(q)}(0)|, \|\mathfrak{f}_\ell^{(q)}\|_{\mathrm{Lip}}, \|\bar{\mathfrak{f}}_j^{(q)}\|_{\mathrm{Lip}} \leq C$, which along with Lemma 3.3.8 implies that $\|\big(\int_0^T |X_{\ell,t}^i|^q \, \mathrm{d}t\big)^{\frac{1}{q}}\|_{\Psi_1} \leq C$ and $\|\big(\int_0^T |\psi^\theta(t, X_t^i)_j|^q \, \mathrm{d}t\big)^{\frac{1}{q}}\|_{\Psi_1} \leq C$, uniformly with

respect to i, ℓ, j, q, θ. Hence, we can obtain from Lemmas 3.3.5 and 3.3.6 a uniform bound for the $\|\cdot\|_{\Psi_{1/(3+\vartheta)}}$-norms of all the terms in (3.3.10) and (3.3.11).

Consequently, we can deduce the desired concentration inequality by applying Lemma 3.6.4 (with $\alpha = 1/(3+\vartheta)$, $N = m$ and $\varepsilon' = m\varepsilon$) to each component of the zero-mean random varables $\big(\int_0^T Z_t^i (Z_t^i)^\intercal \,\mathrm{d}t - \overline{U}\big)_{i=1}^m$ and $\big(\int_0^T Z_t^i (\mathrm{d}X_t^i)^\intercal - \overline{V}\big)_{i=1}^m$. □

The following proposition improves the concentration inequality in Proposition 3.3.9 for the case without jumps.

Proposition 3.3.10. *Suppose (H.4) holds with $\gamma_{\max} = 0$ and let $\mathcal{K} \subset \mathbb{R}^{n\times(n+k)}$ be a bounded set. Then there exist constants $C_1, C_2 \geq 0$ such that (3.3.8) holds for all $\varepsilon \geq 0$, $m \in \mathbb{N}$ and $\theta \in \mathcal{K}$ with $\beta = 1$.*

Proof. We first refine the result of Lemma 3.3.8 and prove Lipschitz functionals of the state process $X^{x_0,\theta}$ is sub-Gaussian. By [49, Theorem 1.1 and Corollary 4.1], there exists $C \geq 0$ such that for all $x_0 \in \mathbb{R}^n$ and for every Lipschitz continuous function $\mathfrak{f} : (\mathbb{D}([0,T];\mathbb{R}^n), d_\infty) \to \mathbb{R}$ with $\|\mathfrak{f}\|_{\mathrm{Lip}} \leq 1$, $\mathbb{E}\big[\exp\big(\lambda(\mathfrak{f}(X^{x_0,\theta}) - \mathbb{E}[\mathfrak{f}(X^{x_0,\theta})])\big)\big] \leq \exp\big(C^2\lambda^2\big)$ for all $\lambda > 0$, which along with [156, Proposition 2.5.2 (v)] implies that $\|\mathfrak{f}(X^{x_0,\theta}) - \mathbb{E}[\mathfrak{f}(X^{x_0,\theta})]\|_{\Psi_2} \leq C$ for some constant C, uniformly with respect to $x_0 \in \mathbb{R}^n$, $\theta \in \mathcal{K}$ and $\mathfrak{f} : (\mathbb{D}([0,T];\mathbb{R}^n), d_\infty) \to \mathbb{R}$ satisfying $\|\mathfrak{f}\|_{\mathrm{Lip}} \leq 1$. Then, we can deduce from the fact that $\|\cdot\|_{\Psi_2}$ is a norm that $\|\mathfrak{f}(X^{x_0,\theta})\|_{\Psi_2} \leq C(\|\mathfrak{f}\|_{\mathrm{Lip}} + |\mathbb{E}[\mathfrak{f}(X^{x_0,\theta})]|)$ for all $x_0 \in \mathbb{R}^n$, $\theta \in \mathcal{K}$ and Lipschitz continuous functions $\mathfrak{f}$.

We then proceed along the proof of Proposition 3.3.9. For each $i = 1, \ldots, m$, all entries of $\int_0^T Z_t^i (Z_t^i)^\intercal \,\mathrm{d}t$ are given in (3.3.10), and all entries of $\int_0^T Z_t^i (\mathrm{d}X_t^i)^\intercal$ are given by (cf. (3.3.11)):

$$\begin{aligned}
&\int_0^T X_{\ell,t}^i (A^\star X_t^i)_j \,\mathrm{d}t + \int_0^T X_{\ell,t}^i (B^\star \psi^\theta(t, X_t^i))_j \,\mathrm{d}t + \int_0^T X_{\ell,t}^i \sigma_j \,\mathrm{d}W_t^i,\\
&\int_0^T \psi^\theta(t, X_t^i)_\ell (A^\star X_t^i)_j \,\mathrm{d}t + \int_0^T \psi^\theta(t, X_t^i)_\ell (B^\star \psi^\theta(t, X_t^i))_j \,\mathrm{d}t + \int_0^T \psi^\theta(t, X_t^i)_\ell \sigma_j \,\mathrm{d}W_t^i,
\end{aligned} \tag{3.3.12}$$

for all $\ell = 1, \ldots, n$ and $j = 1, \ldots, k$, where we have omitted the dependence on (x_0, θ) in the subscripts for notational simplicity. Hence, by following the same argument as in Proposition 3.3.9, we can show there exists a constant C, such that for all $i = 1, \ldots, m$, $\ell = 1, \ldots, n$, $j = 1, \ldots, k$ and $\theta \in \mathcal{K}$, we have $\|\big(\int_0^T |X_{\ell,t}^i|^2 \,\mathrm{d}t\big)^{\frac{1}{2}}\|_{\Psi_2} \leq C$ and $\|\big(\int_0^T |\psi^\theta(t, X_t^i)_j|^2 \,\mathrm{d}t\big)^{\frac{1}{2}}\|_{\Psi_2} \leq C$. Then, we can obtain from Lemmas 3.3.5 and 3.3.7 a uniform bound for the $\|\cdot\|_{\Psi_1}$-norms of all entires of $\int_0^T Z_t^i (Z_t^i)^\intercal \,\mathrm{d}t$ and $\int_0^T Z_t^i (\mathrm{d}X_t^i)^\intercal$. Consequently, we can apply Lemma 3.6.4 (with $\alpha = 1$, $N = m$ and $\varepsilon' = m\varepsilon$) to each entry of $\big(\int_0^T Z_t^i (Z_t^i)^\intercal \,\mathrm{d}t - \overline{U}\big)_{i=1}^m$ and $\big(\int_0^T Z_t^i (\mathrm{d}X_t^i)^\intercal - \overline{V}\big)_{i=1}^m$, and deduce the desired concentration inequality. □

3.3.4.3 Step 3: Proof of general regret bounds

After demonstrating how to verify (3.3.8) based on the precise jump sizes in the state dynamics, it remains to establish the general regret result in Theorem 3.3.2 under the assumption that (3.3.8) holds for some $\beta \geq 1$.

We start by showing that under (H.4(1)) and (H.5), the expression (3.3.5) is well-defined if θ is a sufficiently accurate estimation of the true parameter $\theta^\star$.

Lemma 3.3.11. *Suppose (H.4(1)) and (H.5) hold. Then there exist constants $\varepsilon_0 > 0$ and $\tau_0 > 0$, such that for all $\theta \in \mathcal{K}_0 := \{\theta \in \mathbb{R}^{n\times(n+k)} \mid |\theta - \theta^\star| \leq \varepsilon_0\}$, we have $\lambda_{\min}(\overline{U}^{x_0,\theta}) \geq \tau_0$, where $\overline{U}^{x_0,\theta}$ is defined as in (3.3.7) and $\lambda_{\min}(A)$ is the smallest eigenvalue of a symmetric matrix A.*

Proof. Since $\overline{U}^{x_0,\theta^\star}$ is positive semidefinite, we shall prove $\lambda_{\min}(\overline{U}^{x_0,\theta^\star}) > 0$ by assuming that $\lambda_{\min}(\overline{U}^{x_0,\theta^\star}) = 0$. Then we see there exists a non-zero vector $u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \in \mathbb{R}^{n+k}$ with $u_1 \in \mathbb{R}^n$ and $u_2 \in \mathbb{R}^k$, such that $u^\intercal \overline{U}^{x_0,\theta^\star} u = 0$. By the definition of $\overline{U}^{x_0,\theta^\star}$ in (3.3.7), we can deduce that $\mathbb{E}[\int_0^T |u^\intercal Z_t^{x_0,\theta^\star}|^2 \,\mathrm{d}t] = 0$, which along with the definition of $Z_t^{x_0,\theta^\star}$ in (3.3.5) implies for $\mathrm{d}\mathbb{P} \otimes \mathrm{d}t$ a.e. that $u_1^\intercal X_t^{x_0,\theta^\star,\alpha^\star} + u_2^\intercal \alpha_t^{x_0,\star} = 0$. This contradicts to (H.5), which leads to the desired inequality that $\lambda_{\min}(\overline{U}^{x_0,\theta^\star}) > 0$.

We then show that the map $\mathbb{R}^{n\times(n+k)} \ni \theta \mapsto \overline{U}^{x_0,\theta} \in \mathbb{R}$ is continuous. Theorem 3.2.7 shows that the map $\mathbb{R}^{n\times(n+k)} \ni \theta \mapsto X^{x_0,\theta} \in \mathcal{H}^2(\mathbb{R}^n)$ is continuous. Moreover, Theorems 3.2.5 and 3.2.6 imply that there exists a constant $C \geq 0$, such that for all $\theta \in \mathbb{R}^{n\times(n+k)}$ satisfying $|\theta - \theta^\star| \leq 1$, $t \in [0,T]$ and $x, x' \in \mathbb{R}^n$, we have that $|\psi^\theta(t,0)| \leq C$, $|\psi^\theta(t,x) - \psi^\theta(t,x')| \leq C|x-x'|$ and $|\psi^\theta(t,x) - \psi^{\theta^\star}(t,x)| \leq C(1+|x|)|\theta-\theta^\star|$, from which we can deduce that

$$\begin{aligned} |\psi^\theta(t,x) - \psi^{\theta^\star}(t,x')| &\leq |\psi^\theta(t,x) - \psi^{\theta^\star}(t,x)| + |\psi^{\theta^\star}(t,x) - \psi^{\theta^\star}(t,x')| \\ &\leq C(1+|x|)|\theta-\theta^\star| + C|x-x'|. \end{aligned}$$

Hence, for all $\theta \in \mathbb{R}^{n\times(n+k)}$ with $|\theta - \theta^\star| \leq 1$,

$$\|\psi^\theta(\cdot, X^{x_0,\theta}) - \psi^{\theta^\star}(\cdot, X^{x_0,\theta^\star})\|_{\mathcal{H}^2} \leq C(1 + \|X^{x_0,\theta}\|_{\mathcal{H}^2})|\theta-\theta^\star| + C\|X^{x_0,\theta} - X^{x_0,\theta^\star}\|_{\mathcal{H}^2},$$

which along with the continuity of the map $\mathbb{R}^{n\times(n+k)} \ni \theta \mapsto X^{x_0,\theta} \in \mathcal{H}^2(\mathbb{R}^n)$ implies that the map $\mathbb{R}^{n\times(n+k)} \ni \theta \mapsto \psi^\theta(\cdot, X^{x_0,\theta}) \in \mathcal{H}^2(\mathbb{R}^k)$ is continuous. Since the entires of $\overline{U}^{x_0,\theta}$ involve only the expectations of products of $X^{x_0,\theta}$ and $\psi^\theta(\cdot, X^{x_0,\theta})$, the desired continuity of the map $\mathbb{R}^{n\times(n+k)} \ni \theta \mapsto \overline{U}^{x_0,\theta} \in \mathbb{R}$ follows.

Finally, by the continuity of the minimum eigenvalue function, clearly $\mathbb{R}^{n\times(n+k)} \ni \theta \mapsto \lambda_{\min}(\overline{U}^{x_0,\theta}) \in \mathbb{R}$ is continuous, which along with the fact that $\lambda_{\min}(\overline{U}^{x_0,\theta^\star}) > 0$ leads to the desired result. □

We then quantify the estimation error of the least-squares estimator (3.3.6) by assuming the concentration inequality (3.3.8) holds for the compact set $\mathcal{K}_0$ in Lemma 3.3.11.

Proposition 3.3.12. *Suppose (H.4(1)) and (H.5) hold. Let $\mathcal{K}_0$ be the set in Lemma 3.3.11. Assume further that there exist constants $C_1, C_2 > 0$ and $\beta \geq 1$ such that (3.3.8) holds for all $\varepsilon \geq 0$, $m \in \mathbb{N}$ and $\theta \in \mathcal{K}_0$. Then there exist constants $\bar{C}_1, \bar{C}_2 \geq 0$, such that for all $\theta \in \mathcal{K}_0$ and $\delta \in (0, 1/2)$, if $m \geq \bar{C}_1(-\ln \delta)^\beta$, then we have with probability at least $1 - 2\delta$ that*

$$|\hat{\theta} - \theta^\star| \leq \bar{C}_2 \left(\sqrt{\frac{-\ln \delta}{m}} + \frac{(-\ln \delta)^\beta}{m} + \frac{(-\ln \delta)^{2\beta}}{m^2} \right), \tag{3.3.13}$$

where $\hat{\theta}$ denotes the transpose of the left-hand side of (3.3.6) associated with θ.

Proof. Throughout the proof, let $\delta \in (0, 1/2)$ and $\theta \in \mathcal{K}_0$ be fixed and let $\|\cdot\|_2$ be the matrix norm induced by Euclidean norms. The invertibility of $\overline{U}^{x_0,\theta}$ (see Lemma 3.3.11) implies that (3.3.5) is well-defined, which along with (3.3.6) leads to

$$\begin{aligned} \|\hat{\theta} - \theta^\star\|_2 &= \|(U^{x_0,\theta,m} + \tfrac{1}{m}\mathbb{I})^{-1} V^{x_0,\theta,m} - (\overline{U}^{x_0,\theta})^{-1} \overline{V}^{x_0,\theta}\|_2 \\ &\leq \|(U^{x_0,\theta,m} + \tfrac{1}{m}\mathbb{I})^{-1} - (\overline{U}^{x_0,\theta})^{-1}\|_2 \|V^{x_0,\theta,m}\|_2 \\ &\quad + \|(\overline{U}^{x_0,\theta})^{-1}\|_2 \|V^{x_0,\theta,m} - \overline{V}^{x_0,\theta}\|_2. \end{aligned} \tag{3.3.14}$$

We now estimate each term in the right-hand side of (3.3.14). By Lemma 3.3.11, $\lambda_{\min}(\overline{U}^{x_0,\theta}) \geq \tau_0$ for some $\tau_0 > 0$, which implies that $\|(\overline{U}^{x_0,\theta})^{-1}\|_2 \leq 1/\tau_0$. Moreover, by setting the right-hand side of (3.3.8) to be δ, we can deduce with probability at least $1 - 2\delta$ that $|U^{x_0,\theta,m} - \overline{U}^{x_0,\theta}| \leq \delta_m$ and $|V^{x_0,\theta,m} - \overline{V}^{x_0,\theta}| \leq \delta_m$ with the constant δ_m given by

$$\delta_m := \max\left\{ \left(\frac{C_2^2}{C_1 m} \ln\left(\frac{C_2}{\delta}\right) \right)^{\frac{1}{2}}, \frac{C_2}{m} \left(\frac{1}{C_1} \ln\left(\frac{C_2}{\delta}\right) \right)^{\beta} \right\}, \tag{3.3.15}$$

where we have assumed without loss of generality that $C_2 \geq 1$.

Let m be a sufficiently large constant satisfying $\delta_m + 1/m \leq \tau_0/2$. The fact that $\|\cdot\|_2 \leq |\cdot|$ indicates with probability at least $1 - 2\delta$ that $\|U^{x_0,\theta,m} + \frac{1}{m}\mathbb{I} - \overline{U}^{x_0,\theta}\|_2 \leq \frac{1}{m} + \delta_m \leq \frac{\tau_0}{2}$, which in turn yields

$$\lambda_{\min}(U^{x_0,\theta,m} + \tfrac{1}{m}\mathbb{I}) \geq \lambda_{\min}(\overline{U}^{x_0,\theta}) - \|U^{x_0,\theta,m} + \tfrac{1}{m}\mathbb{I} - \overline{U}^{x_0,\theta}\|_2 \geq \tfrac{\tau_0}{2},$$

or equivalently $\|(U^{x_0,\theta,m} + \frac{1}{m}\mathbb{I})^{-1}\|_2 \leq 2/\tau_0$. Then, since $A^{-1} - (A+B)^{-1} = (A+B)^{-1}BA^{-1}$ for all nonsingular matrices A and $A + B$, we have with probability at least $1 - 2\delta$ that,

$$\begin{aligned} \|(U^{x_0,\theta,m} + \tfrac{1}{m}\mathbb{I})^{-1} - (\overline{U}^{x_0,\theta})^{-1}\|_2 &= \|(\overline{U}^{x_0,\theta} + U^{x_0,\theta,m} + \tfrac{1}{m}\mathbb{I} - \overline{U}^{x_0,\theta})^{-1} - (\overline{U}^{x_0,\theta})^{-1}\|_2 \\ &\leq \|(U^{x_0,\theta,m} + \tfrac{1}{m}\mathbb{I})^{-1}\|_2 \|(\overline{U}^{x_0,\theta})^{-1}\|_2 \|(U^{x_0,\theta,m} + \tfrac{1}{m}\mathbb{I}) - \overline{U}^{x_0,\theta}\|_2 \\ &\leq \tfrac{2}{\tau_0^2}(\tfrac{1}{m} + \delta_m), \end{aligned}$$

which along with the inequality that $\|V^{x_0,\theta,m}\|_2 \le \|\overline{V}^{x_0,\theta}\|_2 + |V^{x_0,\theta,m} - \overline{V}^{x_0,\theta}|$ allows us to derive the following estimate from (3.3.14):

$$\|\hat{\theta} - \theta^\star\|_2 \le \tfrac{2}{\tau_0^2}(\tfrac{1}{m} + \delta_m)(\|\overline{V}^{x_0,\theta}\|_2 + \delta_m) + \tfrac{\delta_m}{\tau_0}.$$

Note that $\|\overline{V}^{x_0,\theta}\|_2$ is uniformly bounded for all $\theta \in \mathcal{K}_0$ by the compactness of $\mathcal{K}_0$ and the continuity of the map $\theta \mapsto \overline{V}^{x_0,\theta}$ (cf. Lemma 3.3.11). Thus, by the condition that $\beta \ge 1$ and the definition of δ_m in (3.3.15), we see that there exists a constant $\bar{C}_2$, depending only on C_1, C_2, β, τ_0, and the constants in (H.4(1)), such that the desired estimate (3.3.13) holds with probability at least $1 - 2\delta$, provided that m satisfies $\delta_m + 1/m \le \tau_0/2$. Since $\beta \ge 1$ and $\delta \le 1/2$, we see there exists $\bar{C}_1 \ge 0$, independent of m, δ and θ, such that the inequality (3.3.13) holds for all m satisfying $m \ge \bar{C}_1(-\ln\delta)^\beta$. □

Now we are ready to present the proof of Theorem 3.3.2.

Proof of Theorem 3.3.2. We start by proving Item (1). Then, by the assumptions that $\lambda_{\min}(\overline{U}^{x_0,\theta_0}) > 0$ and (3.3.8) holds for $\mathcal{K} = \overline{\mathcal{K}}_0 := \mathcal{K}_0 \cup \{\theta_0\}$ with $\mathcal{K}_0$ from Lemma 3.3.11, we can extend Proposition 3.3.12 to show that (3.3.13) holds for all $\theta \in \overline{\mathcal{K}}_0$, $\delta \in (0, 1/2)$ and $m \ge \bar{C}_1(-\ln\delta)^\beta$, with some constants $\bar{C}_1, \bar{C}_2 \ge 1$ depending on $\overline{\mathcal{K}}_0$. In the subsequent analysis, we fix $\delta \in (0, 1/4)$ and for all $\ell \in \mathbb{N} \cup \{0\}$, we define $\delta_\ell = 2^{-\ell}\delta$, and let $\theta_{\ell+1}$ be generated by using (3.3.6) with $m = m_\ell$ and $\theta = \theta_\ell$. We shall specify the precise choice of m_0 later.

In the sequel, we assume without loss of generality that $\varepsilon_0/(3\bar{C}_2) \le 1$ and $\bar{C}_2/\varepsilon_0 \ge \bar{C}_1$, where $\varepsilon_0 > 0$ is the constant in the definition of $\mathcal{K}_0$ (see Lemma 3.3.11). We first show that there exists $\hat{C}_0 > 0$, independent of δ, such that if $m_0 \ge \hat{C}_0(-\ln\delta)^\beta$, then for all $\ell \in \mathbb{N}\cup\{0\}$,

$$\bar{C}_2\left(\sqrt{\frac{-\ln\delta_\ell}{m_\ell}} + \frac{(-\ln\delta_\ell)^\beta}{m_\ell} + \frac{(-\ln\delta_\ell)^{2\beta}}{m_\ell^2}\right) \le \varepsilon_0. \tag{3.3.16}$$

By the assumption that $\varepsilon_0/(3\bar{C}_2) \le 1$, it suffices to show that for all $\ell \in \mathbb{N}\cup\{0\}$, $-\ln\delta_\ell/m_\ell \le (\varepsilon_0/(3\bar{C}_2))^2$ and $(-\ln\delta_\ell)^\beta/m_\ell \le \varepsilon_0/(3\bar{C}_2)$. Given $\beta \ge 1$ and $\delta_\ell < 1/4$, it suffices to ensure $m_\ell \ge C(-\ln\delta_\ell)^\beta$ for all $\ell \in \mathbb{N}\cup\{0\}$, where C is a sufficiently large constant independent of δ and ℓ. By the definitions of $(\delta_\ell)_{\ell\in\mathbb{N}}$ and $(m_\ell)_{\ell\in\mathbb{N}}$ and the fact that $\delta < 1/4$, the desired condition can be achieved by choosing $m_0 \ge \hat{C}_0(-\ln\delta)^\beta$, for a sufficiently large constant $\hat{C}_0$ satisfying

$$\sup_{\ell\in\mathbb{N}\cup\{0\},\delta\in(0,\frac{1}{4})} \frac{(-\ln(2^{-\ell}\delta))^\beta}{2^\ell(-\ln\delta)^\beta} = \sup_{\ell\in\mathbb{N}\cup\{0\},\delta\in(0,\frac{1}{4})} 2^{-\ell}\Big(\frac{\ell\ln 2}{-\ln\delta} + 1\Big)^\beta \le \sup_{\ell\in\mathbb{N}\cup\{0\}} 2^{-\ell}\Big(\frac{\ell}{2} + 1\Big)^\beta \le \hat{C}_0 < \infty.$$

Now we choose $m_0 \ge \max(\hat{C}_0, \bar{C}_1)(-\ln\delta)^\beta$, and show by induction that for all $k \in \mathbb{N}\cup\{0\}$, it holds with probability at least $1 - 2\sum_{\ell=0}^{k-1}\delta_\ell$ that $\theta_\ell \in \overline{\mathcal{K}}_0$ for all $\ell = 0, \dots, k$ and

$$|\theta_k - \theta^\star|_2 \le \begin{cases} |\theta_0 - \theta^\star|_2, & k = 0, \\ \bar{C}_2\left(\sqrt{\frac{-\ln\delta_k}{m_k}} + \frac{(-\ln\delta_k)^\beta}{m_k} + \frac{(-\ln\delta_k)^{2\beta}}{m_k^2}\right), & k \in \mathbb{N}. \end{cases} \tag{3.3.17}$$

The statement clearly holds for $k = 0$. Now suppose that the induction statement holds for some $k \in \mathbb{N} \cup \{0\}$. Conditioning on $\theta_k \in \overline{\mathcal{K}}_0$, we can apply (3.3.13) with $\theta = \theta_k$, $\delta = \delta_k < 1/2$ and $m = m_k \geq \bar{C}_1(-\ln \delta_k)^\beta$ (see (3.3.16) and $\bar{C}_2/\varepsilon_0 \geq \bar{C}_1$), and deduce with probability at least $1 - 2\delta_k$ that (3.3.17) holds for the index $k+1$, which along with (3.3.16) shows that $\theta_{k+1} \in \mathcal{K}_0 \subset \overline{\mathcal{K}}_0$. Since the induction hypothesis implies that $\theta_k \in \overline{\mathcal{K}}_0$ holds with probability at least $1 - 2\sum_{\ell=0}^{k-1} \delta_\ell$, one can deduce that the induction statement also holds $k+1$.

The above induction argument shows that if $m_0 = C(-\ln \delta)^\beta$ for any constant $C \geq C_0 := \max(\hat{C}_0, \bar{C}_1)$, then with probability at least $1 - 2\sum_{\ell=0}^{\infty} \delta_\ell = 1 - 4\delta$, $\theta_k \in \overline{\mathcal{K}}_0$ and (3.3.17) holds for all $k \in \mathbb{N} \cup \{0\}$. Now let us assume such a setting, and observe that the i-th trajectory is generated with control ψ^{θ_ℓ} if $i \in (\sum_{j=0}^{\ell-1} m_j, \sum_{j=0}^{\ell} m_j] = (m_0(2^\ell - 1), m_0(2^{\ell+1} - 1)]$ for $\ell \in \mathbb{N} \cup \{0\}$ (cf. Algorithm 10). Then we can apply Theorem 3.2.7 and deduce for all $N \in \mathbb{N}$ that

$$\begin{aligned} R(N) &\leq \sum_{\ell=0}^{\lceil \log_2(\frac{N}{m_0}+1)\rceil - 1} m_\ell \Big(J^{\theta^\star}(\psi^{\theta_\ell}; x_0) - V(x_0; \theta^\star) \Big) \leq C' \sum_{\ell=0}^{\lceil \log_2(\frac{N}{m_0}+1)\rceil - 1} m_\ell |\theta_\ell - \theta^\star| \\ &\leq C' m_0 + C' \sum_{\ell=1}^{\lceil \log_2(\frac{N}{m_0}+1)\rceil - 1} \left(\sqrt{(-\ln \delta_\ell) m_\ell} + (-\ln \delta_\ell)^\beta \Big(1 + \frac{(-\ln \delta_\ell)^\beta}{m_\ell} \Big) \right) \\ &\leq C'(-\ln \delta)^\beta + C' \sum_{\ell=1}^{\lceil \log_2(\frac{N}{m_0}+1)\rceil - 1} \left(\sqrt{(-\ln \delta_\ell) m_\ell} + (-\ln \delta_\ell)^\beta \right), \end{aligned} \tag{3.3.18}$$

where we have denoted by C' a generic constant independent of ℓ, N, δ, and used the fact that $(-\ln \delta_\ell)^\beta / m_\ell \leq C'$ for the last inequality (cf. the choice of $\hat{C}_0$). We then derive an upper bound of (3.3.18). By virtue of the inequality that $\sqrt{(-\ln \delta_\ell) m_\ell} = \sqrt{(\ell \ln 2 - \ln \delta) 2^\ell m_0} \leq C' \sqrt{(\ell - \ln \delta) m_0} \sqrt{2}^\ell$ for all $\ell \in \mathbb{N}$, we have

$$\begin{aligned} \sum_{\ell=1}^{\lceil \log_2(\frac{N}{m_0}+1)\rceil - 1} \sqrt{(-\ln \delta_\ell) m_\ell} &\leq C' \sqrt{(\ln N - \ln \delta) m_0} \sqrt{2}^{\log_2(\frac{N}{m_0}+1)} \\ &\leq C' \sqrt{(\ln N - \ln \delta)(N + (-\ln \delta)^\beta)}. \end{aligned}$$

Moreover, by $\ln \delta_\ell = -\ell \ln 2 + \ln \delta$ and Hölder's inequality,

$$\sum_{\ell=1}^{\lceil \log_2(\frac{N}{m_0}+1)\rceil - 1} (-\ln \delta_\ell)^\beta \leq \sum_{\ell=1}^{C' \ln N} C' \big((\ell \ln 2)^\beta + (-\ln \delta)^\beta \big) \leq C' \big((\ln N)^{\beta+1} + \ln N (-\ln \delta)^\beta \big).$$

Consequently, from (3.3.18), $\beta \geq 1$ and the inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for all $x, y \geq 0$, it is clear for all $N \in \mathbb{N}$, $R(N) \leq C' \big(\sqrt{N} \sqrt{\ln N} + \sqrt{-\ln \delta} \sqrt{N} + (-\ln \delta)^\beta \ln N \big)$ for some constant C' independent of β and N, which finishes the proof of Item (1).

We are ready to show Item (2). For each $N \in \mathbb{N} \cap [3, \infty)$, we define $\delta_N = 1/N^2$ and the event $A_N = \{R(N) > C'(\sqrt{N}\sqrt{\ln N} + \sqrt{-\ln \delta_N}\sqrt{N} + (-\ln \delta_N)^\beta \ln N)\}$. Item (1) shows that $\sum_{N=3}^\infty \mathbb{P}(A_N) \le 4 \sum_{N=3}^\infty \delta_N < \infty$. Hence, from the Borel-Cantelli lemma, $\mathbb{P}(\limsup_{N\to\infty} A_N) = 0$, which along with the definition of δ_N implies the desired conclusion. □

3.4 Extension: RL problems with controlled diffusion

In this section, we extend our framework to analyze the regret order of learning algorithms for general continuous-time RL problems, whose state dynamics involves controlled diffusion. To simplify the presentation, we focus on entropy-regularized problems studied in [159, 143, 135, 152] and outline the essential steps of the argument.

For each $\theta = (A, B) \in \mathbb{R}^{n\times(n+k)}$, define $V(\cdot;\theta) : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ by

$$V(t,x;\theta) \coloneqq \inf_{\alpha\in\mathcal{H}^2(\mathbb{R}^k)} \mathbb{E}\left[\int_t^T f(s, X_s^{t,x,\alpha}, \alpha_s)\,\mathrm{d}s + g(X_T^{t,x,\alpha})\right], \quad \forall (t,x) \in [0,T]\times\mathbb{R}^n, \tag{3.4.1}$$

where for each $\alpha \in \mathcal{H}^2(\mathbb{R}^k)$, $X^{t,x,\alpha} \in \mathcal{S}^2(\mathbb{R}^n)$ satisfies the controlled dynamics:

$$\mathrm{d}X_s = (AX_s + B\alpha_s)\,\mathrm{d}s + \sigma(s, X_s, \alpha_s)\,\mathrm{d}W_s, \quad s \in [t,T], \quad X_t = x. \tag{3.4.2}$$

The functions $f : [0,T]\times\mathbb{R}^n\times\mathbb{R}^k \to \mathbb{R}\cup\{\infty\}$ and $\sigma : [0,T]\times\mathbb{R}^n\times\mathbb{R}^k \to \mathbb{R}^{n\times d}$ are such that for all $(t,x) \in [0,T]\times\times\mathbb{R}^n$ and $a = (a_i)_{i=1}^k \in \mathbb{R}^k$,

$$f(t,x,a) = \sum_{i=1}^k \overline{f}_i(t,x)a_i + \mathcal{R}_{\mathrm{en}}(a), \quad \sigma(t,x,a)\sigma(t,x,a)^\mathsf{T} = \sum_{i=1}^k \overline{\sigma}_i(t,x)\overline{\sigma}_i(t,x)^\mathsf{T} a_i, \tag{3.4.3}$$

where for each $i = 1,\dots k$, $\overline{f}_i : [0,T]\times\mathbb{R}^n \to \mathbb{R}$, $\overline{\sigma}_i : [0,T]\times\mathbb{R}^n \to \mathbb{R}^{n\times d}$ are some given functions and $\mathcal{R}_{\mathrm{en}} : \mathbb{R}^k \to \mathbb{R}\cup\{\infty\}$ is Shannon's entropy function (cf. Example 3.2.2) such that

$$\mathcal{R}_{\mathrm{en}}(a) = \begin{cases} \sum_{i=1}^k a_i \ln(a_i), & a \in \Delta_k \coloneqq \{a \in [0,1]^k \mid \sum_{i=1}^k a_i = 1\}, \\ \infty, & a \in \mathbb{R}^k \setminus \Delta_k. \end{cases} \tag{3.4.4}$$

To avoid needless technicalities, we assume $(\overline{f}_i)_{i=1}^k$, $(\overline{\sigma}_i)_{i=1}^k$ and g to be bounded and sufficiently regular as in Proposition 3.4.1.

Note that (3.4.4) restricts control processes to those taking values in Δ_k. Hence, if $\overline{\sigma}_\ell(t,x) \equiv \overline{\sigma}$ for some $\overline{\sigma} \in \mathbb{R}^{n\times d}$, then (3.4.1)-(3.4.2) is a special case of the linear-convex model studied in Sections 3.2-3.3. Consequently, Theorem 3.3.2 can be applied to study the regret order of GLS algorithms for (3.4.1)-(3.4.2) with given initial time and state $(t,x) \in [0,T]\times\mathbb{R}^n$ but with unknown parameter θ.

To analyze the regret order of learning algorithms with general σ, a crucial step is to extend Theorem 3.2.7 and quantify the performance of a greedy policy from an incorrect model.

The fact that control affects the diffusion coefficients complicates the stability analysis of optimal feedback controls (i.e., Theorem 3.2.6) for (3.4.1)-(3.4.2). The following proposition proves a linear performance gap under the condition that the value function $V(t,x;\theta)$ in (3.4.1) is sufficiently regular in t,x and θ. Recall that the first and second-order derivatives of a sufficiently regular value function can be represented by solutions to the associated FB-SDE (3.2.11) (see e.g., [169, Theorem 4.1, p. 250]). Hence the linear performance gap can also be established by assuming sufficient regularity of the solution process $(Y^{t,x},Z^{t,x})$ to (3.2.11), whose details are omitted here.

Proposition 3.4.1. *For each $\theta\in\mathbb{R}^{n\times(n+k)}$, let $V(\cdot;\theta):[0,T]\times\mathbb{R}^n\to\mathbb{R}$ be defined by (3.4.1). Suppose that $(\overline{f}_i)_{i=1}^k,(\overline{\sigma}_i)_{i=1}^k\subset C^{0,1}([0,T]\times\mathbb{R}^n)$, $g\in C^1(\mathbb{R}^n)$, and there exists $\mathfrak{m}:[0,\infty)\to[0,\infty)$ such that for all $(t,x)\in[0,T]\times\mathbb{R}^n$ and $\theta,\theta'\in\mathbb{R}^{n\times(n+k)}$, $\frac{\partial}{\partial t}V(\cdot;\theta)$ is continuous, $\|V(\cdot;\theta)\|_{C^{0,3}([0,T]\times\mathbb{R}^n)}\le\mathfrak{m}(|\theta|)$,*

$$\begin{aligned}&|\nabla_x V(t,x;\theta)-\nabla_x V(t,x;\theta')|+|\operatorname{Hess}_x V(t,x;\theta)-\operatorname{Hess}_x V(t,x;\theta')|\\&\qquad\le(\mathfrak{m}(|\theta|)+\mathfrak{m}(|\theta|'))|\theta-\theta'|(1+|x|).\end{aligned}$$

Then for all $\theta\in\mathbb{R}^{n\times(n+k)}$, there exists $\psi^\theta\in\mathcal{V}$ such that

(1) ψ^θ is an optimal feedback control of (3.4.1)-(3.4.2) satisfying for all $x_0\in\mathbb{R}^n$ and $\theta\in\mathbb{R}^{n\times(n+k)}$, $V(0,x_0;\theta)=J(\psi^\theta;x_0,\theta)$, where for each $\psi\in\mathcal{V}$,

$$J(\psi;x_0,\theta):=\mathbb{E}\left[\int_0^T f(t,X_t^{x_0,\theta,\psi},\psi(t,X_t^{x_0,\theta,\psi}))\,\mathrm{d}t+g(X_T^{x_0,\theta,\psi})\right],$$

and $X^{x_0,\theta,\psi}\in\mathcal{S}^2(\mathbb{R}^n)$ satisfies the following dynamics:

$$\mathrm{d}X_t=\theta\begin{pmatrix}X_t\\ \psi(t,X_t)\end{pmatrix}\mathrm{d}t+\sigma(t,X_t,\psi(t,X_t))\mathrm{d}W_t,\quad t\in[0,T],\quad X_0=x_0,$$

(2) for all $x_0\in\mathbb{R}^n$ and $R\ge0$ there exists a constant C such that for all $\theta,\theta'\in\mathbb{R}^{n\times(n+k)}$ with $|\theta|,|\theta'|\le R$,

$$|J(\psi^{\theta'};x_0,\theta)-J(\psi^\theta;x_0,\theta)|\le C|\theta'-\theta|.$$

Proposition 3.4.1 relies on the regularity and Lipschitz stability of the value function V. For instance, if all coefficients are bounded and sufficiently smooth, and σ satisfies the uniform parabolicity condition, then $C^{2+\alpha}$ regularity results for fully nonlinear parabolic PDEs (see e.g., the Evan-Kryolv theorem in [97, Theorems 6.4.3 and 6.4.4, p. 301]) and a bootstrap argument would ensure that for any given θ, the function $V(\cdot,\theta)$ is continuously differentiable in t and three-time continuously differentiable in x. Due to the unbounded drift coefficient of (3.4.2), the boundedness in the $C^{0,3}([0,T]\times\mathbb{R}^n)$-norm and the locally Lipschitz continuity of V in θ follow from an extension of the Schauder estimate (see e.g., [96]) to nonlinear parabolic equations with unbounded coefficients in the whole space.

With Proposition 3.4.1, we can then quantify the regrets of GLS algorithms (see Algorithm 10) for (3.4.1)-(3.4.2) with unknown drift parameter θ and known diffusion coefficient σ. By the boundedness of σ and the regularity of ψ^θ, one can prove Proposition 3.3.10 in the present setting. Hence, Theorem 3.3.2 (with $\beta = 1$ in (3.3.8)) shows that Algorithm 10 enjoys a sublinear regret as shown in Theorem 3.3.4.

Proof of Proposition 3.4.1. For any given $\theta = (A, B) \in \mathbb{R}^{n\times(n+k)}$, the regularity of $V(\cdot;\theta)$ and [169, Proposition 3.5, p. 182] imply that $V(\cdot;\theta)$ is the unique classical solution to the associated HJB equation. That is, for all $(t,x) \in [0,T) \times \mathbb{R}^d$,

$$\tfrac{\partial}{\partial t}V(t,x)+\inf_{a\in\Delta_k}\Big(\tfrac{1}{2}\mathrm{tr}(\sigma(t,x,a)\sigma(t,x,a)^\intercal \operatorname{Hess}_x V(t,x))+\langle Ax+Ba, \nabla_x V(t,x)\rangle+f(t,x,a)\Big)=0,$$

and $V(T,x) = g(x)$ for all $x \in \mathbb{R}^d$. By (3.4.3), for all $(t,x) \in [0,T] \times \mathbb{R}^n$,

$$\begin{aligned}\psi^\theta(t,x) &\coloneqq \arg\min_{a\in\Delta_k}\big(\tfrac{1}{2}\mathrm{tr}(\sigma(t,x,a)\sigma(t,x,a)^\intercal \operatorname{Hess}_x V(t,x;\theta)) + \langle Ba, \nabla_x V(t,x;\theta)\rangle + f(t,x,a)\big)\\ &= \nabla\mathcal{R}^*_{\mathrm{en}}\big(-\tfrac{1}{2}\mathrm{tr}(\overline{\sigma}(t,x)\overline{\sigma}(t,x)^\intercal \operatorname{Hess}_x V(t,x;\theta)) - B^\intercal\nabla_x V(t,x;\theta) - \overline{f}(t,x)\big),\end{aligned}$$

where for all $z \in \mathbb{R}^k$, $\mathcal{R}^*_{\mathrm{en}}(z) = \sup_{a\in\Delta_k}(\langle a, z\rangle - \mathcal{R}_{\mathrm{en}}(a)) = \ln\sum_{i=1}^k \exp(z_i)$, and

$$\mathrm{tr}(\overline{\sigma}(t,x)\overline{\sigma}(t,x)^\intercal \operatorname{Hess}_x V(t,x;\theta)) = \begin{pmatrix}\mathrm{tr}(\overline{\sigma}_1(t,x)\overline{\sigma}_1(t,x)^\intercal \operatorname{Hess}_x V(t,x;\theta))\\ \vdots\\ \mathrm{tr}(\overline{\sigma}_k(t,x)\overline{\sigma}_k(t,x)^\intercal \operatorname{Hess}_x V(t,x;\theta))\end{pmatrix},$$

$$\overline{f}(t,x) = \begin{pmatrix}\overline{f}_1(t,x)\\ \vdots\\ \overline{f}_k(t,x)\end{pmatrix}.$$

The Lipschitz continuity of $\nabla\mathcal{R}^*_{\mathrm{en}}$ and the regularity assumptions imply that $\psi^\theta \in \mathcal{V}$ and the corresponding state process $X^{x_0,\theta,\psi^\theta}$ is well defined. Then a standard verification argument (see e.g., [169, Theorem 6.6, p. 278]) shows ψ^θ is an optimal feedback control and finishes the proof of Item (1).

To prove Item (2), Fix $x_0 \in \mathbb{R}^n$ and $R \geq 0$ and let C be a generic constant independent of θ. Note that the Fenchel-Young identity gives that $\mathcal{R}^*_{\mathrm{en}}(\nabla\mathcal{R}^*_{\mathrm{en}}(z)) = \langle z, \nabla\mathcal{R}^*_{\mathrm{en}}(z)\rangle - \mathcal{R}^*_{\mathrm{en}}(z)$ for all $z \in \mathbb{R}^k$, which along with (3.4.3) implies that for all $(t,x,\theta) \in [0,T] \times \mathbb{R}^n \times \mathbb{R}^{n\times(n+k)}$,

$$\begin{aligned}f(t,x,\psi^\theta(t,x)) = &-\langle\tfrac{1}{2}\mathrm{tr}(\overline{\sigma}(t,x)\overline{\sigma}(t,x)^\intercal \operatorname{Hess}_x V(t,x;\theta)) + B^\intercal\nabla_x V(t,x;\theta), \psi^\theta(t,x)\rangle\\ &- \mathcal{R}^*_{\mathrm{en}}(-\tfrac{1}{2}\mathrm{tr}(\overline{\sigma}(t,x)\overline{\sigma}(t,x)^\intercal \operatorname{Hess}_x V(t,x;\theta)) - B^\intercal\nabla_x V(t,x;\theta) - \overline{f}(t,x)).\end{aligned}$$

By the regularity assumptions of the coefficients and the function V, for all $t \in [0,T]$, $x, x' \in \mathbb{R}^n$ and $\theta, \theta' \in \mathbb{R}^{n\times(n+k)}$ with $|\theta|, |\theta'| \leq R$, there exists $C \geq 0$ such that

$$\begin{aligned}&|\psi^{\theta'}(t,x') - \psi^\theta(t,x)| + |f(t,x',\psi^{\theta'}(t,x')) - f(t,x,\psi^\theta(t,x))|\\ &\qquad\qquad \leq C\big(|x-x'| + (1+|x'|+|x|)|\theta-\theta'|\big).\end{aligned}$$

Proceeding along the lines of the proof of Theorem 3.2.7 leads to the desired estimate in Item (2). □

3.5 Numerical experiments

In this section, we test the theoretical findings and Algorithm 10 through numerical experiment on a three-dimensional LQ RL problem considered in [45, 46]. Our experiments show the convergence of the least-squares estimations to the true parameters as the number of episodes increases, as well as the sublinear cumulative regret as indicated in Theorem 3.3.4. Moreover, it confirms that the state coefficient $A^\star$ is easier to learn than the control coefficient $B^\star$, consistent with the observations in [46]. Our numerical result shows that a rough estimation of the control parameter $B^\star$ is often sufficient to design a nearly optimal feedback control, and that the Algorithm 10 is robust with respect to the initial batch size m_0.

Problem setup. We consider a three-dimensional LQ RL problems over the time horizon $[0, T]$ with $T = 1.5$, where the linear state dynamics (3.3.2) has the initial state x_0 and unknown coefficients $\theta^\star = (A^\star, B^\star) \in \mathbb{R}^{3\times(3+3)}$ chosen as in [45, 46]:

$$A^\star = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix}, \; B^\star = \mathbb{I}_3, \; \sigma = \mathbb{I}_3, \; \gamma \equiv 0, \; x_0 = 0,$$

with $\mathbb{I}_3$ being the 3×3 identity matrix, and the cost functional (3.3.1) involves quadratic functions $g \equiv 0$ and $f(t, x, a) = (x^\mathsf{T} Q x + a^\mathsf{T} R a)/2$, with $Q = 0.1\mathbb{I}_3$ and $R = \mathbb{I}_3$. As mentioned in [45, 46], this state dynamics corresponds to a marginally unstable graph Laplacian system where adjacent nodes are weakly connected, which arises naturally from consensus and distributed averaging problems. Since the cost penalizes the control inputs more than the states, it is essential to learn the unstable components of $A^\star$ and perform control on these components in order to achieve an optimal cost. Note that this LQ RL problem satisfies (H.4); see the last paragraph of Remark 3.3.1.

The numerical experiments are coded using Python. Algorithm 10 is initialized with $m_0 = 4$ and the initial guess $A_0 = \begin{bmatrix} 1.6243 & -0.6118 & -0.5282 \\ -1.0730 & 0.8654 & -2.3015 \\ 1.7448 & -0.7612 & 0.3190 \end{bmatrix}$ and $B_0 = \begin{bmatrix} -0.2494 & 1.4621 & -2.0601 \\ -0.3224 & -0.3841 & 1.1338 \\ -1.0999 & -0.1724 & -0.8779 \end{bmatrix}$, whose entries are sampled independently from the standard normal distribution. For each $\ell \in \mathbb{N} \cup \{0\}$, given the current estimate $\theta_\ell = (A_\ell, B_\ell)$ of $\theta^\star$, classical LQ control theory (see e.g., [169]) shows that solutions to (3.2.11) can be found analytically via Riccati equations, and the greedy policy ψ^{θ_ℓ} is given by $\psi^{\theta_\ell}(t, x) = -R^{-1} B^\mathsf{T} P_t^{\theta_\ell} x$, where P^{θ_ℓ} is the unique positive semidefinite solution to

$$\tfrac{\mathrm{d}}{\mathrm{d}t} P_t + A_\ell^\mathsf{T} P_t + P_t A_\ell - P_t (B_\ell R^{-1} B_\ell^\mathsf{T}) P_t + Q = 0, \; t \in (0, T); \quad P_T = 0. \tag{3.5.1}$$

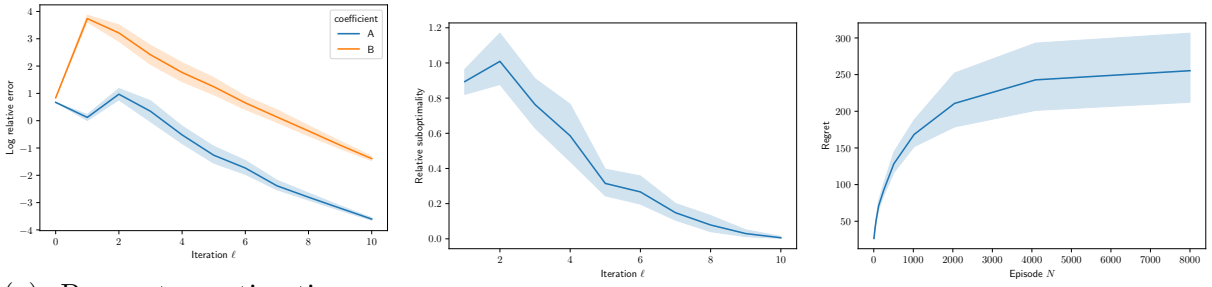
We solve (3.5.1) numerically via a high-order Runge-Kutta method on a uniform time grid with stepsize $T/100$, and then simulate $m_\ell = 2^\ell m_0$ independent trajectories of the state dynamics (3.3.4) (controlled by ψ^{θ_ℓ}) using the Euler-Maruyama method on the same time grid. To estimate statistical properties of the algorithm regret (3.3.3), we execute Algorithm 10 for 100 independent runs, where among different executions, the observed state trajectories are simulated based on independent Brownian motion increments.

Performance with $m_0 = 4$. Figure 3.1 exhibits the performance of Algorithm 10 for this LQ-RL problem, where the solid lines and the shallow areas indicate the sample mean and the 95% confidence interval over 100 repeated experiments. The numerical results indicate that algorithm 10 manages to learn the parameters over time while incurring a desirable sublinear regret, which is consistent with our theoretical result in Theorem 3.3.4. More precisely,

- Figure 3.1a presents the logarithmic relative error of the estimate (A_ℓ, B_ℓ) (in the Frobenius norm) after the ℓ-th update for $\ell \in \{0, \dots, 10\}$. One can observe that the estimate $(A_\ell, B_\ell)_\ell$ converge to the true parameter $(A^\star, B^\star)$ as the number of episodes increases. Our experiment shows that it is much easier to learn the state coefficient $A^\star$ than the control coefficient $B^\star$, which is consistent with the observation in [46] for other adaptive control schemes.

- Figure 3.1b presents the relative error between the expected cost $J^{\theta^\star}(\psi^{\theta_\ell}; x_0)$ and the optimal expected cost $J^{\theta^\star}(\psi^{\theta^\star}; x_0)$. One can see that a rough estimate of the control parameter $B^\star$ is often sufficient to design a nearly optimal feedback control. In particular, after the 10-th update ($\ell = 10$), although the relative approximation errors of A_ℓ and B_ℓ are 2.7% and 24.9%, respectively, the cost of ψ^{θ_ℓ} approximates the optimal cost accurately with a relative error 0.6%.

- Figure 3.1c presents the cumulative regret over episodes. One can see that the small performance gap results in a slowly growing algorithm regret. In fact, performing a linear regression for logarithms of expected regret and episode shows that the regret after the N-th episode is of the magnitude $\mathcal{O}(N^{0.34})$, which is slightly better than the theoretical upper bound in Theorem 3.3.4.

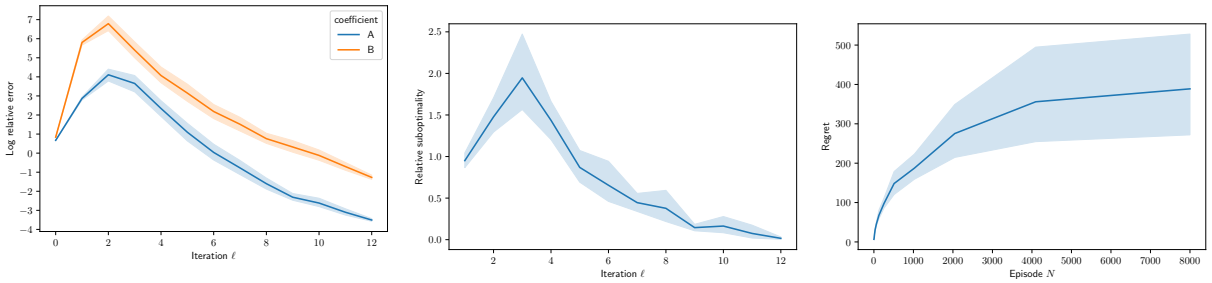


(a) Parameter estimation errors. (b) Cost suboptimality gap. (c) Algorithm regret.

Figure 3.1: Performance of Algorithm 10 for the LQ-RL problem ($m_0 = 4$).

Robustness with respect to the initial batch size m_0. We next demonstrate the robustness of Algorithm 10 by performing computations with $m_0 = 1$ and fixing other settings as above. The results are shown in Figure 3.2. Note that the smaller initial batch size m_0 makes the learning more challenging. By comparing the results against those with $m_0 = 4$, one can see that our algorithm is robust and performs well with the small m_0. In particular, we see that

- Estimating parameters with fewer sample trajectories leads to larger parameter estimation errors with suboptimality gaps, especially for the first few iterations. It also leads to a wider range of $(A_\ell, B_\ell)_\ell$ among different algorithm executions and hence a larger variance of the algorithm regret.
- As the number of episodes increases, the estimate $(A_\ell, B_\ell)_\ell$ converge to the true parameter $(A^\star, B^\star)$ and the suboptimality gap quickly converges to 0, see Figures 3.2a and 3.2b. The algorithm regret grows sublinearly (see Figure 3.2c), and the regret after the N-th episode is of the magnitude $\mathcal{O}(N^{0.51})$. This confirms the theoretical results in Theorem 3.3.4 even for a small m_0.



(a) Parameter estimation errors. (b) Cost suboptimality gap. (c) Algorithm regret.

Figure 3.2: Performance of Algorithm 10 for the LQ-RL problem ($m_0 = 1$).

3.6 Appendix

3.6.1 Preliminaries

Here, we collect some fundamental results which are used for our analysis.

We start with a stability result for coupled FBSDEs under a generalized monotonicity condition, which is crucial for our stability analysis of feedback controls. For any given $t \in [0,T]$ and $\lambda \in [0,1]$, we consider the following FBSDE defined on $[t,T]$: for $s \in [t,T]$,

$$\mathrm{d}X_s = (\lambda \bar{b}(s, X_s, Y_s) + \mathcal{I}_s^b)\,\mathrm{d}s + \bar{\sigma}(s)\,\mathrm{d}W_s + \int_{\mathbb{R}_0^p} \bar{\gamma}(s,u)\,\tilde{N}(\mathrm{d}s,\mathrm{d}u), \quad X_t = \xi, \tag{3.6.1a}$$

$$\mathrm{d}Y_s = -(\lambda \bar{f}(s, X_s, Y_s) + \mathcal{I}_s^f)\,\mathrm{d}t + Z_s\,\mathrm{d}W_s + \int_{\mathbb{R}_0^p} M_s\,\tilde{N}(\mathrm{d}s,\mathrm{d}u), \qquad Y_T = \lambda \bar{g}(X_T) + \mathcal{I}^g, \tag{3.6.1b}$$

with given $\xi \in L^2(\mathcal{F}_t;\mathbb{R}^n)$, $(\mathcal{I}^b, \mathcal{I}^f) \in \mathcal{H}^2(\mathbb{R}^n \times \mathbb{R}^n)$, $\mathcal{I}^g \in L^2(\mathcal{F}_T;\mathbb{R}^m)$ and measurable functions $\bar{\sigma} : [0,T] \to \mathbb{R}^{n\times d}$, $\bar{\gamma} : [0,T] \times \mathbb{R}_0^p \to \mathbb{R}^n$, $\bar{b}, \bar{f} : [0,T] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ and $\bar{g} : \mathbb{R}^n \to \mathbb{R}^n$.

Lemma 3.6.1. *Let $K \geq 0$, for each $i \in \{1,2\}$, let $\bar{b}_i, \bar{f}_i : [0,T] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$, $\bar{g}_i : \mathbb{R}^n \to \mathbb{R}^n$ satisfy $\int_0^T (|\bar{b}_i(t,0,0)|^2 + |\bar{f}_i(t,0,0)|^2)\,\mathrm{d}t < \infty$ and for all $t \in [0,T]$ that $\bar{f}_i(t,\cdot), \bar{g}_i$ are K-Lipschitz continuous, let $\bar{\sigma}_i : [0,T] \to \mathbb{R}^{n\times d}$ satisfy $\int_0^T |\bar{\sigma}_i(t)|^2\,\mathrm{d}t < \infty$ and let $\bar{\gamma}_i : [0,T] \times \mathbb{R}_0^p \to \mathbb{R}^n$ satisfy $\int_0^T \int_{\mathbb{R}_0^p} |\bar{\gamma}_i(t,u)|^2\,\nu(\mathrm{d}u)\mathrm{d}t < \infty$. Assume further that there exists $\tau > 0$ and a measurable function $\eta : [0,T] \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to [0,\infty)$ such that for all $t \in [0,T], (x,y), (x',y') \in \mathbb{R}^n \times \mathbb{R}^n$,*

$$\langle \bar{b}_1(t,x,y) - \bar{b}_1(t,x',y'), y - y' \rangle + \langle -\bar{f}_1(t,x,y) + \bar{f}_1(t,x',y'), x - x' \rangle \leq -\tau \eta(t,x,y,x',y'), \tag{3.6.2}$$

$$|\bar{b}_1(t,x,y) - \bar{b}_1(t,x',y')| \leq K(|x-x'| + \eta(t,x,y,x',y')), \tag{3.6.3}$$

$$\langle \bar{g}(x) - \bar{g}(x'), x - x' \rangle \geq 0. \tag{3.6.4}$$

Then there exists $C > 0$, depending only on T, K, λ and the dimensions, such that for all $t \in [0,T]$, $\lambda_0 \in [0,1]$, $i \in \{1,2\}$, for every $(X_i, Y_i, Z_i, M_i) \in \mathcal{S}^2(t,T;\mathbb{R}^n) \times \mathcal{S}^2(t,T;\mathbb{R}^n) \times \mathcal{H}^2(t,T;\mathbb{R}^{n\times d}) \times \mathcal{H}_\nu^2(t,T;\mathbb{R}^n)$ satisfying (3.6.1) with $\lambda = \lambda_0$, $(\bar{b}, \bar{\sigma}, \bar{\gamma}, \bar{f}, \bar{g}) = (\bar{b}_i, \bar{\sigma}_i, \bar{\gamma}_i, \bar{f}_i, \bar{g}_i)$, $\xi = \xi_i \in L^2(\mathcal{F}_t;\mathbb{R}^n)$, $\mathcal{I}^g = \mathcal{I}_i^g \in L^2(\mathcal{F}_T;\mathbb{R}^n)$ and $(\mathcal{I}^b, \mathcal{I}^f) = (\mathcal{I}_i^b, \mathcal{I}_i^f) \in \mathcal{H}^2(\mathbb{R}^n \times \mathbb{R}^n)$, we have that

$$\begin{aligned}
&\|X_1 - X_2\|_{\mathcal{S}^2}^2 + \|Y_1 - Y_2\|_{\mathcal{S}^2}^2 + \|Z_1 - Z_2\|_{\mathcal{H}^2}^2 + \|M_1 - M_2\|_{\mathcal{H}_\nu^2}^2 \\
&\leq C\Big\{ \|\xi_1 - \xi_2\|_{L^2}^2 + \|\lambda_0(\bar{g}_1(X_{2,T}) - \bar{g}_2(X_{2,T})) + \mathcal{I}_1^g - \mathcal{I}_2^g\|_{L^2}^2 + \|\bar{\sigma}_1 - \bar{\sigma}_2\|_{\mathcal{H}^2}^2 + \|\bar{\gamma}_1 - \bar{\gamma}_2\|_{\mathcal{H}_\nu^2}^2 \\
&\quad + \|\lambda_0(\bar{b}_1(\cdot, X_2, Y_2) - \bar{b}_2(\cdot, X_2, Y_2)) + \mathcal{I}_1^b - \mathcal{I}_2^b\|_{\mathcal{H}^2}^2 \\
&\quad + \|\lambda_0(\bar{f}_1(\cdot, X_2, Y_2) - \bar{f}_2(\cdot, X_2, Y_2)) + \mathcal{I}_1^f - \mathcal{I}_2^f\|_{\mathcal{H}^2}^2 \Big\}.
\end{aligned} \tag{3.6.5}$$

Proof. Throughout this proof, let C be a generic constant depending only on T, K, λ and the dimensions, let $t \in [0,T]$, $\lambda_0 \in [0,1]$, let $(\delta X, \delta Y, \delta Z, \delta M) = (X_1 - X_2, Y_1 - Y_2, Z_1 - Z_2, M_1 - M_2)$, $\delta\xi = \xi_1 - \xi_2$, $\delta\sigma = \bar{\sigma}_1 - \bar{\sigma}_2$, $\delta\gamma = \bar{\gamma}_1 - \bar{\gamma}_2$, $\delta\mathcal{I}^g = \mathcal{I}^g_1 - \mathcal{I}^g_2$, and for each $s \in [t,T]$ let $\delta\mathcal{I}^b_s = \mathcal{I}^b_{1,s} - \mathcal{I}^b_{2,s}$, $\delta\mathcal{I}^f_s = \mathcal{I}^f_{1,s} - \mathcal{I}^f_{2,s}$, $\bar{b}_1(\Theta_{1,s}) = \bar{b}_1(t, X_{1,s}, Y_{1,s})$, $\bar{b}_1(\Theta_{2,s}) = \bar{b}_1(t, X_{2,s}, Y_{2,s})$ and $\bar{b}_2(\Theta_{2,s}) = \bar{b}_2(t, X_{2,s}, Y_{2,s})$. Similarly, we introduce the notation $\bar{f}_1(\Theta_{1,s})$, $\bar{f}_1(\Theta_{2,s})$, $\bar{f}_2(\Theta_{2,s})$ for $s \in [t,T]$.

By applying Itô's formula to $(\langle Y_{1,s} - Y_{2,s}, X_{1,s} - X_{2,s}\rangle)_{s\in[t,T]}$, we can obtain from (3.6.1) that

$$
\begin{aligned}
&\mathbb{E}[\langle \lambda_0(\bar{g}_1(X_{1,T}) - \bar{g}_2(X_{2,T})) + \delta\mathcal{I}^g, \delta X_T\rangle - \langle \delta Y_t, \delta\xi\rangle] \\
&= \mathbb{E}\bigg[\int_t^T \Big(\langle \lambda_0(\bar{b}_1(\Theta_{1,s}) - \bar{b}_2(\Theta_{2,s})) + \delta\mathcal{I}^b_s, \delta Y_s\rangle - \langle \lambda_0(\bar{f}_1(\Theta_{1,s}) - \bar{f}_2(\Theta_{2,s})) + \delta\mathcal{I}^f_s, \delta X_s\rangle \\
&\quad + \langle \delta\sigma(s), \delta Z_s\rangle + \int_{\mathbb{R}^p_0} \langle \delta\gamma(s,u), \delta M_s\rangle\, \nu(\mathrm{d}u)\Big)\,\mathrm{d}s\bigg] \\
&\le \mathbb{E}\bigg[\int_t^T \Big(- \lambda_0\tau\eta(s, X_{1,s}, Y_{1,s}, X_{2,s}, Y_{2,s}) + \langle \lambda_0(\bar{b}_1(\Theta_{2,s}) - \bar{b}_2(\Theta_{2,s})) + \delta\mathcal{I}^b_s, \delta Y_s\rangle \\
&\quad - \langle \lambda_0(\bar{f}_1(\Theta_{2,s}) - \bar{f}_2(\Theta_{2,s})) + \delta\mathcal{I}^f_s, \delta X_s\rangle + \langle \delta\sigma(s), \delta Z_s\rangle + \int_{\mathbb{R}^p_0} \langle \delta\gamma(s,u), \delta M_s\rangle\, \nu(\mathrm{d}u)\Big)\,\mathrm{d}s\bigg],
\end{aligned}
$$

where for the last inequality, we have added and subtracted the terms $\lambda_0\bar{b}_1(\Theta_{2,s})$ and $-\lambda_0\bar{f}_1(\Theta_{2,s})$, and applied (3.6.2). Then, we can further deduce from (3.6.4) that

$$
\begin{aligned}
&\lambda_0\tau\mathbb{E}\bigg[\int_t^T \eta(s, X_{1,s}, Y_{1,s}, X_{2,s}, Y_{2,s})\,\mathrm{d}s\bigg] \\
&\le -\mathbb{E}[\langle \lambda_0(\bar{g}_1(X_{2,T}) - \bar{g}_2(X_{2,T})) + \delta\mathcal{I}^g, \delta X_T\rangle - \langle \delta Y_t, \delta\xi\rangle] \\
&\quad + \mathbb{E}\bigg[\int_t^T \Big(\langle \lambda_0(\bar{b}_1(\Theta_{2,s}) - \bar{b}_2(\Theta_{2,s})) + \delta\mathcal{I}^b_s, \delta Y_s\rangle - \langle \lambda_0(\bar{f}_1(\Theta_{2,s}) - \bar{f}_2(\Theta_{2,s})) + \delta\mathcal{I}^f_s, \delta X_s\rangle \\
&\quad + \langle \delta\sigma(s), \delta Z_s\rangle + \int_{\mathbb{R}^p_0} \langle \delta\gamma(s,u), \delta M_s\rangle\, \nu(\mathrm{d}u)\Big)\,\mathrm{d}s\bigg],
\end{aligned}
$$

from which we can apply Young's inequality and obtain for all $\varepsilon > 0$ that

$$
\begin{aligned}
&\lambda_0\mathbb{E}\bigg[\int_t^T \eta(s, X_{1,s}, Y_{1,s}, X_{2,s}, Y_{2,s})\,\mathrm{d}s\bigg] \\
&\le \varepsilon\Big(\|\delta X_T\|^2_{L^2} + \|\delta Y_t\|^2_{L^2} + \|\delta X\|^2_{\mathcal{H}^2} + \|\delta Y\|^2_{\mathcal{H}^2} + \|\delta Z\|^2_{\mathcal{H}^2} + \|\delta M\|^2_{\mathcal{H}^2_\nu}\Big) + C\mathrm{RHS}/\varepsilon,
\end{aligned}
$$

where RHS denotes the terms at the right-hand side of (3.6.5).

By (3.6.3) and a standard stability estimate of (3.6.1a), we can deduce that

$$
\begin{aligned}
\|\delta X\|^2_{\mathcal{S}^2} &\le C\bigg(\lambda_0\mathbb{E}\bigg[\int_t^T \eta(s, X_{1,s}, Y_{1,s}, X_{2,s}, Y_{2,s})\,\mathrm{d}s\bigg] + \mathrm{RHS}\bigg) \\
&\le \varepsilon C\Big(\|\delta Y_t\|^2_{L^2} + \|\delta Y\|^2_{\mathcal{H}^2} + \|\delta Z\|^2_{\mathcal{H}^2} + \|\delta M\|^2_{\mathcal{H}^2_\nu}\Big) + C\mathrm{RHS}/\varepsilon
\end{aligned}
\tag{3.6.6}
$$

for all small enough $\varepsilon > 0$. Moreover, by the Lipschitz continuity of $\bar{f}_1$, $\bar{g}_1$ and the stability estimate of (3.6.1b) (see e.g. [133, Proposition A4]), we have that

$$\begin{aligned}
&\|\delta Y\|^2_{\mathcal{S}^2} + \|\delta Z\|^2_{\mathcal{H}^2} + \|\delta M\|^2_{\mathcal{H}^2_\nu} \\
&\leq C\Big(\|\lambda_0(\bar{g}_1(X_{1,T}) - \bar{g}_2(X_{2,T})) + \delta\mathcal{I}^g\|^2_{L^2} + \|\lambda_0(\bar{f}_1(\cdot, X_1, Y_2) - \bar{f}_2(\cdot, X_2, Y_2)) + \delta\mathcal{I}^f\|^2_{\mathcal{H}^2} \Big) \\
&\leq C\Big(\|\delta X\|^2_{\mathcal{S}^2} + \|\lambda_0(\bar{g}_1(X_{2,T}) - \bar{g}_2(X_{2,T})) + \delta\mathcal{I}^g\|^2_{L^2} + \|\lambda_0(\bar{f}_1(\cdot, X_2, Y_2) - \bar{f}_2(\cdot, X_2, Y_2)) + \delta\mathcal{I}^f\|^2_{\mathcal{H}^2} \Big) \\
&\leq C\mathrm{RHS},
\end{aligned}$$

where we have applied (3.6.6) with a sufficiently small ε for the last inequality. This completes the desired stability estimate. □

We then present a version of Burkholder's inequality for the $\|\cdot\|_{L^q}$-norm of stochastic integrals, which not only extends [27, Corollary 2.2] to stochastic integrals with respect to general Poisson random measures on $[0,T] \times \mathbb{R}^p_0$, but also improves the bounding constants there with a sharper dependence on the index q.

Lemma 3.6.2. *For all $v \in \mathcal{H}^2(0,T;\mathbb{R}^d)$, $w \in \mathcal{H}^2_\nu(0,T;\mathbb{R})$ and $q \geq 2$, we have*

$$\mathbb{E}\bigg[\bigg|\int_0^T v_t^\intercal \,\mathrm{d}W_t\bigg|^q\bigg] \leq C_q \mathbb{E}\bigg[\bigg(\int_0^T |v_t|^2 \,\mathrm{d}t\bigg)^{q/2}\bigg], \tag{3.6.7}$$

$$\begin{aligned}
\mathbb{E}\bigg[\bigg|\int_0^T \int_{\mathbb{R}^p_0} w(t,u)\,\tilde{N}(\mathrm{d}t,\mathrm{d}u)\bigg|^q\bigg] \leq \tilde{C}_q\bigg(&\mathbb{E}\bigg[\int_0^T \int_{\mathbb{R}^p_0} |w(t,u)|^q\,\nu(\mathrm{d}u)\mathrm{d}t\bigg] \\
&+ \mathbb{E}\bigg[\bigg(\int_0^T \int_{\mathbb{R}^p_0} |w(t,u)|^2\,\nu(\mathrm{d}u)\mathrm{d}t\bigg)^{q/2}\bigg]\bigg),
\end{aligned} \tag{3.6.8}$$

where $C_q = (\sqrt{e/2}q)^q$ and $\tilde{C}_q = 21e^q q^{2q}$.

Proof. Recall that Burkholder's inequality in [23, Theorem 4.2.12] shows for all $1 \leq q < \infty$ and for every local martingale $(M_t)_{t\in[0,T]}$ that $\mathbb{E}[|M_T^*|^q] \leq C_q \mathbb{E}[[M,M]_T^{q/2}]$, where $M_T^* = \sup_{t\in[0,T]} |M_t|$, $[M,M]$ is the quadratic variation of M, and $C_q = (\sqrt{10q})^q$ for $q \in [1,2)$ and $C_q = (\sqrt{e/2}q)^q$ for $q \in [2,\infty)$. Hence we can obtain (3.6.7) by setting $M_t = \int_0^t v_s^\intercal \,\mathrm{d}W_s$ for all $t \in [0,T]$, whose quadratic variation process is given by $[M,M]_T = \int_0^T |v|^2 \,\mathrm{d}t$.

We proceed to establish (3.6.8) by following the arguments of [27, Lemma 2.1]. For all $r \geq 1$ and $t \in [0,T]$, let $K_t^{(r)} = \int_0^t \int_{\mathbb{R}^p_0} w(s,u)^r \,\tilde{N}(\mathrm{d}s,\mathrm{d}u)$. For any given $r \geq 1$ and $q \geq 2$, we

can obtain from Burkholder's inequality and $\tilde{N}(\mathrm{d}t, \mathrm{d}u) = N(\mathrm{d}t, \mathrm{d}u) - \nu(\mathrm{d}u)\mathrm{d}t$ that

$$\begin{aligned}
\mathbb{E}[|(K^{(r)})_T^*|^q] &\leq C_q \mathbb{E}\left[\left(\int_0^T \int_{\mathbb{R}_0^p} |w(t,u)|^{2r} \, N(\mathrm{d}t, \mathrm{d}u)\right)^{q/2}\right] \\
&= C_q \mathbb{E}\left[\left(\int_0^T \int_{\mathbb{R}_0^p} |w(t,u)|^{2r} \, \tilde{N}(\mathrm{d}t, \mathrm{d}u) + \int_0^T \int_{\mathbb{R}_0^p} |w(t,u)|^{2r} \, \nu(\mathrm{d}u)\mathrm{d}t\right)^{q/2}\right] \\
&\leq 2^{\frac{q}{2}-1} C_q \mathbb{E}[|(K^{(2r)})_T^*|^{q/2}] + 2^{\frac{q}{2}-1} C_q \mathbb{E}\left[\left|\int_0^T \int_{\mathbb{R}_0^p} |w(t,u)|^{2r} \, \nu(\mathrm{d}u)\mathrm{d}t\right|^{q/2}\right].
\end{aligned}$$

Hence, recursively applying the above estimate yields for all $q \geq 2$ and $n \in \mathbb{N}$ with $q/2^{n-1} \geq 2$ that

$$\begin{aligned}
\mathbb{E}\left[\left|\int_0^T \int_{\mathbb{R}_0^p} w(t,u) \, \tilde{N}(\mathrm{d}t, \mathrm{d}u)\right|^q\right] &\leq \left(\prod_{j=1}^n 2^{\frac{q}{2^j}-1} C_{\frac{q}{2^{j-1}}}\right) \mathbb{E}[|(K^{(2^n)})_T^*|^{q/2^n}] \\
&\quad + \sum_{k=1}^n \left(\prod_{j=1}^k 2^{\frac{q}{2^j}-1} C_{\frac{q}{2^{j-1}}}\right) \mathbb{E}\left[\left|\int_0^T \int_{\mathbb{R}_0^p} |w(t,u)|^{2^k} \, \nu(\mathrm{d}u)\mathrm{d}t\right|^{q/2^k}\right].
\end{aligned} \tag{3.6.9}$$

Now let $q \geq 2$ be fixed and set $n = \lfloor \log_2 q \rfloor$ such that $q \in [2^n, 2^{n+1})$. Since $q/2^n \in [1, 2)$, the constant $C_{q/2^n}$ in Burkholder's inequality satisfies $C_{q/2^n} \leq 20$, from which we can show that (see [27, Lemma 2.1]):

$$\mathbb{E}[|(K^{(2^n)})_T^*|^{q/2^n}] \leq 20 \mathbb{E}\left[\int_0^T \int_{\mathbb{R}_0^p} |w(t,u)|^q \, \nu(\mathrm{d}u)\mathrm{d}t\right].$$

Moreover, by proceeding along the lines of [27, Corollary 2.2], we obtain for all $k = 1, \ldots, n$ that

$$\begin{aligned}
&\mathbb{E}\left[\left|\int_0^T \int_{\mathbb{R}_0^p} |w(t,u)|^{2^k} \, \nu(\mathrm{d}u)\mathrm{d}t\right|^{q/2^k}\right] \\
&\leq \mathbb{E}\left[\int_0^T \int_{\mathbb{R}_0^p} |w(t,u)|^q \, \nu(\mathrm{d}u)\mathrm{d}t\right] + \mathbb{E}\left[\left|\int_0^T \int_{\mathbb{R}_0^p} |w(t,u)|^2 \, \nu(\mathrm{d}u)\mathrm{d}t\right|^{q/2}\right].
\end{aligned}$$

Hence, we can deduce from (3.6.9) that

$$\begin{aligned}
\mathbb{E}\left[\left|\int_0^T \int_{\mathbb{R}_0^p} w(t,u) \, \tilde{N}(\mathrm{d}t, \mathrm{d}u)\right|^q\right] &\leq 21 \sum_{k=1}^{\lceil \log_2 q \rceil} \left(\prod_{j=1}^k 2^{\frac{q}{2^j}-1} C_{\frac{q}{2^{j-1}}}\right) \left(\mathbb{E}\left[\int_0^T \int_{\mathbb{R}_0^p} |w(t,u)|^q \, \nu(\mathrm{d}u)\mathrm{d}t\right]\right. \\
&\quad \left. + \mathbb{E}\left[\left|\int_0^T \int_{\mathbb{R}_0^p} |w(t,u)|^2 \, \nu(\mathrm{d}u)\mathrm{d}t\right|^{q/2}\right]\right).
\end{aligned}$$

We now obtain an upper bound of the constant $21\sum_{k=1}^{\lfloor \log_2 q \rfloor}\left(\prod_{j=1}^{k} 2^{\frac{q}{2^j}-1} C_{\frac{q}{2^{j-1}}}\right)$ as follows:

$$
\begin{aligned}
&21\sum_{k=1}^{\lfloor \log_2 q \rfloor}\left(\prod_{j=1}^{k} 2^{\frac{q}{2^j}-1} C_{\frac{q}{2^{j-1}}}\right) = 21\sum_{k=1}^{\lfloor \log_2 q \rfloor}\prod_{j=1}^{k} 2^{\frac{q}{2^j}-1}\left(\sqrt{\frac{e}{2}}\frac{q}{2^{j-1}}\right)^{\frac{q}{2^{j-1}}} \\
&\leq 21\left(\sum_{k=1}^{\lfloor \log_2 q \rfloor} 2^{-k}\right) e^{\sum_{j=1}^{\lfloor \log_2 q \rfloor}\frac{q}{2^j}} \prod_{j=1}^{\lfloor \log_2 q \rfloor}\left(\frac{q}{2^{j-1}}\right)^{\frac{q}{2^{j-1}}} \leq 21e^q 2^{\sum_{j=1}^{\lfloor \log_2 q \rfloor}\frac{q}{2^{j-1}}\log_2(\frac{q}{2^{j-1}})} \\
&\leq 21e^q 2^{\sum_{j=1}^{\lfloor \log_2 q \rfloor}\frac{q}{2^{j-1}}\log_2 q} \leq 21e^q 2^{2q\log_2 q} = 21e^q q^{2q} \coloneqq \tilde{C}_q,
\end{aligned}
$$

which leads us to the desired conclusion. □

The following lemma estimates the tail behaviors of solutions to SDEs with jumps. The result has been established in Lemma 2.1 and Theorem 2.8 of [111] for SDEs with time homogenous coefficients and bounded Lipschitz continuous functions $\mathfrak{f}$ via Malliavin Calculus, which can be extended to SDEs with time inhomogeneous coefficients and unbounded $\mathfrak{f}$ (via Fatou's lemma) in a straightforward manner.

Lemma 3.6.3. *Let $T \geq 0$ and $b : [0,T]\times\mathbb{R}^n \to \mathbb{R}^n$, $\sigma : [0,T] \to \mathbb{R}^{n\times d}$, $\gamma : [0,T]\times\mathbb{R}_0^p \to \mathbb{R}^n$ be measurable functions such that there exist $K, \sigma_{\max} \geq 0$ and a measurable function $\bar{\gamma} : \mathbb{R}_0^p \to \mathbb{R}$ satisfying for all $(t,u) \in [0,T]\times\mathbb{R}_0^p$, $x, x' \in \mathbb{R}^n$ that $|b(t,0)| \leq K$, $|b(t,x) - b(t,x')| \leq K|x-x'|$, $|\sigma(t)| \leq \sigma_{\max}$ and $|\gamma(t,u)| \leq \bar{\gamma}(u)$, ν-a.e.. Let $\beta : [0,\infty) \to [0,\infty]$ be defined by $\beta(\lambda) \coloneqq \int_{\mathbb{R}_0^p}\left(e^{\lambda\bar{\gamma}(u)} - \lambda\bar{\gamma}(u) - 1\right)\nu(\mathrm{d}u)$ for any $\lambda \geq 0$. Assume that $\beta(\lambda) < \infty$ for some $\lambda > 0$.*

Then there exists a constant $C > 0$, depending only on K and T, such that for all $x \in \mathbb{R}^n$, the unique solution $X^x \in \mathcal{S}^2(\mathbb{R}^n)$ to the following SDE

$$
\mathrm{d}X_t = b(t,X_t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}W_t + \int_{\mathbb{R}_0^p}\gamma(t,u)\,\tilde{N}(\mathrm{d}t,\mathrm{d}u), \quad t \in [0,T], \quad X_0 = x
$$

satisfies for every Lipschitz continuous function $\mathfrak{f} : (\mathbb{D}([0,T];\mathbb{R}^n), d_\infty) \to \mathbb{R}$ that

$$
\mathbb{E}\left[e^{\lambda(\mathfrak{f}(X^x) - \mathbb{E}[\mathfrak{f}(X^x)])}\right] \leq e^{C\eta\left(C\lambda\|\mathfrak{f}\|_{\mathrm{Lip}}\right)} \quad \forall\lambda > 0, \tag{3.6.10}
$$

where $\mathbb{D}([0,T];\mathbb{R}^n)$ is the space of $\mathbb{R}^n$-valued càdlàg functions on $[0,T]$, d_∞ is the uniform metric defined by $d_\infty(\rho_1,\rho_2) \coloneqq \sup_{t\in[0,T]}|\rho_1(t) - \rho_2(t)|$ for any $\rho_1, \rho_2 \in \mathbb{D}([0,T];\mathbb{R}^n)$, $\|\mathfrak{f}\|_{\mathrm{Lip}}$ is the constant defined by $\|\mathfrak{f}\|_{\mathrm{Lip}} \coloneqq \sup_{\rho_1\neq\rho_2}\frac{|\mathfrak{f}(\rho_1) - \mathfrak{f}(\rho_2)|}{d_\infty(\rho_1,\rho_2)}$, and $\eta : [0,\infty) \to [0,\infty]$ is the function defined by $\eta(\lambda) \coloneqq \beta(\lambda) + \sigma_{\max}^2\lambda^2/2$ for any $\lambda \geq 0$.

The next lemma presents a concentration inequality for the sum of independent sub-Weibull random variables, which follows directly from Theorem 3.1 and Proposition A3 in [98].

Lemma 3.6.4. *Let $\alpha \in (0,1]$, $N \in \mathbb{N}$ and $X_1, \dots, X_N \in \mathrm{subW}(\alpha)$ be independent random variables satisfying $\mathbb{E}[X_i] = 0$ for all $i = 1, \dots, N$. Then there exists a constant $C \geq 0$, depending only on α, such that*

$$\mathbb{P}\left(\left|\sum_{i=1}^N X_i\right| \geq \varepsilon'\right) \leq 2\exp\left(-C\min\left\{\frac{(\varepsilon')^2}{\sum_{i=1}^N \|X_i\|_{\Psi_\alpha}^2}, \left(\frac{\varepsilon'}{\max_i \|X_i\|_{\Psi_\alpha}}\right)^\alpha\right\}\right), \quad \forall \varepsilon' \geq 0.$$

3.6.2 Proofs of Lemmas 3.2.8, 3.3.5, 3.3.6, 3.3.7, 3.3.8

Proof of Lemma 3.2.8. We start by establishing the regularity of $\mathbb{R}^n \times \mathbb{R}^k \ni (x,z) \mapsto f(t,x,\partial_z f^*(t,x,z)) \in \mathbb{R} \cup \{\infty\}$ for a given $t \in [0,T]$. Observe that for all $(t,x) \in [0,T] \times \mathbb{R}^n$, $f(t,x,\cdot)$ is proper, convex, and lower semicontinuous, which along with the Fenchel-Young identity implies

$$f(t,x,\partial_z f^*(t,x,z)) = \langle z, \partial_z f^*(t,x,z)\rangle - f^*(t,x,z) \in \mathbb{R}, \quad \forall (t,x,z) \in [0,T] \times \mathbb{R}^n \times \mathbb{R}^k. \tag{3.6.11}$$

Given $t \in [0,T]$ and $(x_1,z_1),(x_2,z_2) \in \mathbb{R}^n \times \mathbb{R}^k$, by (3.6.11),

$$\begin{aligned}
&|f(t,x_1,\partial_z f^*(t,x_1,z_1)) - f(t,x_2,\partial_z f^*(t,x_2,z_2))| \\
&\leq |\langle z_1, \partial_z f^*(t,x_1,z_1)\rangle - \langle z_2, \partial_z f^*(t,x_2,z_2)\rangle| + |f^*(t,x_1,z_1) - f^*(t,x_2,z_1)| \\
&\quad + |f^*(t,x_2,z_1) - f^*(t,x_2,z_2)|.
\end{aligned}$$

We now estimate all the terms on the right hand side of the above inequality. By the Lipchitz continuity and local boundedness of $\partial_z f^*(t,\cdot)$ (see the proof of Lemma 3.2.3), we can obtain the following upper bound for the first and third terms:

$$\begin{aligned}
&|\langle z_1, \partial_z f^*(t,x_1,z_1)\rangle - \langle z_2, \partial_z f^*(t,x_2,z_2)\rangle| + |f^*(t,x_2,z_1) - f^*(t,x_2,z_2)| \\
&\leq |\langle z_1 - z_2, \partial_z f^*(t,x_1,z_1)\rangle| + |\langle z_2, \partial_z f^*(t,x_1,z_1) - \partial_z f^*(t,x_2,z_2)\rangle| \\
&\quad + |f^*(t,x_2,z_1) - f^*(t,x_2,z_2)| \\
&\leq C(1 + |x_1| + |x_2| + |z_1| + |z_2|)(|x_1 - x_2| + |z_1 - z_2|),
\end{aligned} \tag{3.6.12}$$

where the last inequality is by the mean value theorem. Moreover, by applying (3.6.11) to $f^*(t,x_1,z_1)$ and by the definition of $f^*(t,x_2,z_1)$ in (3.2.4), (H.2(3)), the linear growth of $\partial_z f^*(t,\cdot)$,

$$\begin{aligned}
f^*(t,x_1,z_1) - f^*(t,x_2,z_1) &\leq \langle z_1, \partial_z f^*(t,x_1,z_1)\rangle - f(t,x_1,\partial_z f^*(t,x_1,z_1)) \\
&\quad - (\langle z_1, \partial_z f^*(t,x_1,z_1)\rangle - f(t,x_2,\partial_z f^*(t,x_1,z_1))) \\
&= -f_0(t,x_1,\partial_z f^*(t,x_1,z_1)) + f_0(t,x_2,\partial_z f^*(t,x_1,z_1)) \\
&\leq C(1 + |x_1| + |x_2| + |z_1|)|x_1 - x_2|.
\end{aligned} \tag{3.6.13}$$

Then, by interchanging the roles of x_1, x_2 in (3.6.13) and taking account of (3.6.12), we can obtain the following estimate for all $t \in [0,T]$, $(x_1,z_1),(x_2,z_2) \in \mathbb{R}^n \times \mathbb{R}^k$:

$$\begin{aligned}
&|f(t,x_1,\partial_z f^*(t,x_1,z_1)) - f(t,x_2,\partial_z f^*(t,x_2,z_2))| \\
&\leq C(1 + |x_1| + |x_2| + |z_1| + |z_2|)(|x_1 - x_2| + |z_1 - z_2|).
\end{aligned}$$

Therefore, by (3.2.10), (3.2.17) and (3.2.19), for all $t \in [0,T]$ and $x, x' \in \mathbb{R}^n$,

$$\begin{aligned}
&|f(t,x,\psi(t,x)) - f(t,x',\tilde{\psi}(t,x'))| = |f(t,x,\phi(t,x,Y_t^{t.x})) - f(t,x',\tilde{\phi}(t,x',\tilde{Y}_t^{t.x'}))| \\
&= |f(t,x,\partial_z f^*(t,x,-b_2(t)^\mathsf{T} Y_t^{t.x})) - f(t,x',\partial_z f^*(t,x',-\tilde{b}_2(t)^\mathsf{T}\tilde{Y}_t^{t.x'}))| \\
&\le C(1+|x|+|x'|+|b_2(t)^\mathsf{T} Y_t^{t.x}| + |\tilde{b}_2(t)^\mathsf{T}\tilde{Y}_t^{t.x'}|)(|x-x'| + |b_2(t)^\mathsf{T} Y_t^{t.x} - \tilde{b}_2(t)^\mathsf{T}\tilde{Y}_t^{t.x'}|) \\
&\le C(1+|x|+|x'|+\|Y^{t.x}\|_{\mathcal{S}^2} + \|\tilde{Y}^{t.x'}\|_{\mathcal{S}^2}) \\
&\quad \times (|x-x'| + \|b_2 - b_2\|_{L^\infty}\|Y^{t.x}\|_{\mathcal{S}^2} + \|Y^{t.x} - \tilde{Y}^{t.x'}\|_{\mathcal{S}^2}) \\
&\le C(1+|x|+|x'|)(|x-x'| + \mathcal{E}_{\mathrm{per}}(1+|x|)),
\end{aligned}$$

which along with Young's inequality leads to the desired conclusion. □

Proof of Lemma 3.3.5. It suffices to show the statement for processes X, Y such that

$$\|X\|_{L^2(0,T)}, \|Y\|_{L^2(0,T)} \in \mathrm{subW}(\alpha)$$

with $\|X\|_{L^2(0,T)} \coloneqq (\int_0^T |X|^2 \,\mathrm{d}t)^{\frac{1}{2}}$ and $\|Y\|_{L^2(0,T)} \coloneqq (\int_0^T |Y|^2 \,\mathrm{d}t)^{\frac{1}{2}}$, as otherwise the right-hand side of the inequality would be infinity. Since $\|\cdot\|_{\Psi_\alpha}$ is a quasi-norm for any $\alpha > 0$, we shall assume without loss of generality that $\|\|X\|_{L^2(0,T)}\|_{\Psi_\alpha} = \|\|Y\|_{L^2(0,T)}\|_{\Psi_\alpha} = 1$. Then, we can deduce from Hölder's inequality and Young's inequality that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\left|\int_0^T XY \,\mathrm{d}t\right|^{\frac{\alpha}{2}}\right)\right] \le \mathbb{E}\left[\exp\left(\left|\|X\|_{L^2(0,T)}\|Y\|_{L^2(0,T)}\right|^{\frac{\alpha}{2}}\right)\right] \\
&\le \mathbb{E}\left[\exp\left(\frac{1}{2}\|X\|_{L^2(0,T)}^\alpha + \frac{1}{2}\|Y\|_{L^2(0,T)}^\alpha\right)\right] = \mathbb{E}\left[\exp\left(\frac{1}{2}\|X\|_{L^2(0,T)}^\alpha\right)\exp\left(\frac{1}{2}\|Y\|_{L^2(0,T)}^\alpha\right)\right] \\
&\le \left(\mathbb{E}\left[\exp\left(\|X\|_{L^2(0,T)}^\alpha\right)\right]\right)^{\frac{1}{2}}\left(\mathbb{E}\left[\exp\left(\|Y\|_{L^2(0,T)}^\alpha\right)\right]\right)^{\frac{1}{2}} \le 2,
\end{aligned}$$

which implies that $\|\int_0^T XY \,\mathrm{d}t\|_{\Psi_{\alpha/2}} \le 1$ and finishes the proof. □

Proof of Lemma 3.3.6. Note that (3.3.9) and Hölder's inequality suggest that it suffices to estimate the growth of $\|\cdot\|_{L^q}$-norms of the stochastic integrals for $q \ge 2$. Hence, by (3.6.7), there exists a constant C such that for all $q \ge 2$,

$$\begin{aligned}
q^{-2}\left\|\int_0^T X_t\sigma^\mathsf{T} \,\mathrm{d}W_t\right\|_{L^q} &\le q^{-2}Cq\left\|\left(\int_0^T |X_t\sigma|^2 \,\mathrm{d}t\right)^{\frac{1}{2}}\right\|_{L^q} \le C|\sigma| \sup_{q\ge 1}\left(q^{-1}\left\|\left(\int_0^T |X_t|^2 \,\mathrm{d}t\right)^{\frac{1}{2}}\right\|_{L^q}\right) \\
&\le C|\sigma|\left\|\left(\int_0^T |X_t|^2 \,\mathrm{d}t\right)^{\frac{1}{2}}\right\|_{\Psi_1},
\end{aligned}$$

which along with (3.3.9) leads to the desired estimate for $\|\int_0^T X_t\sigma^\mathsf{T} \,\mathrm{d}W_t\|_{\Psi_{1/2}}$.

Similarly, by (3.6.8), there exists a constant C satisfying for all $q \geq 2$ that

$$\begin{aligned}
&\left\| \int_0^T \int_{\mathbb{R}_0^p} X_t \gamma(u)\, \tilde{N}(\mathrm{d}t, \mathrm{d}u) \right\|_{L^q} \\
&\leq Cq^2 \left\{ \left(\mathbb{E}\left[\int_0^T \int_{\mathbb{R}_0^p} |X_t \gamma(u)|^q \, \nu(\mathrm{d}u)\mathrm{d}t \right] \right)^{\frac{1}{q}} + \left(\mathbb{E}\left[\left(\int_0^T \int_{\mathbb{R}_0^p} |X_t \gamma(u)|^2 \, \nu(\mathrm{d}u)\mathrm{d}t \right)^{\frac{q}{2}} \right] \right)^{\frac{1}{q}} \right\} \\
&\leq Cq^2 \left\{ \left(\int_{\mathbb{R}_0^p} |\gamma(u)|^q \, \nu(\mathrm{d}u) \mathbb{E}\left[\int_0^T |X_t|^q \mathrm{d}t \right] \right)^{\frac{1}{q}} + \left(\int_{\mathbb{R}_0^p} |\gamma(u)|^2 \, \nu(\mathrm{d}u) \right)^{\frac{1}{2}} \left(\mathbb{E}\left[\left(\int_0^T |X_t|^2 \, \mathrm{d}t \right)^{\frac{q}{2}} \right] \right)^{\frac{1}{q}} \right\} \\
&\leq Cq^2 \left\{ \left(\int_{\mathbb{R}_0^p} |\gamma(u)|^q \, \nu(\mathrm{d}u) \right)^{\frac{1}{q}} \left\| \left(\int_0^T |X_t|^q \mathrm{d}t \right)^{\frac{1}{q}} \right\|_{L^q} + \left(\int_{\mathbb{R}_0^p} |\gamma(u)|^2 \, \nu(\mathrm{d}u) \right)^{\frac{1}{2}} \left\| \left(\int_0^T |X_t|^2 \, \mathrm{d}t \right)^{\frac{1}{2}} \right\|_{L^q} \right\}.
\end{aligned}$$

Hence by (H.4(2)) and (3.3.9), for all $q \geq 2$,

$$\begin{aligned}
&q^{-(3+\vartheta)} \left\| \int_0^T \int_{\mathbb{R}_0^p} X_t \gamma(u)\, \tilde{N}(\mathrm{d}t, \mathrm{d}u) \right\|_{L^q} \\
&\leq C \left(\sup_{q \geq 2} q^{-\vartheta} \left(\int_{\mathbb{R}_0^p} |\gamma(u)|^q \, \nu(\mathrm{d}u) \right)^{\frac{1}{q}} \right) \left\{ q^{-1} \left\| \left(\int_0^T |X_t|^q \mathrm{d}t \right)^{\frac{1}{q}} \right\|_{L^q} + q^{-1} \left\| \left(\int_0^T |X_t|^2 \, \mathrm{d}t \right)^{\frac{1}{2}} \right\|_{L^q} \right\} \\
&\leq C\gamma_{\max} \left(\left\| \left(\int_0^T |X_t|^q \mathrm{d}t \right)^{\frac{1}{q}} \right\|_{\Psi_1} + \left\| \left(\int_0^T |X_t|^2 \, \mathrm{d}t \right)^{\frac{1}{2}} \right\|_{\Psi_1} \right).
\end{aligned}$$

Therefore, taking the supremum over $q \geq 2$ in the above inequality leads to the desired estimate of $\| \int_0^T \int_{\mathbb{R}_0^p} X_t \gamma(u)\, \tilde{N}(\mathrm{d}t, \mathrm{d}u) \|_{\Psi_{1/(3+\vartheta)}}$ from (3.3.9). □

Proof of Lemma 3.3.7. Let us assume without loss of generality that $|\sigma| > 0$ and $\tau := \|(\int_0^T |X_t|^2 \, \mathrm{d}t)^{1/2}\|_{\Psi_2} < \infty$, which implies that $\|(\int_0^T 2|X_t\sigma|^2 \, \mathrm{d}t)^{1/2}\|_{\Psi_2} \leq \sqrt{2}|\sigma|\tau$. Then, by the characterization of sub-Gaussian random variable in [156, Proposition 2.5.2(iii)], there exists $C \geq 0$ such that

$$\mathbb{E}\left[\exp\left(2\lambda^2 \int_0^T |X_t\sigma|^2 \, \mathrm{d}t \right) \right] \leq \exp(2C^2\lambda^2|\sigma|^2\tau^2) < \infty \quad \forall |\lambda| \leq \frac{1}{\sqrt{2}C|\sigma|\tau}.$$

Hence, it holds for all $|\lambda| \leq 1/(\sqrt{2}C|\sigma|\tau)$ that the process $(M_{\lambda,t})_{t\in[0,T]}$ defined by:

$$M_{\lambda,t} := \exp\left(\int_0^t 2\lambda X_s \sigma^\intercal \mathrm{d}W_s - \frac{1}{2} \int_0^t 4\lambda^2 |X_s\sigma|^2 \mathrm{d}s \right) \quad \forall t \in [0,T]$$

is a martingale, since Novikov's condition is satisfied, which implies that $\mathbb{E}[M_{\lambda,T}] = 1$. Thus, for any given $|\lambda| \leq 1/(\sqrt{2}C|\sigma|\tau)$, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}\bigg[\exp\bigg(\lambda\int_0^T X_t\sigma^\intercal \mathrm{d}W_t\bigg)\bigg] \\
&= \mathbb{E}\bigg[\exp\bigg(\int_0^T \lambda X_t\sigma^\intercal \mathrm{d}W_t - \frac{(2\lambda)^2}{4}\int_0^T |X_t\sigma|^2\,\mathrm{d}t\bigg)\exp\bigg(\frac{(2\lambda)^2}{4}\int_0^T |X_t\sigma|^2\,\mathrm{d}t\bigg)\bigg] \\
&\leq \mathbb{E}[M_{\lambda,T}]^{1/2}\mathbb{E}\bigg[\exp\bigg(2\lambda^2\int_0^T |X_t\sigma|^2\,\mathrm{d}t\bigg)\bigg]^{1/2} \leq \exp(C^2\lambda^2|\sigma|^2\tau^2),
\end{aligned}$$

which along with the fact that $\mathbb{E}[\int_0^T X_t\sigma^\intercal \mathrm{d}W_t] = 0$ and the characterization of sub-exponential random variable [156, Proposition 2.7.1(v)] yields that $\|\int_0^T X_t\sigma^\intercal \mathrm{d}W_t\|_{\Psi_1} \leq C|\sigma|\tau$. □

Proof of Lemma 3.3.8. Throughout this proof, let $x_0 \in \mathbb{R}^n$ and $\theta \in \mathbb{R}^{n\times(n+k)}$ be given constants satisfying $|\theta| \leq K$, and let C be a generic constant depend on K, T and the constants in (H.4), but independent of x_0 and θ.

By (3.3.4), we see that the process $X^{x_0,\theta}$ satisfies the SDE:

$$\mathrm{d}X_t = b^\theta(t, X_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t + \int_{\mathbb{R}_0^p} \gamma(u)\,\tilde{N}(\mathrm{d}t, \mathrm{d}u), \quad t \in [0,T], \quad X_0 = x_0,$$

where $b^\theta(t,x) = A^\star x + B^\star\psi^\theta(t,x)$ for all $(t,x) \in [0,T]\times\mathbb{R}^n$. The definition of the feedback control ψ^θ, (H.4(1)) and Theorem 3.2.5 show that there exists $C \geq 0$ such that $|\psi^\theta(t,0)| \leq C$ and $|\psi^\theta(t,x) - \psi^\theta(t,x')| \leq C|x - x'|$ for all $t \in [0,T]$, $x, x' \in \mathbb{R}^n$, which implies the same properties for the function b^θ. Then, by Lemma 3.6.3, for every Lipschitz continuous function $\mathfrak{f} : (\mathbb{D}([0,T];\mathbb{R}^n), d_\infty) \to \mathbb{R}$, $\mathbb{E}\big[\exp\big(\lambda(\mathfrak{f}(X^{x_0,\theta}) - \mathbb{E}[\mathfrak{f}(X^{x_0,\theta})])\big)\big] \leq \exp\big(C\eta\big(C\lambda\|\mathfrak{f}\|_{\mathrm{Lip}}\big)\big)$ for all $\lambda > 0$, with the function $\eta : [0,\infty] \to [0,\infty]$ defined by:

$$\eta(\lambda) := \int_{\mathbb{R}_0^p} \big(e^{\lambda\gamma(u)} - \lambda\gamma(u) - 1\big)\,\nu(\mathrm{d}u) + \frac{\sigma^2}{2}\lambda^2 \quad \forall\lambda > 0. \tag{3.6.14}$$

By (H.4(2)) and Stirling's approximation $q! \geq (q/e)^q$ for all $q \geq 2$, we have for each $\lambda \in [0, 1/(2\gamma_{\max}e))$,

$$\begin{aligned}
&\int_{\mathbb{R}_0^p} \big(e^{\lambda\gamma(u)} - \lambda\gamma(u) - 1\big)\,\nu(\mathrm{d}u) \\
&= \int_{\mathbb{R}_0^p} \sum_{q=2}^\infty \frac{|\lambda\gamma(u)|^q}{q!}\,\nu(\mathrm{d}u) = \sum_{q=2}^\infty \frac{\lambda^q}{q!}\int_{\mathbb{R}_0^p} |\gamma(u)|^q\,\nu(\mathrm{d}u) \leq \sum_{q=2}^\infty \frac{\lambda^q}{q!}\gamma_{\max}^q q^{\vartheta q} \\
&\leq \sum_{q=2}^\infty \frac{(\lambda\gamma_{\max}e)^q}{q^{(1-\vartheta)q}} \leq \frac{(\lambda\gamma_{\max}e)^2}{1 - \lambda\gamma_{\max}e} \leq 2(\lambda\gamma_{\max}e)^2,
\end{aligned}$$

which implies for all $0 \leq \lambda \leq 1/C$ and $\mathfrak{f} : (\mathbb{D}([0,T];\mathbb{R}^n), d_\infty) \to \mathbb{R}$ satisfying $\|\mathfrak{f}\|_{\text{Lip}} \leq 1$ that $\mathbb{E}\big[\exp\big(\lambda(\mathfrak{f}(X^{x_0,\theta}) - \mathbb{E}[\mathfrak{f}(X^{x_0,\theta})])\big)\big] \leq \exp(C^2\lambda^2)$. Replacing $\mathfrak{f}$ with $-\mathfrak{f}$ shows that the same estimate holds for all for all $|\lambda| \leq 1/C$, which, along with the characterization of sub-exponential random variable in [156, Proposition 2.7.1(v)], leads to $\|\mathfrak{f}(X^{x_0,\theta}) - \mathbb{E}[\mathfrak{f}(X^{x_0,\theta})]\|_{\Psi_1} \leq C$ for some constant C, uniformly with respect to $x_0 \in \mathbb{R}^n$, $|\theta| \leq K$ and $\mathfrak{f} : (\mathbb{D}([0,T];\mathbb{R}^n), d_\infty) \to \mathbb{R}$ satisfying $\|\mathfrak{f}\|_{\text{Lip}} \leq 1$.

Since $\|\cdot\|_{\Psi_1}$ is a norm and $\|\mathbb{E}[\mathfrak{f}(X^x)]\|_{\Psi_1} \leq |\mathbb{E}[\mathfrak{f}(X^x)]|/\ln 2$, $\|\mathfrak{f}(X^{x_0,\theta})\|_{\Psi_1} \leq C(1 + |\mathbb{E}[\mathfrak{f}(X^{x_0,\theta})]|)$ for all $\mathfrak{f}$ with $\|\mathfrak{f}\|_{\text{Lip}} \leq 1$. The estimate for a general Lipschitz continuous function $\mathfrak{f}$ follows by considering $\mathfrak{f}/\|\mathfrak{f}\|_{\text{Lip}}$ and by using the fact that $\|\cdot\|_{\Psi_1}$ is a norm. $\square$

Bibliography

[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

[2] Marc Abeille and Alessandro Lazaric. Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pages 1–9. PMLR, 2018.

[3] B. Acciaio, J. Backhoff, and R. Carmona. Extended mean field control problems: stochastic maximum principle and transport perspective. *Arxiv Preprint:1802.05754*, 2018.

[4] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[5] Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] David J Aldous. *Weak convergence and the general theory of processes*. Editeur inconnu, 1981.

[7] Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Q-learning in regularized mean-field games. *arXiv preprint arXiv:2003.12151*, 2020.

[8] N. Andelman and Y. Mansour. Auctions with budget constraints. In *Scandinavian Workshop on Algorithm Theory*, pages 26–38. Springer, 2004.

[9] Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. Unified reinforcement q-learning for mean field game and control problems. *arXiv preprint arXiv:2006.13912*, 2020.

[10] Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Lauriere. Reinforcement learning for mean field games, with applications to economics. *arXiv preprint arXiv:2106.13755*, 2021.

[11] K. Asadi and M. L. Littman. An alternative softmax operator for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 243–252, 2017.

[12] Karl J Åström and Björn Wittenmark. *Adaptive control*. Courier Corporation, 2013.

[13] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 89–96, 2009.

[14] Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 49–56, 2007.

[15] J Backhoff and Francisco José Silva. Sensitivity results in stochastic optimal control: A Lagrangian perspective. *ESAIM: Control, Optimisation and Calculus of Variations*, 23(1):39–70, 2017.

[16] Rafael Bailo, Mattia Bongini, José A Carrillo, and Dante Kalise. Optimal consensus control of the Cucker-Smale model. *IFAC-PapersOnLine*, 51(13):1–6, 2018.

[17] Matteo Basei, Xin Guo, Anran Hu, and Yufei Zhang. Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *arXiv preprint arXiv:2006.15316*, 2020.

[18] Erhan Bayraktar, Yan Dolinsky, and Jia Guo. Continuity of utility maximization under weak convergence. *Mathematics and Financial Economics*, 14(4):725–757, 2020.

[19] M. G. Bellemare, G. Ostrovski, A. Guez, P. S. Thomas, and R. Munos. Increasing the action gap: new operators for reinforcement learning. In *AAAI Conference on Artificial Intelligence*, pages 1476–1483, 2016.

[20] M. Benaim and J. Y. Le Boudec. A class of mean field interaction models for computer and communication systems. *Performance evaluation*, 65(11-12):823–838, 2008.

[21] Alain Bensoussan, SP Sethi, Raymond Vickson, and N Derzko. Stochastic production planning with production constraints. *SIAM Journal on Control and Optimization*, 22(6):920–935, 1984.

[22] Mark A Bernstein and James Griffin. Regional differences in the price-elasticity of demand for energy. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2006.

[23] Klaus Bichteler and Bichteler Klaus. *Stochastic integration with jumps*. Number 89. Cambridge University Press, 2002.

[24] Robert R Bitmead and Michel Gevers. Riccati difference and differential equations: Convergence, monotonicity and stability. In *The Riccati Equation*, pages 263–291. Springer, 1991.

[25] F. Bolley. Separability and completeness for the Wasserstein distance. *Séminaire de Probabilités XLI*, pages 371–377, 2008.

[26] J Frédéric Bonnans and Alexander Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

[27] Jean-Christophe Breton and Nicolas Privault. Integrability and regularity of the flow of stochastic differential equations with jumps. *arXiv preprint arXiv:1902.03542*, 2019.

[28] H. Cai, K. Ren, W. Zhang, K. Malialis, J. Wang, Y. Yu, and D. Guo. Real-time bidding by reinforcement learning in display advertising. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 661–670. ACM, 2017.

[29] P. E. Caines, M. Huang, and R. P. Malhamé. Mean field games. In T. Basar and G. Zaccour, editors, *Handbook of Dynamic Game Theory*. Springer, Berlin, 2017.

[30] Marco C Campi and PR Kumar. Adaptive linear quadratic gaussian control: the cost-biased approach revisited. *SIAM Journal on Control and Optimization*, 36(6):1890–1907, 1998.

[31] R. Carmona, M. Laurière, and Z. Tan. Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*, 2019.

[32] R. Carmona, M. Laurière, and Z Tan. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. *arXiv preprint arXiv:1910.12802*, 2019.

[33] Alvaro Cartea, Sebastian Jaimungal, and Jason Ricci. Algorithmic trading, stochastic control, and mutually exciting processes. *SIAM Review*, 60(3):673–703, 2018.

[34] Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning linear quadratic regulators efficiently. In *International Conference on Machine Learning*, pages 1328–1337. PMLR, 2020.

[35] Xinyi Chen and Elad Hazan. Black-box control for linear dynamical systems. In *Conference on Learning Theory*, pages 1114–1143. PMLR, 2021.

[36] Yang Chen, Jiamou Liu, and Bakhadyr Khoussainov. Maximum entropy inverse reinforcement learning for mean field games. *arXiv preprint arXiv:2104.14654*, 2021.

[37] Patrick Cheridito, H. Mete Soner, and Nizar Touzi. Small time path behavior of double stochastic integrals and applications to stochastic control. *The Annals of Applied Probability*, 15(4):2472–2495, 2005.

[38] Philippe G Ciarlet. *Linear and nonlinear functional analysis with applications*, volume 130. Siam, 2013.

[39] Christian Clason, Armin Rund, Karl Kunisch, and Richard C Barnard. A convex penalty for switching control of partial differential equations. *Systems & Control Letters*, 89:66–73, 2016.

[40] Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear quadratic regulators efficiently with only $\sqrt{T}$ regret. *arXiv preprint arXiv:1902.06223*, 2019.

[41] Kai Cui and Heinz Koeppl. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1909–1917. PMLR, 2021.

[42] Kai Cui, Anam Tahir, Mark Sinzger, and Heinz Koeppl. Discrete-time mean field control with environment states. *arXiv preprint arXiv:2104.14900*, 2021.

[43] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. pages 5713–5723, 2017.

[44] Martino Bernasconi de Luca, Edoardo Vittori, Francesco Trovò, and Marcello Restelli. Dealer markets: a reinforcement learning mean field approach. 2021.

[45] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.

[46] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

[47] François Delarue and Athanasios Vasileiadis. Exploration noise for learning linear-quadratic mean field games. *arXiv preprint arXiv:2107.00839*, 2021.

[48] E. Derman and S. Mannor. Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894*, 2020.

[49] Hacene Djellout, Arnaud Guillin, and Liming Wu. Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *The Annals of Probability*, 32(3B):2702–2732, 2004.

[50] Tyrone E Duncan, Lei Guo, and Bozenna Pasik-Duncan. Adaptive continuous-time linear quadratic Gaussian control. *IEEE Transactions on Automatic Control*, 44(9):1653–1662, 1999.

[51] Tyrone E Duncan, Petr Mandl, and Bożenna Pasik-Duncan. On least squares estimation in continuous time linear stochastic systems. *Kybernetika*, 28(3):169–180, 1992.

[52] Romuald Elie, Julien Perolat, Mathieu Laurière, Matthieu Geist, and Olivier Pietquin. On the convergence of model free learning in mean field games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7143–7150, 2020.

[53] E. Even-Dar and Y. Mansour. Learning rates for Q-learning. *Journal of Machine Learning Research*, 5(Dec):1–25, 2003.

[54] Ioannis Exarchos, Evangelos A Theodorou, and Panagiotis Tsiotras. Stochastic L^1-optimal control via forward and backward sampling. *Systems & Control Letters*, 118:101–108, 2018.

[55] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite-time adaptive stabilization of linear systems. *IEEE Transactions on Automatic Control*, 64(8):3498–3505, 2018.

[56] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On adaptive linear-quadratic regulators. *arXiv e-prints*, pages arXiv–1806, 2018.

[57] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Randomized algorithms for data-driven stabilization of stochastic linear systems. In *2019 IEEE Data Science Workshop (DSW)*, pages 170–174. IEEE, 2019.

[58] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Input perturbations for adaptive control and learning. *Automatica*, 117:108950, 2020.

[59] Dylan Foster and Max Simchowitz. Logarithmic regret for adversarial online control. In *International Conference on Machine Learning*, pages 3211–3221. PMLR, 2020.

[60] Zuyue Fu, Zhuoran Yang, Yongxin Chen, and Zhaoran Wang. Actor-critic provably finds nash equilibria of linear-quadratic mean-field games. *arXiv preprint arXiv:1910.07498*, 2019.

[61] Mukul Gagrani, Sagar Sudhakara, Aditya Mahajan, Ashutosh Nayyar, and Yi Ouyang. Thompson sampling for linear quadratic mean-field teams. *arXiv preprint arXiv:2011.04686*, 2020.

[62] B. Gao and L. Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *Arxiv Preprint:1704.00805*, 2017.

[63] M. Geist, B. Scherrer, and O. Pietquin. A theory of regularized Markov decision processes. *arXiv preprint arXiv:1901.11275*, 2019.

[64] Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: the mean-field game viewpoint. *arXiv preprint arXiv:2106.03787*, 2021.

[65] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[66] Arnob Ghosh and Vaneet Aggarwal. Model free reinforcement learning algorithm for stationary mean field equilibrium for multiple types of agents. *arXiv preprint arXiv:2012.15377*, 2020.

[67] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[68] D. A. Gomes, J. Mohr, and R. R. Souza. Discrete time, finite state space mean field games. *Journal de mathmatiques pures et appliques*, 93(3):308–328, 2010.

[69] Graham C Goodwin, Peter J Ramadge, and Peter E Caines. Discrete time stochastic adaptive control. *SIAM Journal on Control and Optimization*, 19(6):829–853, 1981.

[70] Friedrich Götze, Holger Sambale, and Arthur Sinulis. Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability*, 26:1–22, 2021.

[71] P Jameson Graber. Linear quadratic mean field type control and mean field games with common noise, with application to production of an exhaustible resource. *Applied Mathematics & Optimization*, 74(3):459–486, 2016.

[72] Michael Green and John B Moore. Persistence of excitation in linear systems. *Systems & control letters*, 7(5):351–360, 1986.

[73] Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field controls with q-learning for cooperative marl: Convergence and complexity analysis. *arXiv preprint arXiv:2002.04131*, 2020.

[74] Olivier Guéant, Jean-Michel Lasry, and Pierre-Louis Lions. Mean field games and applications. In *Paris-Princeton lectures on mathematical finance 2010*, pages 205–266. Springer, 2011.

[75] R. Gummadi, P. Key, and A. Proutiere. Repeated auctions under budget constraints: Optimal bidding strategies and equilibria. In *the Eighth Ad Auction Workshop*, 2012.

[76] X. Guo, A. Hu, R. Xu, and J. Zhang. Learning mean-field games. In *Advances in Neural Information Processing Systems*, pages 4967–4977, 2019.

[77] Xin Guo, Renyuan Xu, and Thaleia Zariphopoulou. Entropy regularization for mean field games with learning. *arXiv preprint arXiv:2010.00145*, 2020.

[78] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. Reinforcement learning with deep energy-based policies. *Arxiv Preprint:1702.08165*, 2017.

[79] J. Hamari, M. Sjöklint, and A. Ukkonen. The sharing economy: Why people participate in collaborative consumption. *Journal of the Association for Information Science and Technology*, 67(9):2047–2059, 2016.

[80] Ben M Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *Available at SSRN*, 2020.

[81] UG Haussmann and JP Lepeltier. On the existence of optimal controls. *SIAM Journal on Control and Optimization*, 28(4):851–902, 1990.

[82] P. Hernandez-Leal, B. Kartal, and M. E. Taylor. Is multiagent deep reinforcement learning the answer or the question? A brief survey. *Arxiv Preprint:1810.05587*, 2018.

[83] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

[84] J. Hu and M. P. Wellman. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4(Nov):1039–1069, 2003.

[85] Shuyue Hu, Chin-Wing Leung, Ho-fung Leung, and Harold Soh. The evolutionary dynamics of independent learning agents in population games. *arXiv preprint arXiv:2006.16068*, 2020.

[86] M. Huang and Y. Ma. Mean field stochastic games with binary action spaces and monotone costs. *ArXiv Preprint:1701.06661*, 2017.

[87] M. Huang, R. P. Malhamé, and P. E. Caines. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems*, 6(3):221–252, 2006.

[88] Petros Ioannou and Bariş Fidan. *Adaptive control tutorial*. SIAM, 2006.

[89] Petros A Ioannou and Jing Sun. *Robust adaptive control*. Courier Corporation, 2012.

[90] K. Iyer, R. Johari, and M. Sundararajan. Mean field equilibria of dynamic auctions with learning. *ACM SIGecom Exchanges*, 10(3):10–14, 2011.

[91] S. H. Jeong, A. R. Kang, and H. K. Kim. Analysis of game bot's behavioral characteristics in social interaction networks of MMORPG. *ACM SIGCOMM Computer Communication Review*, 45(4):99–100, 2015.

[92] J. Jin, C. Song, H. Li, K. Gai, J. Wang, and W. Zhang. Real-time bidding with multi-agent reinforcement learning in display advertising. *Arxiv Preprint:1802.09756*, 2018.

[93] S. Kapoor. Multi-agent reinforcement learning: A report on challenges and approaches. *Arxiv Preprint:1807.09427*, 2018.

[94] Ali D Kara and Serdar Yuksel. Robustness to incorrect system models in stochastic control. *SIAM Journal on Control and Optimization*, 58(2):1144–1182, 2020.

[95] A. C Kizilkale and P. E Caines. Mean field stochastic adaptive control. *IEEE Transactions on Automatic Control*, 58(4):905–920, 2013.

[96] Nicolai V Krylov and Enrico Priola. Elliptic and parabolic second-order PDEs with growing coefficients. *Communications in Partial Differential Equations*, 35(1):1–22, 2009.

[97] Nikolai Vladimirovich Krylov. *Nonlinear elliptic and parabolic equations of the second order*. Springer, 1987.

[98] Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.

[99] PR Kumar. Optimal adaptive control of linear-quadratic-gaussian systems. *SIAM Journal on Control and Optimization*, 21(2):163–178, 1983.

[100] Daniel Lacker. Mean field games via controlled martingale problems: existence of Markovian equilibria. *Stochastic Processes and their Applications*, 125(7):2856–2894, 2015.

[101] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Explore more and improve regret in linear quadratic regulators. *arXiv preprint arXiv:2007.12291*, 2020.

[102] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 33:20876–20888, 2020.

[103] Ioan Doré Landau, Rogelio Lozano, Mohammed M'Saad, and Alireza Karimi. *Adaptive control: algorithms, analysis and applications*. Springer Science & Business Media, 2011.

[104] J-M. Lasry and P-L. Lions. Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260, 2007.

[105] Kiyeob Lee, Desik Rengarajan, Dileep Kalathil, and Srinivas Shakkottai. Reinforcement learning for mean field games with strategic complementarities. In *International Conference on Artificial Intelligence and Statistics*, pages 2458–2466. PMLR, 2021.

[106] C-A. Lehalle and C. Mouzouni. A mean field game of portfolio trading and its consequences on perceived correlations. *ArXiv Preprint:1902.09606*, 2019.

[107] Chen Li and Georg Stadler. Sparse solutions in optimal control of PDEs with uncertain parameters: The linear case. *SIAM Journal on Control and Optimization*, 57(1):633–658, 2019.

[108] Juan Li and Qingmeng Wei. L^p estimates for fully coupled FBSDEs with jumps. *Stochastic Processes and their Applications*, 124(4):1582–1611, 2014.

[109] J. P. M. López. Discrete time mean field games: The short-stage limit. *Journal of Dynamics & Games*, 2(1):89–101, 2015.

[110] Yuwei Luo, Zhuoran Yang, Zhaoran Wang, and Mladen Kolar. Natural actor-critic converges globally for hierarchical linear quadratic regulator. *arXiv preprint arXiv:1912.06875*, 2019.

[111] Yutao Ma. Transportation inequalities for stochastic differential equations with jumps. *Stochastic processes and their applications*, 120(1):2–21, 2010.

[112] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32:10154–10164, 2019.

[113] Xuerong Mao. *Stochastic differential equations and applications*. Elsevier, 2007.

[114] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

[115] D. Mguni, J. Jennings, and E. M. de Cote. Decentralised learning in systems with many, many strategic agents. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[116] V. M. Minh, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, 2016.

[117] Rajesh K Mishra, Deepanshu Vasal, and Sriram Vishwanath. Model-free reinforcement learning for non-stationary mean field games. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 1032–1037. IEEE, 2020.

[118] Hamidreza Modares and Frank L. Lewis. Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning. *IEEE Transactions on Automatic Control*, 59(11):3051–3056, 2014.

[119] Rémi Munos. A study of reinforcement learning in the continuous case by the means of viscosity solutions. *Machine Learning*, 40(3):265–299, 2000.

[120] Rémi Munos. Policy gradient in continuous time. *Journal of Machine Learning Research*, 7(May):771–791, 2006.

[121] Rémi Munos and Paul Bourgine. Reinforcement learning for continuous stochastic control problems. In *Advances in Neural Information Processing Systems*, pages 1029–1035, 1998.

[122] G. Neu, A. Jonsson, and V. Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

[123] Karoui Nicole el, Nguyen Du'hŪŪ, and Jeanblanc-Picqué Monique. Compactification methods in the control of degenerate diffusions: existence of an optimal control. *Stochastics: an international journal of probability and stochastic processes*, 20(3):169–219, 1987.

[124] Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

[125] Yi Ouyang, Mukul Gagrani, and Rahul Jain. Learning-based control of unknown linear systems with thompson sampling. *arXiv preprint arXiv:1709.04047*, 2017.

[126] C. H. Papadimitriou and T. Roughgarden. Computing equilibria in multi-player games. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 82–91, 2005.

[127] Barna Pasztor, Ilija Bogunovic, and Andreas Krause. Efficient model-based multi-agent mean-field reinforcement learning. *arXiv preprint arXiv:2107.04050*, 2021.

[128] J. Pérolat, B. Piot, and O. Pietquin. Actor-critic fictitious play in simultaneous move multistage games. In *International Conference on Artificial Intelligence and Statistics*, 2018.

[129] Julien Perolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. Scaling up mean field games with online mirror descent. *arXiv preprint arXiv:2103.00623*, 2021.

[130] Sarah Perrin, Mathieu Laurière, Julien Pérolat, Matthieu Geist, Romuald Élie, and Olivier Pietquin. Mean field games flock! the reinforcement learning way. *arXiv preprint arXiv:2105.07933*, 2021.

[131] Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. *arXiv preprint arXiv:2007.03458*, 2020.

[132] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[133] Marie-Claire Quenez and Agnès Sulem. BSDEs with jumps, optimization and applications to dynamic risk measures. *Stochastic Processes and their Applications*, 123(8):3328–3357, 2013.

[134] Christoph Reisinger, Wolfgang Stockinger, and Yufei Zhang. Path regularity of coupled McKean-Vlasov FBSDEs. *arXiv preprint arXiv:2011.06664*, 2020.

[135] Christoph Reisinger and Yufei Zhang. Regularity and stability of feedback relaxed controls. *SIAM Journal on Control and Optimization*, 59(5):3118–3151, 2021.

[136] Syed Ali Asad Rizvi and Zongli Lin. Output feedback reinforcement learning control for the continuous-time linear quadratic regulator problem. In *2018 Annual American Control Conference (ACC)*, pages 3417–3422. IEEE, 2018.

[137] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[138] N. Saldi, T. Basar, and M. Raginsky. Markov–Nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018.

[139] Shankar Sastry and Marc Bodson. *Adaptive control: stability, convergence and robustness*. Courier Corporation, 2011.

[140] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

[141] L. Shani, Y. Efroni, and S. Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. *arXiv preprint arXiv:1909.02769*, 2019.

[142] Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

[143] David Šiška and Łukasz Szpruch. Gradient flows for regularized stochastic control problems. *arXiv preprint arXiv:2006.05956*, 2020.

[144] J. Subramanian and A. Mahajan. Reinforcement learning in stationary mean-field games. In *18th International Conference on Autonomous Agents and Multiagent Systems*, pages 251–259, 2019.

[145] Sriram Ganapathi Subramanian, Pascal Poupart, Matthew E Taylor, and Nidhi Hegde. Multi type mean field reinforcement learning. *arXiv preprint arXiv:2002.02513*, 2020.

[146] Sriram Ganapathi Subramanian, Matthew E Taylor, Mark Crowley, and Pascal Poupart. Partially observable mean field reinforcement learning. *arXiv preprint arXiv:2012.15791*, 2020.

[147] Richard S. Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[148] Lukasz Szpruch, Tanut Treetanthiploet, and Yufei Zhang. Exploration-exploitation trade-off for continuous-time episodic reinforcement learning with linear-convex models. *arXiv preprint arXiv:2112.10264*, 2021.

[149] Corentin Tallec, Léonard Blier, and Yann Ollivier. Making Deep Q-learning methods robust to time discretization. *arXiv preprint arXiv:1901.09732*, 2019.

[150] M. Tan. Multi-agent reinforcement learning: independent vs. cooperative agents. In *International Conference on Machine Learning*, pages 330–337, 1993.

[151] Shanjian Tang and Xunjing Li. Necessary conditions for optimal control of stochastic systems with random jumps. *SIAM Journal on Control and Optimization*, 32(5):1447–1475, 1994.

[152] Wenpin Tang, Paul Yuming Zhang, and Xun Yu Zhou. Exploratory HJB equations and their convergence. *arXiv preprint arXiv:2109.10269*, 2021.

[153] Muhammad Aneeq uz Zaman, Kaiqing Zhang, Erik Miehling, and Tamer Başar. Approximate equilibrium computation for discrete-time linear-quadratic mean-field games. In *2020 American Control Conference (ACC)*, pages 333–339. IEEE, 2020.

[154] Muhammad Aneeq uz Zaman, Kaiqing Zhang, Erik Miehling, and Tamer Baar. Reinforcement learning in non-stationary discrete-time linear-quadratic mean-field games. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 2278–2284. IEEE, 2020.

[155] Mark Veraar. The stochastic Fubini theorem revisited. *Stochastics An International Journal of Probability and Stochastic Processes*, 84(4):543–551, 2012.

[156] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[157] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[158] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[159] Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Exploration versus exploitation in reinforcement learning: a stochastic control approach. *Available at SSRN 3316387*, 2019.

[160] Haoran Wang and Xun Yu Zhou. Continuous-time mean–variance portfolio selection: A reinforcement learning framework. *Mathematical Finance*, 30(4):1273–1308, 2020.

[161] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

[162] Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Breaking the curse of many agents: Provable mean embedding q-iteration for mean-field reinforcement learning. In *International Conference on Machine Learning*, pages 10092–10103. PMLR, 2020.

[163] Weichen Wang, Jiequn Han, Zhuoran Yang, and Zhaoran Wang. Global convergence of policy gradient for linear-quadratic mean-field control/game in continuous time. In *International Conference on Machine Learning*, pages 10772–10782. PMLR, 2021.

[164] Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. Provable fictitious play for general mean-field games. *arXiv preprint arXiv:2010.04211*, 2020.

[165] Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. Learning while playing in mean-field games: Convergence and optimality. In *International Conference on Machine Learning*, pages 11436–11447. PMLR, 2021.

[166] J. Yang, X. Ye, R. Trivedi, H. Xu, and H. Zha. Deep mean field games for learning optimal behavior policy of large populations. *Arxiv Preprint:1711.03156*, 2017.

[167] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang. Mean field multi-agent reinforcement learning. *Arxiv Preprint:1802.05438*, 2018.

[168] H. Yin, P. G. Mehta, S. P. Meyn, and U. V. Shanbhag. Learning in mean-field games. *IEEE Transactions on Automatic Control*, 59(3):629–644, 2013.

[169] Jiongmin Yong and Xun Yu Zhou. *Stochastic Controls: Hamiltonian Systems and HJB Equations*, volume 43. Springer Science & Business Media, 1999.