An α-Potential Game Framework for Non-Cooperative Dynamic Games: Theory and Algorithms

By

Xinyu Li

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 Steven N. Evans
Assistant Professor Thibaut Mastrolia

Summer 2025

An α-Potential Game Framework for Non-Cooperative Dynamic Games:
Theory and Algorithms

Copyright 2025
by
Xinyu Li

Abstract

An α-Potential Game Framework for Non-Cooperative Dynamic Games: Theory and Algorithms

by

Xinyu Li

Doctor of Philosophy in Engineering – Industrial Engineering and Operations Research

University of California, Berkeley

Professor Xin Guo, Chair

Multi-agent systems naturally arise in many real-world scenarios where multiple decision-makers interact, such as autonomous driving, network design, and financial markets. Analyzing multi-agent non-cooperative games is inherently challenging due to the asymmetry among players and the diverse structures of the games. This thesis introduces a unified framework to analyze N-player non-cooperative games in dynamic settings, considering both discrete-time and continuous-time state transitions. Additionally, efficient reinforcement learning (RL) algorithms and stochastic control techniques are proposed to identify strategies for each player that lead to approximate Nash equilibria (NE).

The first part of this thesis introduces and analyzes a general class of dynamic N-player non-cooperative games called α-potential games. In this framework, the change in a player's objective function resulting from a unilateral deviation from her strategy is equal to the change in a common function, called the α-potential function, up to an error α. The existence of an α-potential function simplifies the challenging task of finding α-Nash equilibria in dynamic games to minimizing the α-potential function, as the optimizer of the α-potential function is shown to be an α-Nash equilibrium of the game.

In Chapter 2, we focus on Markov games with finite state space, finite action space, and Markovian policy. The state transition follows a discrete-time Markov decision process. In this case, we establish the existence of an associated α-potential function. Additionally, we provide a semi-infinite linear program to find α and its corresponding α-potential function for any Markov game. We study two important classes of practically significant Markov games, Markov congestion games and the perturbed Markov team games, via the framework of Markov α-potential games, with explicit characterization of an upper bound for α and its relation to game parameters. Furthermore, we study two equilibrium approximation algorithms, namely the projected gradient-ascent algorithm and the sequential maximum

improvement algorithm, along with their Nash regret analysis, and corroborate the results with numerical experiments using model-free RL algorithms.

In Chapter 3, we study dynamic games with continuous-time state transitions, focusing on general classes of states, actions, and controls/policies, with a particular emphasis on stochastic differential games. An analytical characterization of the α-potential function is established, with α represented in terms of the magnitude of the asymmetry of the second-order derivatives of the players' objective functions. For stochastic differential games in which the state dynamic is a controlled diffusion, α is characterized in terms of the number of players, the choice of admissible strategies, and the intensity of interactions and the level of heterogeneity among players. Two classes of stochastic differential games, namely distributed games and games with mean field interactions, are analyzed to highlight the dependence of α on general game characteristics. To analyze the α-NE, the associated optimization problem is embedded into a conditional McKean-Vlasov control problem. A verification theorem is established to construct α-NE based on solutions to an infinite-dimensional Hamilton-Jacobi-Bellman equation, which is reduced to a system of ordinary differential equations for linear-quadratic (LQ) games. We conclude by case-studying an N-player LQ game on a graph network, analyzing α under different graph structures, and deriving the explicit solutions to the α-NE.

Since our framework reduces multi-agent games to a single optimization problem, the second part of this thesis focuses on designing efficient algorithms for single-agent reinforcement learning (RL). While much progress has been made in RL for discrete Markov decision processes, continuous RL remains less explored. Therefore, in Chapter 4, we propose and analyze two new policy learning methods: regularized policy gradient (RPG) and iterative policy optimization (IPO), for a class of discounted linear-quadratic control (LQC) problems with continuous state space and continuous action space, over an infinite time horizon with entropy regularization. Assuming access to the exact policy evaluation, both proposed approaches are proved to converge linearly in finding optimal policies. Moreover, the IPO method can achieve a super-linear convergence rate once it enters a local region around the optimal policy. Finally, when the optimal policy for an RL problem in a known environment is appropriately transferred as the initial policy to an RL problem in an unknown environment, the IPO method is shown to converge at a super-linear rate if the two environments are sufficiently close. A model-free version of the policy-based methods is also discussed. Performances of these proposed algorithms are supported by numerical examples.

Contents

List of Figures

Chapter 1

Introduction

Game theory has framed our understanding of strategic interaction since the seminal work of Von Neumann and Morgenstern [147] and the equilibrium concept of Nash [115]. Over the decades, it has become a foundational analytical tool across a wide range of fields, including economics [127, 7, 123], finance [143, 3, 35, 34, 70], transportation systems [159, 54, 83], and evolutionary biology [79, 20, 42].

In parallel, reinforcement learning (RL) has emerged as a powerful data-driven framework for sequential decision-making, particularly in settings where full information may not be available [138, 92]. Many successful applications of RL such as autonomous driving, the game of Go, and algorithmic trading, involve the interactions of multiple agents, which naturally fall into the realm of multi-agent RL (MARL). While classical MARL methods have achieved promising empirical results [104, 62, 162], the theoretical understanding of aspects such as convergence, stability, and equilibrium properties remain relatively limited in the literature.

In general-sum or unstructured environments, learning dynamics may fail to converge, show cyclic behaviors, or exhibit high sensitivity to parameter tuning [102, 109, 12]. Therefore, theoretical studies are often conducted on more structured classes of games, such as two-player zero-sum games [130], mean-field games and their variants [70, 4, 68, 5, 120], and Markov potential games [101, 131, 106, 163, 114]. While these formulations offer theoretical tractability, they may not fully capture the diversity of strategic interactions arising in more complex or heterogeneous real-world environments.

The gap between the rapid advancement of MARL algorithms and the limitations of existing game-theoretic frameworks motivates the central objective of this thesis. Specifically, this thesis proposes a new framework called α-potential games that aims to strike a balance between generality and tractability. This framework enables the analysis of general-sum games while supporting gradient-based learning and guaranteeing equilibrium convergence under suitable conditions. The remainder of this introduction first reviews the foundations of static, Markov, and stochastic differential games, and then introduces α-potential games in the context of recent theoretical and algorithmic advances.

1.1 Static Games

Definition of Static Games. Static games, also known as *one-shot simultaneous-move* or *normal-form games*, are among the most fundamental models in game theory. In these games, a finite set of players each selects a strategy simultaneously and independently, and each player's payoff depends on the full profile of strategies chosen by all participants.

A game $\mathcal{G}$ is defined by a tuple $(N, (A_i)_{i\in[N]}, (u_i)_{i\in[N]})$, where N is the number of players and $[N] = (1, 2, \cdots, N)$ denotes the set of player indices. For each player $i \in [N]$, A_i denotes the action space of player i. We denote the joint action profile of all players as $a \coloneqq (a_1, \cdots, a_N) \in A \coloneqq \prod_{i\in[N]} A_i$, and the joint action profile of all players except i as $a_{-i} \in A_{-i} \coloneqq \prod_{j\in[N],j\neq i} A_j$. The payoff function for player i is given by $u_i : A \to \mathbb{R}$. Each player i aims to maximize their own payoff u_i (or equivalently, minimize a cost function c_i, defined by $c_i = -u_i$).

Nash equilibrium. The central solution concept in game theory is the Nash equilibrium (NE). Intuitively, a Nash equilibrium is an action profile in which no player can unilaterally improve their payoff by deviating from their current action. In this sense, it characterizes a stationary point of the game where each player's choice is optimal given the choices of others. We now introduce the formal definition of a Nash equilibrium in both pure and mixed strategies.

Definition 1.1.1. *A pure strategy Nash equilibrium (NE) is a joint action profile $a^* = (a_1^*, \ldots, a_N^*) \in A$ such that for every player $i \in [N]$,*

$$u_i(a_i^*, a_{-i}^*) \geq u_i(a_i, a_{-i}^*) \quad \text{for all } a_i \in A_i.$$

When pure strategy equilibria do not exist or are not guaranteed, one may consider mixed strategy Nash equilibria instead, where each player randomizes over their action space.

Definition 1.1.2. *Let $\mathcal{P}(A_i)$ denote the set of probability distributions over A_i. A mixed strategy profile $\sigma^* = (\sigma_1^*, \ldots, \sigma_N^*) \in \prod_{i\in[N]} \mathcal{P}(A_i)$ is a Nash equilibrium if for every player $i \in [N]$,*

$$\mathbb{E}_{a\sim\sigma^*}[u_i(a)] \geq \mathbb{E}_{a\sim(\sigma_i,\sigma_{-i}^*)}[u_i(a)] \quad \text{for all } \sigma_i \in \mathcal{P}(A_i),$$

where the expectation is taken over the joint distribution induced by the mixed strategies.

The classical result of Nash [115] guarantees the existence of at least one mixed strategy Nash equilibrium in any static game with a finite number of players and finite action sets. However, computing and analyzing Nash equilibria can be challenging, especially with high-dimensional action spaces. In particular, Daskalakis et al. [46] shows that even for two-player games, the problem of finding a Nash equilibrium is PPAD-complete (short for Polynomial Parity Argument on Directed graphs), a complexity class of problems whose solutions are guaranteed to exist but are hard to find. This suggests that, without further structure, no polynomial-time algorithm is known for computing equilibria.

1.1.1 Static Potential Games

The computational complexity barriers in finding NE motivate the study of structured classes of games, where equilibria are not only guaranteed to exist but can also be efficiently learned via decentralized dynamics.

A prominent example of a tractable class of N-player games is the class of potential games, introduced by Monderer and Shapley [113]. The key feature of potential games is that any unilateral deviation in a player's payoff aligns exactly with the change in a common scalar function, known as the potential function.

Definition 1.1.3. *A static game is called a static potential game if there exists a function $\phi : A \to \mathbb{R}$ such that for all players i and for any $a_i, a_i' \in A_i, a_{-i} \in A_{-i}$, we have*

$$u_i(a_i', a_{-i}) - u_i(a_i, a_{-i}) = \phi(a_i', a_{-i}) - \phi(a_i, a_{-i}).$$

The function ϕ is referred to as the (static) potential function.

Characterization of the potential function. Under sufficient regularity assumptions, i.e., when each payoff function u_i is twice continuously differentiable, $u_i \in \mathcal{C}^2$, and each action set $A_i \subseteq \mathbb{R}$, the verification of whether a game is a potential game, as well as the construction of a corresponding potential function if one exists, can be formulated in terms of first- and second-order derivatives.

Theorem 1.1.1. *[113, Theorem 4.5] Let $\mathcal{G}$ be a game in which $u_i \in \mathcal{C}^2$ and $A_i \subseteq \mathbb{R}$. Then $\mathcal{G}$ is a potential game if and only if*

$$\frac{\partial^2 u_i}{\partial a_i \partial a_j} = \frac{\partial^2 u_j}{\partial a_i \partial a_j}, \quad \text{for every } i, j \in [N]. \tag{1.1}$$

Moreover, if the payoff functions satisfy (1.1), let $z = (z_1, \cdots, z_N)$ be an arbitrary (but fixed) action profile in $\prod_{i\in[N]} A_i$, and let $p_i : [0,1] \times A_i \mapsto A_i$ be a continuously differentiable reparameterization of A_i such that for all $a_i \in \mathcal{A}_i$, $p_i(0, a_i) = z_i$ and $p_i(1, a_i) = a_i$. Then a potential function for $\mathcal{G}$ is given by

$$\phi(a) = \int_0^1 \sum_{i=1}^N (\partial_{a_i} u_i)(p(r,a)) \partial_r p_i(r, a_i) \mathrm{d}r,$$

where $p(r,a) \coloneqq (p_i(r,a_i))_{i\in[N]}$.

Definitions 1.1.3 and 1.1.1 immediately imply that any local (player-wise) maximizer (defined below) of ϕ is a pure strategy NE:

Proposition 1.1.1. *Let $\mathcal{G}$ be a potential game with potential function ϕ. If ϕ has a local (player-wise) maximum at a^*, namely*

$$\phi(a^*) \geq \phi(a_i, a_{-i}^*), \ \text{for any } a_i \in A_i, i \in [N],$$

then a^* is an NE to $\mathcal{G}$.

As a result, it reduces the challenging task of finding NE to maximizing a single function. The structure of potential games ensures that various learning processes, such as best response and fictitious play, are guaranteed to converge to NE [113, 112, 57, 124, 157].

1.1.2 Near-Potential Games

A key appeal of potential games is that many adaptive dynamics converge to a Nash equilibrium. This raises the question of whether such convergence extends to games that are close to potential games. Candogan et al. [27] formalize the notion of near-potential games and analyze the convergence behavior of various learning dynamics.

To quantify the distance between two static games $\hat{\mathcal{G}}$ and $\mathcal{G}$, Candogan et al. [27] proposes the following measure of "closeness" between the games:

Definition 1.1.4 (Maximum pairwise difference)**.** *Let $\mathcal{G}$ and $\hat{\mathcal{G}}$ be two games with N players, set of action profiles A, and collections of payoff functions $\{u_i\}_{i\in[N]}$ and $\{\hat{u}_i\}_{i\in[N]}$ respectively. The maximum pairwise difference (MPD) between these games is defined as*

$$\mathrm{d}(\mathcal{G},\hat{\mathcal{G}}) := \max_{\substack{a_i,a_i'\in A_i,\\ a_{-i}\in A_{-i}, i\in[N]}} \left|\left(u_i\left(a_i',a_{-i}\right)-u_i\left(a_i,a_{-i}\right)\right)-\left(\hat{u}_i\left(a_i',a_{-i}\right)-\hat{u}_i\left(a_i,a_{-i}\right)\right)\right|.$$

We now formulate the problem of finding the closest potential game to a given game in terms of the maximum pairwise difference defined in Definition 1.1.4. Suppose we are given a game $\mathcal{G}$ with payoff functions $u_{ii\in[N]}$. We seek a potential game $\hat{\mathcal{G}}$ with payoff functions $\hat{u}ii \in [N]$ and a potential function ϕ, such that the MPD from the original game is minimized. This leads to the following optimization formulation:

$$\begin{aligned} \min_{\phi,\{\hat{u}_i\}_{i\in[N]}} \quad \max_{\substack{a_i,a_i'\in A_i,\\ a_{-i}\in A_{-i}, i\in[N]}} \quad & |(u_i(a_i',a_{-i})-u_i(a_i,a_{-i}))-(\hat{u}_i(a_i',a_{-i})-\hat{u}_i(a_i,a_{-i}))| \\ \text{subject to} \quad & \phi(a_i',a_{-i})-\phi(a_i,a_{-i})=\hat{u}_i(a_i',a_{-i})-\hat{u}_i(a_i,a_{-i}), \\ & \text{for all } i\in[N],\ a_i,a_i'\in A_i,\ a_{-i}\in A_{-i}. \end{aligned} \tag{1.2}$$

The optimization problem mentioned above is convex with linear constraints and can be reformulated as a linear program. Therefore, when the action space is finite, the closest potential game can be computed efficiently in polynomial time.

The near-potential games preserve the key properties of potential games, where efficient algorithms are shown to converge to the NE. Candogan et al. [25, 26] show that several learning dynamics—such as best response, logit dynamics, and fictitious play—continue to exhibit desirable convergence behavior in near-potential games. For example, for a game $\mathcal{G}$ such that $\mathrm{d}(\mathcal{G},\hat{\mathcal{G}}) \leq \delta$ with $\hat{\mathcal{G}}$ being a potential game, the trajectory of action profiles using the best response converges to a bounded neighborhood of a Nash equilibrium, where the size of the neighborhood depends on δ continuously.

1.2 Dynamic Games

This section reviews the setup and key models in dynamic games, where players' instantaneous payoffs depend on an underlying state transition process. We consider both discrete-time and continuous-time state dynamics, with a focus on structured game classes that exhibit potential-like properties. These properties, analogous to static potential games but extended to dynamic settings, provide a foundation for our development of α-potential games, a central contribution of this thesis.

Open-loop, closed-loop controls, and Markovian policies. We consider a stochastic dynamic system defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, equipped with a filtration $(\mathcal{F}_t)_{t\geq 0}$ that satisfies the usual conditions. This framework accommodates both discrete-time Markov games, modeled via Markov decision processes (MDPs), and continuous-time stochastic differential games, formulated through stochastic differential equations (SDEs).

Depending on the structure of available information, players may adopt different types of control strategies. Let $\mathcal{F}_t^W := \sigma(W_s^j, s \leq t, j \in [N])$ denote the filtration generated by the exogenous noise (such as Brownian motions in SDE), and let $\mathcal{F}_t^X := \sigma(X_s^j, s \leq t, j \in [N])$ be the filtration generated by the state process. A strategy is said to be *open-loop* if the control process is adapted to $\mathcal{F}_t^W$, i.e., the agent's decision at time t can depend on the initial condition and the noise history. In contrast, a strategy is *closed-loop* if it is adapted to $\mathcal{F}_t^X$, meaning that decisions can depend on the observed state trajectory [153]. A particularly important subclass of closed-loop controls is *feedback control*, in which the decision at time t depends only on the current state X_t, typically taking the form $\alpha_i(t) = \phi_i(t, X_t)$, where ϕ_i is a measurable function [55]. Such a mapping ϕ_i is usually called a *feedback policy*.

In discrete-time settings such as Markov decision processes (MDPs) and Markov games, *Markovian policies* constitute one of the most important classes of policies in the discrete-time RL literature, due to their tractability and practical implementability. A Markovian policy for player i is a mapping $\pi_i : \mathcal{S} \to \mathcal{P}(A_i)$, where the action distribution at time k depends only on the current state s^k, rather than on the full history. That is, for all k, $\pi_i^k(a_i \mid s^0, \ldots, s^k) = \pi_i(a_i \mid s^k)$. The analogue of Markovian policies in the continuous-time RL literature can be found in a recent line of work on *relaxed control*, which is used to model exploration and exploitation in continuous-time reinforcement learning [148, 89, 90, 142, 69].

For an optimal control problem governed by an Itô-type SDE, it is well established that, under appropriate regularity conditions, open-loop and closed-loop controls are equivalent in the sense that they yield the same optimal value function, which coincides with the unique viscosity solution to the associated Hamilton–Jacobi–Bellman (HJB) equation [153, 55, 154]. However, this equivalence may fail in general path-dependent control problems [153] or in game-theoretic settings; see, for example, Carmona et al. [34] for an N-player systemic risk problem and Sun and Yong [132] for a two-player zero-sum linear-quadratic game. In an N-player game, a closed-loop control of the form $\phi_i(t, X_t^1, \ldots, X_t^N)$ introduces interdependence between the players: if player j adjusts her strategy ϕ_j, this affects her state dynamics X_t^j, which in turn influences player i's control through its dependence on the full state vector. In contrast, under open-loop control, where strategies are adapted to the underlying Brownian motions, such interdependence is absent. Therefore, in the context of N-player games, the open-loop and closed-loop control strategies will lead to different equilibrium characterizations. In Chapter 3 of this thesis, we investigate and quantify the impact of control structures on the game dynamics and the equilibrium outcomes.

1.2.1 Discrete-Time Markov Games

For discrete-time dynamic games, we focus on Markov games, a foundational class of games where both state transitions and players' payoffs depend only on the current state and joint actions, rather than the full history of play. Markov games serve as the primary framework for multi-agent reinforcement learning (MARL) algorithms. Methods such as Nash Q-learning [85] and policy gradient approaches [104] leverage the Markov structure to enable scalable learning in dynamic environments.

We consider a Markov game defined by the tuple $\mathcal{G} = \langle N, S, (A_i)_{i\in[N]}, (u_i)_{i\in[N]}, P, \gamma\rangle$, where N is the number of players, S is a finite state space, A_i is the finite action set of player i, and $u_i : S \times A \to \mathbb{R}$ is the one-stage payoff function. The state transition kernel $P(s' \mid s, a)$ determines the probability of moving from state s to s' under joint action a, and $\gamma \in [0, 1)$ is the discount factor.

At each time step k, the system is in state $s^k \in S$, each player i selects an action $a_i^k \sim \pi_i(\cdot \mid s^k)$, and the joint action $a^k = (a_i^k)_{i\in[N]}$ determines the next state $s^{k+1} \sim P(\cdot \mid s^k, a^k)$. The players follow stationary Markov policies, with $\pi_i : S \to \mathcal{P}(A_i) \in \Pi_i$, and the joint policy is denoted by $\pi = (\pi_i)_{i\in[N]} \in \Pi = \prod_{i\in[N]} \Pi_i$. Each player i wants to maximize her expected discounted return (value function), which is defined as

$$V_i(s, \pi) := \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k u_i(s^k, a^k) \mid s^0 = s \right],$$

under policy π and initial state s, and $V_i(\mu, \pi) := \mathbb{E}_{s\sim\mu}[V_i(s, \pi)]$ when the initial state is drawn from distribution $\mu \in \mathcal{P}(S)$.

Markov Potential Games. Leonardos et al. [101] introduces a direct extension of static potential games to dynamic settings, called Markov potential game (MPG), by assuming the existence of a potential function that globally characterizes unilateral deviations in agents' value functions.

Definition 1.2.1. *A Markov game is called a Markov potential game if there exists a function* $\Phi : S \times \Pi \to \mathbb{R}$ *such that for any* $s \in S, i \in [N], \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$,

$$V_i(s, \pi_i, \pi_{-i}) - V_i(s, \pi_i', \pi_{-i}) = \Phi(s, \pi_i, \pi_{-i}) - \Phi(s, \pi_i', \pi_{-i}). \tag{1.3}$$

The MPG structure has enabled learning algorithms with convergence guarantees to NE, including policy gradient-based methods [101, 49, 56, 136, 163, 164] and best-response-based methods [106]. However, two main challenges remain: (1) the lack of real-world examples that can be provably shown to be MPGs, and games where each state corresponds to a static potential game may fail to be MPGs (see examples provided in Leonardos et al. [101]); and (2) the difficulty of certifying games as MPGs and constructing potential functions, except in special cases (e.g., state-independent transitions or identical payoffs [114, 101]).

Our solution: Markov α-Potential Games. In Chapter 2, we propose a framework called Markov α-potential games that extends the notion of potential structure. Markov α-potential games allow misalignment between incentive differences in the value functions and those in a common function called the α-potential function. This misalignment is measured through the notion of maximum pairwise distance between a Markov game and a real-valued function defined on $S \times \Pi$:

Definition 1.2.2. *Given any Markov game $\mathcal{G}$ and a function $\Psi : S \times \Pi \to \mathbb{R}$, the maximum pairwise distance $\widehat{\mathbf{d}}$ between Ψ and $\mathcal{G}$ is defined as*

$$\widehat{\mathbf{d}}(\Psi, \mathcal{G}) := \sup_{\substack{s \in S, i \in [N], \\ \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}}} \left| \Psi\left(s, \pi_i', \pi_{-i}\right) - \Psi\left(s, \pi_i, \pi_{-i}\right) - \left(V_i\left(s, \pi_i', \pi_{-i}\right) - V_i\left(s, \pi_i, \pi_{-i}\right)\right) \right|.$$

Let $\mathcal{F}^{\mathcal{G}}$ be a suitable class of bounded uniformly equi-continuous function. The precise definition is deferred to Chapter 2. We then define Markov α-potential games:

Definition 1.2.3 (Markov α-potential games)**.** *A Markov game $\mathcal{G}$ is a Markov α-potential game if*

$$\alpha = \inf_{\Psi \in \mathcal{F}^{\mathcal{G}}} \widehat{\mathbf{d}}(\Psi, \mathcal{G}).$$

Furthermore, any $\Phi \in \mathcal{F}^{\mathcal{G}}$ such that $\widehat{\mathbf{d}}(\Phi, \mathcal{G}) = \alpha$ is called an α-potential function of $\mathcal{G}$.

Next, we present a useful property due to Definition 1.2.3.

Corollary 1.2.1. *Let $\mathcal{G}$ be a Markov α-potential game with α-potential function Φ. Then, for any $s \in S, i \in [N], \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$,*

$$|V_i(s, \pi_i, \pi_{-i}) - V_i(s, \pi_i', \pi_{-i}) - (\Phi(s, \pi_i, \pi_{-i}) - \Phi(s, \pi_i', \pi_{-i}))| \leq \alpha.$$

Corollary 1.2.1 shows that, compared to MPGs, Markov α-potential games allow for a misalignment between V_i and Φ of at most α. In other words, the α-potential function Φ approximately captures the change in player incentives when they unilaterally deviate, with an error bounded by α. For any given game, upon identifying its α-potential function and the corresponding α, we can conclude that any optimizer of Φ induces an α-Nash equilibrium. In Chapter 3, we prove the existence of the α-potential function. Moreover, similar to

Section 1.1.2, the value of α and the corresponding α-potential function can be obtained by formulating the problem as an optimization problem:

$$\begin{aligned}\min_{\substack{y\in\mathbb{R}\\ \phi:S\times A\to\mathbb{R}}} \quad & y \\ \text{s.t.} \quad & \Big|\sum_{s',a'}(d^s(s',a';\pi_i,\pi_{-i}) - d^s(s',a';\pi_i',\pi_{-i}))\cdot(\phi-u_i)(s',a')\Big| \le y, \\ & \forall s\in S,\ \forall i\in[N],\ \forall \pi_i,\pi_i'\in\Pi_i,\ \forall\pi_{-i}\in\Pi_{-i}, \\ & |\phi(s,a)| \le N\max_{i\in[N]}\|u_i\|_\infty, \quad \forall s\in S, a\in A.\end{aligned} \tag{1.4}$$

where $d^s(s',a';\pi)$ is the occupation measure, defined as

$$d^s(s',a';\pi) \coloneqq \pi(a'|s')\mathbb{E}_\pi\left[\sum_{k=0}^\infty \gamma^k \mathbb{1}(s^k=s')\Big|s^0=s\right].$$

Compared to (1.2), which has finitely many constraints and can be reformulated as a linear program, (1.4) features infinitely many constraints defined over the policy space. Therefore, (1.4) is a semi-infinite linear program. Several algorithmic methods have been developed to solve such problems [141, 82].

In Chapter 2, we present two algorithms, one based on the policy-gradient method and the other on the best-response method, to find approximate Nash equilibria for Markov α-potential games. The resulting Nash regret explicitly depends on α.

1.2.2 Continuous-Time Stochastic Differential Games

Continuous-time games model systems where strategies and states evolve continuously. These games are especially relevant in control-theoretic contexts such as differential games, and have a wide application in financial modeling [77, 69, 70], trajectory planning [134, 135], and autonomous systems [104, 103].

Consider an N-player stochastic game $\mathcal{G}$ over the time horizon $[0,T]$, defined on a complete probability space $(\Omega,\mathcal{F},\mathbb{P})$ with an m-dimensional Brownian motion $W=(W^k)_{k=1}^m$. For each player $i\in[N]$, let $\mathcal{A}_i$ be the set of admissible controls u_i taking values in $A_i\subseteq\mathbb{R}^n$. Denote the joint control by $\boldsymbol{u}=(u_1,\ldots,u_N)\in\mathcal{A}^{(N)}=\prod_{i\in[N]}\mathcal{A}_i$, and let $\mathbf{X}^{\boldsymbol{u}}=(X_i^{\boldsymbol{u}})_{i=1}^N$ be the associated state process satisfying, for all $i\in[N]$,

$$\mathrm{d}X_{t,i} = b_i(t,\mathbf{X}_t,\boldsymbol{u}_t)\mathrm{d}t + \sigma_i(t,\mathbf{X}_t,\boldsymbol{u}_t)\mathrm{d}W_t, \quad X_{0,i}=x_i,$$

where $x_i\in\mathbb{R}^d$, $b_i:[0,T]\times\mathbb{R}^{Nd}\times\mathbb{R}^{Nn}\to\mathbb{R}^d$, and $\sigma_i:[0,T]\times\mathbb{R}^{Nd}\times\mathbb{R}^{Nn}\to\mathbb{R}^{d\times m}$ are measurable. The objective of player i is to minimize the cost functional

$$V_i(\boldsymbol{u}) = \mathbb{E}\left[\int_0^T f_i(t,\mathbf{X}_t^{\boldsymbol{u}},\boldsymbol{u}_t)\mathrm{d}t + g_i(\mathbf{X}_T^{\boldsymbol{u}})\right],$$

where $f_i:[0,T]\times\mathbb{R}^{Nd}\times\mathbb{R}^{Nn}\to\mathbb{R}$ and $g_i:\mathbb{R}^{Nd}\to\mathbb{R}$ are measurable.

Tractable stochastic differential games. One classical approach is through *mean-field games* (MFG), which simplify large-population stochastic differential games by assuming homogeneity among agents and taking the limit as the number of players $N \to \infty$. This framework reduces the problem to a representative agent coupled with a distributional flow, typically characterized by a system of Hamilton–Jacobi–Bellman and Fokker–Planck equations [100, 87]. Another line of work studies *graphon games*, which model heterogeneous interactions between agents through weighted graphs or graphons. These methods analyze the limit behavior of games with complex network structures as $N \to \infty$ [9, 23, 15, 52, 36]. For some more recent works with sparse graph and finite players, we refer to [96, 86].

α-potential games. Guo and Zhang [66] study potential games within the framework of stochastic differential games, focusing on closed-loop controls in feedback form. They provide two characterizations of continuous-time potential games: a probabilistic characterization and a PDE characterization. Moreover, Guo and Zhang [66] establish an analogous characterization of the potential function Φ, comparable to the static potential function characterization in Theorem 1.1.1, by employing linear derivatives:

Definition 1.2.4. *Let $\mathcal{A}^{(N)} = \prod_{i \in [N]} \mathcal{A}_i$ be a convex set and $f : \mathcal{A}^{(N)} \to \mathbb{R}$. For each $i \in [N]$, we say f has a linear derivative with respect to $\mathcal{A}_i$, if there exists $\frac{\delta f}{\delta a_i} : \mathcal{A}^{(N)} \times \operatorname{span}(\mathcal{A}_i) \to \mathbb{R}$, such that for all $\boldsymbol{a} = (a_i, a_{-i}) \in \mathcal{A}^{(N)}$, $\frac{\delta f}{\delta a_i}(\boldsymbol{a}; \cdot)$ is linear and*

$$\lim_{\varepsilon \searrow 0} \frac{f\left(\left(a_i + \varepsilon\left(a_i' - a_i\right), a_{-i}\right)\right) - f(\boldsymbol{a})}{\varepsilon} = \frac{\delta f}{\delta a_i}\left(\boldsymbol{a}; a_i' - a_i\right), \quad \forall a_i' \in \mathcal{A}_i.$$

Linear differentiability, as defined in Definition 1.2.4, is weaker than Fréchet or Gâteaux differentiability because it avoids imposing a topology on the strategy classes $\mathcal{A}_i$. This provides greater flexibility in handling different types of control classes.

In Chapter 3, we extend Guo and Zhang [66] to propose α-potential games in stochastic differential game settings.

Definition 1.2.5 (Guo et al. [74])**.** *We call a game an α-potential game, if there exists $\alpha \geq 0$ and $\Phi : \mathcal{A}^{(N)} \to \mathbb{R}$ such that for all $i \in [N]$, $a_i, a_i' \in \mathcal{A}_i$ and $a_{-i} \in \mathcal{A}_{-i}^{(N)}$,*

$$\left|V_i\left(\left(a_i', a_{-i}\right)\right) - V_i\left(\left(a_i, a_{-i}\right)\right) - \left(\Phi\left(\left(a_i', a_{-i}\right)\right) - \Phi\left(\left(a_i, a_{-i}\right)\right)\right)\right| \leq \alpha,$$

with $\mathcal{A}^{(N)} = \prod_{i \in [N]} \mathcal{A}_i$ the set of strategy profiles for all players, and $\mathcal{A}_{-i}^{(N)} = \prod_{j \in [N] \setminus \{i\}} \mathcal{A}_j$ the set of strategy profiles of all players except player i.

Such Φ is called an α-potential function for the game $\mathcal{G}$. In the case of $\alpha = 0$, we simply call the game $\mathcal{G}$ a potential game and Φ a potential function for $\mathcal{G}$.

Theorem 1.2.1. *If the objective functions $\{V_i\}_{i\in[N]}$ of a game $\mathcal{G}$ admit second-order linear derivatives, then under some mild regularity conditions, for any fixed $\boldsymbol{z} \in \mathcal{A}^{(N)}$, the function*

$$\Phi(\boldsymbol{a}) := \int_0^1 \sum_{j=1}^{N} \frac{\delta V_j}{\delta a_j} \left(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_j - z_j\right) \mathrm{d}r$$

is an α-potential function of $\mathcal{G}$, with

$$\alpha \leq 2 \sup_{i\in[N], a_i' \in \mathcal{A}_i, \boldsymbol{a}, \boldsymbol{a}'' \in \mathcal{A}^{(N)}} \sum_{j=1}^{N} \left| \frac{\delta^2 V_i}{\delta a_i \delta a_j} \left(\boldsymbol{a}; a_i', a_j''\right) - \frac{\delta^2 V_j}{\delta a_j \delta a_i} \left(\boldsymbol{a}; a_j'', a_i'\right) \right|.$$

This characterization generalizes existing results of potential games with finite-dimensional strategy classes [113, 101, 84] to general dynamic games with arbitrary convex strategy classes. In particular, it replaces the Fréchet derivatives used in earlier works with linear derivatives, without requiring a topological structure on $\mathcal{A}^{(N)}$. Moreover, it quantifies the performance of the α-potential function (3.2) in terms of the difference between the second-order linear derivatives of the objective functions.

We characterize games under both open-loop and closed-loop control structures and identify the corresponding value of α. Notably, due to the interdependence introduced by closed-loop control, the α values for open-loop games are usually smaller than those for the corresponding closed-loop games. We illustrate this in Chapter 3 through several examples, including network games, distributed games, and games with mean-field interactions.

With the α-potential function defined in Theorem 1.2.1, we can find an NE by minimizing the α-potential function, $\boldsymbol{a}^* = \inf_{\boldsymbol{a}} \Phi(\boldsymbol{a})$, which can be written as an equivalent conditional McKean-Vlasov problem. In an open-loop linear-quadratic setting, it can be solved explicitly by solving a set of ODEs.

Learning in α-potential games. In Chapter 2, we present two RL algorithms to learn the NE in discrete-time Markov α-potential games, demonstrating their effectiveness through model-free numerical examples. Learning in continuous-time α-potential games is more sophisticated than in discrete-time games, and designing efficient RL algorithms for continuous-time games remains an open and challenging problem.

Since α-potential games reduce the task of finding a Nash equilibrium to solving a minimization problem, a natural starting point is to draw inspiration from single-agent continuous RL methods. As a preliminary step, we consider a linear-quadratic control problem in Chapter 4, and study the convergence rate of policy gradient based and policy optimization based methods in this setting. Several continuous-time RL methods have been proposed in recent literature, including [148, 89, 90]. However, it remains an open and interesting question how to effectively integrate RL methods to establish efficient algorithms with convergence guarantees for continuous-time α-potential games.

Chapter 2

Markov α-Potential Games

2.1 Introduction

Designing non-cooperative multi-agent systems interacting within a shared dynamic environment is a central challenge in many existing and emerging autonomy applications, including autonomous driving, smart grid management, and e-commerce. Markov game, proposed in [129], provide a mathematical framework for studying such interactions [162]. A primary objective in these systems is for agents to reach a *Nash equilibrium*, where no agent benefits from changing its strategy unilaterally. However, designing algorithms for approximating or computing Nash equilibrium is generally intractable [117], unless certain structure of underlying multi-agent interactions are exploited. There is a rich line of literature on equilibrium computation and approximation algorithms for Nash equilibrium in Markov zero-sum games (see [125] and references therein), Markov team games (see [14] and references therein), symmetric Markov games (see [156]), and in particular, Markov potential games (see [106, 163, 101, 114] and references therein) and its generalization to weakly acyclic games (see [6, 155] and references therein).

In this paper, we propose the *Markov α-potential game* framework, where changes in an agent's long-run utility from unilateral policy deviations are captured by an "*α-potential function*" and a parameter α (Definition 2.2.4). We establish that any finite-state, finite-action Markov game is a Markov α-potential game for some $\alpha \geq 0$, and there exists an α-potential function (Theorem 2.2.1). Furthermore, we show that any optimizer of an α-potential function, if it exists, is an α-stationary Nash equilibrium (Proposition 2.2.1).

Markov α-potential games generalize the framework of Markov potential games (MPGs). MPGs, originally proposed in [105] and [101], correspond to the special case of $\alpha = 0$ and extend a rich body of literature on static potential games (or static congestion games) [113]. The MPG structure has enabled learning algorithms with convergence guarantees to Nash equilibrium (e.g., [106, 49]). However, two main challenges remain: (1) the lack of real-world

[1] This chapter is mainly based on work [71] entitled *Markov α-Potential Games*, coauthored with Xin Guo, Chinmay Maheshwari, Shankar Sastry, and Manxi Wu from UC Berkeley.

examples that can be provably shown to be MPGs, and (2) the difficulty of certifying games as MPGs and constructing potential functions, except in special cases (e.g., state-independent transitions or identical payoffs [114, 101]). Our α-potential game framework addresses both challenges: it shows that any finite-state, finite-action Markov game is a Markov α-potential game and provides a semi-infinite linear programming approach to certify MPGs (Section 2.4).

Our Markov α-potential games framework extends the static near-potential games, proposed in [27, 24], to Markov games. Unlike static games, where the nearest potential function always exists, the existence of an α-potential function requires additional analysis (Theorem 2.2.1). Moreover, while finding the nearest static potential function involves finite-dimensional linear programming, computing the α and its potential function requires solving a semi-infinite linear programming problem, as the α-potential function spans both state and policy spaces, the latter being uncountable. We derive explicit upper bounds on the parameter α for two classes of relevant games. First, we consider *Markov congestion games (MCGs)*, where each stage game is a congestion game (proposed in [121]) and the state transition depends on agents' aggregate resource utilization. This is equivalent to Markov games where each stage is a static potential game, as static congestion games and static potential games are equivalent [113]. This class of games models applications like dynamic routing, communication networks, and robotic interactions [45, 134, 135, 94]. We show that the upper bound on α for MCGs scales linearly with the state and resource set sizes, and inversely with the number of agents (Proposition 2.3.2). Second, we consider *perturbed Markov team games (PMTGs)*, which generalize Markov team games by allowing utility deviations from the team objective. We provide an upper bound for PMTGs that scales with the magnitude of these deviations (Proposition 2.3.3). For both MCGs and PMTGs, we calculate an upper bound on α by using a *specific* candidate α-potential function to compute an analytical upper bound on α. However, this upper bound can be loose. In such cases, the semi-infinite linear programming method described in Section 2.4 can be used to obtain tighter numerical estimates of α.

We propose two algorithms to approximate stationary Nash equilibrium in Markov α-potential games. We study the Nash-regret of both algorithms and characterize its dependence on α (Theorems 2.5.1 and 2.5.2). First, we analyze the *projected gradient-ascent algorithm* (Algorithm 2), originally proposed in [49] for MPGs, in the context of Markov α-potential games by bounding the path length of policy updates using changes in the α-potential function and α. Following our proof technique, the analysis of many existing algorithms for MPGs can be extended similarly to Markov α-potential games. Second, we propose a *new algorithm* called the *sequential maximum improvement algorithm* (Algorithm 3) and derive its Nash-regret. The main technical novelty in the analysis is to bound the maximum improvement of a "smoothed" Q-functions with respect to change in policies (aka "path length of policies"), which in turn is bounded by cumulative change in α-potential function (Lemma 2.5.5). For $\alpha = 0$, this algorithm and its analysis are independently relevant to MPGs. We numerically validate these algorithms on examples of MCGs and PMTGs.

2.1.1 Additional Related Works

Our work on Markov α-potential games is related to the literature on weakly acyclic Markov games, proposed in [6]. Weakly acyclic Markov games extend weakly acyclic static games to Markov games, encompassing MPGs as a special case. Unlike MPGs, weakly acyclic Markov games do not require the existence of an exact potential function, instead retain many key properties of potential games, such as the existence of pure equilibria and finite strict best-response paths. Just as MPGs, most games are not weakly acyclic, and determining whether a game is weakly acyclic remains an open problem. On one hand, the introduction of a Markov α-potential games allows for design and analysis of algorithms as a game diverges from a MPG. On the other hand, if a game is weakly acyclic, it is an α-potential game with the value of α not necessarily zero. Exploring the connection between these two approaches and how they might be used together to analyze general Markov games is an interesting and open direction for future research.

Our Algorithm 2 for Markov α-potential games is connected with a substantial body of work on learning approximate Nash equilibria (NEs) in MPGs (see [106, 49, 108, 131, 56, 136, 165]). The first global convergence result for the policy gradient method in MPGs was established in [101]. Additionally, these algorithms have been studied in both discounted infinite horizon settings [49, 56] and finite horizon episodic settings [108, 131]. Other methods, such as natural policy gradient [56, 136, 165] and best-response based methods [106], have also been explored.

Our Algorithm 3 is reminiscent of the "Nash-CA" algorithm developed for MPGs in [131], which requires each player to sequentially compute the best response policy using an RL algorithm in each iteration; in contrast, our algorithm only computes a smoothed *one-step* optimal deviation. One-step optimal deviation based algorithms has also been studied for MPGs [106, 38]. Additionally, incorporating smoothness for better performance in Markov games is also studied in [39, 51, 111].

Finally, a recent work [45] introduces an approximation algorithm for MCGs and investigates the Nash-regret. Their results and approach are tailored exclusively for congestion games, whereas our work focuses on a broader framework of Markov α-potential games.

2.1.2 Notations

For any $n \in \mathbb{N}$, $[n] := \{1, 2, 3, ..., n\}$. For a finite set X, $\mathcal{P}(X)$ denotes the set of probability distributions over X. For any function $f : X \to \mathbb{R}$, the L_∞-norm is defined by $\|f\|_\infty = \max_{x \in X} |f(x)|$, the L_1-norm is $\|f\|_1 = \sum_{x \in X} |f(x)|$, and the L_2-norm is $\|f\| = \sqrt{\sum_{x \in X} |f(x)|^2}$.

2.2 Framework of Markov α-Potential Games

2.2.1 Setup: Markov Games

Consider a Markov game $\mathcal{G}$ as characterized by the tuple $\langle N, S, (A_i)_{i\in[N]}, (u_i)_{i\in[N]}, P, \gamma\rangle$, where N is the number of players, S is the finite set of states, A_i is the finite set of actions of player $i \in [N]$ and $A := \times_{i\in[N]} A_i$ is the set of joint actions of all players, $u_i : S \times A \to \mathbb{R}$ is the one-stage payoff function of player $i \in [N]$, $P = (P(s'|s,a))_{s,s'\in S, a\in A}$ is the probability transition kernel such that $P(s'|s,a)$ is the probability of transitioning to state $s' \in S$ given the current state $s \in S$ and action profile $a \in A$, and $\gamma \in [0,1)$ is the discount factor.

The game proceeds in discrete time steps. At each time step $k = 0, 1, 2, \cdots$, the state of the game is $s^k \in S$, the action taken by player $i \in [N]$ is $a_i^k \in A_i$, and the joint action of all players is $a^k = (a_i^k)_{i\in[N]} \in A$. Once players select their actions, each player $i \in [N]$ observes her one-stage payoff $u_i(s^k, a^k) \in \mathbb{R}$, and the system transits to state s^{k+1}, where $s^{k+1} \sim P(\cdot|s^k, a^k)$. In this study, we assume that the action taken by any player is based on a randomized stationary Markov policy, as in the Markov games literature [47, 101, 49]. That is, for any player $i \in [N]$, the action selected at time step k is $a_i^k \sim \pi_i(\cdot|s^k)$, and the joint policy of all players is $\pi = (\pi_i)_{i\in[N]} \in \Pi := \times_{i\in[N]} \Pi_i$, with $\Pi_i := \{\pi_i : S \to \mathcal{P}(A_i)\}$. The joint policy of all players except player i is denoted as $\pi_{-i} = (\pi_j)_{j\in[N]\setminus\{i\}} \in \Pi_{-i} := \times_{j\in[N]\setminus\{i\}} \Pi_j$. Given $\pi \in \Pi$, the probability of the system transiting from s to s' is denoted as $P^\pi(s'|s) := \mathbb{E}_{a\sim\pi}[P(s'|s,a)]$.

Each player i aims to maximize the accumulated reward (a.k.a. the *utility function*), given the initial state $s \in S$ and the joint policy $\pi \in \Pi$,

$$V_i(s,\pi) := \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k u_i\left(s^k, a^k\right) \mid s^0 = s \right], \tag{2.1}$$

where $\gamma \in [0,1)$ is the discount factor, $a^k \sim \pi\left(\cdot|s^k\right)$, and $s^{k+1} \sim P\left(\cdot|s^k, a^k\right)$. Denote also $V_i(\mu,\pi) := \mathbb{E}_{s\sim\mu}[V_i(s,\pi)]$, if the initial state follows a distribution $\mu \in \mathcal{P}(S)$. Additionally, define the discounted state visitation distribution as $d_\mu^\pi(s) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^k P(s^k = s | s^0 \sim \mu)$. To analyze this game, we adopt the solution concept of ϵ-stationary Nash equilibrium (NE).

Definition 2.2.1. *(ϵ-stationary Nash equilibrium). For any $\epsilon \geq 0$, a policy profile $\pi^* = (\pi_i^*, \pi_{-i}^*)$ is an* ϵ-stationary Nash equilibrium *of the Markov game $\mathcal{G}$ if for any $i \in [N]$, any $\pi_i \in \Pi_i$, and any $\mu \in \mathcal{P}(S)$, $V_i(\mu, \pi_i^*, \pi_{-i}^*) \geq V_i(\mu, \pi_i, \pi_{-i}^*) - \epsilon$.*

When $\epsilon = 0$, it is simply called a *stationary NE*, which always exists in our setup [57].

2.2.2 Markov α-Potential Games

In this section, we introduce the framework of Markov α-potential games. We show that any Markov game can be analyzed under this framework. First, we introduce some preliminaries.

We define a metric $\mathbf{d}$ on Π as follows: for any $\pi, \tilde{\pi} \in \Pi$,

$$\begin{aligned} \mathbf{d}_i\left(\pi_i, \tilde{\pi}_i\right) &:= \max_{s \in S, a_i \in A_i}\left|\pi_i\left(a_i \mid s\right)-\tilde{\pi}_i\left(a_i \mid s\right)\right|, \quad \forall i \in [N], \\ \mathbf{d}(\pi, \tilde{\pi}) &:= \max_{i \in [N]} \mathbf{d}_i\left(\pi_i, \tilde{\pi}_i\right). \end{aligned} \tag{2.2}$$

Evidently, the sets of policies $\{\Pi_i\}_{i\in[N]}$ are compact in the topology induced by the metrics $\{\mathbf{d}_i\}_{i\in[N]}$, Π is compact in the topology induced by $\mathbf{d}$, and the utility functions are continuous with respect to π under the metric $\mathbf{d}$ [156]. Next, we introduce the notion of maximum pairwise distance between a Markov game and a real-valued function defined on $S \times \Pi$.

Definition 2.2.2. *(Maximum pairwise distance). Given any Markov game $\mathcal{G}$ and a function $\Psi : S \times \Pi \rightarrow \mathbb{R}$, the* maximum pairwise distance $\widehat{\mathbf{d}}$ *between Ψ and $\mathcal{G}$ is defined as*

$$\widehat{\mathbf{d}}(\Psi, \mathcal{G}) := \sup_{\substack{s \in S, i \in [N], \\ \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}}} \left|\Psi\left(s, \pi_i', \pi_{-i}\right)-\Psi\left(s, \pi_i, \pi_{-i}\right)-\left(V_i\left(s, \pi_i', \pi_{-i}\right)-V_i\left(s, \pi_i, \pi_{-i}\right)\right)\right|.$$

Definition 2.2.2 generalizes the concept of maximum pairwise distance from [27, Definition 2.3], extending it from static games (action profiles) to Markov games, where the distance is measured over policies that map states to action distributions. Next, we introduce the notion of a game elasticity parameter, which is useful for defining Markov α-potential games. Intuitively, this parameter captures the smallest value of the maximum pairwise distance between any function in a set $\mathcal{F}^{\mathcal{G}}$ (to be defined shortly) and $\mathcal{G}$.

Definition 2.2.3. *(Game elasticity parameter). Given any game $\mathcal{G}$, its* game elasticity parameter *α is defined as*

$$\alpha := \inf_{\Psi \in \mathcal{F}^{\mathcal{G}}} \widehat{\mathbf{d}}(\Psi, \mathcal{G}), \tag{2.3}$$

where $\mathcal{F}^{\mathcal{G}} := \{\Psi : S \times \Pi \to \mathbb{R}$ s.t. $\|\Psi\|_\infty \leq \frac{2N}{1-\gamma} \max_{i\in[N]} \|u_i\|_\infty\}$ is a class of bounded uniformly equi-continuous function on Π. [1]

Our choice of the specific value of the upper bound on functions in $\mathcal{F}^{\mathcal{G}}$ is useful for the proof of Proposition 4.1.

Clearly $\alpha < \infty$ as one can take $\Psi = 0$ in (2.3) to ensure $\alpha \leq 2\|V_i\|_\infty < \infty$.

Furthermore, the game elasticity parameter depends on variety of game parameters, including the number of players, the action and state sets, the utility function values, the Markov state transition dynamics, and the discount factor.

Next, we define Markov α-potential games.

[1] A set $\mathcal{F}$ of functions $f : S \times \Pi \rightarrow \mathbb{R}$ is called *uniformly equi-continuous on* Π, if there exists $\delta_{\mathcal{F}} : \mathbb{R}_+ \to \mathbb{R}_+$ such that for every $\epsilon > 0$, $|f(s,\pi) - f(s,\pi')| \leq \epsilon$ for all $f \in \mathcal{F}, s \in S$, $\pi, \pi' \in \Pi$ such that $\mathbf{d}(\pi, \pi') \leq \delta_{\mathcal{F}}(\epsilon)$.

Definition 2.2.4. *(Markov α-potential game). A Markov game $\mathcal{G}$ is a* Markov α-potential game *if α is the game elasticity parameter. Furthermore, any $\Phi \in \mathcal{F}^{\mathcal{G}}$ such that $\widehat{\mathbf{d}}(\Phi, \mathcal{G}) = \alpha$ is called an* α-potential function *of $\mathcal{G}$.*

Next, we present a useful property due to Definition 2.2.4.

Corollary 2.2.1. *Let $\mathcal{G}$ be a Markov α-potential game with α-potential function Φ. Then, for any $s \in S, i \in [N], \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$,*

$$|V_i(s, \pi_i, \pi_{-i}) - V_i(s, \pi_i', \pi_{-i}) - (\Phi(s, \pi_i, \pi_{-i}) - \Phi(s, \pi_i', \pi_{-i}))| \leq \alpha. \tag{2.4}$$

Next, we show the existence of an α-potential function.

Theorem 2.2.1. *(Existence of α-potential function). For any Markov game $\mathcal{G}$, there exists $\Phi \in \mathcal{F}^{\mathcal{G}}$ such that $\widehat{\mathbf{d}}(\Phi, \mathcal{G}) = \inf_{\Psi \in \mathcal{F}^{\mathcal{G}}} \widehat{\mathbf{d}}(\Psi, \mathcal{G})$.*

Proof. Define a mapping $\mathcal{F}^{\mathcal{G}} \times \Pi \times \Pi \ni (\Psi, \pi, \pi') \mapsto h(\Psi, \pi, \pi') := \max_{s \in S, i \in [N]} \big| \Psi\left(s, \pi_i', \pi_{-i}\right) - \Psi\left(s, \pi_i, \pi_{-i}\right) - \left(V_i\left(s, \pi_i', \pi_{-i}\right) - V_i\left(s, \pi_i, \pi_{-i}\right)\right) \big| \in \mathbb{R}$. Note that such h is continuous under the standard topology induced by sup-norm on $\mathcal{F}^{\mathcal{G}} \times \Pi \times \Pi$. By Berge's maximum theorem, $g(\Psi) := \max_{\pi, \pi' \in \Pi} h(\Psi, \pi, \pi')$ is continuous with respect to Ψ. Since $\mathcal{F}^{\mathcal{G}}$ is uniformly bounded and uniformly equi-continuous, the Arzelà–Ascoli theorem implies that $\mathcal{F}^{\mathcal{G}}$ is relatively compact in $\mathcal{C}^{\Pi}$, where $\mathcal{C}^{\Pi} := \{f : S \times \Pi \to \mathbb{R} \mid \forall s \in S, f(s, \cdot) \text{ is a continuous function}\}$ [122]. Finally, by the extreme-value theorem [122], there exists a function $\Phi \in \mathcal{F}^{\mathcal{G}}$ such that $\widehat{\mathbf{d}}(\Phi, \mathcal{G}) = \inf_{\Psi \in \mathcal{F}^{\mathcal{G}}} \widehat{\mathbf{d}}(\Psi, \mathcal{G})$. □

Corollary 2.2.1 and Theorem 2.2.1 jointly show that for any Markov game $\mathcal{G}$, an α-potential function exists such that the gap between the change in the utility function of any agent due to a unilateral change in its policy and the change in α-potential function is at most α. Next, we show that any optimizer of the α-potential function with respect to policy π yields an α-Nash equilibrium (NE) of game $\mathcal{G}$.

Proposition 2.2.1. *Given a Markov α-potential game $\mathcal{G}$ with an α-potential function Φ, for any $\epsilon > 0$, if there exists a $\pi^* \in \Pi$ such that for every $s \in S$, $\Phi(s, \pi^*) + \epsilon \geq \sup_{\pi \in \Pi} \Phi(s, \pi)$, then $\pi^* \in \Pi$ is an $(\alpha + \epsilon)$-stationary NE of $\mathcal{G}$.*

Remark 2.2.1. *Note that Proposition 2.2.1 holds for any function $\Psi \in \mathcal{F}^{\mathcal{G}}$ that yields an upper bound for α. That is, given a Markov α-potential game $\mathcal{G}$ and a function Ψ satisfying*

$$\begin{aligned} &|V_i(s, \pi_i, \pi_{-i}) - V_i(s, \pi_i', \pi_{-i}) - (\Psi(s, \pi_i, \pi_{-i}) - \Psi(s, \pi_i', \pi_{-i}))| \leq \bar{\alpha}, \\ &\quad \forall s \in S, \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}, \end{aligned}$$

for some $\bar{\alpha} \in [\alpha, \infty)$, then for any $\pi^* \in \Pi$ such that for every $s \in S$, $\Psi(s, \pi^*) + \epsilon \geq \sup_{\pi \in \Pi} \Psi(s, \pi)$, π^* is an $(\bar{\alpha} + \epsilon)$-stationary NE of $\mathcal{G}$.

2.3 Examples of Markov α-Potential Games

In this section, we present three important classes of games, Markov potential games, Markov congestion games, and perturbed Markov team games, which can be analyzed through the framework of Markov α-potential games.

2.3.1 Markov Potential Game

A game is a Markov potential game if there exists an auxiliary function (a.k.a. potential function) such that when a player unilaterally deviates from her policy, the change of the potential function is equal to the change of her utility function.

Definition 2.3.1 (Markov potential games [101])**.** *A Markov game $\mathcal{G}$ is a* Markov potential game *(MPG) if there exists a potential function $\Phi : S \times \Pi \rightarrow \mathbb{R}$ such that for any $i \in [N]$, $s \in S$, $\pi_i, \pi_i' \in \Pi_i$, and $\pi_{-i} \in \Pi_{-i}$, $\Phi(s, \pi_i', \pi_{-i}) - \Phi(s, \pi_i, \pi_{-i}) = V_i(s, \pi_i', \pi_{-i}) - V_i(s, \pi_i, \pi_{-i})$.*

Proposition 2.3.1. *An MPG is a Markov α-potential game with $\alpha = 0$.*

2.3.2 Markov Congestion Game

The Markov congestion game (MCG) $\mathcal{G}_{\mathsf{mcg}}$ is a dynamic counterpart to the static congestion game introduced by [113], involving a finite number of players using a finite set of resources. Each stage of $\mathcal{G}_{\mathsf{mcg}}$ is a static congestion game with a state-dependent reward function for each resource, and the state transition depends on the aggregated usage of each resource by the players. Specifically, let the finite set of resources in the one-stage congestion game be denoted as E. The action $a_i \in A_i \subseteq 2^E$ of each player $i \in [N]$ represents the set of resources chosen by player i. Here, the action set A_i is the set of all resource combinations that are feasible for player i. The total usage demand of all players is 1, and each player's demand is assumed to be $1/N$.

Given an action profile $a = (a_i)_{i\in[N]}$, the aggregated usage demand of each resource $e \in E$ is given by

$$w_e(a) = \frac{1}{N} \sum_{i\in[N]} \mathbb{1}(e \in a_i). \tag{2.5}$$

In each state s, the reward for using resource e is denoted as $(1/N) \cdot c_e(s, w_e(a))$. Thus, the one-stage payoff for player $i \in [N]$ in state $s \in S$, given the joint action profile $a \in A$, is $u_i(s, a) = (1/N) \cdot \sum_{e\in a_i} c_e(s, w_e(a))$. The state transition probability, denoted as $P(s'|s, w)$, depends on the aggregate usage vector $w = (w_e)_{e\in E}$, which is induced by the players' action profile as in (2.5). The set of all feasible aggregate usage demands is denoted by W.

The next proposition shows that, under a regularity condition on the state transition probability, $\mathcal{G}_{\mathsf{mcg}}$ is a Markov α-potential game such that the upper bound of α scales linearly with respect to the Lipschitz constant ζ, the size of state space $|S|$, resource set $|E|$, and decreases as N increases.

Proposition 2.3.2. *If there exists some $\zeta > 0$ such that for any $s, s' \in S, w, w' \in W$, $|P(s'|s,w) - P(s'|s,w')| \leq \zeta \|w - w'\|_1$, then the congestion game $\mathcal{G}_{\mathsf{mcg}}$ is a Markov α-potential game with $\alpha \leq 2\zeta\gamma|S||E| \sup_{s,\pi} \Psi(s,\pi)/(N(1-\gamma))$, where*

$$\Psi(\mu, \pi) := \frac{1}{N}\mathbb{E}_{\mu,\pi}\left[\sum_{k=0}^{\infty} \gamma^k \Big(\sum_{e \in E} \sum_{j=1}^{w_e^k N} c_e\Big(s^k, \frac{j}{N}\Big)\Big)\right], \tag{2.6}$$

such that $s^0 \sim \mu$, the aggregate usage vector $w^k = (w_e^k)_{e \in E}$ is induced by $a^k \sim \pi(s^k)$, and $s^k \sim P(\cdot|s^{k-1}, w^{k-1})$.

2.3.3 Perturbed Markov Team Game

A Markov game is called a perturbed Markov team game (PMTG) $\mathcal{G}_{\mathsf{pmtg}}$ if the payoff function for each player $i \in [N]$ can be decomposed as $u_i(s,a) = r(s,a) + \xi_i(s,a)$. Here, $r(s,a)$ represents the common interest of the team, and $\xi_i(s,a)$ represents player i's heterogeneous preference, such that $\|\xi_i\|_{L_\infty} \leq \kappa$, where $\kappa \geq 0$ measures each player's deviation from the team's common interest. As $\kappa \to 0$, $\mathcal{G}_{\mathsf{pmtg}}$ becomes a Markov team game, which is an MPG [101].

The next proposition shows that a $\mathcal{G}_{\mathsf{pmtg}}$ is a Markov α-potential game, and the upper bound of α decreases as the magnitude of the payoff perturbation κ decreases.

Proposition 2.3.3. *A perturbed Markov team game $\mathcal{G}_{\mathsf{pmtg}}$ is a Markov α-potential game with $\alpha \leq \frac{2\kappa}{(1-\gamma)^2}$.*

2.4 Finding an Upper Bound of α

The analysis of MCG and PMTG in Section 2.3 utilizes a specific form of the Markov α-potential function to obtain an upper bound on α. In this section, we provide an optimization-based procedure to find an upper bound on α by also computing the α-potential function.

Our approach is based on changing the feasible set of the optimization problem in (2.3) to $\tilde{\mathcal{F}}^{\mathcal{G}}$, defined as follows:

$$\begin{aligned}\tilde{\mathcal{F}}^{\mathcal{G}} := \Bigg\{ \Psi(s,\pi) = \sum_{s' \in S, a' \in A} d^s(s', a'; \pi)\phi(s', a'), \forall s \in S, \pi \in \Pi \,\Bigg|\, \phi : S \times A \to \mathbb{R} \\ \text{s.t. } \|\phi\|_\infty \leq N \max_{i \in [N]} \|u_i\|_\infty \Bigg\},\end{aligned} \tag{2.7}$$

where, for any $s \in S$, $d^s(\cdot; \pi) : S \times A \to \mathbb{R}$ is the *state-action occupancy measure* induced due to π, defined as follows:

$$d^s(s', a'; \pi) := \pi(a'|s')\mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \mathbb{1}(s^k = s') \Big| s^0 = s\right],$$

where $a^k \sim \pi\left(s^k\right)$, and $s^{k+1} \sim P\left(\cdot|s^k, a^k\right)$. Intuitively, for any $\Psi \in \tilde{\mathcal{F}}^{\mathcal{G}}$, there exists $\phi : S \times A \to \mathbb{R}$ such that $\Psi(s, \pi)$ represents the long-horizon discounted value of a Markov decision process with state transition P, starting from state s, using policy $\pi \in \Pi$, and one-step utility ϕ.

Proposition 2.4.1. *For any Markov α-potential game $\mathcal{G}$, $\tilde{\mathcal{F}}^{\mathcal{G}} \subseteq \mathcal{F}^{\mathcal{G}}$. That is, $\bar{\alpha} \geq \alpha$ with*

$$\bar{\alpha} \coloneqq \inf_{\Psi \in \tilde{\mathcal{F}}^{\mathcal{G}}} \widehat{\mathbf{d}}(\Psi, \mathcal{G}). \tag{2.8}$$

Using Remark 2.2.1, we can conclude that any optimizer of Ψ, where $\widehat{\mathbf{d}}(\Psi, \mathcal{G}) = \bar{\alpha}$, can be used to find a $\bar{\alpha}$-stationary NE for the game $\mathcal{G}$.

Next, we provide an optimization based method to compute $\bar{\alpha}$. Note that (2.8) can be reformulated as follows:

$$\begin{aligned}
\min_{\substack{y \in \mathbb{R} \\ \phi: S \times A \to \mathbb{R}}} \quad & y && (2.9)\\
\text{s.t.} \quad & \Big|\sum_{s', a'} (d^s(s', a'; \pi_i, \pi_{-i}) - d^s(s', a'; \pi_i', \pi_{-i})) \cdot (\phi - u_i)(s', a')\Big| \leq y, && \text{(C1)}\\
& \forall s \in S,\ \forall i \in [N],\ \forall \pi_i, \pi_i' \in \Pi_i,\ \forall \pi_{-i} \in \Pi_{-i}, \\
& |\phi(s, a)| \leq N \max_{i \in [N]} \|u_i\|_\infty, \quad \forall s \in S, a \in A.
\end{aligned}$$

Here, we use

$$V_i(s, \pi) = \sum_{s' \in S, a' \in A} d^s(s', a'; \pi) u_i(s', a'), \text{ and } \Psi(s, \pi) = \sum_{s' \in S, a' \in A} d^s(s', a'; \pi) \phi(s', a'),$$

for some $\phi : S \times A \to \mathbb{R}$.

Note that (2.9) is a semi-infinite linear programming where the objective is a linear function with an uncountable number of linear constraints. Particularly, in (C1) there is one linear constraint corresponding to each tuple $(s, i, \pi_i, \pi_i', \pi_{-i})$. Moreover, the coefficients of each linear constraint in (C1) are composed of state-action occupancy measures which are computed by solving a Bellman equation. There are a number of algorithmic approaches to solve semi-infinite linear programming problems [141, 82].

2.4.1 Algorithms to Solve Semi-Infinite Linear Programming

In this section, we present an algorithm based on the stochastic gradient method from [141] to solve the semi-infinite linear programming problem (2.9). Denote $C \coloneqq N \max_{i \in [N]} \|u_i\|_\infty$ and define

$$g(\phi, y; \pi, \pi') = \max\left\{ \max_{i\in I} \Big| \sum_{s',a'} (d^s(s', a'; \pi_i, \pi_{-i}) - d^s(s', a'; \pi'_i, \pi_{-i}))(\phi - u_i)(s', a') \Big| - y, \right.$$
$$\left. \max_{s\in S, a\in A} |\phi(s, a)| - C \right\}, \tag{2.10}$$

which ensures that constraint (C1) in (2.9) can be rewritten as $g(\phi, y; \pi, \pi') \leq 0, \forall \pi, \pi' \in \Pi$. Let $h : \mathbb{R} \to \mathbb{R}$ be a convex differentiable function such that

$$h(x) = 0 \text{ for all } x \leq 0, \text{ and } h(x) > 0 \text{ for all } x > 0.$$

A candidate choice of h is $h(x) = (\max\{0, x\})^2$. Finally, we consider step-size schedules $\{\eta_t\}_{t=1}^{\infty}$ and $\{\beta_t\}_{t=1}^{\infty}$ such that

$$\lim_{t\to\infty} \beta_t = \infty, \sum_{t=1}^{\infty} \eta_t^2 \beta_t^2 < \infty, \sum_{t=1}^{\infty} \eta_t = \infty, \text{ and } \eta_t > 0, \beta_t < \beta_{t+1} \text{ for all } t \geq 0. \tag{2.11}$$

Theorem 4 in [141] shows that with probability 1, $(y^{(t)}, \phi^{(t)})$ almost surely converges to a solution of (2.9).

Algorithm 1 Algorithm to solve (2.9) [141]

```
**Input:** $y^{(0)} \in \mathbb{R}_+, \phi^{(0)} \in \mathbb{R}^{S\times A}$, $\{\eta_t\}_{t=1}^{\infty}$ and $\{\beta_t\}_{t=1}^{\infty}$ satisfying (2.11).
for $t = 0, 1, 2, ..., T-1$ do
    Sample $\pi$, $\pi'$ in $\Pi$ from uniform distribution and calculate $g(\phi^{(t)}, y^{(t)}; \pi, \pi')$ in (2.10).
    Update $\phi^{(t)}$ with
        $\phi^{(t+1)} = \phi^{(t)} - \eta_{t+1}\beta_{t+1} h'\left(g\left(\phi^{(t)}, y^{(t)}; \pi, \pi'\right)\right) \cdot \nabla_\phi g\left(\phi^{(t)}, y^{(t)}; \pi, \pi'\right),$   (2.12)
    and update $y^{(t)}$ with
        $y^{(t+1)} = y^{(t)} - \eta_{t+1}\Big(1 + \beta_{t+1} h'\left(g\left(\phi^{(t)}, y^{(t)}; \pi, \pi'\right)\right) \cdot \nabla_y g\left(\phi^{(t)}, y^{(t)}; \pi, \pi'\right)\Big).$
end for
```

State-wise potential games. Algorithm 1 iteratively updates the variables $y \in \mathbb{R}$ and $\phi \in \mathbb{R}^{S\times A}$. However, this method may be slow as the dimension of ϕ scales with $|S| \cdot |A|$. For MCGs, where each state is a static potential game, one can utilize the game structure to accelerate the convergence of algorithm.

For an MCG $\mathcal{G}_{\mathsf{mcg}}$, there exists a function $\phi^* : S \times A \to \mathbb{R}$ such that for every $i \in [N], s \in S, a_i, a'_i \in A_i, a_{-i} \in A_{-i}$, $|\phi^*(s, a_i, a_{-i}) - \phi^*(s, a'_i, a_{-i}) - (u_i(s, a_i, a_{-i}) - u_i(s, a'_i, a_{-i}))| = 0$.

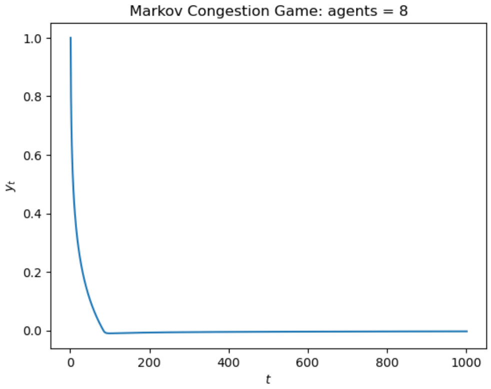


Figure 2.1: Estimating α in a Markov congestion game

Note. The value of α is computed using Algorithm 1 with $\phi^{(0)} = \phi^*$, $\eta_t = \frac{1}{t}, \beta_t = t^{0.4999}$, $\forall t \geq 1$.

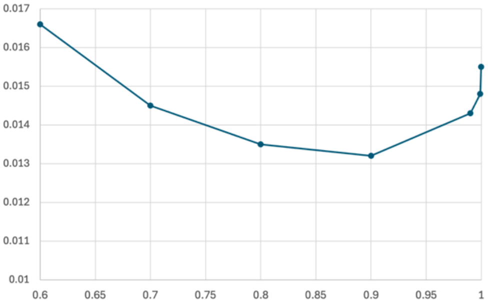


Figure 2.2: Variation of α with the discount factor in the perturbed Markov team game

Note. The number of players is $N = 3$, and perturbation parameter is $\kappa = 0.1$; The setup of this game is same as that in Section 2.6 with $\lambda_1 = \lambda_3 = 0.8, \lambda_2 = \lambda_4 = 0.2$.

Then one can input $\phi^{(0)} = \phi^*$ and omit the update of $\phi^{(t)}$ in (2.12) in Algorithm (1). Figure 2.1 shows the empirical performance of Algorithm 1 for the Markov congestion game. Note that with the setting in Section 2.6, $y^{(t)}$ converges to 0, which suggests that $\mathcal{G}_{\mathsf{mcg}}$ may be an MPG, at least for some model parameters.

Perturbed team games. Figure 2.2 illustrates how α varies with different discount factors γ in a PMTG using Algorithm 1. Note that the growth of the numerical estimate of α is much more benign than the analytical characterization obtained in Proposition 2.3.3.

2.5 Approximation Algorithms and Nash-Regret Analysis

In this section, we present two equilibrium approximation algorithms for Markov α-potential games: the *projected gradient-ascent algorithm*, proposed in [49] for MPGs, and the *sequential maximum improvement algorithm*, where each player's strategy is updated based on a one-stage smoothed best response. We also derive non-asymptotic convergence rates for these algorithms in terms of Nash-regret, defined as Nash-regret$(T) := \frac{1}{T}\sum_{t=1}^{T} \max_{i\in[N]} R_i^{(t)}$, where $R_i^{(t)} := \max_{\pi_i'\in\Pi_i} V_i\left(\mu, \pi_i', \pi_{-i}^{(t)}\right) - V_i\left(\mu, \pi^{(t)}\right)$, and $\pi^{(t)}$ denotes the t-th iterate. Note that Nash-regret is always non-negative; if Nash-regret$(T) \leq \epsilon$ for some $\epsilon > 0$, then there exists $t^\dagger$ such that $\pi^{(t^\dagger)}$ is an ϵ-stationary NE.

2.5.1 Projected Gradient-Ascent Algorithm

First, we define some useful notations. Given a joint policy $\pi \in \Pi$, define player i's *Q-function* as $Q_i^\pi(s, a_i) = \mathbb{E}_{a_{-i}\sim\pi_{-i}(s)}\left[u_i(s, a_i, a_{-i}) + \gamma \sum_{s'\in S} P(s'|s, a_i, a_{-i}) V_i(s', \pi)\right]$, and denote $\mathbf{Q}_i^\pi(s) = (Q_i^\pi(s, a_i))_{a_i\in A_i}$. Let κ_μ denote the maximum distribution mismatch of π relative to μ, and let $\tilde{\kappa}_\mu$ denote the minimax value of the distribution mismatch of π relative to μ. That is,

$$\kappa_\mu := \sup_{\pi\in\Pi} \left\|d_\mu^\pi/\mu\right\|_\infty, \quad \tilde{\kappa}_\mu := \inf_{\nu\in\mathcal{P}(S)} \sup_{\pi\in\Pi} \|d_\mu^\pi/\nu\|_\infty, \tag{2.13}$$

where the division d_μ^π/ν is evaluated in a component-wise manner. The algorithm iterates for T steps. We abuse the notation to use $Q_i^{(t)}$ to denote $Q_i^{\pi^{(t)}}$, and $\mathbf{Q}_i^{(t)}$ to denote $\mathbf{Q}_i^{\pi^{(t)}}$. In every step $t \in [T-1]$, each player $i \in [N]$ updates her policy following a projected gradient-ascent algorithm as in (2.14).

Algorithm 2 Projected Gradient-Ascent Algorithm

```
  Input: Step size $\eta$, for every $i \in [N], a_i \in A_i, s \in S$, set $\pi_i^{(0)}(a_i|s) = 1/|A_i|$.
  for $t = 0, 1, 2, ..., T-1$ do
      For every $i \in [N], s \in S$, update the policies as follows
          $\pi_i^{(t+1)}(s) = \texttt{Proj}_{\Pi_i}\left(\pi_i^{(t)}(s) + \eta\mathbf{Q}_i^{(t)}(s)\right),$    (2.14)
      where $\texttt{Proj}_{\Pi_i}$ denotes the orthogonal projection on $\Pi_i$.
  end for
```

Remark 2.5.1. *Algorithm 1 is not the standard policy gradient algorithm. The standard policy gradient is given by* $\frac{\partial V_i^\pi(\rho)}{\partial \pi_i(a_i|s)} = 1/(1-\gamma)\cdot d_\rho^\pi(s) Q_i^\pi(s, a_i)$ *[101]. The RHS in the this equation scales with the state visitation frequency* $d_\rho^\pi(s)$*, which results in slow learning rate for states with low visitation frequencies under the current policy. To address this issue,* *[49] proposed to remove the term $d_\rho^\pi(s)/(1-\gamma)$ from the standard policy gradient update, which accelerates the learning for states with low visitation probabilities. We adopted the convention of [49] to call it "policy gradient-ascent algorithm".*

Theorem 2.5.1. *Given a Markov α-potential game with an α-potential function Φ and an initial state distribution μ, the policy updates generated from Algorithm 2 satisfies*

(i) Nash-regret$(T) \leq \mathcal{O}\left(\frac{\sqrt{\tilde{\kappa}_\mu \bar{A} N}}{(1-\gamma)^{\frac{9}{4}}}\left(\frac{C_\Phi}{T}+N^2\alpha\right)^{\frac{1}{4}}\right)$ with $\eta = \frac{(1-\gamma)^{2.5}\sqrt{C_\Phi+N^2\alpha T}}{2N\bar{A}\sqrt{T}}$;

(ii) Nash-regret$(T) \leq \mathcal{O}\left(\sqrt{\frac{\min(\kappa_\mu,|S|)^4 N\bar{A}}{(1-\gamma)^6}}\left(\frac{C_\Phi}{T}+N^2\alpha\right)^{\frac{1}{2}}\right)$ with $\eta = \frac{(1-\gamma)^4}{8\min(\kappa_\mu,|S|)^3 N\bar{A}}$,

where $\bar{A} := \max_{i\in[N]}|A_i|$, κ_μ and $\tilde{\kappa}_\mu$ are defined in (2.13)*, and $C_\Phi > 0$ is a constant satisfying $|\Phi(\mu,\pi)-\Phi(\mu,\pi')| \leq C_\Phi$ for any $\pi,\pi' \in \Pi, \mu \in \mathcal{P}(S)$.*

We emphasize that the Nash-regret bounds in Theorem 2.5.1 (also Theorem 2.5.2 in the next section) will hold even without knowing the exact form of Φ and the game elasticity parameter α. It is sufficient to have an upper bound $\bar{\alpha}$ for α and an associated function Ψ for which this upper bound holds. In the special case of $\alpha = 0$, the Nash-regret bound in Theorem 2.5.1 recovers the Nash-regret bound from [49] for MPG.

The proof of Theorem 2.5.1 is inspired by [49] for the Nash-regret analysis of MPGs. First, we state multi-player performance difference lemma (Lemma 2.5.1), which enables bounding the Nash-regret of an algorithm by summing the norms of policy updates, denoted as $\|\pi_i^{(t+1)}-\pi_i^{(t)}\|$. The main modification for our analysis is to bound the sum of these policy update differences by the game elasticity parameter α and the change in the α-potential function Φ (Lemma 2.5.2).

Lemma 2.5.1 (Performance difference (Lemma 1 in [49]))**.** *For any $i \in [N]$, $\mu \in \mathcal{P}(S)$, $\pi_i', \pi_i \in \Pi_i$, and $\pi_{-i} \in \Pi_{-i}$,*

$$V_i(\mu,\pi_i',\pi_{-i}) - V_i(\mu,\pi_i,\pi_{-i}) = \frac{1}{1-\gamma}\sum_{s,a_i} d_\mu^{\pi_i',\pi_{-i}}(s)\cdot\left(\pi_i'(a_i|s)-\pi_i(a_i|s)\right) Q_i^{\pi_i,\pi_{-i}}(s,a_i).$$

Lemma 2.5.2 (Policy improvement)**.** *For Markov α-potential game* (2.3) *with any state distribution $\nu \in \mathcal{P}(S)$, the α-potential function $\Phi(\nu,\pi)$ at two consecutive policies $\pi^{(t+1)}$ and $\pi^{(t)}$ in Algorithm 2 satisfies*

$$\text{(i)}\,\Phi(\nu,\pi^{(t+1)}) - \Phi(\nu,\pi^{(t)}) + N^2\alpha \geq -\frac{4\eta^2\bar{A}^2N^2}{(1-\gamma)^5} + \frac{1}{2\eta(1-\gamma)}\sum_{i\in[N],s\in S} d_\nu^{\pi_i^{(t+1)},\pi_{-i}^{(t)}}(s)\left\|\pi_i^{(t+1)}(s)-\pi_i^{(t)}(s)\right\|^2;$$

$$\text{(ii)}\,\Phi(\nu,\pi^{(t+1)}) - \Phi(\nu,\pi^{(t)}) + N^2\alpha \geq \frac{1}{2\eta(1-\gamma)}\left(1-\frac{4\eta\kappa_\nu^3\bar{A}N}{(1-\gamma)^4}\right)\sum_{i\in[N],s\in S} d_\nu^{\pi_i^{(t+1)},\pi_{-i}^{(t)}(s)}\cdot\left\|\pi_i^{(t+1)}(s)-\pi_i^{(t)}(s)\right\|^2.$$

Proof of Theorem 2.5.1 Using the variational characterization of projection operation in (2.14), we note that for any $\pi_i' \in \Pi_i$,

$$\left\langle \pi_i'(s) - \pi_i^{(t+1)}(s), \eta \mathbf{Q}_i^{(t)}(s) - \pi_i^{(t+1)}(s) + \pi_i^{(t)}(s) \right\rangle_{A_i} \leq 0.$$

Therefore, for any $\pi_i' \in \Pi_i$,

$$\begin{aligned}\left\langle \pi_i'(s) - \pi_i^{(t)}(s), \mathbf{Q}_i^{(t)}(s) \right\rangle_{A_i} &= \left\langle \pi_i'(s) - \pi_i^{(t+1)}(s), \mathbf{Q}_i^{(t)}(s) \right\rangle_{A_i} + \left\langle \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s), \mathbf{Q}_i^{(t)}(s) \right\rangle_{A_i} \\ &\leq \frac{1}{\eta} \left\langle \pi_i'(s) - \pi_i^{(t+1)}(s), \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\rangle_{A_i} \\ &\quad + \left\langle \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s), \mathbf{Q}_i^{(t)}(s) \right\rangle_{A_i}.\end{aligned}$$

Note that for any two probability distributions p_1 and p_2, $\|p_1 - p_2\| \leq \|p_1 - p_2\|_1 \leq 2$. Therefore,

$$\begin{aligned}\left\langle \pi_i'(s) - \pi_i^{(t)}(s), \mathbf{Q}_i^{(t)}(s) \right\rangle_{A_i} &\leq \frac{2}{\eta} \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\| + \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\| \left\| \mathbf{Q}_i^{(t)}(s) \right\| \\ &\leq \frac{3}{\eta} \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|, \qquad (2.15)\end{aligned}$$

where the last inequality is due to $\left\| \mathbf{Q}_i^{(t)}(s) \right\| \leq \frac{\sqrt{A}}{1-\gamma}$ and $\eta \leq \frac{1-\gamma}{\sqrt{A}}$. Hence, by Lemma 2.5.1 and (2.15),

$$\begin{aligned}T \cdot \text{Nash-regret}(T) &= \sum_{t=1}^{T} \max_{i \in [N], \pi_i'} V_i(\mu, \pi_i', \pi_{-i}^{(t)}) - V_i(\mu, \pi^{(t)}) \\ =& \sum_{t=1}^{T} \max_{\pi_i'} \sum_{s, a_i} \frac{d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s)}{1-\gamma} (\pi_i'(a_i|s) - \pi_i^{(t)}(a_i|s)) \mathbf{Q}_i^{(t)}(s, a_i) \\ \leq& \frac{3}{\eta(1-\gamma)} \sum_{t=1}^{T} \sum_{s} d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|,\end{aligned}$$

where in the second line we slightly abuse the notation i to represent $\arg\max_i$ and in the last line we slightly abuse the notation π_i' to represent $\arg\max_{\pi_i'}$. Now, continuing the above calculation with an arbitrary $\nu \in \mathcal{P}(S)$ and using

$$\frac{d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s)}{d_\nu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)} \leq \frac{d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s)}{(1-\gamma)\nu(s)} \leq \frac{\sup_{\pi \in \Pi} \left\| d_\mu^\pi / \nu \right\|_\infty}{1-\gamma}$$

to get:

$$\begin{aligned}
&T \cdot \text{Nash-regret}(T) \\
\leq&\frac{3\sqrt{\sup_{\pi\in\Pi}\left\|d_\mu^\pi/\nu\right\|_\infty}}{\eta(1-\gamma)^{\frac{3}{2}}}\sum_{t=1}^{T}\sum_{s}\sqrt{d_\mu^{\pi_i',\pi_{-i}^{(t)}}(s)d_\nu^{\pi_i^{(t+1)},\pi_{-i}^{(t)}}(s)}\cdot\left\|\pi_i^{(t+1)}(s)-\pi_i^{(t)}(s)\right\| \qquad (2.16)\\
\leq&\frac{3\sqrt{\sup_{\pi\in\Pi}\left\|d_\mu^\pi/\nu\right\|_\infty}}{\eta(1-\gamma)^{\frac{3}{2}}}\sqrt{\sum_{t=1}^{T}\sum_{s}d_\mu^{\pi_i',\pi_{-i}^{(t)}}(s)}\cdot\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{s}d_\nu^{\pi_i^{(t+1)},\pi_{-i}^{(t)}(s)}\left\|\pi_i^{(t+1)}(s)-\pi_i^{(t)}(s)\right\|^2},
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality and replacing i ($\arg\max_i$) by the sum over all players. There are two choices to proceed beyond (2.16):
1) Fix $\epsilon > 0$. Take $\nu_\epsilon^* \in \mathcal{P}(S)$ such that $\sup_{\pi\in\Pi}\left\|d_\mu^\pi/\nu_\epsilon^*\right\|_\infty - \epsilon \leq \inf_{\nu\in\mathcal{P}(S)}\sup_{\pi\in\Pi}\left\|d_\mu^\pi/\nu_\epsilon^*\right\|_\infty$. Then apply Lemma 2.5.2 (i) and the fact $|\Phi(\nu,\pi)-\Phi(\nu,\pi')| \leq C_\Phi$ for any $\pi,\pi'\in\Pi, \nu\in\mathcal{P}(S)$ to get

$$\text{Nash-regret}(T) \leq \frac{3}{T}\left(\frac{2(\widetilde{\kappa}_\mu+\epsilon)T(C_\Phi+N^2\alpha\cdot T)}{\eta(1-\gamma)^2}+\frac{8(\widetilde{\kappa}_\mu+\epsilon)\eta T^2\bar{A}^2N^2}{(1-\gamma)^7}\right)^{\frac{1}{2}}.$$

By letting ϵ to 0 and taking step size $\eta = \frac{(1-\gamma)^{2.5}\sqrt{C_\Phi+N^2\alpha T}}{2N\bar{A}\sqrt{T}}$, we have

$$\text{Nash-regret}(T) \leq \frac{3\cdot 2^{\frac{3}{2}}\sqrt{\widetilde{\kappa}_\mu\bar{A}N}}{(1-\gamma)^{\frac{9}{4}}}\left(\frac{C_\Phi}{T}+N^2\alpha\right)^{\frac{1}{4}}.$$

2) We can also proceed (2.16) with Lemma 2.5.2 (ii) and $\eta \leq \frac{(1-\gamma)^4}{8\kappa_\nu^3N\bar{A}}$ to get

$$\text{Nash-regret}(T) \leq 6\sqrt{\frac{\sup_{\pi\in\Pi}\left\|\frac{d_\mu^\pi}{\nu}\right\|_\infty(C_\Phi+N^2\alpha\cdot T)}{\eta T(1-\gamma)^2}}.$$

We next discuss two special choices of ν for proving our bound. First, if $\nu = \mu$, then $\eta \leq \frac{(1-\gamma)^4}{8\kappa_\mu^3N\bar{A}}$. By letting $\eta = \frac{(1-\gamma)^4}{8\kappa_\mu^3N\bar{A}}$, the last square root term can be bounded by $\mathcal{O}\left(\sqrt{\frac{\kappa_\mu^4N\bar{A}(C_\Phi+N^2\alpha\cdot T)}{T(1-\gamma)^6}}\right)$. Second, if $\nu = \frac{1}{|S|}\mathbf{1}$, the uniform distribution over S, then $\kappa_\nu \leq \frac{1}{S}$, which allows a valid choice $\eta = \frac{(1-\gamma)^4}{8|S|^3N\bar{A}} \leq \frac{(1-\gamma)^4}{8\kappa_\nu^3N\bar{A}}$. Hence, we can bound the last square root term by $\mathcal{O}\left(\sqrt{\frac{|S|^4N\bar{A}(C_\Phi+N^2\alpha\cdot T)}{T(1-\gamma)^6}}\right)$. Since ν is arbitrary, combining these two special choices completes the proof.

2.5.2 Sequential Maximum Improvement Algorithm

Let us first fix some notations. Associated with any Markov game $\mathcal{G}$, we define *smoothed* (or regularized) Markov game $\tilde{\mathcal{G}}$, where the expected one-stage payoff of each player i with state s under the joint policy π is $\tilde{u}_i(s,\pi) = \mathbb{E}_{a\sim\pi(s)}[u_i(s,a)] - \tau\sum_{j\in[N]}\nu_j(s,\pi_j)$, where $\nu_j(s,\pi_j) := \sum_{a_j\in A_j}\pi_j(a_j|s)\log(\pi_j(a_j|s))$ is the entropy function, and $\tau > 0$ denotes the regularization parameter. With the smoothed one-stage payoffs, the expected total discounted infinite horizon payoff of player i under policy π is given by

$$\tilde{V}_i(s,\pi) = \mathbb{E}_\pi\Big[\sum_{k=0}^{\infty}\gamma^k\big(u_i(s^k,a^k) - \tau\sum_{j\in[N]}\nu_j(s^k,\pi_j)\big)|s^0 = s\Big], \tag{2.17}$$

for every $s\in S$. The *smoothed* (or entropy-regularized) Q-function is given by

$$\tilde{Q}_i^\pi(s,a_i) = \sum_{a_{-i}\in A_{-i}}\pi_{-i}(a_{-i}|s)\Big(u_i(s,a_i,a_{-i}) - \tau\sum_{j\in[N]}\nu_j(s,\pi_j) + \gamma\sum_{s'\in S}P(s'|s,a)\tilde{V}_i(s',\pi)\Big). \tag{2.18}$$

Algorithm 3 has two main components: first, it computes the optimal one-stage policy update using the smoothed Q-function. Here, the vector of smoothed Q-functions is denoted by $\tilde{\mathbf{Q}}_i^\pi(s) = (\tilde{Q}_i^\pi(s,a_i))_{a_i\in A_i}$. Second, it selects the player who achieves the maximum improvement in the current state to adopt her one-stage policy update, with the policy for the remaining players and the remaining states unchanged. More specifically, the algorithm iterates for T time steps. In every time step $t\in[T-1]$, based on the current policy profile $\pi^{(t)}$, we abuse the notation to use $\tilde{Q}_i^{(t)}$ to denote $\tilde{Q}_i^{\pi^{(t)}}$ and $\tilde{\mathbf{Q}}_i^{(t)}$ to denote $\tilde{\mathbf{Q}}_i^{\pi^{(t)}}$. The expected smoothed Q-function of player i is computed as $\tilde{Q}_i^{(t)}(s,\pi_i) = \sum_{a_i\in A_i}\pi_i(a_i|s)\tilde{Q}_i^{(t)}(s,a_i)$ for all $s\in S$ and all $i\in[N]$. Then, each player computes her one-stage best response strategy by maximizing the smoothed Q-function: for every $i\in[N], a_i\in A_i, s\in S$,

$$\begin{aligned}\mathrm{BR}_i^{(t)}(a_i|s) &= \left(\operatorname*{arg\,max}_{\pi_i'\in\Pi_i}\ \Big(\tilde{Q}_i^{(t)}(s,\pi_i') - \tau\nu_i(s,\pi_i')\Big)\right)_{a_i}\\ &= \frac{\exp(\tilde{Q}_i^{(t)}(s,a_i)/\tau)}{\sum_{a_i'\in A_i}\exp(\tilde{Q}_i^{(t)}(s,a_i')/\tau)},\end{aligned} \tag{2.19}$$

and its maximum improvement of smoothed Q-function value in comparison to current policy is

$$\Delta_i^{(t)}(s) = \max_{\pi_i'\in\Pi_i}\Big(\tilde{Q}_i^{(t)}(s,\pi_i') - \tau\nu_i(s,\pi_i')\Big) - \Big(\tilde{Q}_i^{(t)}(s,\pi_i^{(t)}) - \tau\nu_i(s,\pi_i^{(t)})\Big), \quad \forall s\in S. \tag{2.20}$$

Note that computing $\Delta_i^{(t)}$ is straightforward as the maximization in (2.20) is attained at $\mathrm{BR}_i^{(t)}(s)$ (cf. (2.19)).

If the maximum improvement $\Delta_i^{(t)}(s) \leq 0$ for all $i \in [N]$ and all $s \in S$, then the algorithm terminates and returns the current policy profile $\pi^{(t)}$. Otherwise, the algorithm chooses a tuple of player and state $(\bar{i}^{(t)}, \bar{s}^{(t)})$ associated with the maximum improvement value $\Delta_i^{(t)}(s)$, and updates the policy of player $\bar{i}^{(t)}$ in state $\bar{s}^{(t)}$ with her one-stage best response strategy[2]. The policies of all other players and other states remain unchanged.

Remark 2.5.2. *Using entropy regularization in* (2.19) *has several advantages: (i) unlike Algorithm 2, it avoids projection over simplex which can be costly in large-scale problems; (ii) it ensures that the optimizer is unique.*

Remark 2.5.3. *Algorithm 3 is reminiscent of the "Nash-CA" algorithm*[3] *proposed in [131], which requires each player to sequentially compute the best response policy using an RL algorithm in each iteration, while keeping the strategies of other players fixed. Such sequential best response algorithms are known to ensure finite improvement in the potential function value in potential games [113], which ensures convergence. Meanwhile, Algorithm 3 does not compute the best response strategy in the updates. Instead, it only computes a smoothed* one-step *optimal deviation, as per* (2.19)*, for the current state. The policies for the remaining states and other players are unchanged. The analysis of such one-step deviation-based dynamics is non-trivial and requires new techniques, as highlighted in the next section.*

Remark 2.5.4. *While Algorithm 2 can be run independently by each player in a decentralized fashion, Algorithm 3 is centralized as players do not update their policies simultaneously. Comparing Nash regret in Theorems 2.5.1 and 2.5.2, it is evident that the coordination in Algorithm 3 ensures better scaling of regret with respect to the number of players.*

Theorem 2.5.2. *Consider a Markov α-potential game with an α-potential function Φ and initial state distribution μ such that $\bar{\mu} := \min_{s\in S} \mu(s) > 0$. Denote $\bar{A} := \max_{i\in[N]} |A_i|$ and $C := \max_{i\in[N]} \|u_i\|_\infty$. Then the policy updates generated from Algorithm 3 with parameter*

$$\tau = \frac{1}{N}\left(\log(\bar{A}) + \frac{\log(\bar{A})}{\sqrt{\alpha + \frac{C_\Phi}{T}}} \sqrt{\frac{2\log(\bar{A})}{(1-\gamma)}} \sqrt{\frac{N}{T}} + \frac{2\sqrt{\bar{\mu}}(1-\gamma)\log(\bar{A})}{8C\sqrt{\bar{A}}\sqrt{\alpha + \frac{C_\Phi}{T}}} \right)^{-1} \tag{2.23}$$

has the Nash-regret(T) bounded by

$$\mathcal{O}\left(\frac{\sqrt{N^{3/2}\bar{A}}\log(\bar{A})}{(1-\gamma)^{5/2}\sqrt{\bar{\mu}}} \max\left\{ \left(\alpha + \frac{C_\Phi}{T}\right)^{\frac{1}{2}}, \left(\alpha + \frac{C_\Phi}{T}\right)^{\frac{1}{4}} \right\} \right),$$

where $C_\Phi > 0$ is a constant satisfying $|\Phi(\mu,\pi) - \Phi(\mu,\pi')| \leq C_\Phi$ for any $\pi, \pi' \in \Pi, \mu \in \mathcal{P}(S)$.

[2] Any tie-breaking rule can be used here if the maximum improvement is achieved by more than one tuple.

[3] Unlike this paper, the Nash-CA Algorithm in [131] was proposed in the context of finite horizon Markov potential games.

Algorithm 3 Sequential Maximum Improvement Algorithm

```
Input: Smoothness parameter $\tau$, for every $i \in [N], a_i \in A_i, s \in S$, set $\pi_i^{(0)}(a_i|s) = 1/|A_i|$.
for $t = 0, 1, 2, ..., T-1$ do
  Compute the maximum improvement of smoothed $Q$-function $\{\Delta_i^{(t)}(s)\}_{i\in[N],s\in S}$ as in (2.20).
  if $\Delta_i^{(t)}(s) \leq 0$ for all $i \in [N]$ and all $s \in S$ then
    return $\pi^{(t)}$.
  else
    Choose the tuple $(\bar{i}^{(t)}, \bar{s}^{(t)})$ with the maximum improvement
        $$(\bar{i}^{(t)}, \bar{s}^{(t)}) \in \arg\max_{i\in[N],s\in S} \Delta_i^{(t)}(s), \qquad (2.21)$$
    and update policy
        $$\begin{aligned} \pi_{\bar{i}^{(t)}}^{(t+1)}(a|\bar{s}^{(t)}) =& \mathrm{BR}_{\bar{i}^{(t)}}^{(t)}(a|\bar{s}^{(t)}),\ \forall a \in A_{\bar{i}^{(t)}}, \\ \pi_i^{(t+1)}(s) =& \pi_i^{(t)}(s)\ \forall (i,s) \neq (\bar{i}^{(t)}, \bar{s}^{(t)}). \end{aligned} \qquad (2.22)$$
  end if
end for
```

In the special case of $\alpha = 0$, Theorem 2.5.2 provides a Nash-regret bound of Algorithm 3 for the case of MPGs.

To prove Theorem 2.5.2, we first develop a smoothed version of the multi-agent performance difference lemma (Lemma 2.5.3). This lemma bounds the difference in the smoothed value function $\tilde{V}_i$ by the changes in policy π_i, which is further bounded by the maximum improvements $\Delta_i^{(t)}$. Lemma 2.5.4 bounds the discrepancy between the value function V_i and the smoothed value function $\tilde{V}_i$. Lemma 2.5.3 and 2.5.4 together implies that the Nash-regret of Algorithm 3 is bounded by $\Delta_i^{(t)}$ (2.20). Finally, Lemma 2.5.5 establishes $\Delta_i^{(t)}$ can be bounded by policy updates, which in turn, are bounded by α and the difference in the α-potential function Φ.

Lemma 2.5.3 (Smoothed performance difference)**.** *For any $i \in [N]$, $\mu \in \mathcal{P}(S)$, $\pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$,*

$$\tilde{V}_i(\mu,\pi) - \tilde{V}_i(\mu,\pi') = \frac{1}{1-\gamma}\sum_{s'\in S} d_\mu^\pi(s')\Big((\pi_i(s') - \pi_i'(s'))^\top \cdot \tilde{\mathbf{Q}}_i^{\pi'}(s') + \tau\nu_i(s',\pi_i') - \tau\nu_i(s',\pi_i)\Big),$$

where $\pi = (\pi_i, \pi_{-i})$, and $\pi' = (\pi_i', \pi_{-i})$.

Lemma 2.5.4. *For any $i \in [N], \mu \in \mathcal{P}(S), \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$,* $\Big|V_i(\mu,\pi_i,\pi_{-i}) - V_i(\mu,\pi_i',\pi_{-i}) - (\tilde{V}_i(\mu,\pi_i,\pi_{-i}) - \tilde{V}_i(\mu,\pi_i',\pi_{-i}))\Big| \leq \frac{2\tau N\log(\bar{A})}{1-\gamma}$.

Lemma 2.5.5. *The following inequalities hold:*

(1) $\Delta^{(t)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}) \leq \frac{4C\sqrt{\bar{A}}(1+\tau N\log(\bar{A}))}{1-\gamma}\|\pi^{(t+1)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}) - \pi^{(t)}_{\bar{i}^{(t)}}(\bar{s}^{(t)})\|_2$, *for any* $t \in [T]$.

(2) $\sum_{t=0}^{T-1} \|\pi^{(t+1)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}) - \pi^{(t)}_{\bar{i}^{(t)}}(\bar{s}^{(t)})\|_2^2 \leq \frac{2}{\tau\bar{\mu}}\Big(|\Phi(\mu, \pi^{(T)}) - \Phi(\mu, \pi^{(0)})| + \alpha T + \frac{2\tau N\log(\bar{A})}{1-\gamma}\Big)$.

Proof of Theorem 2.5.2. First, we bound the instantaneous regret $R_i^{(t)}$ for any arbitrary player $i \in [N]$ at time $t \in [T]$. Recall that $R_i^{(t)} = V_i(\mu, \pi_i^\dagger, \pi_{-i}^{(t)}) - V_i(\mu, \pi^{(t)})$, where $\pi_i^\dagger \in \arg\max_{\pi_i' \in \Pi_i} V_i(\mu, \pi_i', \pi_{-i}^{(t)})$. By Lemma 2.5.4,

$$R_i^{(t)} \leq \tilde{V}_i(\mu, \pi_i^\dagger, \pi_{-i}^{(t)}) - \tilde{V}_i(\mu, \pi^{(t)}) + \frac{2\tau N\log(\bar{A})}{(1-\gamma)}.$$

Next, note that for any $i \in [N], \mu \in \mathcal{P}(S)$, by Lemma 2.5.3,

$$\begin{aligned}
\tilde{V}_i\left(\mu, \pi_i^\dagger, \pi_{-i}^{(t)}\right) - \tilde{V}_i\left(\mu, \pi_i^{(t)}, \pi_{-i}^{(t)}\right) &\leq \frac{1}{1-\gamma}\sum_{s\in S} d_\mu^{\pi_i^\dagger, \pi_{-i}^{(t)}}(s)\Bigg(\tau(\nu_i(s, \pi_i^{(t)}) - \nu_i(s, \pi_i'))\\
&\quad + \max_{\pi_i'} \sum_{a_i \in A_i}\left(\left(\pi_i'(a_i|s) - \pi_i^{(t)}(a_i|s)\right)\tilde{Q}_i^{(t)}(s, a_i)\right)\Bigg)\\
&\overset{(a)}{=} \frac{1}{1-\gamma}\sum_{s\in S} d_\mu^{\pi_i^\dagger, \pi_{-i}^{(t)}}(s)\Delta_i^{(t)}(s)\\
&\overset{(b)}{\leq} \frac{1}{1-\gamma}\sum_{s\in S} d_\mu^{\pi_i^\dagger, \pi_{-i}^{(t)}}(s)\Delta^{(t)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}) = \frac{1}{1-\gamma}\left(\Delta^{(t)}_{\bar{i}^{(t)}}(\bar{s}^{(t)})\right),
\end{aligned}$$

where (a) is by (2.20), (b) holds since $\Delta_i^{(t)}(s) \leq \Delta^{(t)}_{\bar{i}^{(t)}}(\bar{s}^{(t)})$ for all $i \in [N], s \in S$. To summarize,

$$R_i^{(t)} \leq \frac{1}{1-\gamma}\left(\Delta^{(t)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}) + 2\tau N\log(\bar{A})\right).$$

Then by Lemma 2.5.5 (1),

$$\begin{aligned}
\text{Nash-regret}(T) &\leq \frac{1}{T(1-\gamma)}\sum_{t\in[T]}\left(\Delta^{(t)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}) + 2\tau N\log(\bar{A})\right)\\
&\leq \frac{2\tau N\log(\bar{A})}{(1-\gamma)} + \frac{4C\sqrt{\bar{A}}(1+\tau N\log(\bar{A}))}{T(1-\gamma)^2}\cdot\sum_{t\in[T]}\left\|\pi^{(t+1)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}) - \pi^{(t)}_{\bar{i}^{(t)}}(\bar{s}^{(t)})\right\|_2\\
&\leq \frac{2\tau N\log(\bar{A})}{(1-\gamma)} + \frac{4C\sqrt{\bar{A}}(1+\tau N\log(\bar{A}))}{\sqrt{T}(1-\gamma)^2}\cdot\Big(\sum_{t\in[T]}\left\|\pi^{(t+1)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}) - \pi^{(t)}_{\bar{i}^{(t)}}(\bar{s}^{(t)})\right\|_2^2\Big)^{\frac{1}{2}}, \qquad (2.24)
\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz inequality. For ease of exposition, define $D_1 := \frac{8C\sqrt{\bar{A}}}{\sqrt{\bar{\mu}}(1-\gamma)^2}$, $D_2 := \sqrt{\alpha + \frac{C_\Phi}{T}}$, and $D_3 := \sqrt{\frac{2\log(\bar{A})}{(1-\gamma)}}$. Then by Lemma 2.5.5 (2),

$$\begin{aligned}
(2.24) \leq & \frac{D_1(1+\tau N\log(\bar{A}))}{\sqrt{\tau}}\sqrt{D_2^2 + \frac{\tau N}{T}D_3^2} + \tau N D_3^2 \\
\leq & \frac{D_1(1+\tau N\log(\bar{A}))}{\sqrt{\tau}}\left(D_2 + \sqrt{\frac{\tau N}{T}}D_3\right) + \tau N D_3^2,
\end{aligned}$$

where the last inequality follows from the fact that for any two positive scalars x, y, $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$. Setting τ as per (2.23) ensures that $\tau < \sqrt{\tau}$ as $\tau \leq 1$. Thus,

$$\text{Nash-regret}(T) \leq \frac{D_1 D_2}{\sqrt{\tau}} + \frac{D_1 D_3\sqrt{N}}{\sqrt{T}} + \sqrt{\tau}N\left(D_1 D_2\log(\bar{A}) + D_1 D_3\log(\bar{A})\sqrt{\frac{N}{T}} + D_3^2\right).$$

Plugging in the value of τ as per (2.23) we obtain,

$$\begin{aligned}
\text{Nash-regret}(T) &\leq \sqrt{N}\left(D_1^2 D_2^2\log(\bar{A}) + D_1^2 D_2 D_3\log(\bar{A})\sqrt{\frac{N}{T}} + D_1 D_2 D_3^2\right)^{\frac{1}{2}} + \frac{D_1 D_3\sqrt{N}}{\sqrt{T}} \\
&\leq D_1 D_2\sqrt{N}\sqrt{\log(\bar{A})} + D_1\sqrt{D_2 D_3\log(\bar{A})}\frac{N^{\frac{3}{4}}}{T^{\frac{1}{4}}} + \sqrt{D_1 D_2}D_3\sqrt{N} + \frac{D_1 D_3\sqrt{N}}{\sqrt{T}}.
\end{aligned}$$

Note that $D_3 \geq 1$ and additionally, we assume that $D_1 \geq 1$ (choose large enough C that ensures this). Then,

$$\begin{aligned}
&\text{Nash-regret}(T) \\
&\leq D_1 D_2 D_3\sqrt{N}\sqrt{\log(\bar{A})} + D_1 D_3\sqrt{D_2\log(|\bar{A}|)}\frac{N^{\frac{3}{4}}}{T^{\frac{1}{4}}} + \sqrt{D_2}D_1 D_3\sqrt{N} + \frac{D_1 D_3\sqrt{N}}{\sqrt{T}} \\
&\leq D_1 D_3\sqrt{N\log(\bar{A})}\left(D_2 + \sqrt{D_2}\left(1 + \left(\frac{N}{T}\right)^{\frac{1}{4}}\right) + \sqrt{\frac{1}{T}}\right) \\
&\leq D_1 D_3\sqrt{\log(\bar{A})}N^{\frac{3}{4}}\mathcal{O}(\max\{D_2, \sqrt{D_2}\}).
\end{aligned}$$

The proof is finished by plugging in D_1, D_2 and D_3.

2.6 Numerical Experiments

This section studies the empirical performance of Algorithms 2 and 3 for Markov congestion game (MCG) and perturbed Markov team game (PMTG) discussed in Section 2.2.2.

Although Section 2.5 focuses on model-based algorithms, in our numerical study both Algorithm 2 and Algorithm 3 are implemented in a model-free manner, where the Q-functions are estimated from samples [49, 101]. Below are the details of the setup of the experiments.

MCG: Consider MCG with $N = 8$ players, where there are $|E| = 4$ facilities A, B, C, D that each player can select from, i.e., $|A_i| = 4$. For each facility j, there is an associated state s_j: *normal* ($s_j = 0$) or *congested* ($s_j = 1$) state, and the state of the game is $s = (s_j)_{j\in E}$. The reward for each player being at facility k is equal to w_k^{safe} times the number of players at $k = A, B, C, D$. We set $w_A^{\text{safe}} = 1 < w_B^{\text{safe}} = 2 < w_C^{\text{safe}} = 4 < w_D^{\text{safe}} = 6$, i.e., facility D is most preferable by all players. However, if more than $N/2$ players find themselves in the same facility, then this facility transits to the *congested* state, where the reward for each player is reduced by a large constant $c = -100$. To return to the *normal* state, the facility should contain no more than $N/4$ players.

PMTG: Consider a game where each player votes for approving or disapproving a project, which is only conducted if a majority of players vote for approval. The state of excitement about the project changes between different rounds depending on the number of players approving it. Mathematically, consider a game with $N = 16$ players, where there are two actions per player: *approve* ($a_i = 1$) or *disapprove* ($a_i = 0$). There can be two states of the project: *high* ($s = 1$) and *low* ($s = 0$) levels of excitement for the project.

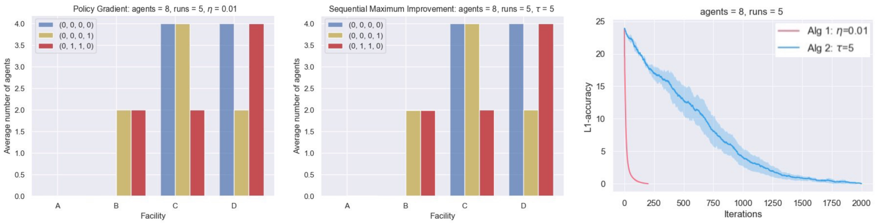
The individual reward of player i is given by $u_i(s, a) = \mathbf{1}_{\{\sum_i a_i \geq N/2\}} + w_i \mathbf{1}_{\{a_i = s\}} - w_i' a_i$, where the first term represents the common utility derived by everyone if the project is approved, the second term represents the utility derived by a player in approving a high-priority project or disapproving a low-priority project, and the third term corresponds to the cost of approving the project. Here, we set $w_i = 10\kappa \cdot \frac{N+1-i}{N}$ and $w_i' = \kappa \cdot \frac{i+1}{N}$. Here, parameter κ captures the magnitude of perturbation.

The state transitions from the *high excitement state* to itself with probability λ_1 if more than $N/4$ players approve it; otherwise, it transitions to itself with probability λ_2. In contrast, the state transitions from the *low excitement state* to high with probability λ_3 if there are at least $N/2$ approvers; if there are $N/2$ or fewer approvers, it transitions to high with probability λ_4.

For both games, we perform episodic updates with 20 steps and a discount factor $\gamma = 0.99$. We estimate the Q-functions and the utility functions using the average of mini-batches of size 10. For MCG, Figures 2.3a and 2.3b illustrate the average number of players taking particular action in different states at the converged values of policy. For example, in the state $(0, 0, 0, 1)$ (denoted by the yellow label in Figure 2.3a and 2.3b), facility D is congested, while the other facilities remain in a normal state. In this scenario, only $N/4 = 2$ players select facility D to restore it to a normal state. Simultaneously, $N/2$ players choose facility C, which provides the second-highest reward after D. The number of players at C is within the congestion threshold ($N/2$), thus ensuring that it remains in a normal state.

For PMTG, we set $\lambda_1 = \lambda_3 = 1$, $\lambda_2 = \lambda_4 = 0$ and $\kappa = 0.1$. Figures 2.4a and 2.4b illustrate the average number of players taking particular action in different states at the converged values of policy. For example, in the "high" state of excitement about project (denoted by the red label in Figure 2.4a and 2.4b), almost all players will select to approve as it will always remain in high state thereon. Meanwhile, if the state of excitement is “low”, then at least half of the players select to approve it so that it transitions to “high” state in future.

Figures 2.3c and 2.4c depict the L_1-accuracy in the policy space at each iteration, defined as the average distance between the current policy and the final policy of all players, i.e., L_1-accuracy $= \frac{1}{N}\sum_{i\in I}\|\pi_i - \pi_i^{(T)}\|_1$. Figures 2.3c and 2.4c show that Algorithm 2 converges faster for PMTG, while Algorithm 3 converges faster for MCG.

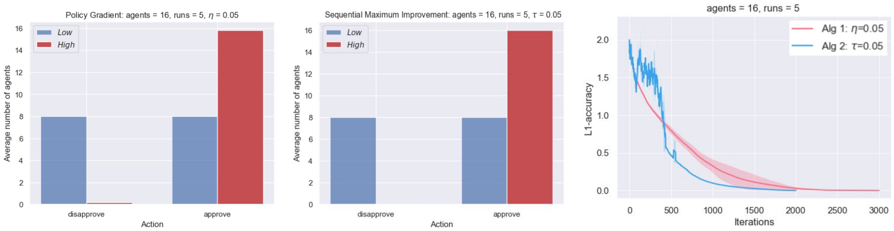


(a) (b) (c)

Figure 2.3: Markov congestion game

Note. (a) and (b) are distributions of players taking four actions in representative states using $\pi^{(T)}$ given by (a) Algorithm 2 with step-size $\eta = 0.01$; (b) Algorithm 3 with regularizer $\tau_t = 0.999^t \cdot 5$. (c) is mean L1-accuracy with shaded region of one standard deviation over all runs

Remark 2.6.1. *We note that the regret bound proposed in our analysis can be loose. In Figure 2.5, we compare growth of regret bound obtained in our theoretical results with that obtained in experiments, where we observe significant gap between the two quantities. This suggests an interesting direction of future research to develop tighter regret bounds.*



(a) (b) (c)

Figure 2.4: Perturbed Markov team game

Note. (a) and (b) are distributions of players taking actions in all states: (a) using Algorithm 2 with step-size $\eta = 0.05$; (b) using Algorithm 3 with regularizer $\tau_t = 0.9975^t \cdot 0.05$. (c) is mean L1-accuracy with shaded region of one standard deviation over all runs.

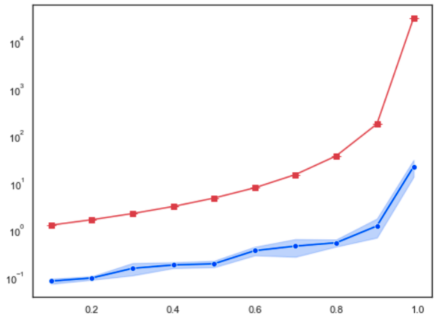


Figure 2.5: Variation of Nash regret with the discount factor for perturbed Markov team game

Note. The perturbation parameter is $\kappa = 0.1$ The red curve plots the function $1/(1-\gamma)^{9/4}$ (as stated in Theorem 6.1) and the blue shaded region show the Nash regret computed through 10 rounds of experiments with random initialization. Note that the scale on y-axis is in log.

2.7 Proofs of Main Results

2.7.1 Proofs in Section 2.3

Proof of Proposition 2.3.1

Let Φ be a potential function of MPG $\mathcal{G}$. Using Definition 2.3.1, it suffices to show $\Phi \in \mathcal{F}^{\mathcal{G}}$. First, we claim that for every $s \in S, \pi, \pi' \in \Pi$,

$$|\Phi(s,\pi) - \Phi(s,\pi')| \leq \sum_{i=1}^{N} |V_i(s,\tilde{\pi}^{(i)}) - V_i(s,\tilde{\pi}^{(i+1)})|, \tag{2.25}$$

where for any $i \in [N]$, $\tilde{\pi}^{(i)} = (\pi'_1, \pi'_2, ..\pi'_{i-1}, \pi_i, \pi_{i+1}, ..., \pi_N)$ with the understanding that $\tilde{\pi}^{(1)} = \pi$ and $\tilde{\pi}^{(N+1)} = \pi'$. To prove this claim, note that

$$|\Phi(s,\pi) - \Phi(s,\pi')| = \left| \sum_{i=1}^{N} \Phi(s,\tilde{\pi}^{(i)}) - \Phi(s,\tilde{\pi}^{(i+1)}) \right| \leq \sum_{i=1}^{N} \left| V_i(s,\tilde{\pi}^{(i)}) - V_i(s,\tilde{\pi}^{(i+1)}) \right|,$$

which follows from Definition 2.3.1 as $\tilde{\pi}^{(i)}$ and $\tilde{\pi}^{(i+1)}$ only differ at player i's policy. By (2.25), for any $s \in S$, $\pi, \pi' \in \Pi$,

$$|\Phi(s,\pi) - \Phi(s,\pi')| \leq 2N \max_{i\in[N]} \|V_i\|_\infty \leq \frac{2N}{1-\gamma} \max_{i\in[N]} \|u_i\|_\infty.$$

Without loss of generality, we have $\min_{\pi\in\Pi} \Phi(s,\pi) = 0$ for every $s \in S$. Therefore, $\|\Phi\|_\infty \leq \frac{2N}{1-\gamma} \max_{i\in[N]} \|u_i\|_\infty$.

To show that Φ lies in a uniformly equi-continuous set $\mathcal{F}_{\mathcal{G}}$, we next show that Φ is uniformly continuous. Note that for each $s \in S$ and $i \in [N]$, $V_i(s,\cdot) : \Pi \to \mathbb{R}$ is a continuous function [156, Lemma 2.10]. Given that Π is compact and $|S| < \infty$, for every $\epsilon > 0$ there exists $\bar{\delta}(\epsilon) > 0$ such that $\max_{i\in[N],s\in S}|V_i(s,\pi) - V_i(s,\pi')| \leq \epsilon/N$ for any $\pi, \pi' \in \Pi$ satisfying $\mathbf{d}(\pi,\pi') \leq \bar{\delta}(\epsilon)$. Consequently, from (2.25), we conclude that for any $\epsilon > 0$, $|\Phi(s,\pi) - \Phi(s,\pi')| \leq \epsilon$ for any $\pi, \pi' \in \Pi$ satisfying $\mathbf{d}(\pi,\pi') \leq \bar{\delta}(\epsilon)$.

Proof of Proposition 2.3.2

The proof of Proposition 2.3.2 relies on the following lemma.

Lemma 2.7.1. *If there exists some $\zeta > 0$ such that for all $s, s' \in S$, $|P(s'|s,w) - P(s'|s,w')| \leq \zeta\|w - w'\|_1$. Then for any $i \in [N], \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$,*

$$\|P^{\pi_i,\pi_{-i}} - P^{\pi_i',\pi_{-i}}\|_\infty \leq 2\zeta|S| \max_{a_i\in A_i}|a_i|/N. \tag{2.26}$$

Proof. For any $i \in [N]$, $\pi \in \Pi$, $\pi_i' \in \Pi_i$, and $s, s' \in S$,

$$\begin{aligned} &P^{\pi_i,\pi_{-i}}(s'|s) - P^{\pi_i',\pi_{-i}}(s'|s) \\ &= \mathbb{E}_{\substack{a_{-i}\sim\pi_{-i}\\ a_i\sim\pi_i}}\Big[P(s'|s,w(a_i,a_{-i})) - P(s'|s,w(a_i,a_{-i}))\Big] \\ &\leq \mathbb{E}_{a_{-i}\sim\pi_{-i}}\Big[P(s'|s,w(\bar{a}_i,a_{-i})) - P(s'|s,w(\underline{a}_i,a_{-i}))\Big], \end{aligned} \tag{2.27}$$

where the first equation is due to the structure of transition function,

$$\bar{a}_i \in \arg\max\nolimits_{a_i\in A_i}P(s'|s,w(a_i,a_{-i})), \text{ and } \underline{a}_i \in \arg\min\nolimits_{a_i\in A_i}P(s'|s,w(a_i,a_{-i})).$$

By (2.27) and the Lipschitz property of the transition matrix in Lemma 2.7.1,

$$\begin{aligned} \sum_{s'\in S}|P^{\pi_i,\pi_{-i}}(s'|s) - P^{\pi_i',\pi_{-i}}(s'|s)| &\overset{(a)}{\leq} \frac{\zeta|S|}{N}\mathop{\mathbb{E}}_{a_{-i}\sim\pi_{-i}}\left[\sum_{e\in E}|\mathbb{1}(e\in\bar{a}_i) - \mathbb{1}(e\in\underline{a}_i)|\right] \\ &\leq \frac{2\zeta|S|\max_{a_i\in A_i}|a_i|}{N}, \quad \forall\, s\in S, \end{aligned}$$

where (a) follows by (2.5). □

Proof of Proposition 2.3.2. Recall that for any $s \in S$, the stage game is a potential game with a potential function $\varphi(s,a) = 1/N\sum_{e\in E}\sum_{j=1}^{w_e(a)N} c_e(s, j/N)$. Under this notation, we can equivalently write (2.6) as

$$\Psi(s,\pi) = \varphi(s,\pi) + \gamma\sum_{s'\in S}P^\pi(s'|s)\Psi(s',\pi). \tag{2.28}$$

For the rest of the proof, fix arbitrary $\pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$ and denote $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi_i', \pi_{-i})$. By (2.28),

$$\Psi(s,\pi) - \Psi(s,\pi') = \varphi(s,\pi) - \varphi(s,\pi') + \gamma \sum_{s' \in S} \left(P^{\pi}(s'|s)\Psi(s',\pi) - P^{\pi'}(s'|s)\Psi(s',\pi') \right). \tag{2.29}$$

Additionally, recall that $V_i(s,\pi) = u_i(s,\pi) + \gamma \sum_{s' \in S} P^{\pi}(s'|s) V_i(s',\pi)$. Consequently,

$$\begin{aligned} V_i(s,\pi) - V_i(s,\pi') &= u_i(s,\pi) - u_i(s,\pi') \\ &+ \gamma \sum_{s' \in S} \left(P^{\pi}(s'|s) V_i(s',\pi) - P^{\pi'}(s'|s) V_i(s',\pi') \right). \end{aligned} \tag{2.30}$$

Subtracting (2.29) from (2.30), we obtain

$$\begin{aligned} &V_i(s,\pi) - V_i(s,\pi') - (\Psi(s,\pi) - \Psi(s,\pi')) \\ &= \gamma \sum_{s' \in S} P^{\pi}(s'|s) \left(V_i(s',\pi) - \Psi(s',\pi) \right) - \gamma \sum_{s' \in S} P^{\pi'}(s'|s) \left(V_i(s',\pi') - \Psi(s',\pi') \right) \\ &= \gamma \sum_{s' \in S} P^{\pi}(s'|s) \left(V_i(s',\pi) - V_i(s',\pi') + \Psi(s',\pi') - \Psi(s',\pi) \right) \\ &\quad - \gamma \sum_{s' \in S} \left(P^{\pi'}(s'|s) - P^{\pi}(s'|s) \right) \left(V_i(s',\pi') - \Psi(s',\pi') \right). \end{aligned}$$

Thus,

$$\begin{aligned} &\max_{s \in S} |V_i(s,\pi) - V_i(s,\pi') - (\Psi(s,\pi) - \Psi(s,\pi'))| \\ &\leq \gamma \max_{s \in S} |V_i(s,\pi) - V_i(s,\pi') - (\Psi(s,\pi) - \Psi(s,\pi'))\,| \\ &+ \gamma \max_{s' \in S} |\Psi(s',\pi') - V_i(s',\pi')| \max_{s \in S} \sum_{s' \in S} \left| P^{\pi}(s'|s) - P^{\pi'}(s'|s) \right|. \end{aligned} \tag{2.31}$$

Rearranging terms leads to

$$\begin{aligned} (2.31) &\leq \frac{\gamma}{1-\gamma} \max_{s' \in S} |\Psi(s',\pi') - V_i(s',\pi')| \| P^{\pi} - P^{\pi'} \|_{\infty} \\ &\leq \frac{2\gamma\zeta |S| \max_{a_i \in A_i} |a_i|}{(1-\gamma)N} \max_{s' \in S} |\Psi(s',\pi') - V_i(s',\pi')|. \end{aligned} \tag{2.32}$$

where the last inequality follows from Lemma 2.7.1. Finally, since

$$u_i(s^k, a^k) = \sum_{e \in E} c_e(s^k, w_e^k) \mathbb{1}(e \in a_i^k) \leq \sum_{e \in E} c_e(s^k, w_e^k) \leq \varphi(s^k, a^k),$$

then for any $s' \in S$,

$$
\begin{aligned}
|\Psi(s',\pi') - V_i(s',\pi')| &\leq \mathbb{E}_{\pi'}\left[\sum_{k=0}^{\infty} \gamma^k \left|\varphi(s^k,a^k) - u_i(s^k,a^k)\right|\right] \\
&\leq \left|\mathbb{E}_{\pi'}\left[\sum_{k=0}^{\infty} \gamma^k \varphi(s^k,a^k)\right]\right| \leq \sup_{s,\pi} \Psi(s,\pi).
\end{aligned}
$$

Plugging the above inequality into (2.32) finishes the proof.

Proof of Proposition 2.3.3

Throughout the proof, let us fix arbitrary $i \in [N], \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$, and define $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi_i', \pi_{-i})$. We show that for every $i \in [N], \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$,

$$
\max_{s \in S} |V_i(s,\pi) - V_i(s,\pi') - (\Psi(s,\pi) - \Psi(s,\pi'))| \leq \frac{2\kappa}{(1-\gamma)^2},
$$

where $\Psi(s,\pi) \coloneqq \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r(s^k,a^k) | s^0 = s\right]$. Note that

$$
\Psi(s,\pi) = r(s,\pi) + \gamma \sum_{s' \in S} P^\pi(s'|s)\Psi(s',\pi). \tag{2.33}
$$

By (2.33), for any $s \in S$,

$$
\Psi(s,\pi) - \Psi(s,\pi') = r(s,\pi) - r(s,\pi') + \gamma \sum_{s' \in S} \left(P^\pi(s'|s)\Psi(s',\pi) - P^{\pi'}(s'|s)\Psi(s',\pi')\right). \tag{2.34}
$$

Similarly, for any $s \in S$,

$$
V_i(s,\pi) - V_i(s,\pi') = u_i(s,\pi) - u_i(s,\pi') + \gamma \sum_{s' \in S} P^\pi(s'|s)V_i(s',\pi) - P^{\pi'}(s'|s)V_i(s',\pi'). \tag{2.35}
$$

Consequently,

$$
\begin{aligned}
&V_i(s,\pi) - V_i(s,\pi') - (\Psi(s,\pi) - \Psi(s,\pi')) \\
=&u_i(s,\pi) - u_i(s,\pi') - (r(s,\pi) - r(s,\pi')) \\
&- \gamma \sum_{s' \in S} \left(P^{\pi'}(s'|s) - P^\pi(s'|s)\right)(V_i(s',\pi') - \Psi(s',\pi')) \\
&+ \gamma \sum_{s' \in S} P^\pi(s'|s)\left(V_i(s',\pi) - V_i(s',\pi') + \Psi(s',\pi') - \Psi(s',\pi)\right).
\end{aligned}
$$

Since $|u_i(s,\pi) - u_i(s,\pi') - (r(s,\pi) - r(s,\pi'))| \leq 2\|\xi_i\|_\infty \leq 2\kappa$, then

$$\begin{aligned}
&\max_{s\in S} |V_i(s,\pi) - V_i(s,\pi') - (\Psi(s,\pi) - \Psi(s,\pi'))\,| \qquad (2.36)\\
\leq& 2\kappa + 2\gamma \max_{s'\in S} |\Psi(s',\pi') - V_i(s',\pi')|\\
&\quad + \gamma \max_{s\in S} |V_i(s,\pi) - V_i(s,\pi') - (\Psi(s,\pi) - \Psi(s,\pi'))\,|.
\end{aligned}$$

Rearranging terms in above inequality, we obtain

$$(2.36) \leq \frac{2\kappa}{1-\gamma} + \frac{2\gamma}{1-\gamma} \max_{s'\in S} |\Psi(s',\pi') - V_i(s',\pi')|. \qquad (2.37)$$

Note that $|\Psi(s',\pi') - V_i(s',\pi')| = |\sum_{k=0}^{\infty} \gamma^k \xi_i(s,\pi'(s^k))| \leq \kappa/(1-\gamma)$. Plugging this inequality into (2.37) completes the proof.

2.7.2 Proofs in Section 2.4

Proof of Proposition 2.4.1

To prove Proposition 2.4.1, we first need the following lemma.

Lemma 2.7.2 (Lemma B.1 in [156])**.** *Fix $i \in [N]$ and $K \in \mathbb{N}$. For any $s \in S$ and $\omega = \left(\tilde{s}^k, \tilde{a}^k\right)_{k=0}^K \in (S \times A)^{K+1}$, the mapping $\Pi \ni \pi \mapsto \mathbb{E}_\pi\left[\mathbb{1}\left((s^k,a^k)_{k=0}^K = \omega\right) \mid s^0 = s\right]$ is continuous.*

Proof of Proposition 2.4.1. Fix $\epsilon > 0$ and define $M := N \max_{i\in[N]} \|u_i\|_\infty$. Choose $K \in \mathbb{N}$ large enough that $\frac{\gamma^K \cdot M}{1-\gamma} < \frac{\epsilon}{4}$ and $\tilde{\epsilon} := \frac{(1-\gamma)\epsilon}{2M|S|^{K+1}|A|^{K+1}}$. Since Π is compact and $S \times A$ is finite, Lemma 2.7.2 ensures that there exists $\delta(\epsilon)$ such that for any $\pi, \pi' \in \Pi$ with $\mathbf{d}(\pi,\pi') \leq \delta(\epsilon)$, and $\omega \in (S\times A)^{K+1}, s \in S$,

$$\left|\mathbb{E}_\pi\left[\mathbb{1}\left((s^k,a^k)_{k=0}^K = \omega\right) \mid s^0 = s\right] - \mathbb{E}_{\pi'}\left[\mathbb{1}\left((s^k,a^k)_{k=0}^K = \omega\right) \mid s^0 = s\right]\right| \leq \tilde{\epsilon}. \qquad (2.38)$$

From (2.7), we note that for any $\Psi \in \tilde{\mathcal{F}}^{\mathcal{G}}$, there exists $\phi : S \times A \to \mathbb{R}$ such that for any $\pi, \pi' \in \Pi, s \in S$,

$$\begin{aligned}
&|\Psi(s,\pi) - \Psi(s,\pi')|\\
\leq& \left|\mathbb{E}_\pi\left[\sum_{k=0}^K \gamma^k \phi\left(s^k,a^k\right) \mid s_0 = s\right] - \mathbb{E}_{\pi'}\left[\sum_{k=0}^K \gamma^k \phi\left(s^k,a^k\right) \mid s_0 = s\right]\right| + \frac{\epsilon}{2}. \qquad (2.39)
\end{aligned}$$

Define a function $\varphi : (S \times A)^{K+1} \to \mathbb{R}$ such that for every $\left(\tilde{s}^k, \tilde{a}^k\right)_{k=0}^K \in (S\times A)^{K+1}$, $\varphi\left(\tilde{s}^0, \tilde{a}^0, \cdots, \tilde{s}^K, \tilde{a}^K\right) := \sum_{k=0}^K \gamma^k \phi\left(\tilde{s}^k, \tilde{a}^k\right)$. Thus, for any $\pi \in \Pi$,

$$\mathbb{E}_\pi\left[\sum_{k=0}^K \gamma^k \phi\left(s^k,a^k\right) \mid s^0 = s\right] = \sum_{\omega\in(S\times A)^{K+1}} \varphi(\omega)\mathbb{E}_\pi\left[\mathbb{1}\left(\left(s^k,a^k\right)_{t=0}^K = \omega\right) \middle| s^0 = s\right].$$

Thus, by applying the above equation and (2.38) to (2.39), we obtain that for any $s \in S, \pi, \pi' \in \Pi$ satisfying $\mathbf{d}(\pi, \pi') \leq \delta(\epsilon)$,

$$|\Psi(s,\pi) - \Psi(s,\pi')| \leq \|\varphi\|_\infty |S|^{K+1}|A|^{K+1}\tilde{\epsilon} + \frac{\epsilon}{2} \leq \frac{M|S|^{K+1}|A|^{K+1}\tilde{\epsilon}}{1-\gamma} + \frac{\epsilon}{2} \leq \epsilon.$$

Since we chose arbitrary $\Psi \in \tilde{\mathcal{F}}^{\mathcal{G}}$, and δ is independent of the choice of Ψ, then $\tilde{\mathcal{F}}^{\mathcal{G}}$ is equi-continuous. Thus, $\tilde{\mathcal{F}}^{\mathcal{G}} \subseteq \mathcal{F}^{\mathcal{G}}$.

2.7.3 Proofs in Section 2.5.1

Proof of Lemma 2.5.2

To prove Lemma 2.5.2, we define $\pi_{i\sim j} \coloneqq \{\pi_k\}_{k=i+1}^{j-1}$ as the joint policy for players from $i+1$ to $j-1$; $\pi_{<i} \coloneqq \{\pi_k\}_{k=1}^{i-1}$, and $\pi_{>j} \coloneqq \{\pi_k\}_{k=j+1}^{N}$ are defined similarly. Next, we recall a useful result from [49].

Lemma 2.7.3 (Lemma 2 in [49]). *For any function $f : \Pi \to \mathbb{R}$, and any two policies $\pi, \pi' \in \Pi$,*

$$\begin{aligned} f(\pi') - f(\pi) = &\sum_{i=1}^{N} \left(f(\pi_i', \pi_{-i}) - f(\pi)\right) \\ &+ \sum_{i=1}^{N}\sum_{j=i+1}^{N} \Big(f(\pi_{<i,i\sim j}, \pi'_{>j}, \pi'_i, \pi'_j) - f(\pi_{<i,i\sim j}, \pi'_{>j}, \pi_i, \pi'_j) \\ &- f(\pi_{<i,i\sim j}, \pi'_{>j}, \pi'_i, \pi_j) + f(\pi_{<i,i\sim j}, \pi'_{>j}, \pi_i, \pi_j)\Big). \end{aligned} \tag{2.40}$$

Next, we state a result that lower bounds the improvement in value function of each player in each step of Algorithm 2.

Lemma 2.7.4. *Consider a Markov game $\mathcal{G}$ with initial state distribution ν, let $\pi^{(t+1)}$ and $\pi^{(t)}$ be consecutive policies in Algorithm 2. Then we have,*

$$\begin{aligned} &\textit{(i)}\, V_i(\nu, \pi^{(t+1)}) - V_i(\nu, \pi^{(t)}) \\ &\quad \geq -\frac{4\eta^2 \bar{A}^2 N^2}{(1-\gamma)^5} + \frac{1}{2\eta(1-\gamma)} \cdot \sum_{i\in[N], s\in S} d_\nu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|^2; \end{aligned}$$

$$\begin{aligned} &\textit{(ii)}\, V_i(\nu, \pi^{(t+1)}) - V_i(\nu, \pi^{(t)}) \\ &\quad \geq \frac{1}{2\eta(1-\gamma)} \left(1 - \frac{4\eta\kappa_\nu^3 \bar{A} N}{(1-\gamma)^4}\right) \cdot \sum_{i=1}^{N}\sum_{s\in S} d_\nu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|^2. \end{aligned}$$

Proof. This result directly follows from [49, Lemma 3]. Specifically, the proof of [49, Lemma 3] is established by lower-bounding the difference $\Phi(\nu, \pi^{(t+1)}) - \Phi(\nu, \pi^{(t)})$ for a Markov potential game with potential function Φ. At its core, the proof relies on the key property of Markov potential games, which allows the difference in potential functions to be expressed as the difference in value functions for each player. The remainder of the proof focuses on lower-bounding the difference in value functions at each step of the policy update process in Algorithm 2, which is precisely what we require. We omit details due to space constraints. □

Proof of Lemma 2.5.2. For ease of exposition, let $\pi' = \pi^{(t+1)}$ and $\pi = \pi^{(t)}$. By Definition 2.2.3, $|V_i(\nu, \pi_i', \pi_{-i}) - V_i(\nu, \pi_i, \pi_{-i}) - (\Phi(\nu, \pi_i', \pi_{-i}) - \Phi(\nu, \pi_i, \pi_{-i}))| \leq \alpha$ for any $\nu, i \in [N], \pi_i, \pi_i' \in \Pi_i$ and $\pi_{-i} \in \Pi_{-i}$. Apply Lemma 2.7.3 with $f(\cdot) = V_i(\nu, \cdot) - \Phi(\nu, \cdot)$ respectively. Since each term in (2.40) only differs in one player's policy, we obtain

$$|V_i(\nu, \pi') - V_i(\nu, \pi) - (\Phi(\nu, \pi') - \Phi(\nu, \pi'))| \leq \sum_{i=1}^{N} \alpha + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \alpha \leq N^2 \alpha.$$

The proof follows by the above inequality and Lemma 2.7.4.

2.7.4 Proofs in Section 2.5.2

Proof of Lemma 2.5.3

Fix arbitrary $i \in [N], \mu \in \mathcal{P}(S), \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$. We define $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi_i', \pi_{-i}) \in \Pi$. Note that

$$\begin{aligned}
&\tilde{V}_i(\mu, \pi) - \tilde{V}_i(\mu, \pi') \\
&= \mathbb{E}_\pi\Big[\sum_{k=0}^{\infty} \gamma^k \Big(u_i(s^k, a^k) - \tau \sum_{j \in [N]} \nu_j(s^k, \pi_j) - \tilde{V}_i(s^k, \pi') + \tilde{V}_i(s^k, \pi')\Big)\Big] - \tilde{V}_i(\mu, \pi') \\
&= \mathbb{E}_\pi\Big[\sum_{k=0}^{\infty} \gamma^k \Big(u_i(s^k, a^k) - \tau \sum_{j \in [N]} \nu_j(s^k, \pi_j) - \tilde{V}_i(s^k, \pi')\Big)\Big] + \mathbb{E}_\pi\left[\sum_{k=1}^{\infty} \gamma^k \tilde{V}_i(s^k, \pi')\right]. \quad (2.41)
\end{aligned}$$

Note that

$$\mathbb{E}_\pi\big[\sum_{k=1}^{\infty} \gamma^k \tilde{V}_i(s^k, \pi')\big] = \gamma \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \tilde{V}_i(s^{k+1}, \pi')\right],$$

thus,

$$
\begin{aligned}
(2.41) &= \mathbb{E}_{\pi}\Big[\sum_{k=0}^{\infty}\gamma^k\big(u_i\left(s^k,a^k\right)-\tau\sum_{j\in[N]}\nu_j\left(s^k,\pi_j\right)-\tilde{V}_i\left(s^k,\pi'\right)+\gamma\tilde{V}_i\left(s^{k+1},\pi'\right)\big)\Big] \\
&= \mathbb{E}_{\pi}\Big[\sum_{k=0}^{\infty}\gamma^k\bigg(u_i(s^k,a^k)-\tau\sum_{j\in[N]}\nu_j(s^k,\pi'_j)+\gamma\sum_{s'\in S}P(s'|s^k,a^k)\tilde{V}_i(s',\pi')-\tilde{V}_i(s^k,\pi') \\
&\qquad +\tau\sum_{j\in[N]}\nu_j(s^k,\pi'_j)-\tau\sum_{j\in[N]}\nu_j(s^k,\pi_j)\bigg)\Big]. \qquad (2.42)
\end{aligned}
$$

We can continue the above calculations by applying smoothed Q-function and noting that $\pi'_j=\pi_j$ for all $j\neq i$ and $\tilde{V}_i(s',\pi')=\pi'_i(s')^\top\tilde{\mathbf{Q}}_i^{\pi'}(s')$,

$$
\begin{aligned}
(2.42) &= \mathbb{E}_{\pi_i}\left[\sum_{k=0}^{\infty}\gamma^k\bigg(\tilde{Q}_i^{\pi'}(s^k,a_i^k)-\tilde{V}_i(s^k,\pi')+\tau\sum_{j\in[N]}\nu_j(s^k,\pi'_j)-\tau\sum_{j\in[N]}\nu_j(s^k,\pi_j)\bigg)\right] \\
&=\frac{1}{1-\gamma}\sum_{s'\in S}d_\mu^\pi(s')\Big(\pi_i(s')-\pi'_i(s')\Big)^\top\tilde{\mathbf{Q}}_i^{\pi'}(s')+\tau\nu_i(s',\pi'_i)-\tau\nu_i(s',\pi_i)\Big).
\end{aligned}
$$

Proof of Lemma 2.5.4

From the definition of smoothed infinite horizon utility (2.17), we note that for every $i\in[N],\pi_i\in\Pi_i,\pi_{-i}\in\Pi_{-i},s\in S$,

$$
\tilde{V}_i(s,\pi_i,\pi_{-i})=V_i(s,\pi_i,\pi_{-i})-\tau\mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty}\gamma^k\sum_{j\in[N]}\nu_j(s^k,\pi_j)|s_0=s\right]. \qquad (2.43)
$$

Using (2.43), it holds that for any $\mu\in\mathcal{P}(S)$ and $\pi\in\Pi$,

$$
\begin{aligned}
|\tilde{V}_i(\mu,\pi)-V_i(\mu,\pi)| &= \tau\left|\mathbb{E}_{\mu,\pi}\left[\sum_{k=0}^{\infty}\gamma^k\sum_{j\in[N]}\nu_j(s^k,\pi_j)\right]\right| \\
&\leq\frac{\tau N\max_{s,\pi_i}\nu_i(s,\pi_i)}{1-\gamma}=\frac{\tau N\log(\bar{A})}{1-\gamma}. \qquad (2.44)
\end{aligned}
$$

The desired result follows from triangle inequality and (2.44).

Proof of Lemma 2.5.5

The proof of Lemma 2.5.5 requires the following technical lemmas.

Lemma 2.7.5. *If $\mathcal{G}$ is a Markov α-potential game with Φ as its α-potential function, then for any $s \in S, i \in [N], \pi_i', \pi_i \in \Pi_i, \pi_{-i} \in \Pi_{-i}$, $\Big|(\tilde{\Psi}(s, \pi_i', \pi_{-i}) - \tilde{\Psi}\,(s, \pi_i, \pi_{-i})) - (\tilde{V}_i(s, \pi_i', \pi_{-i}) - \tilde{V}_i(s, \pi_i, \pi_{-i}))\Big| \leq \alpha$, where*

$$\tilde{\Psi}(s, \pi) := \Phi(s, \pi) - \tau \mathbb{E}_\pi[\sum_{j\in[N]} \sum_{k=0}^{\infty} \gamma^k \nu_j(s^k, \pi_j) \mid s^0 = s].$$

Proof. To ease the notation, for function $f : S \times \Pi \to \mathbb{R}$, we write $f(s, \cdot)$ as $f^s(\cdot)$. By (2.43) and the definition of $\tilde{\Psi}$ in Lemma 2.5.5, we have for all $s \in S, i \in [N], \pi_i', \pi_i \in \Pi_i, \pi_{-i} \in \Pi_{-i}$,

$$\begin{aligned}&|\tilde{\Psi}^s(\pi_i', \pi_{-i}) - \tilde{\Psi}^s(\pi_i, \pi_{-i}) - (\tilde{V}_i^s(\pi_i', \pi_{-i}) - \tilde{V}_i^s(\pi_i, \pi_{-i}))|\\ =&|\Phi^s(\pi_i', \pi_{-i}) - \Phi^s(\pi_i, \pi_{-i}) - (V_i^s(\pi_i', \pi_{-i}) - V_i^s(\pi_i, \pi_{-i}))|,\end{aligned}$$

which is bounded by α using Definition 2.2.4. □

Lemma 2.7.6. *For any $i \in [N], s \in S, \pi_i' \in \Pi_i, t \in [T]$, it hold that*

$$\begin{aligned}&\sum_{a_i\in A_i} \tilde{Q}_i^{(t)}(s, a_i) \left(BR_i^{(t)}(a_i|s) - \pi_i'(a_i|s)\right)\\ &\geq \tau \sum_{a_i\in A_i} \log\left(BR_i^{(t)}(a_i|s)\right)\left(BR_i^{(t)}(a_i|s) - \pi_i'(a_i|s)\right).\end{aligned}$$

Proof. Fix arbitrary $i \in [N], s \in S$, and $t \in [T]$. Next, note that the optimization problem in (2.19) is a strongly concave optimization problem. By the first order conditions of constrained optimality, for all $\pi_i' \in \Pi_i$,

$$\left(\tilde{\mathbf{Q}}_i^{(t)}(s) - \tau \nabla_{\pi_i(s)} \nu_i(s, \mathrm{BR}_i^{(t)}(s))\right)^\top (\mathrm{BR}_i^{(t)}(s) - \pi_i'(s)) \geq 0.$$

Note that $\nabla_{\pi_i(a_i|s)} \nu_i(s, \pi_i) = 1 + \log(\pi_i(a_i|s))$ for every $a_i \in A_i$. Therefore, for every $\pi_i' \in \Pi_i$,

$$\begin{aligned}&\sum_{a_i\in A_i} \tilde{Q}_i^{(t)}(s, a_i) \left(\mathrm{BR}_i^{(t)}(a_i|s) - \pi_i'(a_i|s)\right)\\ &\geq \tau \sum_{a_i\in A_i} \left(1 + \log\left(\mathrm{BR}_i^{(t)}(a_i|s)\right)\right)\left(\mathrm{BR}_i^{(t)}(a_i|s) - \pi_i'(a_i|s)\right).\end{aligned}$$

The result follows by noting that $\sum_{a_i\in A_i} \mathrm{BR}_i^{(t)}(a_i|s) = \sum_{a_i\in A_i} \pi_i'(a_i|s) = 1$. □

Lemma 2.7.7. *For any $i \in [N], s \in S, \pi_i, \pi_i' \in \Pi_i$,*

$$\nu_i(s, \pi_i) - \nu_i(s, \pi_i') \geq \frac{1}{2}\|\pi_i(s) - \pi_i'(s)\|^2 + \sum_{a_i\in A_i} (\log(\pi_i'(a_i|s)))\,(\pi_i(a_i|s) - \pi_i'(a_i|s))\,.$$

Proof. Fix arbitrary $i \in [N], s \in S$. To prove the lemma, we first claim that the mapping $\mathcal{P}(A_i) \ni \pi \mapsto \nu_i(s, \pi)$ is 1-strongly convex. This can be observed by computing the Hessian, which is a $\mathbb{R}^{A_i \times A_i}$ diagonal matrix with (a_i, a_i) entry as $1/\pi(a_i|s)$. Since $\pi(a_i|s) \leq 1$, it follows that the diagonal entries of the Hessian matrix are all greater than 1. Thus, $\nu_i(s, \cdot)$ is 1-strongly convex function. The result follows by noting that for any κ-strongly convex function f, $f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\kappa}{2}\|y - x\|^2$. □

Lemma 2.7.8. *For any $i \in [N], t \in [T], a \in A_{\bar{i}^{(t)}}$, there exists $0 \leq t^* \leq t$ such that*

$$\tau |\log(\pi^{(t)}_{\bar{i}^{(t)}}(a|\bar{s}^{(t)}))| \leq 2\|\tilde{\mathbf{Q}}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)})\|_\infty + \tau \log(|A_{\bar{i}^{(t)}}|).$$

Proof. Recall that in Algorithm 3, at any time step $t \in [T]$, player $\bar{i}^{(t)}$ updates her policy at time $t+1$ in the state $\bar{s}^{(t)}$, while policies for other players and other states remain unchanged. Fix arbitrary $t \in [T]$. Let $0 \leq t^* \leq t$ be the latest time step when player $\bar{i}^{(t)}$ updated its policy in state $\bar{s}^{(t)}$ before time t. Note that $t^* = 0$ if t is the first time when player $\bar{i}^{(t)}$ is updating its policy in state $\bar{s}^{(t)}$. Naturally, $\bar{i}^{(t)} = \bar{i}^{(t^*)}$ and $\bar{s}^{(t)} = \bar{s}^{(t^*)}$. Consequently, for every $a \in A_{\bar{i}^{(t)}}$,

$$\pi^{(t)}_{\bar{i}^{(t)}}(a|\bar{s}^{(t)}) = \mathrm{BR}^{(t^*)}_{\bar{i}^{(t)}}(a|\bar{s}^{(t)}) = \frac{\exp(\tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, a))}{\sum_{a' \in A_{\bar{i}^{(t)}}} \exp(\tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, a'))}.$$

Consequently, for every $a \in A_{\bar{i}^{(t)}}$,

$$\begin{aligned}\pi^{(t)}_{\bar{i}^{(t)}}(a|\bar{s}^{(t)}) &\geq \frac{\exp(\tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, \underline{a})/\tau)}{|A_{\bar{i}^{(t)}}| \exp(\tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, \bar{a})/\tau)} \\ &= \frac{1}{|A_{\bar{i}^{(t)}}|} \exp\left(\left(\tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, \underline{a}) - \tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, \bar{a})\right)/\tau\right),\end{aligned}$$

with $\bar{a} \in \underset{a \in A_{\bar{i}^{(t)}}}{\arg\max} \tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, a)$ and $\underline{a} \in \underset{a \in A_{\bar{i}^{(t)}}}{\arg\min} \tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, a)$. Since $\pi^{(t)}_{\bar{i}^{(t)}}(a|\bar{s}^{(t)}) \leq 1$, it follows that for every $a \in A_{\bar{i}^{(t)}}$,

$$\begin{aligned}|\log(\pi^{(t)}_{\bar{i}^{(t)}}(a|\bar{s}^{(t)}))| &\leq \log(|A_{\bar{i}^{(t)}}|) + \frac{1}{\tau}\left(\tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, \bar{a}) - \tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, \underline{a})\right) \\ &\leq \log(|A_{\bar{i}^{(t)}}|) + \frac{2}{\tau}\|\tilde{\mathbf{Q}}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)})\|_\infty.\end{aligned}$$

Lemma 2.7.9. *For any $t \in [T], i \in [N], s \in S$, it holds that $\|\tilde{\mathbf{Q}}^{(t)}_i(s)\|_\infty \leq C\frac{1+\tau N \log(\bar{A})}{1-\gamma}$, where $C := \max_{i \in [N]} \|u_i\|_\infty$.*

Proof. First, we note that for any $s \in S$, $\pi \in \Pi$,

$$\begin{aligned}|\tilde{V}_i(s,\pi)| &\leq \mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k|u_i(s^k,a^k)-\tau\sum_{j\in[N]}\nu_j(s^k,\pi_j)|\right]\\ &\leq \mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k\left(|u_i(s^k,a^k)|+\tau N\log(\bar{A})\right)\right]\leq C\frac{1+\tau N\log(\bar{A})}{(1-\gamma)}.\end{aligned}$$

By (2.18), we note that for every $i \in [N], s \in S, a_i \in A_i$,

$$\begin{aligned}|\tilde{Q}_i^{(t)}(s,a_i)| &\leq \mathop{\mathbb{E}}_{a_{-i}\sim\pi_{-i}}\left[|u_i(s,a_i,a_{-i})-\tau\sum_{j\in[N]}\nu_j(s,\pi_j)|+\gamma\sum_{s'\in S}P(s'|s,a_i,a_{-i})\left|\tilde{V}_i(s',\pi)\right|\right]\\ &\leq C\mathop{\mathbb{E}}_{a_{-i}\sim\pi_{-i}}\left[(1+\tau N\log(\bar{A}))\left(1+\frac{\gamma}{1-\gamma}\right)\right].\end{aligned}$$

Proof of Lemma 2.5.5. (1) Fix $t \in [T]$. To ease the notation, let $\pi'_* := \pi^{(t+1)}_{\bar{i}(t)}$, $\pi_* := \pi^{(t)}_{\bar{i}(t)}$, $\pi_{-*} := \pi^{(t)}_{-\bar{i}(t)}$, $\nu_* := \nu_{\bar{i}(t)}$, Q_* denote $\tilde{Q}^{(t)}_{\bar{i}(t)}$, $\mathbf{Q}_*$ denote $\tilde{\mathbf{Q}}^{(t)}_{\bar{i}(t)}$. Note that by (2.20) and (2.22),

$$\begin{aligned}&\Delta^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)}) = \sum_{a\in A_{\bar{i}(t)}}\left(\pi'_*(a|\bar{s}^{(t)})-\pi_*(a|\bar{s}^{(t)})\right)Q_*(\bar{s}^{(t)},a)+\tau\nu_*(\bar{s}^{(t)},\pi_*)-\tau\nu_*(\bar{s}^{(t)},\pi'_*)\\ &\leq \sum_{a\in A_{\bar{i}(t)}}\left(\pi'_*(a|\bar{s}^{(t)})-\pi_*(a|\bar{s}^{(t)})\right)Q_*(\bar{s}^{(t)},a)+\tau\sum_{a\in A_{\bar{i}(t)}}\log(\pi_*(a|\bar{s}^{(t)}))\left(\pi_*(a|\bar{s}^{(t)})-\pi'_*(a|\bar{s}^{(t)})\right)\\ &\leq \sum_{a\in A_{\bar{i}(t)}}\left(\left|\pi'_*(a|\bar{s}^{(t)})-\pi'_*(a|\bar{s}^{(t)})\right|\cdot\left|Q_*(\bar{s}^{(t)},a)-\tau\log(\pi_*(a|\bar{s}^{(t)}))\right|\right),\end{aligned}\tag{2.45}$$

where the first inequality follows from convexity of $\nu_i(s,\cdot)$. By Cauchy-Schwarz inequality and noting that $\max_{i\in[N]}|A_i| \leq \bar{A}$,

$$\begin{aligned}(2.45)&\leq\sqrt{\bar{A}}\max_{a\in A_{\bar{i}(t)}}\left|Q_*(\bar{s}^{(t)},a)-\tau\log(\pi_*(a|\bar{s}^{(t)}))\right|\cdot\left\|\pi'_*(\bar{s}^{(t)})-\pi_*(\bar{s}^{(t)})\right\|_2\\ &\leq\sqrt{\bar{A}}\left(\max_{a\in A_{\bar{i}(t)}}\left|Q_*(\bar{s}^{(t)},a)\right|+\max_{a\in A_{\bar{i}(t)}}\tau\left|\log(\pi_*(a|\bar{s}^{(t)}))\right|\right)\cdot\left\|\pi'_*(\bar{s}^{(t)})-\pi_*(\bar{s}^{(t)})\right\|_2.\end{aligned}$$

Note that Lemma 2.7.8 implies that there exists $\hat{t} \leq t$ such that $\max_{a\in A_{\bar{i}(t)}}\tau\left|\log(\pi_*(a|\bar{s}^{(t)}))\right| \leq 2\|\tilde{\mathbf{Q}}^{(\hat{t})}_{\bar{i}(t)}(\bar{s}^{(t)})\|_\infty+\tau\log(\bar{A})$. Consequently, it follows that

$$\begin{aligned}\Delta^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)}) &\leq \sqrt{\bar{A}}\left(\|\mathbf{Q}_*(\bar{s}^{(t)})\|_\infty+2\|\tilde{\mathbf{Q}}^{(\hat{t})}_{\bar{i}(t)}(\bar{s}^{(t)})\|_\infty+\tau\log(\bar{A})\right)\cdot\left\|\pi'_*(\bar{s}^{(t)})-\pi_*(\bar{s}^{(t)})\right\|_2\\ &\leq 4C\frac{1+\tau N\log(\bar{A})}{1-\gamma}\sqrt{\bar{A}}\left\|\pi'_*(\bar{s}^{(t)})-\pi_*(\bar{s}^{(t)})\right\|_2,\end{aligned}$$

where the last inequality follows from Lemma 2.7.9. This concludes the proof for Lemma 2.5.5 1).

(2) Here, we show that

$$\sum_{t=1}^{T-1} \|\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\|_2^2 \leq \frac{2}{\tau\bar{\mu}} \left(\tilde{\Psi}(\mu, \pi^{(T)}) - \tilde{\Psi}(\mu, \pi^{(0)}) + \alpha T\right).$$

To see this, note that for any $t \in [T]$,

$$\begin{aligned} &\tilde{\Psi}(\mu, \pi^{(t+1)}) - \tilde{\Psi}(\mu, \pi^{(t)}) = \tilde{\Psi}(\mu, \pi'_*, \pi_{-*}) - \tilde{\Psi}(\mu, \pi_*, \pi_{-*}) \qquad (2.46)\\ &\overset{(i)}{\geq} \tilde{V}_{\bar{i}^{(t)}}(\mu, \pi'_*, \pi_{-*}) - \tilde{V}_{\bar{i}^{(t)}}(\mu, \pi_*, \pi_{-*}) - \alpha \\ &\overset{(ii)}{=} \frac{1}{1-\gamma} \sum_{s\in S} d_\mu^{\pi'_*,\pi_{-*}}(s) \Big(\big(\pi'_*(s) - \pi_*(s)\big)^\top \mathbf{Q}_*(s) + \tau\nu_*(s, \pi_*) - \tau\nu_*(s, \pi'_*) \Big) - \alpha \\ &\overset{(iii)}{=} \frac{1}{1-\gamma} d_\mu^{\pi'_*,\pi_{-*}}(\bar{s}^{(t)}) \Big((\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})^\top) \cdot \mathbf{Q}_*(\bar{s}^{(t)}) + \tau\nu_*(\bar{s}^{(t)}, \pi_*) - \tau\nu_*(\bar{s}^{(t)}, \pi'_*) \Big) - \alpha, \end{aligned}$$

where (i) follows from Lemma 2.7.5, (ii) follows from Lemma 2.5.3, and (iii) holds because $\pi'_*(s) = \pi_*(s)$ for all $s \neq \bar{s}^{(t)}$. Next, from Algorithm 3, note that $\pi'_*(\bar{s}^{(t)}) = \mathrm{BR}^{(t)}_{\bar{i}^{(t)}}(\bar{s}^{(t)})$. Consequently, using Lemma 2.7.6, we obtain

$$\begin{aligned} (2.46) \geq \frac{\tau d_\mu^{\pi'_*,\pi_{-*}}(\bar{s}^{(t)})}{1-\gamma} \Big(&\log(\pi'_*(\bar{s}^{(t)}))^\top \cdot \big(\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\big) \\ &+ \nu_*(\bar{s}^{(t)}, \pi_*) - \nu_*(\bar{s}^{(t)}, \pi'_*) \Big) - \alpha. \qquad (2.47) \end{aligned}$$

Furthermore, using Lemma 2.7.7, we obtain

$$\begin{aligned} (2.47) \geq & \frac{\tau}{2(1-\gamma)} d_\mu^{\pi'_*,\pi'_{-*}(\bar{s}^{(t)})} \|\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\|_2^2 - \alpha \\ &\overset{(a)}{\geq} \frac{\tau\bar{\mu}}{2} \|\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\|_2^2 - \alpha, \end{aligned}$$

where (a) follows from $d_\mu^{\pi_i^{(t+1)},\pi_{-i}^{(t)}}(\bar{s}^{(t)}) \geq (1-\gamma)\bar{\mu}$. Summing the above inequality over all $t \in [T]$ yields:

$$\begin{aligned} &\tilde{\Psi}(\mu, \pi^{(T)}) - \tilde{\Psi}(\mu, \pi^{(0)}) = \sum_{t\in[T]} \tilde{\Psi}(\mu, \pi^{(t+1)}) - \tilde{\Psi}(\mu, \pi^{(t)}) \\ &\geq \frac{\tau\bar{\mu}}{2} \sum_{t\in[T]} \|\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\|_2^2 - \alpha T. \end{aligned}$$

Finally to conclude Lemma 2.5.5 (2), note that

$$\begin{aligned}\sum_{t\in[T]} \|\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\|_2^2 &\leq \frac{2}{\tau\bar{\mu}}\left(\tilde{\Psi}(\mu,\pi^{(T)}) - \tilde{\Psi}(\mu,\pi^{(0)}) + \alpha T\right) \\ &\leq \frac{2}{\tau\bar{\mu}}\left(|\Phi(\mu,\pi^{(T)}) - \Phi(\mu,\pi^{(0)})| + \frac{2\tau N\log(\bar{A})}{1-\gamma} + \alpha T\right),\end{aligned}$$

where the last inequality follows by noting that for any $\pi,\pi' \in \Pi$ and any $\mu \in \mathcal{P}(S)$,

$$\begin{aligned}|\tilde{\Psi}(\mu,\pi) - \tilde{\Psi}(\mu,\pi')| &\leq |\Phi(\mu,\pi) - \Phi(\mu,\pi')| \\ &\quad + \tau\left|\mathbb{E}_{\pi}\left[\sum_{\substack{j\in[N]\\ t\in\mathbb{N}}}\gamma^t\nu_j(s^t,\pi_j)\right]\right| + \tau\left|\mathbb{E}_{\pi'}\left[\sum_{\substack{j\in[N]\\ t\in\mathbb{N}}}\gamma^t\nu_j(s^t,\pi_j)\right]\right| \\ &\leq |\Phi(\mu,\pi) - \Phi(\mu,\pi')| + 2\tau\frac{N\log(\bar{A})}{1-\gamma}.\end{aligned}$$

Chapter 3

Continuous-Time α-Potential Game

3.1 Introduction

3.1.1 Overview

Static potential games, introduced by Monderer and Shapley in [113], are non-cooperative games where any player's change in utility function upon unilaterally deviating from her policy can be evaluated through the change of an auxiliary function called potential function. The introduction of the potential function is powerful as it simplifies the otherwise challenging task of finding Nash equilibria in N-player non-cooperative games to optimizing a single function. Static potential games and their variants have been a popular framework for studying N-player static games, especially with heterogeneous players.

In the dynamic setting with Markovian state transitions and Markov policies, direct generalization of the static potential game called Markov potential game is proposed in [105]. Unfortunately, most dynamic games are not Markov potential games. In fact, [101] shows that even a Markov game where the game at each state is a static potential game may not be a Markov potential game. In practice, Markov potential game framework imposes restrictive assumptions for various applied problems, such as state transitions being of distributed types for multi-agent robotics [94, 134, 135] and instantaneous reward functions being separable for resource allocation [114].

Recently, a more general form of dynamic game called Markov α-potential game is proposed by [71] (see also Chapter 2) for N-player non-cooperative Markov games with finite-state, finite-action, and discrete-time state transition. The introduction of a parameter α and an associated α-potential function enables capturing the interactions of players and their heterogeneity. They establish the existence of α-potential function for discrete-time Markov games, and show that maximizing the α-potential function yields an α-Nash equilibrium (NE). Meanwhile, they identify several important classes of dynamic α-potential

[0] This chapter is mainly based on work [74] entitled *An α-potential game framework for N-player dynamic games*, coauthored with Xin Guo (UC Berkeley) and Yufei Zhang (Imperial College London).

games. These present new potential applications, in addition to various potential games explored earlier in transportation systems [167], power networks [88], and multi-agent robotics [94, 134, 135], along with more recent studies [101, 108, 131, 163, 49, 56, 106, 114, 67].

In this chapter, we propose and study general dynamic α-potential games, including stochastic differential games with continuous state-action space, and with continuous-time state transition. Similar to the α-potential game in the discrete-time setting in [71], this general α-potential game framework reduces the challenging task of finding approximate NE in a dynamic game to a (simpler) optimization problem of minimizing a single function.

In the framework of α-potential games, there are two key mathematical questions: finding and optimizing the α-potential function, and analyzing the magnitude of α. In the discrete-time setting with finite state and finite action, these two questions have been answered in [71] by formulating a semi-infinite linear programming (SLP) problem such that its optimal solution is the α-potential function and its minimum yields the α. However, this SLP approach does not apply to continuous-time and arbitrary state-action spaces.

Instead, in this chapter, we adopt the tool of linear derivatives developed in [67] to construct the α-potential function Φ, and to characterize α in terms of the magnitude of the asymmetry of objective functions' second-order derivatives. For stochastic differential games where the state dynamic is a controlled diffusion, the α-potential function is expressed via the sensitivity processes of the controlled diffusion, and α is explicitly characterized in terms of the game structure including the number of players, the choice of strategy classes, and the intensity of interactions and the level of heterogeneity among players. To analyze the α-NE, our approach is to show that minimizing Φ is equivalent to solving a conditional McKean-Vlasov control problem: we first develop the dynamic programming principle (DPP), and then establish a verification theorem to construct a minimizer of the α-potential function Φ based on solutions to an infinite-dimensional Hamilton-Jacobi-Bellman (HJB) equation. To the best of our knowledge, this is the first result establishing a DPP for dynamic potential games. Prior to our work, the only known approach for α-NE is the policy-gradient algorithm in [71] for finite-state discrete-time α-potential games. Our approach is illustrated through a linear-quadratic network game, where the α-NE and the associated HJB equation are explicitly solved.

3.1.2 Outline of Main Results

α-potential games and approximate Nash equilibria. Consider a general N-player game $\mathcal{G}$ characterized by $\mathcal{G} = ([N], S, (\mathcal{A}_i)_{i\in[N]}, (V_i)_{i\in[N]})$,[1]where $[N] = \{1, \ldots, N\}$ is the set of players, S is the state space of the underlying dynamics, $\mathcal{A}_i$ is the set of admissible strategies of player i, and $V_i : \prod_{i\in[N]} \mathcal{A}_i \to \mathbb{R}$ is the total cost function of player i, with $V_i(\boldsymbol{a})$ being player i's expected accumulated cost if the state dynamics starts with a fixed initial state $s_0 \in S$ and all players take the strategy profile $\boldsymbol{a}$. For each $i \in [N]$, player i aims to minimize her objective function V_i over all admissible strategies in $\mathcal{A}_i$.

[1] For notational simplicity, we do not write explicitly the dependence of $\mathcal{G}$ on the fixed initial state s_0.

Here we focus on a class of games called α-potential games, where there exists $\alpha \geq 0$ and $\Phi : \mathcal{A}^{(N)} \to \mathbb{R}$ such that for all $i \in [N]$, $a_i, a_i' \in \mathcal{A}_i$ and $a_{-i} \in \mathcal{A}_{-i}^{(N)}$,

$$|V_i\left((a_i', a_{-i})\right) - V_i\left((a_i, a_{-i})\right) - \left(\Phi\left((a_i', a_{-i})\right) - \Phi\left((a_i, a_{-i})\right)\right)| \leq \alpha, \tag{3.1}$$

with $\mathcal{A}^{(N)} = \prod_{i\in[N]} \mathcal{A}_i$ the set of strategy profiles for all players, and $\mathcal{A}_{-i}^{(N)} = \prod_{j\in[N]\setminus\{i\}} \mathcal{A}_j$ the set of strategy profiles of all players except player i. Such Φ is called an α-potential function for the game $\mathcal{G}$. In the case of $\alpha = 0$, we simply call the game $\mathcal{G}$ a potential game and Φ a potential function for $\mathcal{G}$.

Equation (3.1) relaxes the notion of potential games in [113, 105] by introducing a parameter α. That is, a game $\mathcal{G}$ is an α-potential game if the change of a player's objective function upon her unilateral deviation from her strategy is equal to the change of the α-potential function up to an error α. This additional parameter α enables capturing important information regarding the interaction between players' state dynamics and strategies, beyond the number of players which has been the primary focus of approximate Nash equilibrium approach such as mean field games.

Similar to potential games, an α-potential game $\mathcal{G}$ has an important property: any minimizer of an α-potential function of $\mathcal{G}$ is an α-NE of the game $\mathcal{G}$ (Proposition 3.2.1). Proposition 3.2.1 suggests three key components in applying the α-potential game framework to analyze general non-cooperative games: constructing an α-potential function, characterizing (upper bounds of) the associated parameter α, and developing a solution technique for minimizing the α-potential function over admissible strategy sets.

Characterizing general α-potential games. We start by constructing the α-potential function and characterizing the associated parameter α for a given game $\mathcal{G}$, where all players' strategy classes are convex. Specifically, for each $i \in [N]$, denote by $\operatorname{span}(\mathcal{A}_i)$ the vector space of all linear combinations of strategies in $\mathcal{A}_i$. The concept of linear derivative of V_i with respect to $\mathcal{A}_i$, introduced in [67] for arbitrary convex strategy classes, enables us to establish Theorem 3.2.1: if the objective functions of a game $\mathcal{G}$ admit second-order linear derivatives, then under some mild regularity conditions, for any fixed $\boldsymbol{z} \in \mathcal{A}^{(N)}$, the function

$$\Phi(\boldsymbol{a}) := \int_0^1 \sum_{j=1}^N \frac{\delta V_j}{\delta a_j}\left(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_j - z_j\right) \mathrm{d}r \tag{3.2}$$

is an α-potential function of $\mathcal{G}$, with

$$\alpha \leq 2 \sup_{i\in[N], a_i' \in \mathcal{A}_i, \boldsymbol{a}, \boldsymbol{a}'' \in \mathcal{A}^{(N)}} \sum_{j=1}^N \left| \frac{\delta^2 V_i}{\delta a_i \delta a_j}\left(\boldsymbol{a}; a_i', a_j''\right) - \frac{\delta^2 V_j}{\delta a_j \delta a_i}\left(\boldsymbol{a}; a_j'', a_i'\right) \right|. \tag{3.3}$$

This characterization generalizes existing results of potential games with finite-dimensional strategy classes [113, 101, 84] to general dynamic games with arbitrary convex strategy classes. In particular, it replaces the Fréchet derivatives used in earlier works with linear derivatives, without requiring a topological structure on $\mathcal{A}^{(N)}$. Moreover, it quantifies the performance of the α-potential function (3.2) in terms of the difference between the second-order linear derivatives of the objective functions.

Constructing α-potential function for stochastic differential game. The main contribution of this chapter is to develop the criteria (3.2) and (3.3) for stochastic differential games in which the state dynamic is a controlled diffusion. Specifically, let $T \in (0, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space supporting an m-dimensional Brownian motion $W = (W^k)_{k=1}^m$, and let $\mathbb{F}$ be the natural filtration of W. Let $\mathcal{H}^2(\mathbb{R}^n)$ be the space of $\mathbb{R}^n$-valued square integrable $\mathbb{F}$-adapted processes, and for each $i \in [N]$, $\mathcal{A}_i$ be a convex subset of $\mathcal{H}^2(\mathbb{R}^n)$ representing player i's admissible controls. For each $\boldsymbol{u} \in \mathcal{A}^{(N)}$, let $\mathbf{X}^{\boldsymbol{u}}$ be the associated state process satisfying for all $i \in [N]$ and $t \in [0, T]$,

$$\mathrm{d}X_{t,i} = b_i(t, \mathbf{X}_t, \boldsymbol{u}_t)\mathrm{d}t + \sum_{k=1}^{m} \sigma_{ik}(t, \mathbf{X}_t, \boldsymbol{u}_t)\mathrm{d}W_t^k, \quad X_{0,i} = x_i, \tag{3.4}$$

where $x_i \in \mathbb{R}$ is a given initial state, $b_i : [0,T] \times \mathbb{R}^{Nd} \times \mathbb{R}^{Nn} \to \mathbb{R}^d$ and $\sigma_i = (\sigma_{i1}, \ldots, \sigma_{im}) : [0,T] \times \mathbb{R}^{Nd} \times \mathbb{R}^{Nn} \to \mathbb{R}^{d\times m}$ are given functions. The objective function $V_i : \mathcal{A}^{(N)} \to \mathbb{R}$ of player i is

$$V_i(\boldsymbol{u}) = \mathbb{E}\left[\int_0^T f_i(t, \mathbf{X}_t^{\boldsymbol{u}}, \boldsymbol{u}_t)\, \mathrm{d}t + g_i(\mathbf{X}_T^{\boldsymbol{u}})\right], \tag{3.5}$$

where $f_i : [0,T] \times \mathbb{R}^{Nd} \times \mathbb{R}^{Nn} \to \mathbb{R}$ and $g_i : \mathbb{R}^{Nd} \to \mathbb{R}$ are given functions. Precise assumptions on x_i, b_i, σ_i, f_i and g_i are given in Assumption 3.3.1.

We characterize the linear derivative of V_i and the function Φ in (3.2) through the sensitivity processes of the state process with respect to controls (Theorem 3.3.1). In particular, assuming $0 \in \mathcal{A}^{(N)}$, the function Φ (with $\boldsymbol{z} = 0$) can be expressed as

$$\Phi(\boldsymbol{u}) = \int_0^1 \sum_{i=1}^{N} \mathbb{E}\left[\int_0^T \begin{pmatrix} \mathbf{Y}_t^{r\boldsymbol{u},u_i} \\ u_{t,i} \end{pmatrix}^{\top} \begin{pmatrix} \partial_x f_i \\ \partial_{u_i} f_i \end{pmatrix} (t, \mathbf{X}_t^{r\boldsymbol{u}}, r\boldsymbol{u}_t)\, \mathrm{d}t + (\partial_x g_i)^{\top}(\mathbf{X}_T^{r\boldsymbol{u}})\mathbf{Y}_T^{r\boldsymbol{u},u_i}\right]\mathrm{d}r, \tag{3.6}$$

where for each $\boldsymbol{u} \in \mathcal{A}^{(N)}$ and $u_i' \in \mathcal{H}^2(\mathbb{R}^n)$, the sensitivity process $\mathbf{Y}^{\boldsymbol{u},u_i'}$ is the derivative (in the L^2 sense) of the state $\mathbf{X}^{\boldsymbol{u}}$ when player i varies her control in the direction u_i', and satisfies a controlled linear stochastic differential equation (as in (3.16)). See Theorem 3.3.1 for the expression of Φ with general $\mathcal{A}^{(N)}$ for open-loop controls and see Theorem 3.4.1 for the expression of Φ for closed-loop controls.

Quantifying α for stochastic differential game. Using the bound (3.3), we then quantify the parameter α for the game (3.4)-(3.5) based on the structure of the game. A key technical step is to characterize and estimate the second-order linear derivative of V_i through second-order sensitivity processes, representing the derivative of $\mathbf{Y}^{\boldsymbol{u},u_i'}$. Under suitable structural assumptions on the coefficients of (3.4), we establish precise estimates for these sensitivity processes, and obtain the following upper bound (stated more precisely in Theorem 3.3.2): for all $i, j \in [N]$,

$$\left| \frac{\delta^2 V_i}{\delta u_i \delta u_j} \left(\boldsymbol{u}; u_i', u_j''\right) - \frac{\delta^2 V_j}{\delta u_j \delta u_i} \left(\boldsymbol{u}; u_j'', u_i'\right) \right| \leq C C_{f,g,N}, \tag{3.7}$$

where $C \geq 0$ is a constant depending only on the state coefficients and time horizon, and $C_{f,g,N}$ is a constant depending explicitly on the number of players N and the sup-norms of the partial derivatives of $f_i - f_j$ and $g_i - g_j$.

This analysis of α shows its general dependence on game characteristics, including possibly asymmetric and heterogeneous forms of cost functions and state dynamics, and is not limited to the scale of N as in the mean-field paradigm. To highlight this distinction, we specialize the above bound (3.7) of α to two classes of stochastic games:

- For distributed games where players only interact through their cost functions and not the state and control processes, we prove that if a static potential function can be derived from the cost functions, then dynamics games are potential games (with $\alpha = 0$), regardless of the number of players (Example 3.3.2).

- For games with mean field type interactions, if each player's dependence on others' states and actions is solely through her empirical measures, then α is of the magnitude $\mathcal{O}(1/N)$ as $N \to \infty$ (Example 3.3.3). Note that α decays to zero as the number of players increases, even with heterogeneity in state dynamics, in cost functions, and in admissible strategy classes. This is in contrast to the classical mean field games with homogeneous players; see Remark 3.3.3 for a more detailed comparison.

α-NE via a McKean-Vlasov control problem. We further develop a dynamic programming approach to minimize the function Φ over $\mathcal{A}^{(N)}$. The main difficulty is that the objective (3.6) depends on the aggregated behavior of the state and sensitivity processes with respect to $r \in [0, 1]$, which acts as an additional noise independent of the Brownian motion W. Meanwhile, the admissible controls in $\mathcal{A}^{(N)}$ are adapted to a smaller filtration $\mathbb{F}$ that depends only on W. To recover the dynamic programming principle, we embed the optimization problem into a conditional McKean–Vlasov control problem. This is achieved by treating r in (3.6) as a uniform random variable $\mathfrak{r}$ independent of W, and expressing the objective $\Phi(\boldsymbol{u})$ in terms of $\boldsymbol{u}$ and the conditional law of $(\mathbf{X}^{\mathfrak{r}\boldsymbol{u}}, \mathbf{Y}^{\mathfrak{r}\boldsymbol{u},u_1}, \ldots, \mathbf{Y}^{\mathfrak{r}\boldsymbol{u},u_N}, \mathfrak{r})$ given W. This approach allows us to embed the minimization of $\boldsymbol{u} \mapsto \Phi(\boldsymbol{u})$ into a control problem, where the state space is a subset of the Wasserstein space of probability measures (Proposition 3.5.1). Moreover, by Itô's formula along a flow of conditional measures, we establish a verification theorem to construct a minimizer of Φ based on solutions to an infinite-dimensional Hamilton-Jacobi-Bellman (HJB) equation (Theorem 3.5.1).

A toy example of dynamic games on graph. We illustrate our results through a simple linear-quadratic game on a undirected graph, whose the vertices represent players, and edges indicate dependencies between them. We show that this game is an α-potential game, and characterize α explicitly in terms of N the number of players, and q_{ij} the strength and the degree of heterogeneous interaction between players i and j (see Section 3.6.1). We further construct an α-NE of the game analytically through a system of ordinary differential equations (Theorem 3.6.1). This is accomplished by solving the associated HJB equation for the α-potential function (3.6) by utilizing the game's linear-quadratic structure.

The asymptotic limit derived from our analysis allows for general asymmetric interactions and heterogeneity among players, in contrast to existing works on the mean field approximation for both differential games (e.g., [99, 100, 32]) and games on graphs (e.g., [60, 97, 15]). It shows that the α-potential game framework enables differentiating game characteristics and player interactions which are hard to quantify under previous more restrictive game frameworks, such as Markov potential games [105, 101, 108, 131, 49] and near potential games [24, 27, 146, 145].

3.2 Analytical Framework for General α-Potential Games

3.2.1 α-Potential Games and Approximate Nash Equilibria

This section introduces the mathematical framework for α-potential games, starting by some basic notions for the game and associated strategies.

Consider a game $\mathcal{G} = ([N], S, (\mathcal{A}_i)_{i\in[N]}, (V_i)_{i\in[N]})$ defined as follows: $[N] = \{1, \dots, N\}$, $N \in \mathbb{N}$, is a finite set of players, S is a set representing the state space of the underlying dynamics, $\mathcal{A}_i$ is a subset of a real vector space representing all admissible strategies of player i, and $\mathcal{A}^{(N)} = \prod_{i\in[N]} \mathcal{A}_i$ is the set of strategy profiles for all players. For each $i \in [N]$, $V_i : \mathcal{A}^{(N)} \to \mathbb{R}$ is the objective function of player i, where $V_i(\boldsymbol{a})$ is player i's expected cost if the state dynamics starts with a fixed initial state $s_0 \in S$ and all players take the strategy profile $\boldsymbol{a} \in \mathcal{A}^{(N)}$. For any $i \in [N]$, player i aims to minimize the objective function V_i over all admissible strategies in $\mathcal{A}_i$. We denote by $\mathcal{A}_{-i}^{(N)} = \prod_{j\in[N]\setminus\{i\}} \mathcal{A}_j$ the set of strategy profiles of all players except player i, and by $\boldsymbol{a}$ and a_{-i} a generic element of $\mathcal{A}^{(N)}$ and $\mathcal{A}_{-i}^{(N)}$, respectively.

Note that this game framework includes static games, and discrete-time and continuous-time dynamic games. Moreover, depending on the precise definitions of strategy classes, this framework also accommodates stochastic differential games with either open-loop controls in Section 3.3.1 or closed-loop controls in Section 3.4. The focus of this chapter is on a class of games $\mathcal{G}$ called α-potential games, defined as follows.

Definition 3.2.1 (α-potential game)**.** *Given a game $\mathcal{G} = ([N], S, (\mathcal{A}_i)_{i\in[N]}, (V_i)_{i\in[N]})$, if there exists $\alpha \geq 0$ and $\Phi : \mathcal{A}^{(N)} \to \mathbb{R}$ such that for all $i \in [N]$, $a_i, a_i' \in \mathcal{A}_i$ and $a_{-i} \in \mathcal{A}_{-i}^{(N)}$,*

$$|V_i\left((a_i', a_{-i})\right) - V_i\left((a_i, a_{-i})\right) - \left(\Phi\left((a_i', a_{-i})\right) - \Phi\left((a_i, a_{-i})\right)\right)| \leq \alpha, \tag{3.8}$$

then we say $\mathcal{G}$ is an α-potential game, and Φ is an α-potential function for $\mathcal{G}$. In the case where $\alpha = 0$, we simply call the game $\mathcal{G}$ a potential game and Φ a potential function for $\mathcal{G}$.

Intuitively, a game $\mathcal{G}$ is an α-potential game if there exists an α-potential function such that whenever one player unilaterally deviates from her strategy, the change of that player's objective function is equal to the change of the α-potential function up to an error α. This definition generalizes the notion of potential games in [113] by allowing for a positive α. Such a relaxation is essential for dynamic games, as many dynamic games that are not potential games are, in fact, α-potential games for some $\alpha > 0$; see [71] and also Sections 3.3.1. Indeed, it is clear that if $\hat{\alpha} := \sup_{i\in[N],\boldsymbol{a}\in\mathcal{A}^{(N)}} |V_i(\boldsymbol{a})| < \infty$, then $\mathcal{G}$ is a $2\hat{\alpha}$-potential game and a $2\hat{\alpha}$-potential function $\Phi = 0$.

For a given game $\mathcal{G}$, there can be multiple parameters α satisfying the condition (3.8). In [71], an α-potential game is defined with the optimal α determined by

$$\alpha^* = \inf_{\Phi\in\mathscr{F}} \sup_{\substack{i\in[N],a_i,a_i'\in\mathcal{A}_i,\\ a_{-i}\in\mathcal{A}_{-i}^{(N)}}} |V_i\left((a_i', a_{-i})\right) - V_i\left((a_i, a_{-i})\right) - \left(\Phi\left((a_i', a_{-i})\right) - \Phi\left((a_i, a_{-i})\right)\right)|, \tag{3.9}$$

where $\mathscr{F}$ contains suitable functions $\Phi : \mathcal{A}^{(N)} \to \mathbb{R}$. For discrete games with finite states and actions, [71] shows that selecting $\mathscr{F}$ as the set of uniformly equi-continuous functions on Markov policies ensures a well-defined α^* and also the existence of an α^*-potential function within $\mathscr{F}$.

However, in continuous-time games with continuous state and action spaces, computing the optimal α^* in (3.9) is challenging, and selecting a suitable set $\mathscr{F}$ for the existence of an α^*-potential function remains unclear. Most critically, as shown in [71], having an appropriate upper bound α of α^* and an associated α-potential function Φ is sufficient for the key analysis. Therefore, we adopt Definition 3.2.1, which frees us to focus on characterizing some upper bound of α^* in terms of the number of players, the set of admissible strategies, and the game structure.

For an α-potential game, computing an approximate Nash equilibrium reduces to an optimization problem. To see it, we first recall the solution concept of ε-Nash equilibrium.

Definition 3.2.2. *For any $\varepsilon \geq 0$, a strategy profile $\boldsymbol{a} = (a_i)_{i\in[N]} \in \mathcal{A}^{(N)}$ is an ε-Nash equilibrium of the game $\mathcal{G}$ if $V_i\left((a_i, a_{-i})\right) \leq V_i\left((a_i', a_{-i})\right) + \varepsilon$, for any $i \in [N], a_i' \in \mathcal{A}_i$.*

Definition 3.2.2 provides a unified definition of approximate Nash equilibrium for a general game $\mathcal{G}$. When $\mathcal{G}$ is a stochastic differential game and the set $\mathcal{A}^{(N)}$ of admissible strategy profiles contains the set of open-loop controls or closed-loop controls, Definition 3.2.2 is consistent with the concepts of open-loop Nash equilibrium or closed-loop Nash equilibrium described in [32, Chapter 2].

The following proposition shows that an approximate Nash equilibrium of an α-potential game can be obtained by optimizing its corresponding α-potential function. This is analogous to static potential games with potential functions. The proof follows directly from Definitions 3.2.1 and 3.2.2 and hence is omitted.

Proposition 3.2.1. *Let $\mathcal{G}$ be an α-potential game for some α and Φ be an α-potential function. For each $\varepsilon \geq 0$, if there exists $\overline{\boldsymbol{a}} \in \mathcal{A}^{(N)}$ such that $\Phi(\overline{\boldsymbol{a}}) \leq \inf_{\boldsymbol{a}\in\mathcal{A}^{(N)}} \Phi(\boldsymbol{a}) + \varepsilon$, then $\overline{\boldsymbol{a}}$ is an $(\alpha+\varepsilon)$-Nash equilibrium of $\mathcal{G}$.*

3.2.2 Characterization of α-Potential Games via Linear Derivatives

Proposition 3.2.1 highlights the importance of explicitly characterizing an α-potential function for a given game and the parameter α.

For the special class of potential games (i.e., $\alpha = 0$) with finite-dimensional strategy class [113, 101, 84], it is well known that a game is a potential game if the objective functions are twice continuously (Fréchet) differentiable in policy parameters and have symmetric second-order derivatives. More precisely, consider a game $\mathcal{G} = ([N], (\mathcal{A}_i)_{i\in[N]}, (V_i)_{i\in[N]})$ where for all $i \in [N]$, $\mathcal{A}_i$ is an interval. Suppose that for all $i \in [N]$, $V_i : \mathcal{A}^{(N)} \to \mathbb{R}$ is twice continuously differentiable. Then by [113, Theorem 4.5], $\mathcal{G}$ is a potential game if and only if $\partial^2_{a_i a_j} V_i = \partial^2_{a_j a_i} V_j$ for all $i, j \in [N]$, and a form of potential function is given.

In this section, we will provide an analytical framework to construct the parameter α and the associated α-potential functions based on linear derivatives of the objective functions with respect to strategies as introduced in [67]. Let us start by recalling the linear derivative of a scalar-valued function with respect to unilateral deviations of strategies. For each $i \in [N]$, we denote by $\operatorname{span}(\mathcal{A}_i)$ the vector space of all linear combinations of strategies in $\mathcal{A}_i$, i.e., $\operatorname{span}(\mathcal{A}_i) = \left\{ \sum_{\ell=1}^m c_\ell a_i^{(\ell)} \mid c_\ell \in \mathbb{R}, a_i^{(\ell)} \in \mathcal{A}_i, \text{ for any } l = 1, 2, \cdots, m, \text{ and } m \in \mathbb{N} \right\}$.

Definition 3.2.3. *Let $\mathcal{A}^{(N)} = \prod_{i\in[N]} \mathcal{A}_i$ be a convex set and $f : \mathcal{A}^{(N)} \to \mathbb{R}$. For each $i \in [N]$, we say f has a linear derivative with respect to $\mathcal{A}_i$, if there exists $\frac{\delta f}{\delta a_i} : \mathcal{A}^{(N)} \times \operatorname{span}(\mathcal{A}_i) \to \mathbb{R}$, such that for all $\boldsymbol{a} = (a_i, a_{-i}) \in \mathcal{A}^{(N)}$, $\frac{\delta f}{\delta a_i}(\boldsymbol{a}; \cdot)$ is linear and*

$$\lim_{\varepsilon \searrow 0} \frac{f\left(\left(a_i + \varepsilon\left(a_i' - a_i\right), a_{-i}\right)\right) - f(\boldsymbol{a})}{\varepsilon} = \frac{\delta f}{\delta a_i}\left(\boldsymbol{a}; a_i' - a_i\right), \quad \forall a_i' \in \mathcal{A}_i. \tag{3.10}$$

Moreover, for each $i, j \in [N]$, we say f has second-order linear derivatives with respect to $\mathcal{A}_i \times \mathcal{A}_j$, if (i) for all $k \in \{i, j\}$, f has a linear derivative $\frac{\delta f}{\delta a_k}$ with respect to $\mathcal{A}_k$, and (ii) for all $(k, \ell) \in \{(i, j), (j, i)\}$, there exists $\frac{\delta^2 f}{\delta a_k \delta a_\ell} : \mathcal{A}^{(N)} \times \operatorname{span}(\mathcal{A}_k) \times \operatorname{span}(\mathcal{A}_\ell) \to \mathbb{R}$ such that for all $\boldsymbol{a} \in \mathcal{A}^{(N)}$, $\frac{\delta^2 f}{\delta a_k \delta a_\ell}(\boldsymbol{a}, \cdot, \cdot)$ is bilinear and for all $a_k' \in \operatorname{span}(\mathcal{A}_k)$, $\frac{\delta^2 f}{\delta a_k \delta a_\ell}(\cdot; a_k', \cdot)$ is a linear derivative of $\frac{\delta f}{\delta a_k}(\cdot; a_k')$ with respect to $\mathcal{A}_\ell$. We refer to $\frac{\delta^2 f}{\delta a_i \delta a_j}$ and $\frac{\delta^2 f}{\delta a_j \delta a_i}$ as second-order linear derivatives of f with respect to $\mathcal{A}_i \times \mathcal{A}_j$.

Remark 3.2.1. *Linear differentiability, as defined in Definition 3.2.3, is weaker than Fréchet or Gâteaux differentiability, as it avoids introducing a topology on the strategy classes* $\mathcal{A}_i$*.*

Recall that a function $f : O \subset X \to \mathbb{R}$ *defined on an* open subset O *of a locally convex topological vector space* X *is Gâteaux differentiable if, for all* $u \in V$, $Df(u; v) = \lim_{\varepsilon\to 0} \frac{f(u+\varepsilon v)-f(u)}{\varepsilon}$ *exists for all* $v \in V$*. If in addition* $(X, \|\cdot\|_X)$ *is a normed vector space,* $X \ni v \mapsto Df(u; v) \in \mathbb{R}$ *is a bounded linear operator, and* $\lim_{\|v\|_X\to 0} \frac{|f(u+v)-f(u)-Df(u;v)|}{\|v\|_X} = 0$*, then* f *is Fréchet differentiable. Note that both Fréchet and Gâteaux derivatives are defined only in the interior of a set* O*, as their definitions require that* $u + \varepsilon v$ *remains within the domain* O *for all sufficiently small* ε*. This necessitates a topological structure on* O*.*

Definition 3.2.3 defines derivatives using convex combinations within the strategy class, without the need of a topology. Therefore, it can be applied to analyze games with any convex strategy class. Moreover, if f *has a Gâteaux derivative* Df *with respect to* $\mathcal{A}_i$*, then* f *also has a linear derivative given by* $\frac{\delta f}{\delta a_i}(\boldsymbol{a}; a_i' - a_i) = Df(\boldsymbol{a}; a_i' - a_i)$*.*

Note that Definition 3.2.3 generalizes the notion of linear derivative for functions of Markov policies introduced in [67] to functions defined on arbitrary convex strategy classes. It enables us to construct an α-potential function for a game $\mathcal{G}$ using the linear derivative of its objective functions, with α bounded by the difference between the second-order linear derivatives of the objective functions.

Theorem 3.2.1. *Let* $\mathcal{G}$ *be a game whose set of strategy profiles* $\mathcal{A}^{(N)}$ *is convex. Suppose that for all* $i, j \in [N]$*, the objective function* V_i *has second-order linear derivatives with respect to* $\mathcal{A}_i \times \mathcal{A}_j$ *such that for all* $\boldsymbol{z} = (z_j)_{j\in[N]} \in \mathcal{A}^{(N)}$, $\boldsymbol{a} = (a_j)_{j\in[N]} \in \mathcal{A}^{(N)}$, $a_i', \tilde{a}_i' \in \mathcal{A}_i$ *and* $a_j'' \in \mathcal{A}_j$*,*

(1) $\sup_{r,\varepsilon\in[0,1]} \left| \frac{\delta^2 V_i}{\delta a_i \delta a_j} \left(\boldsymbol{z} + r\left(\boldsymbol{a}^\varepsilon - \boldsymbol{z}\right); a_i', a_j''\right) \right| < \infty$*, where* $\boldsymbol{a}^\varepsilon := \left(a_i + \varepsilon\left(\tilde{a}_i' - a_i\right), a_{-i}\right)$*;*

(2) $[0, 1]^N \ni \varepsilon \mapsto \frac{\delta^2 V_i}{\delta a_i a_j} \left(\boldsymbol{z} + \varepsilon \cdot (\boldsymbol{a} - \boldsymbol{z}); a_i', a_j''\right)$ *is continuous at 0 , where* $\boldsymbol{z} + \varepsilon \cdot (\boldsymbol{a} - \boldsymbol{z}) := \left(z_i + \varepsilon_i\left(a_i - z_i\right)\right)_{i\in[N]}$*.*

Fix $\boldsymbol{z} \in \mathcal{A}^{(N)}$ *and define* $\Phi : \mathcal{A}^{(N)} \to \mathbb{R}$ *by*

$$\Phi(\boldsymbol{a}) = \int_0^1 \sum_{j=1}^N \frac{\delta V_j}{\delta a_j} \left(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_j - z_j\right) \mathrm{d}r. \tag{3.11}$$

Then Φ *is an* α*-potential function of* $\mathcal{G}$ *with*

$$\alpha \leq 2 \sup_{i\in[N], a_i'\in\mathcal{A}_i, \boldsymbol{a}, \boldsymbol{a}''\in\mathcal{A}^{(N)}} \sum_{j=1}^N \left| \frac{\delta^2 V_i}{\delta a_i \delta a_j} \left(\boldsymbol{a}; a_i', a_j''\right) - \frac{\delta^2 V_j}{\delta a_j \delta a_i} \left(\boldsymbol{a}; a_j'', a_i'\right) \right|. \tag{3.12}$$

Theorem 3.2.1 constructs an α-potential function using the linear derivatives of objective functions, which exist for general strategy classes without requiring a topological structure. The corresponding α is quantified explicitly in terms of the magnitude of the asymmetry of the second-order linear derivatives, and $\alpha = 0$ recovers the symmetric case in these earlier works [113, 101, 84]. The base-point action $\boldsymbol{z}$ ensures that $\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z})$ remains in $\mathcal{A}^{(N)}$, so that the linear derivatives are well-defined. The specific choice of $\boldsymbol{z}$ will not change the upper bound of α, as (3.12) takes the supremum over all strategies, but it may lead to different minima of α-potential functions Φ. The proof of Theorem 3.2.1 is given in Section 3.7.1.

The next proposition shows that under sufficient regularity, the partial derivative of V_i and Φ are close given the α-potential function in (3.11):

Proposition 3.2.2. *Suppose that for all $i \in [N]$, $\mathcal{A}_i$ is an open subset of an inner product space, $C := \sup_{j\in[N], a_j \in \mathcal{A}_j} \|a_j\| < \infty$, and the objective function V_i is twice continuously Fréchet differentiable in $\mathcal{A}_i$. Then for any $i \in [N]$, $a_i, a_i' \in \mathcal{A}_i$, and $a_{-i} \in \mathcal{A}_{-i}$,*

$$|V_i((a_i', a_{-i})) - V_i((a_i, a_{-i})) - (\Phi((a_i', a_{-i})) - \Phi((a_i, a_{-i})))| \le \tilde{\alpha}\|a_i' - a_i\|,$$

with $\tilde{\alpha} \le C \sup_{i,j\in[N], \boldsymbol{a}\in\mathcal{A}} \sum_{j=1}^N \left\| \partial^2_{a_i a_j} V_i(\boldsymbol{a}) - \partial^2_{a_j a_i} V_j(\boldsymbol{a}) \right\|$.

The definition of α-potential game in (3.8) does not necessarily guarantee that the partial derivative of V_i and the partial derivative of Φ are close. However, with Φ construct in (3.11), Proposition 3.2.2 shows under sufficient regularity, the derivative of V_i and the derivative of Φ are close, up to an error bounded by $\tilde{\alpha}$, and $\tilde{\alpha}$ is determined by the asymmetry in the second-order derivatives of the value functions.

Existing literature in multi-agent reinforcement learning (MARL) on Markov potential games (MPG) replies on the fact that the partial derivatives $\partial_{a_i} V_i$ and $\partial_{a_i}\Phi$ are identical in an MPG. In α-potential games, Proposition 3.2.2 demonstrates that these partial derivatives are close to each other, which can facilitate the regret analysis by allowing one to leverage tools developed for MPGs while accounting for the approximation error introduced by the discrepancy in first-order derivatives.

The α-potential function (3.11) involves aggregating all players' strategies and the derivatives of their objective functions linearly through the parameter r. When the objective functions are sufficiently regular, analogue α-potential functions can be constructed through nonlinear aggregation of all players' strategies. Indeed, we have

Proposition 3.2.3. *Suppose that for all $i \in [N]$, $\mathcal{A}_i$ is an open subset of an inner product space, and the objective function V_i is continuously Fréchet differentiable in $\mathcal{A}_i$. Fix $\boldsymbol{z} \in \mathcal{A}^{(N)}$, and for all $i \in [N]$, let $p_i : [0,1] \times \mathcal{A}_i \to \mathcal{A}_i$ be a continuously differentiable reparameterization of $\mathcal{A}_i$ such that for all $a_i \in \mathcal{A}_i$, $p_i(0, a_i) = z_i$ and $p_i(1, a_i) = a_i$. Then one can define*

$$\Phi(\boldsymbol{a}) = \int_0^1 \sum_{i=1}^N \langle \partial_{a_i} V_i(p(r, \boldsymbol{a})), \partial_r p_i(r, a_i) \rangle \, \mathrm{d}r, \tag{3.13}$$

where $p(r, \boldsymbol{a}) := (p_i(r, a_i))_{i\in[N]}$, and $\partial_{a_i} V_i$ is the Fréchet derivative of V_i.

Consequently, if we assume further regularity of $(V_i)_{i\in[N]}$, the corresponding α for (3.13) can be quantified in terms of the asymmetry in second-order derivatives of objective functions and the derivatives of the parameterization p as in Theorem 3.2.1.

It is worth noting that this α-potential function (3.13) extends the characterization of potential functions for static games with finite-dimensional strategy spaces as established in [113, Theorem 4.5], and coincides the expression (3.11) by setting $\frac{\delta V_i}{\delta a_i}(\boldsymbol{a}; a_i') = \langle \partial_{a_i} V_i(\boldsymbol{a}), a_i' \rangle$ and $p_i(r, a_i) = z_i + r(a_i - z_i)$. When the game $\mathcal{G}$ is a potential game (i.e., $\alpha = 0$), any potential function is given by (3.13) (or (3.11)) up to an additive constant, as all potential functions share the same gradient and are therefore equivalent up to a constant.

For ease of exposition and clarity, we focus on the α-potential function given in (3.11) in the subsequent analysis. As we will see, for stochastic differential games, the adoption of linear derivatives in (3.11) simplifies the analysis and avoids the tedious verification of the Fréchet differentiability of $(V_i)_{i\in[N]}$. Moreover, minimizing (3.11) will be shown as a class of conditional McKean-Vlasov control problem.

3.3 Open-Loop Stochastic Differential Games

This section characterizes α-potential function (3.11) given in Theorem 3.2.1 for stochastic differential games whose state dynamics is a controlled diffusion with open-loop controls. Under suitable regularity conditions, the linear derivative of objective functions are characterized through the sensitivity processes of the state dynamics with respect to controls.

Let $T \in (0, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space on which an m-dimensional Brownian motion $W = (W_t)_{t\geq 0}$ is defined, and let $\mathbb{F}$ be the $\mathbb{P}$-completion of the filtration generated by W. For each $p \geq 1$ and Euclidean space $(E, |\cdot|)$, let $\mathcal{S}^p(E)$ be the space of E-valued $\mathbb{F}$-progressively measurable processes $X : \Omega \times [0,T] \to E$ satisfying $\|X\|_{\mathcal{S}^p(E)} = \mathbb{E}[\sup_{s\in[0,T]} |X_s|^p]^{1/p} < \infty$, and let $\mathcal{H}^p(E)$ be the space of E-valued $\mathbb{F}$-progressively measurable processes $X : \Omega \times [0,T] \to E$ satisfying $\|X\|_{\mathcal{H}^p(E)} = \mathbb{E}[\int_0^T |X_s|^p \mathrm{d}s]^{1/p} < \infty$. With a slight abuse of notation, for any $m, n \in \mathbb{N}$, we identify the product spaces $\mathcal{S}^p(\mathbb{R}^n)^m$ and $\mathcal{H}^p(\mathbb{R}^n)^m$ with $\mathcal{S}^p(\mathbb{R}^{mn})$ and $\mathcal{H}^p(\mathbb{R}^{mn})$, respectively.

Consider the open-loop differential game $\mathcal{G}^{\mathsf{op}}$ defined as follows: let $[N] = \{1, \dots, N\}$, and for each $i \in [N]$, let $A_i \subset \mathbb{R}^n$ be a convex set, and let $\mathcal{A}^i$ be the set of processes $u_i \in \mathcal{H}^2(\mathbb{R}^n)$ taking values in A_i, representing the set of admissible (open-loop) controls of player i. For each $\boldsymbol{u} = (u_i)_{i\in[N]} \in \mathcal{H}^2(\mathbb{R}^{Nn})$, let $\mathbf{X}^{\boldsymbol{u}} = (X_i^{\boldsymbol{u}})_{i=1}^N$ be the associated state process governed by the following dynamics: for all $i \in [N]$ and $t \in [0,T]$,

$$\mathrm{d}X_{t,i} = b_i(t, \mathbf{X}_t, \boldsymbol{u}_t)\mathrm{d}t + \sum_{k=1}^m \sigma_{ik}(t, \mathbf{X}_t, \boldsymbol{u}_t)\mathrm{d}W_t^k, \quad X_{0,i} = x_i, \tag{3.14}$$

where $x_i \in \mathbb{R}^d$ is a given initial state, $b_i : [0,T] \times \mathbb{R}^{Nd} \times \mathbb{R}^{Nn} \to \mathbb{R}^d$ and $\sigma_i = (\sigma_{i1}, \dots, \sigma_{im}) : [0,T] \times \mathbb{R}^{Nd} \times \mathbb{R}^{Nn} \to \mathbb{R}^{d\times m}$ are given measurable functions, and $W = (W^k)_{k=1}^m$ is an m-dimensional $\mathbb{F}$-Brownian motion on the space $(\Omega, \mathcal{F}, \mathbb{P})$. The objective function $V_i : \mathcal{A}^{(N)} \subset \mathcal{H}^2(\mathbb{R}^{Nn}) \to \mathbb{R}$ of player i is given by

$$V_i(\boldsymbol{u}) = \mathbb{E}\left[\int_0^T f_i(t, \mathbf{X}_t^{\boldsymbol{u}}, \boldsymbol{u}_t)\,\mathrm{d}t + g_i(\mathbf{X}_T^{\boldsymbol{u}})\right], \tag{3.15}$$

where $f_i : [0,T]\times\mathbb{R}^{Nd}\times\mathbb{R}^{Nn}\to\mathbb{R}$ and $g_i:\mathbb{R}^{Nd}\to\mathbb{R}$ are given measurable functions. Player i aims to minimize (3.15) over all admissible controls in $\mathcal{A}_i$.

We impose the following regularity condition on the coefficients of (3.14)-(3.15). It guarantees for each $\boldsymbol{u}\in\mathcal{H}^2(\mathbb{R}^{Nn})$, (3.14) admits a unique strong solution $\mathbf{X}^{\boldsymbol{u}}\in\mathcal{S}^2(\mathbb{R}^{Nd})$, and (3.15) is well-defined.

Assumption 3.3.1. *For all $i\in[N]$, A_i is a nonempty convex subset of $\mathbb{R}^n$.*

(1) For all $t\in[0,T]$, $(x,u)\mapsto(b_i(t,x,u),\sigma_i(t,x,u),f_i(t,x,u),g_i(x))$ is twice continuously differentiable.

(2) For all $\varphi\in\{b_i,\sigma_i\}$, $\sup_{t\in[0,T]}|\varphi(t,0,0)|<\infty$, and $(x,u)\mapsto\varphi(t,x,u)$ has bounded first and second derivatives (uniformly in t).

(3) $\sup_{t\in[0,T]}(|f_i(t,0,0)|+|(\partial_{(x,u)}f_i)(t,0,0)|)<\infty$, and $(x,u)\mapsto(f_i(t,x,u),g_i(x))$ has bounded second derivatives (uniformly in t).

We proceed to characterize the α-potential function (3.11) for the game $\mathcal{G}^{\mathsf{op}}$. This is achieved by expressing the linear derivatives of the objective function (3.15) using the sensitivity processes of the state dynamics (3.14). In this following, we present only the first-order linear derivatives, as these are sufficient to characterize the α-potential function. The second-order linear derivatives are given in Section 3.3.1, which will be used to quantify the constant α defined in (3.12).

We start by introducing the sensitivity of the controlled state with respect to a single player's control. For each $\boldsymbol{u}\in\mathcal{H}^2(\mathbb{R}^{Nn})$, let $\mathbf{X}^{\boldsymbol{u}}$ be the state process satisfying (3.14). For each $h\in[N]$ and $u'_h\in\mathcal{H}^2(\mathbb{R}^n)$, define $\mathbf{Y}^{\boldsymbol{u},u'_h}\in\mathcal{S}^2(\mathbb{R}^{Nd})$ as the solution of the following dynamics: for all $t\in[0,T]$ and $i\in[N]$,

$$\begin{aligned}\mathrm{d}Y^h_{t,i}&=\left((\partial_x b_i)(t,\mathbf{X}^{\boldsymbol{u}}_t,\boldsymbol{u}_t)\,\mathbf{Y}^h_t+(\partial_{u_h}b_i)(t,\mathbf{X}^{\boldsymbol{u}}_t,\boldsymbol{u}_t)\,u'_{t,h}\right)\mathrm{d}t\\&\quad+\sum_{k=1}^m\left((\partial_x\sigma_{ik})(t,\mathbf{X}^{\boldsymbol{u}}_t,\boldsymbol{u}_t)\,\mathbf{Y}^h_t+(\partial_{u_h}\sigma_{ik})(t,\mathbf{X}^{\boldsymbol{u}}_t,\boldsymbol{u}_t)\,u'_{t,h}\right)\mathrm{d}W^k_t,\quad Y^h_{0,i}=0.\end{aligned}\tag{3.16}$$

By [31, Lemma 4.7], for all $u'_h\in\mathcal{H}^2(\mathbb{R}^n)$, $\lim_{\varepsilon\searrow0}\mathbb{E}\Big[\sup_{t\in[0,T]}\Big|\frac{1}{\varepsilon}(\mathbf{X}^{\boldsymbol{u}^\varepsilon}_t-\mathbf{X}^{\boldsymbol{u}}_t)-\mathbf{Y}^{\boldsymbol{u},u'_h}_t\Big|^2\Big]=0$, where $\boldsymbol{u}^\varepsilon=(u_h+\varepsilon u'_h,u_{-h})$ for all $\varepsilon\in(0,1)$. That is, in the L^2 sense, $\mathbf{Y}^{\boldsymbol{u},u'_h}$ is the derivative of the controlled state $\mathbf{X}^{\boldsymbol{u}}$ when player h varies her control in the direction u'_h.

Now, the linear derivatives of V_i in (3.15) can be represented using the sensitivity processes given by (3.16). Indeed, for all $i,h\in[N]$, define the map $\frac{\delta V_i}{\delta u_h}:\mathcal{A}^{(N)}\times\mathcal{H}^2(\mathbb{R}^n)\to\mathbb{R}$ such that for all $\boldsymbol{u}\in\mathcal{A}^{(N)}$ and $u'_h\in\mathcal{H}^2(\mathbb{R}^n)$,

$$\frac{\delta V_i}{\delta u_h}(\boldsymbol{u};u'_h):=\mathbb{E}\left[\int_0^T\begin{pmatrix}\mathbf{Y}^{\boldsymbol{u},u'_h}_t\\u'_{t,h}\end{pmatrix}^\top\begin{pmatrix}\partial_x f_i\\\partial_{u_h}f_i\end{pmatrix}(t,\mathbf{X}^{\boldsymbol{u}}_t,\boldsymbol{u}_t)\,\mathrm{d}t+(\partial_x g_i)^\top(\mathbf{X}^{\boldsymbol{u}}_T)\mathbf{Y}^{\boldsymbol{u},u'_h}_T\right].\tag{3.17}$$

By the convexity of $\mathcal{A}_h$ and [31, Lemma 4.8], $\lim_{\varepsilon\searrow 0}\frac{V_i(\boldsymbol{u}^\varepsilon)-V_i(\boldsymbol{u})}{\varepsilon}=\frac{\delta V_i}{\delta u_i}(\boldsymbol{u};u_h'-u_h)$, for all $u_h'\in\mathcal{A}_h$, where $\boldsymbol{u}^\varepsilon=(u_h+\varepsilon(u_h'-u_h),u_{-h})$ for all $\varepsilon\in(0,1)$. That is, $\frac{\delta V_i}{\delta u_h}$ is the linear derivative of V_i with respect to $\mathcal{A}_h$.

Using the expression (3.17) of $(\frac{\delta V_i}{\delta u_i})_{i\in[N]}$, the following theorem characterizes the α-potential function for the open-loop differential game $\mathcal{G}^{\mathsf{op}}$.

Theorem 3.3.1. *Consider the game $\mathcal{G}^{op}$ defined by (3.14)-(3.15). Suppose Assumption 3.3.1 holds. For any fixed $\boldsymbol{z}=(z_i)_{i\in[N]}\in\mathcal{A}^{(N)}$, the function $\Phi:\mathcal{A}^{(N)}\to\mathbb{R}$ in (3.11) can be expressed as*

$$\Phi(\boldsymbol{u})=\int_0^1\sum_{i=1}^N\mathbb{E}\left[\int_0^T\begin{pmatrix}\boldsymbol{Y}_t^{\boldsymbol{u}^r,u_i-z_i}\\ u_{t,i}-z_{t,i}\end{pmatrix}^\top\begin{pmatrix}\partial_x f_i\\ \partial_{u_i}f_i\end{pmatrix}(t,\boldsymbol{X}_t^{\boldsymbol{u}^r},\boldsymbol{u}_t^r)\,\mathrm{d}t+(\partial_x g_i)^\top(\boldsymbol{X}_T^{\boldsymbol{u}^r})\,\boldsymbol{Y}_T^{\boldsymbol{u}^r,u_i-z_i}\right]\mathrm{d}r \tag{3.18}$$

with $\boldsymbol{u}^r:=\boldsymbol{z}+r(\boldsymbol{u}-\boldsymbol{z})$.

The expression (3.18) follows directly from (3.11) for $\Phi(\boldsymbol{u})$ and (3.17) for $\frac{\delta V_i}{\delta u_i}$, by substituting h with i, $\boldsymbol{u}$ with $\boldsymbol{z}+r(\boldsymbol{u}-\boldsymbol{z})$, and u_h' with u_i-z_i.

The α-potential function in (3.18) can be alternatively expressed using backward stochastic differential equations (BSDEs). The proof follows directly from [31, Corollary 4.11].

Proposition 3.3.1. *Under the setting of Theorem 3.3.1, Φ defined in (3.18) can be equivalently written as*

$$\Phi(\boldsymbol{u})=\int_0^1\sum_{i=1}^N\mathbb{E}\left[\int_0^T(\partial_{u_i}\mathbb{H}^i)^\top(t,\boldsymbol{X}_t^{\boldsymbol{u}^r},\boldsymbol{u}_t^r,\boldsymbol{G}_t^{i,\boldsymbol{u}^r},\boldsymbol{H}_t^{i,\boldsymbol{u}^r})(u_{t,i}-z_{t,i})\mathrm{d}t\right]\mathrm{d}r, \tag{3.19}$$

with $\boldsymbol{u}^r:=\boldsymbol{z}+r(\boldsymbol{u}-\boldsymbol{z})$, where for each $i\in[N]$,

$$\mathbb{H}^i(t,x,u,\mathfrak{g},\mathfrak{h}):=b^\top(t,x,u)\mathfrak{g}+\operatorname{tr}\left((\sigma\sigma^\top)(t,x,u)\mathfrak{h}\right)+f_i(t,x,u)$$

with $b=\mathsf{vcat}(b_1,\ldots,b_N)$ and $\sigma=\mathsf{vcat}(\sigma_1,\ldots,\sigma_N)$,[2] *and for each $\boldsymbol{u}\in\mathcal{H}^2(\mathbb{R}^{Nn})$, $(\boldsymbol{G}^{i,\boldsymbol{u}},\boldsymbol{H}^{i,\boldsymbol{u}})\in\mathcal{S}^2(\mathbb{R}^{Nd})\times\mathcal{H}^2(\mathbb{R}^{Nd\times m})$ satisfies*

$$\mathrm{d}\boldsymbol{G}_t^i=-(\partial_x\mathbb{H}^i)(t,\boldsymbol{X}_t^{\boldsymbol{u}},\boldsymbol{u}_t,\boldsymbol{G}_t^i,\boldsymbol{H}_t^i)\mathrm{d}t+\boldsymbol{H}_t^i\mathrm{d}W_t,\quad\forall t\in[0,T];\quad\boldsymbol{G}_T^i=(\partial_x g_i)(\boldsymbol{X}_T^{\boldsymbol{u}}).$$

In the sequel, we adopt the representation (3.18) of α-potential function in terms of the sensitivity processes. A detailed exploration of the BSDE approach for α-potential games is left for future work.

[2] We denote by $\mathsf{vcat}(A_1,\ldots,A_N):=(A_1^\top,\ldots,A_N^\top)^\top$ the vertical concentration of matrices $A_i\in\mathbb{R}^{m_i\times n}$, $1\le i\le N$.

3.3.1 Quantifying α for open-loop stochastic differential game

In this section, we quantify α in (3.12) for stochastic differential games based on the structure of the game. The analysis relies on characterizing the second-order derivatives of objective functions by utilizing the second-order sensitivity of the state dynamics with respect to the controls.

Let $\mathcal{G}^{\mathrm{op}}$ be the differential game defined in Section 3.3. For ease of exposition, in this section, we assume that each player has one-dimensional state and control processes, with the drift of (3.14) depending linearly on the control and the diffusion of (3.14) being independent of both the state and control. Similar analysis can be extended to sufficiently regular nonlinear drift and diffusion coefficients in a multidimensional setting. More precisely, for each $\boldsymbol{u} = (u_i)_{i\in[N]} \in \mathcal{H}^2(\mathbb{R}^N)$, let $\mathbf{X}^{\boldsymbol{u}} = (X_i^{\boldsymbol{u}})_{i=1}^N \in \mathcal{S}^2(\mathbb{R}^N)$ be the associated state process governed by the following dynamics: for all $i \in [N]$ and $t \in [0,T]$,

$$\mathrm{d}X_{t,i} = (b_i(t, X_{t,i}, \mathbf{X}_t) + u_{t,i})\,\mathrm{d}t + \sigma_i(t)\mathrm{d}W_t^i, \quad X_{0,i} = x_i, \tag{3.20}$$

where $x_i \in \mathbb{R}$, $b_i : [0,T]\times\mathbb{R}\times\mathbb{R}^N \to \mathbb{R}$ is a given sufficiently regular function, $\sigma_i : [0,T] \to \mathbb{R}$ is a given measurable function, and $W = (W^i)_{i\in[N]}$ is an N-dimensional $\mathbb{F}$-Brownian motion. Let $\mathcal{A}^{(N)} \subset \mathcal{H}^2(\mathbb{R}^N)$ be a nonempty convex set, representing the joint control profiles of all players. Player i's objective function $V_i : \mathcal{A}^{(N)} \to \mathbb{R}$ is given as in (3.15):

$$V_i(\boldsymbol{u}) = \mathbb{E}\left[\int_0^T f_i(t, \mathbf{X}_t^{\boldsymbol{u}}, \boldsymbol{u}_t)\,\mathrm{d}t + g_i(\mathbf{X}_T^{\boldsymbol{u}})\right], \tag{3.21}$$

where $f_i : [0,T]\times\mathbb{R}^N\times\mathbb{R}^N \to \mathbb{R}$ and $g_i : \mathbb{R}^N \to \mathbb{R}$ are given measurable functions.

Note that in (3.20), we have expressed the dependence of b_i on the private state $X_i^{\boldsymbol{u}}$ and the population state $\mathbf{X}^{\boldsymbol{u}}$ separately. This separation allows for specifying the structure of the drift coefficient. To this end, let $\mathscr{F}^{0,2}([0,T]\times\mathbb{R}\times\mathbb{R}^N;\mathbb{R})$ be the vector space of measurable functions $\psi : [0,T]\times\mathbb{R}\times\mathbb{R}^N \to \mathbb{R}$ such that

(1) for all $t \in [0,T]$, $(x,y) \mapsto \psi(t,x,y)$ is twice continuously differentiable,

(2) there exists $L^\psi, L_y^\psi \geq 0$ such that for all $(t,x,y) \in [0,T]\times\mathbb{R}\times\mathbb{R}^N$ and $i,j \in [N]$, $|\psi(t,0,0)| \leq L^\psi$, $|(\partial_x\psi)(t,x,y)| \leq L^\psi$, $|(\partial_{xx}^2\psi)(t,x,y)| \leq L^\psi$, $|(\partial_{y_i}\psi)(t,x,y)| \leq L_y^\psi/N$, $|(\partial_{xy_i}^2\psi)(t,x,y)| \leq L_y^\psi/N$, and $|(\partial_{y_iy_j}^2\psi)(t,x,y)| \leq \frac{1}{N}L_y^\psi \mathbb{1}_{i=j} + \frac{1}{N^2}L_y^\psi \mathbb{1}_{i\neq j}$.

For any $\psi = (\psi_i)_{i\in[N]} \in \mathscr{F}^{0,2}([0,T]\times\mathbb{R}\times\mathbb{R}^N;\mathbb{R})^N$, we write $L^\psi = \max_{i\in[N]} L^{\psi_i}$ and $L_y^\psi = \max_{i\in[N]} L_y^{\psi_i}$.

In the sequel, we impose the following regularity conditions on the coefficients of (3.20)-(3.21).

Assumption 3.3.2. *For all $i \in [N]$, $b_i \in \mathscr{F}^{0,2}([0,T]\times\mathbb{R}\times\mathbb{R}^N;\mathbb{R})$, $\sigma_i \in L^\infty([0,T];\mathbb{R})$, and f_i and g_i satisfy the conditions in Assumption 3.3.1(3).*

Remark 3.3.1. *For each $i \in [N]$, the condition $b_i \in \mathscr{F}^{0,2}([0,T]\times\mathbb{R}\times\mathbb{R}^N;\mathbb{R})$ implies the partial derivatives of $y \mapsto b_i(t,x,y)$ admit explicit decay rates in terms of N. This assumption naturally holds if each player's state depends on the empirical measure of the joint state process, i.e., the mean field interaction. To see it, suppose that $b_i(t,x,y) = h\big(t,x,\frac{1}{N}\sum_{j=1}^N \delta_{y_j}\big)$ with $(t,x,y) \in [0,T]\times\mathbb{R}\times\mathbb{R}^N$, for a measurable function $h : [0,T]\times\mathbb{R}\times\mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$, where $\mathcal{P}_2(\mathbb{R})$ is the space of probability measures on $\mathbb{R}$ with second moments. If $(x,\mu) \mapsto h(t,x,\mu)$ is sufficiently regular, then by [32, Propositions 5.35 and 5.91], $(\partial_{y_i} b_i)(t,x,y) = \frac{1}{N}(\partial_\mu h)(t,x,\frac{1}{N}\sum_{j=1}^N \delta_{y_j})(y_i)$, $(\partial^2_{xy_i} b_i)(t,x,y) = \frac{1}{N}(\partial_\mu\partial_x h)(t,x,\frac{1}{N}\sum_{j=1}^N \delta_{y_j})(y_i)$, and*

$$(\partial^2_{y_iy_j} b_i)(t,x,y) = \frac{1}{N}(\partial_v\partial_\mu h)\left(t,x,\frac{1}{N}\sum_{\ell=1}^N \delta_{y_\ell}\right)(y_i)\delta_{i,j} + \frac{1}{N^2}(\partial^2_\mu h)\left(t,x,\frac{1}{N}\sum_{\ell=1}^N \delta_{y_\ell}\right)(y_i,y_j),$$

where $(\partial_\mu h)(t,x,\mu)(\cdot)$ (resp. $(\partial_\mu\partial_x h)(t,x,\mu)(\cdot)$) is Lions derivative of $\mu \mapsto h(t,x,\mu)$ (resp. $\mu \mapsto (\partial_x h)(t,x,\mu)$), $(\partial_v\partial_\mu h)(t,x,\mu)(\cdot)$ is the derivative of $v \mapsto (\partial_\mu h)(t,x,\mu)(v)$, and $(\partial^2_\mu h)(t,x,\mu)(v,\cdot)$ is the Lions derivative of $\mu \mapsto (\partial_\mu h)(t,x,\mu)(v)$. Hence if $\partial_\mu h$, $\partial_\mu\partial_x h$, $\partial_v\partial_\mu h$ and $\partial^2_\mu h$ are continuous and uniformly bounded, then $b_i \in \mathscr{F}^{0,2}([0,T]\times\mathbb{R}\times\mathbb{R}^N;\mathbb{R})$ with a constant $L^{b_i}_y$ depending on the upper bounds of the Lions derivatives but independent of N.

The dependence of the constant $L^b_y = \max_{i\in[N]} L^{b_i}_y$ on N reflects the degree of coupling among all players' state dynamics. For instance, if L^b_y remains bounded as $N \to \infty$, then the state dynamics can have mean field type interactions. Alternatively, if $L^b_y = 0$, then all players' states are decoupled.

To quantify the magnitude of the asymmetry of the second-order linear derivatives of objective functions (3.21), hence α in (3.12), we characterize the linear derivatives using the sensitivity processes of (3.20). Observe that for the state dynamics (3.20), the dynamics (3.16) for the first-order sensitivity process $\mathbf{Y}^{\boldsymbol{u},u'_h} \in \mathcal{S}^2(\mathbb{R}^N)$ simplifies into for all $t \in [0,T]$,

$$\mathrm{d}Y^h_{t,i} = \left[(\partial_x b_i)(t,X^{\boldsymbol{u}}_{t,i},\mathbf{X}^{\boldsymbol{u}}_t)Y^h_{t,i} + \sum_{j=1}^N (\partial_{y_j} b_i)(t,X^{\boldsymbol{u}}_{t,i},\mathbf{X}^{\boldsymbol{u}}_t)Y^h_{t,j} + \delta_{h,i}u'_{t,h}\right]\mathrm{d}t, \quad Y^h_{0,i} = 0; \quad \forall i \in [N], \tag{3.22}$$

where $\delta_{i,j}$ denotes the Kronecker delta such that $\delta_{i,j} = 0$ if $i = j$ and 0 otherwise. We now characterize the second-order sensitivity of the state process with respect to the changes in two players' controls. For each $h,\ell \in [N]$ with $h \neq \ell$, and each $u'_h, u''_\ell \in \mathcal{H}^4(\mathbb{R})$, define $\mathbf{Z}^{\boldsymbol{u},u'_h,u''_\ell} \in \mathcal{S}^2(\mathbb{R}^N)$ as the solution of the following dynamics: for all $i \in [N]$ and $t \in [0,T]$,

$$\mathrm{d}Z^{h,\ell}_{t,i} = \left[(\partial_x b_i)(t,X^{\boldsymbol{u}}_{t,i},\mathbf{X}^{\boldsymbol{u}}_t)Z^{h,\ell}_{t,i} + \sum_{j=1}^N (\partial_{y_j} b_i)(t,X^{\boldsymbol{u}}_{t,i},\mathbf{X}^{\boldsymbol{u}}_t)Z^{h,\ell}_{t,j} + \mathfrak{f}^{\boldsymbol{u},u'_h,u''_\ell}_{t,i}\right]\mathrm{d}t, \quad Z^{h,\ell}_{0,i} = 0, \tag{3.23}$$

where $\mathfrak{f}_i^{\boldsymbol{u},u'_h,u''_\ell} : \Omega \times [0,T] \to \mathbb{R}$ is defined by

$$\mathfrak{f}_{t,i}^{\boldsymbol{u},u'_h,u''_\ell} := \begin{pmatrix} Y_{t,i}^{\boldsymbol{u},u'_h} \\ \mathbf{Y}_t^{\boldsymbol{u},u'_h} \end{pmatrix}^\top \begin{pmatrix} \partial^2_{xx} b_i & \partial^2_{xy} b_i \\ \partial^2_{yx} b_i & \partial^2_{yy} b_i \end{pmatrix} (t, X_{t,i}^{\boldsymbol{u}}, \mathbf{X}_t^{\boldsymbol{u}}) \begin{pmatrix} Y_{t,i}^{\boldsymbol{u},u''_\ell} \\ \mathbf{Y}_t^{\boldsymbol{u},u''_\ell} \end{pmatrix}, \tag{3.24}$$

and $\mathbf{Y}^{\boldsymbol{u},u'_h}$ and $\mathbf{Y}^{\boldsymbol{u},u''_\ell}$ are defined as in (3.22). Similar arguments as that for [31, Lemma 4.7] show that for all $u'_h, u''_\ell \in \mathcal{H}^4(\mathbb{R})$, $\lim_{\varepsilon \searrow 0} \mathbb{E}\left[\sup_{t\in[0,T]} \left|\frac{1}{\varepsilon}(\mathbf{Y}_t^{\boldsymbol{u}^\varepsilon,u'_h} - \mathbf{Y}_t^{\boldsymbol{u},u'_h}) - \mathbf{Z}_t^{\boldsymbol{u},u'_h,u''_\ell}\right|^2\right] = 0$, where $\boldsymbol{u}^\varepsilon = (u_\ell + \varepsilon u''_\ell, u_{-\ell})$ for all $\varepsilon \in (0,1)$. That is, $\mathbf{Z}^{\boldsymbol{u},u'_h,u''_\ell}$ is the second-order derivative of the state $\mathbf{X}^{\boldsymbol{u}}$ when player h first varies her control in the direction u'_h, and then player ℓ varies her control in the direction u''_ℓ.

Now, the linear derivatives of V_i in (3.21) can be represented using the sensitivity processes satisfying (3.22) and (3.23). The first order linear derivative $\frac{\delta V_i}{\delta u_h}$ of V_i is given as in (3.17). For the second-order linear derivatives, for all $h, \ell \in [N]$, define the map $\frac{\delta^2 V_i}{\delta u_h \delta u_\ell} : \mathcal{A}^{(N)} \times \mathcal{H}^4(\mathbb{R}) \times \mathcal{H}^4(\mathbb{R}) \to \mathbb{R}$ such that for all $\boldsymbol{u} \in \mathcal{A}^{(N)}$ and $u'_h, u''_\ell \in \mathcal{H}^4(\mathbb{R}^4)$,

$$\begin{aligned} \frac{\delta^2 V_i}{\delta u_h \delta u_\ell}(\boldsymbol{u}; u'_h, u''_\ell) = \mathbb{E}\Bigg[\int_0^T \Bigg(\begin{pmatrix} \mathbf{Y}_t^{\boldsymbol{u},u'_h} \\ u'_{t,h} \end{pmatrix}^\top \begin{pmatrix} \partial^2_{xx} f_i & \partial^2_{xu_\ell} f_i \\ \partial^2_{u_h x} f_i & \partial^2_{u_h u_\ell} f_i \end{pmatrix} \Big(t, \mathbf{X}_t^{\boldsymbol{u}}, \boldsymbol{u}_t\Big) \begin{pmatrix} \mathbf{Y}_t^{\boldsymbol{u},u''_\ell} \\ u''_{t,\ell} \end{pmatrix} \\ + (\partial_x f_i)^\top (t, \mathbf{X}_t^{\boldsymbol{u}}, \boldsymbol{u}_t)\, \mathbf{Z}_t^{\boldsymbol{u},u'_h,u''_\ell} \Bigg) \mathrm{d}t \Bigg] \\ + \mathbb{E}\left[\left(\mathbf{Y}_T^{\boldsymbol{u},u'_h}\right)^\top \left(\partial^2_{xx} g_i\right) (\mathbf{X}_T^{\boldsymbol{u}}) \mathbf{Y}_T^{\boldsymbol{u},u''_\ell} + (\partial_x g_i)^\top (\mathbf{X}_T^{\boldsymbol{u}})\, \mathbf{Z}_T^{\boldsymbol{u},u'_h,u''_\ell} \right]. \end{aligned} \tag{3.25}$$

Consequently, if $\mathcal{A}_h$ and $\mathcal{A}_\ell$ are convex subsets of $\mathcal{H}^4(\mathbb{R})$, then by [31, Lemma 4.8],

$$\lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon}\Big(\frac{\delta V_i}{\delta u_h}(\boldsymbol{u}^\varepsilon; u'_h) - \frac{\delta V_i}{\delta u_h}(\boldsymbol{u}; u'_h)\Big) = \frac{\delta^2 V_i}{\delta u_h \delta u_\ell}(\boldsymbol{u}; u'_h, u''_\ell - u_\ell)$$

for all $u'_h \in \mathcal{A}_h$ and $u''_\ell \in \mathcal{A}_\ell$, where $\boldsymbol{u}^\varepsilon = (u_\ell + \varepsilon(u''_\ell - u_\ell), u_{-\ell})$ for all $\varepsilon \in (0,1)$. That is, $\frac{\delta^2 V_i}{\delta u_h \delta u_\ell}(\boldsymbol{u}; u'_h, \cdot)$ is the linear derivative of $\boldsymbol{u} \mapsto \frac{\delta V_i}{\delta u_h}(\boldsymbol{u}; u'_h)$ with respect to $\mathcal{A}_\ell$, and hence the second-order linear derivative of V_i.

Before stating the theorem, we introduce a few constants that will be used in the analysis. For any $i, j \in [N]$ with $i \neq j$, we define $\Delta^f_{i,j} = f_i - f_j$, $\Delta^g_{i,j} = g_i - g_j$, as well as the following three constants $C^{i,j}_{V,1}$ $C^{i,j}_{V,2}$ $C^{i,j}_{V,3}$, depending on the upper bounds of the first- and second-order derivatives of $\Delta^f_{i,j}$ and $\Delta^g_{i,j}$ in (x, u):

$$C^{i,j}_{V,1} := \|\partial^2_{x_i x_j} \Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{x_i u_j} \Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_i x_j} \Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_i u_j} \Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{x_i x_j} \Delta^g_{i,j}\|_{L^\infty}, \tag{3.26}$$

$$
\begin{aligned}
C^{i,j}_{V,2} \coloneqq & \sum_{\ell\in[N]\setminus\{j\}} \|\partial^2_{u_i x_\ell}\Delta^f_{i,j}\|_{L^\infty} + \sum_{h\in[N]\setminus\{i\}} \|\partial_{x_h u_j}\Delta^f_{i,j}\|_{L^\infty} \\
& + \sum_{h\in\{i,j\}} \Big(\|(\partial_{x_h}\Delta^f_{i,j})(\cdot,0,0)\|_{L^2} + |(\partial_{x_h}\Delta^g_{i,j})(0)| \Big) \\
& + \sum_{h\in\{i,j\},\ell\in[N]} \Big(\|\partial^2_{x_h x_\ell}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{x_h u_\ell}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{x_h x_\ell}\Delta^g_{i,j}\|_{L^\infty} \Big), \qquad (3.27)
\end{aligned}
$$

$$
\begin{aligned}
C^{i,j}_{V,3} \coloneqq & \sum_{h\in[N]\setminus\{i,j\}} \Big(\|(\partial_{x_h}\Delta^f_{i,j})(\cdot,0,0)\|_{L^2} + |(\partial_{x_h}\Delta^g_{i,j})(0)| \Big) \\
& + \sum_{\substack{h\in[N]\setminus\{i,j\} \\ \ell\in[N]\setminus\{i,j\}}} \Big(\|\partial^2_{x_h x_\ell}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{x_h u_\ell}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{x_h x_\ell}\Delta^g_{i,j}\|_{L^\infty} \Big), \qquad (3.28)
\end{aligned}
$$

where $\|\cdot\|_\infty$ denotes the sup-norm norm.

We are ready to present the upper bound of the α defined in (3.12) for general cost functions $(f_i, g_i)_{i\in[N]}$, *without imposing any structural assumptions.*

Theorem 3.3.2. *Suppose Assumption 3.3.2 holds. Then for all $\boldsymbol{u} \in \mathcal{H}^2(\mathbb{R}^N)$ and $u_i', u_j'' \in \mathcal{H}^4(\mathbb{R})$,*

$$
\begin{aligned}
& \left| \frac{\delta^2 V_i}{\delta u_i \delta u_j}\left(\boldsymbol{u}; u_i', u_j''\right) - \frac{\delta^2 V_j}{\delta u_j \delta u_i}\left(\boldsymbol{u}; u_j'', u_i'\right) \right| \\
& \leq C\|u_i'\|_{\mathcal{H}^4(\mathbb{R})}\|u_j''\|_{\mathcal{H}^4(\mathbb{R})} \left(C^{i,j}_{V,1} + L^b_y \left(\frac{1}{N} C^{i,j}_{V,2} + \frac{1}{N^2} C^{i,j}_{V,3} \right) \right),
\end{aligned}
\qquad (3.29)
$$

where the constant L^b_y represents the coupling in the state dynamics (see Remark 3.3.1), the constants $C^{i,j}_{V,1}$, $C^{i,j}_{V,2}$ and $C^{i,j}_{V,3}$, defined in (3.26), (3.27) *and* (3.28), *respectively, and the constant $C \geq 0$ depends only on the upper bounds of T, $\max_{i\in[N]}|x_i|$, $\max_{i\in[N]}\|\sigma_i\|_{L^2}$, L^b and L^b_y.*

Consequently, if $\sup_{i\in[N],u_i\in\mathcal{A}_i}\|u_i\|_{\mathcal{H}^4(\mathbb{R})} < \infty$ and $0 \in \mathcal{A}_i$, then $\mathcal{G}^{op}$ is an α-potential game with an α-potential function Φ given by (3.18) *with $\boldsymbol{z} = 0$, and a constant α satisfying*

$$
\alpha \leq C \max_{i\in[N]} \sum_{j\in[N]\setminus\{i\}} \left(C^{i,j}_{V,1} + L^b_y \big(\frac{1}{N} C^{i,j}_{V,2} + \frac{1}{N^2} C^{i,j}_{V,3} \big) \right) \qquad (3.30)
$$

for a constant $C \geq 0$ independent of the cost functions.

Remark 3.3.2. *Since the minimizer of the function Φ given in Theorem 3.5.1 is an ϵ-Nash equilibrium of the game* (3.20)-(3.21), *with $\epsilon \leq \alpha$ in* (3.30). *One can construct approximate Nash equilibria for N-player games without the symmetry and homogeneity conditions among players imposed for mean field approximations [32]. Moreover, the upper bound* (3.30) *is expressed in terms of the number of players, the strength of interactions, and the degree of heterogeneity among the players, proving rich insights for assessing the approximate Nash* *equilibria in relation to the game structure, compared to the classical mean field approximation, which typically bounds the approximation error solely based on the number of players* N.

The condition $\sup_{i\in[N],u_i\in\mathcal{A}_i}\|u_i\|_{\mathcal{H}^4(\mathbb{R})} < \infty$ *for the estimate* (3.30) *can be relaxed to* $\sup_{i\in[N],u_i\in\mathcal{A}_i}\|u_i\|_{\mathcal{H}^2(\mathbb{R})} < \infty$ *if the state dynamics* (3.20) *is decoupled, i.e., if the drift* b_i *is independent of* $(X_j)_{j\neq i}$*; see Example 3.3.2 and Section 3.6. Indeed, the appearance of* $\|u_i'\|_{\mathcal{H}^4}$ *and* $\|u_j''\|_{\mathcal{H}^4}$ *in the estimate* (3.29) *is due to the* L^2*-estimate of the process* $\boldsymbol{Z}^{\boldsymbol{u},u_i',u_j''}$*; see Proposition 3.7.5 and* (3.86)*. If the state dynamics is decoupled, then* $\boldsymbol{Z}^{\boldsymbol{u},u_i',u_j''} = 0$ *for* $i \neq j$*, and the additional condition on the* $\|\cdot\|_{\mathcal{H}^4(\mathbb{R})}$ *is unnecessary. The uniform integrability condition* $\sup_{i\in[N],u_i\in\mathcal{A}_i}\|u_i\|_{\mathcal{H}^2(\mathbb{R})} < \infty$ *comes from estimating the* $\|\cdot\|_{\mathcal{H}^2(\mathbb{R})}$*-norm of* $\boldsymbol{Y}^{\boldsymbol{u},u_i'}$ *uniformly over* $u_i' \in \mathcal{A}_i$ *and* $i \in [N]$*. This highlights the dependence of* α *on the choice of admissible control classes* $(\mathcal{A}_i)_{i\in[N]}$*. This dependence may be useful for analyzing the sensitivity of* α*-NE with respect to design of game strategies.*

A similar estimate of α *can be established if the drift coefficient in* (3.20) *depends nonlinearly on* u_i*. In such cases, the sensitivity equations* (3.22) *and* (3.23) *will incorporate the derivatives of the drift coefficient with respect to* u_i*, and the constant* C *in Theorem 3.3.2 will depend on the upper bounds of these derivatives.*

The proof of Theorem 3.3.2 is given in Section 3.7.2. The essential step is to establish precise estimates of the sensitivity processes $\mathbf{Y}^{\boldsymbol{u},u_h'}$ and $\mathbf{Z}^{\boldsymbol{u},u_h',u_\ell''}$ in terms of the number of players N and the indices h,ℓ. These estimates quantify the dependence of each player's state process on the changes in other players' controls, with a constant depending explicitly on the coupling strength L^b_y in the drift coefficients (see Remark 3.3.1) and the number of players.

3.3.2 Examples of open-loop α-potential games

Distributed games. Theorem 3.3.2 simplifies the task of quantifying the constant α in (3.12) to bounding the difference of derivatives of the cost functions. For instance, the following example presents a special case where the state dynamics (3.20) are decoupled. It shows that under suitable conditions of the cost functions, a distributed open-loop game (3.20)-(3.21) is an α-potential game, with α decaying to 0 as the number of players $N \to \infty$. It generalizes [66, Theorem 3.2] by allowing the cost functions f_i and f_j to have asymmetric second-order derivatives.

Example 3.3.1 (Distributed games)**.** *Consider the game* $\mathcal{G}^{op}$ *defined as in* (3.20)-(3.21)*. Suppose Assumption 3.3.2 holds, for all* $i \in [N]$, $(t,x,y) \mapsto b_i(t,x,y)$ *is independent of* y*, and there exists* $L, L^c \geq 0$ *and* $\beta \geq 1/2$ *such that* $\sup_{i\in[N],u\in\mathcal{A}_i}\|u\|_{\mathcal{H}^4(\mathbb{R})} \leq L$, $\max_{i\in[N]}\mathbb{E}[|\xi_i|^2] \leq L$, $\max_{i\in[N]} L^{b_i} \leq L$, $\max_{i\in[N]}\|\sigma_i\|_{L^\infty} \leq L$*, and for all* $i,j \in [N]$, $\Delta^f_{i,j} := f_i - f_j$ *and* $\Delta^g_{i,j} := g_i - g_j$ *satisfy for all* $(t,x,u) \in [0,T]\times\mathbb{R}^N\times\mathbb{R}^N$, $|(\partial^2_{x_ix_j}\Delta^f_{i,j})(t,x,u)| + |(\partial^2_{x_iu_j}\Delta^f_{i,j})(t,x,u)| + |(\partial^2_{u_ix_j}\Delta^f_{i,j})(t,x,u)| + |(\partial^2_{u_iu_j}\Delta^f_{i,j})(t,x,u)| + |(\partial^2_{x_ix_j}\Delta^g_{i,j})(x)| \leq L^cN^{-2\beta}$*. Then* $\mathcal{G}^{op}$ *is an* α*-*

potential game with $\alpha \leq CL^c N^{-(2\beta-1)}$, where $C \geq 0$ is a constant independent of N and β.

Example 3.3.1 follows directly from Theorems 3.2.1 and 3.3.2 (with $L_y^b = 0$).

The following example illustrates a special case where the distributed game is in fact a potential game ($\alpha = 0$). As a result, the minimizer of the potential function Φ given in Theorem 3.5.1 is a Nash equilibrium of the N-player game (3.20)-(3.21).

Example 3.3.2 (Distributed games with $\alpha = 0$). *Consider the game $\mathcal{G}^{\mathrm{op}}$ defined as in (3.20)-(3.21). Suppose Assumption 3.3.2 holds, and for all $i \in [N]$, $(t,x,y) \mapsto b_i(t,x,y)$ is independent of y, and f_i and g_i are of the form*

$$f_i(t,x,u) = c_i(t,x_i,u_i) + \tilde{f}_i\left(t,x_i,u_i,\bar{\mu}_{(x,u)_{-i}}\right), \quad g_i(x) = \bar{g}_i(x_i) + \tilde{g}_i\left(x_i,\bar{\mu}_{x_{-i}}\right). \tag{3.31}$$

where $\bar{\mu}_{(x,u)_{-i}} = \frac{1}{N-1}\sum_{j \in I_N \setminus \{i\}} \delta_{(x_j,u_j)}$, $\bar{\mu}_{x_{-i}} = \frac{1}{N-1}\sum_{j \in I_N \setminus \{i\}} \delta_{x_j}$, and $c_i : [0,T] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $\tilde{f}_i : [0,T] \times \mathbb{R} \times \mathbb{R} \times \mathcal{P}_2(\mathbb{R} \times \mathbb{R}) \to \mathbb{R}$, $\bar{g}_i : \mathbb{R} \to \mathbb{R}$, and $\tilde{g}_i : \mathbb{R} \times \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ are twice continuously differentiable. Assume further that there exist twice continuously differentiable functions $F : [0,T] \times \mathcal{P}_2(\mathbb{R} \times \mathbb{R}) \to \mathbb{R}$ and $G : \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ such that for all $i \in [N]$, $(t,x,u) \in [0,T] \times \mathbb{R}^N \times \mathbb{R}^N$ and $(x_i', u_i') \in \mathbb{R} \times \mathbb{R}$,

$$\begin{aligned}
&\tilde{f}_i(t,x_i,u_i,\bar{\mu}_{(x,u)_{-i}}) - \tilde{f}_i(t,x_i',u_i',\bar{\mu}_{(x,u)_{-i}}) \\
&= F\Big(t, \frac{1}{N}\delta_{(x_i,u_i)} + \frac{N-1}{N}\bar{\mu}_{(x,u)_{-i}}\Big) - F\Big(t, \frac{1}{N}\delta_{(x_i',u_i')} + \frac{N-1}{N}\bar{\mu}_{(x,u)_{-i}}\Big), \\
&\tilde{g}_i\left(x_i,\bar{\mu}_{x_{-i}}\right) - \tilde{g}_i\left(x_i',\bar{\mu}_{x_{-i}}\right) = G\Big(\frac{1}{N}\delta_{x_i} + \frac{N-1}{N}\bar{\mu}_{x_{-i}}\Big) - G\Big(\frac{1}{N}\delta_{x_i'} + \frac{N-1}{N}\bar{\mu}_{x_{-i}}\Big).
\end{aligned} \tag{3.32}$$

Then $\alpha = 0$ and $\mathcal{G}^{\mathrm{op}}$ is a potential game.

Example 3.3.2 extends [32, Proposition 2.24] from cost functions dependent solely on state variables to those dependent on both state and control variables. It follows from the fact that by (3.32), $h_i^f := \tilde{f}_i - F$ and $h_i^g := \tilde{g}_i - G$ are independent of (x_i,u_i). This implies $C_{V,1}^{i,j} = 0$ as defined in (3.26). As the states are decoupled, $L_y^b = 0$ and hence $\alpha = 0$ by Theorem 3.3.2.

Games with mean field interactions. When all players' state dynamics (3.20) are coupled, a stronger condition on the cost functions is needed to ensure the constant α in (3.30) decays to zero as the number of players $N \to \infty$. The following example shows that if the cost functions in (3.21) depend on the joint states and controls only through their empirical measures, then the N-player game (3.20)-(3.21) is an α-potential game with $\alpha = \mathcal{O}(1/N)$ as $N \to \infty$.

Example 3.3.3 (games with mean field interactions). *Consider the game $\mathcal{G}^{\mathrm{op}}$ defined by (3.20)-(3.21). Suppose Assumption 3.3.2 holds and there exists $L \geq 0$ such that*

$$\sup_{i \in [N], u \in \mathcal{A}_i} \|u\|_{\mathcal{H}^4(\mathbb{R})} \leq L, \quad \max_{i \in [N]} |x_i| \leq L, \text{ and } \max_{i \in [N]} \|\sigma_i\|_{L^\infty} \leq L.$$

Assume further that there exists $f_0 : [0,T] \times \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ and $g_0 : \mathbb{R}^N \to \mathbb{R}$ such that for all $i \in [N]$, b_i, f_i and g_i are of the following form:

$$b_i(t, x_i, x) = \bar{b}_i \left(t, x_i, \frac{1}{N} \sum_{\ell=1}^{N} \delta_{x_\ell} \right), \tag{3.33}$$

$$f_i(t, x, u) = f_0(t, x, u) + c_i(u_i) + \bar{f}_i \left(t, \frac{1}{N} \sum_{\ell=1}^{N} \delta_{(x_\ell, u_\ell)} \right), \quad g_i(x) = g_0(x) + \bar{g}_i \left(\frac{1}{N} \sum_{\ell=1}^{N} \delta_{x_\ell} \right), \tag{3.34}$$

where $\bar{b}_i : [0,T] \times \mathbb{R} \times \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$, $c_i : \mathbb{R} \to \mathbb{R}$, $\bar{f}_i : [0,T] \times \mathcal{P}_2(\mathbb{R} \times \mathbb{R}) \to \mathbb{R}$, and $\bar{g}_i : \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ are twice continuously differentiable with bounded second-order derivatives (uniformly in N). Then $\mathcal{G}^{op}$ is an α-potential game with $\alpha \le C/N$, for a constant $C \ge 0$ independent of N.

Example 3.3.3 follows from the fact that by (3.33) and (3.34),

$$|(\partial_{x_h} \Delta^f_{i,j})(t, 0, 0)| + |(\partial_{x_h} \Delta^g_{i,j})(0)| \le C/N,$$

and

$$|(\partial^2_{x_h x_\ell} \Delta^f_{i,j})(t, x, u)| + |(\partial^2_{x_h u_\ell} \Delta^f_{i,j})(t, x, u)| + |(\partial^2_{x_h x_\ell} \Delta^g_{i,j})(x)| \le C \left(\frac{1}{N} \mathbb{1}_{h=\ell} + \frac{1}{N^2} \mathbb{1}_{h \ne \ell} \right)$$

for some constant $C \ge 0$ independent of N (see Remark 3.3.1), which yields the bound of α due to Theorem 3.3.2.

Remark 3.3.3. *Example 3.3.3 allows all players to have different admissible control sets $\mathcal{A}_i$, and heterogeneous dependencies on the empirical measures of the joint state and control profiles. This is in contrast to the classical N-player mean field games with symmetric and homogeneous players (see [32]).*

Note that even if all players have homogeneous coefficients, the conditions in Example 3.3.3 differ from those for potential mean field games (MFGs) introduced in [99, 100, 30]. An MFG is considered potential if there exists an optimal control problem whose optimal trajectories coincide with the equilibria of the MFG. This is a weaker condition than the notion of N-player potential game described in [113], as it is a local property that concerns only the minimizer of the potential function.

In contrast, Example 3.3.3 allows the α-potential function to control the derivatives of each player's objective function globally, with an error of order $\mathcal{O}(1/N)$ as $N \to \infty$. This property is crucial for ensuring the convergence of gradient-based learning algorithms (see [71] and references therein). Consequently, when b_i depends on the empirical measure of the states, we require the cost functions $\bar{f}_i$ and $\bar{g}_i$ in (3.34) to depend on the state and controls only through their empirical measures. Assuming the uniqueness of Nash equilibria in MFGs, the minimum of the α-potential function (with appropriate scaling) converges to the minimum of the mean field potential function as $N \to \infty$, provided that sufficient conditions are met to allow the interchangeability of minimization and the limit.

3.4 Closed-Loop Stochastic Differential Games

This section studies the stochastic differential games with closed-loop controls. Compared with the open-loop games studied in Section 3.3.1, characterizing closed-loop α-potential games is more technically involved. The increased complexity arises from the fact that if one player changes her policy, this change is likely to impact the state trajectory of the system. Even if other players continue employing the same policy, their actions and value functions will change due to this change of the system state. This additional interdependence through policies must be incorporated into the analysis.

Consider the closed-loop differential game $\mathcal{G}^{\mathsf{cl}} = ([N], (\mathcal{A}_i)_{i\in[N]}, (V_i)_{i\in[N]})$ defined as follows: let $[N] = \{1, \ldots, N\}$, let $\Pi = \mathscr{F}^{0,2}([0,T] \times \mathbb{R} \times \mathbb{R}^N; \mathbb{R})$ be the vector space defined in Section 3.3.1, and let $\Pi^N = \mathscr{F}^{0,2}([0,T] \times \mathbb{R} \times \mathbb{R}^N; \mathbb{R})^N$. For each $i \in [N]$, let $\mathcal{A}^i$ be a convex subset of Π representing the set of admissible closed-loop policies of player i. For each $\phi = (\phi_i)_{i\in[N]} \in \Pi^N$, let $\mathbf{X}^\phi = (X_i^\phi)_{i=1}^N \in \mathcal{S}^2(\mathbb{R}^N)$ be the associated state process governed by the following dynamics (cf. (3.20)): for all $i \in [N]$ and $t \in [0,T]$,

$$\mathrm{d}X_{t,i} = \bar{b}_i(t, \mathbf{X}_t, u_{t,i}^\phi)\mathrm{d}t + \sigma_i(t)\mathrm{d}W_t^i, \quad X_{0,i} = \xi_i, \quad \text{with } u_{t,i}^\phi = \phi_i(t, X_{t,i}, \mathbf{X}_t), \tag{3.35}$$

where ξ_i is a given square integrable $\mathcal{F}_0$-measurable random variable, and $\bar{b}_i$ and σ_i are given measurable functions. Define the value function $V_i : \mathcal{A}^{(N)} \subset \Pi^N \to \mathbb{R}$ of player i by

$$V_i(\phi) = \mathbb{E}\left[\int_0^T f_i(t, \mathbf{X}_t^\phi, \boldsymbol{u}_t^\phi)\, \mathrm{d}t + g_i(\mathbf{X}_T^\phi)\right], \tag{3.36}$$

where $\boldsymbol{u}_t^\phi \coloneqq (\phi_i(t, X_{t,i}^\phi, \mathbf{X}_t^\phi))_{i\in[N]}$ is the joint closed-loop control profile[3] at t, and f_i and g_i are given measurable function of quadratic growth in (x,u). We assume that the coefficients $(\xi_i)_{i\in[N]}$, $(\bar{b}_i)_{i\in[N]}$, $(\sigma_i)_{i\in[N]}$, $(f_i)_{i\in[N]}$ and $(g_i)_{i\in[N]}$ satisfy Assumption 3.3.2. In particular, for all $i \in [N]$, $\bar{b}_i$ satisfies $\bar{b}_i(t,x,u) = b_i(t, x_i, x) + u$ for some $b_i \in \mathscr{F}^{0,2}([0,T] \times \mathbb{R} \times \mathbb{R}^N; \mathbb{R})$.

Note that the closed-loop game $\mathcal{G}^{\mathsf{cl}}$ introduced above allows player i's admissible policies to depend nonlinearly on both the private state X_i^ϕ and the population state $\mathbf{X}^\phi$ in a heterogeneous manner. To simplify the notation, we restrict the admissible policies to be elements of the space $\mathscr{F}^{0,2}([0,T] \times \mathbb{R} \times \mathbb{R}^N; \mathbb{R})$, which implies that each policy's dependence on the population state admits suitable a-priori bounds in terms of N. As highlighted in Remark 3.3.1, these policies include those that depend on the empirical measure of the joint state $\mathbf{X}^\phi$ as special cases. For each $\phi \in \mathcal{A}_i$, we express the dependence of ϕ on the private state and the population state separately. This separation allows for quantifying the constant α in (3.12) in terms of the strength of the dependence of admissible policies on the population state, namely, the constant L_y^ϕ.

Similar to the open-loop games, the linear derivatives of the value function (3.36) for closed-loop games $\mathcal{G}^{\mathsf{cl}}$ can also be represented and analyzed using appropriate sensitivity

[3] In the sequel, a (closed-loop) policy refers to a deterministic function ϕ that maps time and state variables to an action, and a closed-loop control refers to the stochastic process $\boldsymbol{u}^\phi$ generated by a certain policy ϕ.

processes of the state dynamic (3.35). These sensitivity processes are more complex than those for open-loop controls (see (3.22) and (3.23)) because of the additional interdependence created by the closed-loop policies.

For each $\phi \in \Pi^N$, let $\mathbf{X}^\phi$ be the state process satisfying (3.35). For each $h \in [N]$ and $\phi'_h \in \Pi$, let $\mathbf{Y}^{\phi,\phi'_h} \in \mathcal{S}^2(\mathbb{R}^N)$ be the solution to the following dynamics: for all $t \in [0,T]$,

$$\begin{aligned} \mathrm{d}Y^h_{t,i} &= \Bigg[(\partial_x(b_i+\phi_i))(t, X^\phi_{t,i}, \mathbf{X}^\phi_t)Y^h_{t,i} \\ &\qquad + \sum_{j=1}^N (\partial_{y_j}(b_i+\phi_i))(t, X^\phi_{t,i}, \mathbf{X}^\phi_t)Y^h_{t,j} + \delta_{h,i}\phi'_h(t, X^\phi_{t,i}, \mathbf{X}^\phi_t)\Bigg]\mathrm{d}t, \\ Y^h_{0,i} &= 0, \quad \forall i \in [N]. \end{aligned} \tag{3.37}$$

Compared with (3.22), (3.37) has additional terms involving partial derivatives of ϕ_i, due to the feedback structure of the closed-loop policy. Moreover, as shown in [66, Lemma 5.1], for any $q \geq 2$, if $\xi_0 \in L^q(\Omega;\mathbb{R})$ for all $i \in [N]$, then for all $\phi'_h \in \Pi$,

$$\lim_{\varepsilon \searrow 0} \mathbb{E}\left[\sup_{t\in[0,T]} \left| \frac{1}{\varepsilon}(\mathbf{X}^{\phi^\varepsilon}_t - \mathbf{X}^\phi_t) - \mathbf{Y}^{\phi,\phi'_h}_t \right|^q \right] = 0,$$

where $\phi^\varepsilon = (\phi_h + \varepsilon\phi'_h, \phi_{-h})$ for all $\varepsilon \in (0,1)$. This implies that $\mathbf{Y}^{\phi,\phi'_h}$ is the derivative of the controlled state $\mathbf{X}^\phi$ when player h varies her policy in the direction ϕ'_h.

Moreover, for each $h, \ell \in [N]$ and $\phi'_h, \phi''_\ell \in \Pi$, let $\mathbf{Z}^{\phi,\phi'_h,\phi''_\ell}$ be the solution to the following dynamics: for all $i \in [N]$ and $t \in [0,T]$,

$$\begin{aligned} \mathrm{d}Z^{h,\ell}_{t,i} &= \Bigg[(\partial_x(b_i+\phi_i))(t, X^\phi_{t,i}, \mathbf{X}^\phi_t)Z^{h,\ell}_{t,i} \\ &\qquad + \sum_{j=1}^N (\partial_{y_j}(b_i+\phi_i))(t, X^\phi_{t,i}, \mathbf{X}^\phi_t)Z^{h,\ell}_{t,j} + \mathfrak{f}^{\phi,\phi'_h,\phi''_\ell}_{t,i}\Bigg]\mathrm{d}t, \\ Z^{h,\ell}_{0,i} &= 0, \end{aligned} \tag{3.38}$$

where $\mathfrak{f}^{\phi,\phi'_h,\phi''_\ell}_i : \Omega \times [0,T] \to \mathbb{R}$ is defined by

$$\begin{aligned} \mathfrak{f}^{\phi,\phi'_h,\phi''_\ell}_{t,i} &:= \begin{pmatrix} Y^{\phi,\phi'_h}_{t,i} \\ \mathbf{Y}^{\phi,\phi'_h}_t \end{pmatrix}^\top \begin{pmatrix} \partial^2_{xx}(b_i+\phi_i) & \partial^2_{xy}(b_i+\phi_i) \\ \partial^2_{yx}(b_i+\phi_i) & \partial^2_{yy}(b_i+\phi_i) \end{pmatrix} (t, X^\phi_{t,i}, \mathbf{X}^\phi_t) \begin{pmatrix} Y^{\phi,\phi''_\ell}_{t,i} \\ \mathbf{Y}^{\phi,\phi''_\ell}_t \end{pmatrix} \\ &\quad + \delta_{h,i}\Big((\partial_x\phi'_h)(t, X^\phi_{t,i}, \mathbf{X}^\phi_t)Y^{\phi,\phi''_\ell}_{t,i} + (\mathbf{Y}^{\phi,\phi''_\ell}_t)^\top(\partial_y\phi'_h)(t, X^\phi_{t,i}, \mathbf{X}^\phi_t)\Big) \\ &\quad + \delta_{\ell,i}\Big((\partial_x\phi''_\ell)(t, X^\phi_{t,i}, \mathbf{X}^\phi_t)Y^{\phi,\phi'_h}_{t,i} + (\mathbf{Y}^{\phi,\phi'_h}_t)^\top(\partial_y\phi''_\ell)(t, X^\phi_{t,i}, \mathbf{X}^\phi_t)\Big), \end{aligned} \tag{3.39}$$

and $\mathbf{Y}^{\phi,\phi'_h}$ and $\mathbf{Y}^{\phi,\phi''_\ell}$ are defined as in (3.37). By [66, Lemma 5.2], for all $\phi'_h, \phi''_\ell \in \Pi$, if $\xi_0 \in L^4(\Omega;\mathbb{R})$ for all $i \in [N]$, then $\lim_{\varepsilon\searrow 0} \mathbb{E}\left[\sup_{t\in[0,T]} \left|\frac{1}{\varepsilon}(\mathbf{Y}_t^{\phi^\varepsilon,\phi'_h} - \mathbf{Y}_t^{\phi,\phi'_h}) - \mathbf{Z}_t^{\phi,\phi'_h,\phi''_\ell}\right|^2\right] = 0$, where $\phi^\varepsilon = (\phi_\ell + \varepsilon\phi''_\ell, \phi_{-\ell})$ for all $\varepsilon \in (0,1)$. This proves that $\mathbf{Z}^{\phi,\phi'_h,\phi''_\ell}$ is the second-order derivative of the state $\mathbf{X}^\phi$ when player h first varies her policy in the direction ϕ'_h, and then player ℓ varies her policy in the direction ϕ''_ℓ.

In addition to the sensitivity processes for the controlled state $\mathbf{X}^\phi$, we also require the sensitivity of the control process $\boldsymbol{u}^\phi$ with respect to players' policies. These processes capture the change in each player's control due to the change in the system state. More precisely, let $\phi \in \Pi^N$, let $\boldsymbol{u}^\phi = (\phi_i(\cdot, X_i^\phi, \mathbf{X}^\phi))_{i\in[N]}$. For each $h, \ell \in [N]$ and each $\phi'_h, \phi''_\ell \in \Pi$, define $\boldsymbol{v}^{\phi,\phi'_h} = (v_i^{\phi,\phi'_h})_{i\in[N]}$ such that for all $i \in [N]$,

$$v_i^{\phi,\phi'_h} = (\partial_x\phi_i)(\cdot, X_i^\phi, \mathbf{X}^\phi)Y_i^{\phi,\phi'_h} + (\mathbf{Y}^{\phi,\phi'_h})^\top(\partial_y\phi_i)(\cdot, X_i^\phi, \mathbf{X}^\phi) + \delta_{h,i}\phi'_h(\cdot, X_i^\phi, \mathbf{X}^\phi), \tag{3.40}$$

and define $\boldsymbol{w}^{\phi,\phi'_h,\phi''_\ell} = (w_i^{\phi,\phi'_h,\phi''_\ell})_{i\in[N]}$ such that for all $i \in [N]$,

$$\begin{aligned} w_i^{\phi,\phi'_h,\phi''_\ell} = &\begin{pmatrix} Y_i^{\phi,\phi'_h} \\ \mathbf{Y}^{\phi,\phi'_h} \end{pmatrix}^\top \begin{pmatrix} \partial^2_{xx}\phi_i & \partial^2_{xy}\phi_i \\ \partial^2_{yx}\phi_i & \partial^2_{yy}\phi_i \end{pmatrix} (\cdot, X_i^\phi, \mathbf{X}^\phi) \begin{pmatrix} Y_i^{\phi,\phi''_\ell} \\ \mathbf{Y}^{\phi,\phi''_\ell} \end{pmatrix} \\ &+ (\partial_x\phi_i)(\cdot, X_i^\phi, \mathbf{X}^\phi)Z_i^{\phi,\phi'_h,\phi''_\ell} + (\mathbf{Z}^{\phi,\phi'_h,\phi''_\ell})^\top(\partial_y\phi_i)(\cdot, X_i^\phi, \mathbf{X}^\phi) \\ &+ \delta_{h,i}\left((\partial_x\phi'_h)(\cdot, X_i^\phi, \mathbf{X}^\phi)Y_i^{\phi,\phi''_\ell} + (\mathbf{Y}^{\phi,\phi''_\ell})^\top(\partial_y\phi'_h)(\cdot, X_i^\phi, \mathbf{X}^\phi)\right) \\ &+ \delta_{\ell,i}\left((\partial_x\phi''_\ell)(\cdot, X_i^\phi, \mathbf{X}^\phi)Y_i^{\phi,\phi'_h} + (\mathbf{Y}^{\phi,\phi'_h})^\top(\partial_y\phi''_\ell)(\cdot, X_i^\phi, \mathbf{X}^\phi)\right). \end{aligned} \tag{3.41}$$

As shown in [66, Lemma 5.2], if $\xi_i \in L^4(\Omega;\mathbb{R})$ for all $i \in [N]$, then

$$\lim_{\varepsilon\searrow 0} \mathbb{E}\left[\int_0^T \left|\frac{1}{\varepsilon}(\boldsymbol{u}_t^{\phi^\varepsilon} - \boldsymbol{u}_t^\phi) - \boldsymbol{v}_t^{\phi,\phi'_h}\right|^4 \mathrm{d}t\right] = 0,$$

where $\phi^\varepsilon = (\phi_h + \varepsilon\phi'_h, \phi_{-h})$ for all $\varepsilon \in (0,1)$, and

$$\lim_{\varepsilon\searrow 0} \mathbb{E}\left[\int_0^T \left|\frac{1}{\varepsilon}(\boldsymbol{v}_t^{\widetilde{\phi}^\varepsilon,\phi'_h} - \boldsymbol{v}_t^{\phi,\phi'_h}) - \boldsymbol{w}_t^{\phi,\phi'_h,\phi''_\ell}\right|^2 \mathrm{d}t\right] = 0,$$

where $\widetilde{\phi}^\varepsilon = (\phi_\ell + \varepsilon\phi''_\ell, \phi_{-\ell})$ for all $\varepsilon \in (0,1)$. This shows that $\boldsymbol{v}^{\phi,\phi'_h}$ is the derivative of the joint control process $\boldsymbol{u}^\phi$ with respect to player h's policy, and $\boldsymbol{w}^{\phi,\phi'_h,\phi''_\ell}$ is the derivative of $\boldsymbol{u}^\phi$ with respect to player h and player ℓ's policies.

We now characterize the linear derivatives of V_i in (3.36) with respect to policies as in [66, Theorem 4.4]. For all $i, h \in [N]$, the linear derivative $\frac{\delta V_i}{\delta\phi_h} : \mathcal{A}^{(N)} \times \Pi \to \mathbb{R}$ of V_i with respect to $\mathcal{A}_h$ is given by

$$\frac{\delta V_i}{\delta\phi_h}(\phi;\phi'_h) = \mathbb{E}\left[\int_0^T \begin{pmatrix} \mathbf{Y}_t^{\phi,\phi'_h} \\ \boldsymbol{v}_t^{\phi,\phi'_h} \end{pmatrix}^\top \begin{pmatrix} \partial_x f_i \\ \partial_u f_i \end{pmatrix} (t, \mathbf{X}_t^\phi, \boldsymbol{u}_t^\phi)\mathrm{d}t + (\mathbf{Y}_T^{\phi,\phi'_h})^\top(\partial_x g_i)(\mathbf{X}_T^\phi)\right]. \tag{3.42}$$

Moreover, for all $i, h, \ell \in [N]$, define $\frac{\delta^2 V_i}{\delta\phi_h\phi_\ell} : \mathcal{A}^{(N)} \times \Pi \times \Pi \to \mathbb{R}$ by

$$\begin{aligned}\frac{\delta^2 V_i}{\delta\phi_h\delta\phi_\ell}(\phi; \phi'_h, \phi''_\ell) = \mathbb{E}\Bigg[\int_0^T \Bigg\{ &\begin{pmatrix} \mathbf{Y}_t^{\phi,\phi'_h} \\ \boldsymbol{v}_t^{\phi,\phi'_h} \end{pmatrix}^\top \begin{pmatrix} \partial^2_{xx} f_i & \partial^2_{xu} f_i \\ \partial^2_{ux} f_i & \partial^2_{uu} f_i \end{pmatrix} (t, \mathbf{X}_t^\phi, \boldsymbol{u}_t^\phi) \begin{pmatrix} \mathbf{Y}_t^{\phi,\phi''_\ell} \\ \boldsymbol{v}_t^{\phi,\phi''_\ell} \end{pmatrix} \\ &+ \begin{pmatrix} \mathbf{Z}_t^{\phi,\phi'_h,\phi''_\ell} \\ \boldsymbol{w}_t^{\phi,\phi'_h,\phi''_\ell} \end{pmatrix}^\top \begin{pmatrix} \partial_x f_i \\ \partial_u f_i \end{pmatrix} (t, \mathbf{X}_t^\phi, \boldsymbol{u}_t^\phi) \Bigg\} \mathrm{d}t \Bigg] \\ &+ \mathbb{E}\left[(\mathbf{Y}_T^{\phi,\phi'_h})^\top (\partial^2_{xx} g_i)(\mathbf{X}_T^\phi) \mathbf{Y}_T^{\phi,\phi''_\ell} + (\mathbf{Z}_T^{\phi,\phi'_h,\phi''_\ell})^\top (\partial_x g_i)(\mathbf{X}_T^\phi) \right], \qquad (3.43)\end{aligned}$$

which is the linear derivative of $\phi \mapsto \frac{\delta V_i}{\delta\phi_h}(\phi; \phi'_h)$ with respect to $\mathcal{A}_\ell$.

An explicit representation of the α-potential function Φ in (3.11) can be obtained for the closed-loop game $\mathcal{G}^{\mathsf{cl}}$ based on the expression of $(\frac{\delta V_i}{\delta\phi_i})_{i\in[N]}$ in (3.42):

Theorem 3.4.1. *Consider the game $\mathcal{G}^{cl}$ defined by* (3.35)-(3.36). *Suppose Assumption 3.3.2 holds. For any fixed $z = (z_i)_{i\in[N]} \in \Pi^{(N)}$, the function $\Phi : \mathcal{A}^{(N)} \to \mathbb{R}$ in* (3.11) *can be expressed as*

$$\Phi(\boldsymbol{u}^\phi) = \int_0^1 \sum_{h=1}^N \mathbb{E}\left[\int_0^T \begin{pmatrix} \boldsymbol{Y}_t^{\phi^r,\phi_h - z_h} \\ \boldsymbol{v}_t^{\phi^r,\phi_h - z_h} \end{pmatrix}^\top \begin{pmatrix} \partial_x f_i \\ \partial_u f_i \end{pmatrix} (t, \boldsymbol{X}_t^{\phi^r}, \boldsymbol{u}_t^{\phi^r}) \mathrm{d}t + (\boldsymbol{Y}_T^{\phi^r,\phi_h - z_h})^\top (\partial_x g_i)(\boldsymbol{X}_T^{\phi^r})\right] \mathrm{d}r.$$

with $\phi^r := \boldsymbol{z} + r(\phi - \boldsymbol{z})$.

The above expression follows directly from (3.11) for $\Phi(\boldsymbol{u})$ and (3.42) for $\frac{\delta V_i}{\delta u_i}$, by substituting ϕ with $\boldsymbol{z} + r(\phi - \boldsymbol{z})$, and ϕ'_h with $\phi_h - z_h$.

3.4.1 Quantifying α for closed-loop stochastic differential game

The following theorem is an analog of Theorem 3.3.2 and characterizes the constant α in (3.12) for closed-loop games with general cost functions.

Before stating the theorem, we introduce a few constants that will be used in the analysis. Let $i, j \in [N]$ with $i \neq j$, and define $\Delta^f_{i,j} = f_i - f_j$ and $\Delta^g_{i,j} = g_i - g_j$. Three constants $C^{i,j}_{V,1}, C^{i,j}_{V,2}, C^{i,j}_{V,3}$ are given by:

$$C^{i,j}_{V,1} := \|\partial^2_{x_i x_j} \Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{x_i u_j} \Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_i x_j} \Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_i u_j} \Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{x_i x_j} \Delta^g_{i,j}\|_{L^\infty}, \qquad (3.44)$$

$$\begin{aligned}C^{i,j}_{V,2} := &\sum_{h\in\{i,j\}} \Big(\|(\partial_{x_h} \Delta^f_{i,j})(\cdot, 0, 0)\|_{L^2} + \|(\partial_{u_h} \Delta^f_{i,j})(\cdot, 0, 0)\|_{L^2} + |(\partial_{x_h} \Delta^g_{i,j})(0)| \Big) \\ &+ \sum_{h\in\{i,j\},\ell\in[N]} \Big(\|\partial^2_{x_h x_\ell} \Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{x_h u_\ell} \Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_h x_\ell} \Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_h u_\ell} \Delta^f_{i,j}\|_{L^\infty} \\ &\quad + \|\partial^2_{x_h x_\ell} \Delta^g_{i,j}\|_{L^\infty} \Big), \qquad (3.45)\end{aligned}$$

$$
\begin{aligned}
C^{i,j}_{V,3} :=& \sum_{h\in[N]\setminus\{i,j\}} \Big(\|(\partial_{x_h}\Delta^f_{i,j})(\cdot,0,0)\|_{L^2} + \|(\partial_{u_h}\Delta^f_{i,j})(\cdot,0,0)\|_{L^2} + |(\partial_{x_h}\Delta^g_{i,j})(0)| \Big) \\
&+ \sum_{\substack{h\in[N]\setminus\{i,j\} \\ \ell\in[N]\setminus\{i,j\}}} \Big(\|\partial^2_{x_h x_\ell}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{x_h u_\ell}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_h x_\ell}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_h u_\ell}\Delta^f_{i,j}\|_{L^\infty} \\
&\qquad + \|\partial^2_{x_h x_\ell}\Delta^g_{i,j}\|_{L^\infty} \Big).
\end{aligned}
\tag{3.46}
$$

Theorem 3.4.2. *Suppose Assumption 3.3.2 holds and $\xi_i \in L^4(\Omega;\mathbb{R})$ for all $i \in [N]$. Then for all $\phi \in \Pi^N$ and $\phi'_i, \phi''_j \in \Pi$, it holds with $\overline{L}_y = \max\{L^b_y, L^\phi_y, L^{\phi'_i}_y, L^{\phi''_j}_y\}$ that*

$$
\left| \frac{\delta^2 V_i}{\delta\phi_i \delta\phi_j}(\phi;\phi'_i,\phi''_j) - \frac{\delta^2 V_j}{\delta\phi_j \delta\phi_i}(\phi;\phi''_j,\phi'_i) \right| \le C\left(C^{i,j}_{V,1} + \overline{L}_y \left(\frac{1}{N} C^{i,j}_{V,2} + \frac{1}{N^2} C^{i,j}_{V,3} \right) \right),
$$

where the constant $C \ge 0$ depends only on the upper bounds of T, $\max_{i\in[N]} \mathbb{E}[|\xi_i|^4]$, $\max_{i\in[N]} \|\sigma_i\|_{L^4}$, L^b, L^ϕ, $L^{\phi'_i}$, $L^{\phi''_j}$ and $\overline{L}_y$, and the constants $C^{i,j}_{V,1}$, $C^{i,j}_{V,2}$ and $C^{i,j}_{V,3}$, defined in (3.44), (3.45) *and* (3.46), *respectively, depend only on the upper bounds of the first- and second-order derivatives of $\Delta^f_{i,j}$ and $\Delta^g_{i,j}$ in (x,u).*

Consequently, if $\sup_{i\in[N],\phi_i\in\mathcal{A}_i}(L^{\phi_i} + L^{\phi_i}_y) < \infty$, then the game $\mathcal{G}^{cl}$ is an α-potential game with

$$
\alpha = C \max_{i\in[N]} \sum_{j\in[N]\setminus\{i\}} \left(C^{i,j}_{V,1} + L^b_y \left(\frac{1}{N} C^{i,j}_{V,2} + \frac{1}{N^2} C^{i,j}_{V,3} \right) \right),
$$

where $C \ge 0$ is a constant independent of the cost functions, and an α-potential function Φ is given by (3.11), *with $\frac{\delta V_i}{\delta \phi_i}$, $i \in [N]$, specified in* (3.42).

The proof of Theorem 3.4.2 is given in Section 3.7.3. The key step is to precisely quantify the moment of the sensitivity processes $\mathbf{Y}^{\phi,\phi'_h}$, $\mathbf{Z}^{\phi,\phi'_h,\phi''_\ell}$, $\boldsymbol{v}^{\phi,\phi'_h}$ and $\boldsymbol{w}^{\phi,\phi'_h,\phi''_\ell}$ in terms of the number of players N, the indices h, ℓ, and the coupling strength in the drift coefficients and the policies; see Propositions 3.7.7, 3.7.8 and 3.7.9. These moment estimates are more complex compared to those for the open-loop sensitivity processes (3.22) and (3.23), due to the coupling present in the closed-loop policies.

3.4.2 Examples of closed-loop α-potential games

We illustrate the application of Theorem 3.4.2 through two examples, i.e., distributed games and games with mean-field type interactions. In both cases, we obtain more explicit non-asymptotic bounds of α by leveraging structural conditions of the drift coefficients and cost functions.

Distributed games. If each player's admissible closed-loop policies depend only on her own state, then the closed-loop distributed game is an α-potential game under the same condition as the open-loop distributed game.

Example 3.4.1 (Distributed games)**.** *Consider the game $\mathcal{G}^{cl}$ defined as in* (3.35)-(3.36)*. Suppose Assumption 3.3.2 holds, and for all $i \in [N]$ and $\phi_i \in \mathcal{A}_i$, $(t,x,y) \mapsto b_i(t,x,y)$ and $(t,x,y) \mapsto \phi_i(t,x,y)$ are independent of y. Assume further that there exists $L, L^c \geq 0$ and $\beta \geq 1/2$ such that $\sup_{i\in[N],\phi_i\in\mathcal{A}_i} L^{\phi_i} \leq L$, $\max_{i\in[N]} \mathbb{E}[|\xi_i|^4] \leq L$, $\max_{i\in[N]} L^{b_i} \leq L$, and for all $i,j \in [N]$, $\Delta^f_{i,j} := f_i - f_j$ and $\Delta^g_{i,j} := g_i - g_j$ satisfy for all $(t,x,u) \in [0,T] \times \mathbb{R}^N \times \mathbb{R}^N$, $|(\partial^2_{x_i x_j}\Delta^f_{i,j})(t,x,u)| + |(\partial^2_{x_i u_j}\Delta^f_{i,j})(t,x,u)| + |(\partial^2_{u_i x_j}\Delta^f_{i,j})(t,x,u)| + |(\partial^2_{u_i u_j}\Delta^f_{i,j})(t,x,u)| + |(\partial^2_{x_i x_j}\Delta^g_{i,j})(x)| \leq L^c N^{-2\beta}$. Then $\mathcal{G}^{cl}$ is an α-potential game with $\alpha \leq C L^c N^{-(2\beta-1)}$, where $C \geq 0$ is a constant independent of N and β.*

We present a closed-loop analog of Example 3.3.2 for completeness. The result follows directly from Theorem 3.4.1 and the same argument as in Example 3.3.2. Additionally, it can be seen as a special case of the general characterization of closed-loop distributed potential games in [66, Theorem 3.1].

Example 3.4.2 (Distributed games with $\alpha = 0$)**.** *Consider the game $\mathcal{G}^{cl}$ defined as in* (3.35)-(3.36)*. Suppose Assumption 3.3.2 holds, and for all $i \in [N]$ and $\phi_i \in \mathcal{A}_i$, $(t,x,y) \mapsto b_i(t,x,y)$ and $(t,x,y) \mapsto \phi_i(t,x,y)$ are independent of y. Assume further that for all $i \in [N]$, f_i and g_i satisfy the conditions* (3.31) *and* (3.32)*. Then $\mathcal{G}^{cl}$ is a potential game.*

Games with mean-field type interactions. Building upon Theorem 3.4.2, we establish an analog of Example 3.3.3 for closed-loop games with mean-field type interactions. Due to the additional coupling introduced by the closed-loop controls, stronger conditions on the cost functions are required to ensure that the N-player game is an α-potential game with a decaying α.

The following example provides a sufficient condition for the closed-loop game $\mathcal{G}^{\mathrm{cl}}$ to be an α-potential game with $\alpha = \mathcal{O}(1/N)$. In contrast to the open-loop game described in Example 3.3.3, each player in general cannot have separate dependence on her individual behavior (i.e., without the function c_i in (3.34)), due to the coupling of closed-loop policies.

Example 3.4.3. *Consider the game $\mathcal{G}^{cl}$ defined as in* (3.35)-(3.36)*. Suppose Assumption 3.3.2 holds and there exists $L \geq 0$ such that $\sup_{i\in[N],\phi_i\in\mathcal{A}_i}(L^{\phi_i} + L^{\phi_i}_y) \leq L$, $\max_{i\in[N]} \mathbb{E}[|\xi_i|^4] \leq L$, $\max_{i\in[N]}(L^{b_i} + L^{b_i}_y) \leq L$ and $\max_{i\in[N]} \|\sigma_i\|_{L^\infty} \leq L$. Assume further that there exists $f_0 : [0,T] \times \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ and $g_0 : \mathbb{R}^N \to \mathbb{R}$ such that for all $i \in [N]$, f_i and g_i are of the form*

$$f_i(t,x,u) = f_0(t,x,u) + \bar{f}_i\left(t, \frac{1}{N}\sum_{\ell=1}^N \delta_{(x_\ell,u_\ell)}\right), g_i(x) = g_0(x) + \bar{g}_i\left(\frac{1}{N}\sum_{\ell=1}^N \delta_{x_\ell}\right),$$

where $\bar{f}_i : [0,T] \times \mathcal{P}_2(\mathbb{R} \times \mathbb{R}) \to \mathbb{R}$ and $\bar{g}_i : \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ are twice continuously differentiable with uniformly bounded second-order derivatives. Then $\mathcal{G}^{cl}$ is an α-potential game with $\alpha = C/N$ and the constant $C \geq 0$ independent of N.

3.5 Conditional McKean-Vlasov Control Problem for α-NE

Given an α-potential function $\mathcal{G}_{\text{diff}}$, this section establishes a dynamic programming approach to minimize the α-potential function Φ in (3.18) over $\mathcal{A}^{(N)}$, where the state process is governed by (3.14). For notational simplicity, we focus on open-loop games in this section. An analogous analysis can be applied to closed-loop games. The main difficulty is that the objective (3.18) depends on the aggregated behavior of the state processes with respect to $r \in [0,1]$, which acts as an additional noise independent of the Brownian motion W. Meanwhile, the admissible controls in $\mathcal{A}^{(N)}$ are adapted to a smaller filtration $\mathbb{F}$ that depends only on W but is independent of r. To apply the dynamic programming approach, we embed the optimization problem into a suitable conditional McKean–Vlasov (MKV) control problem.

3.5.1 Conditional MKV control problem

We start with some necessary notation: For each $t, s \in [0,T]$, let $W_s^t := W_{s\vee t} - W_t$ be the Brownian increment after time t, and let the filtration $\mathbb{F}^t$ be the $\mathbb{P}$-complement of the filtration generated by $W^t = (W_s^t)_{s\geq 0}$. Note that $\mathbb{F}^0$ coincides with $\mathbb{F}$. For each Euclidean space E, we denote by $\mathcal{P}_2(E)$ the set of probability measures μ on E with finite second moment, i.e., $\|\mu\|_2^2 := \int_E |x|^2 \mu(\mathrm{d}x) < \infty$. The space $\mathcal{P}_2(E)$ is equipped with the 2-Wasserstein distance. We assume without loss of generality (see e.g., [44]) that there exists a sub-σ-field $\mathcal{G} \subset \mathcal{F}$, which is independent of W and is "rich enough" in the sense that $\mathcal{P}_2(\mathcal{S}) = \{\mathcal{L}(\xi) \mid \xi \in L^2(\mathcal{G};\mathcal{S})\}$, where $\mathcal{L}(\xi)$ denotes the distribution of ξ under $\mathbb{P}$, $\mathcal{S} := \mathbb{R}^{(N+1)Nd} \times [0,1]$, and $L^2(\mathcal{G};\mathcal{S})$ is the set of $\mathcal{S}$-valued $\mathcal{G}$-measurable square integrable random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We define $\mathbb{G} := (\mathcal{G}_t)_{t\in[0,T]}$ to be the filtration generated by W, augmented with $\mathcal{G}$ and $\mathbb{P}$-null sets.

Now we introduce the MKV control problem associated with (3.18) and (3.14). The state process of the MKV control problem takes values in $\mathcal{S} := \mathbb{R}^{(N+1)Nd} \times [0,1]$, encompassing the original state process $\mathbf{X}^{\boldsymbol{u}}$, the sensitivity processes $(\mathbf{Y}^{\boldsymbol{u},u_i})_{i\in[N]}$, and the additional parameter r. More precisely, let $A = \prod_{i=1}^N A_i$, and fix $z = (z_i)_{i\in[N]} \in A$. For each $t \in [0,T]$, let $\mathcal{A}^t$ be the set of $\mathbb{F}^t$-progressively measurable square integrable processes taking values in A. Let $\mathcal{P}_2^{\text{Unif}}(\mathcal{S})$ be the space of measures $\nu \in \mathcal{P}_2(\mathcal{S})$ whose marginal $\nu|_{[0,1]}$ on $[0,1]$ is the uniform distribution:

$$\mathcal{P}_2^{\text{Unif}}(\mathcal{S}) := \{\nu \in \mathcal{P}_2(\mathcal{S}) \mid \nu|_{[0,1]} = \text{Unif}(0,1)\}.$$

For each $\nu \in \mathcal{P}_2^{\text{Unif}}(\mathcal{S})$, $\boldsymbol{u} \in \mathcal{A}^t$, and $(\xi, \mathfrak{r}) \in L^2(\mathcal{G};\mathcal{S})$ with $\mathcal{L}(\xi, \mathfrak{r}) = \nu$, consider the process $\mathbb{X}^{t,\xi,\mathfrak{r},\boldsymbol{u}}$ governed by the following dynamics, which concentrates the state process (3.14) and the sensitivity processes (3.16):

$$\mathbb{X}_s^{t,\xi,\mathfrak{r},\boldsymbol{u}} = \xi + \int_t^s B(v, \mathbb{X}_v^{t,\xi,\mathfrak{r},\boldsymbol{u}}, \mathfrak{r}, \boldsymbol{u}_v)\mathrm{d}v + \int_t^s \Sigma(v, \mathbb{X}_v^{t,\xi,\mathfrak{r},\boldsymbol{u}}, \mathfrak{r}, \boldsymbol{u}_v)\mathrm{d}W_v^t, \quad s \in [t,T], \tag{3.47}$$

where $B = \mathsf{vcat}(B_1, \dots, B_{N+1}) : [0,T] \times \mathcal{S} \times A \to \mathbb{R}^{(N+1)Nd}$ is defined by: for all $t \in [0,T]$, $\mathbb{x} = \mathsf{vcat}(x, y_1, \dots, y_N) \in \mathbb{R}^{(N+1)Nd}$, $r \in [0,1]$, and $u = (u_i)_{i\in[N]} \in A$, $B_1(t, \mathbb{x}, r, u) := \mathsf{vcat}(b_1(t, x, z + r(u - z)), \dots, b_N(t, x, z + r(u - z)))$ and for any $i \in [N]$,

$$B_{i+1}(t, \mathbb{x}, r, u) := \begin{pmatrix} (\partial_x b_1)(t, x, z + r(u-z))\, y_i + (\partial_{u_i} b_1)(t, x, z + r(u-z))\,(u_i - z_i) \\ \vdots \\ (\partial_x b_N)(t, x, z + r(u-z))\, y_i + (\partial_{u_i} b_N)(t, x, z + r(u-z))\,(u_i - z_i) \end{pmatrix},$$

and $\Sigma = (\Sigma_1, \dots, \Sigma_m) : [0,T] \times \mathcal{S} \times A \to \mathbb{R}^{(N+1)Nd\times m}$ is defined such that for all $k = 1, \dots, m$, Σ_k is defined in the same way as B, but with b_i replaced by σ_{ik} for all $i \in [N]$. Under Assumption 3.3.1, $(\mathbb{X}^{t,\xi,\mathfrak{r},\boldsymbol{u}}, \mathfrak{r})$ is a uniquely defined $\mathcal{S}$-valued $\mathbb{G}$-adapted square integrable process. Moreover, as $\mathfrak{r}$ is independent of $\mathcal{F}^t_s$ and is stationary in time, the conditional law $\mu^{t,\xi,\mathfrak{r},\boldsymbol{u}}_s := \mathcal{L}(\mathbb{X}^{t,\xi,\mathfrak{r},\boldsymbol{u}}_s, \mathfrak{r}|\mathcal{F}^t_s)$, $s \in [t,T]$, is a $\mathcal{P}^{\mathrm{Unif}}_2(\mathcal{S})$-valued $\mathbb{G}$-optional process (see [50, Lemma A.1]).

Consider the following cost functional, which is a dynamic version of the α-potential function (3.12):

$$J(t, \xi, \mathfrak{r}, \boldsymbol{u}) := \mathbb{E}\left[\int_t^T \left\langle F(s, \cdot, \cdot, \boldsymbol{u}_s), \mu^{t,\xi,\mathfrak{r},\boldsymbol{u}}_s \right\rangle \mathrm{d}s + \left\langle G, \mu^{t,\xi,\mathfrak{r},\boldsymbol{u}}_T \right\rangle\right], \tag{3.48}$$

where $\mu^{t,\xi,\mathfrak{r},\boldsymbol{u}}_s = \mathcal{L}(\mathbb{X}^{t,\xi,\mathfrak{r},\boldsymbol{u}}_s, \mathfrak{r}|\mathcal{F}^t_s)$, $F : [0,T] \times \mathcal{S} \times A \to \mathbb{R}$ and $G : \mathcal{S} \to \mathbb{R}$ are defined by: for all $t \in [0,T]$, $\mathbb{x} = \mathsf{vcat}(x, y_1, \dots, y_N) \in \mathbb{R}^{(N+1)Nd}$, $r \in [0,1]$ and $u = (u_i)_{i\in[N]} \in A$,

$$\begin{aligned} F(t, \mathbb{x}, r, u) &:= \sum_{j=1}^N \begin{pmatrix} y_j \\ u_j - z_j \end{pmatrix}^\top \begin{pmatrix} \partial_x f_j \\ \partial_{u_j} f_j \end{pmatrix} (t, x, z + r(u - z)), \\ G(\mathbb{x}, r) &:= \sum_{j=1}^N y_j^\top (\partial_x g_j)(x), \end{aligned} \tag{3.49}$$

where $\langle h, \mu \rangle$ denotes the integral of the function h with respect to the measure μ.

The following proposition identifies minimizing the α-potential function Φ in (3.18) as solving an MKV control problem with a specific initial condition. The result relies on the crucial observation that the cost functional J in (3.48) satisfies the law invariance property [43, 50], i.e., it depends on the law of $(\xi, \mathfrak{r})$ instead of the specific choice of the random variable $(\xi, \mathfrak{r})$ itself.

Proposition 3.5.1. *Suppose Assumption 3.3.1 holds. Let J be defined in* (3.48)*. For all $(t, \nu) \in [0,T] \times \mathcal{P}^{\mathrm{Unif}}_2(\mathcal{S})$, $\boldsymbol{u} \in \mathcal{A}^t$, and $(\xi, \mathfrak{r}), (\xi', \mathfrak{r}') \in L^2(\mathcal{G}; \mathcal{S})$ with law ν, $J(t, \xi, \mathfrak{r}, \boldsymbol{u}) = J(t, \xi', \mathfrak{r}', \boldsymbol{u})$.*

Consequently, the optimal value function for minimizing (3.48) *can be identified as $V : [0,T] \times \mathcal{P}^{\mathrm{Unif}}_2(\mathcal{S}) \to \mathbb{R} \cup \{-\infty, \infty\}$:*

$$V(t, \nu) := \inf_{\boldsymbol{u}\in\mathcal{A}^t} J(t, \xi, \mathfrak{r}, \boldsymbol{u}), \tag{3.50}$$

for any $(\xi, \mathfrak{r}) \in L^2(\mathcal{G}; \mathcal{S})$ *with* $\mathcal{L}(\xi, \mathfrak{r}) = \nu$. *Moreover, let* Φ *be defined in* (3.18)*, we have*

$$V\left(0, \delta_{\mathsf{vcat}(x_1,\dots,x_N,0_{N^2d})} \otimes \mathrm{Unif}(0,1)\right) = \inf_{\boldsymbol{u} \in \mathcal{A}^{(N)}} \Phi(\boldsymbol{u}), \tag{3.51}$$

where x_i *is the initial state of* (3.14) *and* $0_{N^2d} \in \mathbb{R}^{N^2d}$ *is the zero vector.*

The law invariance of J follows from the fact that each $\boldsymbol{u} \in \mathcal{A}^t$ is adapted to the filtration of W, and is independent of $\mathcal{G}$. Then by the strong uniqueness of (3.47), it holds that for all $(\xi, \mathfrak{r}), (\xi', \mathfrak{r}') \in L^2(\mathcal{G}; \mathcal{S})$ with law ν, $\mathcal{L}(\mathbb{X}^{t,\xi,\mathfrak{r},\boldsymbol{u}}, \boldsymbol{u}) = \mathcal{L}(\mathbb{X}^{t,\xi',\mathfrak{r}',\boldsymbol{u}}, \boldsymbol{u})$, and hence $J(t, \xi, \mathfrak{r}, \boldsymbol{u}) = J(t, \xi', \mathfrak{r}', \boldsymbol{u})$ (see [50, Proposition 2.4]). The identity (3.51) follows from $\mathcal{A}^0 = \mathcal{A}^{(N)}$ and by the law of iterated conditional expectations, $\Phi(\boldsymbol{u}) = J(t, \xi, \mathfrak{r}, \boldsymbol{u})$ with $\xi = \mathsf{vcat}(x_1, \dots, x_N, 0_{N^2d})$ and a uniform random variable $\mathfrak{r} \in L^2(\mathcal{G}; [0,1])$.

We remark that (3.50) is a specific stochastic control problem with conditional MKV dynamics, where the state (3.47) does not involve law dependence, and the cost functions (3.48) depend linearly on the conditional distribution. As a result, the dynamic programming approach, developed for general MKV control problems in [119, 50], can be applied to minimize the α-potential function Φ.

3.5.2 HJB equation for the α-potential function

In the section, we identify the optimal value function (3.50) as a solution of an HJB equation. We will adopt the notion of linear derivative with respect to probability measures as in [73, 48, 65], as it allows for a clear distinction between the derivatives with respect to the marginal laws of $\mathbb{X}^{t,\xi,\mathfrak{r},\boldsymbol{u}}$ and $\mathfrak{r}$; see Remark 3.5.1.

Specially, we say a function $\phi : [0,T] \times \mathcal{P}_2(E) \to \mathbb{R}$ is in $C^{1,2}([0,T] \times \mathcal{P}_2(E))$ if there exist continuous functions $\frac{\delta \phi}{\delta \mu} : [0,T] \times \mathcal{P}_2(E) \times E \to \mathbb{R}$ and $\frac{\delta^2 \phi}{\delta \mu^2} : [0,T] \times \mathcal{P}_2(E) \times E \times E \to \mathbb{R}$ such that $\frac{\delta^2 \phi}{\delta \mu^2}$ is symmetric in its last two arguments and the following properties hold:

- continuously differentiable: $\partial_t \phi(t, \mu)$, $\partial_v \frac{\delta \phi}{\delta \mu}(t, \mu, v)$, $\partial^2_{vv} \frac{\delta \phi}{\delta \mu}(t, \mu, v)$ and $\partial^2_{vv'} \frac{\delta^2 \phi}{\delta^2 \mu}(t, \mu, v, v')$ exist and are continuous in (t, μ, v, v').
- locally uniform bound: for any compact $K \subset \mathcal{P}_2(E)$, there exists $c_K \geq 0$ such that for all $(t, \mu) \in [0,T] \times K$ and $v, v' \in E$, $|\partial_v \frac{\delta \phi}{\delta \mu}(t, \mu, v)| \leq c_K(1 + |v|)$, $|\partial^2_{vv} \frac{\delta \phi}{\delta \mu}(t, \mu, v)| + |\partial^2_{vv'} \frac{\delta^2 \phi}{\delta^2 \mu}(t, \mu, v, v')| \leq c_K$.
- fundamental theorem of calculus: for all $\mu, \nu \in \mathcal{P}_2(E)$ and $t \in [0,T]$,

$$\begin{aligned}
\phi(t, \mu) - \phi(t, \nu) &= \int_0^1 \int_E \frac{\delta \phi}{\delta \mu}(t, \lambda \mu + (1-\lambda)\nu, v)(\mu - \nu)(\mathrm{d}v)\mathrm{d}\lambda, \\
\phi(t, \mu) - \phi(t, \nu) &- \int_E \frac{\delta \phi}{\delta \mu}(t, \nu, v)(\mu - \nu)(\mathrm{d}v) \\
&= \int_0^1 \int_0^r \int_{E \times E} \frac{\delta^2 \phi}{\delta^2 \mu}(t, s\mu + (1-s)\nu, v, v')(\mu - \nu)(\mathrm{d}v)(\mu - \nu)(\mathrm{d}v')\mathrm{d}s\mathrm{d}r.
\end{aligned}$$

For each $u \in A$, $\phi \in C^{1,2}([0,T] \times \mathcal{P}_2(\mathcal{S}))$, $t \in [0,T]$, and $\mu \in \mathcal{P}_2(\mathcal{S})$, define the function $\mathbb{L}^u\phi(t,\mu) : \mathcal{S} \to \mathbb{R}$ by

$$\mathbb{L}^u\phi(t,\mu)(\mathbb{x},r) \coloneqq B(t,\mathbb{x},r,u)^\top \partial_{\mathbb{x}} \frac{\delta\phi}{\delta\mu}(t,\mu,\mathbb{x},r) + \frac{1}{2}\operatorname{tr}\left((\Sigma\Sigma^\top)(t,\mathbb{x},r,u)\partial^2_{\mathbb{x}\mathbb{x}}\frac{\delta\phi}{\delta\mu}(t,\mu,\mathbb{x},r)\right) \tag{3.52}$$

with B and Σ in (3.47), and define the function $\mathbb{M}^u\phi(t,\mu) : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ by

$$\mathbb{M}^u\phi(t,\mu)(\mathbb{x},r,\mathbb{x}',r') \coloneqq \frac{1}{2}\operatorname{tr}\left(\Sigma(t,\mathbb{x},r,u)\Sigma(t,\mathbb{x}',r',u)^\top \partial^2_{\mathbb{x}\mathbb{x}'}\frac{\delta^2\phi}{\delta^2\mu}(t,\mu,\mathbb{x},r,\mathbb{x}',r')\right). \tag{3.53}$$

Note that under Assumption 3.3.1, $\mathbb{L}^u\phi(t,\mu) \in L^1(\mathcal{S},\mu)$ and $\mathbb{M}^u\phi(t,\mu) \in L^1(\mathcal{S}\times\mathcal{S},\mu\otimes\mu)$. Define the Hamiltonian

$$\hat{H}(t,\mu,\phi,u) \coloneqq \langle \mathbb{L}^u\phi(t,\mu),\mu\rangle + \langle\ \mathbb{M}^u\phi(t,\mu),\mu\otimes\mu\rangle + \langle F(t,\cdot,\cdot,u),\mu\rangle, \tag{3.54}$$

with F defined in (3.49).

Remark 3.5.1. *As $\mathfrak{r}$ is stationary in* (3.47)*, the operators $\mathbb{L}^u$ and $\mathbb{M}^u$ only involve the partial derivative with respect to the $\mathbb{x}$-component, and not the derivative with respect to the r-component. One can equivalently express these operators using the Lions derivatives as in [119]. Indeed, let $\partial_\mu\phi$ be the Lions derivative of ϕ,*

$$\begin{aligned}\mathbb{L}^u\phi(t,\mu)(\mathbb{x},r) = &\begin{pmatrix} B(t,\mathbb{x},r,u) \\ 0 \end{pmatrix}^\top \partial_\mu\phi(t,\mu)(\mathbb{x},r) \\ &+ \frac{1}{2}\operatorname{tr}\left(\begin{pmatrix} (\Sigma\Sigma^\top)(t,\mathbb{x},r,u) & 0_{N(N+1)d} \\ 0^\top_{N(N+1)d} & 0 \end{pmatrix}\partial_{(\mathbb{x},r)}\partial_\mu\phi(t,\mu)(\mathbb{x},r)\right),\end{aligned}$$

due to the relation $\partial_\mu\phi = \partial_{(\mathbb{x},r)}\frac{\delta\phi}{\delta\mu}$ (see [32, Proposition 5.48]). Similar expression holds for $\mathbb{M}^u\phi$. We adopt the expressions (3.52) *and* (3.53) *to simplify the notation.*

We now present a verification theorem, which constructs an optimal control of (3.50) (and (3.18)) in an analytic feedback form using a smooth solution to an HJB equation in the Wasserstein space.

Theorem 3.5.1. *Suppose Assumption 3.3.1 holds. Let $v \in C^{1,2}([0,T]\times\mathcal{P}_2(\mathcal{S}))$ be such that for a constant $C \geq 0$,*

$$|v(t,\mu)| \leq C(1+\|\mu\|_2^2), \qquad \left|\partial_{(\mathbb{x},r)}\frac{\delta v}{\delta\mu}(t,\mu,\mathbb{x},r)\right| \leq C(1+|\mathbb{x}|+\|\mu\|_2),$$

for any $(t,\mu) \in [0,T]\times\mathcal{P}_2^{\mathrm{Unif}}(\mathcal{S}), (\mathbb{x},r) \in \mathcal{S}$. Assume that $\inf_{u\in A}\hat{H}(t,\mu,v,u) \in \mathbb{R}$ for all (t,μ), and v satisfies the following HJB equation:

$$\begin{cases} \partial_t w(t,\mu) + \min\limits_{u\in A}\hat{H}(t,\mu,w,u) = 0, \quad (t,\mu) \in [0,T)\times\mathcal{P}_2^{\mathrm{Unif}}(\mathcal{S}), \\ \qquad\qquad\qquad\qquad w(T,\mu) = \langle G,\mu\rangle, \quad \mu \in \mathcal{P}_2^{\mathrm{Unif}}(\mathcal{S}). \end{cases} \tag{3.55}$$

Assume further that there exists a measurable map $\hat{a} : [0,T] \times \mathcal{P}_2^{\mathrm{Unif}}(\mathcal{S}) \to A$ such that for all $(t,\mu) \in [0,T] \times \mathcal{P}_2^{\mathrm{Unif}}(\mathcal{S})$,

$$\hat{a}(t,\mu) \in \arg\min_{u \in A} \hat{H}(t,\mu,v,u), \tag{3.56}$$

for any $(\xi,\mathfrak{r}) \in L^2(\mathcal{G};\mathcal{S})$ with law μ, the following equation

$$\begin{aligned} \hat{\mathbb{X}}_s =& \xi + \int_t^s B\left(v, \hat{\mathbb{X}}_v, \mathfrak{r}, \hat{a}\big(v, \mathcal{L}(\hat{\mathbb{X}}_v, \mathfrak{r} \mid \mathcal{F}_v^t)\big)\right) \mathrm{d}v \\ &+ \int_t^s \Sigma\left(v, \hat{\mathbb{X}}_v, \mathfrak{r}, \hat{a}\big(v, \mathcal{L}(\hat{\mathbb{X}}_v, \mathfrak{r} \mid \mathcal{F}_v^t)\big)\right) \mathrm{d}W_v^t, \quad s \in [t,T] \end{aligned} \tag{3.57}$$

admits a square integrable solution $\hat{\mathbb{X}}^{t,\xi,\mathfrak{r}}$, and the feedback control $\hat{\boldsymbol{u}}_s^{t,\xi,\mathfrak{r}} := \hat{a}(s, \mathcal{L}(\hat{\mathbb{X}}_s^{t,\xi,\mathfrak{r}}, \mathfrak{r} \mid \mathcal{F}_s^t))$, $s \in [t,T]$, is in $\mathcal{A}^t$. Then $v = V$ on $[0,T] \times \mathcal{P}_2^{\mathrm{Unif}}(\mathcal{S})$, and for all $(t,\mu) \in [0,T] \times \mathcal{P}_2^{\mathrm{Unif}}(\mathcal{S})$, $\hat{\boldsymbol{u}}^{t,\xi,\mathfrak{r}} \in \mathcal{A}^t$ with $\mathcal{L}(\xi,\mathfrak{r}) = \mu$ is an optimal control for $V(t,\mu)$.

Consequently, given $\xi = \mathsf{vcat}(x_1, \dots, x_N, 0_{N^2 d})$ and a uniform random variable $\mathfrak{r} \in L^2(\mathcal{G};[0,1])$, the control $\hat{\boldsymbol{u}}^{0,\xi,\mathfrak{r}} \in \mathcal{A}^{(N)}$ minimizes the α-potential function Φ given in (3.18), thus is an α-Nash equilibrium of the game $\mathcal{G}^{op}$ in Section 3.3.

Theorem 3.5.1 only requires the function v to satisfy the HJB equation (3.55) on the subspace $\mathcal{P}_2^{\mathrm{Unif}}(\mathcal{S})$, rather than on the entire space $\mathcal{P}_2(\mathcal{S})$ as is the case for general MKV control problems [119, 50]. This is due to the fact that the flow $(\mu_s^{t,\xi,\mathfrak{r},\boldsymbol{u}})_{s\in[t,T]} = (\mathcal{L}(\mathbb{X}_s^{t,\xi,\mathfrak{r},\boldsymbol{u}}, \mathfrak{r} | \mathcal{F}_s^t))_{s\in[t,T]}$ remains in $\mathcal{P}_2^{\mathrm{Unif}}(\mathcal{S})$ for any control $\boldsymbol{u} \in \mathcal{A}^t$. Restricting the domain of (3.55) to $\mathcal{P}_2^{\mathrm{Unif}}(\mathcal{S})$ is essential for minimizing (3.18) analytically in linear-quadratic games; see Section 3.6.

Theorem 3.5.1 adapts [119, Theorem 4.2] to the present setting. Compared with [119], since we do not assume the compactness of the action space, we introduce the additional assumption on the finiteness of $\inf_{u\in A} \hat{H}(t,\mu,v,u)$. With this assumption in place, the proof follows directly along the same lines as the same lines of the verification theorem [119, Theorem 4.2].

Indeed, fix $(t,\mu) \in [0,T] \times \mathcal{P}_2^{\mathrm{Unif}}(\mathcal{S})$ and $(\xi,\mathfrak{r}) \in L^2(\mathcal{G};\mathcal{S})$ with law μ. For any $\boldsymbol{u} \in \mathcal{A}^t$, applying Itô's formula in [65] (see also [33, Theorem 4.17]) to $s \mapsto v(s, \mathcal{L}(\mathbb{X}_s^{t,\xi,\mathfrak{r},\boldsymbol{u}}, \mathfrak{r} \mid \mathcal{F}_s^t))$ and using the fact that $\mathcal{L}(\mathbb{X}_s^{t,\xi,\mathfrak{r},\boldsymbol{u}}, \mathfrak{r} \mid \mathcal{F}_s^t)$ lies in $\mathcal{P}_2^{\mathrm{Unif}}(\mathcal{S})$ and the condition (3.55) of v yield that $v(t,\mu) \le J(t,\xi,\tau,\boldsymbol{u})$, which implies that $v(t,\mu) \le V(t,\mu)$. The condition (3.56) of $\hat{a}$ and the assumption $\hat{\boldsymbol{u}}^{t,\xi,\mathfrak{r}} \in \mathcal{A}^t$ imply the optimality of $\hat{\boldsymbol{u}}^{t,\xi,\mathfrak{r}}$. A special case of Theorem 3.5.1 for a class of linear-quadratic games is presented in Theorem 3.6.1, with a detailed proof provided.

3.6 A Toy Example: Linear-Quadratic α-Potential Games

In this section, we illustrate our results through a simple open-loop linear-quadratic (LQ) game $\mathcal{G}_{\mathrm{LQ}}$ on an undirected graph $G = (V,E)$. The vertex of the graph is the set of players

$V = [N]$, and each edge between the vertices represents a dependency between the associated players. The objective function of player i in this game is given by

$$V_i(\boldsymbol{u}) = \mathbb{E}\left[\int_0^T \left(u_{i,t}^2 + \frac{1}{N}\sum_{j=1}^N q_{ij}(X_{i,t}^{\boldsymbol{u}} - X_{j,t}^{\boldsymbol{u}})^2\right) \mathrm{d}t + \gamma_i (X_{i,T}^{\boldsymbol{u}} - d_i)^2\right], \tag{3.58}$$

where $q_{ij}, \gamma_i \geq 0$, $d_i \in \mathbb{R}$, and for any $\boldsymbol{u} = (u_i)_{i\in[N]} \in \mathcal{H}^2(\mathbb{R}^N)$, the state process $\mathbf{X}_t^{\boldsymbol{u}}$ is governed by:

$$\mathrm{d}X_{i,t} = (a_i(t)X_{i,t} + u_{i,t})\,\mathrm{d}t + \sigma_i(t)\mathrm{d}W_t^i, \quad X_{i,0} = x_i, \quad t \in [0,T],\ i \in [N], \tag{3.59}$$

where $x_i \in \mathbb{R}$, $a_i, \sigma_i : [0,T] \to \mathbb{R}$ are (possibly distinct) continuous functions. Player i's aims to minimize (3.58) over the control set

$$\mathcal{A}_i = \{u_i \in \mathcal{H}^2(\mathbb{R}) | \|u\|_{\mathcal{H}^2(\mathbb{R})} \leq L\}, \tag{3.60}$$

where $L > 0$ is a given sufficiently large constant.

The above game can be viewed as a crowd flocking game [95, 8, 37]. The goal is for all players to reach their respective destinations by a specified terminal time. During the game, players exhibit a tendency to group together, mimicking the collective behavior observed in natural flocks or herds. This phenomenon, known as flocking, is driven by factors such as safety, efficiency, and social interaction.

3.6.1 Quantifying α for $\mathcal{G}_{\mathrm{LQ}}$

Since the dynamics (3.59) is decoupled, Theorem 3.3.2 and Remark 3.3.2 imply that $\mathcal{G}_{\mathrm{LQ}}$ is an α_N-potential game with

$$\alpha_N \leq C\frac{1}{N}\max_{i\in[N]}\sum_{j\neq i}|q_{ji} - q_{ij}|. \tag{3.61}$$

Suppose that the constants $(q_{ij}, \gamma_i, d_i, x_i)_{i,j\in[N]}, L$ and the sup-norms of $(a_i, \sigma_i)_{i\in[N]}$ are uniformly bounded in N. Then, an explicit bound for α_N in terms of the number of players N and the strength and symmetry of players' interactions can be obtained, as illustrated below:

- **Symmetric interaction.** If the interaction weights $(q_{ij})_{i,j\in[N]}$ satisfy the *pairwise* symmetry condition $q_{ij} = q_{ji}$ for all $i, j \in [N]$, then $\mathcal{G}_{\mathrm{LQ}}$ is a potential game, i.e., $\alpha_N \equiv 0$ regardless of the number of players N. This symmetry condition is common in many interaction kernels, where player i's influence on player j depends only on the distance between them [8, 32, 10].

- **Asymmetric interaction.** Suppose that the graph G has a bounded degree $k := \max_{i\in V}\deg(i)$ for some $k \geq 2$, i.e., each vertex is connected to at most k vertices. Assume further that the interaction weights $(q_{ij})_{i,j\in[N]}$ exhibit an exponential decay of the form
$$q_{ij} = w_i \eta^{c(i,j)}, \quad \forall i, j \in [N], \tag{3.62}$$
where $(w_i)_{i\in[N]}$ are distinct positive constants that are uniformly bounded in N, $\eta \in (0, 1)$ is a given constant, and $c(i, j)$ is the (shortest-path) distance between vertices i and j. Such a structure models localized interactions, where a player's impact is strongest on their immediate neighbors and diminishes further away [58, 59, 64]. For clarity of exposition, we assume a sufficiently fast decay rate η satisfying $\eta < 1/k$.

  In this setting, by (3.61), there exists a constant $C \geq 0$, independent of η, k and N, such that
$$\alpha_N \leq \frac{C}{N}\max_{i\in[N]}\sum_{j\neq i}\eta^{c(i,j)} \leq \frac{C}{N}\sum_{\ell=1}^{\infty}\eta^{\ell}k^{\ell} = C\frac{\eta k}{(1-\eta k)N}, \tag{3.63}$$
where the second inequality used the fact that, for any vertex $v \in V$, the number of vertices at distance ℓ from v is at most k^{ℓ}. The bound (3.63) demonstrates that α_N decays to zero as the number of players increases. Additionally, α_N vanishes as $\eta \to 0$, reflecting the weakening interactions among players.

3.6.2 Constructing α-NE for $\mathcal{G}_{\mathrm{LQ}}$

An α_N-NE of $\mathcal{G}_{\mathrm{LQ}}$ can be constructed by minimizing the corresponding α_N-potential function (3.18). The structure of $\mathcal{G}_{\mathrm{LQ}}$ significantly simplifies the α_N-potential function compared to the general case studied in Sections 3.3 and 3.5. Indeed, as $X_i^{\boldsymbol{u}}$ depends only on u_i, the sensitivity processes $Y_{t,j}^{\boldsymbol{u},u_i'} \equiv 0$ for $i \neq j$, reducing the dimension of the state process in (3.18) from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$. Moreover, due to the LQ structure (3.58)-(3.59), the α-potential function becomes a LQ control problem, whose minimizer can be solved analytically.

We consider an extended state dynamics including both the original state dynamics (3.59) for $\mathbf{X}^{\boldsymbol{u}}$, and the dynamics for the sensitivity processes $(Y_i^{\boldsymbol{u},u_i})_{i\in[N]}$. Specifically, fix a uniform random variable $\mathfrak{r} \in L^2(\mathcal{G}; [0, 1])$, and for each $\boldsymbol{u} \in \mathcal{H}^2(\mathbb{R}^N)$, consider the $\mathbb{R}^{2N}$-valued $\mathbb{G}$-adapted square integrable process $\mathbb{X}^{\mathfrak{r},\boldsymbol{u}}$ governed by
$$\mathrm{d}\mathbb{X}_t = (A(t)\mathbb{X}_t + \mathcal{I}_{\mathfrak{r}}\boldsymbol{u}_t)\,\mathrm{d}t + \Sigma(t)\mathrm{d}W_t, \quad \mathbb{X}_0 = \mathsf{vcat}(x_1, \cdots, x_N, 0_N), \tag{3.64}$$
where $\mathcal{I}_{\mathfrak{r}} := \mathsf{vcat}(\mathfrak{r}\mathbb{I}_N, \mathbb{I}_N) \in \mathbb{R}^{2N\times N}$, $A(t) := \mathrm{diag}(\tilde{A}(t), \tilde{A}(t)) \in \mathbb{S}^{2N}$ with $\tilde{A}(t) := \mathrm{diag}(a_1(t), \cdots, a_N(t))$, and $\Sigma(t) = \mathsf{vcat}(\sigma(t), 0_{N\times N}) \in \mathbb{R}^{2N\times N}$ with $\sigma(t) := \mathrm{diag}(\sigma_1(t), \cdots, \sigma_N(t))$. The α-potential function Φ for $\mathcal{G}_{\mathrm{LQ}}$ is given by (see (3.12) and (3.48) with $z = 0$):
$$\Phi(\boldsymbol{u}) = \mathbb{E}\left[\int_0^T \int_{\mathcal{S}} \left(\mathbb{x}^\top Q\mathbb{x} + 2r\boldsymbol{u}_t^\top \boldsymbol{u}_t\right)\mathrm{d}\mu_t^{\mathfrak{r},\boldsymbol{u}}\mathrm{d}t + \int_{\mathcal{S}}\left(\mathbb{x}^\top \bar{Q}\mathbb{x} + 2\mathfrak{p}^\top\mathbb{x}\right)\mathrm{d}\mu_T^{\mathfrak{r},\boldsymbol{u}}\right], \tag{3.65}$$

where $\mathcal{S} := \mathbb{R}^{2N} \times [0,1]$, $\mu_t^{\mathfrak{r},\boldsymbol{u}} := \mathcal{L}(\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}}, \mathfrak{r} | \mathcal{F}_t)$ for all t, $Q := \begin{pmatrix} \mathbb{0}_N & \tilde{Q}^\top \\ \tilde{Q} & \mathbb{0}_N \end{pmatrix} \in \mathbb{S}^{2N}$ with $\tilde{Q} \in \mathbb{R}^{N\times N}$ given by $\tilde{Q}_{i,i} = \frac{1}{N}\sum_{k\neq i, k\in[N]} q_{ik}$ and $\tilde{Q}_{i,j} = -\frac{q_{ij}}{N}$ for all $i \neq j$, $\bar{Q} := \begin{pmatrix} \mathbb{0}_N & \Gamma \\ \Gamma & \mathbb{0}_N \end{pmatrix} \in \mathbb{S}^{2N}$ with $\Gamma := \operatorname{diag}(\gamma_1, \cdots, \gamma_N) \in \mathbb{S}^N$, and $\mathfrak{p} := -\mathsf{vcat}(0_N, \gamma_1 d_1, \cdots, \gamma_N d_N) \in \mathbb{R}^{2N}$. Above and hereafter, for each $n \in \mathbb{N}$, we denote by $\mathbb{S}^n$ the space of $n \times n$ symmetric matrices, by $\mathbb{0}_n$ the $n \times n$ zero matrix, and by $\operatorname{diag}(a_1, \ldots, a_n)$ the diagonal matrix with diagonal elements $(a_1, \ldots, a_n)$.

The minimizer of (3.65) can be characterized with suitable ordinary differential equations (ODEs). These ODEs differ from the Riccati equations for usual LQ control problems studied in [133], due to the additional dependence on the parameter r in (3.64) and (3.65). To see this, let $M_0 \in C^1([0,T]; \mathbb{S}^{2N})$ satisfy the following linear ODE:

$$\dot{M}_0 + A^\top M_0 + M_0 A + Q = 0; \quad M_0(T) = \bar{Q}, \tag{3.66}$$

where the dot denotes the time derivative. Consider the following Riccati equation for $M_1 \in C^1([0,T]; \mathbb{S}^{4N})$:

$$\dot{M}_1 + \begin{pmatrix} A & \mathbb{0}_{2N} \\ \mathbb{0}_{2N} & A \end{pmatrix} M_1 + M_1 \begin{pmatrix} A & \mathbb{0}_{2N} \\ \mathbb{0}_{2N} & A \end{pmatrix} - K_{M_0,M_1}^\top K_{M_0,M_1} = 0; \quad M_1(T) = \mathbb{0}_{4N}, \tag{3.67}$$

with $K_{M_0,M_1} : [0,T] \to \mathbb{R}^{N\times 4N}$ defined by

$$K_{M_0,M_1} := \begin{pmatrix} \begin{pmatrix} \mathbb{0}_N & \mathbb{I}_N \end{pmatrix} M_0 & \begin{pmatrix} \mathbb{I}_N & \mathbb{0}_N \end{pmatrix} M_0 \end{pmatrix} + \tilde{I} M_1, \quad \tilde{I} := \begin{pmatrix} \frac{1}{2}\mathbb{I}_N & \mathbb{I}_N & \frac{1}{3}\mathbb{I}_N & \frac{1}{2}\mathbb{I}_N \end{pmatrix} \in \mathbb{R}^{N\times 4N}. \tag{3.68}$$

The constants in $\tilde{I}$ correspond to $\mathbb{E}[\mathfrak{r}]$ and $\mathbb{E}[\mathfrak{r}^2]$ for the uniform random variable $\mathfrak{r}$ in (3.64). Given a solution M_1 to (3.67), consider the following linear ODE for $M_2 \in C^1([0,T]; \mathbb{R}^{4N})$:

$$\dot{M}_2 + \begin{pmatrix} A & \mathbb{0}_{2N} \\ \mathbb{0}_{2N} & A \end{pmatrix} M_2 - K_{M_0,M_1}^\top \tilde{I} M_2 = 0; \quad M_2(T) = \begin{pmatrix} \mathfrak{p} \\ 0_{2N} \end{pmatrix}. \tag{3.69}$$

The following theorem constructs a minimizer of $\mathcal{H}^2(\mathbb{R}^N) \ni \boldsymbol{u} \mapsto \Phi(\boldsymbol{u}) \in \mathbb{R}$ based on solutions of (3.66), (3.67) and (3.69), which subsequently yields an α_N-NE of the game $\mathcal{G}_{\mathrm{LQ}}$. The proof is given in Section 3.7.4.

Theorem 3.6.1. *Suppose that $M_0 \in C^1([0,T]; \mathbb{S}^{2N})$, $M_1 \in C^1([0,T]; \mathbb{S}^{4N})$, and $M_2 \in C^1([0,T]; \mathbb{R}^{4N})$ satisfy* (3.66), (3.67), *and* (3.69), *respectively. Define*

$$\boldsymbol{u}_t^* = -K_{M_0,M_1}(t) \begin{pmatrix} \mathbb{E}[\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}^*} | \mathcal{F}_t] \\ \mathbb{E}[\mathfrak{r}\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}^*} | \mathcal{F}_t] \end{pmatrix} - \tilde{I} M_2(t)$$

for all $t \in [0,T]$. Assume that $\boldsymbol{u}^* = (u_i^*)_{i\in[N]}$ satisfies $\|u_i^*\|_{\mathcal{H}^2(\mathbb{R})} \leq L$ for all $i \in [N]$, with $L > 0$ in (3.60). *Then $\boldsymbol{u}^*$ is an α_N-NE of $\mathcal{G}_{LQ}$, with α_N satisfying* (3.61). *Moreover, the* *process* $F_t := \begin{pmatrix} \mathbb{E}[\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}^*}|\mathcal{F}_t] \\ \mathbb{E}[\mathfrak{r}\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}^*}|\mathcal{F}_t] \end{pmatrix}$, $t \in [0,T]$, *satisfies the linear SDE*

$$\begin{aligned} \mathrm{d}F_t &= \left[\left(\begin{pmatrix} A(t) & \mathbb{0}_{2N} \\ \mathbb{0}_{2N} & A(t) \end{pmatrix} - \tilde{I}^\top K_{M_0,M_1}(t)\right) F_t - \tilde{I}^\top \tilde{I} M_2(t)\right] \mathrm{d}t + \begin{pmatrix} \Sigma(t) \\ \frac{1}{2}\Sigma(t) \end{pmatrix} \mathrm{d}W_t, \\ F_0 &= \begin{pmatrix} \mathbb{X}_0 \\ \frac{1}{2}\mathbb{X}_0 \end{pmatrix}. \end{aligned} \tag{3.70}$$

Remark 3.6.1. *Theorem 3.6.1 leverages the LQ structure of $\mathcal{G}_{LQ}$ to characterize the α_N-NE $\boldsymbol{u}^*$ as a feedback function of F, which involves finite conditional moments of $(\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}^*}, \mathfrak{r})$. These moments serve as sufficient statistics for the infinite-dimensional conditional law $\mathcal{L}(\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}^*}, \mathfrak{r}|\mathcal{F}_t)$. Notably, the process F is Markovian and satisfies the linear SDE* (3.70), *enabling the efficient computation of the α_N-NE. We remark that the solvability of* (3.66) *and* (3.69) *follows from linear ODE theory, and the solvability of* (3.67) *can be ensured at least for sufficiently small T.*

3.7 Proofs of Main Results

3.7.1 Proof of Theorem 3.2.1

The following lemmas regarding the linear derivative are given in [67, Lemmas 4.1 and 4.2], and will be used in the proof of Theorem 3.2.1.

Lemma 3.7.1. *Suppose $\mathcal{A}^{(N)}$ is convex, $i \in [N]$, and $f : \mathcal{A}^{(N)} \to \mathbb{R}$ has a linear derivative $\frac{\delta f}{\delta a_i}$ with respect to $\mathcal{A}_i$. Let $\boldsymbol{a} = (a_i, a_{-i}) \in \mathcal{A}^{(N)}$, $a'_i \in \mathcal{A}_i$, and for each $\varepsilon \in [0,1]$, let $\boldsymbol{a}^\varepsilon = (a_i + \varepsilon(a'_i - a_i), a_{-i})$. Then the function $[0,1] \ni \varepsilon \mapsto f(\boldsymbol{a}^\varepsilon) \in \mathbb{R}$ is differentiable and $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f(\boldsymbol{a}^\varepsilon) = \frac{\delta f}{\delta a_i}(\boldsymbol{a}^\varepsilon; a'_i - a_i)$ for all $\varepsilon \in [0,1]$.*

Lemma 3.7.2. *Suppose $\mathcal{A}^{(N)}$ is convex and for all $i \in [N]$, $f : \mathcal{A}^{(N)} \to \mathbb{R}$ has a linear derivative $\frac{\delta f}{\delta a_i}$ with respect to $\mathcal{A}_i$ such that for all $\boldsymbol{z}, \boldsymbol{a} \in \mathcal{A}^{(N)}$ and $a'_i \in \mathcal{A}_i$, $[0,1]^N \ni \varepsilon \mapsto \frac{\delta f}{\delta a_i}(\boldsymbol{z} + \varepsilon \cdot (\boldsymbol{a} - \boldsymbol{z}); a'_i)$ is continuous at 0, where $\boldsymbol{z} + \varepsilon \cdot (\boldsymbol{a} - \boldsymbol{z}) := (z_i + \varepsilon_i(a_i - z_i))_{i \in [N]}$. Then for all $\boldsymbol{z}, \boldsymbol{a} \in \mathcal{A}^{(N)}$, the map $[0,1] \ni r \mapsto f(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z})) \in \mathbb{R}$ is differentiable and $\frac{\mathrm{d}}{\mathrm{d}r} f(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z})) = \sum_{j=1}^N \frac{\delta f}{\delta a_j}(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_j - z_j)$.*

Proof of Theorem 3.2.1. By Condition 2 and Lemma 3.7.2, $[0,1] \ni r \mapsto \frac{\delta V_j}{\delta a_j}(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_j - z_j) \in \mathbb{R}$ is differentiable, and hence Φ in (3.11) is well-defined.

We now prove that Φ has a linear derivative with respect to $\mathcal{A}_i$ for all $i \in [N]$. To this end, let $i \in [N]$, $\boldsymbol{a} \in \mathcal{A}^{(N)}$ and $a'_i \in \mathcal{A}_i$. For all $\varepsilon \in (0,1]$, let $\boldsymbol{a}^\varepsilon := (a_i + \varepsilon\,(a'_i - a_i)\,, a_{-i})$.

By the definition of Φ in (3.11),

$$\begin{aligned}\Phi\left(\boldsymbol{a}^{\varepsilon}\right)-\Phi(\boldsymbol{a}) = & \int_0^1 \sum_{j=1}^N \frac{\delta V_j}{\delta a_j}\left(\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right); a_j+\varepsilon \delta_{ji}\left(a_i^{\prime}-a_i\right)-z_j\right) \mathrm{d} r \\ & -\int_0^1 \sum_{j=1}^N \frac{\delta V_j}{\delta a_j}\left(\boldsymbol{z}+r(\boldsymbol{a}-\boldsymbol{z}) ; a_j-z_j\right) \mathrm{d} r .\end{aligned}$$

Then by $\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right) \in \mathcal{A}^{(N)}$, for all $\varepsilon \in(0,1]$,

$$\begin{aligned}\frac{\Phi\left(\boldsymbol{a}^{\varepsilon}\right)-\Phi(\boldsymbol{a})}{\varepsilon} & =\frac{1}{\varepsilon} \int_0^1 \sum_{j=1}^N\left(\frac{\delta V_j}{\delta a_j}\left(\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right) ; a_j-z_j\right)-\frac{\delta V_j}{\delta a_j}\left(\boldsymbol{z}+r(\boldsymbol{a}-\boldsymbol{z}) ; a_j-z_j\right)\right) \mathrm{d} r \\ & +\frac{1}{\varepsilon} \int_0^1 \sum_{j=1}^N \varepsilon \delta_{ji} \frac{\delta V_j}{\delta a_j}\left(\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right) ; a_j^{\prime}-a_j\right) \mathrm{d} r \\ & =\int_0^1 \sum_{j=1}^N \frac{1}{\varepsilon}\left(\frac{\delta V_j}{\delta a_j}\left(\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right) ; a_j-z_j\right)-\frac{\delta V_j}{\delta a_j}\left(\boldsymbol{z}+r(\boldsymbol{a}-\boldsymbol{z}) ; a_j-z_j\right)\right) \mathrm{d} r \\ & +\int_0^1 \frac{\delta V_i}{\delta a_i}\left(\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right) ; a_i^{\prime}-a_i\right) \mathrm{d} r .\end{aligned} \tag{3.71}$$

To send $\varepsilon \rightarrow 0$ in the above equation, note that for all $\varepsilon \in[0,1], r \in[0,1],\left(\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right)\right)_{-i}=z_{-i}+r\left(a_{-i}-z_{-i}\right)$ and $\left(\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right)\right)_i=z_i+r\left(a_i+\varepsilon\left(a_i^{\prime}-a_i\right)-z_i\right)=z_i+r\left(a_i-z_i\right)+\varepsilon\left(\left(z_i+r\left(a_i^{\prime}-z_i\right)\right)-\left(z_i+r\left(a_i-z_i\right)\right)\right)$ with $z_i+r\left(a_i-z_i\right), z_i+r\left(a_i^{\prime}-z_i\right) \in \mathcal{A}_i$. Thus for all $j \in[N]$, the twice differentiability of V_j and Lemma 3.7.1 imply that $\varepsilon \mapsto \frac{\delta V_j}{\delta a_j}(\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right);$ $\left.a_j-z_j\right)$ is differentiable on $[0,1]$ and

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d} \varepsilon} \frac{\delta V_j}{\delta a_j}\left(\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right) ; a_j-z_j\right) & =\frac{\delta^2 V_j}{\delta a_j \delta a_i}\left(\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right) ; a_j-z_j, r\left(a_i^{\prime}-a_i\right)\right) \\ & =\frac{\delta^2 V_j}{\delta a_j \delta a_i}\left(\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right) ; a_j-z_j, a_i^{\prime}-a_i\right) r,\end{aligned}$$

where the last identity used the linearity of $\frac{\delta^2 V_j}{\delta a_j \delta a_i}$ in its last component. Hence, by the mean value theorem and Condition 1, for all $\varepsilon \in(0,1]$,

$$\begin{aligned}& \left|\frac{1}{\varepsilon}\left(\frac{\delta V_j}{\delta a_j}\left(\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right) ; a_j-z_j\right)-\frac{\delta V_j}{\delta a_j}\left(\boldsymbol{z}+r(\boldsymbol{a}-\boldsymbol{z}) ; a_j-z_j\right)\right)\right| \\ & \quad \leq \sup_{r, \varepsilon \in[0,1]}\left|\frac{\delta^2 V_j}{\delta a_j \delta a_i}\left(\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right) ; a_j-z_j, a_i^{\prime}-a_i\right) r\right|<\infty .\end{aligned}$$

Similarly, as $a_i^{\prime}-a_i \in \operatorname{span}\left(\mathcal{A}_i\right)$, by the twice differentiability of V_i, for all $r \in(0,1)$, $\lim_{\varepsilon \downarrow 0} \frac{\delta V_i}{\delta a_i}(\boldsymbol{z}+r\left(\boldsymbol{a}^{\varepsilon}-\boldsymbol{z}\right) ; a_i^{\prime}-a_i)=\frac{\delta V_i}{\delta a_i}(\boldsymbol{z}+r(\boldsymbol{a}-\boldsymbol{z}) ; a_i^{\prime}-a_i)$, and for all $r, \varepsilon \in[0,1]$, by the mean value theorem, there exists $\tilde{\varepsilon} \in [0,1]$ such that

$$\begin{aligned}\left|\frac{\delta V_i}{\delta a_i}\left(\boldsymbol{z} + r\left(\boldsymbol{a}^{\varepsilon} - \boldsymbol{z}\right); a_i' - a_i\right)\right| \leq & \left|\frac{\delta V_i}{\delta a_i}\left(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_i' - a_i\right)\right| \\ & + \left|\frac{\delta^2 V_i}{\delta a_i \delta a_i}\left(z + r\left(\boldsymbol{a}^{\tilde{\varepsilon}} - \boldsymbol{z}\right); a_i' - a_i, a_i' - a_i\right) r\right|.\end{aligned} \tag{3.72}$$

Using Lemma 3.7.2, for all $a_i' \in \mathcal{A}_i$,

$$\frac{\mathrm{d}}{\mathrm{d}r}\frac{\delta V_i}{\delta a_i}\left(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_i'\right) = \sum_{j=1}^{N}\frac{\delta^2 V_i}{\delta a_i \delta a_j}\left(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_i', a_j - z_j\right), \tag{3.73}$$

which along with (3.72) and Condition 1 implies that

$$\sup_{(r,\varepsilon)\in[0,1]^2}\left|\frac{\delta V_i}{\delta a_i}\left(\boldsymbol{z} + r\left(\boldsymbol{a}^{\varepsilon} - \boldsymbol{z}\right); a_i' - a_i\right)\right| < \infty.$$

Hence, letting $\varepsilon \to 0$ in (3.71) and using Lebesgue's dominated convergence theorem give

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Phi\left(\boldsymbol{a}^{\varepsilon}\right)\Big|_{\varepsilon=0} = & \int_0^1 \sum_{j=1}^{N}\frac{\delta^2 V_j}{\delta a_j \delta a_i}\left(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_j - z_j, a_i' - a_i\right) r\mathrm{d}r \\ & + \int_0^1 \frac{\delta V_i}{\delta a_i}\left(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_i' - a_i\right)\mathrm{d}r.\end{aligned}$$

Let $\mathcal{E} : [0,1] \to \mathbb{R}$ be given by

$$\mathcal{E}_r := \sum_{j=1}^{N}\left(\frac{\delta^2 V_j}{\delta a_j \delta a_i}\left(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_j - z_j, a_i' - a_i\right) - \frac{\delta^2 V_i}{\delta a_i \delta a_j}\left(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_i' - a_i, a_j - z_j\right)\right).$$

Then by (3.73),

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Phi\left(\boldsymbol{a}^{\varepsilon}\right)\Big|_{\varepsilon=0} &= \int_0^1\left(\sum_{j=1}^{N}\frac{\delta^2 V_i}{\delta a_i \delta a_j}\left(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_i' - a_i, a_j - z_j\right) + \mathcal{E}_r\right) r\mathrm{d}r \\ &\quad + \int_0^1 \frac{\delta V_i}{\delta a_i}\left(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_i' - a_i\right)\mathrm{d}r \\ &= \int_0^1 r\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{\delta V_i}{\delta a_i}\left(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_i' - a_i\right)\right)\mathrm{d}r \\ &\quad + \int_0^1 \frac{\delta V_i}{\delta a_i}\left(\boldsymbol{z} + r(\boldsymbol{a} - \boldsymbol{z}); a_i' - a_i\right)\mathrm{d}r + \int_0^1 \mathcal{E}_r r\mathrm{d}r \\ &= \frac{\delta V_i}{\delta a_i}\left(\boldsymbol{a}; a_i' - a_i\right) + \int_0^1 \mathcal{E}_r r\mathrm{d}r,\end{aligned} \tag{3.74}$$

where the last line uses the integration by part formula. This proves the linear differentiability of Φ.

Now we prove Φ is an α-potential function of $\mathcal{G}$. Let $i \in [N], a_i' \in \mathcal{A}_i$ and $\boldsymbol{a} \in \mathcal{A}^{(N)}$. For each $\varepsilon \in [0,1]$, let $\boldsymbol{a}^\varepsilon = (a_i + \varepsilon(a_i' - a_i), a_{-i}) \in \mathcal{A}^{(N)}$. By the differentiability of V_i and Lemma 3.7.1, $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} V_i(\boldsymbol{a}^\varepsilon) = \frac{\delta V_i}{\delta a_i}(\boldsymbol{a}^\varepsilon; a_i' - a_i)$ for all $\varepsilon \in [0,1]$, and $\varepsilon \mapsto \frac{\delta V_i}{\delta a_i}(\boldsymbol{a}^\varepsilon; a_i' - a_i)$ is differentiable on $[0,1]$. This implies that $\varepsilon \mapsto V_i(\boldsymbol{a}^\varepsilon)$ is continuously differentiable on $[0,1]$. Similarly, by Lemma 3.7.1 and (3.74) and the continuity assumption, $[0,1] \ni \varepsilon \mapsto \Phi(\boldsymbol{a}^\varepsilon) \in \mathbb{R}$ is also continuously differentiable with $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Phi(\boldsymbol{a}^\varepsilon) = \frac{\delta V_i}{\delta a_i}(\boldsymbol{a}^\varepsilon; a_i' - a_i) + \int_0^1 \mathcal{E}_{r,\varepsilon} r \mathrm{d}r$, where $\mathcal{E}_{r,\varepsilon}$ is given by

$$\begin{aligned}\mathcal{E}_{r,\varepsilon} = \sum_{j=1}^N \Bigg(&\frac{\delta^2 V_j}{\delta a_j \delta a_i}(\boldsymbol{z} + r(\boldsymbol{a}^\varepsilon - \boldsymbol{z}); a_j - z_j, a_i' - a_i) \\ &- \frac{\delta^2 V_i}{\delta a_i \delta a_j}(\boldsymbol{z} + r(\boldsymbol{a}^\varepsilon - \boldsymbol{z}); a_i' - a_i, a_j - z_j)\Bigg).\end{aligned} \tag{3.75}$$

Hence by the fundamental theorem of calculus,

$$\begin{aligned} V_i((a_i', a_{-i})) - V_i((a_i, a_{-i})) &= \int_0^1 \frac{\delta V_i}{\delta a_i}(\boldsymbol{a}^\varepsilon; a_i' - a_i)\mathrm{d}\varepsilon \\ &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Phi(\boldsymbol{a}^\varepsilon)\mathrm{d}\varepsilon - \int_0^1\int_0^1 \mathcal{E}_{r,\varepsilon} r \mathrm{d}r \mathrm{d}\varepsilon \\ &= \Phi((a_i', a_{-i})) - \Phi((a_i, a_{-i})) - \int_0^1\int_0^1 \mathcal{E}_{r,\varepsilon} r \mathrm{d}r \mathrm{d}\varepsilon. \end{aligned} \tag{3.76}$$

Finally, the desired upper bound of α follows from the fact that

$$\left|\int_0^1\int_0^1 \mathcal{E}_{r,\varepsilon} r \mathrm{d}r \mathrm{d}\varepsilon\right| \le 2 \sup_{i\in[N], a_i'\in\mathcal{A}_i, \boldsymbol{a}, \boldsymbol{a}''\in\mathcal{A}^{(N)}} \sum_{j=1}^N \left| \frac{\delta^2 V_i}{\delta a_i \delta a_j}(\boldsymbol{a}; a_i', a_j'') - \frac{\delta^2 V_j}{\delta a_j \delta a_i}(\boldsymbol{a}; a_j'', a_i')\right|. \tag{3.77}$$

due to the bilinearity of $\frac{\delta^2 V_j}{\delta a_j \delta a_i}$ and $\frac{\delta^2 V_i}{\delta a_i \delta a_j}$, and the fact that $\int_0^1\int_0^1 r \mathrm{d}r \mathrm{d}\epsilon = \frac{1}{2}$. This finishes the proof.

□

Proof of Proposition 3.2.2. As we assume that the second-order Fréchet derivative of V_i for any $i \in [N]$ exists, one can define

$$\frac{\delta^2 V_i}{\delta a_i \delta a_j}(\boldsymbol{z}; \tilde{a}_i', \tilde{a}_j'') = \langle \tilde{a}_i', \partial^2_{a_i a_j} V_i(\boldsymbol{z})\tilde{a}_j''\rangle. \tag{3.78}$$

Following the same notation as in the proof of Theorem 3.2.1, by (3.75) and (3.78), as well as V_j is twice continuously differentiable for any $j \in [N]$,

$$
\begin{aligned}
&|\mathcal{E}_{r,\varepsilon}| \\
&= \left| \left\langle a_i' - a_i, \sum_{j=1}^{N} \left(\partial^2_{a_i a_j} V_j \left(\boldsymbol{z} + r(\boldsymbol{a}^\varepsilon - \boldsymbol{z}) \right) (a_j - z_j) - \partial^2_{a_i a_j} V_i \left(\boldsymbol{z} + r(\boldsymbol{a}^\varepsilon - \boldsymbol{z}) \right) (a_j - z_j) \right) \right\rangle \right| \\
&\le \|a_i' - a_i\| \cdot 2 \sup_{j \in [N], a_j \in \mathcal{A}_j} \|a_j\| \cdot \sum_{j=1}^{N} \left| \partial^2_{a_j a_i} V_j \left(\boldsymbol{z} + r(\boldsymbol{a}^\varepsilon - \boldsymbol{z}) \right) - \partial^2_{a_i a_j} V_i \left(\boldsymbol{z} + r(\boldsymbol{a}^\varepsilon - \boldsymbol{z}) \right) \right|
\end{aligned}
$$

Then combining (3.76) and (3.77) implies that

$$
\begin{aligned}
&\left| V_i \left((a_i', a_{-i}) \right) - V_i \left((a_i, a_{-i}) \right) - \left(\Phi \left((a_i', a_{-i}) \right) - \Phi \left((a_i, a_{-i}) \right) \right) \right| \\
&\le \left| \int_0^1 \int_0^1 \mathcal{E}_{r,\varepsilon} r \mathrm{d}r \mathrm{d}\varepsilon \right| \le \|a_i' - a_i\| \cdot \sup_{j \in [N], a_j \in \mathcal{A}_j} \|a_j\| \cdot \sup_{\boldsymbol{a}} \sum_{j=1}^{N} \left\| \partial^2_{a_i a_j} V_i (\boldsymbol{a}) - \partial^2_{a_j a_i} V_j (\boldsymbol{a}) \right\|.
\end{aligned}
$$

3.7.2 Proof of Theorem 3.3.2

The following propositions estimate the moments of the state process $\mathbf{X}^{\boldsymbol{u}}$ and the sensitivity processes $\mathbf{Y}^{\boldsymbol{u},u_h'}$ and $\mathbf{Z}^{\boldsymbol{u},u_h',u_\ell''}$. The proofs of these propositions are included in Section 3.8.1.

Proposition 3.7.3. *Suppose Assumption 3.3.2 holds. For each $\boldsymbol{u} \in \mathcal{H}^p(\mathbb{R}^N)$, the solution $\boldsymbol{X}^{\boldsymbol{u}} \in \mathcal{H}^p(\mathbb{R}^N)$ to* (3.20) *satisfies for all $i \in [N]$, $\sup_{t\in[0,T]} \mathbb{E}[|X^{\boldsymbol{u}}_{t,i}|^p] \le C_X^{i,p}$, with the constant $C_X^{i,p}$ defined by $C_X^{i,p} := \Big(|x_i|^p + (p-1)\|\sigma_i\|^p_{L^p} + L^b T + \|u_i\|^p_{\mathcal{H}^p(\mathbb{R})} + \frac{L^b_y}{N} \sum_{k=1}^N \big(|x_k|^p + (p-1)\|\sigma_k\|^p_{L^p} + L^b T + \|u_k\|^p_{\mathcal{H}^p(\mathbb{R})} \big) \Big) e^{c_p (L^b + L^b_y + 1)T}$, and $c_p \ge 1$ is a constant depending only on p.*

Proposition 3.7.4. *Suppose Assumption 3.3.2 holds and let $p \ge 2$. For all $\boldsymbol{u} \in \mathcal{H}^p(\mathbb{R}^N)$, $h \in [N]$ and $u_h' \in \mathcal{H}^p(\mathbb{R})$, the solution $\boldsymbol{Y}^{\boldsymbol{u},u_h'} \in \mathcal{H}^p(\mathbb{R}^N)$ of* (3.22) *satisfies for all $i \in [N]$,*

$$
\sup_{t\in[0,T]} \mathbb{E}[|Y^{\boldsymbol{u},u_h'}_{t,i}|^p] \le \left(\delta_{h,i} C_Y^{h,p} + \frac{(L^b_y)^p}{N^p} \bar{C}_Y^{h,p} \right) \|u_h'\|^p_{\mathcal{H}^p(\mathbb{R})},
$$

where $C_Y^{h,p} := (2T)^{p-1} e^{pL^b T}$ and $\bar{C}_Y^{h,p} := (2T)^{2p-1} e^{p(L^b + L^b_y)T} e^{pL^b T}$.

Proposition 3.7.5. *Suppose Assumption 3.3.2 holds. For all $\boldsymbol{u} \in \mathcal{H}^p(\mathbb{R}^N)$, $h, \ell \in [N]$ with $h \ne \ell$, and all $u_h', u_\ell'' \in \mathcal{H}^2(\mathbb{R})$, the solution $\boldsymbol{Z}^{\boldsymbol{u},u_h',u_\ell''} \in \mathcal{H}^p(\mathbb{R}^N)$ of* (3.23) *satisfies for all $i \in [N]$,*

$$
\sup_{t\in[0,T]} \mathbb{E}\left[|Z^{\boldsymbol{u},u_h',u_\ell''}_{t,i}|^2 \right] \le C (L^b_y)^2 \left((\delta_{h,i} + \delta_{\ell,i}) \frac{1}{N^2} + \frac{1}{N^4} \right) \|u_h'\|^2_{\mathcal{H}^4(\mathbb{R})} \|u_\ell''\|^2_{\mathcal{H}^4(\mathbb{R})},
$$

where $C \ge 0$ is a constant depending only on the upper bounds of T, L^b, L^b_y.

We now prove Theorem 3.3.2 based on Propositions 3.7.3, 3.7.4 and 3.7.5.

Proof of Theorem 3.3.2. To simplify the notation, we omit the dependence on $\boldsymbol{u}$ in the superscript of all processes, i.e., $\mathbf{X} = \mathbf{X}^{\boldsymbol{u}}, \mathbf{Y}^i = \mathbf{Y}^{\boldsymbol{u},u_i'}$. We denote by $C \geq 0$ a generic constant depending only on the upper bounds of T, $\max_{i\in[N]} |x_i|^2$, $\max_{i\in[N]} \|\sigma_i\|_{L^2}$, L^b, L^b_y, $\max_{k\in[N]} \|u_k\|_{\mathcal{H}^2(\mathbb{R})}$.

By the definition of $\frac{\delta^2 V_j}{\delta u_j \delta u_i}(\boldsymbol{u}; u_j'', u_i')$ in (3.25) and the fact that $\mathbf{Z}^{\boldsymbol{u},u_i',u_j''} = \mathbf{Z}^{\boldsymbol{u},u_j'',u_i'}$,

$$
\begin{aligned}
&\left| \frac{\delta^2 V_i}{\delta u_i \delta u_j}(\boldsymbol{u}; u_i', u_j'') - \frac{\delta^2 V_j}{\delta u_j \delta u_i}(\boldsymbol{u}; u_j'', u_i') \right| \\
&= \mathbb{E}\left[\int_0^T \left\{ \begin{pmatrix} \mathbf{Y}_t^i \\ u_{t,i}' \end{pmatrix}^\top \begin{pmatrix} \partial^2_{xx}\Delta^f_{i,j} & \partial^2_{xu_j}\Delta^f_{i,j} \\ \partial^2_{u_i x}\Delta^f_{i,j} & \partial^2_{u_i u_j}\Delta^f_{i,j} \end{pmatrix} (t,\cdot) \begin{pmatrix} \mathbf{Y}_t^j \\ u_{t,j}'' \end{pmatrix} + \left(\mathbf{Z}_t^{i,j}\right)^\top \left(\partial_x \Delta^f_{i,j}\right)(t,\cdot) \right\} \mathrm{d}t \right] \\
&\quad + \mathbb{E}\left[(\mathbf{Y}_T^i)^\top (\partial^2_{xx}\Delta^g_{i,j})(\mathbf{X}_T)\mathbf{Y}_T^j + (\mathbf{Z}_T^{i,j})^\top (\partial_x \Delta^g_{i,j})(\mathbf{X}_T) \right],
\end{aligned}
\tag{3.79}
$$

where we write for simplicity $\partial^2_{xx}\Delta^f_{i,j}(t,\cdot) = \partial^2_{xx}(f_i - f_j)(t, \mathbf{X}_t, \boldsymbol{u}_t)$ and similarly for other derivatives. In the sequel, we derive upper bounds for all terms on the right-hand side of (3.79).

To estimate the term involving the Hessian of $\Delta^f_{i,j}$ in (3.79), observe that for all $t \in [0,T]$,

$$
\begin{aligned}
&\begin{pmatrix} \mathbf{Y}_t^i \\ u_{t,i}' \end{pmatrix}^\top \begin{pmatrix} \partial^2_{xx}\Delta^f_{i,j} & \partial^2_{xu}\Delta^f_{i,j} \\ \partial^2_{u_i x}\Delta^f_{i,j} & \partial^2_{u_i u_j}\Delta^f_{i,j} \end{pmatrix} (t,\cdot) \begin{pmatrix} \mathbf{Y}_t^j \\ u_{t,j}'' \end{pmatrix} \\
&= \sum_{h,\ell=1}^N (\partial^2_{x_h x_\ell}\Delta^f_{i,j})(t,\cdot) Y^i_{t,h} Y^j_{t,\ell} + u_{t,j}'' \sum_{h}^N (\partial^2_{x_h u_j}\Delta^f_{i,j})(t,\cdot) Y^i_{t,h} \\
&\quad + u_{t,i}' \sum_{\ell=1}^N (\partial^2_{u_i x_\ell}\Delta^f_{i,j})(t,\cdot) Y^j_{t,\ell} + (\partial^2_{u_i u_j}\Delta^f_{i,j})(t,\cdot) u_{t,i}' u_{t,j}''.
\end{aligned}
\tag{3.80}
$$

The first term on the right-hand side of (3.80) satisfies the identity:

$$
\begin{aligned}
\sum_{h,\ell=1}^N (\partial^2_{x_h x_\ell}\Delta^f_{i,j})(t,\cdot) Y^i_{t,h} Y^j_{t,\ell} &= (\partial^2_{x_i x_j}\Delta^f_{i,j})(t,\cdot) Y^i_{t,i} Y^j_{t,j} + \sum_{\ell\in[N]\setminus\{j\}} (\partial^2_{x_i x_\ell}\Delta^f_{i,j})(t,\cdot) Y^i_{t,i} Y^j_{t,\ell} \\
&\quad + \sum_{h\in[N]\setminus\{i\}} (\partial^2_{x_h x_j}\Delta^f_{i,j})(t,\cdot) Y^i_{t,h} Y^j_{t,j} + \sum_{h\in[N]\setminus\{i\},\ell\in[N]\setminus\{j\}} (\partial^2_{x_h x_\ell}\Delta^f_{i,j})(t,\cdot) Y^i_{t,h} Y^j_{t,\ell},
\end{aligned}
$$

which yields the following estimate:

$$\begin{aligned}
&\left|\mathbb{E}\left[\int_0^T \sum_{h,\ell=1}^N (\partial^2_{x_h x_\ell}\Delta^f_{i,j})(t,\cdot)Y^i_{t,h}Y^j_{t,\ell}\mathrm{d}t\right]\right| && (3.81)\\
&\leq \|\partial^2_{x_i x_j}\Delta^f_{i,j}\|_{L^\infty}\|Y^i_i Y^j_j\|_{\mathcal{H}^1(\mathbb{R})} + \sum_{\ell\in[N]\backslash\{j\}}\|\partial^2_{x_i x_\ell}\Delta^f_{i,j}\|_{L^\infty}\|Y^i_i Y^j_\ell\|_{\mathcal{H}^1(\mathbb{R})}\\
&\quad + \sum_{h\in[N]\backslash\{i\}}\|\partial^2_{x_h x_j}\Delta^f_{i,j}\|_{L^\infty}\|Y^i_h Y^j_j\|_{\mathcal{H}^1(\mathbb{R})} + \sum_{h\in[N]\backslash\{i\},\ell\in[N]\backslash\{j\}}\|\partial^2_{x_h x_\ell}\Delta^f_{i,j}\|_{L^\infty}\|Y^i_h Y^j_\ell\|_{\mathcal{H}^1(\mathbb{R})}\\
&\leq C\|u'_i\|_{\mathcal{H}^2(\mathbb{R})}\|u''_j\|_{\mathcal{H}^2(\mathbb{R})}\Bigg\{\|\partial^2_{x_i x_j}\Delta^f_{i,j}\|_{L^\infty} + \frac{L^b_y}{N}\Bigg(\sum_{\ell\in[N]\backslash\{j\}}\|\partial^2_{x_i x_\ell}\Delta^f_{i,j}\|_{L^\infty}\\
&\quad + \sum_{h\in[N]\backslash\{i\}}\|\partial^2_{x_h x_j}\Delta^f_{i,j}\|_{L^\infty}\Bigg) + \frac{(L^b_y)^2}{N^2}\Bigg(\sum_{h\in[N]\backslash\{i\}}\sum_{\ell\in[N]\backslash\{j\}}\|\partial^2_{x_h x_\ell}\Delta^f_{i,j}\|_{L^\infty}\Bigg)\Bigg\}. && (3.82)
\end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality and Proposition 3.7.4. Similarly, using Propositions 3.7.4, the second and third terms in (3.80) can be bounded by

$$\begin{aligned}
&\left|\mathbb{E}\left[\int_0^T u''_{t,j}\sum_{h=1}^N(\partial^2_{x_h u_j}\Delta^f_{i,j})(t,\cdot)Y^i_{t,h}\mathrm{d}t\right]\right| + \left|\mathbb{E}\left[\int_0^T u'_{t,i}\sum_{\ell=1}^N(\partial^2_{u_i x_\ell}\Delta^f_{i,j})(t,\cdot)Y^j_{t,\ell}\mathrm{d}t\right]\right|\\
\leq& C\|u'_i\|_{\mathcal{H}^2(\mathbb{R})}\|u''_j\|_{\mathcal{H}^2(\mathbb{R})}\Bigg\{\|\partial^2_{x_i u_j}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_i x_j}\Delta^f_{i,j}\|_{L^\infty} && (3.83)\\
&\quad + \frac{L^b_y}{N}\big(\sum_{h\in[N]\backslash\{i\}}\|\partial^2_{x_h u_j}\Delta^f_{i,j}\|_{L^\infty} + \sum_{\ell\in[N]\backslash\{j\}}\|\partial_{u_i x_\ell}\Delta^f_{i,j}\|_{L^\infty}\big)\Bigg\}
\end{aligned}$$

and the fourth term in (3.80) can be bounded by

$$\left|\mathbb{E}\left[\int_0^T(\partial^2_{u_i u_j}\Delta^f_{i,j})(t,\cdot)u'_{t,i}u''_{t,j}\mathrm{d}t\right]\right| \leq \|u'_i\|_{\mathcal{H}^2(\mathbb{R})}\|u''_j\|_{\mathcal{H}^2(\mathbb{R})}\|\partial^2_{u_i u_j}\Delta^f_{i,j}\|_{L^\infty}. \qquad (3.84)$$

Combining (3.82), (3.83), and (3.84) yield the following bound of (3.80):

$$\begin{aligned}
&\left|\mathbb{E}\left[\int_0^T \begin{pmatrix}\mathbf{Y}_t^i\\ u'_{t,i}\end{pmatrix}^\top \begin{pmatrix}\partial^2_{xx}\Delta^f_{i,j} & \partial^2_{xu_j}\Delta^f_{i,j}\\ \partial^2_{u_ix}\Delta^f_{i,j} & \partial^2_{u_iu_j}\Delta^f_{i,j}\end{pmatrix}(t,\cdot)\begin{pmatrix}\mathbf{Y}_t^j\\ u''_{t,j}\end{pmatrix}\mathrm{d}t\right]\right| \\
&\leq C\|u'_i\|_{\mathcal{H}^2(\mathbb{R})}\|u''_j\|_{\mathcal{H}^2(\mathbb{R})}\Bigg\{\|\partial^2_{x_ix_j}\Delta^f_{i,j}\|_{L^\infty}+\|\partial^2_{x_iu_j}\Delta^f_{i,j}\|_{L^\infty}+\|\partial^2_{u_ix_j}\Delta^f_{i,j}\|_{L^\infty}+\|\partial^2_{u_iu_j}\Delta^f_{i,j}\|_{L^\infty} \\
&\quad+\frac{L^b_y}{N}\Bigg(\sum_{\ell\in[N]\setminus\{j\}}\big(\|\partial^2_{x_ix_\ell}\Delta^f_{i,j}\|_{L^\infty}+\|\partial^2_{u_ix_\ell}\Delta^f_{i,j}\|_{L^\infty}\big)+\sum_{h\in[N]\setminus\{i\}}\big(\|\partial^2_{x_hx_j}\Delta^f_{i,j}\|_{L^\infty} \\
&\quad+\|\partial^2_{x_hu_j}\Delta^f_{i,j}\|_{L^\infty}\big)\Bigg)+\frac{(L^b_y)^2}{N^2}\Bigg(\sum_{h\in[N]\setminus\{i\},\ell\in[N]\setminus\{j\}}\|\partial^2_{x_hx_\ell}\Delta^f_{i,j}\|_{L^\infty}\Bigg)\Bigg\}.
\end{aligned} \tag{3.85}$$

To estimate the term involving the gradient of $\Delta^f_{i,j}$ in (3.79), observe that for all $t\in[0,T]$, $\left(\mathbf{Z}^{i,j}_t\right)^\top\left(\partial_x\Delta^f_{i,j}\right)(t,\cdot)=\sum_{h=1}^N(\partial_{x_h}\Delta^f_{i,j})(t,\cdot)Z^{i,j}_{t,h}$. The fundamental theorem of calculus implies that for all $(t,x,u)=(t,(x_\ell)_{\ell=1}^N,(u_\ell)_{\ell=1}^N)\in[0,T]\in\mathbb{R}^N\times\mathbb{R}^N$ and $h\in[N]$, $|(\partial_{x_h}\Delta^f_{i,j})(t,x,u)|\leq|(\partial_{x_h}\Delta^f_{i,j})(t,0,0)|+\sum_{\ell=1}^N(\|\partial^2_{x_hx_\ell}\Delta^f_{i,j}\|_{L^\infty}|x_\ell|+\|\partial^2_{x_hu_\ell}\Delta^f_{i,j}\|_{L^\infty}|u_\ell|)$, which implies that

$$\begin{aligned}
&\left|\mathbb{E}\left[\int_0^T\sum_{h=1}^N(\partial_{x_h}\Delta^f_{i,j})(t,\cdot)Z^{i,j}_{t,h}\mathrm{d}t\right]\right|\leq\sum_{h\in\{i,j\}}\Big(\|(\partial_{x_h}\Delta^f_{i,j})(\cdot,0,0)\|_{L^2}\|Z^{i,j}_h\|_{\mathcal{H}^2(\mathbb{R})} \\
&\quad+\sum_{\ell=1}^N(\|\partial^2_{x_hx_\ell}\Delta^f_{i,j}\|_{L^\infty}\|X_\ell Z^{i,j}_h\|_{\mathcal{H}^1(\mathbb{R})}+\|\partial^2_{x_hu_\ell}\Delta^f_{i,j}\|_{L^\infty}\|u_\ell Z^{i,j}_h\|_{\mathcal{H}^1(\mathbb{R})})\Big) \\
&\quad+\sum_{h\in[N]\setminus\{i,j\}}\Big(\|(\partial_{x_h}\Delta^f_{i,j})(\cdot,0,0)\|_{L^2}\|Z^{i,j}_h\|_{\mathcal{H}^2(\mathbb{R})} \\
&\qquad+\sum_{\ell=1}^N(\|\partial^2_{x_hx_\ell}\Delta^f_{i,j}\|_{L^\infty}\|X_\ell Z^{i,j}_h\|_{\mathcal{H}^1(\mathbb{R})}+\|\partial^2_{x_hu_\ell}\Delta^f_{i,j}\|_{L^\infty}\|u_\ell Z^{i,j}_h\|_{\mathcal{H}^1(\mathbb{R})})\Big).
\end{aligned}$$

Then by the Cauchy-Schwarz inequality and Propositions 3.7.3 and 3.7.5,

$$\begin{aligned}
&\left|\mathbb{E}\left[\int_0^T\sum_{h=1}^N(\partial_{x_h}\Delta^f_{i,j})(t,\cdot)Z^{i,j}_{t,h}\mathrm{d}t\right]\right|\leq C\|u'_i\|_{\mathcal{H}^4(\mathbb{R})}\|u''_j\|_{\mathcal{H}^4(\mathbb{R})}L^b_y\Bigg\{\frac{1}{N}\sum_{h\in\{i,j\}}\Bigg(\|(\partial_{x_h}\Delta^f_{i,j})(\cdot,0,0)\|_{L^2} \\
&\quad+\sum_{\ell=1}^N\Big(\|\partial^2_{x_hx_\ell}\Delta^f_{i,j}\|_{L^\infty}+\|\partial^2_{x_hu_\ell}\Delta^f_{i,j}\|_{L^\infty}\Big)\Bigg) \\
&\quad+\frac{1}{N^2}\sum_{h\in[N]\setminus\{i,j\}}\Bigg(\|(\partial_{x_h}\Delta^f_{i,j})(\cdot,0,0)\|_{L^2}+\sum_{\ell=1}^N\Big(\|\partial^2_{x_hx_\ell}\Delta^f_{i,j}\|_{L^\infty}+\|\partial^2_{x_hu_\ell}\Delta^f_{i,j}\|_{L^\infty}\Big)\Bigg)\Bigg\}.
\end{aligned} \tag{3.86}$$

Finally, using similar arguments as those for (3.82) and (3.86) allows for estimating the terms involving $\Delta^g_{i,j}$ in (3.79):

$$\begin{aligned}
&\left|\mathbb{E}\left[(\mathbf{Y}^i_T)^\top(\partial^2_{xx}\Delta^g_{i,j})(\mathbf{X}_T)\mathbf{Y}^j_T+(\mathbf{Z}^{i,j}_T)^\top(\partial_x\Delta^g_{i,j})(\mathbf{X}_T)\right]\right| \\
&\leq C\|u'_i\|_{\mathcal{H}^4(\mathbb{R})}\|u''_j\|_{\mathcal{H}^4(\mathbb{R})}\Bigg\{\|\partial^2_{x_ix_j}\Delta^g_{i,j}\|_{L^\infty}+\frac{L^b_y}{N}\Bigg(\sum_{h\in\{i,j\}}|(\partial_{x_h}\Delta^g_{i,j})(0)| \\
&\quad+\sum_{h\in\{i,j\},\ell\in[N]}\|\partial^2_{x_hx_\ell}\Delta^g_{i,j}\|_{L^\infty}\Bigg)+\frac{L^b_y}{N^2}\Bigg(\sum_{h\in[N]\setminus\{i,j\}}|(\partial_{x_h}\Delta^g_{i,j})(0)| \qquad (3.87)\\
&\quad+\sum_{\substack{h\in[N]\setminus\{i,j\},\\ \ell\in[N]}}\|\partial^2_{x_hx_\ell}\Delta^g_{i,j}\|_{L^\infty}+\sum_{h\in[N]\setminus\{i\},\ell\in[N]\setminus\{j\}}\|\partial^2_{x_hx_\ell}\Delta^g_{i,j}\|_{L^\infty}\Bigg)\Bigg\}.
\end{aligned}$$

Note that the last two terms in the last line can be replaced by $\sum_{\substack{h\in[N]\setminus\{i,j\},\\ \ell\in[N]\setminus\{i,j\}}}\|\partial^2_{x_hx_\ell}\Delta^g_{i,j}\|_{L^\infty}$, as the remaining ones can be absorbed in the terms with $1/N$. Consequently, using (3.79) and grouping the terms in the estimates (3.85), (3.86) and (3.87) according to the orders $1/N$ and $1/N^2$ yield

$$\begin{aligned}
&\left|\frac{\delta^2 V_i}{\delta u_i\delta u_j}\left(\boldsymbol{u};u'_i,u''_j\right)-\frac{\delta^2 V_j}{\delta u_j\delta u_i}\left(\boldsymbol{u};u''_j,u'_i\right)\right| \\
&\leq C\|u'_i\|_{\mathcal{H}^4(\mathbb{R})}\|u''_j\|_{\mathcal{H}^4(\mathbb{R})}\left(C^{i,j}_{V,1}+L^b_y\left(\frac{1}{N}C^{i,j}_{V,2}+\frac{1}{N^2}C^{i,j}_{V,3}\right)\right),
\end{aligned}$$

where $C^{i,j}_{V,1}, C^{i,j}_{V,2}$ and $C^{i,j}_{V,3}$ are given (3.26), (3.27), and (3.28) respectively. This finishes the proof. □

3.7.3 Proof of Theorem 3.4.2

Before proving Theorem 3.4.2, we present several propositions regarding moment estimates of the state and control processes and their sensitivity processes. The first proposition estimates the moments of the state process (3.35). The proof follows the exact same line as that for Proposition 3.7.3 and is therefore omitted.

Proposition 3.7.6. *Suppose Assumption 3.3.2 holds and that there exists $p\geq 2$ such that $\xi_i\in L^p(\Omega;\mathbb{R})$ for all $i\in I_N$. For all $\phi\in\Pi^N$, the solution $\boldsymbol{X}^\phi$ to (3.35) satisfies for all $i\in I_N$, $\sup_{t\in[0,T]}\mathbb{E}[|X^\phi_{t,i}|^p]\leq C^{i,p}_X$, where $C^{i,p}_X:=\Big(\mathbb{E}[|\xi_i|^p]+(p-1)\|\sigma_i\|^p_{L^p}+L^{b+\phi}T+\frac{L^{b+\phi}_y}{N}\sum_{i=1}^N(\mathbb{E}[|\xi_i|^p]+(p-1)\|\sigma_i\|^p_{L^p}+L^{b+\phi}T)\Big)e^{c_p(L^{b+\phi}+L^{b+\phi}_y+1)T}$, and $c_p\geq 1$ is a constant depending only on p.*

The following propositions estimate the moments of the sensitivity processes of the state and control variables. The proofs of these propositions are included in Section 3.8.2.

Proposition 3.7.7. *Suppose Assumption 3.3.2 holds and that there exists $p \geq 2$ such that $\xi_i \in L^p(\Omega; \mathbb{R})$ for all $i \in I_N$. For all $\phi \in \Pi^N$, $h \in I_N$ and $\phi'_h \in \Pi$, the solution $\boldsymbol{Y}^{\phi,\phi'_h}$ to* (3.37) *satisfies for all $i \in I_N$,*

$$\sup_{t\in[0,T]} \mathbb{E}[|Y_{t,i}^{\phi,\phi'_h}|^p] \leq \delta_{h,i} C_Y^{h,p} + \frac{1}{N^p} \bar{C}_Y^{h,p},$$

where the constants $C_Y^{h,p}$ and $\bar{C}_Y^{h,p}$ are defined by

$$C_Y^{h,p} := c_p\big((L^{\phi'_h})^p(1 + C_X^{h,p}) + (L_y^{\phi'_h})^p \frac{1}{N}\sum_{k=1}^{N} C_X^{k,p}\big) T^{2(p-1)} e^{pL^{b+\phi}T},$$

and $\bar{C}_Y^{h,p} := c_p(L_y^{b+\phi})^p \left((L^{\phi'_h})^p(1 + C_X^{h,p}) + (L_y^{\phi'_h})^p \frac{1}{N}\sum_{k=1}^{N} C_X^{k,p}\right) T^{3p-2} e^{c_p(L^{b+\phi}+L_y^{b+\phi})T}$, *with $(C_X^{k,p})_{k\in I_N}$ defined in Proposition 3.7.6, and a constant $c_p \geq 1$ depending only on p.*

Proposition 3.7.8. *Suppose Assumption 3.3.2 holds and that $\xi_i \in L^4(\Omega; \mathbb{R})$ for all $i \in I_N$. For all $\phi \in \Pi^N$, $h, \ell \in I_N$ with $h \neq \ell$, and $\phi'_h, \phi''_\ell \in \Pi$, the solution $\boldsymbol{Z}^{\phi,\phi'_h,\phi''_\ell}$ to* (3.38) *satisfies for all $i \in I_N$,*

$$\sup_{t\in[0,T]} \mathbb{E}\left[|Z_{t,i}^{\phi,\phi'_h,\phi''_\ell}|^2\right] \leq C\left((\delta_{h,i} + \delta_{\ell,i})\frac{1}{N^2} + \frac{1}{N^4}\right) \max\{L_y^{b+\phi}, L_y^{\phi'_h}, L_y^{\phi''_\ell}\}^2,$$

where $C \geq 0$ is a constant depending only on the upper bounds of T, $\max_{i\in I_N} \mathbb{E}[|\xi_i|^4]$, $\max_{i\in I_N} \|\sigma_i\|_{L^4}$, $L^{b+\phi}$, $L^{\phi'_h}$, $L^{\phi''_\ell}$, $L_y^{b+\phi}$, $L_y^{\phi'_h}$ and $L_y^{\phi''_\ell}$.

Proposition 3.7.9. *Suppose Assumption 3.3.2 holds that $\xi_i \in L^4(\Omega; \mathbb{R})$ for all $i \in I_N$. For all $\phi \in \Pi^N$, $h, \ell \in I_N$ with $h \neq \ell$ and $\phi'_h, \phi''_\ell \in \Pi$, let $\boldsymbol{u}^\phi = (\phi_i(\cdot, X_i^\phi, \boldsymbol{X}^\phi))_{i\in I_N}$, let $\boldsymbol{v}^{\phi,\phi'_h}$ be defined in* (3.40)*, and let $\boldsymbol{w}^{\phi,\phi'_h,\phi''_\ell}$ be defined in* (3.41)*. Then for all $i \in I_N$,*

$$\|u_i^\phi\|^2_{\mathcal{H}^2(\mathbb{R})} \leq C, \qquad \|v_i^{\phi,\phi'_h}\|^2_{\mathcal{H}^2(\mathbb{R})} \leq C\left(\delta_{h,i} + \frac{1}{N^2}(L_y^\phi)^2\right),$$
$$\|w_i^{\phi,\phi'_h,\phi''_\ell}\|^2_{\mathcal{H}^2(\mathbb{R})} \leq C\left((\delta_{h,i} + \delta_{\ell,i})\frac{1}{N^2} + \frac{1}{N^4}\right) \max\{L_y^b, L_y^\phi, L_y^{\phi'_h}, L_y^{\phi''_\ell}\}^2,$$

where $C \geq 0$ is a constant depending only on the upper bounds of T, $\max_{i\in I_N} \mathbb{E}[|\xi_i|^4]$, $\max_{i\in I_N} \|\sigma_i\|_{L^4}$, L^b, L^ϕ, $L^{\phi'_h}$, $L^{\phi''_\ell}$, L_y^b, L_y^ϕ, $L_y^{\phi'_h}$ and $L_y^{\phi''_\ell}$.

We are now ready to prove Theorem 3.4.2 based on Propositions 3.7.6, 3.7.7, 3.7.8, and 3.7.9.

Proof of Theorem 3.4.2. To simplify the notation, we omit the dependence on ϕ in the superscript of all processes, i.e., $\mathbf{X} = \mathbf{X}^\phi, \mathbf{Y}^i = \mathbf{Y}^{\phi,\phi'_i}$. We denote by $C \geq 0$ a generic constant depending only on the upper bounds of T, $\max_{i\in I_N} \mathbb{E}[|\xi_i|^4]$, $\max_{i\in I_N} \|\sigma_i\|_{L^4}$, L^b, L^ϕ, $L^{\phi'_i}$, $L^{\phi''_j}$, L_y^b, L_y^ϕ, $L_y^{\phi'_i}$ and $L_y^{\phi''_j}$.

By the definition of $\frac{\delta^2 V_j}{\delta\phi_j\delta\phi_i}(\phi;\phi''_j,\phi'_i)$ in (3.43) and the fact that $\mathbf{Z}^{\phi,\phi'_i,\phi''_j}=\mathbf{Z}^{\phi,\phi''_j,\phi'_i}$ and $\boldsymbol{w}^{\phi,\phi'_i,\phi''_j}=\boldsymbol{w}^{\phi,\phi''_j,\phi'_i}$,

$$\begin{aligned}&\left|\frac{\delta^2 V_i}{\delta\phi_i\delta\phi_j}(\phi;\phi'_i,\phi''_j)-\frac{\delta^2 V_j}{\delta\phi_j\delta\phi_i}(\phi;\phi''_j,\phi'_i)\right|\\&=\mathbb{E}\left[\int_0^T\left\{\begin{pmatrix}\mathbf{Y}^i_t\\ \boldsymbol{v}^i_t\end{pmatrix}^\top\begin{pmatrix}\partial^2_{xx}\Delta^f_{i,j} & \partial^2_{xu}\Delta^f_{i,j}\\ \partial^2_{ux}\Delta^f_{i,j} & \partial^2_{uu}\Delta^f_{i,j}\end{pmatrix}(t,\cdot)\begin{pmatrix}\mathbf{Y}^j_t\\ \boldsymbol{v}^j_t\end{pmatrix}+\begin{pmatrix}\mathbf{Z}^{i,j}_t\\ \boldsymbol{w}^{i,j}_t\end{pmatrix}^\top\begin{pmatrix}\partial_x\Delta^f_{i,j}\\ \partial_u\Delta^f_{i,j}\end{pmatrix}(t,\cdot)\right\}\mathrm{d}t\right]\\&\quad+\mathbb{E}\left[(\mathbf{Y}^i_T)^\top(\partial^2_{xx}\Delta^g_{i,j})(\mathbf{X}_T)\mathbf{Y}^j_T+(\mathbf{Z}^{i,j}_T)^\top(\partial_x\Delta^g_{i,j})(\mathbf{X}_T)\right],\end{aligned}\tag{3.88}$$

where we write for simplicity $\partial^2_{xx}\Delta^f_{i,j}(t,\cdot)=\partial^2_{xx}(f_i-f_j)(t,\mathbf{X}_t,\boldsymbol{u}_t)$ and similarly for other derivatives. In the sequel, we derive upper bounds for all terms on the right-hand side of (3.88).

To estimate the term involving the Hessian of $\Delta^f_{i,j}$ in (3.88), observe that for all $t\in[0,T]$,

$$\begin{aligned}&\begin{pmatrix}\mathbf{Y}^i_t\\ \boldsymbol{v}^i_t\end{pmatrix}^\top\begin{pmatrix}\partial^2_{xx}\Delta^f_{i,j} & \partial^2_{xu}\Delta^f_{i,j}\\ \partial^2_{ux}\Delta^f_{i,j} & \partial^2_{uu}\Delta^f_{i,j}\end{pmatrix}(t,\cdot)\begin{pmatrix}\mathbf{Y}^j_t\\ \boldsymbol{v}^j_t\end{pmatrix}\\&=\sum_{h,\ell=1}^N(\partial^2_{x_hx_\ell}\Delta^f_{i,j})(t,\cdot)Y^i_{t,h}Y^j_{t,\ell}+\sum_{h,\ell=1}^N(\partial^2_{x_hu_\ell}\Delta^f_{i,j})(t,\cdot)Y^i_{t,h}v^j_{t,\ell}\\&\quad+\sum_{h,\ell=1}^N(\partial^2_{u_hx_\ell}\Delta^f_{i,j})(t,\cdot)v^i_{t,h}Y^j_{t,\ell}+\sum_{h,\ell=1}^N(\partial^2_{u_hu_\ell}\Delta^f_{i,j})(t,\cdot)v^i_{t,h}v^j_{t,\ell}.\end{aligned}\tag{3.89}$$

The first term on the right-hand side of (3.89) satisfies the identity:

$$\begin{aligned}\sum_{h,\ell=1}^N(\partial^2_{x_hx_\ell}\Delta^f_{i,j})(t,\cdot)Y^i_{t,h}Y^j_{t,\ell}&=(\partial^2_{x_ix_j}\Delta^f_{i,j})(t,\cdot)Y^i_{t,i}Y^j_{t,j}+\sum_{\ell\in I_N\setminus\{j\}}(\partial^2_{x_ix_\ell}\Delta^f_{i,j})(t,\cdot)Y^i_{t,i}Y^j_{t,\ell}\\&\quad+\sum_{h\in I_N\setminus\{i\}}\left((\partial^2_{x_hx_j}\Delta^f_{i,j})(t,\cdot)Y^i_{t,h}Y^j_{t,j}+\sum_{\ell\in I_N\setminus\{j\}}(\partial^2_{x_hx_\ell}\Delta^f_{i,j})(t,\cdot)Y^i_{t,h}Y^j_{t,\ell}\right),\end{aligned}$$

which yields the following estimate:

$$\begin{aligned}\left|\mathbb{E}\left[\int_0^T\sum_{h,\ell=1}^N(\partial^2_{x_hx_\ell}\Delta^f_{i,j})(t,\cdot)Y^i_{t,h}Y^j_{t,\ell}\mathrm{d}t\right]\right|&\le\|\partial^2_{x_ix_j}\Delta^f_{i,j}\|_{L^\infty}\|Y^i_iY^j_j\|_{\mathcal{H}^1(\mathbb{R})}\\&\quad+\sum_{\ell\in I_N\setminus\{j\}}\|\partial^2_{x_ix_\ell}\Delta^f_{i,j}\|_{L^\infty}\|Y^i_iY^j_\ell\|_{\mathcal{H}^1(\mathbb{R})}\\&\quad+\sum_{h\in I_N\setminus\{i\}}\left(\|\partial^2_{x_hx_j}\Delta^f_{i,j}\|_{L^\infty}\|Y^i_hY^j_j\|_{\mathcal{H}^1(\mathbb{R})}+\sum_{\ell\in I_N\setminus\{j\}}\|\partial^2_{x_hx_\ell}\Delta^f_{i,j}\|_{L^\infty}\|Y^i_hY^j_\ell\|_{\mathcal{H}^1(\mathbb{R})}\right),\end{aligned}\tag{3.90}$$

and

$$
\begin{aligned}
(3.90) \leq C\Bigg\{ &\|\partial^2_{x_i x_j}\Delta^f_{i,j}\|_{L^\infty} + \frac{L_y^{b+\phi}}{N}\Bigg(\sum_{\ell\in I_N\setminus\{j\}} \|\partial^2_{x_i x_\ell}\Delta^f_{i,j}\|_{L^\infty} + \sum_{h\in I_N\setminus\{i\}} \|\partial^2_{x_h x_j}\Delta^f_{i,j}\|_{L^\infty}\Bigg) \\
&+ \frac{(L_y^{b+\phi})^2}{N^2} \sum_{h\in I_N\setminus\{i\}} \sum_{\ell\in I_N\setminus\{j\}} \|\partial^2_{x_h x_\ell}\Delta^f_{i,j}\|_{L^\infty}\Bigg\}, \qquad (3.91)
\end{aligned}
$$

where the second inequality follows from using the Cauchy-Schwarz inequality and Proposition 3.7.7.

Similarly, using Propositions 3.7.7 and 3.7.9, the second and third terms in (3.89) can be bounded by

$$
\begin{aligned}
&\left|\mathbb{E}\left[\int_0^T \sum_{h,\ell=1}^N (\partial^2_{x_h u_\ell}\Delta^f_{i,j})(t,\cdot) Y^i_{t,h} v^j_{t,\ell} \mathrm{d}t\right]\right| + \left|\mathbb{E}\left[\int_0^T \sum_{h,\ell=1}^N (\partial^2_{u_h x_\ell}\Delta^f_{i,j})(t,\cdot) v^i_{t,h} Y^j_{t,\ell} \mathrm{d}t\right]\right| \\
\leq & C\Bigg\{ \|\partial^2_{x_i u_j}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_i x_j}\Delta^f_{i,j}\|_{L^\infty} + \frac{\max\{L_y^\phi, L_y^{b+\phi}\}}{N}\Bigg(\sum_{\ell\in I_N\setminus\{j\}} (\|\partial^2_{x_i u_\ell}\Delta^f_{i,j}\|_{L^\infty} \\
&+ \|\partial^2_{u_i x_\ell}\Delta^f_{i,j}\|_{L^\infty}) + \sum_{h\in I_N\setminus\{i\}} (\|\partial^2_{x_h u_j}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_h x_j}\Delta^f_{i,j}\|_{L^\infty})\Bigg) \qquad (3.92) \\
&+ \frac{\max\{L_y^\phi, L_y^{b+\phi}\}^2}{N^2} \sum_{\substack{h\in I_N\setminus\{i\} \\ \ell\in I_N\setminus\{j\}}} (\|\partial^2_{x_h u_\ell}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_h x_\ell}\Delta^f_{i,j}\|_{L^\infty})\Bigg\},
\end{aligned}
$$

and the fourth term in (3.89) can be bounded by

$$
\begin{aligned}
&\left|\mathbb{E}\left[\int_0^T \sum_{h,\ell=1}^N (\partial^2_{u_h u_\ell}\Delta^f_{i,j})(t,\cdot) v^i_{t,h} v^j_{t,\ell} \mathrm{d}t\right]\right| \leq C\Bigg\{ \|\partial^2_{u_i u_j}\Delta^f_{i,j}\|_{L^\infty} + \frac{L_y^\phi}{N}\Bigg(\sum_{\ell\in I_N\setminus\{j\}} \|\partial^2_{u_i u_\ell}\Delta^f_{i,j}\|_{L^\infty} \\
&+ \sum_{h\in I_N\setminus\{i\}} \|\partial^2_{u_h u_j}\Delta^f_{i,j}\|_{L^\infty}\Bigg) + \sum_{\substack{h\in I_N\setminus\{i\} \\ \ell\in I_N\setminus\{j\}}} \|\partial^2_{u_h u_\ell}\Delta^f_{i,j}\|_{L^\infty} \frac{(L_y^\phi)^2}{N^2}\Bigg\}. \qquad (3.93)
\end{aligned}
$$

Combining (3.91), (3.92) and (3.93) and setting $\overline{L}_y = \max\{L_y^b, L_y^\phi, L_y^{\phi'_i}, L_y^{\phi''_j}\}$ yield the bound of (3.89):

$$
\begin{aligned}
&\left|\mathbb{E}\left[\int_0^T \begin{pmatrix}\mathbf{Y}_t^i \\ \boldsymbol{v}_t^i\end{pmatrix}^\top \begin{pmatrix}\partial^2_{xx}\Delta^f_{i,j} & \partial^2_{xu}\Delta^f_{i,j} \\ \partial^2_{ux}\Delta^f_{i,j} & \partial^2_{uu}\Delta^f_{i,j}\end{pmatrix}(t,\cdot)\begin{pmatrix}\mathbf{Y}_t^j \\ \boldsymbol{v}_t^j\end{pmatrix}\mathrm{d}t\right]\right| \qquad (3.94)\\
&\leq C\Bigg\{\|\partial^2_{x_i x_j}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{x_i u_j}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_i x_j}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_i u_j}\Delta^f_{i,j}\|_{L^\infty} \\
&\quad + \frac{\overline{L}_y}{N}\Bigg(\sum_{\ell\in I_N\setminus\{j\}}\Big(\|\partial^2_{x_i x_\ell}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{x_i u_\ell}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_i x_\ell}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_i u_\ell}\Delta^f_{i,j}\|_{L^\infty}\Big) \\
&\quad + \sum_{h\in I_N\setminus\{i\}}\Big(\|\partial^2_{x_h x_j}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{x_h u_j}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_h x_j}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_h u_j}\Delta^f_{i,j}\|_{L^\infty}\Big)\Bigg) \\
&\quad + \frac{\overline{L}_y}{N^2}\sum_{\substack{h\in I_N\setminus\{i\},\\ \ell\in I_N\setminus\{j\}}}\Big(\|\partial^2_{x_h x_\ell}\Delta^f_{i,j}\|_{L^\infty} + |\partial^2_{x_h u_\ell}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_h x_\ell}\Delta^f_{i,j}\|_{L^\infty} + \|\partial^2_{u_h u_\ell}\Delta^f_{i,j}\|_{L^\infty}\Big)\Bigg\}.
\end{aligned}
$$

To estimate the term involving the gradient of $\Delta^f_{i,j}$ in (3.88), note that

$$
\begin{pmatrix}\mathbf{Z}_t^{i,j} \\ \boldsymbol{w}_t^{i,j}\end{pmatrix}^\top \begin{pmatrix}\partial_x\Delta^f_{i,j} \\ \partial_u\Delta^f_{i,j}\end{pmatrix}(t,\cdot) = \sum_{h=1}^N(\partial_{x_h}\Delta^f_{i,j})(t,\cdot)Z^{i,j}_{t,h} + \sum_{h=1}^N(\partial_{u_h}\Delta^f_{i,j})(t,\cdot)w^{i,j}_{t,h},
$$

for all $t\in[0,T]$. The fundamental theorem of calculus implies that for all $h\in I_N$ and $(t,x,u)\in[0,T]\in\mathbb{R}^N\times\mathbb{R}^N$,

$$
|(\partial_{x_h}\Delta^f_{i,j})(t,x,u)| \leq |(\partial_{x_h}\Delta^f_{i,j})(t,0,0)| + \sum_{\ell=1}^N\Big(\|\partial^2_{x_h x_\ell}\Delta^f_{i,j}\|_{L^\infty}|x_\ell| + \|\partial^2_{x_h u_\ell}\Delta^f_{i,j}\|_{L^\infty}|u_\ell|\Big),
$$

which implies that

$$
\begin{aligned}
&\left|\mathbb{E}\left[\int_0^T\sum_{h=1}^N(\partial_{x_h}\Delta^f_{i,j})(t,\cdot)Z^{i,j}_{t,h}\mathrm{d}t\right]\right| \leq \sum_{h\in\{i,j\}}\Bigg(\|(\partial_{x_h}\Delta^f_{i,j})(\cdot,0,0)\|_{L^2}\|Z^{i,j}_h\|_{\mathcal{H}^2(\mathbb{R})} \\
&\quad + \sum_{\ell=1}^N\Big(\|\partial^2_{x_h x_\ell}\Delta^f_{i,j}\|_{L^\infty}\|X_\ell Z^{i,j}_h\|_{\mathcal{H}^1(\mathbb{R})} + \|\partial^2_{x_h u_\ell}\Delta^f_{i,j}\|_{L^\infty}\|u_\ell Z^{i,j}_h\|_{\mathcal{H}^1(\mathbb{R})}\Big)\Bigg) \\
&\quad + \sum_{h\in I_N\setminus\{i,j\}}\Big(\|(\partial_{x_h}\Delta^f_{i,j})(\cdot,0,0)\|_{L^2}\|Z^{i,j}_h\|_{\mathcal{H}^2(\mathbb{R})} + \sum_{\ell=1}^N\big(\|\partial^2_{x_h x_\ell}\Delta^f_{i,j}\|_{L^\infty}\|X_\ell Z^{i,j}_h\|_{\mathcal{H}^1(\mathbb{R})} \\
&\qquad + \|\partial^2_{x_h u_\ell}\Delta^f_{i,j}\|_{L^\infty}\|u_\ell Z^{i,j}_h\|_{\mathcal{H}^1(\mathbb{R})}\big)\Big).
\end{aligned}
$$

Then by the Cauchy-Schwarz inequality and Propositions 3.7.6, 3.7.8 and 3.7.9,

$$\begin{aligned}
&\left|\mathbb{E}\left[\int_0^T \sum_{h=1}^N (\partial_{x_h}\Delta^f_{i,j})(t,\cdot) Z^{i,j}_{t,h}\mathrm{d}t\right]\right| \\
&\le C\overline{L}_y\Bigg\{\frac{1}{N}\sum_{h\in\{i,j\}}\bigg(\|(\partial_{x_h}\Delta^f_{i,j})(\cdot,0,0)\|_{L^2}+\sum_{\ell=1}^N\Big(\|\partial^2_{x_h x_\ell}\Delta^f_{i,j}\|_{L^\infty}+\|\partial^2_{x_h u_\ell}\Delta^f_{i,j}\|_{L^\infty}\Big)\bigg) \\
&\quad+\frac{1}{N^2}\sum_{h\in I_N\setminus\{i,j\}}\bigg(\|(\partial_{x_h}\Delta^f_{i,j})(\cdot,0,0)\|_{L^2}+\sum_{\ell=1}^N\Big(\|\partial^2_{x_h x_\ell}\Delta^f_{i,j}\|_{L^\infty}+\|\partial^2_{x_h u_\ell}\Delta^f_{i,j}\|_{L^\infty}\Big)\bigg)\Bigg\}.
\end{aligned} \tag{3.95}$$

where $\overline{L}_y=\max\{L^b_y, L^\phi_y, L^{\phi'_i}_y, L^{\phi''_j}_y\}$. Similar estimates show that

$$\begin{aligned}
&\left|\mathbb{E}\left[\int_0^T \sum_{h=1}^N (\partial_{u_h}\Delta^f_{i,j})(t,\cdot) w^{i,j}_{t,h}\mathrm{d}t\right]\right| \le C\overline{L}_y\Bigg\{\frac{1}{N}\sum_{h\in\{i,j\}}\bigg(\|(\partial_{u_h}\Delta^f_{i,j})(\cdot,0,0)\|_{L^2} \\
&\quad+\sum_{\ell=1}^N(\|\partial^2_{u_h x_\ell}\Delta^f_{i,j}\|_{L^\infty}+\|\partial^2_{u_h u_\ell}\Delta^f_{i,j}\|_{L^\infty})\bigg) \\
&\quad+\frac{1}{N^2}\sum_{h\in I_N\setminus\{i,j\}}\bigg(\|(\partial_{u_h}\Delta^f_{i,j})(\cdot,0,0)\|_{L^2}+\sum_{\ell=1}^N\Big(\|\partial^2_{u_h x_\ell}\Delta^f_{i,j}\|_{L^\infty}+\|\partial^2_{u_h u_\ell}\Delta^f_{i,j}\|_{L^\infty}\Big)\bigg)\Bigg\}.
\end{aligned} \tag{3.96}$$

Finally, using similar arguments as those for (3.91) and (3.95) allows for estimating the terms involving $\Delta^g_{i,j}$ in (3.88):

$$\begin{aligned}
&\left|\mathbb{E}\left[(\mathbf{Y}^i_T)^\top(\partial^2_{xx}\Delta^g_{i,j})(\mathbf{X}_T)\mathbf{Y}^j_T+(\mathbf{Z}^{i,j}_T)^\top(\partial_x\Delta^g_{i,j})(\mathbf{X}_T)\right]\right| \\
&\le C\Bigg\{\|\partial^2_{x_i x_j}\Delta^g_{i,j}\|_{L^\infty}+\overline{L}_y\Bigg[\frac{1}{N}\Bigg(\sum_{h\in\{i,j\}}|(\partial_{x_h}\Delta^g_{i,j})(0)|+\sum_{h\in\{i,j\},\ell\in I_N}\|\partial^2_{x_h x_\ell}\Delta^g_{i,j}\|_{L^\infty}\Bigg) \\
&\quad+\frac{1}{N^2}\Bigg(\sum_{h\in I_N\setminus\{i,j\}}|(\partial_{x_h}\Delta^g_{i,j})(0)|+\sum_{h\in I_N\setminus\{i,j\},\ell\in I_N}\|\partial^2_{x_h x_\ell}\Delta^g_{i,j}\|_{L^\infty} \\
&\quad+\sum_{h\in I_N\setminus\{i\},\ell\in I_N\setminus\{j\}}\|\partial^2_{x_h x_\ell}\Delta^g_{i,j}\|_{L^\infty}\Bigg)\Bigg]\Bigg\}.
\end{aligned} \tag{3.97}$$

Note that the last two terms in the last line can be replaced by $\sum_{h\in I_N\setminus\{i,j\},\ell\in I_N\setminus\{i,j\}}\|\partial^2_{x_h x_\ell}\Delta^g_{i,j}\|_{L^\infty}$, as the remaining ones can be absorbed in the terms with $1/N$.

Consequently, by using (3.88) and grouping the terms in the estimates (3.94), (3.95), (3.96) and (3.97) according to the orders $1/N$ and $1/N^2$, we have

$$\left|\frac{\delta^2 V_i}{\delta\phi_i\delta\phi_j}(\phi;\phi_i',\phi_j'') - \frac{\delta^2 V_j}{\delta\phi_j\delta\phi_i}(\phi;\phi_j'',\phi_i')\right| \le C\left(C_{V,1}^{i,j} + \overline{L}_y\left(\frac{1}{N}C_{V,2}^{i,j} + \frac{1}{N^2}C_{V,3}^{i,j}\right)\right),$$

where $C_{V,1}^{i,j}$, $C_{V,2}^{i,j}$, and $C_{V,2}^{i,j}$ are given in (3.44), (3.45), and (3.46). This finishes the proof. □

3.7.4 Proof of Theorem 3.6.1

Proof of Theorem 3.6.1. It suffices to show $\boldsymbol{u}^*$ is a minimizer of (3.65) over $\mathcal{H}^2(\mathbb{R}^N)$. Define $\hat{V} : [0,T]\times\mathcal{P}_2(\mathcal{S}) \to \mathbb{R}$ such that for all $(t,\mu) \in [0,T]\times\mathcal{P}_2(\mathcal{S})$,

$$\hat{V}(t,\mu) = \operatorname{tr}(M_0(t)\overline{\mu}_2) + \begin{pmatrix}\overline{\mu}\\ \overline{\mu}_1\end{pmatrix}^\top M_1(t)\begin{pmatrix}\overline{\mu}\\ \overline{\mu}_1\end{pmatrix} + 2M_2(t)^\top\begin{pmatrix}\overline{\mu}\\ \overline{\mu}_1\end{pmatrix} + M_3(t),$$

where $\overline{\mu} := \int_{\mathcal{S}} \mathbb{x}\mu(\mathrm{d}(\mathbb{x},r))$, $\overline{\mu}_1 := \int_{\mathcal{S}} r\mathbb{x}\mu(\mathrm{d}(\mathbb{x},r))$, $\overline{\mu}_2 := \int_{\mathcal{S}} \mathbb{x}\mathbb{x}^\top\mu(\mathrm{d}(\mathbb{x},r))$, and $M_3 \in C([0,T];\mathbb{R})$ satisfies

$$\dot{M}_3 + \operatorname{tr}\left(\Sigma\Sigma^\top\left(M_0 + \begin{pmatrix}\mathbb{I}_{2N}\\ \frac{1}{2}\mathbb{I}_{2N}\end{pmatrix}^\top M_1\begin{pmatrix}\mathbb{I}_{2N}\\ \frac{1}{2}\mathbb{I}_{2N}\end{pmatrix}\right)\right) - (\tilde{I}M_2)^\top\tilde{I}M_2 = 0; \quad M_3(T) = 0.$$

We shall prove $\hat{V}$ satisfies the optimality condition (3.51). In the sequel, the time variable of all coefficients will be dropped when there is no risk of confusion.

Let $\boldsymbol{u} \in \mathcal{H}^2(\mathbb{R}^N)$, let $\mathbb{X}^{\mathfrak{r},\boldsymbol{u}} \in \mathcal{S}^2(\mathbb{R}^{2N})$ satisfy (3.64), and let $\mu_t^{\mathfrak{r},\boldsymbol{u}} := \mathcal{L}(\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}},\mathfrak{r}|\mathcal{F}_t)$ for all t. By Itô's formula in [65] (see also [33, Theorem 4.17]),

$$\begin{aligned}\hat{V}(T,\mu_T^{\mathfrak{r},\boldsymbol{u}}) - \hat{V}(0,\mu_0^{\mathfrak{r},\boldsymbol{u}}) = \bar{\mathbb{E}}\Bigg[\int_0^T \Bigg\{&(\partial_t\hat{V})(t,\mu_t^{\mathfrak{r},\boldsymbol{u}}) + \left(A(t)\tilde{\mathbb{X}}_t^{\tilde{\mathfrak{r}},\boldsymbol{u}} + \mathcal{I}_{\tilde{\mathfrak{r}}}\boldsymbol{u}_t\right)^\top \partial_{\mathbb{x}}\frac{\delta\hat{V}}{\delta\mu}(t,\mu_t^{\mathfrak{r},\boldsymbol{u}},\tilde{\mathbb{X}}_t^{\tilde{\mathfrak{r}},\boldsymbol{u}},\tilde{\mathfrak{r}})\\ &+ \frac{1}{2}\operatorname{tr}\left(\Sigma(t)\Sigma(t)^\top\partial^2_{\mathbb{x}\mathbb{x}}\frac{\delta\hat{V}}{\delta\mu}(t,\mu_t^{\mathfrak{r},\boldsymbol{u}},\tilde{\mathbb{X}}_t^{\tilde{\mathfrak{r}},\boldsymbol{u}},\tilde{\mathfrak{r}})\right)\\ &+ \frac{1}{2}\operatorname{tr}\left(\Sigma(t)\Sigma(t)^\top\partial^2_{\mathbb{x}\mathbb{x}'}\frac{\delta^2\hat{V}}{\delta^2\mu}(t,\mu_t^{\mathfrak{r},\boldsymbol{u}},\tilde{\mathbb{X}}_t^{\tilde{\mathfrak{r}},\boldsymbol{u}},\tilde{\mathfrak{r}},\hat{\mathbb{X}}_t^{\hat{\mathfrak{r}},\boldsymbol{u}},\hat{\mathfrak{r}})\right)\Bigg\}\mathrm{d}t\,\Bigg|\,\mathcal{F}_T\Bigg],\end{aligned} \tag{3.98}$$

where $(\tilde{\mathbb{X}}^{\tilde{\mathfrak{r}},\boldsymbol{u}},\tilde{\mathfrak{r}})$ and $(\hat{\mathbb{X}}^{\hat{\mathfrak{r}},\boldsymbol{u}},\hat{\mathfrak{r}})$ are conditional independent copies of $(\mathbb{X}^{\mathfrak{r},\boldsymbol{u}},\mathfrak{r})$ given $\mathcal{F}_T$ defined on an enlarged probability space $(\bar{\Omega},\bar{\mathcal{F}},\bar{\mathbb{P}})$ with $\mathcal{F}_T \subset \bar{\mathcal{F}}$, and $\bar{\mathbb{E}}[\cdot|\mathcal{F}_T]$ is the conditional expectation in the enlarged probability space.

We now compute the right-hand side of (3.98). Note that $\boldsymbol{u}$ and $\mu^{\mathfrak{r},\boldsymbol{u}}$ are measurable with respect to $\mathcal{F}_T$, and $\mu_t^{\mathfrak{r},\boldsymbol{u}} = \mathcal{L}(\tilde{\mathbb{X}}^{\tilde{\mathfrak{r}},\boldsymbol{u}},\tilde{\mathfrak{r}}_t|\mathcal{F}_T) = \mathcal{L}(\hat{\mathbb{X}}_t^{\hat{\mathfrak{r}},\boldsymbol{u}},\hat{\mathfrak{r}}|\mathcal{F}_T)$ for all $t \in [0,T]$. Hence for all

$t \in [0,T]$, by the symmetry of $M_0(t)$ and $M_1(t)$,

$$
\begin{aligned}
&\bar{\mathbb{E}}\left[\left(A(t)\tilde{\mathbb{X}}_t^{\tilde{\mathfrak{r}},\boldsymbol{u}} + \mathcal{I}_{\tilde{\mathfrak{r}}}\boldsymbol{u}_t\right)^\top \partial_{\mathbb{x}} \frac{\delta \hat{V}}{\delta \mu}(t, \mu_t^{\mathfrak{r},\boldsymbol{u}}, \tilde{\mathbb{X}}_t^{\tilde{\mathfrak{r}},\boldsymbol{u}}, \tilde{\mathfrak{r}}) \middle| \mathcal{F}_T\right] \\
=&2\int_{\mathcal{S}} (A(t)\mathbb{x} + \mathcal{I}_r \boldsymbol{u}_t)^\top \left(M_0(t)\mathbb{x} + \begin{pmatrix} \mathbb{I}_{2N} \\ r\mathbb{I}_{2N} \end{pmatrix}^\top M_1(t) \begin{pmatrix} \overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}} \\ (\overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}})_1 \end{pmatrix} + \begin{pmatrix} \mathbb{I}_{2N} \\ r\mathbb{I}_{2N} \end{pmatrix}^\top M_2(t)\right) \mathrm{d}\mu_t^{\mathfrak{r},\boldsymbol{u}}(\mathbb{x}, r) \\
=&2\Bigg\{ \operatorname{tr}\left(A^\top M_0(\overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}})_2\right) + \begin{pmatrix} \overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}} \\ (\overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}})_1 \end{pmatrix}^\top M_1 \begin{pmatrix} A & \mathbb{0}_{2N} \\ \mathbb{0}_{2N} & A \end{pmatrix} \begin{pmatrix} \overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}} \\ (\overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}})_1 \end{pmatrix} \\
&\quad + M_2^\top \begin{pmatrix} A & \mathbb{0}_{2N} \\ \mathbb{0}_{2N} & A \end{pmatrix} \begin{pmatrix} \overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}} \\ (\overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}})_1 \end{pmatrix} + \boldsymbol{u}_t^\top \left[K_{M_0,M_1} \begin{pmatrix} \overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}} \\ (\overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}})_1 \end{pmatrix} + \tilde{I} M_2\right] \Bigg\},
\end{aligned}
\tag{3.99}
$$

where the last term used the fact that the marginal distribution of $\mu_t^{\mathfrak{r},\boldsymbol{u}}$ on $[0,1]$ is the uniform distribution. Moreover,

$$
\begin{aligned}
\frac{1}{2}\operatorname{tr}\left(\Sigma(t)\Sigma(t)^\top \partial^2_{\mathbb{x}\mathbb{x}} \frac{\delta \hat{V}}{\delta \mu}(t, \mu_t^{\mathfrak{r},\boldsymbol{u}}, \tilde{\mathbb{X}}_t^{\tilde{\mathfrak{r}},\boldsymbol{u}}, \tilde{\mathfrak{r}})\right) &= \operatorname{tr}\left(\Sigma(t)\Sigma(t)^\top M_0(t)\right), \\
\frac{1}{2}\operatorname{tr}\left(\Sigma(t)\Sigma(t)^\top \partial^2_{\mathbb{x}\mathbb{x}'} \frac{\delta^2 \hat{V}}{\delta^2 \mu}(t, \mu_t^{\mathfrak{r},\boldsymbol{u}}, \tilde{\mathbb{X}}_t^{\tilde{\mathfrak{r}},\boldsymbol{u}}, \tilde{\mathfrak{r}}, \hat{\mathbb{X}}_t^{\hat{\mathfrak{r}},\boldsymbol{u}}, \hat{\mathfrak{r}})\right) &= \operatorname{tr}\left(\Sigma(t)\Sigma(t)^\top \begin{pmatrix} \mathbb{I}_{2N} \\ \frac{1}{2}\mathbb{I}_{2N} \end{pmatrix}^\top M_1(t) \begin{pmatrix} \mathbb{I}_{2N} \\ \frac{1}{2}\mathbb{I}_{2N} \end{pmatrix}\right).
\end{aligned}
\tag{3.100}
$$

Observe further that for all $t \in [0,T]$, by completing the squares,

$$
\begin{aligned}
&\boldsymbol{u}_t^\top 2 \left[K_{M_0,M_1} \begin{pmatrix} \overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}} \\ (\overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}})_1 \end{pmatrix} + \tilde{I} M_2\right] \\
&\geq -\boldsymbol{u}_t^\top \boldsymbol{u}_t - \left[K_{M_0,M_1} \begin{pmatrix} \overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}} \\ (\overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}})_1 \end{pmatrix} + \tilde{I} M_2\right]^\top \left[K_{M_0,M_1} \begin{pmatrix} \overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}} \\ (\overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}})_1 \end{pmatrix} + \tilde{I} M_2\right] \\
&\geq -\int_{\mathcal{S}} 2r \boldsymbol{u}_t^\top \boldsymbol{u}_t \mathrm{d}\mu_t^{\mathfrak{r},\boldsymbol{u}} - \begin{pmatrix} \overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}} \\ (\overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}})_1 \end{pmatrix}^\top K_{M_0,M_1}^\top K_{M_0,M_1} \begin{pmatrix} \overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}} \\ (\overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}})_1 \end{pmatrix} \\
&\quad - 2(\tilde{I}M_2)^\top K_{M_0,M_1} \begin{pmatrix} \overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}} \\ (\overline{\mu_t^{\mathfrak{r},\boldsymbol{u}}})_1 \end{pmatrix} - (\tilde{I}M_2)^\top \tilde{I} M_2.
\end{aligned}
\tag{3.101}
$$

Hence combining (3.98), (3.99), (3.100), and (3.101), and using the ODEs for M_0, M_1 and M_2 yield

$$
\hat{V}(T, \mu_T^{\mathfrak{r},\boldsymbol{u}}) - \hat{V}(0, \mu_0^{\mathfrak{r},\boldsymbol{u}}) \geq \int_0^T - \int_{\mathcal{S}} (\mathbb{x}^\top Q \mathbb{x} + 2r \boldsymbol{u}_t^\top \boldsymbol{u}_t) \mathrm{d}\mu_t^{\mathfrak{r},\boldsymbol{u}} \mathrm{d}t,
\tag{3.102}
$$

from which, by using $\hat{V}(T,\mu_T^{\mathfrak{r},\boldsymbol{u}}) = \int_{\mathcal{S}} \left(\mathbb{x}^\top \bar{Q}\mathbb{x} + 2\mathfrak{p}^\top \mathbb{x}\right) \mathrm{d}\mu_T^{\mathfrak{r},\boldsymbol{u}}$ and taking the expectation, we obtain that

$$\hat{V}\left(0, \delta_{\mathsf{vcat}(x_1,\dots,x_N,0_{N^2d})} \otimes \mathrm{Unif}(0,1)\right) \leq \Phi(\boldsymbol{u}), \quad \forall \boldsymbol{u} \in \mathcal{H}^2(\mathbb{R}^N).$$

Finally, consider the feedback map $\hat{a}(t,\mu) \coloneqq -\left[K_{M_0,M_1}(t)\begin{pmatrix}\overline{\mu} \\ \overline{\mu}_1\end{pmatrix} + \tilde{I}M_2(t)\right]$ for all $(t,\mu) \in [0,T]\times\mathcal{P}_2(\mathcal{S})$. Since $\mathfrak{r}$ is fixed, by [50, Theorem A.3] and the boundedness of K_{M_0,M_1} and M_2, the dynamics

$$\mathrm{d}\mathbb{X}_t = \big(A(t)\mathbb{X}_t + \mathcal{I}_{\mathfrak{r}}\hat{a}\big(t,\mathcal{L}(\mathbb{X}_t,\mathfrak{r} \mid \mathcal{F}_t)\big)\big)\,\mathrm{d}t + \Sigma(t)\mathrm{d}W_t, \quad \mathbb{X}_0 = \mathsf{vcat}(x_1,\cdots,x_N,0_N), \tag{3.103}$$

admits a unique $\mathbb{G}$-adapted strong solution $(\hat{\mathbb{X}},\mathfrak{r})$ satisfying $\mathbb{E}[\sup_{t\in[0,T]}||\hat{\mathbb{X}}_t|^p] < \infty$ for any $p \geq 2$. Thus the control $\boldsymbol{u}_t^* = \hat{a}\big(t,\mathcal{L}(\hat{\mathbb{X}}_t,\mathfrak{r} \mid \mathcal{F}_t)\big)$, $t \in [0,T]$, is in $\mathcal{A}^{(N)} = \prod_{i\in[N]}\mathcal{A}_i$, and achieves the equality in (3.101). This implies (3.102) is an equality and hence $\boldsymbol{u}^*$ is the minimizer of Φ. Since Φ is the α_N-potential function of $\mathcal{G}_{\mathrm{LQ}}$, $\boldsymbol{u}^*$ is an α_N-NE of $\mathcal{G}_{\mathrm{LQ}}$ by Proposition 3.2.1.

To derive that dynamics of $F_t \coloneqq \begin{pmatrix}\mathbb{E}[\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}^*}|\mathcal{F}_t] \\ \mathbb{E}[\mathfrak{r}\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}^*}|\mathcal{F}_t]\end{pmatrix}$, $t \in [0,T]$, using (3.103),

$$\begin{aligned}\mathrm{d}\begin{pmatrix}\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}^*} \\ \mathfrak{r}\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}^*}\end{pmatrix} ={}& \left(\begin{pmatrix}A(t) & \mathbb{0}_{2N} \\ \mathbb{0}_{2N} & A(t)\end{pmatrix}\begin{pmatrix}\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}^*} \\ \mathfrak{r}\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}^*}\end{pmatrix} - \begin{pmatrix}\mathcal{I}_{\mathfrak{r}} \\ \mathfrak{r}\mathcal{I}_{\mathfrak{r}}\end{pmatrix}\left[K_{M_0,M_1}(t)\begin{pmatrix}\mathbb{E}[\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}^*}|\mathcal{F}_t] \\ \mathbb{E}[\mathfrak{r}\mathbb{X}_t^{\mathfrak{r},\boldsymbol{u}^*}|\mathcal{F}_t]\end{pmatrix} + \tilde{I}M_2(t)\right]\right)\mathrm{d}t \\ &+ \begin{pmatrix}\Sigma(t) \\ \mathfrak{r}\Sigma(t)\end{pmatrix}\mathrm{d}W_t.\end{aligned}$$

For each $t \in [0,T]$, taking the conditional expectation with respect to $\mathcal{F}_t$, applying the conditional Fubini Theorem and using the independence between $\mathfrak{r}$ and $\mathcal{F}_t$ yield the dynamics (3.70) of F. This completes the proof. □

3.8 Proofs of Technical Lemmas and Propositions

The following lemma quantifies the growth of $f \in \mathscr{F}^{0,2}([0,T]\times\mathbb{R}\times\mathbb{R}^N;\mathbb{R})$ in the space variables, which will be used to prove Proposition 3.7.3 (and also Proposition 3.7.6). The proof follows directly from the mean value theorem and hence is omitted.

Lemma 3.8.1. *Let $f \in \mathscr{F}^{0,2}([0,T]\times\mathbb{R}\times\mathbb{R}^N;\mathbb{R})$. Then for all $t \in [0,T], x \in \mathbb{R}$ and $y \in \mathbb{R}^N$, $|f(t,x,y)| \leq L^f(1+|x|) + \frac{L_y^f}{N}\sum_{i=1}^N |y_i|$.*

Proof of Proposition 3.7.3. Throughout this proof, we write $\mathbf{X} = \mathbf{X}^{\boldsymbol{u}}$ for notational simplicity. By (3.20) and Itô's formula, for all $t \in [0,T]$,

$$\begin{aligned}&\mathrm{d}|X_{t,i}|^p = |X_{t,i}|^{p-2}\left(pX_{t,i}\left(u_{t,i} + b_i(t,X_{t,i},\mathbf{X}_t)\right) + \frac{p(p-1)}{2}\sigma_i^2(t)\right)\mathrm{d}t + p|X_{t,i}|^{p-2}X_{t,i}\sigma_i(t)\mathrm{d}W_t^i, \\ &\quad X_{0,i}^p = x_i^p.\end{aligned}$$

Taking the expectation of both sides and using the fact that $\left(\int_0^t |X_{u,i}|^{p-2}X_{u,i}\sigma_i(u)\mathrm{d}W_u^i\right)_{t\geq 0}$ is a martingale (see [160, Problem 2.10.7]) yield that

$$\begin{aligned}\mathbb{E}[|X_{t,i}|^p] \leq & |x_i|^p + \mathbb{E}\Bigg[\int_0^t \Bigg(p|X_{s,i}|^{p-1}\Bigg(|u_i(s)| + L^b(1+|X_{s,i}|) + \frac{L_y^b}{N}\sum_{k=1}^N |X_{s,k}|\Bigg) \\ & + |X_{s,i}|^{p-2}\frac{p(p-1)}{2}\sigma_i^2(s)\Bigg)\mathrm{d}s\Bigg].\end{aligned}$$

By Young's inequality, for all $a,b\geq 0$, $ab \leq \frac{p-1}{p}a^{p/(p-1)} + \frac{1}{p}b^p$ and $ab \leq \frac{p-2}{p}a^{p/(p-2)} + \frac{2}{p}b^{p/2}$ if $p>2$. Hence

$$\begin{aligned}\mathbb{E}[|X_{t,i}|^p] &\leq |x_i|^p + \mathbb{E}\Bigg[\int_0^t \Bigg((p-1)|X_{s,i}|^p + |u_i(s)|^p + L^b(1+(2p-1)|X_{s,i}|^p) \\ &\quad + \frac{L_y^b}{N}\sum_{k=1}^N ((p-1)|X_{s,i}|^p + |X_{s,k}|^p) + \frac{p(p-1)}{2}\left(\frac{p-2}{p}|X_{s,i}|^p + \frac{2}{p}|\sigma_i(s)|^p\right)\Bigg)\mathrm{d}s\Bigg] \\ &\leq |x_i|^p + (p-1)\|\sigma_i\|_{L^p}^p + L^bT + \|u_i\|_{\mathcal{H}^p(\mathbb{R})}^p \\ &\quad + \int_0^t \Bigg(\left(L^b(2p-1) + L_y^b(p-1) + \frac{(p-1)p}{2}\right)\mathbb{E}[|X_{s,i}|^p] + \frac{L_y^b}{N}\sum_{k=1}^N \mathbb{E}[|X_{s,k}|^p]\Bigg)\mathrm{d}s.\end{aligned} \tag{3.104}$$

Summing up the above equation over the index $i\in[N]$ yields for all $t\in[0,T]$,

$$\begin{aligned}\sum_{i=1}^N \mathbb{E}[|X_{t,i}|^p] \leq & \sum_{i=1}^N \left(|x_i|^p + (p-1)\|\sigma_i\|_{L^p}^p + L^bT + \|u_i\|_{\mathcal{H}^p(\mathbb{R})}^p\right) \\ & + \int_0^t \Bigg(\left(L^b(2p-1) + L_y^b p + \frac{(p-1)p}{2}\right)\sum_{k=1}^N \mathbb{E}[|X_{s,k}|^p]\Bigg)\mathrm{d}s,\end{aligned}$$

which along with Gronwall's inequality implies that

$$\sum_{k=1}^N \mathbb{E}[|X_{t,k}|^p] \leq \sum_{k=1}^N \left(|x_k|^p + (p-1)\|\sigma_k\|_{L^p}^p + L^bT + \|u_k\|_{\mathcal{H}^p(\mathbb{R})}^p\right)e^{c_p(L^b+L_y^b+1)T},$$

for a constant $c_p\geq 1$ depending only on p. Substituting the above inequality into (3.104) and applying Gronwall's inequality yield

$$\begin{aligned}\mathbb{E}[|X_{t,i}|^p] &\leq |x_i|^p + (p-1)\|\sigma_i\|_{L^p}^p + L^b T + \int_0^t c_p\left(L^b + L^b_y + 1\right)\mathbb{E}[|X_{s,i}|^p]\mathrm{d}s \\ &\quad + \frac{L^b_y}{N}\sum_{k=1}^N \left(|x_k|^p + (p-1)\|\sigma_k\|_{L^p}^p + L^b T + \|u_k\|^p_{\mathcal{H}^p(\mathbb{R})}\right) e^{c_p(L^b+L^b_y+1)T} \\ &\leq \Bigg(|x_i|^p + (p-1)\|\sigma_i\|_{L^p}^p + L^b T + \|u_i\|^p_{\mathcal{H}^p(\mathbb{R})} \\ &\quad + \frac{L^b_y}{N}\sum_{k=1}^N \left(|x_k|^p + (p-1)\|\sigma_k\|_{L^p}^p + L^b T + \|u_k\|^p_{\mathcal{H}^p(\mathbb{R})}\right) e^{c_p(L^b+L^b_y+1)T}\Bigg) e^{c_p(L^b+L^b_y+1)T}.\end{aligned}$$

This finishes the proof. □

The following lemma will be used to estimate the sensitivity processes.

Lemma 3.8.2. *Let $p \geq 2$ and for each $i, j \in [N]$, let $B_i, \bar{B}_{i,j} : \Omega \times [0,T] \to \mathbb{R}$ be bounded adapted processes, and $f_i \in \mathcal{H}^p(\mathbb{R})$. Let $\boldsymbol{S} = (S_i)_{i=1}^N \in \mathcal{S}^p(\mathbb{R}^N)$ satisfy the following dynamics: for all $t \in [0,T]$,*

$$\mathrm{d}S_{t,i} = \left(B_i(t)S_{t,i} + \sum_{j=1}^N \bar{B}_{ij}(t)S_{t,j} + f_{t,i}\right)\mathrm{d}t, \quad S_{0,i} = 0; \quad \forall i = 1, \cdots, N. \tag{3.105}$$

Then for all $i \in [N]$,

$$\sup_{t\in[0,T]} \mathbb{E}[|S_{t,i}|^p] \leq (2T)^{p-1}\left(\|f_i\|^p_{\mathcal{H}^p(\mathbb{R})} + \left\|\sum_{k=1}^N |f_k|\right\|^p_{\mathcal{H}^p(\mathbb{R})} \|\bar{B}\|^p_\infty T^p e^{p(\|B\|_\infty + N\|\bar{B}\|_\infty)T}\right) e^{p\|B\|_\infty T}.$$

where $\|B\|_\infty = \max_{i\in[N]}\|B_i\|_{L^\infty}$ and $\|\bar{B}\|_\infty = \max_{i,j\in[N]}\|\bar{B}_{i,j}\|_{L^\infty}$.

Proof. By (3.105), for all $t \in [0,T]$ and $i \in [N]$,

$$|S_{t,i}| \leq \int_0^t \left(\|B\|_\infty |S_{u,i}| + \|\bar{B}\|_\infty \sum_{k=1}^N |S_{u,k}| + |f_{u,i}|\right)\mathrm{d}u. \tag{3.106}$$

Summarizing (3.106) over the index $i \in [N]$ yields for all $t \in [0,T]$,

$$\sum_{k=1}^N |S_{t,k}| \leq \int_0^t \left((\|B\|_\infty + N\|\bar{B}\|_\infty)\sum_{k=1}^N |S_{u,k}| + \sum_{k=1}^N |f_{u,k}|\right)\mathrm{d}u,$$

which along with Gronwall's inequality implies that

$$\sum_{k=1}^N |S_{t,k}| \leq \left(\int_0^T \sum_{k=1}^N |f_{u,i}|\,\mathrm{d}u\right) e^{(\|B\|_\infty + N\|\bar{B}\|_\infty)t}.$$

Substituting the above inequality into (3.106) yields for all $t > 0$,

$$|S_{t,i}| \leq \int_0^t \|B\|_\infty |S_{u,i}| \mathrm{d}u + \left(\left(\int_0^T \sum_{k=1}^N |f_{u,i}| \, \mathrm{d}u \right) \int_0^T \|\bar{B}\|_\infty e^{(\|B\|_\infty + N\|\bar{B}\|_\infty)u} \mathrm{d}u \right) + \int_0^T |f_{u,i}| \mathrm{d}u.$$

This with Gronwall's inequality shows that for all $t > 0$,

$$|S_{t,i}| \leq \left(\int_0^T |f_{u,i}| \mathrm{d}u + \left(\int_0^T \sum_{k=1}^N |f_{u,k}| \, \mathrm{d}u \right) \|\bar{B}\|_\infty T e^{(\|B\|_\infty + N\|\bar{B}\|_\infty)T} \right) e^{\|B\|_\infty T}.$$

Taking the p-th moments of both sides of the above inequality and using the fact that $(a+b)^p \leq 2^{p-1}(a^p + b^p)$ for all $a, b \geq 0$ yield

$$\begin{aligned} \mathbb{E}[|S_{t,i}|^p] &\leq \mathbb{E}\left[\left(\int_0^T |f_{u,i}| \mathrm{d}u + \left(\int_0^T \sum_{k=1}^N |f_{u,k}| \, \mathrm{d}u \right) \|\bar{B}\|_\infty T e^{(\|B\|_\infty + N\|\bar{B}\|_\infty)T} \right)^p \right] e^{p\|B\|_\infty T} \\ &\leq 2^{p-1} \mathbb{E}\left[\left(\int_0^T |f_{u,i}| \mathrm{d}u \right)^p + \left(\int_0^T \sum_{k=1}^N |f_{u,k}| \, \mathrm{d}u \right)^p \|\bar{B}\|_\infty^p T^p e^{p(\|B\|_\infty + N\|\bar{B}\|_\infty)T} \right] e^{p\|B\|_\infty T} \\ &\leq (2T)^{p-1} \left(\|f_i\|^p_{\mathcal{H}^p(\mathbb{R})} + \left\| \sum_{k=1}^N |f_k| \right\|^p_{\mathcal{H}^p(\mathbb{R})} \|\bar{B}\|_\infty^p T^p e^{p(\|B\|_\infty + N\|\bar{B}\|_\infty)T} \right) e^{p\|B\|_\infty T}. \end{aligned}$$

This proves the desired estimate. □

3.8.1 Proofs of Propositions 3.7.4 and 3.7.5

Proof of Proposition 3.7.4. To simplify the notation, we write $\mathbf{X} = \mathbf{X}^{\boldsymbol{u}}$ and $\mathbf{Y}^h = \mathbf{Y}^{\boldsymbol{u},u'_h}$. Applying Lemma 3.8.2 with $\mathbf{S} = \mathbf{Y}^h$, $B_i(t) = \partial_x b_i(t, X_{t,i}, \mathbf{X}_t)$, $\bar{B}_{i,j}(t) = \partial_{y_j} b_i(t, X_{t,i}, \mathbf{X}_t)$ and $f_{t,i} = \delta_{h,i} u'_{t,h}$ yields that for all $i \in [N]$,

$$\begin{aligned} \sup_{t\in[0,T]} \mathbb{E}[|Y^h_{t,i}|^p] &\leq (2T)^{p-1} \left(\|f_i\|^p_{\mathcal{H}^p(\mathbb{R})} + \left\| \sum_{k=1}^N |f_k| \right\|^p_{\mathcal{H}^p(\mathbb{R})} \frac{(L^b_y)^p}{N^p} T^p e^{p(L^b + L^b_y)T} \right) e^{pL^b T} \\ &\leq (2T)^{p-1} \left(\delta_{h,i} + \frac{(L^b_y)^p}{N^p} T^p e^{p(L^b + L^b_y)T} \right) \|u'_h\|^p_{\mathcal{H}^p(\mathbb{R})} e^{pL^b T}. \end{aligned} \tag{3.107}$$

where we used $\|\bar{B}_{i,j}\|_{L^\infty} \leq L^b/N$. □

Finally, to prove Proposition 3.7.5, we estimate the moment of the process $\mathfrak{f}_i^{\boldsymbol{u},u'_h,u''_\ell}$ defined in (3.24).

Lemma 3.8.3. *Suppose Assumption 3.3.2 holds. For all $\boldsymbol{u}\in\mathcal{H}^2(\mathbb{R}^N)$, $i,h,\ell\in[N]$ with $h\neq\ell$, and all $u'_h,u''_\ell\in\mathcal{H}^4(\mathbb{R})$, the process $\mathfrak{f}_i^{\boldsymbol{u},u'_h,u''_\ell}$ defined in* (3.24) *satisfies*

$$\|\mathfrak{f}_i^{\boldsymbol{u},u'_h,u''_\ell}\|_{\mathcal{H}^2(\mathbb{R})}\le C\left((\delta_{h,i}+\delta_{\ell,i})\frac{1}{N}+\frac{1}{N^2}\right)L^b_y\|u'_h\|_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|_{\mathcal{H}^4(\mathbb{R})},$$

where $C\ge 0$ is a constant depending only on the upper bounds of T, L^b and L^b_y.

Proof. Fix $h,\ell\in[N]$ with $h\neq\ell$. To simplify the notation, we write $\mathbf{X}=\mathbf{X}^{\boldsymbol{u}}$, $\mathbf{Y}^h=\mathbf{Y}^{\boldsymbol{u},u'_h}$ and $\mathbf{Y}^\ell=\mathbf{Y}^{\boldsymbol{u},u''_\ell}$. Observe that by Proposition 3.7.4, for all $i,j\in[N]$,

$$\begin{aligned}
&\left\|Y^h_iY^\ell_j\right\|^2_{\mathcal{H}^2(\mathbb{R})}\le T\sup_{t\in[0,T]}\mathbb{E}[|Y^h_{t,i}Y^\ell_{t,j}|^2]\le T\sup_{t\in[0,T]}\mathbb{E}[|Y^h_{t,i}|^4]^{\frac12}\mathbb{E}[|Y^\ell_{t,j}|^4]^{\frac12}\\
&\le T\|u'_h\|^2_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|^2_{\mathcal{H}^4(\mathbb{R})}\left(\delta_{h,i}C^{h,4}_Y+\frac{(L^b_y)^4}{N^4}\bar{C}^{h,4}_Y\right)^{\frac12}\left(\delta_{\ell,j}C^{\ell,4}_Y+\frac{(L^b_y)^4}{N^4}\bar{C}^{\ell,4}_Y\right)^{\frac12}\\
&\le T\|u'_h\|^2_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|^2_{\mathcal{H}^4(\mathbb{R})}\\
&\quad\times\left(\delta_{h,i}\delta_{\ell,j}(C^{h,4}_YC^{\ell,4}_Y)^{\frac12}+\frac{(L^b_y)^2}{N^2}\left(\delta_{h,i}(C^{h,4}_Y\bar{C}^{\ell,4}_Y)^{\frac12}+\delta_{\ell,j}(C^{\ell,4}_Y\bar{C}^{h,4}_Y)^{\frac12}\right)+\frac{(L^b_y)^4}{N^4}(\bar{C}^{h,4}_Y\bar{C}^{\ell,4}_Y)^{\frac12}\right)\\
&\le C\|u'_h\|^2_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|^2_{\mathcal{H}^4(\mathbb{R})}\left(\delta_{h,i}\delta_{\ell,j}+\frac{(L^b_y)^2}{N^2}\left(\delta_{h,i}+\delta_{\ell,j}\right)+\frac{(L^b_y)^4}{N^4}\right),
\end{aligned}\tag{3.108}$$

where the third line follows by noting $\sqrt{a_1+\cdots+a_N}\le\sqrt{a_1}+\cdots+\sqrt{a_N}$ for any $a_1,\cdots,a_N\ge 0$.

We now bound each term in (3.24). Observe that by (3.24), for all $t\in[0,T]$,

$$\begin{aligned}
\mathfrak{f}_{t,i}^{\boldsymbol{u},u'_h,u''_\ell}&=(\partial^2_{xx}b_i)(t,X_{t,i},\mathbf{X}_t)Y^h_{t,i}Y^\ell_{t,i}+\sum_{j=1}^N(\partial^2_{xy_j}b_i)(t,X_{t,i},\mathbf{X}_t)(Y^h_{t,i}Y^\ell_{t,j}+Y^\ell_{t,i}Y^h_{t,j})\\
&\quad+\sum_{j,k=1}^N(\partial^2_{y_jy_k}b_i)(t,X_{t,i},\mathbf{X}_t)Y^h_{t,j}Y^\ell_{t,k}.
\end{aligned}\tag{3.109}$$

Apply (3.108) with the fact that $\delta_{h,i}\delta_{\ell,i}=0$ as $h\neq\ell$ to get

$$\begin{aligned}
&\|(\partial^2_{xx}b_i)(\cdot,X_{\cdot,i},\mathbf{X}_\cdot)Y^h_iY^\ell_i\|_{\mathcal{H}^2(\mathbb{R})}\le L^b\|Y^h_iY^\ell_i\|^2_{\mathcal{H}^2(\mathbb{R})}\\
&\le C\|u'_h\|_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|_{\mathcal{H}^4(\mathbb{R})}\left((\delta_{h,i}+\delta_{\ell,i})\frac{L^b_y}{N}+\frac{(L^b_y)^2}{N^2}\right),
\end{aligned}\tag{3.110}$$

where C is a constant depending on T, $C^{h,4}_Y$ and $\bar{C}^{h,4}_Y$ for any $h\in[N]$.

We then estimate $\sum_{j=1}^N(\partial^2_{xy_j}b_i)(\cdot,X_{\cdot,i},\mathbf{X}_\cdot)(Y_i^hY_j^\ell+Y_i^\ell Y_j^h)$ in (3.109). The fact that $\partial^2_{xy_j}b_i$ is bounded by L^b_y/N and the inequality that $(\sum_{k=1}^N a_k)^2\le N\sum_{k=1}^N a_k^2$ for all $a_1,a_2,\cdots,a_N\in[0,\infty)$ show that

$$
\begin{aligned}
&\left\|\sum_{j=1}^N(\partial^2_{xy_j}b_i)(\cdot,X_{\cdot,i},\mathbf{X}_\cdot)(Y_i^hY_j^\ell+Y_i^\ell Y_j^h)\right\|^2_{\mathcal{H}^2(\mathbb{R})}\le\frac{(L^b_y)^2}{N^2}\left\|\sum_{j=1}^N(|Y_i^hY_j^\ell|+|Y_i^\ell Y_j^h|)\right\|^2_{\mathcal{H}^2(\mathbb{R})}\\
&=\frac{(L^b_y)^2}{N^2}\left\||Y_i^hY_\ell^\ell|+|Y_i^\ell Y_h^h|+\sum_{j\neq\ell}|Y_i^hY_j^\ell|+\sum_{j\neq h}|Y_i^\ell Y_j^h|\right\|^2_{\mathcal{H}^2(\mathbb{R})}\\
&\le 4\frac{(L^b_y)^2}{N^2}\left(\|Y_i^hY_\ell^\ell\|^2_{\mathcal{H}^2(\mathbb{R})}+\|Y_i^\ell Y_h^h\|^2_{\mathcal{H}^2(\mathbb{R})}+\left\|\sum_{j\neq\ell}|Y_i^hY_j^\ell|\right\|^2_{\mathcal{H}^2(\mathbb{R})}+\left\|\sum_{j\neq h}|Y_i^\ell Y_j^h|\right\|^2_{\mathcal{H}^2(\mathbb{R})}\right)\\
&\le 4\frac{(L^b_y)^2}{N^2}\left(\|Y_i^hY_\ell^\ell\|^2_{\mathcal{H}^2(\mathbb{R})}+\|Y_i^\ell Y_h^h\|^2_{\mathcal{H}^2(\mathbb{R})}+(N-1)\left(\sum_{j\neq\ell}\|Y_i^hY_j^\ell\|^2_{\mathcal{H}^2(\mathbb{R})}+\sum_{j\neq h}\|Y_i^\ell Y_j^h\|^2_{\mathcal{H}^2(\mathbb{R})}\right)\right),
\end{aligned}
$$

which along with (3.108) yields

$$
\begin{aligned}
&\left\|\sum_{j=1}^N(\partial^2_{xy_j}b_i)(\cdot,X_{\cdot,i},\mathbf{X}_\cdot)(Y_i^hY_j^\ell+Y_i^\ell Y_j^h)\right\|^2_{\mathcal{H}^2(\mathbb{R})}\\
&\le\frac{4(L^b_y)^2}{N^2}C\|u'_h\|^2_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|^2_{\mathcal{H}^4(\mathbb{R})}\Bigg[\left(\delta_{h,i}+\frac{(L^b_y)^2}{N^2}(\delta_{h,i}+1)+\frac{(L^b_y)^4}{N^4}\right)\\
&\quad+\left(\delta_{\ell,i}+\frac{(L^b_y)^2}{N^2}(1+\delta_{\ell,i})+\frac{(L^b_y)^4}{N^4}\right)+\frac{N-1}{N^2}\left(\delta_{h,i}+\frac{(L^b_y)^2}{N^2}+\delta_{\ell,i}+\frac{(L^b_y)^2}{N^2}\right)\Bigg],\\
&\le C\|u'_h\|^2_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|^2_{\mathcal{H}^4(\mathbb{R})}\left((\delta_{h,i}+\delta_{\ell,i})\frac{(L^b_y)^2}{N^2}+\frac{(L^b_y)^4}{N^4}\right).
\end{aligned}
\tag{3.111}
$$

Finally, we estimate $\sum_{j,k=1}^N(\partial^2_{y_jy_k}b_i)(t,X_i,\mathbf{X})Y_j^hY_k^\ell$ in (3.109). We write for simplicity $(\partial^2_{y_jy_k}b_i)(\cdot)=(\partial^2_{y_jy_k}b_i)(\cdot,X_{\cdot,i},\mathbf{X}_\cdot)$ for all $j,k\in[N]$. Since $h\neq\ell$, $[N]\times[N]=\{(h,\ell)\}\cup\{(h,h)\}\cup\{(\ell,\ell)\}\cup\big\{(h,k)\mid k\in[N]\setminus\{h,\ell\}\big\}\cup\big\{(j,\ell)\mid j\in[N]\setminus\{h,\ell\}\big\}\cup\big\{(j,k)\mid j\in[N]\setminus\{h\},k\in[N]\setminus\{\ell\}\big\}$, and hence

$$
\begin{aligned}
\sum_{j,k=1}^N(\partial^2_{y_jy_k}b_i)(\cdot)Y_j^hY_k^\ell&=(\partial^2_{y_hy_\ell}b_i)(\cdot)Y_h^hY_\ell^\ell+(\partial^2_{y_hy_h}b_i)(\cdot)Y_h^hY_h^\ell+(\partial^2_{y_\ell y_\ell}b_i)(\cdot)Y_\ell^hY_\ell^\ell\\
&\quad+\sum_{k\in[N]\setminus\{\ell,h\}}(\partial^2_{y_hy_k}b_i)(\cdot)Y_h^hY_k^\ell+\sum_{j\in[N]\setminus\{h,\ell\}}(\partial^2_{y_jy_\ell}b_i)(\cdot)Y_j^hY_\ell^\ell+\sum_{j\in[N]\setminus\{h\}}\sum_{k\in[N]\setminus\{\ell\}}(\partial^2_{y_jy_k}b_i)(\cdot)Y_j^hY_k^\ell.
\end{aligned}
\tag{3.112}
$$

To analyze the first line in (3.112), note that by (3.108) and $h \neq \ell$,

$$\begin{aligned}
\|(\partial^2_{y_h y_\ell} b_i)(\cdot) Y^h_h Y^\ell_\ell\|_{\mathcal{H}^2(\mathbb{R})} &\le \frac{L^b_y}{N^2}\|Y^h_h Y^\ell_\ell\|_{\mathcal{H}^2(\mathbb{R})} \le C\|u'_h\|_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|_{\mathcal{H}^4(\mathbb{R})}\frac{L^b_y}{N^2},\\
\|(\partial^2_{y_h y_h} b_i)(\cdot) Y^h_h Y^\ell_h\|_{\mathcal{H}^2(\mathbb{R})} &\le \frac{L^b_y}{N}\|Y^h_h Y^\ell_h\|_{\mathcal{H}^2(\mathbb{R})} \le C\|u'_h\|_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|_{\mathcal{H}^4(\mathbb{R})}\frac{(L^b_y)^2}{N^2},\\
\|(\partial^2_{y_\ell y_\ell} b_i)(\cdot) Y^h_\ell Y^\ell_\ell\|_{\mathcal{H}^2(\mathbb{R})} &\le \frac{L^b_y}{N}\|Y^h_\ell Y^\ell_\ell\|_{\mathcal{H}^2(\mathbb{R})} \le C\|u'_h\|_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|_{\mathcal{H}^4(\mathbb{R})}\frac{(L^b_y)^2}{N^2}.
\end{aligned} \tag{3.113}$$

Moreover, to analyze the first two terms in the second line of (3.112), by (3.108),

$$\begin{aligned}
&\left\|\sum_{k\in[N]\setminus\{\ell,h\}} (\partial^2_{y_h y_k} b_i)(\cdot) Y^h_h Y^\ell_k\right\|_{\mathcal{H}^2(\mathbb{R})} + \left\|\sum_{j\in[N]\setminus\{h,\ell\}} (\partial^2_{y_j y_\ell} b_i)(\cdot) Y^h_j Y^\ell_\ell\right\|_{\mathcal{H}^2(\mathbb{R})}\\
&\le \frac{L^b_y}{N^2}\sum_{k\in[N]\setminus\{\ell,h\}} \left\|Y^h_h Y^\ell_k\right\|_{\mathcal{H}^2(\mathbb{R})} + \frac{L^b_y}{N^2}\sum_{j\in[N]\setminus\{h,\ell\}} \left\|Y^h_j Y^\ell_\ell\right\|_{\mathcal{H}^2(\mathbb{R})} \le C\|u'_h\|_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|_{\mathcal{H}^4(\mathbb{R})}\frac{L^b_y}{N^2}.
\end{aligned} \tag{3.114}$$

Furthermore, to analyze the last term in (3.112), as $h \neq \ell$,

$$\begin{aligned}
&\sum_{j\in[N]\setminus\{h\}}\sum_{k\in[N]\setminus\{\ell\}} (\partial^2_{y_j y_k} b_i)(\cdot) Y^h_j Y^\ell_k\\
&= \sum_{k\in[N]\setminus\{\ell\}} (\partial^2_{y_\ell y_k} b_i)(\cdot) Y^h_\ell Y^\ell_k + \sum_{j\in[N]\setminus\{h,\ell\}} \left((\partial^2_{y_j y_j} b_i)(\cdot) Y^h_j Y^\ell_j + \sum_{k\in[N]\setminus\{\ell,j\}} (\partial^2_{y_j y_k} b_i)(\cdot) Y^h_j Y^\ell_k\right),
\end{aligned} \tag{3.115}$$

where the first and second terms can be estimated by

$$\begin{aligned}
&\left\|\sum_{k\in[N]\setminus\{\ell\}} (\partial^2_{y_\ell y_k} b_i)(\cdot) Y^h_\ell Y^\ell_k\right\|_{\mathcal{H}^2(\mathbb{R})} + \left\|\sum_{j\in[N]\setminus\{h,\ell\}} (\partial^2_{y_j y_j} b_i)(\cdot) Y^h_j Y^\ell_j\right\|_{\mathcal{H}^2(\mathbb{R})}\\
&\le \sum_{k\in[N]\setminus\{\ell\}} \frac{L^b_y}{N^2}\left\|Y^h_\ell Y^\ell_k\right\|_{\mathcal{H}^2(\mathbb{R})} + \sum_{j\in[N]\setminus\{h,\ell\}} \frac{L^b_y}{N}\left\|Y^h_j Y^\ell_j\right\|_{\mathcal{H}^2(\mathbb{R})}\\
&\le C\left(N\frac{(L^b_y)^3}{N^4} + N\frac{(L^b_y)^3}{N^3}\right)\|u'_h\|_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|_{\mathcal{H}^4(\mathbb{R})} \le C\frac{(L^b_y)^2}{N^2}\|u'_h\|_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|_{\mathcal{H}^4(\mathbb{R})},
\end{aligned} \tag{3.116}$$

and the last term can be estimated by

$$\begin{aligned}
&\left\|\sum_{j\in[N]\setminus\{h,\ell\}}\sum_{k\in[N]\setminus\{\ell,j\}} (\partial^2_{y_j y_k} b_i)(\cdot) Y^h_j Y^\ell_k\right\|_{\mathcal{H}^2(\mathbb{R})} \le \sum_{j\in[N]\setminus\{h,\ell\}}\sum_{k\in[N]\setminus\{\ell,j\}} \frac{L^b_y}{N^2}\left\|Y^h_j Y^\ell_k\right\|_{\mathcal{H}^2(\mathbb{R})}\\
&\le N^2\frac{L^b_y}{N^2}\frac{C(L^b_y)^2\|u'_h\|_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|_{\mathcal{H}^4(\mathbb{R})}}{N^2} = C\frac{(L^b_y)^2}{N^2}\|u'_h\|_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|_{\mathcal{H}^4(\mathbb{R})}.
\end{aligned} \tag{3.117}$$

Hence combining (3.112), (3.113), (3.114) and (3.115) yields

$$\Big\| \sum_{j,k=1}^{N} (\partial^2_{y_j y_k} b_i)(\cdot, X_{\cdot,i}, \mathbf{X}_{\cdot}) Y^h_j Y^\ell_k \Big\|_{\mathcal{H}^2(\mathbb{R})} \le C\|u'_h\|_{\mathcal{H}^4(\mathbb{R})} \cdot \|u''_\ell\|_{\mathcal{H}^4(\mathbb{R})} \frac{L^b_y}{N^2}.$$

This along with (3.110) and (3.111) yield the desired estimate. □

Proof of Proposition 3.7.5. To simplify the notation, we write $\mathbf{X} = \mathbf{X}^{\boldsymbol{u}}$, $\mathbf{Y}^h = \mathbf{Y}^{\boldsymbol{u},u'_h}$, $\mathbf{Y}^\ell = \mathbf{Y}^{\boldsymbol{u},u''_\ell}$, $\mathbf{Z}^{h,\ell} = \mathbf{Z}^{\boldsymbol{u},u'_h,u''_\ell}$ and $\mathfrak{f}^{h,\ell} = \mathfrak{f}^{\boldsymbol{u},u'_h,u''_\ell}$. Applying Lemma 3.8.2 with $\mathbf{S} = \mathbf{Z}^{\ell,h}$, $B_i(t) = \partial_x b_i(t, X_{t,i}, \mathbf{X}_t)$, $\bar{B}_{i,j}(t) = \partial_{y_j} b_i(t, X_{t,i}, \mathbf{X}_t)$ and $f_{t,i} = \mathfrak{f}^{h,\ell}_{t,i}$ yields that for all $i \in [N]$,

$$\sup_{t\in[0,T]} \mathbb{E}[|Z^{h,\ell}_{t,i}|^2] \le 2T \left(\left\|\mathfrak{f}^{h,\ell}_i\right\|^2_{\mathcal{H}^2(\mathbb{R})} + \left\| \sum_{k=1}^{N} |\mathfrak{f}^{h,\ell}_k| \right\|^2_{\mathcal{H}^2(\mathbb{R})} \frac{(L^b_y)^2}{N^2} T^2 e^{2(L^b+L^b_y)T} \right) e^{2L^bT}. \quad (3.118)$$

By Lemma 3.8.3, one can get $|\mathfrak{f}^{h,\ell}_i\|_{\mathcal{H}^2(\mathbb{R})} \le C\|u'_h\|_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|_{\mathcal{H}^4(\mathbb{R})} L^b_y \left((\delta_{h,i} + \delta_{\ell,i})\frac{1}{N} + \frac{1}{N^2}\right)$, where $C \ge 0$ is a constant, which depends on the upper bounds of T, L^b, L^b_y. Moreover,

$$\begin{aligned} \left\| \sum_{k=1}^{N} |\mathfrak{f}^{h,\ell}_k| \right\|_{\mathcal{H}^2(\mathbb{R})} &\le \sum_{k=1}^{N} \left\|\mathfrak{f}^{h,\ell}_k\right\|_{\mathcal{H}^2(\mathbb{R})} = \sum_{k\in\{h,\ell\}} \left\|\mathfrak{f}^{h,\ell}_k\right\|_{\mathcal{H}^2(\mathbb{R})} + \sum_{k\in[N]\setminus\{h,\ell\}} \left\|\mathfrak{f}^{h,\ell}_k\right\|_{\mathcal{H}^2(\mathbb{R})} \\ &\le C\left(\frac{1}{N} + (N-2)\frac{1}{N^2}\right) \|u'_h\|_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|_{\mathcal{H}^4(\mathbb{R})} L^b_y \le \frac{C\|u'_h\|_{\mathcal{H}^4(\mathbb{R})}\|u''_\ell\|_{\mathcal{H}^4(\mathbb{R})} L^b_y}{N}. \end{aligned}$$

Summarizing the above estimates yields the desired conclusion. □

3.8.2 Proofs of Propositions 3.7.7, 3.7.8 and 3.7.9

Proof of Proposition 3.7.7. To simplify the notation, we write $\mathbf{X} = \mathbf{X}^\phi$ and $\mathbf{Y}^h = \mathbf{Y}^{\phi,\phi'_h}$. Applying Lemma 3.8.2 with $\mathbf{S} = \mathbf{Y}^h$, $B_i(t) = (\partial_x(b_i + \phi_i))(t, X_{t,i}, \mathbf{X}_t)$, $\bar{B}_{i,j}(t) = (\partial_{y_j}(b_i + \phi_i))(t, X_{t,i}, \mathbf{X}_t)$ and $f_{t,i} = \delta_{h,i}\phi'_h(t, X_{t,i}, \mathbf{X}_t)$ yields that for all $i \in I_N$,

$$\sup_{t\in[0,T]} \mathbb{E}[|Y^h_{t,i}|^p] \le (2T)^{p-1} \left(\|f_i\|^p_{\mathcal{H}^p(\mathbb{R})} + \left\| \sum_{k=1}^{N} |f_k| \right\|^p_{\mathcal{H}^p(\mathbb{R})} \frac{(L^{b+\phi}_y)^p}{N^p} T^p e^{p(L^{b+\phi}+L^{b+\phi}_y)T} \right) e^{pL^{b+\phi}T}. \quad (3.119)$$

where we used $\|\bar{B}_{i,j}\|_{L^\infty} \le L^{b+\phi}/N$. Note that using the Cauchy-Schwarz inequality, Fubini's theorem, and the definition of $f_{t,i}$,

$$\|f_i\|^p_{\mathcal{H}^p(\mathbb{R})} \le T \sup_{t\in[0,T]} \mathbb{E}\left[|f_{t,i}|^p\right] \le \delta_{h,i}(2T)^{p-1} \left((L^{\phi'_h})^p \mathbb{E}[(1 + |X_{t,i}|)^p] + \left(\frac{L^{\phi'_h}_y}{N}\right)^p \left(\sum_{k=1}^{N} |X_{t,k}|\right)^p \right).$$

Since $(\sum_{k=1}^N a_k)^p \le N^{p-1}\sum_{k=1}^N a_k^p$ for all $(a_k)_{k=1}^N \in [0,\infty)$, by Proposition 3.7.6,

$$\begin{aligned}\|f_i\|^p_{\mathcal{H}^p(\mathbb{R})} &\le \delta_{h,i}(2T)^{p-1}\left(2^{p-1}(L^{\phi'_h})^p\mathbb{E}[1+|X_{t,i}|^p] + \left(\frac{L_y^{\phi'_h}}{N}\right)^p\left(\sum_{k=1}^N |X_{t,k}|\right)^p\right)\\ &\le \delta_{h,i}(2T)^{p-1}\left(2^{p-1}(L^{\phi'_h})^p(1+C_X^{h,p}) + \left(L_y^{\phi'_h}\right)^p \frac{1}{N}\sum_{k=1}^N C_X^{k,p}\right).\end{aligned} \tag{3.120}$$

Similarly, using the definition of $f_{t,k}$,

$$\left\|\sum_{k=1}^N |f_k|\right\|^p_{\mathcal{H}^p(\mathbb{R})} \le \|f_h\|^p_{\mathcal{H}^p(\mathbb{R})} \le (2T)^{p-1}\left(2^{p-1}(L^{\phi'_h})^p(1+C_X^{h,p}) + \left(L_y^{\phi'_h}\right)^p \frac{1}{N}\sum_{k=1}^N C_X^{k,p}\right). \tag{3.121}$$

Combining (3.119), (3.120) and (3.121) yields the desired estimate. □

To prove Proposition 3.7.8, we estimate the process $\mathfrak{f}^{\phi,\phi'_h,\phi''_\ell}$ defined in (3.39) .

Lemma 3.8.4. *Suppose Assumption 3.3.2 holds and that $\xi_i \in L^4(\Omega;\mathbb{R})$ for all $i \in I_N$. For all $\phi \in \Pi^N$, $h, \ell \in I_N$ with $h \neq \ell$ and $\phi'_h, \phi''_\ell \in \Pi$, the process $\mathfrak{f}^{\phi,\phi'_h,\phi''_\ell}$ defined in* (3.39) *satisfies for all $i \in I_N$,*

$$\|\mathfrak{f}_i^{\phi,\phi'_h,\phi''_\ell}\|_{\mathcal{H}^2(\mathbb{R})} \le C\left((\delta_{h,i}+\delta_{\ell,i})\frac{1}{N} + \frac{1}{N^2}\right)\max\{L_y^{b+\phi}, L_y^{\phi'_h}, L_y^{\phi''_\ell}\},$$

where $C \ge 0$ is a constant depending only on the upper bounds of T, $\max_{i\in I_N}\mathbb{E}[|\xi_i|^4]$, $\max_{i\in I_N}\|\sigma_i\|_{L^4}$, $L^{b+\phi}$, $L^{\phi'_h}$, $L^{\phi''_\ell}$, $L_y^{b+\phi}$, $L_y^{\phi'_h}$ and $L_y^{\phi''_\ell}$.

Proof. To simplify the notation, we write $\mathbf{X} = \mathbf{X}^\phi$, $\mathbf{Y}^h = \mathbf{Y}^{\phi,\phi'_h}$ and $\mathbf{Y}^\ell = \mathbf{Y}^{\phi,\phi''_\ell}$. Observe that by Proposition 3.7.7, for all $i, j \in I_N$,

$$\begin{aligned}&\left\|Y_i^h Y_j^\ell\right\|^2_{\mathcal{H}^2(\mathbb{R})} \le T\sup_{t\in[0,T]}\mathbb{E}[|Y_{t,i}^h Y_{t,j}^\ell|^2] \le T\sup_{t\in[0,T]}\mathbb{E}[|Y_{t,i}^h|^4]^{\frac{1}{2}}\mathbb{E}[|Y_{t,j}^\ell|^4]^{\frac{1}{2}}\\ &\le T\left(\delta_{h,i}C_Y^{h,4} + \frac{1}{N^4}\bar{C}_Y^{h,4}\right)^{\frac{1}{2}}\left(\delta_{\ell,j}C_Y^{\ell,4} + \frac{1}{N^4}\bar{C}_Y^{\ell,4}\right)^{\frac{1}{2}}\\ &\le T\bigg(\delta_{h,i}\delta_{\ell,j}(C_Y^{h,4}C_Y^{\ell,4})^{\frac{1}{2}} + \frac{1}{N^2}\left(\delta_{h,i}(C_Y^{h,4}\bar{C}_Y^{\ell,4})^{\frac{1}{2}} + \delta_{\ell,j}(C_Y^{\ell,4}\bar{C}_Y^{h,4})^{\frac{1}{2}}\right) + \frac{1}{N^4}(\bar{C}_Y^{h,4}\bar{C}_Y^{\ell,4})^{\frac{1}{2}}\bigg).\end{aligned} \tag{3.122}$$

In the sequel, we denote by $C \ge 0$ a generic constant, which depends on the upper bounds of T, $\max_{i\in I_N}(\mathbb{E}[|\xi_i|^4] + \|\sigma_i\|_{L^4})$, $L^{b+\phi}$, $L_y^{b+\phi}$, $L^{\phi'_h}$, $L_y^{\phi'_h}$, $L^{\phi''_\ell}$ and $L_y^{\phi''_\ell}$, and may take a different value at each occurrence.

We now bound each term in (3.39). Observe that for all $t \in [0,T]$,

$$\begin{aligned} &\begin{pmatrix} Y^h_{t,i} \\ \mathbf{Y}^h_t \end{pmatrix}^\top \begin{pmatrix} \partial^2_{xx}(b_i+\phi_i) & \partial^2_{xy}(b_i+\phi_i) \\ \partial^2_{yx}(b_i+\phi_i) & \partial^2_{yy}(b_i+\phi_i) \end{pmatrix} (t, X_{t,i}, \mathbf{X}_t) \begin{pmatrix} Y^\ell_{t,i} \\ \mathbf{Y}^\ell_t \end{pmatrix} = (\partial^2_{xx}(b_i+\phi_i))(t, X_{t,i}, \mathbf{X}_t) Y^h_{t,i} Y^\ell_{t,i} \\ &\quad + \sum_{j=1}^N (\partial^2_{xy_j}(b_i+\phi_i))(t, X_{t,i}, \mathbf{X}_t)(Y^h_{t,i}Y^\ell_{t,j} + Y^\ell_{t,i}Y^h_{t,j}) + \sum_{j,k=1}^N (\partial^2_{y_j y_k}(b_i+\phi_i))(t, X_{t,i}, \mathbf{X}_t) Y^h_{t,j} Y^\ell_{t,k}. \end{aligned} \tag{3.123}$$

The upper bound of (3.123) can be obtained by similar arguments as for (3.109), with (3.108) replaced by (3.122), b_i replaced by $b_i + \phi_i$, and Proposition 3.7.4 replaced by Proposition 3.7.7. Hence it holds that

$$\left\| \begin{pmatrix} Y^h_i \\ \mathbf{Y}^h \end{pmatrix}^\top \begin{pmatrix} \partial^2_{xx}(b_i+\phi_i) & \partial^2_{xy}(b_i+\phi_i) \\ \partial^2_{yx}(b_i+\phi_i) & \partial^2_{yy}(b_i+\phi_i) \end{pmatrix} (\cdot, X_i, \mathbf{X}) \begin{pmatrix} Y^\ell_i \\ \mathbf{Y}^\ell \end{pmatrix} \right\|^2_{\mathcal{H}^2(\mathbb{R})} \leq C(L^{b+\phi}_y)^2 \left(\frac{\delta_{h,i} + \delta_{\ell,i}}{N^2} + \frac{1}{N^4} \right). \tag{3.124}$$

Moreover, as $\sum_{j=1}^N (\partial_{y_j}\phi'_h)(\cdot, X_i, \mathbf{X}) Y^\ell_j = (\partial_{y_\ell}\phi'_h)(\cdot, X_i, \mathbf{X}) Y^\ell_\ell + \sum_{j\neq\ell} (\partial_{y_j}\phi'_h)(\cdot, X_i, \mathbf{X}) Y^\ell_j$, using the upper bounds of $\partial_x \phi'_h$ and $\partial_{y_j}\phi'_h$, the identity $\delta_{h,i}\|Y^\ell_i\|^2_{\mathcal{H}^2(\mathbb{R})} = \delta_{h,i}\|Y^\ell_h\|^2_{\mathcal{H}^2(\mathbb{R})}$ and Proposition 3.7.7,

$$\begin{aligned} &\delta_{h,i} \left\| (\partial_x \phi'_h)(\cdot, X_i, \mathbf{X}) Y^\ell_i + \sum_{j=1}^N (\partial_{y_j}\phi'_h)(\cdot, X_i, \mathbf{X}) Y^\ell_j \right\|^2_{\mathcal{H}^2(\mathbb{R})} \\ &\leq \delta_{h,i} 3 \left((L^{\phi'_h})^2 \|Y^\ell_i\|^2_{\mathcal{H}^2(\mathbb{R})} + \frac{(L^{\phi'_h}_y)^2}{N^2} \|Y^\ell_\ell\|^2_{\mathcal{H}^2(\mathbb{R})} + \left\| \sum_{j\neq\ell} (\partial_{y_j}\phi'_h)(\cdot, X_i, \mathbf{X}) Y^\ell_j \right\|^2_{\mathcal{H}^2(\mathbb{R})} \right) \\ &\leq \delta_{h,i} 3 \left((L^{\phi'_h})^2 \|Y^\ell_h\|^2_{\mathcal{H}^2(\mathbb{R})} + \frac{(L^{\phi'_h}_y)^2}{N^2} \|Y^\ell_\ell\|^2_{\mathcal{H}^2(\mathbb{R})} + (N-1)\frac{(L^{\phi'_h}_y)^2}{N^2} \sum_{j\neq\ell} \|Y^\ell_j\|^2_{\mathcal{H}^2(\mathbb{R})} \right) \\ &\leq \delta_{h,i} 3T \left((L^{\phi'_h})^2 \frac{1}{N^2} \bar{C}^{\ell,2}_Y + \frac{(L^{\phi'_h}_y)^2}{N^2} \left(C^{\ell,2}_Y + \frac{1}{N^2}\bar{C}^{\ell,2}_Y \right) + (N-1)\frac{(L^{\phi'_h}_y)^2}{N^2} \sum_{j\neq\ell} \frac{1}{N^2} \bar{C}^{\ell,2}_Y \right) \\ &\leq \delta_{h,i} \frac{3T}{N^2} \left((L^{\phi'_h})^2 \bar{C}^{\ell,2}_Y + (L^{\phi'_h}_y)^2 (C^{\ell,2}_Y + \bar{C}^{\ell,2}_Y) \right), \end{aligned}$$

where the last inequality used $(1 + (N-1)^2)/N^2 \leq 1$ for all $N \geq 1$. Hence

$$\delta_{h,i} \left\| (\partial_x \phi'_h)(\cdot, X_i, \mathbf{X}) Y^\ell_i + \sum_{j=1}^N (\partial_{y_j}\phi'_h)(\cdot, X_i, \mathbf{X}) Y^\ell_j \right\|^2_{\mathcal{H}^2(\mathbb{R})} \leq \delta_{h,i} \frac{C}{N^2} \left((L^{\phi'_h})^2 (L^{b+\phi}_y)^2 + (L^{\phi'_h}_y)^2 \right), \tag{3.125}$$

Similarly,

$$\delta_{\ell,i}\left\|(\partial_x\phi''_\ell)(\cdot,X_i,\mathbf{X})Y_i^h+\sum_{j=1}^N(\partial_{y_j}\phi''_\ell)(\cdot,X_i,\mathbf{X})Y_j^h\right\|^2_{\mathcal{H}^2(\mathbb{R})}\le\delta_{\ell,i}\frac{C}{N^2}\left((L^{\phi''_\ell})^2(L_y^{b+\phi})^2+(L_y^{\phi''_\ell})^2\right). \tag{3.126}$$

Combining (3.124), (3.125), (3.126) yields the desired estimate. □

Proof of Proposition 3.7.8. To simplify the notation, we write $\mathbf{X}=\mathbf{X}^\phi$, $\mathbf{Y}^h=\mathbf{Y}^{\phi,\phi'_h}$, $\mathbf{Y}^\ell=\mathbf{Y}^{\phi,\phi''_\ell}$, $\mathbf{Z}^{h,\ell}=\mathbf{Z}^{\phi,\phi'_h,\phi''_\ell}$ and $\mathfrak{f}^{h,\ell}=\mathfrak{f}^{\phi,\phi'_h,\phi''_\ell}$. Applying Lemma 3.8.2 with $\mathbf{S}=\mathbf{Z}^{\ell,h}$, $B_{t,i}=(\partial_x(b_i+\phi_i))(t,X_{t,i},\mathbf{X}_t)$, $\bar{B}_{i,j}=(\partial_{y_j}(b_i+\phi_i))(t,X_{t,i},\mathbf{X}_t)$ and $f_{t,i}=\mathfrak{f}^{h,\ell}_{t,i}$ yields that for all $i\in I_N$,

$$\sup_{t\in[0,T]}\mathbb{E}[|Z^{h,\ell}_{t,i}|^2]\le 2T\left(\left\|\mathfrak{f}^{h,\ell}_i\right\|^2_{\mathcal{H}^2(\mathbb{R})}+\left\|\sum_{k=1}^N|\mathfrak{f}^{h,\ell}_k|\right\|^2_{\mathcal{H}^2(\mathbb{R})}\frac{(L_y^{b+\phi})^2}{N^2}T^2e^{2(L^{b+\phi}+L_y^{b+\phi})T}\right)e^{2L^{b+\phi}T}. \tag{3.127}$$

By Lemma 3.8.4, there exists a constant $C\ge 0$, depending on the upper bounds of T, $\max_{i\in I_N}\mathbb{E}[|\xi_i|^4]$, $\max_{i\in I_N}\|\sigma_i\|_{L^4}$, $L^{b+\phi}$, $L^{\phi'_h}$, $L^{\phi''_\ell}$, $L_y^{b+\phi}$, $L_y^{\phi'_h}$ and $L_y^{\phi''_\ell}$, such that for all $i\in I_N$, $\|\mathfrak{f}^{h,\ell}_i\|_{\mathcal{H}^2(\mathbb{R})}\le C\left((\delta_{h,i}+\delta_{\ell,i})\frac{1}{N}+\frac{1}{N^2}\right)\max\{L_y^{b+\phi},L_y^{\phi'_h},L_y^{\phi''_\ell}\}$, which along with $h\neq\ell$ implies that

$$\begin{aligned}\left\|\sum_{k=1}^N|\mathfrak{f}^{h,\ell}_k|\right\|_{\mathcal{H}^2(\mathbb{R})}&\le\sum_{k=1}^N\left\|\mathfrak{f}^{h,\ell}_k\right\|_{\mathcal{H}^2(\mathbb{R})}=\sum_{k\in\{h,\ell\}}\left\|\mathfrak{f}^{h,\ell}_k\right\|_{\mathcal{H}^2(\mathbb{R})}+\sum_{k\in I_N\setminus\{h,\ell\}}\left\|\mathfrak{f}^{h,\ell}_k\right\|_{\mathcal{H}^2(\mathbb{R})}\\&\le C\max\{L_y^{b+\phi},L_y^{\phi'_h},L_y^{\phi''_\ell}\}\left(\frac{1}{N}+(N-2)\frac{1}{N^2}\right)\le\frac{C}{N}\max\{L_y^{b+\phi},L_y^{\phi'_h},L_y^{\phi''_\ell}\}.\end{aligned}$$

Summarizing the above estimates yields the desired conclusion. □

Proof of Proposition 3.7.9 . To simplify the notation, we write $\mathbf{X}=\mathbf{X}^\phi$, $\mathbf{Y}^h=\mathbf{Y}^{\phi,\phi'_h}$, $\mathbf{Y}^\ell=\mathbf{Y}^{\phi,\phi''_\ell}$ and $\mathbf{Z}^{h,\ell}=\mathbf{Z}^{\phi,\phi'_h,\phi''_\ell}$. We denote by $C\ge 0$ a generic constant depending only on the upper bounds of T, $\max_{i\in I_N}\mathbb{E}[|\xi_i|^4]$, $\max_{i\in I_N}\|\sigma_i\|_{L^4}$, $L^{b+\phi}$, $L^{\phi'_h}$, $L^{\phi''_\ell}$, L_y^b, L_y^ϕ, $L_y^{\phi'_h}$ and $L_y^{\phi''_\ell}$ and may take a different value at each occurrence.

Fix $t\in[0,T]$. By Lemma 3.8.1 and Proposition 3.7.6,

$$\mathbb{E}[|u^\phi_{t,i}|^2]^{\frac{1}{2}}\le C\left(L^\phi(1+\mathbb{E}[|X^\phi_{t,i}|^2]^{\frac{1}{2}})+\frac{L_y^\phi}{N}\sum_{j=1}^N\mathbb{E}[|X^\phi_{t,j}|^2]^{\frac{1}{2}}\right)\le C.$$

Moreover, by (3.40) and Lemma 3.8.1,

$$\begin{aligned}\mathbb{E}[|v_{t,i}^{\phi,\phi'_h}|^2] &\le 3\Bigg(\mathbb{E}[|(\partial_x\phi_i)(t,X_{t,i},\mathbf{X}_t)Y_{t,i}^h|^2] + \mathbb{E}\left[\left|\sum_{k=1}^N (\partial_{y_k}\phi_i)(t,X_{t,i},\mathbf{X}_t)Y_{t,k}^h\right|^2\right] \\ &\qquad + \delta_{h,i}\mathbb{E}[|\phi'_h(t,X_{t,i},\mathbf{X}_t)|^2]\Bigg) \\ &\le 3\Bigg\{(L^\phi)^2\mathbb{E}[|Y_{t,i}^h|^2] + 2\frac{(L_y^\phi)^2}{N^2}\Bigg(\mathbb{E}[|Y_{t,h}^h|^2] + (N-1)\sum_{k\neq h}\mathbb{E}\left[\left|Y_{t,k}^h\right|^2\right]\Bigg) \\ &\qquad + \delta_{h,i}\mathbb{E}\left[\left|L^{\phi'_h}(1+|X_{t,i}|) + \frac{L_y^{\phi'_h}}{N}\sum_{k=1}^N |X_{t,k}|\right|^2\right]\Bigg\},\end{aligned} \tag{3.128}$$

which along with Propositions 3.7.6 and 3.7.7 shows that $\mathbb{E}[|v_{t,i}^{\phi,\phi'_h}|^2] \le C\left(\delta_{h,i} + \frac{1}{N^2}(L_y^\phi)^2\right)$. This proves the desired estimate of $v_{t,i}^{\phi,\phi'_h}$.

We then estimate $w_i^{\phi,\phi'_h,\phi''_\ell}$. Similar to (3.124),

$$\left\|\begin{pmatrix} Y_i^h \\ \mathbf{Y}^h\end{pmatrix}^\top \begin{pmatrix} \partial_{xx}^2\phi_i & \partial_{xy}^2\phi_i \\ \partial_{yx}^2\phi_i & \partial_{yy}^2\phi_i\end{pmatrix}(\cdot,X_i,\mathbf{X})\begin{pmatrix} Y_i^\ell \\ \mathbf{Y}^\ell\end{pmatrix}\right\|_{\mathcal{H}^2(\mathbb{R})} \le C\left((\delta_{h,i}+\delta_{\ell,i})\frac{1}{N} + \frac{1}{N^2}\right)\max\{L_y^{b+\phi}, L_y^\phi\}. \tag{3.129}$$

By Proposition 3.7.8,

$$\begin{aligned}&\|(\partial_x\phi_i)(\cdot,X_i,\mathbf{X})Z_i^{h,\ell}\|_{\mathcal{H}^2(\mathbb{R})} + \left\|\sum_{k=1}^N (\partial_{y_k}\phi_i)(\cdot,X_i,\mathbf{X})Z_k^{h,\ell}\right\|_{\mathcal{H}^2(\mathbb{R})} \\ &\le (L^\phi)^2\|Z_i^{h,\ell}\|_{\mathcal{H}^2(\mathbb{R})} + \sum_{k\in\{h,\ell\}}\left\|(\partial_{y_k}\phi_i)(\cdot,X_i,\mathbf{X})Z_k^{h,\ell}\right\|_{\mathcal{H}^2(\mathbb{R})} \\ &\qquad + \sum_{k\in I_N\setminus\{h,\ell\}}\left\|(\partial_{y_k}\phi_i)(\cdot,X_i,\mathbf{X})Z_k^{h,\ell}\right\|_{\mathcal{H}^2(\mathbb{R})} \\ &\le C\max\{L_y^{b+\phi}, L_y^{\phi'_h}, L_y^{\phi''_\ell}\}\left((\delta_{h,i}+\delta_{\ell,i})\frac{1}{N} + \frac{1}{N^2} + \frac{L_y^\phi}{N}\frac{1}{N} + (N-2)\frac{L_y^\phi}{N}\frac{1}{N^2}\right) \\ &\le C\left((\delta_{h,i}+\delta_{\ell,i})\frac{1}{N} + \frac{1}{N^2}\right)\max\{L_y^{b+\phi}, L_y^{\phi'_h}, L_y^{\phi''_\ell}\}.\end{aligned} \tag{3.130}$$

Moreover, by similar arguments as that for (3.128) and using $h \neq \ell$,

$$\begin{aligned}
&\delta_{h,i} \left\| (\partial_x \phi'_h)(\cdot, X_i^\phi, \mathbf{X}^\phi) Y_i^{\phi,\phi''_\ell} + \sum_{k=1}^N (\partial_{y_k} \phi'_h)(\cdot, X_i^\phi, \mathbf{X}^\phi) Y_k^{\phi,\phi''_\ell} \right\|_{\mathcal{H}^2(\mathbb{R})}^2 + \delta_{\ell,i} \left\| (\partial_x \phi''_\ell)(\cdot, X_i^\phi, \mathbf{X}^\phi) Y_i^{\phi,\phi'_h} \right. \\
&\quad \left. + \sum_{k=1}^N (\partial_{y_k} \phi''_\ell)(\cdot, X_i^\phi, \mathbf{X}^\phi) Y_k^{\phi,\phi'_h} \right\|_{\mathcal{H}^2(\mathbb{R})}^2 \leq (\delta_{h,i} + \delta_{\ell,i}) \frac{C}{N^2} \max\{L_y^{b+\phi}, L_y^\phi, L_y^{\phi'_h}\}^2.
\end{aligned} \tag{3.131}$$

Summarizing (3.129), (3.130) and (3.131) and using $L_y^{b+\phi} \leq L_y^b + L_y^\phi$ yield the desired estimate of $w_i^{\phi,\phi'_h,\phi''_\ell}$. □

Chapter 4

Policy Gradient and Policy Optimization methods

4.1 Introduction

Reinforcement Learning (RL) is a powerful framework for solving sequential decision-making problems, where a learning agent interacts with an unknown environment to improve her performance through trial and error [137]. In RL, an agent takes an action and receives a reinforcement signal in terms of a reward, which encodes the outcome of her action. In order to maximize the accumulated reward over time, the agent learns to select her actions based on her past experiences (exploitation) and/or by making new choices (exploration). Exploration and exploitation are the essence of RL, and entropy regularization has shown to be effective to balance the exploration-exploitation in RL, and more importantly to enable fast convergence [40, 75, 110].

Fast convergence and sample efficiency are critical for many applied RL problems, such as financial trading [77] and healthcare treatment recommendations [158], where acquiring new samples is costly or the chance of exploring new actions in the system is limited. In such cases, the cost of making incorrect decisions can be prohibitively high.

Our work. This paper proposes and analyzes two new policy learning methods: regularized policy gradient (RPG) and iterative policy optimization (IPO), for a class of discounted entropy-regularized linear-quadratic control (LQC) problems over an infinite time horizon. Assuming access to the exact policy evaluation, both approaches are shown to converge linearly in finding optimal policies of the regularized LQC (Theorem 4.4.1 and 4.5.1). Moreover, the IPO method can achieve a super-linear convergence rate (on the order of one and a half) once it enters a local region around the optimal policy. Finally, when the optimal policy from an RL problem with a known environment is appropriately transferred as the initial policy

[1] This chapter is mainly based on work [72] entitled *Fast Policy Learning for Linear Quadratic Control with Entropy Regularization*, coauthored with Xin Guo (UC Berkeley) and Renyuan Xu (NYU)

to an RL problem with an unknown environment, the IPO algorithm is shown to enable a super-linear convergence rate if the two environments are sufficiently close (Theorem 4.6.1).

Our analysis approach is inspired by [53] to establish the gradient dominance condition within the linear-quadratic structure. Unlike theirs, our framework incorporates entropy regularization and state transition noise (Section 4.2). Therefore, in contrast to their deterministic and linear policies, our policies are of Gaussian type. Consequently, the gradient dominance condition involves both the gradient of the mean and the gradient of the covariance (Lemma 4.3.2). Accordingly, to establish the convergence of the covariance update in RPG, the smoothness of the objective function for bounded covariance is exploited, which is ensured with proper learning rate (Lemma 4.4.1).

Different from the first-order gradient descent update in most existing literature, our proposed IPO method requires solving an optimization problem at each step. This yields faster (super-linear) local convergence, established by bounding the differences between two discounted state correlation matrices with respect to the change in policy parameters (Lemma 4.5.2 and Theorem 4.5.2). This approach is connected intriguingly with [40], where the bound for the difference between discounted state visitation measures yielded the local quadratic convergence in Markov Decision Processes (MDPs).

Related works of policy gradient methods in LQC. As a cornerstone in optimal control theory, the LQC problem is to find an optimal control in a linear dynamical system with a quadratic cost. LQC is popular due to its analytical tractability and its approximation power to nonlinear problems [13]. Until recently, most works on the LQC problem assumed that the model parameters are fully known. The first global convergence result for the policy gradient method to learn the optimal policy for LQC problems was developed in [53] for an infinite time horizon and with deterministic dynamics. Their work was extended in [19] to give global optimality guarantees of policy gradient methods for a larger class of control problems, which satisfy a closure condition under policy improvement and convexity of policy improvement steps. More progress has been made for policy gradient methods in other settings as well, including [21] for a real-valued matrix function, [22] for a continuous-time setting, [61] for multiplicative noise, [91], [107] for additive noise, and [76] for finite time horizon with an additive noise, [144] and [166] for time-average costs with risk constraints, [80] for nearly-linear dynamic system, [150] for distributional LQC to find the distribution of the return, and [78] for nonlinear stochastic control with exit time. Our work establishes *fast* convergence for both policy gradient based and policy optimization based algorithms for an infinite time horizon LQC with entropy regularization.

Related works of entropy regularization. Entropy regularization has been frequently adopted to encourage exploration and improve convergence [71, 81, 89, 90, 118, 128, 148, 149, 152, 142, 140, 139, 161]. In particular, [2] showed that entropy regularization induces a smoother landscape that allows for the use of larger learning rates, and hence, faster convergence. Convergence rate analysis has been established when the underlying dynamic is an MDP with finite states and finite actions. For instance, [1] and [110] developed convergence guarantees for regularized policy gradient methods, with relative entropy regularization considered in [1] and entropy regularization in [110]. Both papers suggest the role of regularization in guaranteeing *faster* convergence for the tabular setting. For the natural policy gradient method, [40] established a global linear convergence rate and a local quadratic convergence rate.

For system with infinite number of states and actions, the closest to our work in terms of model setup is [149]. Our work replaces their aggregated control setup with controls that are randomly sampled from the policy, which are more realistic in handling real-world systems. The focuses of these two papers are also different: theirs explained the exploitation–exploration trade-off with entropy regularization from a continuous-time stochastic control perspective and provided theoretical support for Gaussian exploration for LQC; while ours is on algorithms design and the convergence analysis. To the best of our knowledge, our work is the first non-asymptotic convergence result for LQC under entropy regularization.

Related works of transfer learning. Transfer learning, *a.k.a.* knowledge transfer, is a technique to utilize external expertise from other domains to benefit the learning process of a new task [28, 29, 116, 151]. It has gained popularity in many areas to improve the efficiency of learning. However, transfer learning in the RL framework is decisively more complicated and remains largely unexplored, as the knowledge to transfer involves a controlled stochastic process [168]. The transfer learning scheme proposed here is the first known theoretical development of transfer learning in the context of RL.

Notations and organization. Throughout the paper, we will denote, for any matrix $Z \in \mathbb{R}^{m\times d}$, $Z^\top$ for the transpose of Z, $\|Z\|$ for the spectral norm of Z, $\|Z\|_F$ for the Frobenius norm of Z, $\mathrm{tr}(Z)$ for the trace of a square matrix Z, and $\sigma_{\min}(Z)$ (*resp.* $\sigma_{\max}(Z)$) for the minimum (*resp.* maximum) singular value of a square matrix Z. Let $\mathbb{S}_+^d$ denote the set of symmetric positive semi-definite matrices in $\mathbb{R}^{d\times d}$ and $\mathbb{S}_{++}^d$ for the subset of $\mathbb{S}_+^d$ consisting of symmetric positive definite matrices. We will adopt $\mathcal{N}(\mu,\Sigma)$ for a Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{S}_+^d$.

The rest of the paper is organized as follows. Section 4.2 introduces the problem and provides its theoretical solution using the dynamic programming principle. Section 4.3 presents the gradient dominance condition and related smoothness property. Section 4.4 introduces the RPG method and provides the global linear convergence result, and Section 4.5 proposes the IPO method, along with its global linear convergence and local super-linear convergence results. Section 4.6 shows that IPO leads to an efficient transfer learning scheme. A model-free version of the policy-based method is discussed in Section 4.7. Numerical examples are presented in Section 4.8.

4.2 Regularized LQC Problem and Solution

4.2.1 Problem Formulation

We consider an entropy-regularized LQC problem over an infinite time horizon with a constant discounted rate.

Randomized policy and entropy regularization. To enable entropy regularization for exploration in the context of learning, we focus on randomized Markovian policies that are stationary. Namely, define the admissible policy set as $\Pi := \{\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})\}$, with $\mathcal{X}$ the state space, $\mathcal{A}$ the action space, and $\mathcal{P}(\mathcal{A})$ the space of probability measures on action space $\mathcal{A}$. Here each admissible policy $\pi \in \Pi$ maps a state $x \in \mathcal{X}$ to a randomized action in $\mathcal{A}$.

For a given admissible policy $\pi \in \Pi$, the corresponding Shannon's entropy is defined as [69, 90]:

$$\mathcal{H}(\pi(\cdot \mid x)) := -\int_{\mathcal{A}} \log \pi(u \mid x) \pi(u \mid x) \mathrm{d}u.$$

The Shannon entropy quantifies the information gain from exploring the unknown environment. We incorporate this entropy term in the objective function as a regularization to encourage collecting information in the unknown environment and performing exploration.

Objective function and dynamics. The decision maker aims to find an optimal policy by minimizing the following objective function

$$\min_{\pi \in \Pi} \mathbb{E}_{x \sim \mathcal{D}}[J_\pi(x)], \tag{4.1}$$

with value function J_π given by

$$J_\pi(x) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(x_t^\top Q x_t + u_t^\top R u_t + \tau \log \pi(u_t|x_t) \right) \middle| x_0 = x \right], \tag{4.2}$$

and such that for $t = 0, 1, 2, \cdots$,

$$x_{t+1} = A x_t + B u_t + w_t, \; x_0 \sim \mathcal{D}. \tag{4.3}$$

Here $x_t \in \mathcal{X} := \mathbb{R}^n$ is the state of the system and the initial state x_0 follows an initial distribution $\mathcal{D}$. Here $u_t \in \mathcal{A} := \mathbb{R}^k$ is the control at time t following a policy π. In addition, $\{w_t\}_{t=0}^{\infty}$ are zero-mean independent and identically distributed (i.i.d) noises. We assume that $\{w_t\}_{t=0}^{\infty}$ have finite second moments. That is, $\mathrm{tr}(W) < \infty$ with $W := \mathbb{E}[w_t w_t^\top]$ for any $t = 0, 1, 2, \cdots$. The matrices $A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times k}$ define the system's (transition) dynamics. $Q \in \mathbb{S}_+^n$ and $R \in \mathbb{S}_{++}^k$ are matrices that parameterize the quadratic costs. $\gamma \in (0, 1)$ denotes the discount factor and $\tau > 0$ denotes the regularization parameter. The expectation in (4.1) is taken with respect to the control $u_t \sim \pi(\cdot|x_t)$ and system noise w_t for $t \geq 0$.

4.2.2 Optimal Value Function and Policy

While the optimal solution to the LQC problem is a well-explored topic, it is worth noting that, to the best of our knowledge, no prior work has presented a solution to the entropy-regularized LQC problem in the form of (4.1). Additionally, in the study of [149] for entropy-regularized LQC with continuous-time state dynamics, they focused on the state transitions with aggregated controls. This differs from the state transitions considered in (4.3), where the controls in the state transitions are randomly sampled from the policy π.

Optimal value function. The optimal value function $J^* : \mathcal{X} \to \mathbb{R}$ is defined as

$$J^*(x) = \min_{\pi\in\Pi} J_\pi(x). \tag{4.4}$$

The following theorem establishes the explicit expression for the optimal control policy and the corresponding optimal value function: the optimal policy is characterized as a multivariate Gaussian distribution, with the mean linear in the state x and a constant covariance matrix.

Theorem 4.2.1 (Optimal value functions and optimal policy)**.** *The optimal policy π^* to* (4.4) *is a Gaussian policy: $\pi^*(\cdot|x) = \mathcal{N}(-K^*x, \Sigma^*), \forall x \in \mathcal{X}$, where*

$$K^* = \gamma(R + \gamma B^\top PB)^{-1}B^\top PA, \quad \Sigma^* = \frac{\tau}{2}(R + \gamma B^\top PB)^{-1}, \tag{4.5}$$

with P and q satisfying

$$P = Q + K^{*\top}RK^* + \gamma(A - BK^*)^\top P(A - BK^*), \tag{4.6a}$$

$$q = \frac{1}{1-\gamma}\Big(\operatorname{tr}(\Sigma^*(R + \gamma B^\top PB)) - \frac{\tau}{2}\left(k + \log\left((2\pi)^k \det \Sigma^*\right)\right) + \gamma \operatorname{tr}(WP)\Big). \tag{4.6b}$$

The optimal value function J^* in (4.4) *can be expressed as $J^*(x) = x^\top Px + q$.*

Proof of Theorem 4.2.1 relies on the following lemma, which establishes the optimal solution for the one-step reward function with entropy regularization in the reward.

Lemma 4.2.1. *For any given matrix $M \in \mathbb{S}^k_+$ and vector $b \in \mathbb{R}^k$, the optimal solution $p^* \in \mathcal{P}(\mathcal{A})$ to the following optimization problem* (P) *is a multivariate Gaussian distribution with covariance $\frac{\tau}{2}M^{-1}$ and mean $-\frac{1}{2}M^{-1}b$:*

$$\begin{aligned} &\min_{p:\mathcal{A}\mapsto[0,\infty)} \mathbb{E}_{u\sim p(\cdot)}\left[u^\top Mu + b^\top u + \tau \log p(u)\right], \\ &\textit{subject to} \int_{\mathcal{A}} p(u)du = 1. \end{aligned} \tag{P}$$

Proof. (of Theorem 4.2.1). By the definition of J^* in (4.4),

$$J^*(x) = \min_{\pi\in\Pi} \mathbb{E}_\pi \Big\{ x^\top Qx + u^\top Ru + \tau \log(\pi(u|x)) + \gamma J^*(Ax + Bu + w) \Big\}, \tag{4.7}$$

where the expectation is taken with respect to $u \sim \pi(\cdot|x)$ and the noise term w, with mean 0 and covariance W. Stipulating

$$J^*(x) = x^\top P x + q \tag{4.8}$$

for $P \in \mathbb{S}_+^n$ and $q \in \mathbb{R}$ and plugging into (4.7), we can obtain the optimal value function with dynamic programming principle [16, 18, 63, 137]:

$$\begin{aligned} J^*(x) = x^\top Q x + \min_\pi \mathbb{E}_\pi \Big\{ & u^\top R u + \tau \log(\pi(u|x)) \\ & + \gamma \left[(Ax + Bu + w)^\top P (Ax + Bu + w) + q \right] \Big\} \\ = x^\top Q x + {} & \gamma \operatorname{Tr}(WP) + \gamma x^\top A^\top P A x + \gamma q \\ & + \min_\pi \mathbb{E}_\pi \Big\{ u^\top (R + \gamma B^\top P B) u + \tau \log(\pi(u|x)) + 2\gamma u^\top B^\top P A x \Big\}. \end{aligned} \tag{4.9}$$

Now apply Lemma 4.2.1 to (4.9) with $M = R + \gamma B^\top P B$ and $b = 2\gamma B^\top P A x$, one can get the optimal policy at state x:

$$\pi^*(\cdot|x) = \mathcal{N}\left(-\gamma (R + \gamma B^\top P B)^{-1} B^\top P A x, \frac{\tau}{2} (R + \gamma B^\top P B)^{-1} \right) = \mathcal{N}\left(-K^* x, \Sigma^* \right), \tag{4.10}$$

where K^*, Σ^* are defined in (4.5).

To derive the associated optimal value function, we first calculate the negative entropy of policy π^* at any state $x \in \mathcal{X}$:

$$\mathbb{E}_{\pi^*}[\log(\pi^*(u|x))] = \int_{\mathcal{A}} \log(\pi^*(u|x)) \pi^*(u|x) du = -\frac{1}{2} \left(k + \log\left((2\pi)^k \det \Sigma^* \right) \right). \tag{4.11}$$

Plug (4.10) and (4.11) into (4.9) to get

$$\begin{aligned} J^*(x) = {} & x^\top \left(Q + K^{*\top} R K^* + \gamma (A - BK^*)^\top P (A - BK^*) \right) x \\ & + \operatorname{Tr}(\Sigma^* R) - \frac{\tau}{2} \left(k + \log\left((2\pi)^k \det \Sigma^* \right) \right) + \gamma \left(\operatorname{Tr}(\Sigma^* B^\top P B) + \operatorname{Tr}(WP) + q \right). \end{aligned}$$

Combining this with (4.8), we obtain the Riccati equation in (4.6a), which according to [17, Proposition 4.4.1] has a unique solution P. This is because (4.6a) takes the same form as the one for the classical LQR problem (without entropy regularization). With the unique solution P, we can then define the unique q as in (4.6b), which finishes the proof. □

4.3 Analysis of Value Function and Policy Gradient

In this section, we analyze the expression of the policy gradient, the gradient dominance condition, and the smoothness property of the value function. These properties are necessary for studying the algorithms proposed in Section 4.4 and Section 4.5.

Throughout the analysis, we assume that there exists $\rho \in (0, \frac{1}{\sqrt{\gamma}})$ satisfying $\|A - BK^*\| \leq \rho$ where K^* is the optimal solution in Theorem 4.2.1. We consider the following domain Ω (*i.e.,* the admissible control set) for (K, Σ): $\Omega = \{K \in \mathbb{R}^{k\times n}, \Sigma \in \mathbb{S}^k_+\}$. For any $x_0 \in \mathcal{X}$ following the initial distribution $\mathcal{D}$, we assume that $\mathbb{E}_{x_0\sim\mathcal{D}}[x_0 x_0^\top]$ exists and $\mu := \sigma_{\min}\left(\mathbb{E}_{x_0\sim\mathcal{D}}[x_0x_0^\top]\right) > 0$. For any $(K, \Sigma) \in \Omega$, define $S_{K,\Sigma}$ as the discounted state correlation matrix, *i.e.,*

$$S_{K,\Sigma} := \mathbb{E}_{\pi_{K,\Sigma}}\left[\sum_{i=0}^{\infty} \gamma^i x_i x_i^\top\right]. \tag{4.12}$$

According to Theorem 4.2.1, the optimal policy of (4.1) is a Gaussian policy with a mean following a linear function of the state and a constant covariance matrix. In the remainder of the paper, we look for a parameterized policy of the form

$$\pi_{K,\Sigma}(\cdot|x) := \mathcal{N}(-Kx, \Sigma), \tag{4.13}$$

for any $x \in \mathcal{X}$. With a slight abuse of notation, we use $J_{K,\Sigma}$ to denote $J_{\pi_{K,\Sigma}}$ and denote the objective in (4.1) as a function of (K, Σ), given by

$$C(K, \Sigma) := \mathbb{E}_{x\sim\mathcal{D}}\left[J_{K,\Sigma}(x)\right]. \tag{4.14}$$

To analyze the dependence of Σ in the objective function (4.14) for any fixed K, we also define $f_K : \mathbb{R}^{k\times k} \to \mathbb{R}$ as

$$f_K(\Sigma) = \frac{\tau}{2(1-\gamma)} \log\det(\Sigma) - \frac{1}{1-\gamma}\operatorname{Tr}\left(\Sigma(R + \gamma B^\top P_K B)\right), \quad \forall\, \Sigma \succ 0. \tag{4.15}$$

By applying the Bellman equation, we can get $J_{K,\Sigma}(x) = x^\top P_K x + q_{K,\Sigma}$ with matrix $P_K \in \mathbb{S}^n_+$ and $q_{K,\Sigma} \in \mathbb{R}$ satisfying

$$\begin{aligned} P_K &= Q + K^\top R K + \gamma(A - BK)^\top P_K (A - BK), \\ q_{K,\Sigma} &= \frac{1}{1-\gamma}\Big(\operatorname{tr}(\Sigma(R + \gamma B^\top P_K B) - \frac{\tau}{2}\left(k + \log\left((2\pi)^k \det\Sigma\right)\right) + \gamma \operatorname{tr}(W P_K)\Big). \end{aligned} \tag{4.16}$$

Note that (4.16) differs from equations (4.6a) in terms of their solutions. In (4.6a), the values of K^* and Σ^* are explicitly defined by $K^* = \gamma(R + \gamma B^\top P B)^{-1} B^\top P A$ and $\Sigma^* = \frac{\tau}{2}(R + \gamma B^\top PB)^{-1}$. By substituting these values into (4.6a), one can obtain the solutions for P and q, which define the optimal value function $J^*(x) = x^\top P x + q$.

Meanwhile, K and Σ in (4.16) can take any admissible policy parameter values in Ω, and the resulting P_K and $q_{K,\Sigma}$ are functions of these policy parameters. The value function $J_{K,\Sigma}(x)$ derived from (4.16) represents the value starting from state x with policy parameters (K, Σ), which may or may not correspond to an optimal policy.

We now provide an explicit form for the gradient of the cost function $C(K, \Sigma)$ with respect to K and Σ. This explicit form will be used to show the gradient dominance condition in Lemma 4.3.2 and also in analyzing the algorithms in Sections 4.4 and 4.5.

Lemma 4.3.1 (Explicit form for the policy gradient)**.** *Take a policy in the form of* (4.13) *with parameter* $(K,\Sigma)\in\Omega$*, then the policy gradient has an explicit form:*

$$\nabla_K C(K,\Sigma) = 2E_K S_{K,\Sigma},\ \nabla_\Sigma C(K,\Sigma) = \frac{1}{1-\gamma}\left(R - \frac{\tau}{2}\Sigma^{-1} + \gamma B^\top P_K B\right), \tag{4.17}$$

where $E_K := -\gamma B^\top P_K(A-BK) + RK$ *and* $S_{K,\Sigma}$ *is defined in* (4.12)*.*

Gradient dominance. To prove the global convergence of policy gradient methods, the key idea is to show the gradient dominance condition, which states that $C(K,\Sigma)-C(K^*,\Sigma^*)$ can be bounded by $\|\nabla_K C(K,\Sigma)\|_F^2$ and $\|\nabla_\Sigma C(K,\Sigma)\|_F^2$ for any $(K,\Sigma)\in\Omega$. This suggests that when the gradient norms are sufficiently small, the cost function of the given policy is sufficiently close to the optimal cost function.

Lemma 4.3.2 (Gradient dominance of $C(K,\Sigma)$)**.** *Let* π^* *be the optimal policy with parameters* K^*,Σ^**. Suppose policy* π *with parameter* $(K,\Sigma)\in\Omega$ *satisfying* $\Sigma\preceq I$ *has a finite expected cost, i.e.,* $C(K,\Sigma)<\infty$*. Then*

$$C(K,\Sigma) - C(K^*,\Sigma^*) \le \frac{\|S_{K^*,\Sigma^*}\|}{4\mu^2\sigma_{\min}(R)}\|\nabla_K C(K,\Sigma)\|_F^2 + \frac{1-\gamma}{\sigma_{\min}(R)}\|\nabla_\Sigma C(K,\Sigma)\|_F^2. \tag{4.18}$$

For a lower bound, with E_K *defined in Lemma 4.3.1,*

$$C(K,\Sigma) - C(K^*,\Sigma^*) \ge \frac{\mu}{\|R+\gamma B^\top P_K B\|}\operatorname{Tr}(E_K^\top E_K).$$

"Almost" smoothness. Next, we will develop a smoothness property for the cost objective $C(K,\Sigma)$, which is necessary for establishing the convergence algorithms proposed in Section 4.4 and Section 4.5.

A function $f:\mathbb{R}^n\to\mathbb{R}$ is considered *smooth* if the following condition is satisfied: $\left|f(x)-f(y)+\nabla f(x)^\top(y-x)\right| \le \frac{m}{2}\|x-y\|^2,\quad \forall x,y\in\mathbb{R}^n$, with m a finite constant [53, 76] . In general, characterizing the smoothness of $C(K,\Sigma)$ is challenging, as it may become unbounded when the eigenvalues of $A-BK$ exceed $\frac{1}{\sqrt{\gamma}}$ or when $\sigma_{\min}(\Sigma)$ is close to 0. Nevertheless, in Lemma 4.3.3, we will see that if $C(K,\Sigma)$ is "almost" smooth, then the difference $C(K,\Sigma)-C(K',\Sigma')$ can be bounded by the sum of linear and quadratic terms involving $K-K'$ and $\Sigma-\Sigma'$.

Lemma 4.3.3 ("Almost" smoothness of $C(K,\Sigma)$)**.** *Fix* $0<a\le 1$ *and define* $M_a = \frac{\tau(-\log(a)+a-1)}{2(1-\gamma)(a-1)^2}$*. For any* K,Σ *and* K',Σ' *satisfying* $aI\preceq\Sigma\preceq I$ *and* $aI\preceq\Sigma'\preceq I$*,*

$$\begin{aligned}
C(K',\Sigma') - C(K,\Sigma) &= \operatorname{Tr}\left(S_{K',\Sigma'}(K'-K)^\top(R+\gamma B^\top P_K B)(K'-K)\right)\\
&\quad + 2\operatorname{Tr}\left(S_{K',\Sigma'}(K'-K)^\top E_K\right) + f_K(\Sigma) - f_K(\Sigma')\\
&\le \operatorname{Tr}\left(S_{K',\Sigma'}(K'-K)^\top(R+\gamma B^\top P_K B)(K'-K)\right) + 2\operatorname{Tr}\left(S_{K',\Sigma'}(K'-K)^\top E_K\right)\\
&\quad + \frac{1}{1-\gamma}\operatorname{tr}\left(\left(R+\gamma B^\top P_K B - \frac{\tau}{2}\Sigma^{-1}\right)(\Sigma'-\Sigma)\right) + M_a\operatorname{tr}\left((\Sigma^{-1}\Sigma'-I)^2\right),
\end{aligned}$$

where f_K *is defined in* (4.15)*.*

4.4 Regularized Policy Gradient Method

In this section, we propose a new regularized policy gradient (RPG) update for the parameters K and Σ:

$$\begin{aligned} K^{(t+1)} &= K^{(t)} - \eta_1 \nabla_K C\left(K^{(t)}, \Sigma^{(t)}\right) \cdot \left(S_{K^{(t)},\Sigma^{(t)}}\right)^{-1}, \\ \Sigma^{(t+1)} &= \Sigma^{(t)} - \eta_2 \Sigma^{(t)} \nabla_\Sigma C\left(K^{(t)}, \Sigma^{(t)}\right) \Sigma^{(t)}. \end{aligned}$$

RPG takes into account the inherent structure of the parameter space, which can accelerate convergence. By the explicit expressions of $\nabla_K C(K,\Sigma)$ and $\nabla_\Sigma C(K,\Sigma)$ in (4.17), the above update is equivalent to

$$\begin{aligned} K^{(t+1)} &= K^{(t)} - 2\eta_1 E_{K^{(t)}}, \\ \Sigma^{(t+1)} &= \Sigma^{(t)} - \frac{\eta_2}{1-\gamma} \Sigma^{(t)} \left(R + \gamma B^\top P_{K^{(t)}} B - \frac{\tau}{2}\left(\Sigma^{(t)}\right)^{-1}\right) \Sigma^{(t)}. \end{aligned} \tag{RPG}$$

From (RPG), one can see that the update of parameter K does not depend on the covariance matrix Σ. However, the update of Σ does depend on K through P_K.

Remark 4.4.1 (Comparison to natural policy gradient). *Assume that the covariance matrix Σ is parameterized as scalar multiplication of an identity matrix, i.e., $\Sigma = sI$ for some $s > 0$ and $\pi_{K,s}(x,u) = \mathcal{N}(Kx, sI)$. Then the natural policy gradient follows the update [93]:*

$$K' \;= K - \eta G_K^{-1} \nabla C(K, sI), \;\; s' = s - \eta G_s^{-1} \partial_s C(K, sI), \tag{4.19}$$

where G_K and G_s are the Fischer information matrices under policy $\pi_{K,s}$, i.e.,

$$\begin{aligned} G_K &= \mathbb{E}\left[\sum_{t=0}^{\infty} \nabla \log \pi_{K,s}(u_t|x_t) \nabla \log \pi_{K,s}(u_t|x_t)^\top\right], \\ G_s &= \mathbb{E}\left[\sum_{t=0}^{\infty} \partial_s \log \pi_{K,s}(u_t|x_t) \partial_s \log \pi_{K,s}(u_t|x_t)^\top\right]. \end{aligned}$$

When the covariance matrix of the Gaussian policy takes a diagonal form as in $\pi_{K,s}$, (4.19) *are equivalent to*

$$K' = K - \eta \nabla C(K, sI)(S_{K,sI})^{-1}, \quad s' = s - \tilde{\eta}\; \partial_s C(K, sI) s^2, \tag{4.20}$$

for some constant $\tilde{\eta} > 0$.

Even though (RPG) *is similar to* (4.20)*, there is no corresponding Fisher information form associated with* (RPG) *because of the additional step that simultaneously updates the covariance matrix Σ, which may not necessarily be diagonal.*

We next show that (RPG) achieves a linear convergence rate. The covariance matrices $\left\{\Sigma^{(t)}\right\}_{t=0}^{\infty}$ using (RPG) remain bounded, provided that the initial covariance matrix $\Sigma^{(0)}$ is appropriately bounded.

Theorem 4.4.1 (Global convergence of (RPG))**.** *Given $\tau \in (0, 2\sigma_{\min}(R)]$, take $(K^{(0)}, \Sigma^{(0)}) \in \Omega$ such that $\Sigma^{(0)} \preceq I$. Define $M_\tau := \frac{\tau k}{2(1-\gamma)} \log\left(\frac{\sigma_{\min}(R)}{\pi\tau}\right)$ and*

$$r_0 := \max\left\{\frac{2}{\tau\sigma_{\min}\left(\Sigma^{(0)}\right)},\ \|R\| + \gamma\frac{\|B^\top B\|\left(C(K^{(0)}, \Sigma^{(0)}) - M_\tau\right)}{\mu + \frac{\gamma}{1-\gamma}\sigma_{\min}(W)}\right\}.$$

Then for $\eta_1, \eta_2 = \frac{1}{2r_0}, \frac{\tau(1-\gamma)}{2r_0^2}$ and for $N \geq \max\left\{\frac{\|S_{K^*,\Sigma^*}\| r_0}{\mu\sigma_{\min}(R)}, \frac{8r_0^3}{\tau^2\sigma_{\min}(R)}\right\} \log\frac{C(K^{(0)},\Sigma^{(0)}) - C(K^*,\Sigma^*)}{\varepsilon}$, the regularized policy gradient descent (RPG) *has the following performance bound:*

$$C\left(K^{(N)}, \Sigma^{(N)}\right) - C(K^*, \Sigma^*) \leq \varepsilon.$$

Remark 4.4.2. *Theorem 4.4.1 shows that in order to achieve an ϵ-optimal value function, the number of iterations required is at least $\mathcal{O}\left(\frac{1}{\tau^5}\right)$. Thus, the larger the value of regularization τ, the smaller the number of iterations, and the faster the convergence.*

Remark 4.4.3. *In the RPG update, we only need an upper bound and a lower bound for Σ, namely, $aI \preceq \Sigma \preceq bI$ for some $0 < a < b$. Different choices of b may lead to different (admissible) ranges for τ. For ease of exposition, we set $b = 1$ in Theorem 4.4.1, and the results can be easily extended to the general case of any $b \geq a > 0$.*

To prove Theorem 4.4.1, we will first need the boundedness of the one-step update of the covariance Σ, in order to guarantee the well-definedness of the cost function along the trajectory when performing (RPG) (Lemma 4.4.1). Additionally, we will bound the one-step update of (RPG) (Lemma 4.4.2), and provide an upper bound of $\|P_K\|$ in terms of the objective function $C(K, \Sigma)$ (Lemma 4.4.3).

Lemma 4.4.1 (Boundedness of the update of Σ in (RPG))**.** *Let $(K, \Sigma) \in \Omega$ be given such that $0 \prec \Sigma \preceq I$. Fix $\tau \in (0,\ 2\sigma_{\min}(R)]$, $a \in \left(0,\ \min\left\{\frac{\tau}{2\|R+\gamma B^\top P_K B\|}, \sigma_{\min}(\Sigma)\right\}\right)$. Let K', Σ' be the one-step update of K, Σ using* (RPG) *with $\eta_2 \leq \frac{2(1-\gamma)a^2}{\tau}$. Then $aI \preceq \Sigma' \preceq I$.*

Lemma 4.4.2 (Contraction of (RPG))**.** *Let $(K, \Sigma) \in \Omega$ be given such that $0 \prec \Sigma \preceq I$. Assume $\tau \in (0, 2\sigma_{\min}(R)]$. Fix $a \in \left(0,\ \min\left\{\frac{\tau}{2\|R+\gamma B^\top P_K B\|}, \sigma_{\min}(\Sigma)\right\}\right)$. Let K', Σ' be the one-step update of K, Σ using* (RPG) *with $\eta_1 \leq \frac{1}{2\|R+\gamma B^\top P_K B\|}, \eta_2 \leq \frac{2(1-\gamma)a^2}{\tau}$. Then $aI \preceq \Sigma' \preceq I$ and*

$$C(K', \Sigma') - C(K^*, \Sigma^*) \leq (1 - \zeta)(C(K, \Sigma) - C(K^*, \Sigma^*)),$$

with $0 < \zeta = \min\left\{\frac{2\mu\eta_1\sigma_{\min}(R)}{\|S_{K^*,\Sigma^*}\|}, \frac{\eta_2 a\sigma_{\min}(R)}{2(1-\gamma)}\right\} < 1$.

Lemma 4.4.3 (Lower bound for $C(K, \Sigma)$)**.** *Let M_τ be defined in the same way as in Theorem 4.4.1. Then for all $(K, \Sigma) \in \Omega$, $C(K, \Sigma) \geq \left(\mu + \frac{\gamma}{1-\gamma}\sigma_{\min}(W)\right)\|P_K\| + M_\tau$.*

Proof. (of Theorem 4.4.1). Using Lemma 4.4.3 for any $t \geq 0$,

$$\frac{1}{\|R + \gamma B^\top P_{K^{(t)}} B\|} \geq \frac{1}{\|R\| + \gamma \|B^\top B\| \, \|P_{K^{(t)}}\|} \geq \frac{1}{\|R\| + \gamma \frac{\|B^\top B\| \left(C(K^{(t)}, \Sigma^{(t)}) - M_\tau \right)}{\mu + \frac{\gamma}{1-\gamma} \sigma_{\min}(W)}}. \tag{4.21}$$

Let $a = \frac{\tau}{2r_0}$, $\zeta = \min \left\{ \frac{2\mu\eta_1\sigma_{\min}(R)}{\|S_{K^*,\Sigma^*}\|}, \frac{\eta_2 a \sigma_{\min}(R)}{2(1-\gamma)} \right\}$. The proof is completed by induction to show $C(K^{(t+1)}, \Sigma^{(t+1)}) \leq (1-\zeta) C(K^{(t)}, \Sigma^{(t)})$ and $aI \preceq \Sigma^{(t+1)} \preceq I$ holds for all $t \geq 0$: at $t = 0$, apply (4.21) to get $\eta_1 \leq \frac{1}{2\|R + \gamma B^\top P_{K^{(0)}} B\|}$ and $a \leq \frac{\tau}{2\|R + \gamma B^\top P_{K^{(0)}} B\|}$. Additionally with $\eta_2 = \frac{\tau(1-\gamma)}{2r_0^2} = \frac{2(1-\gamma)a^2}{\tau}$ and $aI \preceq \Sigma^{(0)} \preceq I$, Lemma 4.4.2 can be applied to get $C(K^{(1)}, \Sigma^{(1)}) \leq (1-\zeta) C(K^{(0)}, \Sigma^{(0)}) \leq C(K^{(0)}, \Sigma^{(0)})$, and $aI \preceq \Sigma^{(1)} \preceq I$. The proof proceeds by arguing that Lemma 4.4.2 can be applied at every step. If it were the case that $C\left(K^{(t)}, \Sigma^{(t)}\right) \leq (1-\zeta) C\left(K^{(t-1)}, \Sigma^{(t-1)}\right) \leq C\left(K^{(0)}, \Sigma^{(0)}\right)$ and $aI \preceq \Sigma^{(t)} \preceq I$, then

$$2\eta_1 = \frac{1}{\|R\| + \gamma \frac{\|B^\top B\| \left(C(K^{(0)}, \Sigma^{(0)}) - M_\tau \right)}{\mu + \frac{\gamma}{1-\gamma} \sigma_{\min}(W)}} \leq \frac{1}{\|R\| + \gamma \frac{\|B^\top B\| \left(C\left(K^{(t)}, \Sigma^{(t)}\right) - M_\tau \right)}{\mu + \frac{\gamma}{1-\gamma} \sigma_{\min}(W)}},$$

thus by (4.21) $\eta_1 \leq \frac{1}{\left\|R + \gamma B^\top P_{K^{(t)}} B\right\|}$ and in the same way $a \leq \frac{\tau}{2\left\|R + \gamma B^\top P_{K^{(t)}} B\right\|}$. Thus, Lemma 4.4.2 can be applied such that $C\left(K^{(t+1)}, \Sigma^{(t+1)}\right) - C(K^*, \Sigma^*) \leq (1-\zeta)\left(C\left(K^{(t)}, \Sigma^{(t)}\right) - C(K^*, \Sigma^*)\right)$. and $aI \preceq \Sigma^{(t+1)} \preceq I$. Thus, the induction is complete. Finally, observe that $0 < \zeta \leq \frac{2\mu\eta_1\sigma_{\min}(R)}{\|S_{K^*,\Sigma^*}\|} = \frac{\mu\sigma_{\min}(R)}{\|S_{K^*,\Sigma^*}\| r_0} < 1$, and $\zeta \leq \frac{\eta_2 a \sigma_{\min}(R)}{2(1-\gamma)} = \frac{\tau^2 \sigma_{\min}(R)}{8r_0^3}$, the proof is complete. □

4.5 Iterative Policy Optimization Method

In this section, we propose an iterative policy optimization method (IPO), in which we optimize a one-step deviation from the current policy in each iteration. For IPO, one can show both the global convergence with a linear rate and a local super-linear convergence when the initialization is close to the optimal policy. This local super-linear convergence result benefits from the entropy regularization.

By the Bellman equation for the value function $J_{K,\Sigma}$,

$$J_{K,\Sigma}(x) = \mathbb{E}_{u \sim \pi_{K,\Sigma}} \Big[x^\top Q x + u^\top R u + \tau \log \pi_{K,\Sigma}(u|x) + \gamma J_{K,\Sigma}(Ax + Bu + w) \Big].$$

We assume the one-step update (K', Σ') satisfies:

$$K', \Sigma' = \arg\min_{\widetilde{K}, \widetilde{\Sigma}} \mathbb{E}_{u \sim \pi_{\widetilde{K}, \widetilde{\Sigma}}} \Big[x^\top Q x + u^\top R u + \tau \log \pi_{\widetilde{K}, \widetilde{\Sigma}}(u|x) + \gamma J_{K,\Sigma}(Ax + Bu + w) \Big].$$

By direct calculation, we have the following explicit forms for the updates:

$$\begin{aligned} K^{(t+1)} &= K^{(t)} - \left(R + \gamma B^\top P_{K^{(t)}} B\right)^{-1} E_{K^{(t)}}, \\ \Sigma^{(t+1)} &= \frac{\tau}{2} \left(R + \gamma B^\top P_{K^{(t)}} B\right)^{-1}, \end{aligned} \tag{IPO}$$

for $t = 1, 2, \cdots$. The update of K in (IPO) is identical to the Gauss-Newton update when the learning rate is equal to 1 in [53]. The update of Σ in (IPO) is not gradient-based and only depends on the value of K in the previous step.

4.5.1 Global Linear Convergence

In this section, we establish the global convergence for (IPO) with a linear rate.

Theorem 4.5.1 (Global convergence of (IPO))**.** *For*

$$N \geq \frac{\|S_{K^*,\Sigma^*}\|}{\mu} \log \frac{C(K^{(0)}, \Sigma^{(0)}) - C(K^*, \Sigma^*)}{\varepsilon},$$

the iterative policy optimization algorithm (IPO) *has the following performance bound:*

$$C(K^{(N)}, \Sigma^{(N)}) - C(K^*, \Sigma^*) \leq \varepsilon.$$

The proof of Theorem 4.5.1 is immediate given the following lemma, which bounds the one-step progress of (IPO):

Lemma 4.5.1 (Contraction of (IPO))**.** *Suppose* (K', Σ') *follows a one-step updating rule of* (IPO) *from* (K, Σ)*. Then*

$$C(K', \Sigma') - C(K^*, \Sigma^*) \leq \left(1 - \frac{\mu}{\|S_{K^*,\Sigma^*}\|}\right) \left(C(K, \Sigma) - C(K^*, \Sigma^*)\right),$$

with $0 < \frac{\mu}{\|S_{K^*,\Sigma^*}\|} \leq 1$.

Theorem 4.5.1 suggests that (IPO) achieves a global linear convergence. Compared with (RPG), (IPO) exhibits faster convergence in terms of the rate at which the objective function decreases (*cf.* Lemmas 4.4.2 and 4.5.1). Furthermore, the subsequent section demonstrates that (IPO) enjoys a local super-linear convergence when the initial policy parameter is within a neighborhood of the optimal policy parameter.

4.5.2 Local Super-linear Convergence

This section establishes a local super-linear convergence for (IPO). We first introduce some constants used throughout this section:

$$\begin{aligned} \xi_{\gamma,\rho} &:= \frac{1 - \gamma\rho^2 + \gamma}{(1 - \gamma\rho^2)^2}, \qquad \zeta_{\gamma,\rho} := \frac{2 - \rho^2}{(1 - \rho^2)^2(1 - \gamma)} + \frac{1}{(1 - \rho^2)^2(1 - \gamma\rho^2)}, \\ \omega_{\gamma,\rho} &:= \frac{1}{(1 - \rho^2)(1 - \gamma)} + \frac{1}{(1 - \rho^2)(1 - \gamma\rho^2)}. \end{aligned} \tag{4.22}$$

To simplify the exposition, we often make use of the notation $S_{K^{(t)},\Sigma^{(t)}}$ and S_{K^*,Σ^*}, which we abbreviate as $S^{(t)}$ and S^*, respectively, provided that the relevant parameter values are clear from the context. Then, define

$$\begin{aligned}
\kappa :=& \frac{\rho + \|A\|}{|\sigma_{\min}(B)|}, \quad c := 2\rho\xi_{\gamma\rho}\|B\|(\|Q\| + \|R\|\kappa^2) + \frac{1}{\mu}\|S^*\|\|R\| \cdot (\kappa + \|K^*\|), \\
c_1 :=& \left(\xi_{\gamma,\rho}\|\mathbb{E}[x_0x_0^\top]\| + \zeta_{\gamma,\rho}\|B\Sigma^*B^\top + W\|\right) \cdot 2\rho\|B\| \\
&\cdot \Big(1 + \sigma_{\min}(R) \cdot \|R + \gamma B^\top P_{K^*}B\| + c\gamma\sigma_{\min}(R) \cdot \left(\|B\|\|A\| + \|B\|^2\kappa\right)\Big), \\
c_2 :=& \frac{c\tau\gamma\omega_{\gamma,\rho}\|B\|^4}{2\sigma_{\min}(R)^2}.
\end{aligned} \tag{4.23}$$

Note that for any $K \in \Omega$, $\|K\| \leq \frac{\|BK\|}{|\sigma_{\min}(B)|} \leq \frac{\|A-BK\|+\|A\|}{|\sigma_{\min}(B)|} \leq \frac{\rho+\|A\|}{|\sigma_{\min}(B)|} = \kappa$.

We now show that (IPO) achieves a super-linear convergence rate once the policy parameter (K, Σ) enters a neighborhood of the optimal policy parameter (K^*, Σ^*).

Theorem 4.5.2 (local super-linear convergence of (IPO))**.** *Let c_1 and c_2 be defined in* (4.23)*. Let $\delta := \min\left\{\frac{1}{c_1+c_2}\sigma_{\min}(S^*), \frac{\rho-\|A-BK^*\|}{\|B\|}\right\}$. Suppose that the initial policy $(K^{(0)}, \Sigma^{(0)})$ satisfies*

$$C(K^{(0)}, \Sigma^{(0)}) - C(K^*, \Sigma^*) \leq \left(\frac{1}{\mu} - \frac{1}{\|S^*\|}\right)^{-1} \sigma_{\min}\left(R + \gamma BP_{K^*}B\right)\delta^2, \tag{4.24}$$

then the iterative policy optimization algorithm (IPO) *has the following convergence rate: for $t = 0, 1, 2, \cdots$,*

$$C(K^{(t+1)}, \Sigma^{(t+1)}) - C(K^*, \Sigma^*) \leq \frac{(c_1 + c_2)\left(C(K^{(t)}, \Sigma^{(t)}) - C(K^*, \Sigma^*)\right)^{1.5}}{\sigma_{\min}(S^*)\sqrt{\mu\sigma_{\min}(R + \gamma B^\top P_{K^*}B)}}.$$

The following Lemma 4.5.2 is critical for establishing this local super-linear convergence: it shows that there is a contraction if the differences between two discounted state correlation matrices $\|S^{(t+1)} - S^*\|$ is small enough. Then, by the perturbation analysis for $S_{K,\Sigma}$ (Lemma 4.5.3), one can bound $\|S^{(t+1)} - S^*\|$ by $\|K^{(t)} - K^*\|$ (Lemma 4.5.4). The proof of Theorem 4.5.2 follows by ensuring the admissibility of model parameters $\left\{K^{(t)}, \Sigma^{(t)}\right\}_{t=0}^{\infty}$ along all the updates.

Lemma 4.5.2. *Suppose that $\|S^{(t+1)} - S^*\| \leq \sigma_{\min}(S^*)$ for all $t \geq 0$ when updating with* (IPO)*, then*

$$C(K^{(t+1)}, \Sigma^{(t+1)}) - C(K^*, \Sigma^*) \leq \frac{\|S^{(t+1)} - S^*\|}{\sigma_{\min}(S^*)}\left(C(K^{(t)}, \Sigma^{(t)}) - C(K^*, \Sigma^*)\right).$$

Lemma 4.5.3 ($S_{K,\Sigma}$ perturbation)**.** *For any (K_1, Σ_1) and (K_2, Σ_2) satisfying $\|A - BK_1\| \leq \rho$ and $\|A - BK_2\| \leq \rho$,*

$$\begin{aligned}
&\|S_{K_1,\Sigma_1} - S_{K_2,\Sigma_2}\| \\
&\leq \left(\xi_{\gamma,\rho}\|\mathbb{E}[x_0x_0^\top]\| + \zeta_{\gamma,\rho}\|B\Sigma_1B^\top + W\|\right) \cdot 2\rho\|B\|\,\|K_1 - K_2\| + \omega_{\gamma,\rho}\|B\|^2\|\Sigma_1 - \Sigma_2\|.
\end{aligned}$$

Lemma 4.5.4 (Bound of one-step update of $S^{(t)}$). *Assume that the update of parameter* (K, Σ) *follows* (IPO). *Let* c_1, c_2 *be defined in* (4.23). *Then for* $K^{(t+1)}$ *satisfying* $\|A - BK^{(t+1)}\| \leq \rho$, *we have* $\|S^{(t+1)} - S^*\| \leq (c_1 + c_2)\|K^{(t)} - K^*\|$.

Proof. (of Theorem 4.5.2). First, Theorem 4.2.1 shows that for an optimal K^*, $K^* = \gamma(R + \gamma B^\top P_{K^*} B)^{-1} B^\top P_{K^*} A$. Then, by the definition of E_K in Lemma 4.3.1,

$$E_{K^*} = -\gamma B^\top P_{K^*} A + (\gamma B^\top P_{K^*} B + R)K^* = 0.$$

Fix integer $t \geq 0$. Observe that

$$\begin{aligned} &\left(1 - \frac{\mu}{\|S_{K^*,\Sigma^*}\|}\right)\left(C(K^{(t)}, \Sigma^{(t)}) - C(K^*, \Sigma^*)\right) \overset{(a)}{\geq} C(K^{(t+1)}, \Sigma^{(t+1)}) - C(K^*, \Sigma^*) \\ &\overset{(b)}{\geq} \mu\sigma_{\min}\left(R + \gamma B P_{K^*} B\right)\|K^{(t+1)} - K^*\|^2 + f_{K^*}(\Sigma^*) - f_{K^*}(\Sigma^{(t+1)}) \\ &\overset{(c)}{\geq} \mu\sigma_{\min}\left(R + \gamma B P_{K^*} B\right)\|K^{(t+1)} - K^*\|^2. \end{aligned} \tag{4.25}$$

(a) is from the contraction property in Lemma 4.5.1; (b) follows from Lemma 4.3.3 and (4.59); to obtain (c), note that $f_{K^*}\left(\Sigma^*\right) - f_{K^*}\left(\Sigma^{(t+1)}\right) \geq 0$, since Σ^* is the maximizer of f_{K^*}. Thus, (4.25) and (4.24) imply $\|K^{(t+1)} - K^*\| \leq \delta$ which suggests that $\|A - BK^{(t+1)}\| \leq \|A - BK^*\| + \|B\|\|K^{(t+1)} - K^*\| \leq \rho$. Then by Lemma 4.5.4,

$$\|S^{(t+1)} - S^*\| \leq (c_1 + c_2)\|K^{(t)} - K^*\| \leq \sigma_{\min}(S^*). \tag{4.26}$$

Thus, one can apply Lemma 4.5.2 to get:

$$C(K^{(t+1)}, \Sigma^{(t+1)}) - C(K^*, \Sigma^*) \leq \frac{c_1 + c_2}{\sigma_{\min}(S^*)}\|K^{(t)} - K^*\|(C(K^{(t)}, \Sigma^{(t)}) - C(K^*, \Sigma^*)). \tag{4.27}$$

Using the same reasoning as in (4.25) (a) to (c), we have $C(K^{(t)}, \Sigma^{(t)}) - C(K^*, \Sigma^*) \geq \mu\sigma_{\min}\left(R + \gamma B^\top P_{K^*} B\right)\|K^{(t)} - K^*\|^2$ and plug it in (4.27) finishes the proof. □

4.6 Transfer Learning for RL

One can apply the local super-linear convergence result in Theorem 4.5.2 to provide an efficient policy transfer from a well-understood environment to a new yet similar environment. The idea is to use the optimal policy from the well-understood environment as an initialization of the policy update. If this initial policy is within the super-linear convergence region of the new environment, one may efficiently learn the optimal policy in the new environment.

Problem set-up and main results. We analyze two environments $\mathcal{M} := (A, B)$ and $\overline{\mathcal{M}} := (\overline{A}, \overline{B})$, with (K^*, Σ^*) and $(\overline{K}^*, \overline{\Sigma}^*)$ as their respective optimal policies and C and $\overline{C}$ as their respective objective functions. Assume that one has access to the optimal (regularized) policy (K^*, Σ^*) for environment $\mathcal{M}$, called the *well-understood environment*. We use (K^*, Σ^*) as a *policy initialization* for the less understood environment $\overline{\mathcal{M}}$, called the *new environment*. The goal is to investigate under what conditions this initialization enters the super-linear convergence regime of $\overline{\mathcal{M}}$.

Throughout this section, we specify the operator norm $\|\cdot\|$ as the one associated with vector q-norm. Namely, for $q \in (0,1)$ and $A \in \mathbb{R}^{n_1 \times n_2}$ for some positive integers n_1, n_2: $\|A\| := \|A\|_q = \sup_{x \neq 0} \left\{ \frac{\|Ax\|_q}{\|x\|_q},\ x \in \mathbb{R}^{n_2} \right\}$. For ease of the analysis and to make the two environments comparable, the following assumptions are made:

Assumption 4.6.1. *Assume the following conditions hold:*
1. Admissibility: (K^*, Σ^*) *is admissible for* $\overline{\mathcal{M}}$ *and* $\left(\overline{K}^*, \overline{\Sigma}^*\right)$ *is admissible for* $\mathcal{M}$*, i.e.,* $\|A - BK^*\| \leq \rho, \|\overline{A} - \overline{B}\,\overline{K}^*\| \leq \rho$ *with* $\rho \in (0, \frac{1}{\sqrt{\gamma}})$ *and* $\Sigma^* \succeq 0, \overline{\Sigma}^* \succeq 0$.
2. Model parameters: $\|B\|_q, \left\|\overline{B}\right\|_q \leq 1$.
3. Optimal policy: $\|K^*\| \leq 1$.

The first condition ensures that the environments $\mathcal{M}$ and $\overline{\mathcal{M}}$ are comparable. The second and third conditions are for ease of exposition and can be easily relaxed.

Similar to P_K defined in (4.16) for environment $\mathcal{M}$, let us define the Riccati equation for the new environment $\overline{M}$ as:

$$\overline{P}_K \;=\; \gamma(\overline{A} - \overline{B}K)^\top \overline{P}_K(\overline{A} - \overline{B}K) + Q + K^\top RK, \tag{4.28}$$

and define κ', c', c_1', c_2' in the same way as (4.23) with $(\overline{A}, \overline{B})$ replacing (A, B).

The following theorem suggests that if the environments $\mathcal{M}$ and $\overline{\mathcal{M}}$ are sufficiently close in the sense of (4.29), then (K^*, Σ^*) serves an efficient initial policy for $\overline{\mathcal{M}}$ which directly leads to a super-linear convergence for the new learning problem

Theorem 4.6.1. *Let* $c_{\gamma,\rho} := \max\{\frac{\gamma}{1-\gamma}, \frac{\gamma\rho}{1-\gamma\rho^2}\}$ *and* $\delta' := \min\left\{\frac{1}{c_1'+c_2'}\sigma_{\min}(S_{\overline{K}^*,\overline{\Sigma}^*}), \frac{\rho - \|\overline{A} - \overline{BK}^*\|}{\|\overline{B}\|}\right\}$*. If the following condition is satisfied:*

$$\left(\|A - \overline{A}\|_q + \|B - \overline{B}\|_q\right) \leq \frac{\left(\frac{1}{\mu} - \frac{1}{\|S^*\|}\right)^{-1} \sigma_{\min}\left(R + \gamma \overline{BP}_{\overline{K}^*}\overline{B}\right)(\delta')^2}{4c_{\gamma,\rho}\left(\left\|\mathbb{E}_{x_0 \sim \mathcal{D}}[x_0 x_0^\top] + W\right\|_q + \frac{\gamma}{1-\gamma} + 1\right)\frac{\|Q\|+\|R\|}{1-\gamma\rho^2}}, \tag{4.29}$$

then (K^*, Σ^*) *is within the super-linear convergence region of environment* $\overline{\mathcal{M}}$*, i.e.,*

$$\overline{C}(\overline{K}^{(t+1)}, \overline{\Sigma}^{(t+1)}) - \overline{C}(\overline{K}^*, \overline{\Sigma}^*) \leq \frac{(c_1' + c_2') \cdot \left(\overline{C}(\overline{K}^{(t)}, \overline{\Sigma}^{(t)}) - \overline{C}(\overline{K}^*, \overline{\Sigma}^*)\right)^{1.5}}{\sigma_{\min}(S_{\overline{K}^*,\overline{\Sigma}^*})\sqrt{\mu\sigma_{\min}(R + \gamma\overline{B}^\top \overline{P}_{\overline{K}^*}\overline{B})}},$$

for all $t \geq 0$*, if the initial policy follows* $(\overline{K}^{(0)}, \overline{\Sigma}^{(0)}) = (K^*, \Sigma^*)$ *and the policy updates according to* (IPO)*.*

Proof. (of Theorem 4.6.1). It is easy to verify that

$$\left|\overline{C}(\overline{K}^*, \overline{\Sigma}^*) - \overline{C}(K^*, \Sigma^*)\right| \leq \left|\overline{C}(\overline{K}^*, \overline{\Sigma}^*) - C(\overline{K}^*, \overline{\Sigma}^*)\right| + \left|\overline{C}(K^*, \Sigma^*) - C(K^*, \Sigma^*)\right|. \tag{4.30}$$

For any given policy (K, Σ) that is admissible to both $\mathcal{M}$ and $\overline{\mathcal{M}}$, we have $J_{K,\Sigma}(x) = x^\top P_K x + q_{K,\Sigma}$ and $\overline{J}_{K,\Sigma}(x) = x^\top \overline{P}_K x + \overline{q}_{K,\Sigma}$ with some symmetric positive definite matrices $P_K, \overline{P}_K \in \mathbb{R}^{n \times n}$ satisfying (4.16) and (4.28) respectively. In addition, the constants $q_{K,\Sigma}$ takes the form of (4.16) and $\overline{q}_{K,\Sigma} \in \mathbb{R}$ takes the form of $\overline{q}_{K,\Sigma} = \frac{1}{1-\gamma}\Big(- \frac{\tau}{2}\left(k + \log\left((2\pi)^k \det \Sigma\right)\right) + \operatorname{tr}\left(\Sigma R + \gamma(\Sigma(\overline{B})^\top \overline{P}_K \overline{B} + W \overline{P}_K)\right)\Big)$. Note that

$$\begin{aligned}
&P_K - \overline{P}_K = \gamma(A - BK)^\top P_K (A - BK) - \gamma(\overline{A} - \overline{B}K)^\top \overline{P}_K (\overline{A} - \overline{B}K) \\
&= \gamma(A - \overline{A} - (B - \overline{B})K)^\top P_K (A - BK) + \gamma(\overline{A} - \overline{B}K)^\top P_K (A - \overline{A} - (B - \overline{B})K) \\
&\quad + \gamma(\overline{A} - \overline{B}K)^\top (P_K - \overline{P}_K)(\overline{A} - \overline{B}K).
\end{aligned}$$

Hence we have $\|P_K - \overline{P}_K\| \leq 2\gamma(\|A - \overline{A}\| + \|K\|\|B - \overline{B}\|)\|P_K\|\rho + \gamma\rho^2\|P_K - \overline{P}_K\|$, and therefore, since $\gamma\rho^2 < 1$,

$$\|P_K - \overline{P}_K\| \leq \frac{2\gamma\rho}{1 - \gamma\rho^2}(\|A - \overline{A}\| + \|K\| \, \|B - \overline{B}\|)\|P_K\|. \tag{4.31}$$

Similarly,

$$\begin{aligned}
\overline{q}_{K,\Sigma} - q_{K,\Sigma} \;\; = \;\; & \frac{\gamma}{1-\gamma}\Big(\mathrm{Tr}\Big((B - \overline{B})^\top P_K \, B\Big) + \mathrm{Tr}\Big(\overline{B}^\top P_K \, (B - \overline{B})\Big) \\
& \qquad + \mathrm{Tr}\Big(\overline{B}^\top (P_K - \overline{P}_K) \, \overline{B}\Big) + \mathrm{Tr}(W(P_K - \overline{P}_K))\Big).
\end{aligned}$$

Recall that $x_0 \sim \mathcal{D}$ and denote $D_0 = \mathbb{E}_{x_0 \sim \mathcal{D}}[x_0 x_0^T]$. Therefore,

$$\begin{aligned}
& \left|\overline{C}(\overline{K}^*, \overline{\Sigma}^*) - C(\overline{K}^*, \overline{\Sigma}^*)\right| \\
\leq \;\; & \left|\operatorname{tr}\left((P_{\overline{K}^*} - \overline{P}_{\overline{K}^*})(D_0 + W)\right)\right| + \frac{\gamma}{1-\gamma}\left|\mathrm{Tr}\Big((B - \overline{B})^\top P_{K^*} \, B\Big)\right| \\
& + \frac{\gamma}{1-\gamma}\left|\mathrm{Tr}\Big(\overline{B}^\top P_{\overline{K}^*} \, (B - \overline{B})\Big)\right| + \frac{\gamma}{1-\gamma}\left|\mathrm{Tr}\Big(\overline{B}^\top (P_{\overline{K}^*} - \overline{P}_{\overline{K}^*}) \, \overline{B}\Big)\right| \\
\leq \;\; & \|P_{\overline{K}^*} - \overline{P}_{\overline{K}^*}\|_p \|D_0 + W\|_q + \frac{\gamma}{1-\gamma}\|B - \overline{B}\|_p \|P_{\overline{K}^*}\|_q (\|B\|_q + \|\overline{B}\|_q) \\
& + \frac{\gamma}{1-\gamma}\|P_{\overline{K}^*} - \overline{P}_{\overline{K}^*}\|_p \|\overline{B} \, \overline{B}^\top\|_q,
\end{aligned} \tag{4.32}$$

where the last inequality follows from $|\mathrm{Tr}(AB)| \leq \|A'\|_p \, \|B\|_q$ when $1/p + 1/q = 1$. This is a consequence of combining von Neumann's trace inequality with Hölder's inequality for Euclidean space.

(4.16) suggests that, for any admissible K such that $\|K\| \le 1$, we have $\|P_K\| \le \gamma\rho^2\|P_K\| + \|Q\| + \|R\|$. Hence, $\|P_K\| \le \frac{\|Q\|+\|R\|}{1-\gamma\rho^2}$. Combining this bound with (4.31) and the fact that $\|B\|_q, \|\overline{B}\|_q, \|K^*\| \le 1$, (4.32) can be bounded such that

$$\begin{aligned}(4.32) \le &\Big(\|D_0 + W\|_q + \frac{\gamma}{1-\gamma}\Big)\frac{2\gamma\rho}{1-\gamma\rho^2}\frac{\|Q\|+\|R\|}{1-\gamma\rho^2}\Big(\|A-\overline{A}\| + \|B-\overline{B}\|\Big)\\ &+ \frac{2\gamma}{1-\gamma}\frac{\|Q\|+\|R\|}{1-\gamma\rho^2}\|B-\overline{B}\|_q\\ &\le 2c_{\gamma,\rho}\Big(\|D_0+W\|_q + \frac{\gamma}{1-\gamma}+1\Big)\frac{\|Q\|+\|R\|}{1-\gamma\rho^2}\Big(\|A-\overline{A}\|_q + \|B-\overline{B}\|_q\Big). \end{aligned} \tag{4.33}$$

Similarly, we have

$$\begin{aligned}&\Big|\overline{C}(K^*, \Sigma^*) - C(K^*, \Sigma^*)\Big|\\ &\le 2c_{\gamma,\rho}\Big(\|D_0+W\|_q + \frac{\gamma}{1-\gamma}+1\Big)\frac{\|Q\|+\|R\|}{1-\gamma\rho^2}\Big(\|A-\overline{A}\|_q + \|B-\overline{B}\|_q\Big). \end{aligned} \tag{4.34}$$

Finally, plugging (4.33)-(4.34) into (4.30), we have

$$\begin{aligned}&\Big|\overline{C}(\overline{K}^*, \overline{\Sigma}^*) - \overline{C}(K^*, \Sigma^*)\Big|\\ &\le 4c_{\gamma,\rho}\Big(\|D_0+W\|_q + \frac{\gamma}{1-\gamma}+1\Big)\frac{\|Q\|+\|R\|}{1-\gamma\rho^2}\Big(\|A-\overline{A}\|_q + \|B-\overline{B}\|_q\Big). \end{aligned} \tag{4.35}$$

□

4.7 Model-free Extension

Model-based convergence provides a *foundation* for model-free analysis, as demonstrated [53, 76] where the more challenging policy convergence analysis for the model-based setting is followed by a relatively routine sample-based analysis of zeroth-order gradient approximation [126, 98, 11, 41]. Similarly, our analysis of local superlinear convergence and transfer learning applications within a model-based framework can be extended to developing a model-free algorithm, for instance, Algorithm 4.

In the setting with unknown model parameters A, B, Q, R, where the controller has only simulation access to the model, we apply a zeroth-order optimization method to approximate the gradient $\nabla_K \widehat{C(K, \Sigma)}$ and $\nabla_\Sigma \widehat{C(K, \Sigma)}$, as in Algorithm 4.

The updating rule for $\nabla_K \widehat{C(K, \Sigma)}$ is standard [53, 76]. We now explain the expression for $\nabla_\Sigma \widehat{C(K, \Sigma)}$. By Lemma 4.5 and Theorem 4.2, Σ is positive definite (with the time index t is omitted for ease of exposition). Let L denote the Cholesky decomposition of Σ such that $\Sigma = LL^\top$. Let $\text{vec}(\Sigma) \in \mathbb{R}^{D_\Sigma}$ and $\text{vec}(L) \in \mathbb{R}^{D_\Sigma}$ denote the stacked vectors of the

Algorithm 4 Policy Gradient Estimation with Unknown Parameters

```
Input: $K, \Sigma$, the number of trajectories $m$, smoothing parameter $r$, dimension $D_K$ and $D_\Sigma$
Apply Cholesky decomposition to matrix $\Sigma$ to get $L$ such that $\Sigma = LL^\top$.
for $i = 1, 2, \cdots, m$ do
    Sample a policy $\widehat{\Sigma}_i = \widehat{L}_i(\widehat{L}_i)^\top$, where $\mathrm{vec}(\widehat{L}_i) = \mathrm{vec}(L) + U_i$, where $U_i$ is drawn uniformly at random over vectors in $\mathbb{R}^{D_\Sigma}$ such that $\|U_i\|_F = r$. Simulate with policy $(K, \widehat{\Sigma}_i)$ from $x_0 \sim \mathcal{D}$ for $l$ steps. Let $\widehat{C}_i$ denote the empirical estimates:
        $\widehat{C}_i = \sum_{t=1}^{l} \gamma^t c_t,$
    where $c_t$ and $x_t$ are the costs and states on this trajectory.
    Sample a policy $\widehat{K}_i = K + U'_i$, where $U'_i$ is drawn uniformly at random over matrices in $\mathbb{R}^{k \times n}$ such that $\|U'_i\|_F = r$. Simulate with policy $(\widehat{K}_i, \Sigma_i)$ from $x_0 \sim \mathcal{D}$ for $l$ steps and let $\widehat{C'_i}$ and $\widehat{S'_i}$ denote the empirical estimates:
        $\widehat{C'_i} = \sum_{t=1}^{l} \gamma^t c'_t, \quad \widehat{S'_i} = \sum_{t=1}^{l} \gamma^t x'_t x_t^\top.$
end for
return the estimates of $\nabla_K C(K, \Sigma), \nabla_\Sigma C(K, \Sigma), S_{K,\Sigma}$:
    $\nabla_{\mathrm{vec}(\Sigma)} \widehat{C(K, \Sigma)} = \left(\nabla_{\mathrm{vec}(L)} \mathrm{vec}(\widehat{\Sigma}(\widehat{L}))^\top\right)^{-1} \cdot \left(\frac{1}{m} \sum_{i=1}^{m} \frac{D_\Sigma}{r^2} \widehat{C}_i U_i\right),$
    $\nabla_K \widehat{C(K, \Sigma)} = \frac{1}{m} \sum_{m=1}^{l} \gamma^t \frac{D_K}{r^2} \widehat{C'_i} U'_i, \quad \widehat{S_{K,\Sigma}} = \frac{1}{m} \sum_{i=1}^{m} \widehat{S'_i}.$
```

lower-triangular entries in matrices $\Sigma \in \mathbb{R}^{k\times k}$ and $L \in \mathbb{R}^{k\times k}$ respectively, with $D_\Sigma := \frac{k(k+1)}{2}$. Similarly, denote $\mathrm{vec}(K) \in \mathbb{R}^{D_K}$ as the stacked vectors of all entries in $K \in \mathbb{R}^{k\times n}$, with $D_K := k \times n$. Then $\nabla_{\mathrm{vec}(L)} C(K,\Sigma) = \nabla_{\mathrm{vec}(L)} \mathrm{vec}(\Sigma(L))^\top \cdot \nabla_{\mathrm{vec}(\Sigma)} C(K,\Sigma)$. In Algorithm 4, we approximate $\nabla_{\mathrm{vec}(L)} C(K,\Sigma)$ with zeroth-order estimate and the above equation to get $\nabla_{\mathrm{vec}(\Sigma)} \widehat{C(K},\Sigma)$. The estimate $\nabla_\Sigma \widehat{C(K},\Sigma)$ can be obtained by rearranging the entries of $\nabla_{\mathrm{vec}(\Sigma)} \widehat{C(K},\Sigma)$ into a matrix form.

4.8 Numerical Experiments

This section provides numerical experiments using (RPG) and (IPO) to illustrate the results established in Section 4.4, 4.5, and 4.6.

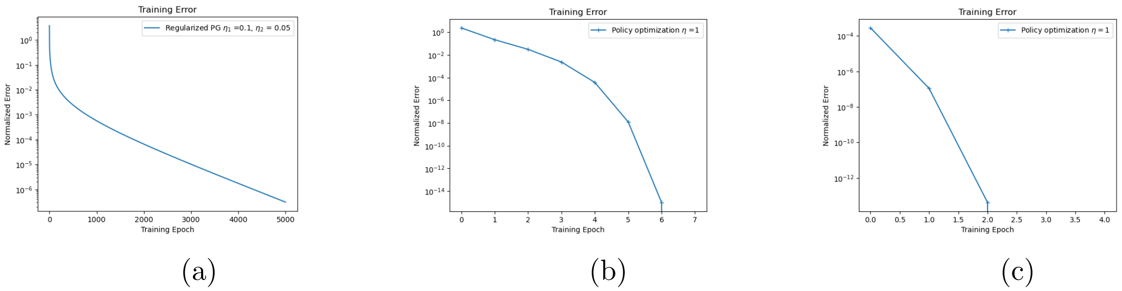
Setup. (1) Parameters: $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, $Q \in \mathbb{S}^n_+$, $R \in \mathbb{S}^k_{++}$ are generated randomly. The scaling of A is chosen so that A is stabilizing with high probability ($\sigma_{\max}(A) < \frac{1}{\sqrt{\gamma}}$). Initialization: $K^{(0)}_{i,j} = 0.01$ for all i, j, $\Sigma^{(0)} = I$.

(2) Transfer learning setup: $\overline{A}$ and $\overline{B}$ are the state transition matrices which are generated by adding a perturbation to A and B: $\overline{A}_{i,j} = A_{i,j} + u_{i,j}$, $\overline{B}_{i,j} = B_{i,j} + u'_{i,j}$, where $u_{i,j}$ and $u'_{i,j}$ are sampled from a uniform distribution on $[0, 10^{-3}]$. The initialization of $\overline{K}$ and $\overline{\Sigma}$ are the optimal solution K^* and Σ^* with state transition matrices A and B.

Performance measure. We use normalized error to quantify the performance of a given policy K, Σ, *i.e.*, NORMALIZED ERROR $= \frac{C(K,\Sigma) - C(K^*,\Sigma^*)}{C(K^*,\Sigma^*)}$, where K^*, Σ^* is the optimal policy defined in Theorem 4.2.1.

(Fast) Convergence. Figure 4.1a shows the linear convergence of (RPG), and Figure 4.1b shows the superior convergence rate of (IPO). The normalized error falls below 10^{-14} within just 6 iterations, and from the third iteration, it enters a region of super-linear convergence. Figure 4.1c shows the result of applying transfer learning using (IPO) in a perturbed environment, when the optimal policy in Figure 4.1a and 4.1b serves as an initialization. Figure 4.1c shows that if the process commences within a super-linear convergence region, then the error falls below 10^{-12} in just two epochs.

Regularization parameter τ. To demonstrate that entropy regularization can accelerate convergence, we conduct experiments with (RPG) under two settings: $n = 200, k = 10$ and $n = 200, k = 50$. We run (RPG) using various values of τ. Figure 4.2 illustrates that, in both settings, a larger value of τ results in a faster linear convergence rate to the optimal solution of (4.14). These results confirm that increasing the regularization parameter enhances the convergence speed, highlighting the practical benefits of entropy regularization in achieving faster optimization.



(a) (b) (c)

Figure 4.1: Performances of RPG, IPO, and transfer learning with IPO

Note. (a) Training error using RPG; (b) Training error using IPO; (c) Training error of transfer learning using IPO with $(\overline{K}^{(0)}, \overline{\Sigma}^{(0)}) = (K^*, \Sigma^*)$ and state transitions $(\overline{A}, \overline{B})$. $n = 400$, $k = 200$. The regularization parameter τ is chosen to be $\sigma_{\min}(R)$.

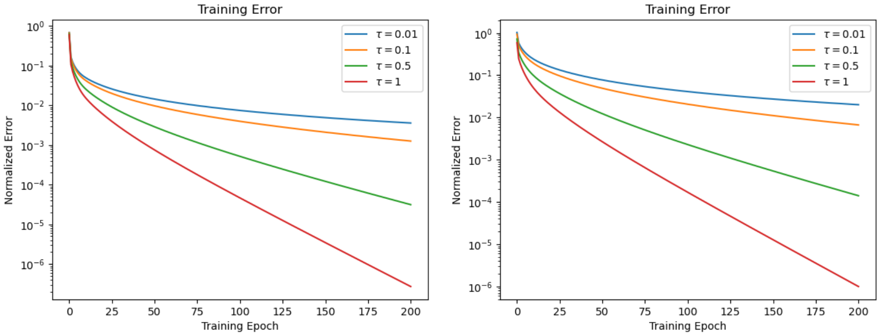


Figure 4.2: RPG with different regularization parameters τ

Note. Left: $n = 200, k = 10$; Right: $n = 200, k = 50$.

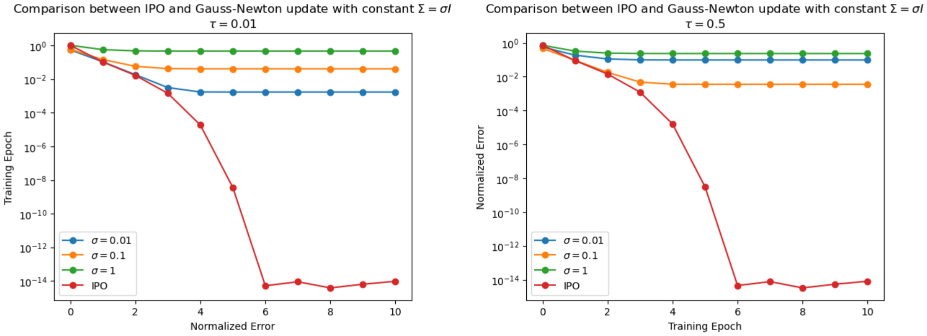


Figure 4.3: Comparison between IPO and Gauss-Newton update

Note. The Gauss-Newton update on K with constant covariance matrix follows (4.36); $n = 200, k = 50$. Left: $\tau = 0.01$; Right: $\tau = 0.5$.

The importance of updating Σ. As discussed in Section 4.5, updating K in (IPO) is identical to updating step of the Gauss-Newton algorithm in an unregularized setting. However, unlike the Gauss-Newton algorithm, (IPO) also updates the covariance matrix Σ simultaneously. To see the effect of dynamic updating of Σ, we compare the performance of (IPO) with the Gauss-Newton update on K with a constant covariance matrix Σ given by:

$$\begin{aligned} K^{(t+1)} &= K^{(t)} - \left(R + \gamma B^\top P_{K^{(t)}} B\right)^{-1} E_{K^{(t)}}, \\ \Sigma^{(t+1)} &= \sigma I, \end{aligned} \tag{4.36}$$

where σ is a fixed positive scalar. Figure 4.3 illustrates that the (IPO) algorithm with updates of Σ achieves a noticeably faster and superlinear convergence rate, compared to the Gauss-Newton update with a fixed Σ. This dynamic update of Σ allows (IPO) to reach the optimal point of Problem (4.14) more efficiently, highlighting the importance of adapting the covariance matrix during iterations.

4.9 Proofs of Main Results

4.9.1 Proofs in Section 4.2

4.9.1.1 Proof of Lemma 4.2.1

Denote the domain of the decision variable as $\mathcal{X} = \{p : \mathcal{A} \mapsto [0, \infty)\}$, and the feasible set as $\mathcal{F} = \{p : \mathcal{A} \mapsto [0, \infty) \mid \int_{\mathcal{A}} p(u)\mathrm{d}u = 1\} \subseteq \mathcal{X}$. Let $f : \mathcal{X} \mapsto \mathbb{R}$ denote the objective function, i.e.,

$$f(p) = \mathbb{E}_{u \sim p(\cdot)} \left[u^\top M u + b^\top u + \tau \log p(u)\right].$$

Let λ be a Lagrangian multiplier to the constraint $\int_{\mathcal{A}} p(u)\mathrm{d}u = 1$. Consider

$$\begin{aligned} \mathcal{L}(p, \lambda) &= \int_{\mathcal{A}} \left(u^\top M u + b^\top u + \tau \log p(u)\right) p(u)\mathrm{d}u + \lambda \left(\int_{\mathcal{A}} p(u)\mathrm{d}u - 1\right) \\ &= \int_{\mathcal{A}} L\left(u, p(u), \lambda\right) \mathrm{d}u - \lambda, \end{aligned}$$

where $L(u, v, \lambda) := (u^\top M u + b^\top u)v + \tau v \log v + \lambda v$. Additionally, define $g(\lambda) = \inf_{\mathcal{X}} \mathcal{L}(p, \lambda)$.

We now show the strong duality result:

$$g(\lambda^*) = \inf_{p \in \mathcal{F}} f(p), \tag{4.37}$$

with $\lambda^* = \arg\max_{\lambda \in \mathbb{R}} g(\lambda)$.

First, the weak duality result follows from

$$g(\lambda) = \inf_{p \in \mathcal{X}} \mathcal{L}(p, \lambda) \leq \inf_{p \in \mathcal{F}} \mathcal{L}(p, \lambda) = \inf_{p \in \mathcal{F}} f(p), \text{ for any } \lambda \in \mathbb{R}. \tag{4.38}$$

Moreover, since $\frac{\partial L}{\partial v}(u,v,\lambda) = \lambda + u^\top Mu + b^\top u + \tau + \tau \log v$, for any $\lambda \in \mathbb{R}, u \in \mathcal{A}$, the minimizer $p_\lambda(u)$ of $L(u,\cdot,\lambda)$ satisfies

$$p_\lambda(u) = \exp\left(-\frac{1}{\tau}(\lambda + u^\top Mu + b^\top u) - 1\right). \tag{4.39}$$

Therefore, by applying (4.39) to (4.38), we have

$$g(\lambda) = \mathcal{L}(p_\lambda, \lambda) = -\tau\left(\exp\left(-\frac{\lambda}{\tau} - 1\right) \cdot C + \frac{\lambda}{\tau}\right), \tag{4.40}$$

where $C := \int_{\mathcal{A}} \exp\left(-\frac{1}{\tau}(u^\top Mu + b^\top u)\right) \mathrm{d}u$. Direct computation yields the maximizer of g in (4.38) as $\lambda^* = \tau \log C - \tau$. Plugging λ^* to (4.39) shows $\int_{\mathcal{A}} p_{\lambda^*}(u)\mathrm{d}u = 1$, implying $p_{\lambda^*} \in \mathcal{F}$ and the strong duality (4.37) holds. Finally, by (4.37) and (4.39), it is clear that the optimal solution is a multivariate Gaussian distribution with $\mathcal{N}\left(-\frac{1}{2}M^{-1}b, \frac{\tau}{2}M^{-1}\right)$.

4.9.2 Proofs in Section 4.3

4.9.2.1 Proof of Lemma 4.3.1

$\nabla_\Sigma C(K,\Sigma)$ in (4.17) can be checked by direct gradient calculation. To verify $\nabla_K C(K,\Sigma)$ in (4.17), first define $f : \mathcal{X} \times \mathbb{R}^{k\times n} \to \mathbb{R}$ by $f(y,K) := y^\top P_K y, \forall y \in \mathcal{X}$ and we aim to find the gradient of f with respect to K. For any $y \in \mathcal{X}$, by the Riccati equation for P_K in (4.16) we have

$$\begin{aligned} f(y,K) &= y^\top \left(\gamma(A-BK)^\top P_K(A-BK) + Q + K^\top RK\right) y \\ &= \gamma f\left((A-BK)y, K\right) + y^\top \left(Q + K^\top RK\right) y. \end{aligned}$$

$\nabla_K f((A-BK)y,K)$ has two terms: one with respect to the input $(A-BK)y$ and one with respect to K in the subscript of P_K. This implies

$$\begin{aligned} \nabla_K f(y,K) &= 2\left(-\gamma B^\top P_K(A-BK) + RK\right) yy^\top + \gamma \nabla_K f(y',K)|_{y'=(A-BK)y} \qquad (4.41) \\ &= 2\left(-\gamma B^\top P_K(A-BK) + RK\right) \sum_{i=0}^{\infty} \gamma^i (A-BK)^i yy^\top (A^\top - K^\top B^\top)^i. \end{aligned}$$

Since $C(K,\Sigma) = \mathbb{E}_{x_0\sim\mathcal{D}}[x_0 P_K x_0] + q_{K,\Sigma}$ with P_K and $q_{K,\Sigma}$ satisfying (4.16), then the gradient of $C(K,\Sigma)$ with respect to K is

$$\begin{aligned}
\nabla_K C(K,\Sigma) =& \mathbb{E}[\nabla_K f(x_0,K)] + \nabla_K q_{K,\Sigma} \\
\overset{(a)}{=}& \mathbb{E}\Big[2\left(-\gamma B^\top P_K(A-BK)+RK\right)x_0x_0^\top + \gamma\nabla_K f(\hat{x}_1,K)|_{\hat{x}_1=(A-BK)x_0} \\
&+\sum_{t=0}\gamma^{t+1}\left(\nabla_K(\Sigma B^\top P_K B)+\nabla_K(w_t^\top P_K w_t)\right)\Big] \\
\overset{(b)}{=}& \mathbb{E}\Big[2\left(-\gamma B^\top P_K(A-BK)+RK\right)x_0x_0^\top + \gamma\nabla_K f(x_1,K)|_{x_1=A+Bu_0+w_0} \\
&+\sum_{t=1}\gamma^{t+1}\left(\nabla_K(\Sigma B^\top P_K B)+\nabla_K(w_t^\top P_K w_t)\right)\Big] \\
\overset{(c)}{=}& 2\left(-\gamma B^\top P_K(A-BK)+RK\right)\mathbb{E}\left[\sum_{i=0}^{\infty}\gamma^i x_i x_i^\top\right],
\end{aligned}$$

where (a) follows from applying (4.41) with $y = x_0$ and taking the gradient of $q_{K,\Sigma}$ in (4.16) with respect to K; (b) follows from

$$\mathbb{E}_{x_0,u_0,w_0}[f(Ax_0+Bu_0+w_0),K] = \mathbb{E}_{x_0,w_0}[f\left((A-BK)x_0,K\right) + w_0^\top P_K w_0] + \Sigma B^\top P_K B.$$

Using recursion to get (c).

4.9.2.2 Proof of Lemma 4.3.2

For a given policy $\pi_{K,\Sigma}(\cdot|x) = \mathcal{N}(-Kx,\Sigma)$ with parameter K and Σ, we define the state-action value function (also known as Q-function) $Q_{K,\Sigma} : \mathcal{X}\times\mathcal{A}\to\mathbb{R}$ as the cost of the policy starting with $x_0 = x$, taking a fixed action $u_0 = u$ and then proceeding with $\pi_{K,\Sigma}$. The Q-function is related to the value function $J_{K,\Sigma}$ defined in (4.1) as

$$Q_{K,\Sigma}(x,u) = x^\top Qx + u^\top Ru + \tau\pi_{K,\Sigma}(u|x) + \gamma\mathbb{E}\left[J_{K,\Sigma}(Ax+Bu+w)\right], \tag{4.42}$$

for any $(x,u)\in\mathcal{X}\times\mathcal{A}$. By definition of the Q-function, we also have the relationship $J_{K,\Sigma}(x) = \mathbb{E}_{u\sim\pi(\cdot|x)}\left[Q_{K,\Sigma}(x,u)\right], \forall x\in\mathcal{X}$. We then introduce the advantage function $A_{K,\Sigma} : \mathcal{X}\times\mathcal{A}\to\mathbb{R}$ of the policy π:

$$A_{K,\Sigma}(x,u) = Q_{K,\Sigma}(x,u) - J_{K,\Sigma}(x), \tag{4.43}$$

which reflects the gain one can harvest by executing control u instead of following the policy $\pi_{K,\Sigma}$ in state x.

With the notations of the Q-function in (4.42) and the advantage function in (4.43), we first provide a convenient form for the difference of the cost functions with respect to two different policies in Lemma 4.9.1. This will be used in the proof of Lemma 4.3.2.

Lemma 4.9.1 (Cost difference)**.** *Suppose policies π and π' are in form of (4.13) with parameters (K', Σ') and (K, Σ). Let $\{x'_t\}_{t=0}^{\infty}$ and $\{u'_t\}_{t=0}^{\infty}$ be state and action sequences generated by π' with noise sequence $\{w'_t\}_{t=0}^{\infty}$ (i.i.d with mean 0 and covariance W), i.e., $x'_{t+1} = Ax'_t + Bu'_t + w'_t$. Then for any $x \in \mathcal{X}$,*

$$J_{K',\Sigma'}(x) - J_{K,\Sigma}(x) = \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t A_{K,\Sigma}(x'_t, u'_t) \middle| x'_0 = x\right], \tag{4.44}$$

where the expectation is taken over $u'_t \sim \pi'(\cdot|x'_t)$ and w_t for all $t = 0, 1, 2, \cdots$.

The expected advantage for any $x \in \mathcal{X}$ by taking expectation over $u \sim \pi'(\cdot|x)$ is:

$$\begin{aligned}\mathbb{E}_{\pi'}\left[A_{K,\Sigma}(x,u)\right] &= x^\top (K'-K)^\top (R + \gamma B^\top P_K B)(K'-K)x \\ &+ 2x^\top (K'-K)^\top \left[(R + \gamma B^\top P_K B)K - \gamma B^\top P_K A\right] x + (1-\gamma)(f_K(\Sigma) - f_K(\Sigma')).\end{aligned} \tag{4.45}$$

Lemma 4.9.1 takes a similar form as [53, Lemma 10], but with an additional terms on Σ. The proof of Lemma 4.9.1 is provided in the online companion.

Proof. (of Lemma 4.3.2). By Lemma 4.9.1, for any π and π' in form of (4.13) with parameter (K, Σ) and (K', Σ') respectively, and for any $x \in \mathcal{X}$,

$$\begin{aligned}\mathbb{E}_{\pi'}[A_{K,\Sigma}(x,u)] &= (1-\gamma)(f_K(\Sigma) - f_K(\Sigma')) - \operatorname{tr}\left(xx^\top E_K^\top \left(R + \gamma B^\top P_K B\right)^{-1} E_K\right) \\ &+ \operatorname{tr}\Big[xx^\top \left(K' - K + \left(R + \gamma B^\top P_K B\right)^{-1} E_K\right)^\top \left(R + \gamma B^\top P_K B\right) \\ &\cdot \left(K' - K + (R + \gamma B^\top P_K B)^{-1} E_K\right)\Big] \\ &\geq -\operatorname{tr}\left(xx^\top E_K^\top \left(R + \gamma B^\top P_K B\right)^{-1} E_K\right) + (1-\gamma)\left(f_K\left(\Sigma\right) - f_K\left(\Sigma'\right)\right),\end{aligned} \tag{4.46}$$

with equality when $K - (R + \gamma B^\top P_K B)^{-1} E_K = K'$ holds. Let $\{x^*_t\}_{t=0}^{\infty}$ and $\{u^*_t\}_{t=0}^{\infty}$ be the state and control sequences generated under $\pi^*(\cdot|x) = \mathcal{N}(-K^*x, \Sigma^*)$ with noise sequence $\{w^*_t\}_{t=0}^{\infty}$ with mean 0 and covariance W. Apply Lemma 4.9.1 and (4.46) with π and π^* to get:

$$\begin{aligned}C(K,\Sigma) - C(K^*, \Sigma^*) &= -\mathbb{E}_{\pi^*}\left[\sum_{t=0}^{\infty} \gamma^t A_{K,\Sigma}(x^*_t, u^*_t)\right] \\ &\leq \operatorname{tr}\left(S_{K^*,\Sigma^*} E_K^\top \left(R + \gamma B^\top P_K B\right)^{-1} E_K\right) + f_K(\Sigma^*) - f_K(\Sigma),\end{aligned} \tag{4.47}$$

where f_K is defined in (4.15). To analyze the first term in (4.47), note that

$$\begin{aligned}&\operatorname{tr}\left(S_{K^*,\Sigma^*} E_K^\top \left(R + \gamma B^\top P_K B\right)^{-1} E_K\right) \leq \frac{\|S_{K^*,\Sigma^*}\|}{\sigma_{\min}(R)} \operatorname{Tr}(E_K^\top E_K) \\ &\leq \frac{\|S_{K^*,\Sigma^*}\|}{\mu^2 \sigma_{\min}(R)} \operatorname{tr}\left(S_{K,\Sigma} E_K^\top E_K S_{K,\Sigma}\right) = \frac{\|S_{K^*,\Sigma^*}\|}{4\mu^2 \sigma_{\min}(R)} \operatorname{tr}(\nabla_K^\top C(K,\Sigma) \nabla_K C(K,\Sigma)),\end{aligned} \tag{4.48}$$

where the last equation follows from (4.17).

To analyze the second terms in (4.47), note that f_K is a concave function, thus we can find its maximizer Σ_K^* by taking the gradient of f_K and setting it to 0, *i.e.,* $\Sigma_K^* = \frac{\tau}{2}(R + \gamma B^\top P_K B)^{-1}$. Thus,

$$\begin{aligned} &f_K(\Sigma^*) - f_K(\Sigma) \leq f_K(\Sigma_K^*) - f_K(\Sigma) \overset{(a)}{\leq} \operatorname{Tr}\left(\nabla f_K(\Sigma)^\top(\Sigma_K^* - \Sigma)\right) \\ &= \operatorname{Tr}\left(\nabla_\Sigma C(K,\Sigma) \cdot (R + \gamma B^\top P_K B)^{-1}\left((R + \gamma B^\top P_K B) - \frac{\tau}{2}\Sigma^{-1}\right)\Sigma\right) \\ &\overset{(b)}{\leq} (1-\gamma) \cdot \|(R + \gamma B^\top P_K B)^{-1}\| \cdot \|\nabla_\Sigma C(K,\Sigma)\|_F^2 \leq \frac{(1-\gamma)\|\nabla_\Sigma C(K,\Sigma)\|_F^2}{\sigma_{\min}(R)}, \end{aligned} \tag{4.49}$$

where (a) follows from the first order concavity condition for f_K and (b) is from $0 \prec \Sigma \preceq I$. Plug (4.48) and (4.49) into (4.47) to get (4.18).

For the lower bound, consider $K' = K - (R + \gamma B^\top P_K B)^{-1} E_K$ and $\Sigma' = \Sigma$ where equality holds in (4.46). Let $\{x_t'\}_{t=0}^\infty$, $\{u_t'\}_{t=0}^\infty$ be the sequence generated with K', Σ'. By $C(K^*, \Sigma^*) \leq C(K', \Sigma')$, we have

$$\begin{aligned} &C(K,\Sigma) - C(K^*,\Sigma^*) \geq C(K,\Sigma) - C(K',\Sigma') = -\mathbb{E}\left[\sum_{t=0}^\infty \gamma^t A_{K,\Sigma}(x_t', u_t')\right] \\ &= \operatorname{tr}\left(S_{K',\Sigma'} E_K^\top \left(R + \gamma B^\top P_K B\right)^{-1} E_K\right) \geq \frac{\mu}{\|R + \gamma B^\top P_K B\|}\operatorname{Tr}(E_K^\top E_K). \end{aligned} \tag{4.50}$$

□

4.9.2.3 Proof of Lemma 4.3.3

Lemma 4.9.2 shows that the cost objective is smooth in Σ when utilizing entropy regularization, given that Σ is bounded.

Lemma 4.9.2 (Smoothness of f_K (4.15))**.** *Let $K \in \mathbb{R}^{k\times n}$ be given and let f_K be defined in* (4.15)*. Fix $0 < a \leq 1$. For any symmetric positive definite matrices $X \in \mathbb{R}^{k\times k}$ and $Y \in \mathbb{R}^{k\times k}$ satisfying $aI \preceq X \preceq I$ and $aI \preceq Y \preceq I$,*

$$\left|f_K(X) - f_K(Y) + \operatorname{Tr}\left(\nabla f_K(X)^\top(Y - X)\right)\right| \leq M_a \operatorname{tr}\left((X^{-1}Y - I)^2\right),$$

where $M_a \in \mathbb{R}$ is defined in Lemma 4.3.3, and $M_a \geq \frac{\tau}{4(1-\gamma)}$.

Proof. (of Lemma 4.9.2). Fix symmetric positive definite matrices X and Y satisfying $aI \preceq X \preceq I$ and $aI \preceq Y \preceq I$. Then f_K being concave implies $f_K(X) - f_K(Y) + \operatorname{Tr}\left(\nabla f_K(X)^\top(Y - X)\right) \geq 0$. To find an upper bound, observe that

$$\begin{aligned} &f_K(X) - f_K(Y) + \operatorname{Tr}\left(\nabla f_K(X)^\top(Y - X)\right) \\ =&\frac{\tau}{2(1-\gamma)}\left(\log\det(X) - \log\det(Y) + \operatorname{Tr}\left(X^{-1}(Y - X)\right)\right). \end{aligned} \tag{4.51}$$

Since $X \succ 0$ and $Y \succ 0$, then all eigenvalues of $X^{-1}Y$ are real and positive and $\sigma_{\min}(X^{-1}Y) \geq \sigma_{\min}(X^{-1})\sigma_{\min}(Y) \geq a$.

Now let us show that there exists $m \in \mathbb{R}^+$ (independent of K) such that for any $Z \in \mathbb{R}^{k\times k}$ with real positive eigenvalues $a \leq z_1 \leq \cdots \leq z_k$, the following holds:

$$-\log(\det(Z)) + \operatorname{tr}(Z - I) \leq m \operatorname{tr}((Z - I)^2). \tag{4.52}$$

Note that (4.52) is equivalent to $\sum_{i=1}^{k} -\log(z_i) + z_i - 1 \leq m \sum_{i=1}^{k}(z_i - 1)^2$. With elementary algebra, one can verify that when $m := (-\log(a) + a - 1)\cdot(a-1)^{-2}$, it holds that $-\log(z) + z - 1 \leq m(z-1)^2$ for all $z \geq a$ and such an m satisfies $m \geq \frac{1}{2}$. Therefore, (4.52) holds. Combining (4.51) and (4.52) with $Z = X^{-1}Y$, we see $f_K(X) - f_K(Y) + \operatorname{tr}\left(\nabla f_K(X)^\top (Y - X)\right) \leq \frac{\tau m}{2(1-\gamma)} \operatorname{tr}\left((X^{-1}Y - I)^2\right)$. □

Proof. (of Lemma 4.3.3). The first equality immediately results from (4.45) in Lemma 4.9.1. The last inequality follows directly from Lemma 4.9.2. □

4.9.3 Proofs in Section 4.4

4.9.3.1 Proofs of Lemma 4.4.1

To ease the exposition, let η denote η_2. The proof is composed of two steps. First, one can show

$$aI \preceq \Sigma - \eta(1-\gamma)^{-1}(R - \frac{\tau}{2}\Sigma^{-1} + \gamma B^\top P_K B) \preceq I. \tag{4.53}$$

Let $g : \mathbb{R}^+ \to \mathbb{R}$ be a function such that $g(x) = x + \frac{\eta\tau}{2(1-\gamma)x}$. Thus, g monotonically increases on $\left[\sqrt{\frac{\eta\tau}{2(1-\gamma)}}, \infty\right)$. Since $\sqrt{\frac{\eta\tau}{2(1-\gamma)}} \leq a \leq \frac{\sigma_{\min}(R)}{\|R+\gamma B^\top P_K B\|} \leq 1$, then

$$\begin{aligned}
&\Sigma + \frac{\eta\tau}{2(1-\gamma)}\Sigma^{-1} - \frac{\eta(R+\gamma B^\top P_K B)}{1-\gamma} \succeq \left(a + \frac{\eta\tau}{2(1-\gamma)a}\right) I - \frac{\eta(R+\gamma B^\top P_K B)}{1-\gamma}\\
&\succeq \left(a + \frac{\eta\|R+\gamma B^\top P_K B\|}{1-\gamma}\right) I - \frac{\eta(R+\gamma B^\top P_K B)}{1-\gamma} \succeq aI, \text{ and}\\
&\Sigma + \frac{\eta\tau}{2(1-\gamma)}\Sigma^{-1} - \frac{\eta(R+\gamma B^\top P_K B)}{1-\gamma} \preceq \left(1 + \frac{\eta\tau}{2(1-\gamma)}\right) I - \frac{\eta(R+\gamma B^\top P_K B)}{1-\gamma}\\
&\preceq \left(1 + \frac{\eta\sigma_{\min}(R)}{1-\gamma}\right) I - \frac{\eta(R+\gamma B^\top P_K B)}{1-\gamma} \preceq I, \text{ hence (4.53).}
\end{aligned}$$

Second, one can show $aI \preceq \Sigma' \preceq I$: observe that (4.53) is equivalent to $aI - \Sigma \preceq -\frac{\eta}{1-\gamma}(R + \gamma B^\top P_K B - \frac{\tau}{2}\Sigma^{-1}) \preceq I - \Sigma$. Then by multiplying Σ to both sides then adding a Σ to each term, we have

$$a\Sigma^2 - \Sigma^3 + \Sigma \preceq \Sigma - \frac{\eta}{1-\gamma}\Sigma(R + \gamma B^\top P_K B - \frac{\tau}{2}\Sigma^{-1})\Sigma \preceq \Sigma^2 - \Sigma^3 + \Sigma.$$

With $aI - \Sigma \preceq 0$, $I - \Sigma \succeq 0$ and $\Sigma \preceq I$, it holds that $aI - \Sigma \preceq (aI - \Sigma)\Sigma^2$, and $I - \Sigma \succeq (I - \Sigma)\Sigma^2$. This implies $aI \preceq a\Sigma^2 - \Sigma^3 + \Sigma \preceq \Sigma' \preceq \Sigma^2 - \Sigma^3 + \Sigma \preceq I$.

4.9.4 Proof of Lemma 4.4.2

For ease of exposition, write $S = S_{K,\Sigma}$, $S' = S_{K',\Sigma'}$, and $S^* = S_{K^*,\Sigma^*}$. Let M_a be defined in the same way as in Lemma 4.3.3. Let f_K be defined as (4.15). Then Lemma 4.3.3 implies

$$\begin{aligned} C(K',\Sigma') - C(K,\Sigma) = \operatorname{Tr}\left(S'(K'-K)^\top(R+\gamma B^\top P_K B)(K'-K)\right) \\ + 2\operatorname{Tr}\left(S'(K'-K)^\top E_K\right) + f_K(\Sigma) - f_K(\Sigma'). \end{aligned} \tag{4.54}$$

By (RPG),

$$\begin{aligned} &\operatorname{tr}\left(S'(K'-K)^\top(R+\gamma B^\top P_K B)(K'-K)\right) + 2\operatorname{Tr}\left(S'(K'-K)^\top E_K\right) \\ \leq &4\eta_1^2\|R+\gamma B^\top P_K B\|\operatorname{Tr}\left(S' E_K^\top E_K\right) - 4\eta_1\operatorname{Tr}\left(S' E_K^\top E_K\right) \overset{(a)}{\leq} -2\eta_1\operatorname{Tr}\left(S' E_K^\top E_K\right) \\ \leq &-2\eta_1\mu\operatorname{Tr}\left(E_K^\top E_K\right) \overset{(b)}{\leq} -2\eta_1\mu\frac{\sigma_{\min}(R)}{\|S^*\|}\operatorname{tr}\left(S_{K^*,\Sigma^*}E_K^\top\left(R+\gamma B^\top P_K B\right)^{-1}E_K\right), \end{aligned} \tag{4.55}$$

where (a) follows from $\eta_1 \leq (2\|R+\gamma B^\top P_K B\|)^{-1}$ and (b) follows from (4.48).

By Lemma 4.4.1, $aI \preceq \Sigma' \preceq I$. Then by Lemma 4.9.2,

$$\begin{aligned} &f_K(\Sigma) - f_K(\Sigma') \\ &\leq -\frac{\eta_2}{(1-\gamma)^2}\operatorname{tr}\left(\left(R+\gamma B^\top P_K B - \frac{\tau}{2}\Sigma^{-1}\right)\Sigma\left(R+\gamma B^\top P_K B - \frac{\tau}{2}\Sigma^{-1}\right)\Sigma\right) \\ &\qquad + \frac{(\eta_2)^2 M_a}{(1-\gamma)^2}\operatorname{tr}\left(\left((R+\gamma B^\top P_K B)\Sigma - \frac{\tau}{2}I\right)^2\right) \\ &\overset{(c)}{\leq} -\frac{\eta_2}{2(1-\gamma)^2}\operatorname{tr}\left(\left((R+\gamma B^\top P_K B)\Sigma - \frac{\tau}{2}I\right)^2\right). \end{aligned}$$

Here (c) follows from the inequality $\eta_2 = \frac{2(1-\gamma)a^2}{\tau} \leq \frac{2(1-\gamma)}{\tau}\left(\frac{\tau}{2\|R+\gamma B^\top P_K B\|}\right)^2 \overset{(d)}{\leq} \frac{2(1-\gamma)}{\tau} \overset{(e)}{\leq} \frac{1}{2M_a}$, where (d) is obtained from the fact that $\tau \leq 2\sigma_{\min}(R) \leq 2\|R+\gamma B^\top P_K B\|$, and (e) follows from Lemma 4.9.2. Meanwhile, observe from (4.49) that

$$\begin{aligned} f_K(\Sigma^*) - f_K(\Sigma) &\leq \frac{1}{(1-\gamma)}\operatorname{tr}\Big(\left((R+\gamma B^\top P_K B)\Sigma - \frac{\tau}{2}I\right)\cdot \\ &\qquad \Sigma^{-1}(R+\gamma B^\top P_K B)^{-1}\left((R+\gamma B^\top P_K B)\Sigma - \frac{\tau}{2}I\right)\Big) \\ &\leq \frac{1}{(1-\gamma)a\sigma_{\min}(R)}\operatorname{tr}\left(((R+\gamma B^\top P_K B)\Sigma - \frac{\tau}{2}I)^2\right), \end{aligned}$$

while implies

$$f_K(\Sigma) - f_K(\Sigma') \leq -\frac{\eta_2 a\sigma_{\min}(R)}{2(1-\gamma)}\left(f_K(\Sigma) - f_K(\Sigma^*)\right). \tag{4.56}$$

Finally, with $\zeta = \min\left\{\frac{2\mu\eta_1\sigma_{\min}(R)}{\|S^*\|}, \frac{\eta_2 a\sigma_{\min}(R)}{2(1-\gamma)}\right\}$, plug (4.55) and (4.56) in (4.54) to get $C(K',\Sigma') - C(K,\Sigma) \leq -\zeta\Big(\operatorname{tr}\left(S_{K^*,\Sigma^*}E_K^\top\left(R+\gamma B^\top P_K B\right)^{-1}E_K\right) + f_K(\Sigma) - f_K(\Sigma^*)\Big)$. The proof is finished by applying (4.47) then adding $C(K,\Sigma) - C(K^*,\Sigma^*)$ to both sides.

4.9.4.1 Proof of Lemma 4.4.3

By (4.14) and (4.16),

$$\begin{aligned} C(K,\Sigma) =& \mathbb{E}_{x_0\sim\mathcal{D}}\left[x_0^\top P_K x_0\right] + \frac{\gamma}{1-\gamma}\operatorname{Tr}(WP_K) \\ &+ \frac{1}{1-\gamma}\operatorname{Tr}\left(\Sigma(R+\gamma B^\top P_K B)\right) - \frac{\tau}{2(1-\gamma)}(k+\log((2\pi)^k\det\Sigma)). \end{aligned}$$

Note that

$$\begin{aligned} &\frac{1}{1-\gamma}\operatorname{Tr}\left(\Sigma(R+\gamma B^\top P_K B)\right) - \frac{\tau}{2(1-\gamma)}(k+\log((2\pi)^k\det\Sigma)) \\ &\geq \frac{1}{1-\gamma}\left(\sigma_{\min}(R)\operatorname{Tr}(\Sigma) - \frac{\tau}{2}\left(k+k\log(2\pi)\right) - \frac{\tau}{2}\log\det\Sigma\right) \\ &\overset{(a)}{\geq} \frac{1}{1-\gamma}\left(\frac{\tau k}{2} - \frac{\tau}{2}\left(k+k\log(2\pi)\right) - \frac{\tau k}{2}\log(\frac{\tau}{2\sigma_{\min}(R)})\right) = M_\tau, \end{aligned}$$

where (a) follows from the fact that $\sigma_{\min}(R)\operatorname{Tr}(\Sigma) - \frac{\tau}{2}\left(k+k\log(2\pi) - \frac{\tau}{2}\log\det\Sigma\right)$ is a convex function with respect to Σ with minimizer $\frac{\tau}{2\sigma_{\min}(R)}I$. Thus, $C(K,\Sigma) \geq \mathbb{E}_{x_0\sim\mathcal{D}}\left[x_0^\top P_K x_0\right] + \frac{\gamma\operatorname{Tr}(WP_K)}{1-\gamma} + M_\tau \geq \left(\mu + \frac{\gamma\sigma_{\min}(W)}{1-\gamma}\right)\|P_K\| + M_\tau$.

4.9.5 Proofs in Section 4.5

4.9.5.1 Proof of Lemma 4.5.1

Fix K,Σ and let f_K be defined in (4.15). Observe that Σ' following (IPO) update is the maximizer of f_K. Then by Lemma 4.3.3,

$$\begin{aligned} C(K',\Sigma') - C(K,\Sigma) &= -\operatorname{Tr}\left(S_{K',\Sigma'}E_K^\top(R+\gamma B^\top P_K B)^{-1}E_K\right) - f_K(\Sigma') + f_K(\Sigma) \\ &\overset{(a)}{\leq} -\frac{\mu}{\|S_{K^*,\Sigma^*}\|}\operatorname{tr}\left(S_{K^*,\Sigma^*}E_K^\top(R+\gamma B^\top P_K B)^{-1}E_K\right) - \frac{\mu}{\|S_{K^*,\Sigma^*}\|}(f_K(\Sigma') - f_K(\Sigma)) \\ &\overset{(b)}{\leq} -\frac{\mu}{\|S_{K^*,\Sigma^*}\|}\left(\operatorname{tr}\left(S_{K^*,\Sigma^*}E_K^\top(R+\gamma B^\top P_K B)^{-1}E_K\right) + f_K(\Sigma^*) - f_K(\Sigma)\right), \end{aligned}$$

where (a) follows the fact that $f_K(\Sigma') - f_K(\Sigma) \geq 0$ and $0 < \frac{\mu}{\|S_{K^*,\Sigma^*}\|} \leq 1$; (b) follows from fact that $f_K(\Sigma') \geq f_K(\Sigma^*)$. Finally, apply (4.47) to get $C(K',\Sigma') - C(K,\Sigma) \leq -\frac{\mu}{\|S_{K^*,\Sigma^*}\|}\left(C(K,\Sigma) - C(K^*,\Sigma^*)\right)$ and add $C(K,\Sigma) - C(K^*,\Sigma^*)$ to both sides finishes the proof.

4.9.5.2 Proof of Lemma 4.5.2

Let K' denote $K^{(t+1)}$, K denote $K^{(t)}$ and S' denote $S^{(t+1)}$. Apply Lemma 4.3.3 with f_K defined in (4.15) to get

$$
\begin{aligned}
& C(K',\Sigma') - C(K,\Sigma) \\
&= -\operatorname{Tr}\left(S^* E_K^\top (R+\gamma B^\top P_K B)^{-1} E_K\right) - \operatorname{Tr}\left((S'-S^*) E_K^\top (R+\gamma B^\top P_K B)^{-1} E_K\right) \\
&\quad + f_K(\Sigma) - f_K(\Sigma') \\
&\overset{(a)}{\leq} \left(-1 + \|(S'-S^*)S^{*-1}\|\right)\left(\operatorname{Tr}\left(S^* E_K^\top (R+\gamma B^\top P_K B)^{-1} E_K\right) + f_K(\Sigma^*) - f_K(\Sigma)\right),
\end{aligned}
$$

where (a) follows from the assumption $\|S'-S^*\| \leq \sigma_{\min}(S^*)$ which implies

$$
-1 + \left\|(S'-S^*)S^{*-1}\right\| \in [-1,0],
$$

and the fact Σ' is the maximizer of f_K. Finally, by (4.47),

$$
\begin{aligned}
C(K',\Sigma') - C(K,\Sigma) &\leq \left(-1 + \left\|(S'-S^*)S^{*-1}\right\|\right)\left(C(K,\Sigma) - C(K^*,\Sigma^*)\right) \\
&\leq \left(-1 + \frac{\|S'-S^*\|}{\sigma_{\min}(S^*)}\right)\left(C(K,\Sigma) - C(K^*,\Sigma^*)\right).
\end{aligned}
$$

Adding $C(K,\Sigma) - C(K^*,\Sigma^*)$ to both sides of the above inequality finishes the proof.

4.9.5.3 Proof of Lemma 4.5.3

This section conducts a perturbation analysis on $S_{K,\Sigma}$ and aims to prove Lemma 4.5.3, which bounds $\|S_{K_1,\Sigma_1} - S_{K_2,\Sigma_2}\|$ by $\|K_1 - K_2\|$ and $\|\Sigma_1 - \Sigma_2\|$.

The proof of Lemma 4.5.3 proceeds with a few technical lemmas. First, define the linear operators on symmetric matrices. For symmetric matrix $X \in \mathbb{R}^{n\times n}$, we set

$$
\begin{aligned}
&\mathcal{F}_K(X) := (A-BK)X(A-BK)^\top, \quad \mathcal{G}_t^K(X) := (A-BK)^t X (A-BK)^{\top t} \\
&\mathcal{T}_K(X) := \sum_{t=0}^{\infty} \gamma^t (A-BK)^t X (A-BK)^{\top t} = \sum_{t=0}^{\infty} \gamma^t \mathcal{G}_t^K(X).
\end{aligned}
\tag{4.57}
$$

Note that when $\|A-BK\| < \frac{1}{\sqrt{\gamma}}$, we have

$$
\mathcal{T}_K = (I - \gamma\mathcal{F}_K)^{-1}. \tag{4.58}
$$

We also define the induced norm for these operators as $\|T\| := \sup_X \frac{\|T(X)\|}{\|X\|}$, where $T = \mathcal{F}_K, \mathcal{G}_t^K, \mathcal{T}_K$ and the supremum is over all symmetric matrix $X \in \mathbb{R}^{n\times n}$ with non-zero spectral norm.

We also define the induced norm for these operators as $\|T\| := \sup_X \frac{\|T(X)\|}{\|X\|}$, where $T = \mathcal{F}_K, \mathcal{G}_t^K, \mathcal{T}_K$ and the supremum is over all symmetric matrix $X \in \mathbb{R}^{n\times n}$ with non-zero spectral norm.

Lemma 4.9.3. $S_{K,\Sigma} = \mathcal{T}_K(\mathbb{E}[x_0x_0^\top]) + \sum_{t=0}^{\infty}\gamma^t\sum_{s=1}^{t}\mathcal{G}_{t-s}^K(B\Sigma B^\top + W)$, *for any* $K, \Sigma \in \Omega$.

Lemma 4.9.4. $\|\mathcal{F}_{K_1} - \mathcal{F}_{K_2}\| \leq (\|A-BK_1\| + \|A-BK_2\|)\|B\|\,\|K_1 - K_2\|$, *for any* K_1 *and* K_2 *in* Ω.

Lemma 4.9.5. $\sum_{t=0}^{T-1}\|(\mathcal{G}_t' - \mathcal{G}_t)(X)\| \leq \frac{2-\rho^2-\rho^{2T}}{(1-\rho^2)^2}\|\mathcal{F}' - \mathcal{F}\|\|X\|, \quad \forall T \geq 1$, *and* $\|(\mathcal{T}_{K_1} - \mathcal{T}_{K_2})(X)\| \leq \sum_{t=0}^{\infty}\gamma^t\|(\mathcal{G}_t^{K_1} - \mathcal{G}_t^{K_2})(X)\| \leq \xi_{\gamma,\rho}\|\mathcal{F}_{K_1} - \mathcal{F}_{K_2}\|\|X\|$, *for any* $K_1 \in \Omega$ *and* $K_2 \in \Omega$, *with* $\xi_{\gamma,\rho}$ *defined in* (4.22).

Proofs of Lemma 4.9.3, 4.9.4, and 4.9.5 can be found in the online companion.

Proof. (of Lemma 4.5.3). Denote $\mathcal{G}_t = \mathcal{G}_t^{K_1}, \mathcal{G}_t' = \mathcal{G}_t^{K_2}, \mathcal{T} = \mathcal{T}_{K_1}, \mathcal{T}' = \mathcal{T}_{K_2}, \mathcal{F} = \mathcal{F}_{K_1}$ and $\mathcal{F}' = \mathcal{F}_{K_2}$ to ease the exposition. Observe that

$$\begin{aligned}
\|S_{K_1,\Sigma_1} - S_{K_2,\Sigma_2}\| &\overset{(a)}{\leq} \|(\mathcal{T} - \mathcal{T}')\mathbb{E}[x_0x_0^\top]\| + \|\sum_{t=0}^{\infty}\gamma^t\sum_{s=1}^{t}(\mathcal{G}_s^{'} - \mathcal{G}_s)(B\Sigma_1 B^\top + W)\| \\
&\quad + \sum_{t=0}^{\infty}\gamma^t\sum_{s=1}^{t}\|\mathcal{G}_{t-s}'(B\Sigma_1 B^\top + W) - \mathcal{G}_{t-s}'(B\Sigma_2 B^\top + W)\| \\
&\overset{(b)}{\leq} \xi_{\gamma,\rho}\|\mathcal{F} - \mathcal{F}'\|\|\mathbb{E}[x_0x_0^\top]\| + \sum_{t=0}^{\infty}\gamma^t\frac{2-\rho^2-\rho^{2t}}{(1-\rho^2)^2}\|\mathcal{F}' - \mathcal{F}\|\|B\Sigma_1 B^\top + W\| \\
&\quad + \sum_{t=0}^{\infty}\gamma^t\sum_{s=1}^{t}\rho^{2(t-s)}\|B\|^2\|\Sigma_1 - \Sigma_2\|,
\end{aligned}$$

where (a) is from (4.12) and (b) follows from Lemma 4.9.5 and $\|\mathcal{G}_t(X) - \mathcal{G}_t(X')\| = \|(A - BK)^t(X - X')(A - BK)^{\top t}\| \leq \rho^{2t}\|X - X'\|, \forall X, X' \in \mathbb{R}^{n\times n}, \forall t \geq 0$. Finally, applying (4.22) and Lemma 4.9.4 finishes the proof. □

4.9.5.4 Proof of Lemma 4.5.4

The objective of this section is to provide a bound for $\|S^{(t+1)} - S^*\|$ in terms of $\|K^{(t)} - K^*\|$, as summarized in Lemma 4.5.4.

Note that Lemma 4.5.3 can be employed to derive a bound on $\|S^{(t+1)} - S^*\|$ in relation to $\|K^{(t+1)} - K^*\|$ and $\|\Sigma^{(t+1)} - \Sigma^*\|$. In this section, we further establish this bound by deriving bounds for $\|K^{(t+1)} - K^*\|$ and $\|\Sigma^{(t+1)} - \Sigma^*\|$ in terms of $\|K^{(t)} - K^*\|$ and $\|P_{K^{(t)}} - P_{K^*}\|$ (*cf.* Lemma 4.9.6 and Lemma 4.9.7). Additionally, the perturbation analysis for P_K in Lemma 4.9.8 demonstrates $\|P_{K^{(t)}} - P_{K^*}\|$ can be bounded by $\|K^{(t)} - K^*\|$, which completes the proof for Lemma 4.5.4.

Lemma 4.9.6 (Bound of one-step update of K)**.** *Assume the update of parameter K follows the updating rule in* (IPO)*. Then it holds that:*

$$\begin{aligned}
\|K^{(t+1)} - K^*\| \leq &\big(1 + \sigma_{\min}(R)\|\gamma B^\top P_{K^*}B + R\|\big)\cdot\|K^{(t)} - K^*\| \\
&+ \gamma\sigma_{\min}(R)\left(\|B\|\|A\| + \|B\|^2\kappa\right)\cdot\|P_{K^{(t)}} - P_{K^*}\|.
\end{aligned}$$

Proof. (of Lemma 4.9.6). Let K' denote $K^{(t+1)}$ and K denote $K^{(t)}$ to ease the notation. Theorem 4.2.1 shows that for an optimal K^*,

$$K^* = \gamma(R + \gamma B^\top P_{K^*} B)^{-1} B^\top P_{K^*} A.$$

Then, by the definition of E_K in Lemma 4.3.1,

$$E_{K^*} = -\gamma B^\top P_{K^*} A + (\gamma B^\top P_{K^*} B + R) K^* = 0. \tag{4.59}$$

Now we bound the difference between $K' - K$:

$$\begin{aligned} \|K' - K^*\| &= \|K - (R + \gamma B^\top P_K B)^{-1} E_K - K^* + (R + \gamma B^\top P_K B)^{-1} E_{K^*}\| \\ &\leq \|K - K^*\| + \|(R + \gamma B^\top P_{K^*} B)^{-1}\| \|E_{K^*} - E_{K^*}\| \\ &\leq \|K - K^*\| + \sigma_{\min}(R) \|E_K - E_{K^*}\|. \end{aligned} \tag{4.60}$$

To bound the difference between E_K and E_{K^*}, observe:

$$\begin{aligned} \|E_K - E_{K^*}\| &\leq \gamma\|B\|\|A\|\|P_K - P_{K^*}\| + \|(\gamma B^\top P_{K^*} B + R)K^* - (\gamma B^\top P_{K^*} B + R)K\| \\ &\quad + \|(\gamma B^\top P_{K^*} B + R)K - (\gamma B^\top P_K B + R)K\| \\ &\leq \gamma\|B\|\|A\|\|P_K - P_{K^*}\| + \|\gamma B^\top P_{K^*} B + R\|\|K^* - K\| \\ &\quad + \gamma\|B\|^2\|P_{K^*} - P_K\|\|K\|. \end{aligned} \tag{4.61}$$

Combining (4.60) and (4.61), then using $\kappa \geq \|K\|$ for any $K \in \Omega$ completes the proof. □

Lemma 4.9.7 (Bound of one-step update of Σ)**.** *Suppose* $\{K^{(t)}, \Sigma^{(t)}\}_{t=0}^\infty$ *follows the update rule in* (IPO)*. Then we have* $\|\Sigma^{(t+1)} - \Sigma^*\| \leq \frac{\tau\gamma\|B\|^2}{\sigma_{\min}(R)} \|P_{K^{(t)}} - P_{K^*}\|$.

Proof. Observe from (IPO) and (4.5) that

$$\begin{aligned} \left\|\Sigma^{(t+1)} - \Sigma^*\right\| &= \frac{\tau}{2} \left\|(R + \gamma B^\top P_{K^{(t)}} B)^{-1} - (R + \gamma B^\top P_{K^*} B)^{-1}\right\| \\ &= \frac{\tau}{2} \left\|(R + \gamma B^\top P_{K^{(t)}} B)^{-1} \cdot \gamma B^\top \left(P_{K^{(t)}} - P_{K^*}\right) B \cdot (R + \gamma B^\top P_{K^*} B)^{-1}\right\| \\ &\leq \frac{\tau\gamma\|B\|^2}{2\sigma_{\min}(R)^2} \left\|P_{K^{(t)}} - P_{K^*}\right\|. \end{aligned}$$

□

Lemma 4.9.8 perform perturbation analysis on P_K and establish bounds for the difference in P_K with respect to the perturbation in K. Consequently, both $\|K^{(t+1)} - K^*\|$ and $\|\Sigma^{(t+1)} - \Sigma^*\|$ (in Lemma 4.9.6 and Lemma 4.9.7) can be bounded in terms of $\|K^{(t)} - K^*\|$.

Lemma 4.9.8 (P_K perturbation)**.** *For any* $K \in \Omega$*, with* c *defined in* (4.23)*,* $\|P_K - P_{K^*}\| \leq c\|K - K^*\|$.

Proof. Fix $K \in \Omega$. By (4.58) and (4.16),

$$\mathcal{T}_K^{-1}(P_K) = (I - \gamma\mathcal{F}_K)(P_K) = P_K - \gamma(A - BK)^\top P_K(A - BK) = Q + K^\top RK,$$

which immediately implies $P_K = \mathcal{T}_K(Q + K^\top RK)$. Similarly $P_{K^*} = \mathcal{T}_{K^*}(Q + {K^*}^\top RK^*)$. To bound the difference between P_K and P_{K^*}, observe that,

$$\|P_K - P_{K^*}\| \leq \left\|\mathcal{T}_K\left(Q + K^\top RK\right) - \mathcal{T}_{K^*}\left(Q + K^\top RK\right)\right\| + \|\mathcal{T}_{K^*}\|\|{K^*}^\top RK^* - K^\top RK\|. \tag{4.62}$$

For the first term in (4.62), we can apply Lemma 4.9.5 and Lemma 4.9.4 to get

$$\begin{aligned} &\left\|\mathcal{T}_K\left(Q + K^\top RK\right) - \mathcal{T}_{K^*}\left(Q + K^\top RK\right)\right\| \\ \leq&\xi_{\gamma\rho}\|\mathcal{F}_{K^*} - \mathcal{F}_K\|\|Q + K^\top RK\| \leq 2\rho\xi_{\gamma\rho}\|B\|\,\|K^* - K\| \cdot \left(\|Q\| + \|R\|\|K\|^2\right). \end{aligned} \tag{4.63}$$

For the second term in (4.62), note that by Lemma 17 in [53], $\|\mathcal{T}_K\| \leq \frac{1}{\mu}\|\mathcal{T}_K(\mathbb{E}[x_0x_0^\top])\|$. Since $S_{K,\Sigma} \succeq \mathcal{T}_K(\mathbb{E}[x_0x_0^\top])$, thus $\|\mathcal{T}_K\| \leq \frac{1}{\mu}\sigma_{max}\left(\mathcal{T}_K(\mathbb{E}[x_0x_0^\top])\right) \leq \frac{1}{\mu}\|S_{K,\Sigma}\|$. Then we have

$$\begin{aligned} &\|\mathcal{T}_{K^*}\|\|{K^*}^\top RK^* - K^\top RK\| \\ =&\|\mathcal{T}_{K^*}\|\|{K^*}^\top RK^* - {K^*}^\top RK + {K^*}^\top RK - K^\top RK\| \\ \leq&\|\mathcal{T}_{K^*}\|\|R\|\|K - K^*\|\left(\|K^*\| + \|K\|\right) \\ \leq&\frac{\|S_{K^*,\Sigma^*}\|}{\mu}\|R\|\|K - K^*\|\left(\|K^*\| + \|K\|\right). \end{aligned} \tag{4.64}$$

Plugging (4.63), (4.64), and (4.23) in (4.62) completes the proof. □

With these lemmas, the proof of Lemma 4.5.4 is completed as follows:

Proof. (of Lemma 4.5.4). Let K' denote $K^{(t+1)}$ and K denote $K^{(t)}$. With the assumption that $\|A - BK'\| \leq \rho$, $\|A - BK^*\| \leq \rho$, we can apply Lemma 4.5.3 to get

$$\begin{aligned} \|S_{K^*,\Sigma^*} - S_{K',\Sigma'}\| \leq& \omega_{\gamma,\rho}\|B\|^2\|\Sigma^* - \Sigma'\| \\ &+ \left(\xi_{\gamma,\rho} \cdot \|\mathbb{E}[x_0x_0^\top]\| + \zeta_{\gamma,\rho} \cdot \|B\Sigma^*B^\top + W\|\right) \cdot 2\rho\|B\|\,\|K^* - K'\|. \end{aligned} \tag{4.65}$$

Apply Lemma 4.9.6 and Lemma 4.9.8 to get

$$\begin{aligned} \|K' - K^*\| \leq& \Big(1 + \sigma_{\min}(R) \cdot \|\gamma B^\top P_{K^*}B + R\| + c\gamma\sigma_{\min}(R)\left(\|B\|\|A\| + \|B\|^2\kappa\right)\Big) \\ &\cdot \|K - K^*\|. \end{aligned}$$

Apply Lemma 4.9.7 and 4.9.8 to get $\|\Sigma' - \Sigma^*\| \leq \frac{\tau\gamma\|B\|^2\|P_K - P_{K^*}\|}{2\sigma_{\min}(R)^2} \leq c\frac{\tau\gamma\|B\|^2\|K - K^*\|}{2\sigma_{\min}(R)^2}$. Finally, plugging the above two inequalities into (4.65) finishes the proof. □

Bibliography

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR, 2020.

[2] Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, pages 151–160. PMLR, 2019.

[3] Franklin Allen and Stephen Morris. Game theory models in finance. In *Game theory and business applications*, pages 17–41. Springer, 2013.

[4] Berkay Anahtarcı, Can Deha Karıksız, and Naci Saldi. Learning in discounted-cost and average-cost mean-field games. *arXiv preprint arXiv:1912.13309*, 2019.

[5] Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Lauriere. Reinforcement learning for mean field games, with applications to economics. *arXiv preprint arXiv:2106.13755*, 2021.

[6] Gürdal Arslan and Serdar Yüksel. Decentralized Q-learning for stochastic teams and games. *IEEE Transactions on Automatic Control*, 62(4):1545–1558, 2016.

[7] Robert J Aumann and Sergiu Hart. *Handbook of game theory with economic applications*, volume 2. Elsevier, 1992.

[8] Alexander Aurell and Boualem Djehiche. Mean-field type modeling of nonlocal crowd aversion in pedestrian crowd dynamics. *SIAM Journal on Control and Optimization*, 56(1):434–455, 2018.

[9] Alexander Aurell, René Carmona, Gökçe Dayanıklı, and Mathieu Lauriere. Finite state graphon games with applications to epidemics. *Dynamic Games and Applications*, 12 (1):49–81, 2022.

[10] Alexander Aurell, René Carmona, and Mathieu Lauriere. Stochastic graphon games: II. the linear-quadratic case. *Applied Mathematics & Optimization*, 85(3):39, 2022.

[11] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. *Advances in Neural Information Processing Systems*, 31, 2018.

[12] David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *International Conference on Machine Learning*, pages 354–363. PMLR, 2018.

[13] Matteo Basei, Xin Guo, Anran Hu, and Yufei Zhang. Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *The Journal of Machine Learning Research*, 23(1):8015–8048, 2022.

[14] Lucas Baudin and Rida Laraki. Best-response dynamics and fictitious play in identical-interest and zero-sum stochastic games. *arXiv preprint arXiv:2111.04317*, 2021.

[15] Erhan Bayraktar, Ruoyu Wu, and Xin Zhang. Propagation of chaos of forward–backward stochastic differential equations with graphon interactions. *Applied Mathematics & Optimization*, 88(1):25, 2023.

[16] Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684, 1957.

[17] Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 4. Athena scientific, 2012.

[18] Dimitri Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming.* Athena Scientific, 1996.

[19] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

[20] Mark Broom and Jan Rychtář. *Game-theoretical models in biology.* Chapman and Hall/CRC, 2022.

[21] Jingjing Bu, Afshin Mesbahi, Maryam Fazel, and Mehran Mesbahi. Lqr through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019.

[22] Jingjing Bu, Afshin Mesbahi, and Mehran Mesbahi. Policy gradient-based algorithms for continuous-time linear quadratic control. *arXiv preprint arXiv:2006.09178*, 2020.

[23] Peter E Caines and Minyi Huang. Graphon mean field games and the gmfg equations: ε-nash equilibria. In *2019 IEEE 58th conference on decision and control (CDC)*, pages 286–292. IEEE, 2019.

[24] Ozan Candogan, Ishai Menache, Asuman Ozdaglar, and Pablo A Parrilo. Flows and decompositions of games: Harmonic and potential games. *Mathematics of Operations Research*, 36(3):474–503, 2011.

[25] Ozan Candogan, Asuman Ozdaglar, and Pablo A Parrilo. Learning in near-potential games. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 2428–2433. IEEE, 2011.

[26] Ozan Candogan, Asuman Ozdaglar, and Pablo A Parrilo. Dynamics in near-potential games. *Games and Economic Behavior*, 82:66–90, 2013.

[27] Ozan Candogan, Asuman Ozdaglar, and Pablo A Parrilo. Near-potential games: Geometry and dynamics. *ACM Transactions on Economics and Computation (TEAC)*, 1(2):1–32, 2013.

[28] Haoyang Cao, Haotian Gu, and Xin Guo. Feasibility of transfer learning: A mathematical framework. *arXiv preprint arXiv:2305.12985*, 2023.

[29] Haoyang Cao, Haotian Gu, Xin Guo, and Mathieu Rosenbaum. Risk of transfer learning and its applications in finance, 2023.

[30] Pierre Cardaliaguet and Saeed Hadikhanloo. Learning in mean field games: the fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations*, 23(2):569–591, 2017.

[31] René Carmona. *Lectures on BSDEs, stochastic control, and stochastic differential games with financial applications*. SIAM, 2016.

[32] René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications I: Mean Field FBSDEs, Control, and Games*, volume 83. Springer, 2018.

[33] René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications II: Mean Field Games with Common Noise and Master Equations*, volume 84. Springer, 2018.

[34] René Carmona, Jean-Pierre Fouque, and Li-Hsien Sun. Mean field games and systemic risk. *arXiv preprint arXiv:1308.2172*, 2013.

[35] René Carmona, Mathieu Laurière, et al. Deep learning for mean field games and mean field control with applications to finance. *arXiv preprint arXiv:2107.04568*, 7, 2021.

[36] René Carmona, Daniel B Cooney, Christy V Graves, and Mathieu Lauriere. Stochastic graphon games: I. the static case. *Mathematics of Operations Research*, 47(1):750–778, 2022.

[37] Rene Carmona, Quentin Cormier, and H Mete Soner. Synchronization in a kuramoto mean field game. *Communications in Partial Differential Equations*, 48(9):1214–1244, 2023.

[38] Álvaro Cartea, Patrick Chang, José Penalva, and Harrison Waldon. Algorithms can learn to collude: A folk theorem from learning with bounded rationality. *Available at SSRN 4293831*, 2022.

[39] Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34:27952–27964, 2021.

[40] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022.

[41] Shuyu Cheng, Guoqiang Wu, and Jun Zhu. On the convergence of prior-guided zeroth-order optimization algorithms. *Advances in Neural Information Processing Systems*, 34:14620–14631, 2021.

[42] Andrew M Colman. *Game theory and its applications: In the social and biological sciences*. Psychology Press, 2013.

[43] Andrea Cosso and Huyên Pham. Zero-sum stochastic differential games of generalized McKean–Vlasov type. *Journal de Mathématiques Pures et Appliquées*, 129:180–212, 2019.

[44] Andrea Cosso, Fausto Gozzi, Idris Kharroubi, Huyên Pham, and Mauro Rosestolato. Optimal control of path-dependent McKean–Vlasov SDEs in infinite-dimension. *The Annals of Applied Probability*, 33(4):2863–2918, 2023.

[45] Qiwen Cui, Zhihan Xiong, Maryam Fazel, and Simon S Du. Learning in congestion games with bandit feedback. *arXiv preprint arXiv:2206.01880*, 2022.

[46] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *Communications of the ACM*, 52(2):89–97, 2009.

[47] Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540, 2020.

[48] Anna De Crescenzo, Marco Fuhrman, Idris Kharroubi, and Huyên Pham. Mean-field control of non exchangeable systems. *arXiv preprint arXiv:2407.18635*, 2024.

[49] Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pages 5166–5220. PMLR, 2022.

[50] Mao Fabrice Djete, Dylan Possamaï, and Xiaolu Tan. Mckean–vlasov optimal control: the dynamic programming principle. *The Annals of Probability*, 50(2):791–833, 2022.

[51] Benjamin Patrick Evans and Sumitra Ganesh. Learning and calibrating heterogeneous bounded rational market behaviour with multi-agent reinforcement learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 2024.

[52] Christian Fabian, Kai Cui, and Heinz Koeppl. Learning sparse graphon mean field games. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 4486–4514. PMLR, 25–27 Apr 2023. URL `https://proceedings.mlr.press/v206/fabian23a.html`.

[53] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pages 1467–1476. PMLR, 2018.

[54] CS Fisk. Game theory and transportation systems modelling. *Transportation Research Part B: Methodological*, 18(4-5):301–313, 1984.

[55] Wendell H. Fleming and H. Mete Soner. *Controlled Markov Processes and Viscosity Solutions*, volume 25 of *Stochastic Modelling and Applied Probability.* Springer, 2nd edition, 2006.

[56] Roy Fox, Stephen M Mcaleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in Markov potential games. In *AISTATS*, pages 4414–4425. PMLR, 2022.

[57] Drew Fudenberg and David K Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

[58] David Gamarnik. Correlation decay method for decision, optimization, and inference in large-scale networks. In *Theory Driven by Influential Applications*, pages 108–121. INFORMS, 2013.

[59] David Gamarnik, David A Goldberg, and Theophane Weber. Correlation decay in random decision networks. *Mathematics of Operations Research*, 39(2):229–261, 2014.

[60] Shuang Gao, Rinel Foguen Tchuendom, and Peter E Caines. Linear quadratic graphon field games. *arXiv preprint arXiv:2006.03964*, 2020.

[61] Benjamin Gravell, Peyman Mohajerin Esfahani, and Tyler Summers. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*, 66(11):5283–5298, 2020.

[62] Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 55(2):895–943, 2022.

[63] Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Dynamic programming principles for mean-field controls with learning. *Operations Research*, 2023.

[64] Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field multiagent reinforcement learning: A decentralized network approach. *Mathematics of Operations Research*, 50(1):506–536, 2025.

[65] Xin Guo and Jiacheng Zhang. Itô's formula for flows of conditional measures on semimartingales. *arXiv:2404.11167*, 2024.

[66] Xin Guo and Yufei Zhang. Towards an analytical framework for dynamic potential games. *arXiv preprint arXiv:2310.02259*, 2023.

[67] Xin Guo and Yufei Zhang. Towards an analytical framework for dynamic potential games. *arXiv:2310.02259*, 2024. URL `https://arxiv.org/abs/2310.02259`.

[68] Xin Guo, Anran Hu, and Jiacheng Zhang. Optimization frameworks and sensitivity analysis of stackelberg mean-field games. *arXiv preprint arXiv:2210.04110*, 2022.

[69] Xin Guo, Renyuan Xu, and Thaleia Zariphopoulou. Entropy regularization for mean field games with learning. *Mathematics of Operations research*, 47(4):3239–3260, 2022.

[70] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. A general framework for learning mean-field games. *Mathematics of Operations Research*, 48(2):656–686, 2023.

[71] Xin Guo, Xinyu Li, Chinmay Maheshwari, Shankar Sastry, and Manxi Wu. Markov α-potential games: Equilibrium approximation and regret analysis. *arXiv preprint arXiv:2305.12553*, 2023.

[72] Xin Guo, Xinyu Li, and Renyuan Xu. Fast policy learning for linear quadratic control with entropy regularization. *arXiv preprint arXiv:2311.14168*, 2023.

[73] Xin Guo, Huyên Pham, and Xiaoli Wei. Itô's formula for flows of measures on semimartingales. *Stochastic Processes and their applications*, 159:350–390, 2023.

[74] Xin Guo, Xinyu Li, and Yufei Zhang. An α-potential game framework for n-player games. *arXiv:2403.16962*, 2024.

[75] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR, 2018.

[76] Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59(5):3359–3391, 2021.

[77] Ben Hambly, Renyuan Xu, and Huining Yang. Recent advances in reinforcement learning in finance. *Mathematical Finance*, 33(3):437–503, 2023.

[78] Mohamed Hamdouche, Pierre Henry-Labordere, and Huyên Pham. Policy gradient learning methods for stochastic control with exit time and applications to share repurchase pricing. *Applied Mathematical Finance*, 29(6):439–456, 2022.

[79] Peter Hammerstein and Reinhard Selten. Game theory and evolutionary biology. *Handbook of game theory with economic applications*, 2:929–993, 1994.

[80] Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Policy gradient converges to the globally optimal policy for nearly linear-quadratic regulators. *arXiv preprint arXiv:2303.08431*, 2023.

[81] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR, 2019.

[82] Rainer Hettich and Kenneth O Kortanek. Semi-infinite programming: theory, methods, and applications. *SIAM review*, 35(3):380–429, 1993.

[83] Yaron Hollander and Joseph N Prashker. The applicability of non-cooperative game theory in transport analysis. *Transportation*, 33:481–496, 2006.

[84] S Hosseinirad, AA Porzani, G Salizzoni, and M Kamgarpour. General-sum finite-horizon potential linear-quadratic games with a convergent policy. *arXiv preprint arXiv:2305.13476*, 2023.

[85] Junling Hu and Michael P Wellman. Nash Q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

[86] Ruimeng Hu, Jihao Long, and Haosheng Zhou. Finite-agent stochastic differential games on large graphs: I. the linear-quadratic case. *arXiv preprint arXiv:2406.09523*, 2024.

[87] Minyi Huang, Roland P Malhamé, and Peter E Caines. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. *COMMUNICATIONS IN INFORMATION AND SYSTEMS*, 6(3):221–252, 2006.

[88] Christian Ibars, Monica Navarro, and Lorenza Giupponi. Distributed demand management in smart grid with a congestion game. In *2010 First IEEE International Conference on Smart Grid Communications*, pages 495–500. IEEE, 2010.

[89] Yanwei Jia and Xun Yu Zhou. Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *The Journal of Machine Learning Research*, 23(1):12603–12652, 2022.

[90] Yanwei Jia and Xun Yu Zhou. q-learning in continuous time. *J. Mach. Learn. Res.*, 24:161–1, 2023.

[91] Zeyu Jin, Johann Michael Schmitt, and Zaiwen Wen. On the analysis of model-free methods for the linear quadratic regulator. *arXiv preprint arXiv:2007.03861*, 2020.

[92] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[93] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[94] Talha Kavuncu, Ayberk Yaraneri, and Negar Mehr. Potential ilqr: A potential-minimizing controller for planning multi-agent interactive trajectories. *arXiv preprint arXiv:2107.04926*, 2021.

[95] Aimé Lachapelle and Marie-Therese Wolfram. On a mean field game approach modeling congestion and aversion in pedestrian crowds. *Transportation research part B: methodological*, 45(10):1572–1589, 2011.

[96] Daniel Lacker and Agathe Soret. A case study on stochastic games on large graphs in mean field and sparse regimes. *Mathematics of Operations Research*, 47(2):1530–1565, 2022.

[97] Daniel Lacker and Agathe Soret. A label-state formulation of stochastic graphon games and approximate equilibria on large networks. *Mathematics of Operations Research*, 2022.

[98] Jeffrey Larson, Matt Menickelly, and Stefan M Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.

[99] Jean-Michel Lasry and Pierre-Louis Lions. Jeux à champ moyen. II–Horizon fini et contrôle optimal. *Comptes Rendus. Mathématique*, 343(10):679–684, 2006.

[100] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

[101] Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in Markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

[102] David S Leslie and Edmund J Collins. Individual q-learning in normal form games. *SIAM Journal on Control and Optimization*, 44(2):495–514, 2005.

[103] Mushuang Liu, Ilya Kolmanovsky, H Eric Tseng, Suzhou Huang, Dimitar Filev, and Anouck Girard. Potential game-based decision-making for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 24(8):8014–8027, 2023.

[104] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[105] Sergio Valcarcel Macua, Javier Zazo, and Santiago Zazo. Learning parametric closed-loop policies for Markov potential games. *arXiv preprint arXiv:1802.00899*, 2018.

[106] Chinmay Maheshwari, Manxi Wu, Druv Pai, and Shankar Sastry. Independent and decentralized learning in markov potential games. *arXiv preprint arXiv:2205.14590*, 2022.

[107] Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *The 22nd international conference on artificial intelligence and statistics*, pages 2916–2925. PMLR, 2019.

[108] Weichao Mao, Tamer Başar, Lin F Yang, and Kaiqing Zhang. Decentralized cooperative multi-agent reinforcement learning with exploration. *arXiv preprint arXiv:2110.05707*, 2021.

[109] Eric Mazumdar, Lillian J. Ratliff, Michael I. Jordan, and S. Shankar Sastry. Policy-gradient algorithms have no guarantees of convergence in linear quadratic games. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pages 860–868, 2020. ISBN 9781450375184.

[110] 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.

[111] Panayotis Mertikopoulos and William H Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016.

[112] Dov Monderer and Lloyd S Shapley. Fictitious play property for games with identical interests. *Journal of economic theory*, 68(1):258–265, 1996.

[113] Dov Monderer and Lloyd S Shapley. Potential games. *Games and economic behavior*, 14(1):124–143, 1996.

[114] Dheeraj Narasimha, Kiyeob Lee, Dileep Kalathil, and Srinivas Shakkottai. Multi-agent learning via Markov potential games in marketplaces for distributed energy resources. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 6350–6357. IEEE, 2022.

[115] John Nash. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

[116] Shuteng Niu, Yongxin Liu, Jian Wang, and Houbing Song. A decade survey of transfer learning (2010–2020). *IEEE Transactions on Artificial Intelligence*, 1(2):151–166, 2020.

[117] Christos H Papadimitriou. The complexity of finding Nash equilibria. *Algorithmic game theory*, 2:30, 2007.

[118] Jan Peters, Katharina Mulling, and Yasemin Altun. Relative entropy policy search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1):1607–1612, 2010.

[119] Huyên Pham and Xiaoli Wei. Dynamic programming for optimal control of stochastic McKean–Vlasov dynamics. *SIAM Journal on Control and Optimization*, 55(2):1069–1101, 2017.

[120] Philipp Plank and Yufei Zhang. Policy optimization for continuous-time linear-quadratic graphon mean field games. *arXiv preprint arXiv:2506.05894*, 2025.

[121] Robert W Rosenthal. A class of games possessing pure-strategy nash equilibria. *International journal of game theory*, 2(1):65–67, 1973.

[122] Walter Rudin et al. *Principles of Mathematical Analysis*, volume 3. McGraw-hill New York, 1976.

[123] Larry Samuelson. Game theory in economics and beyond. *Journal of Economic Perspectives*, 30(4):107–130, 2016.

[124] William H Sandholm. *Population games and evolutionary dynamics*. MIT press, 2010.

[125] Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Basar, and Asuman Ozdaglar. Decentralized q-learning in zero-sum Markov games. *Advances in Neural Information Processing Systems*, 34, 2021.

[126] Katya Scheinberg. Derivative free optimization method. *Department Computing and Software, McMaster University*, 2000.

[127] Andrew Schotter and Gerhard Schwödiauer. Economics and the theory of games: a survey. *Journal of Economic Literature*, 18(2):479–527, 1980.

[128] Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5668–5675, 2020.

[129] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[130] Samuel Sokota, Ryan D'Orazio, J. Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *ICLR*, 2023. URL `https://arxiv.org/abs/2206.05825`.

[131] Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum Markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

[132] Jingrui Sun and Jiongmin Yong. Linear quadratic stochastic differential games: open-loop and closed-loop saddle points. *SIAM Journal on Control and Optimization*, 52 (6):4082–4121, 2014.

[133] Jingrui Sun, Xun Li, and Jiongmin Yong. Open-loop and closed-loop solvabilities for stochastic linear quadratic optimal control problems. *SIAM Journal on Control and Optimization*, 54(5):2274–2308, 2016.

[134] Lingfeng Sun, Pin-Yun Hung, Changhao Wang, Masayoshi Tomizuka, and Zhuo Xu. Distributed multi-agent interaction generation with imagined potential games. *arXiv preprint arXiv:2310.01614*, 2023.

[135] Lingfeng Sun, Yixiao Wang, Pin-Yun Hung, Changhao Wang, Xiang Zhang, Zhuo Xu, and Masayoshi Tomizuka. Imagined potential games: A framework for simulating, learning and evaluating interactive behaviors. *arXiv preprint arXiv:2411.03669*, 2024.

[136] Youbang Sun, Tao Liu, Ruida Zhou, PR Kumar, and Shahin Shahrampour. Provably fast convergence of independent natural policy gradient for Markov potential games. *NeurIPS*, 2023.

[137] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[138] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

[139] 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.

[140] Lukasz Szpruch, Tanut Treetanthiploet, and Yufei Zhang. Optimal scheduling of entropy regulariser for continuous-time linear-quadratic reinforcement learning. *arXiv preprint arXiv:2208.04466*, 2022.

[141] Vladislav B Tadić, Sean P Meyn, and Roberto Tempo. Randomized algorithms for semi-infinite programming problems. In *2003 European Control Conference (ECC)*, pages 3011–3015. IEEE, 2003.

[142] Wenpin Tang, Yuming Paul Zhang, and Xun Yu Zhou. Exploratory hjb equations and their convergence. *SIAM Journal on Control and Optimization*, 60(6):3191–3216, 2022.

[143] Anjan V Thakor. Game theory in finance. *Financial Management*, pages 71–94, 1991.

[144] Anastasios Tsiamis, Dionysios S Kalogerias, Luiz FO Chamon, Alejandro Ribeiro, and George J Pappas. Risk-constrained linear-quadratic regulators. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3040–3047. IEEE, 2020.

[145] Balint Varga. On the upper bound of near potential differential games. *arXiv preprint arXiv:2307.03010*, 2023.

[146] Balint Varga, Jairo Inga, and Sören Hohmann. Limited information shared control: A potential game approach. *IEEE Transactions on Human-Machine Systems*, 53(2): 282–292, 2022.

[147] John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, Princeton, NJ, USA, 1944.

[148] Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *The Journal of Machine Learning Research*, 21(1):8145–8178, 2020.

[149] Haoran Wang, Thaleia Zariphopoulou, and Xunyu Zhou. Exploration versus exploitation in reinforcement learning: A stochastic control approach. *Journal of Machine Learning Research*, 21:1–34, 2020.

[150] Zifan Wang, Yulong Gao, Siyi Wang, Michael M Zavlanos, Alessandro Abate, and Karl Henrik Johansson. Policy evaluation in distributional lqr. In *Learning for Dynamics and Control Conference*, pages 1245–1256. PMLR, 2023.

[151] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[152] Ronald J Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.

[153] Jiongmin Yong and Jianfeng Zhang. Non-equivalence of stochastic optimal control problems with open and closed loop controls. *Systems & Control Letters*, 153:104948, 2021.

[154] Jiongmin Yong and Xun Yu Zhou. *Stochastic Controls: Hamiltonian Systems and HJB Equations*, volume 43 of *Applications of Mathematics.* Springer, 1999.

[155] Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel. Decentralized learning for optimality in stochastic dynamic teams and games with local control and global state information. *IEEE Transactions on Automatic Control*, 67, 2021.

[156] Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel. Satisficing paths and independent multiagent reinforcement learning in stochastic games. *SIAM Journal on Mathematics of Data Science*, 5, 2023.

[157] H Peyton Young. *Strategic learning and its limits*. OUP Oxford, 2004.

[158] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

[159] He Zhang, Yuelong Su, Lihui Peng, and Danya Yao. A review of game theory applications in transportation analysis. In *2010 international conference on computer and information application*, pages 152–157. IEEE, 2010.

[160] Jianfeng Zhang. *Backward stochastic differential equations*. Springer, 2017.

[161] Kaiqing Zhang, Bin Hu, and Tamer Basar. Policy optimization for h_2 linear control with h_∞ robustness guarantee: Implicit regularization and global convergence. *SIAM Journal on Control and Optimization*, 59(6):4081–4109, 2021.

[162] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.

[163] Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*, 2021.

[164] Runyu Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, and Na Li. On the global convergence rates of decentralized softmax gradient play in markov potential games. *Advances in Neural Information Processing Systems*, 35:1923–1935, 2022.

[165] Runyu Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, and Na Li. On the global convergence rates of decentralized softmax gradient play in Markov potential games. In *Advances in Neural Information Processing Systems*, 2022.

[166] Feiran Zhao, Keyou You, and Tamer Başar. Global convergence of policy gradient primal-dual methods for risk-constrained lqrs. *IEEE Transactions on Automatic Control*, 2023.

[167] Zhe Zhou, Scott J Moura, Hongcai Zhang, Xuan Zhang, Qinglai Guo, and Hongbin Sun. Power-traffic network equilibrium incorporating behavioral theory: A potential game perspective. *Applied Energy*, 289:116703, 2021.

[168] Zhuangdi Zhu, Kaixiang Lin, Anil K Jain, and Jiayu Zhou. Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.