Convex Markov Games and Beyond: New Proof of Existence, Characterization and Learning Algorithms for Nash Equilibria

Anas Barakat[1]
SUTD

Ioannis Panageas[2]
UC Irvine

Antonios Varvitsiotis[134]
SUTD, CQT NUS, Archimedes

Abstract

Convex Markov Games (cMGs) were recently introduced as a broad class of multi-agent learning problems that generalize Markov games to settings where strategic agents optimize general utilities beyond additive rewards. While cMGs expand the modeling frontier, their theoretical foundations, particularly the structure of Nash equilibria (NE) and guarantees for learning algorithms, are not yet well understood. In this work, we address these gaps for an extension of cMGs, which we term General Utility Markov Games (GUMGs), capturing new applications requiring coupling between agents' occupancy measures. We prove that in GUMGs, Nash equilibria coincide with the fixed points of projected pseudo-gradient dynamics (i.e., first-order stationary points), enabled by a novel agent-wise gradient domination property. This insight also yields a simple proof of NE existence using Brouwer's fixed-point theorem. We further show the existence of Markov *perfect* equilibria. Building on this characterization, we establish a policy gradient theorem for GUMGs and design a model-free policy gradient algorithm. For potential GUMGs, we establish iteration complexity guarantees for computing approximate-NE under exact gradients and provide sample complexity bounds in both the generative model and on-policy settings. Our results extend beyond prior work restricted to zero-sum cMGs, providing the first theoretical analysis of common-interest cMGs.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) is classically modeled by Markov games (Shapley, 1953; Fink, 1964; Littman, 1994), a direct generalization of Markov Decision Processes (MDPs) (Howard, 1960; Puterman, 2014). While widely used, this additive-reward framework fails to capture key objectives such as risk sensitivity, imitation, exploration, or fairness.

In the last few years, to go beyond additive rewards in MDPs, convex RL (Zhang et al., 2020; Zahavy et al., 2021) has emerged as a general and unifying framework where the goal of an agent is to optimize a concave functional of its occupancy measure (i.e. the frequency of their state-action visitations under the selected policy). In particular, convex RL captures a variety of applications including pure exploration (Hazan et al., 2019), imitation learning (Ho and Ermon, 2016), diverse skill discovery (Eysenbach et al., 2019) and experiment design (Mutny et al., 2023) to name a few.

Inspired by single-agent convex RL, Gemp et al. (2025) recently introduced convex Markov games (cMGs). Extending the scope of classical Markov games for MARL, cMGs provide a flexible unifying framework to model strategic interactions in MARL with general utilities beyond additive rewards. This framework unifies a wide array of MARL problems, from preference alignment (Wu et al., 2024) and task-agnostic exploration (Zamboni et al., 2025) to risk-sensitive sequential games (Bhatt and Sobel, 2025) and fairness-aware MARL (Hughes et al., 2018).

While cMGs extend the modeling frontier, their theoretical foundations remain poorly understood. Gemp et al. (2025) proved existence of Nash equilibria (NE) and proposed approximation methods based on regularization annealing, but required full model knowl-

[1]Singapore University of Technology and Design
[2]University of California Irvine, USA
[3]Centre for Quantum Technologies, National University of Singapore, Singapore
[4]Archimedes/Athena Research Center, Greece

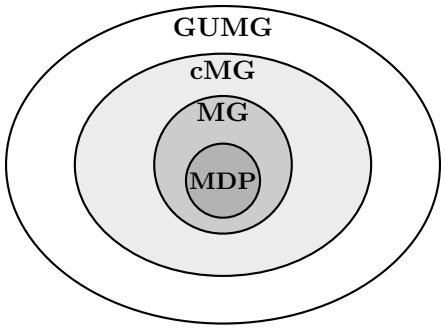

Figure 1: From single-agent RL to General Utility Markov Games (GUMG). MDP: Markov Decision Processes (single-agent RL), MG: Markov Games (MARL), cMG: convex Markov Games.

edge and left equilibrium structure uncharacterized. More recently, Kalogiannis et al. (2025) analyzed the zero-sum subclass of cMGs, but efficient algorithms for general or common-interest cMGs remain largely open.

This raises three fundamental questions: (i) what is the structure of NE in general cMGs beyond existence, (ii) can we design efficient policy gradient algorithms for such games, and (iii) can we learn NE sample-efficiently in the common-interest setting, where the agents' utilities have a potential structure?

1.1 Our Contributions

In this work, we address these questions for an extension of cMGs. Our key contributions are:

- **Framework.** We introduce the new class of General Utility Markov Games (GUMGs) as a natural extension of cMGs. This class strictly generalizes MDPs, Markov games, and cMGs (see Fig. 1).
- **Existence and characterization of NE.** We prove that NE in GUMGs coincide with fixed points of projected pseudo-gradient dynamics. This yields a conceptually simple proof of equilibrium existence via a novel agent-wise gradient domination property of utilities in GUMGs.
- **Policy gradient theorem and algorithm.** Motivated by this characterization, we establish a multi-agent policy gradient theorem for GUMGs and leverage it to design a principled model-free policy gradient (PG) algorithm (Algorithm 1).
- **Learning NE in potential GUMGs.** In potential GUMGs where utilities have a potential structure, our PG algorithm finds an ε-approximate NE in $\mathcal{O}(\varepsilon^{-2})$ iterations with exact gradients. When the transition dynamics are unknown, the PG algorithm learns an ε-approximate NE with total sample complexity of $\mathcal{O}(\varepsilon^{-4})$ under a generative model setting and $\mathcal{O}(\varepsilon^{-5})$ when only having access to on-policy sampled trajectories. To our knowledge, these are the first efficient NE learning guarantees for common-interest cMGs, beyond zero-sum cMGs.

1.2 Technical Overview

Obstacles beyond Markov games. Two obstacles prevent the direct extension of classical MG analysis to cMGs and our GUMGs. First, best response sets are non-convex in the policy space. Second, value functions and Bellman contraction arguments are not immediately available since the utilities are non-additive, unlike in Markov games. This hinders the application of Kakutani's fixed point theorem as used for the existence of NE in Markov games (Fink, 1964). Prior work on cMGs (Gemp et al., 2025) established existence via Debreu's theorem (using contractible sets) (Debreu, 1952) to bypass the lack of convexity of best response sets, but provided no structural characterization of equilibria and assumed full model knowledge for (biased) approximate computation.

Technical tool: per-agent gradient domination. We prove a per-agent gradient domination property which holds for our general class of N-agents GUMGs (hence in cMGs). This per-player property implies that any (approximate) first-order stationary policy is also an (approximate) best response for that player. This generalizes gradient domination arguments used in special cases (MDPs/MGs; Agarwal et al. 2021; Daskalakis et al. 2020; Leonardos et al. 2022; Zhang et al. 2024) to non-additive, multi-agent utilities.

Existence and first-order characterization via Brouwer. Per-agent gradient domination in GUMGs implies the *equivalence* between (a) NE, (b) first-order stationary points and (c) *fixed-points of the projected pseudo-gradient dynamics*. Applying *Brouwer's* fixed point theorem (rather than Kakutani or Debreu) to the projected pseudo-gradient map which is continuous on the compact policy space yields both existence and a structural characterization of NE in GUMGs (for any full-support initial state distribution), addressing the main theoretical gap left open by Gemp et al. (2025). Therefore our gradient domination result allows us to circumvent the first obstacle above to (i) show existence of NE in GUMGs for any full support initial state distribution and (ii) characterize them.

From NE to MPE. We further prove existence of Markov-perfect equilibria (MPE) beyond the mere existence of *non-perfect* NE (which depend on the initial state distribution). First, we show that the NE above

does not depend on the choice of the full-support initial distribution. Then, for initial distributions concentrating mass $1-\varepsilon$ on a single state, continuity of utilities in the initial distribution lets us pass to the limit $\varepsilon \to 0$, yielding an MPE. This generalizes the existence of MPE in Markov games (see e.g. Deng et al. (2023)) via a different proof technique.

PG theorem and model-free algorithm. Despite non-additivity of utilities, we exploit the dynamic programming structure of occupancy measures to establish a PG theorem for N-player GUMGs. This leads to a model-free PG algorithm which does not require full knowledge of the cMG transition model (in contrast with Gemp et al. 2025).

Sample-efficient learning in potential GUMGs. In the common interest setting, our analysis relies on showing smoothness of utilities w.r.t. policies in GUMGs, proving convergence to approximate stationary points using stochastic gradient ascent arguments and exploiting our gradient domination property to show convergence to approximate NE. Our sampling mechanism in the generative model setting differs from Zhang et al. (2024) as we use minibatch trajectory sampling *without* α-greedy policies, leading to an improved sample complexity of $\tilde{\mathcal{O}}(\varepsilon^{-4})$ instead of $\tilde{\mathcal{O}}(\varepsilon^{-6})$. In the on-policy setting, our proof technique relies on an analysis of stochastic projected gradient ascent with minibatch for smooth objectives, differing from the Moreau-envelope-based analysis used for the special case of MPGs in (Leonardos et al., 2022).

1.3 Related Work

Recent years have witnessed active research in theoretical MARL. In particular, an important body of work has focused on learning approximate NE in structured classes of Markov games such as zero-sum Markov games and Markov potential games. Our work extends these research efforts to cMGs and beyond. Recently, Gemp et al. (2025) introduced cMGs, proved existence of NE and proposed to approximate them using a *biased* entropic regularization approach under *full knowledge* of the transition model. In this work, we design a *model-free* PG algorithm for GUMGs and analyze its iteration and *sample complexity*. In a follow-up work, Kalogiannis et al. (2025) studied *2-player zero-sum cMGs*. In contrast, we characterize NE in any N-player general-sum GUMG (with cMGs as particular cases). Moreover, we study the common-interest setting beyond the competitive zero-sum two-player setting. See sections 3 and 4 for more details comparing our theoretical and algorithmic contributions to prior work. In the single-agent setting ($N=1$), convex Markov games reduce to convex RL which is an active research topic (Hazan et al., 2019; Zahavy et al., 2021; Zhang et al., 2020, 2021; Bai et al., 2022; Geist et al., 2022; Mutti et al., 2023; Barakat et al., 2023; Marin Moreno et al., 2024; De Santi et al., 2024; Barakat et al., 2025). A few works consider MARL with general utilities as a *decentralized policy optimization* problem (Zhang et al., 2022a; Ying et al., 2023). Our game-theoretic setting is fundamentally different. See Appendix 7 for an extended literature review.

2 FROM CLASSICAL TO GENERAL UTILITY MARKOV GAMES

In this section, we introduce useful notation, Markov games and cMGs to motivate our class of GUMGs.

Notation. We denote by $[N]$ the set of integers $\{1,\cdots,N\}$. For any vector $v \in \mathbb{R}^n$ where n is any integer larger than one, we denote its infinity norm by $\|v\|_\infty := \max_{i=1,\cdots,N} |v_i|$. For any finite set $\mathcal{X}$ with cardinality $X = |\mathcal{X}|$, we denote by $\Delta(\mathcal{X})$ the set of (discrete) probability distributions on $\mathcal{X}$. This set can be simply identified with the standard simplex, i.e. the set $\{\pi \in \mathbb{R}^X : \pi_i \geq 0, \forall i \in [X], \sum_{i=1}^X \pi_i = 1\}$.

2.1 Markov Games and Notation

A Markov game is defined via a 6-tuple:

$$\mathcal{G} = (\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, P, \{r_i\}_{i\in\mathcal{N}}, \mu, \gamma), \qquad (1)$$

where $\mathcal{N} := \{1,\cdots,N\}$ is a finite set of N agents, $\mathcal{S}$ is a finite set of states, $\mathcal{A}_i$ is a finite set of actions of cardinality $|\mathcal{A}_i|$ for all $i \in \mathcal{N}$. The joint action space is denoted by $\mathcal{A} := \prod_{i\in\mathcal{N}} \mathcal{A}_i$. Moreover, we denote by $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ the probability transition kernel: For any state $s \in \mathcal{S}$ and any joint action $a \in \mathcal{A}$, the game transitions from state s to a state $s' \in \mathcal{S}$ with probability $P(s'|s,a)$. The initial state probability distribution is denoted by $\mu \in \Delta(\mathcal{S})$ and $\gamma \in (0,1)$ is a discount factor. Each agent $i \in \mathcal{N}$ has their own reward function $r_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ assigning a reward $r_i(s,a)$ to any state $s \in \mathcal{S}$ and *joint* action $a \in \mathcal{A}$.

Policies. Each agent $i \in \mathcal{N}$ chooses their actions according to a randomized stationary policy denoted by $\pi_i \in \Pi_i := \Delta(\mathcal{A}_i)^{\mathcal{S}}$. The set of joint policies $\pi = (\pi_i)_{i\in\mathcal{N}}$ is denoted by $\Pi := \times_{i\in\mathcal{N}}\Pi_i$ and we further use the notation $\pi_{-i} = (\pi_j)_{j\in\mathcal{N}\setminus\{i\}} \in \Pi_{-i} := \times_{j\in\mathcal{N}\setminus\{i\}}\Pi_j$ for joint policies of all agents other than i. In this work, we will also be considering α-greedy policies which guarantee a level of exploration by ensuring that each action is selected with positive probability. Define for each agent $i \in \mathcal{N}$ the set of α-greedy policies $\Pi_i^\alpha := \{\pi_i^\alpha \in \Pi_i : \forall s \in \mathcal{S}, a_i \in \mathcal{A}_i, \pi_i^\alpha(a_i|s) := (1-\alpha)\pi_i(a_i|s) + \alpha/|\mathcal{A}_i|, \pi_i \in \Pi_i\}$ for $\alpha > 0$. The set of joint α-greedy policies is denoted by $\Pi^\alpha := \times_{i\in\mathcal{N}}\Pi_i^\alpha$.

At each time step $t \in \mathbb{N}$ in a state $s_t \in \mathcal{S}$, each agent $i \in \mathcal{N}$ chooses an action $a_{i,t} \in \mathcal{A}_i$ with probability $\pi_i(a_{i,t}|s_t)$ and then the environment transitions to a state $s_{t+1} \in \mathcal{S}$ with probability $P(s_{t+1}|s_t, a_t)$ where $a_t = (a_{i,t})_{i\in\mathcal{N}}$ is the joint action. We denote by $\mathbb{P}_{\mu,\pi}$ the probability distribution of the Markov chain $(s_t, a_t)_{t\in\mathbb{N}}$ induced by the joint policy π with initial state distribution μ. We use the notation $\mathbb{E}_{\mu,\pi}$ (or often simply $\mathbb{E}$) for the associated expectation.

Occupancy measures. We define for any policy $\pi \in \Pi$ the (normalized) state and state-joint-action occupancy measures $d^\pi_\mu \in \Delta(\mathcal{S}), \lambda^\pi_\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ which capture the frequency of state-action pairs visited:

$$\lambda^\pi_\mu(s, a) := (1-\gamma) \sum_{t=0}^{+\infty} \gamma^t \mathbb{P}_{\mu,\pi}(s_t = s, a_t = a)\,, \quad (2)$$

and $d^\pi_\mu(s) = \sum_{a\in\mathcal{A}} \lambda^\pi_\mu(s, a)$. Note that λ^π_μ will also be seen as a vector of the Euclidean space $\mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ with a slight abuse of notation. We also define for each agent $i \in \mathcal{N}$ the marginal occupancy measure induced by the joint policy π as follows for every $s \in \mathcal{S}, a_i \in \mathcal{A}_i$,

$$\lambda^\pi_{\mu,i}(s, a_i) := (1-\gamma) \sum_{t=0}^{+\infty} \gamma^t \mathbb{P}_{\mu,\pi}(s_t = s, a_{i,t} = a_i)\,. \quad (3)$$

Note that $\lambda^\pi_{\mu,i}(s, a_i) = \sum_{a_{-i}\in\mathcal{A}_{-i}} \lambda^\pi_\mu(s, a_i, a_{-i})$.

Value functions. Finally, we introduce value functions which inform on the quality of a given initial state(-action) in terms of cumulative discounted rewards under a given policy. For any standard MDP $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, r, \mu, \gamma)$ where $r \in \mathcal{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ refers to a reward function (i.e. $N = 1$ in (1)), we define the (action)-value functions for every $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$Q^\pi_{s,a}(r) := \mathbb{E}_{\rho,\pi}\left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, a_t)|s_0 = s, a_0 = a\right]$, and $V^\pi_\rho(r) = \sum_{a\in\mathcal{A}} \pi(a|s) Q^\pi_{s,a}(r)$.

2.2 Convex Markov Games

As introduced by Gemp et al. (2025), a convex Markov Game (cMG) is a 6-tuple $\mathcal{G} = (\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, P, \{F_i\}_{i\in\mathcal{N}}, \mu, \gamma)$, where $(\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, P, \mu, \gamma)$ is defined as in standard Markov games (see section 2.1 above). However, in contrast to Markov games, utilities of agents are not restricted to be defined as cumulative discounted rewards using some given fixed reward functions. Each agent $i \in \mathcal{N}$ has a utility function $F_i : \Lambda_i \times \Pi_{-i} \to \mathbb{R}$ where $\Lambda_i := \Delta(\mathcal{S} \times \mathcal{A}_i)$. The utility of agent i in the cMG under a joint policy $\pi \in \Pi$ is then given by:

$$u_{\mu,i}(\pi_i, \pi_{-i}) = F_i(\lambda^\pi_{\mu,i}, \pi_{-i})\,, \quad (4)$$

where the marginal occupancy measure $\lambda^\pi_{\mu,i}$ is defined in (3). Each utility function is supposed to be continuous and *concave* in player i's occupancy measure, and continuous in each policy π_j for $j \neq i$. Notice that the utility of agent i *does not* depend explicitly on the individual occupancy measures of *other agents*. Hence, coupling between occupancy marginals of agents is excluded. This motivates our more general definition which captures such couplings.

2.3 General Utility Markov Games

In this section, we introduce our new class of games which generalizes the class of cMGs.

Definition 1 (General Utility Markov Game)**.** *A General Utility Markov Game is defined via a 6-tuple* $(\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, P, \{F_i\}_{i\in\mathcal{N}}, \mu, \gamma)$ *where* $(\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, P, \mu, \gamma)$ *are defined as in section 2.2 and each real-valued function* F_i *is defined over the product space* $\Lambda \times \Pi_{-i}$ *where* $\Lambda := \prod_{i=1}^N \Lambda_i$ *and* $\Lambda_i := \Delta(\mathcal{S} \times \mathcal{A}_i)$*. The utility function* $u_i : \Pi \to \mathbb{R}$ *of agent* $i \in \mathcal{N}$ *is defined for every policy* $\pi = (\pi_1, \cdots, \pi_N) \in \Pi$ *by*

$$u_{\mu,i}(\pi_i, \pi_{-i}) := F_i(\lambda^\pi_{\mu,1}, \cdots, \lambda^\pi_{\mu,N}, \pi_{-i})\,. \quad (5)$$

Throughout this work we will suppose that F_i is differentiable and this implies that u_i is also differentiable. We will also make the following important assumption.

Assumption 1 (Concavity)**.** *For all* $i \in \mathcal{N}$*, the function* F_i *is jointly concave in* $(\lambda_1, \cdots, \lambda_N)$*.*

Why GUMGs? Assumption 1 captures cMGs, and therefore Markov games and convex RL ($N = 1$) as well. Compared to cMGs, GUMGs capture utilities that are concave in per-agent occupancies and *aggregates* thereof. See examples in the next section.

Remark 1. *Another possible and simple definition using the occupancy measure of the joint policy consists of defining* $u_{\mu,i}(\pi) = F_i(\lambda^\pi_\mu)$ *for any policy* $\pi \in \Pi$*. Our results in section 3 will also hold for this definition. However, this is a particular case of cMGs and leads to the need of estimating the joint occupancy measure which cannot be directly estimated by each agent independently (in contrast to marginal occupancies).*

2.4 Examples

Consider the following composite example defined for any $\lambda = (\lambda_1, \cdots, \lambda_N) \in \Lambda, \pi = (\pi_i, \pi_{-i}) \in \Pi$, by:

$$F_i(\lambda, \pi_{-i}) = \underbrace{-\alpha_i \mathrm{KL}(\lambda_i || q_i)}_{\text{imitation}} \underbrace{-\beta_i \mathrm{KL}(\lambda_i || \bar{\lambda}_N)}_{\text{consensus/diversity}} + \underbrace{\gamma_i h_i \left(\sum_{j=1}^N W_{i,j} \lambda_j\right)}_{\text{team aggregate}} + \underbrace{g_i(\pi_{-i})}_{\text{diversity penalty}}\,, \quad (6)$$

where $q_i \in \Lambda_i, i \in \mathcal{N}$ are fixed probability distributions, $\alpha_i, \beta_i, \gamma_i$ are nonnegative reals, $W \in \mathbb{R}^{N\times N}$ is a stochastic matrix of weights, and $\bar{\lambda}_N = \sum_{i=1}^N w_i\lambda_i$ with $w_i \geq 0, \sum_{i=1}^N w_i = 1$. Here, h_i and g_i are concave functions. Note that F_i satisfies Assumption 1 since it is the sum of jointly concave functions, noticing that KL divergence is jointly convex and the composition of a convex function with a linear one is also convex.

Notice that the above composite example for agent i's utility shows dependence on the occupancy measures of *all* agents via the aggregated terms $\bar{\lambda}_N$ and $\sum_{j=1}^N W_{i,j}\lambda_j$. This coupling between agents is *not* captured by the definition of cMGs proposed in Gemp et al. (2025), see (4). We now elaborate on special cases of our composite example (6):

a. Multi-agent imitation learning ($\alpha_i > 0, \beta_i = \gamma_i = 0$). Each agent matches its own occupancy to its expert occupancy q_i. See e.g. Song et al. (2018).

b. Consensus and diversity ($\beta_i > 0, \alpha_i = \gamma_i = 0$). The second term in (6) couples marginals via $\bar{\lambda}_N$ which can be computed from broadcast aggregates or neighbor averages without reconstructing the occupancy of the joint policy. Moreover, the term $g_i(\pi_{-i})$ can be interpreted as a policy diversity penalty. See e.g. Li et al. (2021) for MARL applications using information-theoretical regularization for behavioral diversity.

c. Population-level coordination ($\gamma_i > 0, \alpha_i = \beta_i = 0$). Using per-agent marginals, utilities depend on the population behavior or neighbor average. See mean-field MARL (Yang et al., 2018).

d. Team coverage and distribution matching ($F_i(\lambda_{1:N}, \pi_{-i}) = -\mathrm{KL}(\bar{\lambda}_N || \lambda_{\text{target}})$). This utility encourages matching team visitation with a desired target distribution λ_{target}. Important examples include patrolling, area coverage, search and target localization, see e.g. Mavrommati et al. (2017) in robotics.

e. Collective exploration ($\gamma_i > 0, \alpha_i = \beta_i = 0$). When h_i is the negative entropy, utility maximization enhances exploration. See e.g. Kim and Sung (2023); Zamboni et al. (2025).

Beyond GUMGs? There are individual objectives that do not satisfy our *joint* concavity assumption (e.g. $F_1(\lambda_1, \lambda_2, \pi_{-i}) = \lambda_1\lambda_2$ when $N = 2$). We leave the question of whether our theory can be extended to even more general functions $\{F_i\}_{i\in\mathcal{N}}$ to future work.

3 EXISTENCE AND STRUCTURE OF NASH POLICIES IN GUMG

Recall first the definition of an (approximate)-NE: a joint policy where no agent can improve its utility by deviating unilaterally. For any $\varepsilon \geq 0$, a policy $\pi^\star \in \Pi$ is an ε-approximate NE of a GUMG if for all $i \in \mathcal{N}$, all policies $\pi_i' \in \Pi_i$, $u_{\mu,i}(\pi_i', \pi_{-i}^\star) - u_{\mu,i}(\pi_i^\star, \pi_{-i}^\star) \leq \varepsilon$. When $\varepsilon = 0$, the policy $\pi^\star$ is a NE policy. A Markov perfect equilibrium (MPE) is a Markov policy that constitutes a NE for every initial state $s \in \mathcal{S}$, i.e., the definition of NE holds for any $\mu = \delta_s$ for MPE (where δ_s is the Dirac measure at state s) rather than only for the fixed initial state distribution μ defining the GUMG[1]. This means that the inequalities in the definition of NE hold for every distribution $\mu = \delta_s, s \in \mathcal{S}$ (and hence for any initial state distribution as well). See Appendix 8.3 for more details regarding the difference between NE and MPE and its relevance.

Our goal in this section is to show that in GUMGs, NE exist and admit a first-order characterization under Assumption 1, revealing their connection to projected pseudo-gradient dynamics. First, we make a standard exploration assumption to guarantee that all states are visited with positive probability in the long run.

Assumption 2. *For all $\pi \in \Pi, s \in \mathcal{S}$, $d_\mu^\pi(s) > 0$.*

Note that Assumption 2 is satisfied for instance when the initial state distribution μ has full support since we always have $d_\mu^\pi(s) \geq (1-\gamma)\mu(s)$ for all states $s \in \mathcal{S}$.

The utility of each agent in GUMGs is non-concave in their policy in general, as this is already the case in standard single-agent RL (see e.g. Lemma 1 in Agarwal et al. (2021)). Nevertheless, the following crucial result establishes a gradient domination property which holds per-agent. This structural property of GUMGs relates utility change under unilateral policy deviation to individual policy gradients.

Proposition 1 (Agentwise Gradient Domination)**.** *Let Assumptions 1 and 2 hold. For all $i \in \mathcal{N}$, all policies $\pi = (\pi_i, \pi_{-i}) \in \Pi$, $\pi_i' \in \Pi_i$, we have for any distribution $\mu, \rho \in \Delta(\mathcal{S})$ satisfying Assumption 2,*

$$u_{\mu,i}(\pi_i', \pi_{-i}) - u_{\mu,i}(\pi_i, \pi_{-i}) \leq C_{\mathcal{G}} \cdot \max_{\tilde{\pi}_i \in \Pi_i} \langle \nabla_{\pi_i} u_{\rho,i}(\pi), \tilde{\pi}_i - \pi_i \rangle ,$$

where the minimax distribution mismatch coefficient: $C_{\mathcal{G}} := \max_{i\in\mathcal{N}} \max_{\pi\in\Pi} \min_{\pi_i^\star \in \Pi_i^\star(\pi_{-i})} \left\| \frac{d_\mu^{\pi_i^\star, \pi_{-i}}}{d_\rho^{\pi_i, \pi_{-i}}} \right\|_\infty$, *where $\Pi_i^\star(\pi_{-i})$ denotes the set of best response policies for agent i, is finite.*

Prop. 1 generalizes the per-agent gradient domination property shown in Daskalakis et al. (2020) for 2-player zero-sum Markov games, in Leonardos et al. (2022) for Markov potential games and in Zhang et al. (2024); Giannou et al. (2022) for general-sum Markov games.

[1]Note that utilities are indexed by the initial state distribution to highlight this dependence.

Remark 2. *The mismatch coefficient $C_{\mathcal{G}}$ characterizes hardness of state exploration. This constant can be loosely bounded by $1/((1-\gamma)\min_{s\in\mathcal{S}}\rho(s))$ when ρ has full support, since for any policy $\pi \in \Pi, (1-\gamma)\rho(s) \leq d^{\pi}_{\rho}(s) \leq 1$ for all $s \in \mathcal{S}$. Note also, as mentioned in Daskalakis et al. (2020, Proposition 1), that finiteness of the minimax coefficient $C_{\mathcal{G}}$ is in general weaker than finiteness of the concentrability coefficient $\max_{\pi\in\Pi}\left\|\frac{d^{\pi}_{\mu}}{\mu}\right\|_{\infty}$ which has been used in prior work in Markov games (Leonardos et al., 2022; Ding et al., 2022; Zhang et al., 2024; Alatur et al., 2024).*

Recently, a similar result was proved in (Barakat et al., 2025, Theorem 1) for the *single-agent* case and in (Kalogiannis et al., 2025, Lemma 2.5) for *2-player* convex Markov games. Note that (i) our result holds for any number N of players and for any GUMG beyond 2-player zero-sum cMGs, and (ii) our bound is tighter as it uses a distribution mismatch coefficient which matches the two-player Markov game gradient domination result of Daskalakis et al. (2020). In particular, our proof technique differs from Kalogiannis et al. (2025) which rather relies on hidden convexity (Fatkhullin et al., 2025). Instead, we leverage the single agent gradient domination RL result of Agarwal et al. (2021) in our setting and make use of our concavity assumption. The first step in the proof uses concavity to relate utility deviation to standard single-agent RL value function deviation using pseudo-rewards (by defining a suitable MDP to reduce general utility to standard RL). The second step applies single-agent gradient domination for standard RL when the policies of other players' but i are fixed, and uses our policy gradient expression to recover the first-order variation in the desired inequality. See proof in Appendix 9.1.

In what follows, we use the agent-wise gradient domination property of utilities (Proposition 1) to relate NE to fixed points of the projected pseudo-gradient dynamics. First, we introduce the *pseudo-gradient* vector field notation for every joint policy $\pi \in \Pi$:

$$v_i(\pi) := \nabla_{\pi_i} u_{\mu,i}(\pi)\,, \quad v(\pi) = (v_i(\pi))_{i\in\mathcal{N}}\,. \tag{7}$$

Note that the vector field v is not necessarily a gradient field in general, i.e. the existence of a function f s.t. $v = \nabla f$ is not guaranteed in general. Indeed, each v_i is the gradient of a different function u_i.

It is well-known that Nash equilibria are first-order stationary. However, the converse implication is not true in general. Thanks to agent-wise gradient domination, the next result shows that Nash policies are equivalent to first-order stationary points in GUMGs.

Proposition 2 (First-order Stationary policies are Nash policies)**.** *Let Assumptions 1 and 2 hold. A (joint) policy $\pi^\star \in \Pi$ is a Nash policy in a GUMG if and only if it is first-order stationary, i.e. it satisfies:*

$$\langle v(\pi^\star), \pi - \pi^\star\rangle \leq 0, \quad \forall \pi \in \Pi\,. \tag{FOS}$$

Moreover, for any $\epsilon > 0$, if a policy $\pi^\star$ is a ε-FOS policy, i.e. it satisfies: $\langle v(\pi^\star), \pi - \pi^\star\rangle \leq \varepsilon, \forall \pi \in \Pi$, then $\pi^\star$ is also a $(C_{\mathcal{G}}\,\varepsilon)$-Nash policy, where $C_{\mathcal{G}}$ is the distribution mismatch coefficient defined in Prop. 1.

FOS points coincide with fixed points of the projected pseudo-gradient dynamics. Indeed, a policy $\pi^\star$ is FOS if and only if it is a fixed point of the projected gradient dynamics, i.e., $\pi^\star = \mathrm{Proj}_{\Pi}(\pi^\star + \eta\, v(\pi^\star)), \forall \eta > 0$ (see Lemma 1 in Appendix 9.3 for a proof of this straightforward result). Combining this result with Prop. 2, we obtain a characterization of Nash policies in cMGs.

Theorem 1. *Let Assumptions 1 and 2 hold. A (joint) policy $\pi^\star \in \Pi$ is a NE in a GUMG if and only if it is a fixed point of the projected gradient dynamics, i.e.:*

$$\pi^\star = Proj_{\Pi}(\pi^\star + \eta\, v(\pi^\star)), \quad \forall \eta > 0\,, \tag{8}$$

or equivalently, using the product structure of the joint policy space and the properties of the projection, $\pi^\star_i = Proj_{\Pi_i}(\pi^\star_i + \eta\, v_i(\pi^\star)), \forall i \in \mathcal{N}, \forall \eta > 0$.

Corollary 1. *Under Assumption 1, there exists at least one Nash equilibrium in GUMGs as defined in Definition 1. Moreover, there exists at least one MPE.*

Proof. Fixed points of the projected gradient dynamics described in (8) exist by Brouwer's fixed point theorem under continuity of the mapping v (which follows from Proposition 4 below) observing that the set of policies Π is compact. It follows from the equivalence of Proposition 1 that Nash policies do exist. □

In the special case of Markov games, corollary 1 provides an alternative proof to the original proof of Fink (1964) which uses Kakutani's fixed point theorem (which is for set-valued functions). Our first-order characterization of NE in GUMGs allows us to bypass the need for a set-valued function theorem and directly apply Brouwer's fixed point theorem. Existence of NE for cMGs has been recently proved in Gemp et al. (2025) via Debreu (1952)'s theorem using contractible sets (which are not necessarily convex). Thm. 1 shows that it is sufficient to find a fixed point of the projected pseudo-gradient dynamics to find a NE. This motivates the design of policy gradient algorithms for GUMGs, which we discuss in the next section.

4 POLICY GRADIENT ALGORITHM FOR GUMG

In this section, we design a policy gradient algorithm for GUMGs. We first establish a useful multi-agent

policy gradient theorem. In GUMGs, the gradient of the utility of each player w.r.t. their own policy is given by the following result.

Proposition 3 (Multi-Agent PG Theorem)**.** *For any agent $i \in \mathcal{N}$, any joint policy $\pi = (\pi_1, \cdots, \pi_N) \in \Pi$,*

(i) $\nabla_{\pi_i} u_{\mu,i}(\pi) = \mathbb{E}_{\mu,\pi}\left[\sum_{t=0}^{+\infty} \gamma^t R_i(s_t, a_t)\, \psi^t_{\pi_i}\right]$,
where $R_i(s_t, a_t) := \sum_{j=1}^N \bar{r}^\pi_{ij}(s_t, a_{j,t})$,
$\bar{r}^\pi_{ij} := \nabla_{\lambda_j} F_i(\lambda^\pi_{\mu,1}, \cdots, \lambda^\pi_{\mu,N}, \pi_{-i})$ *for* $j \in \mathcal{N}$,
$\psi^t_{\pi_i} := \sum_{t'=0}^t \nabla_{\pi_i} \log \pi_i(a_{i,t'}|s_{t'})$.

(ii) For any state $s \in \mathcal{S}$ and any action $a_i \in \mathcal{A}_i$,

$$\frac{\partial u_{\mu,i}(\pi)}{\partial \pi_i(a_i|s)} = \frac{1}{1-\gamma} d^\pi_\mu(s) \sum_{j=1}^N \bar{Q}^\pi_{s,a_i}(\bar{r}^\pi_{i,j}), \tag{9}$$

$$\bar{Q}^\pi_{s,a_i}(\bar{r}^\pi_{i,j}) := \sum_{a_{-i} \in \mathcal{A}_{-i}} \pi_{-i}(a_{-i}|s) Q^\pi_{s,(a_i,a_{-i})}(\bar{r}^\pi_{i,j}).$$

The *pseudo-reward* quantity $\bar{r}^\pi_{i,j}$ is an *implicit policy-dependent* reward (in contrast to rewards in classical RL or Markov games). For instance, in pure exploration where F_i is the entropy function, the goal is to explore a state space without reward signals.

Proposition 3 recovers as particular cases recent existing multi-agent policy gradient theorems for Markov games (see e.g. Leonardos et al. (2022), Zhang et al. (2024)), for standard single-agent RL ($N = 1$) (Sutton et al., 1999; Williams, 1992) as well as for (single-agent) RL with general utilities (Zhang et al., 2021). As for the proof, we first use the chain rule to differentiate our GUMG objective and then use the standard policy gradient using pseudo-rewards and the corresponding average Q functions. The proof is inspired from prior work in standard Markov games and single-agent RL with general utilities.

Algorithm design. Our results in section 3 (Thm. 1) suggest to perform projected pseudo-gradient ascent:

$$\pi_i^{t+1} = \mathrm{Proj}_{\Pi_i^\alpha}\left(\pi_i^t + \eta \nabla_{\pi_i} u_{\mu,i}(\pi^t)\right), \tag{10}$$

for each time step t, using a positive step size η. In light of Thm. 3, we replace exact gradients by stochastic policy gradients computed using sampled finite-length trajectories as described in Algorithm 1:

$$\hat{\nabla}_{\pi_i} u_{\mu,i}(\pi) := \sum_{t=0}^{H-1} \gamma^t \hat{R}_i(s_t, a_t) \psi^t_{\pi_i}, \tag{11}$$

where $\hat{R}_i(s_t, a_t) := \sum_{j=1}^N \hat{r}_{i,j}(s_t, a_{t,j})$ and $\psi^t_{\pi_i}$ is defined in Prop. 3. Note that marginal occupancy measures are estimated by $\hat{\lambda}^\pi_{\mu,i}(s, a_i) = \hat{d}^\pi_\mu(s)\pi_i(a_i|s)$ where state occupancy measure estimates are given by:

$$\hat{d}^\pi_\mu(s) = \frac{1}{M} \sum_{k=1}^M \sum_{t=0}^{H-1} \gamma^t \mathbb{1}_{\{s_t^{(k)} = s\}}. \tag{12}$$

Algorithm 1 PG Algorithm for GUMGs (for $i \in \mathcal{N}$)

1: **Input:** $\pi_i^0 \in \Pi_i, T, N \geq 1, \eta > 0, H$.
2: **for** $t = 0, \ldots, T-1$ **do**
3: Sample M trajectories $(\tau_{i,k})_{1 \leq k \leq M}$ of length H: $\tau_{i,k} = \{(s_t^{(k)}, a_{i,t}^{(k)})_{0 \leq t \leq H-1}\}, a_{i,t} \sim \pi_i^t(\cdot|s_t)$.
`//Pseudo-reward estimation`
4: Compute $\hat{\lambda}^{\pi^t}_{\mu,i}$ using Eq. (12).
5: Send $\hat{\lambda}^{\pi^t}_{\mu,i}$ to agents $j \in \mathcal{N}, j \neq i$.
6: Receive $\hat{\lambda}^{\pi^t}_{\mu,j}, j \in \mathcal{N}, j \neq i$.
7: Compute $\hat{r}_{i,j} = \nabla_{\lambda_j} F_i(\hat{\lambda}^{\pi^t}_{\mu,1:N}, \pi^t_{-i}), j \in \mathcal{N}$
`//Policy gradient estimation`
8: Compute PG estimate $\hat{\nabla}_{\pi_i} u_{\mu,i}(\pi^t)$ using (11) (on-policy) or (15) (under a generative model).
`//Policy update`
9: $\pi_i^{t+1} = \mathrm{Proj}_{\Pi_i^\alpha}\left(\pi_i^t + \eta \hat{\nabla}_{\pi_i} u_{\mu,i}(\pi^t)\right)$
10: **end for**

Communication protocol. In a GUMG, to implement the policy gradient Algorithm 1, each agent needs to compute its pseudo-reward ($\bar{r}^\pi_{ij} = \nabla_{\lambda_j} F_i(\lambda^\pi_{\mu,1}, \cdots, \lambda^\pi_{\mu,N}, \pi_{-i})$) which depends on the (marginal) occupancy measures of all agents. A sufficient and practical protocol is for each agent to broadcast their estimated occupancy measure and their policy to the other agents at the beginning of each interaction round (steps 5–6 of Algorithm 1). Each agent can then compute pseudo-rewards locally and update their policy independently. Note that policies of other agents can also be inferred from state-action occupancy measures via marginalization given knowledge of their action spaces. The communication requirement stems directly from the fact that individual GUMG objectives depend on other agents' occupancies and policies in our general model. In particular, communication is not required in special cases such as standard Markov games or settings where each agent's utility depends only on its own occupancy measure. We also emphasize that policies need not be communicated if the utilities F_i do not have a dependence on π_{-i} via their last argument (while they can still depend on π_{-i} via the marginal occupancy measures).

We now highlight some features of our algorithm:

Model-free algorithm. Algorithm 1 is model-free as it only relies on sampled trajectories to estimate policy gradients. In contrast, the algorithm proposed in Gemp et al. (2025) for approximate NE computation in cMGs assumes full knowledge of the transition model. Our algorithm design is rather inspired by PG methods for RL with general utilities (Zhang et al., 2021). Compared to Markov games, an important difference is the need to estimate marginal occupancies $\lambda^\pi_{\mu,j}$ to estimate the pseudo-rewards (which only need to be

estimated at sampled visited state-action pairs).

No regularization. Our algorithm does not use any form of regularization. Prior work uses either entropic regularization with temperature annealing which introduces a bias (see Thm. 2 in Gemp et al. (2025) for cMGs) or quadratic occupancy regularization in zero-sum cMGs (see Thm. 4.5 in Kalogiannis et al. (2025)).

Decentralized play. Algorithm 1 can be executed in a decentralized way: each agent i updates their policy, communicating with other agents $j \neq i$ to obtain estimates of their marginal occupancies $\lambda^{\pi}_{\mu,j}$. When F_i only depends on $\lambda^{\pi}_{\mu,i}$ (and not on the argument π_{-i}), the gradient of agent i simplifies and becomes independent from other agents' occupancies. Agent i only needs to estimate their own marginal occupancy and updates their policy *independently*.

Simultaneous play. In a GUMG, Algorithm 1 can be run simultaneously by N players who do not need to take turns. In contrast, Kalogiannis et al. (2025) propose nested and alternating algorithms for 2-player zero-sum cMGs where players need to take turns, i.e. one player updates their policy while the other waits until the other finishes before updating their own policy, in an asynchronous manner.

5 LEARNING NASH EQUILIBRIA IN POTENTIAL GUMG

In this section, we show how to learn approximate NE efficiently in GUMGs using Algorithm 1 in the fully cooperative setting where the utilities of the agents enjoy a potential structure. We establish iteration and sample complexity results under different model access assumptions. We start by defining potential GUMGs which generalize Markov potential games.

Definition 2 (Potential GUMG)**.** *A GUMG as defined in Definition 1 is said to be potential if there exists a function $\Phi : \Lambda \to \mathbb{R}$ s.t. for every $i \in \mathcal{N}$, for every policy $\pi'_i \in \Pi_i$, joint policy $\pi = (\pi_i, \pi_{-i}) \in \Pi$, $u_{\mu,i}(\pi'_i, \pi_{-i}) - u_{\mu,i}(\pi_i, \pi_{-i}) = \Phi(\lambda^{\pi'_i,\pi_{-i}}_{\mu,1:N}) - \Phi(\lambda^{\pi_i,\pi_{-i}}_{\mu,1:N})$, where $\lambda^{\pi'_i,\pi_{-i}}_{\mu,1:N} = (\lambda^{\pi'_i,\pi_{-i}}_{\mu,1}, \cdots, \lambda^{\pi'_i,\pi_{-i}}_{\mu,N})$.*

Standard Markov potential games (MPGs) are particular cases of GUMGs, including non-identical interest examples. Beyond MPGs, even the common interest setting for GUMGs (where all the functions F_i are identical) captures interesting examples that are not instances of MPGs. These include all convex functions of occupancy measures (e.g. KL divergence in imitation learning, entropy in pure exploration) that are not captured by standard MPGs. See section 2.4 for examples and appendix 8.2 for limitations.

Using the potential function Φ defined in (2), we introduce the function $\phi_\mu : \Pi \to \mathbb{R}$ defined for every joint policy $\pi \in \Pi$ as follows: $\phi_\mu(\pi) := \Phi(\lambda^{\pi}_{\mu,1}, \cdots, \lambda^{\pi}_{\mu,N})$.

For a potential GUMG, it is worth noting that for any joint policy π and any $i \in \mathcal{N}$, $v_i(\pi) = \nabla_{\pi_i} u_{\mu,i}(\pi) = \nabla_{\pi_i}\phi_\mu(\pi)$, i.e. the pseudo-gradient field v defined in (7) is a gradient field ($v = \nabla\phi_\mu$). Using this key observation, we establish the iteration complexity of Algorithm 1. We make the following smoothness assumption on the player utilities F_i which is a multi-player version of the standard smoothness assumption made in the RL with general utility literature (Zhang et al., 2021; Barakat et al., 2023, 2025).

Assumption 3 (Smoothness)**.** *For every $i \in \mathcal{N}$, the function F_i satisfies two requirements. First, there exists $l_\infty > 0$ s.t. for all $j \in [N]$, $\lambda_{1:N} = (\lambda_1, \cdots, \lambda_N) \in \Lambda, \|\nabla_{\lambda_j} F_i(\lambda_{1:N}, \pi_{-i})]\|_\infty \leq l_\infty$. Second, there exists $L > 0$ s.t. for all $\lambda, \lambda' \in \Lambda$, and for all $\pi_{-i}, \pi'_{-i} \in \Pi_{-i}$, $\|\nabla_{\lambda_j} F_i(\lambda, \pi_{-i}) - \nabla_{\lambda_j} F_i(\lambda', \pi'_{-i})\|_\infty \leq L(\|\lambda - \lambda'\|_1 + \|\pi_{-i} - \pi'_{-i}\|_1)$.*

Remark 3. *The first requirement in Assumption 3 follows from compactness of the space of occupancy measures and continuity of all the partial derivatives of the functions F_i. We mainly state it formally to introduce the constant l_∞.*

Under Assumption 3, we obtain smoothness of the utility functions $u_{\mu,i}$ for $i \in \mathcal{N}$.

Proposition 4. *Let Assumption 3 hold. Then, for all $i \in \mathcal{N}$, the utility $u_{\mu,i}$ defined in (5) is smooth, i.e. there exists a constant $\beta > 0$ s.t. for all $\pi, \pi' \in \Pi$, $\|\nabla_\pi u_{\mu,i}(\pi) - \nabla_\pi u_{\mu,i}(\pi')\|_2 \leq \beta\|\pi - \pi'\|_2$.*

5.1 Exact Setting Iteration Complexity

In this section, we suppose that each agent $i \in \mathcal{N}$ has access to exact gradients $\nabla_{\pi_i} u_{\mu,i}(\pi)$ for any joint policy π and we analyze the iteration complexity of Algorithm 1 to obtain an ε-approximate NE.

Define for any joint policy $\pi \in \Pi$ and any $i \in \mathcal{N}$ the individual and maximal NE gaps:

$$\text{NE-Gap}_i(\pi) := \max_{\pi'_i \in \Pi_i} \{u_{\mu,i}(\pi'_i, \pi_{-i}) - u_{\mu,i}(\pi)\}, \quad (13)$$

$$\text{NE-Gap}(\pi) := \max_{i \in \mathcal{N}} \text{NE-Gap}_i(\pi). \quad (14)$$

A policy $\pi^\star$ is an ϵ-NE if and only if $\text{NE-Gap}(\pi^\star) \leq \epsilon$.

Theorem 2 (Deterministic setting)**.** *Let Assumptions 1, 2, and 3 hold. Then, the sequence (π^t) of policies generated by running Algorithm 1 (with exact gradients in step 8 and $\alpha = 0$) simultaneously by all agents $i \in \mathcal{N}$ with exact gradients and stepsize $\eta \leq 1/\beta$ in a potential GUMG satisfies: $\lim_{t\to+\infty}$ NE-Gap$(\pi^t) = 0$. Moreover, for any $\varepsilon > 0$, for $T \geq \frac{32\mathcal{C}_{\mathcal{G}}^2\beta(\Phi_{\max}-\Phi_{\min})}{(1-\gamma)^2\varepsilon^2}$, $\frac{1}{T}\sum_{t=1}^{T}$ NE-Gap$(\pi^t) \leq \varepsilon$.*

Remark 4. *Compared to using a centralized single-agent PG algorithm for convex RL, the iteration complexity only scales with $\sum_{i=1}^{N}|\mathcal{A}_i|$ (which appears in β, see appendix) rather than the cardinality of the joint action space $|\mathcal{A}|$ which scales exponentially with the number N of agents. This additive dependence breaks the so-called curse of multi-agency thanks to decentralization. As for the convergence rate, notice that our gradient domination (Prop. 1) only holds per-agent and is therefore weaker than its single-agent counterpart. Therefore, it cannot be immediately leveraged to obtain faster last-iterate convergence rates like in the single-agent setting (Xiao, 2022).*

5.2 Generative Model Sample Complexity

We now suppose that we only have access to a generative model, i.e. a simulator which we can query and samples a next state $s' \sim P(\cdot|s,a)$ given *any* current state s and any joint action $a \in \mathcal{A}$. Therefore finite-length state-action trajectories can be sampled from *any* state. Using this model and given our PG theorem (Thm. 3-(ii)) we estimate individual utility gradients as follows: Given a current policy π, generate for each $i \in \mathcal{N}, (s,a_i) \in \mathcal{S}\times\mathcal{A}_i$ a set of M independent truncated trajectories of length H: $\tau_{s,a_i}^{(k)} = \{(s_0^{(k)}, a_{i,0}^{(k)}), \cdots, (s_{H-1}^{(k)}, a_{i,H-1}^{(k)})\}, k \in [M]$ with $s_0^{(k)} = s, a_{i,0}^{(k)} = a_i$. Then compute the estimates:

$$[\hat{\nabla}_{\pi_i} u_{\mu,i}(\pi)]_{s,a_i} = \frac{1}{1-\gamma}\hat{d}_\mu^\pi(s)\sum_{j=1}^{N}\hat{q}_{i,j,(s,a_i)}^\pi\,, \qquad (15)$$

where $\hat{q}_{i,j,(s,a_i)}^\pi = \frac{1}{M}\sum_{k=1}^{M}\sum_{t=0}^{H-1}\gamma^t\hat{r}_{i,j}^\pi(s_t^{(k)}, a_{t,j}^{(k)})$, and $\hat{r}_{i,j}^\pi = \nabla_{\lambda_j}F_i(\hat{\lambda}_{\mu,1:N}^\pi)$ and $\hat{d}_\mu^\pi(s), \hat{\lambda}_{\mu,i}^\pi(s,a_i)$ are estimated as in (12). Using the above estimated gradients, we are ready to state our first sample complexity result to find an ε-approximate NE in a potential GUMG.

Theorem 3. *Let Assumptions 1, 2, and 3 hold. For any desired accuracy $\varepsilon > 0$ and any confidence level $\delta \in (0,1)$, if each agent $i \in \mathcal{N}$ runs Algorithm 1 (using (15) in step 8) with stepsize η, greediness parameter $\alpha = 0$, minibatch size $M = \tilde{\mathcal{O}}(\varepsilon^{-2})$ and truncation horizon $H = \mathcal{O}((1-\gamma)^{-1}\log(1/\varepsilon))$ for a number of iterations $T = \tilde{\mathcal{O}}(\varepsilon^{-2})$, then we have $\frac{1}{T}\sum_{t=0}^{T-1}$ NE-Gap$(\pi^{t+1}) \le \varepsilon$ with probability at least $1-\delta$. In particular, the total sample complexity in terms of state-action pairs is $T\times M\times H = \tilde{\mathcal{O}}(\varepsilon^{-4})$.*[2]

5.3 On-policy Learning Sample Complexity

In this last section, we relax the sampling model of section 5.2 and we suppose that each agent can only sample trajectories of finite length using their current policy at each time step t as described in section 4. In particular, agents cannot query the transition model from *any* state, including non-visited ones. This justifies the use of α-greedy policies with $\alpha > 0$ to guarantee state exploration.

In contrast to section 5.2, the algorithm described in this section does not require a generative model. In section 5.2, policy gradients are estimated (using the gradient representation in Prop. 3-(ii)) by computing estimates of the Q functions and occupancy measures *at every state-action pair* (s,a_i) for agent i. This requires being able to sample the next state from any given state–action pair, hence the need for a generative model (see e.g. section 5.1 in Xiao (2022) for a similar model in the single-agent setting). In this section, we only need to estimate the pseudo-rewards at the state-action pairs that were encountered in the sampled trajectories (see Prop. 3-(i)) since only the values of $\bar{r}_{i,j}^\pi(s_t, a_{j,t})$ are required to compute stochastic policy gradients. Computing estimates of the occupancy measures at every state-action pair is not needed. The sampling model in section 5.3 only allows to sample trajectories using the current policy from initial states drawn from the fixed initial state distribution.

The next result shows that we can still learn approximate NE in this case while the sample complexity slightly deteriorates compared to Thm. 3.

Theorem 4. *Let Assumptions 1, 2, and 3 hold. For any desired accuracy $\varepsilon > 0$, if each agent $i \in \mathcal{N}$ runs Algorithm 1 (using (11) in step 8) in a potential GUMG with stepsize η, positive greediness policy parameter $\alpha = \mathcal{O}(\varepsilon)$, minibatch size $M = \tilde{\mathcal{O}}(\varepsilon^{-3})$ and truncation horizon $H = \mathcal{O}((1-\gamma)^{-1}\log(1/\varepsilon))$ for a number of iterations $T = \tilde{\mathcal{O}}(\varepsilon^{-2})$, then we have $\mathbb{E}[$NE-Gap$(\pi^\tau)] \le \varepsilon$ where expectation is w.r.t. both randomness in Algorithm 1 due to trajectory sampling and τ which is a uniform random variable over $[T]$. In particular, the total sample complexity in terms of state-action pairs is $T\times M\times H = \tilde{\mathcal{O}}(\varepsilon^{-5})$.*

6 CONCLUSION

In this work, we established the existence of NE in GUMGs as fixed points of the projected pseudo-gradient dynamics via Brouwer's theorem and a per-agent gradient domination property. We designed a model-free policy gradient method and proved the first efficient NE learning guarantees in potential GUMGs. Future directions include extending our analysis to large state–action spaces with function approximation, and exploring generalizations to general-sum settings, e.g. convergence to correlated equilibria under decentralized dynamics (Erez et al., 2023; Cai et al., 2024).

[2]The notation $\tilde{\mathcal{O}}(\cdot)$ hides polynomial factors in $(1-\gamma)^{-1}, |\mathcal{S}|, \max_{i\in\mathcal{N}}|\mathcal{A}_i|$ and polylogarithmic factors in $1/\varepsilon$ and $|\mathcal{S}|\cdot\max_{i\in\mathcal{N}}|\mathcal{A}_i|/\delta$, see appendix for explicit constants.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback, which helped improve the manuscript. This work is supported by the MOE Tier 2 Grant (MOE-T2EP20223-0018), Ministry of Education Singapore (SRG ESD 2024 174), the CQT++ Core Research Funding Grant (SUTD) (RS-NRCQT-00002), and partially by Project MIS 5154714 of the National Recovery and Resilience Plan, Greece 2.0, funded by the European Union under the NextGenerationEU Program. Ioannis Panageas is supported by NSF grant CCF- 2454115.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76.

Alacaoglu, A., Viano, L., He, N., and Cevher, V. (2022). A natural actor-critic framework for zero-sum markov games. In *International Conference on Machine Learning*.

Alatur, P., Barakat, A., and He, N. (2024). Independent policy mirror descent for markov potential games: Scaling to large number of players. In *IEEE Conference on Decision and Control*.

Aydin, S. and Eksin, C. (2023). Policy gradient play with networked agents in markov potential games. In *Learning for Dynamics and Control Conference*.

Bai, Q., Agarwal, M., and Aggarwal, V. (2022). Joint optimization of concave scalarized multi-objective reinforcement learning with policy gradient based algorithm. *Journal of Artificial Intelligence Research*, 74.

Bai, Y. and Jin, C. (2020). Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*.

Bai, Y., Jin, C., and Yu, T. (2020). Near-optimal reinforcement learning with self-play. In *Advances in Neural Information Processing Systems*.

Barakat, A., Chakraborty, S., Yu, P., Tokekar, P., and Bedi, A. S. (2025). On the global optimality of policy gradient methods in general utility reinforcement learning. In *Annual Conference on Neural Information Processing Systems*.

Barakat, A., Fatkhullin, I., and He, N. (2023). Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. In *International Conference on Machine Learning*.

Bhatt, M. S. and Sobel, M. J. (2025). Risk-sensitive markov-perfect equilibrium. *Annals of Operations Research*, pages 1–17.

Cai, Y., Luo, H., Wei, C.-Y., and Zheng, W. (2023). Uncoupled and convergent learning in two-player zero-sum markov games with bandit feedback. In *Advances in Neural Information Processing Systems*.

Cai, Y., Luo, H., Wei, C.-Y., and Zheng, W. (2024). Near-optimal policy optimization for correlated equilibrium in general-sum markov games. In *International Conference on Artificial Intelligence and Statistics*.

Cheng, M., Zhou, R., Kumar, P., and Tian, C. (2024). Provable policy gradient methods for average-reward markov potential games. In *International Conference on Artificial Intelligence and Statistics*.

Cui, Q. and Du, S. S. (2022). When are offline two-player zero-sum markov games solvable? In *Advances in Neural Information Processing Systems*.

Daskalakis, C., Foster, D. J., and Golowich, N. (2020). Independent policy gradient methods for competitive reinforcement learning.

De Santi, R., Prajapat, M., and Krause, A. (2024). Global reinforcement learning : Beyond linear and convex rewards via submodular semi-gradient methods. In *International Conference on Machine Learning*.

Debreu, G. (1952). A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences*, 38(10):886–893.

Deng, X., Li, N., Mguni, D., Wang, J., and Yang, Y. (2023). On the complexity of computing markov perfect equilibrium in general-sum stochastic games. *National Science Review*, 10(1).

Ding, D., Wei, C.-Y., Zhang, K., and Jovanovic, M. (2022). Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*.

Erez, L., Lancewicki, T., Sherman, U., Koren, T., and Mansour, Y. (2023). Regret minimization and convergence to equilibria in general-sum markov games. In *International Conference on Machine Learning*.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. (2019). Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*.

Fatkhullin, I., He, N., and Hu, Y. (2025). Stochastic optimization under hidden convexity. *SIAM Journal on Optimization*, 35(4):2544–2571.

Fink, A. M. (1964). Equilibrium in a stochastic n-person game. *Journal of Science of the Hiroshima University, Series A-I (mathematics)*, 28(1):89–93.

Fox, R., Mcaleer, S. M., Overman, W., and Panageas, I. (2022). Independent natural policy gradient always converges in markov potential games. In *International Conference on Artificial Intelligence and Statistics*.

Geist, M., Pérolat, J., Laurière, M., Elie, R., Perrin, S., Bachem, O., Munos, R., and Pietquin, O. (2022). Concave utility reinforcement learning: The mean-field game viewpoint. In *International Conference on Autonomous Agents and Multiagent Systems*.

Gemp, I., Haupt, A. A., Marris, L., Liu, S., and Piliouras, G. (2025). Convex markov games: A new frontier for multi-agent reinforcement learning. In *International Conference on Machine Learning*.

Giannou, A., Lotidis, K., Mertikopoulos, P., and Vlatakis-Gkaragkounis, E.-V. (2022). On the convergence of policy gradient methods to nash equilibria in general stochastic games. In *Advances in Neural Information Processing Systems*.

Guo, X., Li, X., Maheshwari, C., Sastry, S., and Wu, M. (2025). Markov α-potential games. *IEEE Transactions on Automatic Control*.

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. (2019). Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR.

Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*.

Howard, R. A. (1960). Dynamic programming and markov processes.

Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., García Castañeda, A., Dunning, I., Zhu, T., McKee, K., Koster, R., et al. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems*, 31.

Kalogiannis, F., Anagnostides, I., Panageas, I., Vlatakis-Gkaragkounis, E.-V., Chatziafratis, V., and Stavroulakis, S. A. (2023). Efficiently computing nash equilibria in adversarial team markov games. In *International Conference on Learning Representations*.

Kalogiannis, F., Vlatakis-Gkaragkounis, E.-V., Gemp, I., and Piliouras, G. (2025). Solving zero-sum convex markov games. In *International Conference on Machine Learning*.

Kalogiannis, F., Yan, J., and Panageas, I. (2024). Learning equilibria in adversarial team markov games: A nonconvex-hidden-concave min-max optimization problem. In *Advances in Neural Information Processing Systems*.

Kim, W. and Sung, Y. (2023). An adaptive entropy-regularization framework for multi-agent reinforcement learning. In *International Conference on Machine Learning*.

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. (2022). Global convergence of multi-agent policy gradient in markov potential games. In *International Conference on Learning Representations*.

Li, C., Wang, T., Wu, C., Zhao, Q., Yang, J., and Zhang, C. (2021). Celebrating diversity in shared multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163.

Maheshwari, C., Wu, M., Pai, D., and Sastry, S. (2025). Independent and decentralized learning in markov potential games. *IEEE Transactions on Automatic Control*.

Marin Moreno, B., Brégère, M., Gaillard, P., and Oudjane, N. (2024). Metacurl: Non-stationary concave utility reinforcement learning. In *Advances in Neural Information Processing Systems*.

Mavrommati, A., Tzorakoleftherakis, E., Abraham, I., and Murphey, T. D. (2017). Real-time area coverage and target localization using receding-horizon ergodic exploration. *IEEE Transactions on Robotics*, 34(1):62–80.

Mutny, M., Janik, T., and Krause, A. (2023). Active exploration via experiment design in markov chains. In *International Conference on Artificial Intelligence and Statistics*.

Mutti, M., De Santi, R., De Bartolomeis, P., and Restelli, M. (2023). Convex reinforcement learning in finite trials. *Journal of Machine Learning Research*, 24(250):1–42.

Park, C., Zhang, K., and Ozdaglar, A. (2023). Multi-player zero-sum markov games with networked separable interactions. In *Advances in Neural Information Processing Systems*.

Perolat, J., Scherrer, B., Piot, B., and Pietquin, O. (2015). Approximate dynamic programming for two-player zero-sum markov games. In *International Conference on Machine Learning*.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Qiu, S., Wei, X., Ye, J., Wang, Z., and Yang, Z. (2021). Provably efficient fictitious play policy optimization for zero-sum markov games with structured transitions. In *International Conference on Machine Learning*.

Sahitaj, R., Sasnauskas, P., Yalın, Y., Mandal, D., and Radanovic, G. (2025). Independent learning in performative markov potential games. In *International Conference on Artificial Intelligence and Statistics*.

Sayin, M., Zhang, K., Leslie, D., Basar, T., and Ozdaglar, A. (2021). Decentralized q-learning in zero-sum markov games. In *Advances in Neural Information Processing Systems*.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100.

Song, J., Ren, H., Sadigh, D., and Ermon, S. (2018). Multi-agent generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*.

Sun, Y., Liu, T., Zhou, R., Kumar, P., and Shahrampour, S. (2023). Provably fast convergence of independent natural policy gradient for markov potential games. In *Advances in Neural Information Processing Systems*.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Wu, Y., Viano, L., Chen, Y., Zhu, Z., Gu, Q., and Cevher, V. (2024). Multi-step preference optimization via two-player markov games. In *Language Gamification - NeurIPS 2024 Workshop*.

Xiao, L. (2022). On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. (2020). Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*.

Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., and Wang, J. (2018). Mean field multi-agent reinforcement learning. In *International Conference on Machine Learning*.

Ying, D., Zhang, Y., Ding, Y., Koppel, A., and Lavaei, J. (2023). Scalable primal-dual actor-critic method for safe multi-agent rl with general utilities. In *Advances in Neural Information Processing Systems*.

Zahavy, T., O'Donoghue, B., Desjardins, G., and Singh, S. (2021). Reward is enough for convex mdps. In *Advances in Neural Information Processing Systems*.

Zamboni, R., Mutti, M., and Restelli, M. (2025). Towards principled unsupervised multi-agent reinforcement learning. In *Annual Conference on Neural Information Processing Systems*.

Zhang, J., Bedi, A. S., Wang, M., and Koppel, A. (2022a). Multi-agent reinforcement learning with general utilities via decentralized shadow reward actor-critic. In *AAAI Conference on Artificial Intelligence*.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. (2020). Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583.

Zhang, J., Ni, C., Yu, Z., Szepesvari, C., and Wang, M. (2021). On the convergence and sample efficiency of variance-reduced policy gradient method. In *Advances in Neural Information Processing Systems*.

Zhang, K., Kakade, S. M., Basar, T., and Yang, L. F. (2023). Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *Journal of Machine Learning Research*, 24(175):1–53.

Zhang, R., Liu, Q., Wang, H., Xiong, C., Li, N., and Bai, Y. (2022b). Policy optimization for markov games: Unified framework and faster convergence. In *Advances in Neural Information Processing Systems*.

Zhang, R., Mei, J., Dai, B., Schuurmans, D., and Li, N. (2022c). On the global convergence rates of decentralized softmax gradient play in markov potential games. In *Advances in Neural Information Processing Systems*.

Zhang, R., Ren, Z., and Li, N. (2024). Gradient play in stochastic games: stationary points, convergence, and sample complexity. *IEEE Transactions on Automatic Control*.

Zhao, Y., Tian, Y., Lee, J., and Du, S. (2022). Provably efficient policy optimization for two-player zero-sum markov games. In *International Conference on Artificial Intelligence and Statistics*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material:

Convex Markov Games and Beyond: New Proof of Existence, Characterization and Learning Algorithms for Nash Equilibria

Contents

1 INTRODUCTION **1**

1.1 Our Contributions . . . 2

1.2 Technical Overview . . . 2

1.3 Related Work . . . 3

2 FROM CLASSICAL TO GENERAL UTILITY MARKOV GAMES **3**

2.1 Markov Games and Notation . . . 3

2.2 Convex Markov Games . . . 4

2.3 General Utility Markov Games . . . 4

2.4 Examples . . . 4

3 EXISTENCE AND STRUCTURE OF NASH POLICIES IN GUMG **5**

4 POLICY GRADIENT ALGORITHM FOR GUMG **6**

5 LEARNING NASH EQUILIBRIA IN POTENTIAL GUMG **8**

5.1 Exact Setting Iteration Complexity . . . 8

5.2 Generative Model Sample Complexity . . . 9

5.3 On-policy Learning Sample Complexity . . . 9

6 CONCLUSION **9**

7 EXTENDED RELATED WORK **15**

8 MORE DETAILS ON EXAMPLES, EQUILIBRIA AND ALGORITHM **16**

8.1 More Details on Examples of Section 2.4 . . . 16

8.2 Limitations of Potential GUMGs . . . 17

8.3 Markov Nash Equilibrium vs Markov Perfect Equilibrium . . . 17

8.4 Initial State Sampling . . . 17

8.5 Utility Function Knowledge . . . 18

8.6 Regularization . . . 18

9 PROOFS FOR SECTION 3 **18**

9.1 Proof of Proposition 1 . . . 18

9.2 Proof of Proposition 2 . . . 19

9.3 Proof of Lemma 1 . . . 20

9.4 Proof of Theorem 1 . . . 20

9.5 Proof of Corollary 1 . . . 20

10 PROOFS FOR SECTION 4: Proof of Proposition 3 **21**

11 PROOFS FOR SECTION 5 **22**

11.1 Proof of Proposition 4 . . . 22

11.2 Proof of Lemma 2 . . . 25

11.3 Proof of Thm. 2 . . . 26

11.4 Proof of Thm. 3 . . . 27

11.5 Proof of Proposition 9 . . . 30

11.6 Proof of Thm. 4 . . . 31

11.6.1 Proof of Proposition 10 . . . 34

11.6.2 Proof of Proposition 11 . . . 35

12 AUXILIARY RESULTS: SINGLE-AGENT RL **36**

13 NUMERICAL SIMULATIONS **38**

7 EXTENDED RELATED WORK

Markov Games. Recent years have witnessed active research in theoretical MARL. In particular, an important body of work has focused on learning approximate Nash equilibria in structured classes of Markov games such as zero-sum Markov games (Perolat et al., 2015; Bai and Jin, 2020; Bai et al., 2020; Xie et al., 2020; Daskalakis et al., 2020; Sayin et al., 2021; Qiu et al., 2021; Alacaoglu et al., 2022; Cui and Du, 2022; Zhang et al., 2022b; Zhao et al., 2022; Park et al., 2023; Zhang et al., 2023; Cai et al., 2023), Markov potential games (Leonardos et al., 2022; Zhang et al., 2024; Ding et al., 2022; Fox et al., 2022; Zhang et al., 2022c; Guo et al., 2025; Sun et al., 2023; Cheng et al., 2024; Maheshwari et al., 2025; Alatur et al., 2024; Aydin and Eksin, 2023; Sahitaj et al., 2025) or adversarial team games (Kalogiannis et al., 2023, 2024). Our work extends these research efforts to cMGs and beyond.

Convex Markov Games. Recently, Gemp et al. (2025) introduced cMGs, proved existence of NE and proposed to approximate them using a *biased* entropic regularization approach under *full knowledge* of the transition model. In this work, we design a *model-free* PG algorithm for GUMGs and analyze its iteration and *sample complexity*. In a follow-up work, Kalogiannis et al. (2025) studied *2-player zero-sum cMGs*. In contrast, we characterize NE in any N-player general-sum GUMG (with cMGs as particular cases). Moreover, we study the common-interest setting beyond the competitive zero-sum two-player setting. See sections 3 and 4 for more details comparing our theoretical and algorithmic contributions to prior work. Recently, Zamboni et al. (2025) studied task-agnostic exploration in MARL, focusing on an identical-interest setting. Their work provides a global optimality iteration complexity for projected gradient ascent leveraging the single-agent analysis of Zhang et al. (2021) under access to exact gradients and assuming smoothness of the objective (see appendix B.1 therein). We note that such a single-agent analysis leads to an exponential dependence on the number of players in general (via the product of action space sizes). In contrast, in our more general setting of GUMGs, we provide a multi-agent decentralized policy gradient sample complexity analysis which only scales with the sum of the action space sizes, using

stochastic gradients rather than exact ones and proving smoothness of the objectives under suitable assumptions on the utility functions. Note that in general global optimality results cannot be established in our general setting using single-agent analysis (Zhang et al., 2020, 2021) since multi-agent gradient domination (Prop. 1) is weaker than the corresponding result for single agents (Xiao, 2022) as it only holds per-player.

Convex MDPs and RL with General Utilities. In the single-agent setting ($N = 1$), convex Markov games reduce to convex RL which is an active research topic (Hazan et al., 2019; Zahavy et al., 2021; Zhang et al., 2020, 2021; Geist et al., 2022; Bai et al., 2022; Mutti et al., 2023; Barakat et al., 2023; Marin Moreno et al., 2024; De Santi et al., 2024). A few works consider MARL with general utilities as a *decentralized policy optimization* problem (Zhang et al., 2022a; Ying et al., 2023). Our setting is fundamentally different: we rather consider strategic agents in a game-theoretic setting where each agent has their own objective. In particular, GUMGs cannot be cast as single-objective policy optimization or consensus problems in general. Furthermore, our state space is not decomposable as a product of local state spaces.

8 MORE DETAILS ON EXAMPLES, EQUILIBRIA AND ALGORITHM

8.1 More Details on Examples of Section 2.4

In this section we provide more details regarding each one of the examples of section 2.4 to show how to apply our framework and demonstrate its versatility, especially for gradient computation in Algorithm 1 (step 7).

Recall from (6) our composite example defined for any $\lambda = (\lambda_1, \cdots, \lambda_N) \in \Lambda, \pi = (\pi_i, \pi_{-i}) \in \Pi$, by:

$$F_i(\lambda_1, \cdots, \lambda_N, \pi_{-i}) \quad = \quad \underbrace{-\alpha_i \mathrm{KL}(\lambda_i || q_i)}_{\text{imitation}} \quad \underbrace{-\beta_i \mathrm{KL}(\lambda_i || \bar{\lambda}_N)}_{\text{consensus/diversity}} + \underbrace{\gamma_i h_i \left(\sum_{j=1}^{N} W_{i,j} \lambda_j \right)}_{\text{team aggregate}} + \underbrace{g_i(\pi_{-i})}_{\text{diversity penalty}} \quad , \tag{16}$$

where $q_i \in \Lambda_i, i \in \mathcal{N}$ are fixed probability distributions, $\alpha_i, \beta_i, \gamma_i$ are nonnegative reals, $W \in \mathbb{R}^{N \times N}$ is a stochastic matrix of weights, and $\bar{\lambda}_N = \sum_{i=1}^{N} w_i \lambda_i$ with $w_i \geq 0, \sum_{i=1}^{N} w_i = 1$. Here, h_i and g_i are concave functions. Note that F_i satisfies Assumption 1 since it is the sum of jointly concave functions, noticing that KL divergence is jointly convex and the composition of a convex function with a linear one is also convex.

We now elaborate on special cases of our composite example:

a. Multi-agent imitation learning. In this case ($\alpha_i > 0, \beta_i = \gamma_i = 0$), the utility function of each agent and its gradient with respect to their own occupancy measure variable are given by:

$$\begin{aligned} F_i(\lambda_1, \cdots, \lambda_N, \pi_{-i}) &:= -\alpha_i \mathrm{KL}(\lambda_i || q_i) + g_i(\pi_{-i}) \,, \\ \nabla_{\lambda_i} F_i(\lambda_1, \cdots, \lambda_N, \pi_{-i}) &= -\alpha_i \left(\log \left(\frac{\lambda_i}{q_i} \right) + \mathbb{1} \right) \,, \\ \nabla_{\lambda_j} F_i(\lambda_1, \cdots, \lambda_N, \pi_{-i}) &= 0 \,, \quad \forall j \neq i \,, \end{aligned} \tag{17}$$

where $\mathbb{1}$ refers to the vector of ones and the notation $\log \left(\frac{\lambda_i}{q_i} \right)$ is to be understood component-wise. It remains to use these gradients to compute pseudo-rewards in Algorithm 1.

b. Consensus and diversity. In this case ($\beta_i > 0, \alpha_i = \gamma_i = 0$), the utility function of each agent and its gradients with respect to the different occupancy measure variables are given by:

$$\begin{aligned} F_i(\lambda_1, \cdots, \lambda_N, \pi_{-i}) &:= -\beta_i \mathrm{KL}(\lambda_i || \bar{\lambda}_N) + g_i(\pi_{-i}) \,, \\ \nabla_{\lambda_i} F_i(\lambda_1, \cdots, \lambda_N, \pi_{-i}) &= -\beta_i \left(\log \left(\frac{\lambda_i}{\bar{\lambda}_N} \right) + \frac{\bar{\lambda}_N - w_i \lambda_i}{\bar{\lambda}_N} \right) \,, \\ \nabla_{\lambda_j} F_i(\lambda_1, \cdots, \lambda_N, \pi_{-i}) &= \beta_i \frac{w_j \lambda_i}{\bar{\lambda}_N} \,, \quad \forall j \neq i \,, \end{aligned} \tag{18}$$

where again log and fractions are to be computed coordinate-wise.

c. Population-level coordination. In this case ($\gamma_i > 0, \alpha_i = \beta_i = 0$), the utility function of each agent and its gradients with respect to the different occupancy measure variables are given by:

$$\begin{aligned} F_i(\lambda_1, \cdots, \lambda_N, \pi_{-i}) &:= \gamma_i h_i \left(\sum_{j=1}^{N} W_{i,j} \lambda_j \right) + g_i(\pi_{-i}) \,, \\ \nabla_{\lambda_i} F_i(\lambda_1, \cdots, \lambda_N, \pi_{-i}) &= \gamma_i W_{i,i}^\top \nabla h_i \left(\sum_{j=1}^{N} W_{i,j} \lambda_j \right) \,, \\ \nabla_{\lambda_j} F_i(\lambda_1, \cdots, \lambda_N, \pi_{-i}) &= \gamma_i W_{i,j}^\top \nabla h_i \left(\sum_{j=1}^{N} W_{i,j} \lambda_j \right) \,, \quad \forall j \neq i \,, \end{aligned} \tag{19}$$

where ∇h_i is the gradient of the function h_i.

d. Team coverage and distribution matching ($F_i(\lambda_{1:N}, \pi_{-i}) = -\mathrm{KL}(\bar{\lambda}_N || \lambda_{\text{target}})$). In this case, the utility function of each agent and its gradients with respect to the different occupancy measure variables are given by:

$$\begin{aligned} F_i(\lambda_1, \cdots, \lambda_N, \pi_{-i}) &:= -\mathrm{KL}(\bar{\lambda}_N || \lambda_{\text{target}}) \,, \\ \nabla_{\lambda_i} F_i(\lambda_1, \cdots, \lambda_N, \pi_{-i}) &= -w_i \left(\log \left(\frac{\bar{\lambda}_N}{\lambda_{\text{target}}} \right) + \mathbb{1} \right) \,, \\ \nabla_{\lambda_j} F_i(\lambda_1, \cdots, \lambda_N, \pi_{-i}) &= -w_j \left(\log \left(\frac{\bar{\lambda}_N}{\lambda_{\text{target}}} \right) + \mathbb{1} \right) \,, \quad \forall j \neq i \,, \end{aligned} \tag{20}$$

where again log and fractions are to be computed coordinate-wise.

e. Collective exploration. When h_i is the negative entropy, utility maximization enhances exploration. In this case ($\gamma_i > 0, \alpha_i = \beta_i = 0$), the utility function of each agent and its gradients with respect to the different occupancy measure variables are given by (19) and the gradient of the negative entropy is straightforward.

8.2 Limitations of Potential GUMGs

The potential assumption is rigid beyond the identical interest setting. However, we point out that this is already the case for MPGs. Beyond the common interest setting, the potential structure imposes important restrictions on the transitions (e.g. transitions independent of actions). In particular, a state-based potential game (for which reward functions induce a normal-form potential game at each state) is not an MPG in general as further restrictive conditions on state transitions are required (see e.g. Prop. 3.2 in Leonardos et al. (2022)). In our case, even the common interest setting is interesting and goes beyond MPGs.

8.3 Markov Nash Equilibrium vs Markov Perfect Equilibrium

A (Markov) Nash equilibrium policy is defined relative to an initial distribution μ and ensures optimality only for the original experiment starting from that distribution. This policy need not be a Nash policy if the system is perturbed to another state. MPE is a more robust and stronger equilibrium notion than NE.

In many dynamic environments, the system can be unexpectedly pushed into states that were extremely unlikely under the initial distribution, for example, a sudden congestion spike in a network, a robot drifting off a planned trajectory, or a firm experiencing an unexpected shock. A Nash equilibrium defined only for the initial state distribution does not guarantee rational behavior outside the initial state distribution. An MPE is independent from the initial state distribution and guarantees that the policy is a best response in every state, ensuring that the policy remains sensible and stable even when the system experiences disturbances or is initialized outside the initially assumed distribution.

8.4 Initial State Sampling

In section 5.3, the initial state in each trajectory is sampled according to the fixed initial state distribution defining the GUMG. This is because the policy gradients are expectations over this initial state distribution

(see Prop. 3-(i)) since utilities are defined as functions of the occupancy measures induced by the same initial distribution. Therefore, sampling starting from the initial state distribution yields an unbiased policy gradient estimate. This is consistent with on-policy policy gradient methods for standard MPGs and single-agent RL with general utilities. Whether it is possible to learn Nash policies using a single trajectory by only sampling starting from the current state at each time step is an interesting question that merits further investigation. We expect that this would induce a bias in the policy gradient that needs to be controlled by carefully bounding the policy drift by tuning the step sizes for instance.

8.5 Utility Function Knowledge

In most applications of (single-agent) RL with general utilities known in the literature, the function F_i is available, and its gradient can therefore be computed directly. Thus, in the majority of scenarios, access to F_i (and its gradient) is not a restrictive requirement. For instance, this includes negative entropy in pure exploration (see e.g. Hazan et al. (2019)), KL divergence in imitation learning (see e.g. Ho and Ermon (2016)), A and D-design objectives in experiment design (see e.g. Mutny et al. (2023)). In addition, utilities are also typically supposed to be known in recent work on RL with general utilities (e.g. Barakat et al. (2023, 2025)) and on cMGs (Gemp et al., 2025; Kalogiannis et al., 2025). Investigating settings where it can be useful to use some weaker utility feedback (such as bandit feedback) or learn such utility functions is an interesting direction for future work.

8.6 Regularization

Regularization has been used in recent work on cMGs in Gemp et al. (2025) and in the follow-up work of Kalogiannis et al. (2025). Gemp et al. (2025) do not adopt our policy gradient viewpoint and assume knowledge of the transition kernel, they use entropic regularization in the utilities and they do not provide convergence guarantees. Kalogiannis et al. (2025) focus on the zero-sum setting and regularization improves conditioning for solving the resulting minmax problem. In contrast, our work provides convergence guarantees in the potential setting which does not lead to a minmax problem (while our existence and characterization results hold for any GUMG).

Our approach is consistent with existing work in Markov games (Markov potential games in particular) where regularization is not necessary to obtain polynomial sample complexity for learning approximate Nash equilibria, even for 2-player zero-sum Markov games (see e.g. Daskalakis et al. (2020)). We believe that we can also prove guarantees for zero-sum GUMGs without regularization using an approach similar to Daskalakis et al. (2020) for 2-player zero-sum Markov games. However, such an analysis only provides one-sided Nash guarantees (for one of the players) when running the algorithm and would lead to iteration and sample complexities which would be worse than the analysis of Kalogiannis et al. (2025) (which require players to take turns on the other hand).

Our approach is also inspired from single-agent RL with general utilities policy optimization (e.g. Zhang et al. (2021); Barakat et al. (2023, 2025)) where policy gradient algorithms have been studied without regularization.

9 PROOFS FOR SECTION 3

9.1 Proof of Proposition 1

Proof. **Step 1. Using concavity to relate utility deviation to standard single-agent RL.** Let $i \in \mathcal{N}$ and let $\pi = (\pi_i, \pi_{-i}) \in \Pi, \pi_i' \in \Pi_i$. By joint concavity of F_i(Assumption 1) w.r.t. $(\lambda_1, \cdots, \lambda_N)$, we have that for any $\lambda = (\lambda_1, \cdots, \lambda_N), \lambda' = (\lambda_1', \cdots, \lambda_N') \in \mathbb{R}^{|\mathcal{S}|\cdot\sum_{j=1}^N |\mathcal{A}_k|}$, and any $\pi_{-i} \in \Pi_{-i}$,

$$F_i(\lambda', \pi_{-i}) - F_i(\lambda, \pi_{-i}) \leq \langle \nabla_\lambda F_i(\lambda, \pi_{-i}), \lambda' - \lambda \rangle = \sum_{j=1}^N \langle \nabla_{\lambda_j} F_i(\lambda_1, \cdots, \lambda_N, \pi_{-i}), \lambda_j' - \lambda_j \rangle \,. \tag{21}$$

Recall now the definition of the (partial) value function defined for any $j \in \mathcal{N}$, any (joint) policy $\pi \in \Pi$ and any reward vector $r \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}_j|}$:

$$v_{\mu,j}^\pi(r) := \langle \lambda_{\mu,j}^\pi, r \rangle = \mathbb{E}_{\mu,\pi}\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_{j,t})\right] \,. \tag{22}$$

We also define for any $i, j \in \mathcal{N}$ and any joint policy $\pi \in \Pi$ the pseudo-reward vector:

$$\bar{r}^{\pi}_{i,j} := \nabla_{\lambda_j} F_i(\lambda^{\pi}_{\mu,1}, \cdots, \lambda^{\pi}_{\mu,N}, \pi_{-i}) \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}_j|}\,. \tag{23}$$

Using this notation and setting $\lambda'_j = \lambda^{\pi'_i,\pi_{-i}}_{\mu,j}$ and $\lambda_j = \lambda^{\pi_i,\pi_{-i}}_{\mu,j}$ in (21) yields

$$\begin{aligned} u_{\mu,i}(\pi'_i, \pi_{-i}) - u_{\mu,i}(\pi_i, \pi_{-i}) &= F_i(\lambda^{\pi'_i,\pi_{-i}}_{\mu,1}, \cdots, \lambda^{\pi'_i,\pi_{-i}}_{\mu,N}, \pi_{-i}) - F_i(\lambda^{\pi_i,\pi_{-i}}_{\mu,1}, \cdots, \lambda^{\pi_i,\pi_{-i}}_{\mu,N}, \pi_{-i}) \\ &\leq \sum_{j=1}^{N} \langle \bar{r}^{\pi}_{i,j}, \lambda^{\pi'_i,\pi_{-i}}_{\mu,j} - \lambda^{\pi_i,\pi_{-i}}_{\mu,j} \rangle \\ &= \sum_{j=1}^{N} v^{\pi'_i,\pi_{-i}}_{\mu,j}(\bar{r}^{\pi}_{i,j}) - v^{\pi_i,\pi_{-i}}_{\mu,j}(\bar{r}^{\pi}_{i,j})\,. \end{aligned} \tag{24}$$

Step 2. Using single-agent gradient domination in an (averaged) MDP. Given the fixed policy π_{-i}, define now for any reward vector $r \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ the following MDP $\mathcal{M} := (\mathcal{S}, \mathcal{A}_i, \bar{P}^{\pi_{-i}}, r, \mu, \gamma)$ where the transition kernel is defined as follows: For any $s, s' \in \mathcal{S}, a_i \in \mathcal{A}_i$,

$$\bar{P}^{\pi_{-i}}(s'|s, a_i) = \sum_{a_{-i}\in\mathcal{A}_{-i}} \pi_{-i}(a_{-i}|s) P(s'|s, a_i, a_{-i})\,, \tag{25}$$

where P is the transition kernel of our initial GUMG.

In this MDP $\mathcal{M}$, observe that $v^{\pi_i,\pi_{-i}}_{\mu,j}(r)$ (as defined in (38)-(41)) is the value function associated to policy π_i and reward r. Using the variational gradient domination property for (single-agent) MDPs (see Proposition 14), we obtain that for every $\pi'_i \in \Pi_i$ and any state distribution $\rho \in \Delta(\mathcal{S})$,

$$v^{\pi^\star_i,\pi_{-i}}_{\mu,j}(r) - v^{\pi_i,\pi_{-i}}_{\mu,j}(r) \leq \left\| \frac{d^{\pi^\star_i,\pi_{-i}}_{\mu}}{d^{\pi_i,\pi_{-i}}_{\rho}} \right\|_\infty \langle \nabla_{\pi_i} v^{\pi}_{\rho,j}(r), \pi'_i - \pi_i \rangle\,, \tag{26}$$

where $\pi^\star_i$ is an optimal policy in the MDP $\mathcal{M}$ (depending on the fixed policy π_{-i}). As a consequence, using (24) and (26) with $r = \bar{r}^{\pi}_{i,j}$, it follows from summing up the above inequality for $j \in \mathcal{N}$ that

$$u_{\mu,i}(\pi'_i, \pi_{-i}) - u_{\mu,i}(\pi_i, \pi_{-i}) \leq C_{\mathcal{G}} \left\langle \sum_{j=1}^{N} \nabla_{\pi_i} v^{\pi}_{\rho,j}(\bar{r}^{\pi}_{i,j}), \pi'_i - \pi_i \right\rangle = C_{\mathcal{G}} \langle \nabla_{\pi_i} u_{\rho,i}(\pi), \pi'_i - \pi_i \rangle\,, \tag{27}$$

where the last identity follows from using the policy gradient expression proved in Proposition 3 (see (40)). To conclude, it remains to take the maximum over π'_i on the right-hand side of the above inequality. □

9.2 Proof of Proposition 2

(i) If $\pi^\star$ is a Nash policy then $\pi^\star$ is FOS.

Proof. This implication is standard in optimization. We provide a proof here for completeness. Let $\pi^\star$ be a Nash policy. By definition of the differentiability of u_i w.r.t π_i at $\pi^\star$, we can write for any $\eta > 0$ and any policy $\pi'_i \in \Pi_i$ (using $\delta_i := \pi'_i - \pi^\star_i$) that

$$u_i(\pi^\star_i + \eta\delta_i, \pi^\star_{-i}) = u_i(\pi^\star_i, \pi^\star_{-i}) + \eta \langle \nabla_{\pi_i} u_i(\pi^\star), \delta_i \rangle + \eta\|\delta_i\| \cdot o_{\eta\to 0}(1)\,. \tag{28}$$

Therefore, we obtain by rearranging the identity

$$\langle \nabla_{\pi_i} u_i(\pi^\star), \pi'_i - \pi^\star_i \rangle = \frac{1}{\eta}(u_i(\pi^\star_i + \eta\,\delta_i, \pi^\star_{-i}) - u_i(\pi^\star_i, \pi^\star_{-i})) - \|\delta_i\| \cdot o_{\eta\to 0}(1)\,. \tag{29}$$

By the definition of a Nash equilibrium policy, it follows that

$$\langle \nabla_{\pi_i} u_i(\pi^\star), \pi'_i - \pi^\star_i \rangle \leq -\|\delta_i\| \cdot o_{\eta\to 0}(1)\,. \tag{30}$$

Taking $\eta \to 0$ yields $\langle \nabla_{\pi_i} u_i(\pi^\star), \pi'_i - \pi^\star_i \rangle \leq 0$, i.e.$\langle v_i(\pi^\star), \pi'_i - \pi^\star_i \rangle \leq 0$. As i is arbitrary in $\mathcal{N}$, the inequality holds for all $i \in \mathcal{N}$ and summing up these inequalities yields $\langle v(\pi^\star), \pi' - \pi^\star \rangle \leq 0$ for all $\pi' \in \Pi$ which means that $\pi^\star$ is a FOS policy. □

(ii) If $\pi^\star$ is a FOS policy then $\pi^\star$ is also a Nash policy.

Proof. The result follows from our gradient domination result of Proposition 1.

Let $\pi^\star = (\pi_i^\star, \pi_{-i}^\star) \in \Pi$ be a FOS policy. By definition it satisfies for all $\pi \in \Pi, \langle v(\pi^\star), \pi - \pi^\star \rangle \leq 0$ for all $\pi \in \Pi$. Choose $\pi = (\pi_i', \pi_{-i}^\star)$ in this FOS inequality to obtain $\langle v_i(\pi^\star), \pi_i' - \pi_i^\star \rangle \leq 0$ for all $i \in \mathcal{N}$ and for all $\pi_i' \in \Pi_i$. By Proposition 1, for all $i \in \mathcal{N}$ and for all $\pi_i' \in \Pi_i$, we have

$$u_i(\pi_i', \pi_{-i}^\star) - u_i(\pi_i^\star, \pi_{-i}^\star) \leq C_{\mathcal{G}} \cdot \max_{\tilde{\pi}_i \in \Pi_i} \langle v_i(\pi^\star), \tilde{\pi}_i - \pi_i^\star \rangle \,. \tag{31}$$

This in turn implies that $u_i(\pi_i', \pi_{-i}^\star) - u_i(\pi_i^\star, \pi_{-i}^\star) \leq 0$ for all $i \in \mathcal{N}$ and for all $\pi_i' \in \Pi_i$ and hence $\pi^\star$ is a Nash policy. □

9.3 Proof of Lemma 1

Lemma 1. *A policy $\pi^\star$ is FOS if and only if it is a fixed point of the projected gradient dynamics, i.e. it satisfies: $\pi^\star = Proj_\Pi(\pi^\star + \eta\, v(\pi^\star)), \forall \eta > 0$.*

Proof. Suppose that $\pi^\star = \mathrm{Proj}_\Pi(\pi^\star + \eta\, v(\pi^\star)), \forall \eta > 0$. By characterization of the projection, we have

$$\forall \pi \in \Pi, \forall \eta > 0, \langle \pi - \pi^\star, (\pi^\star + \eta\, v(\pi^\star)) - \pi^\star \rangle \leq 0 \,. \tag{32}$$

By simplifying and dividing by $\eta > 0$ the above inequality, we obtain the desired variational inequality,

$$\forall \pi \in \Pi, \forall \eta > 0, \langle \pi - \pi^\star, v(\pi^\star) \rangle \leq 0 \,, \tag{33}$$

i.e. $\pi^\star$ is a FOS policy. The converse implication also holds since (32) is a characterization of the projection, i.e. if $\pi^\star \in \Pi$ satisfies (32) then it must be the projection of $\pi^\star + \eta\, v(\pi^\star)$ on the policy space Π. □

9.4 Proof of Theorem 1

The proof follows from combining Prop. 2 and Lem. 1. Prop. 2 shows that a (joint) policy $\pi^\star \in \Pi$ is a Nash policy in a GUMG if and only if it is FOS. Lem. 1 shows that a policy is FOS if and only if it is a fixed point of the projected gradient dynamics.

9.5 Proof of Corollary 1

The proof of existence of a NE is in the main part. We now use this result to further show the existence of Markov perfect equilibrium (MPE).

Let $\varepsilon > 0$. Let $s \in \mathcal{S}$ be an arbitrary state. Consider the mixed initial state distribution $\mu_\varepsilon := (1-\varepsilon)\delta_s + \varepsilon \frac{1}{|\mathcal{S}|}$ where δ_s is the Dirac measure at the state $s \in \mathcal{S}$, defined s.t. $\delta_s(s) = 1$ and $\delta_s(s') = 0$ for any $s' \neq s$. Since μ_ε is fully mixed and hence implies that Assumption 2 is satisfied (i.e. $d_{\mu_\varepsilon}^\pi(s) > 0$ for all $\tilde{s} \in \mathcal{S}, \pi \in \Pi$ since it always holds that $d_{\mu_\varepsilon}^\pi(\tilde{s}) \geq \mu_\varepsilon(\tilde{s})$) We can apply the NE existence result using the initial state distribution μ_ε established in the first part of Corollary 1 to obtain the existence of a NE $\pi_\varepsilon^\star$ (a priori depending on ε). By definition this NE satisfies the following:

$$\forall \pi_i \in \Pi_i, \quad u_{\mu,i}(\pi_i, \pi_{\varepsilon,-i}^\star) \leq u_{\mu,i}(\pi_\varepsilon^\star) \,. \tag{34}$$

Let (ε_n) be a sequence of positive numbers converging to zero. The sequence of policies $(\pi_{\varepsilon_n}^\star)$ evolves in the compact space of policies Π. By the Bolzano-Weierstrass theorem, we can extract from it a converging subsequence $(\pi_{\varepsilon_{\varphi(n)}}^\star)$ where φ is an increasing function over integers. Denote by $\pi^\star$ the limit of the subsequence $(\pi_{\varepsilon_{\varphi(n)}}^\star)$. It follows from (34) that for any integer n,

$$\forall \pi_i \in \Pi_i, \quad u_{\mu_{\varepsilon_{\varphi(n)}},i}(\pi_i, \pi_{\varepsilon_{\varphi(n)},-i}^\star) \leq u_{\mu_{\varepsilon_{\varphi(n)}},i}(\pi_{\varepsilon_{\varphi(n)}}^\star) \,. \tag{35}$$

We now note that the utility function $u_{\mu,i}$ is continuous w.r.t. both its policy argument and the initial distribution μ. Indeed, this follows from observing its definition (5), noting that F_i is supposed to be continuous and

the occupancy measures are continuous functions in both their policy variable (see e.g., Prop. 15) and the initial state distribution μ (by linearity of the occupancy measure w.r.t. μ). Finally, taking the limit when $n \to \infty$ in (35) yields:

$$\forall \pi_i \in \Pi_i, \quad u_{\delta_s,i}(\pi_i, \pi^\star_{-i}) \leq u_{\delta_s,i}(\pi^\star)\,, \tag{36}$$

which proves that $\pi^\star$ is a MPE.

Remark 5. *A stronger result can be shown: an NE for a given full-support initial state distribution is also an MPE. Indeed, it can be proved that a NE π for a given full-support initial state distribution μ is also a NE for any other full-support initial state distribution $\rho \neq \mu$ using per-player gradient domination (Prop. 1) combined with the policy gradient theorem (Prop. 3-(ii)). Therefore a NE is independent of the initial full-support state distribution. Then a similar limit argument as above shows that this same NE policy π is a MPE.*

10 PROOFS FOR SECTION 4: Proof of Proposition 3

Proof. We prove each result separately.

Proof of (i). Using the chain rule, we have for every $i \in \mathcal{N}$,

$$\nabla_{\pi_i} u_{\mu,i}(\pi) = \sum_{j=1}^{N} [\nabla_{\pi_i} \lambda^\pi_{\mu,j}]^T \nabla_{\lambda_j} F_i(\lambda^\pi_{\mu,1}, \cdots, \lambda^\pi_{\mu,N}, \pi_{-i})\,. \tag{37}$$

Now define for all $j \in \mathcal{N}, r_j \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}_j|}$,

$$v^\pi_{\mu,j}(r_j) = \langle \lambda^\pi_{\mu,j}, r_j \rangle = \mathbb{E}_{\mu,\pi}\left[\sum_{t=0}^{+\infty} \gamma^t r_j(s_t, a^j_t)\right]\,. \tag{38}$$

It follows that

$$\nabla_{\pi_i} v^\pi_{\mu,j}(r_j) = \sum_{s\in\mathcal{S}, a_j\in\mathcal{A}_j} \nabla_{\pi_i} \lambda^\pi_{\mu,j}(s,a_j) r_j(s,a_j) = [\nabla_{\pi_i} \lambda^\pi_{\mu,j}]^\top r_j\,. \tag{39}$$

Therefore, we obtain from (37) and (39) together that

$$\nabla_{\pi_i} u_{\mu,i}(\pi) = \sum_{j=1}^{N} \nabla_{\pi_i} v^\pi_{\mu,j}(r_j)\big|_{r_j = \nabla_{\lambda_j} F_i(\lambda^\pi_{\mu,1},\cdots,\lambda^\pi_{\mu,N},\pi_{-i})} = \sum_{j=1}^{N} \nabla_{\pi_i} v^\pi_{\mu,j}(\bar{r}^\pi_{i,j})\,, \tag{40}$$

where we recall that $\bar{r}^\pi_{ij} = \nabla_{\lambda_j} F_i(\lambda^\pi_{\mu,1}, \cdots, \lambda^\pi_{\mu,N}, \pi_{-i})$ for any $j \in \mathcal{N}$. Now define the value function

$$\tilde{v}^\pi_{\mu,j}(\tilde{r}_j) = \mathbb{E}_{\mu,\pi}\left[\sum_{t=0}^{+\infty} \gamma^t \tilde{r}_j(s_t, a_{j,t}, a_{-j,t})\right]\,, \tag{41}$$

where $\tilde{r}_j(s_t, a_t) = r_j(s_t, a_{j,t})$ for all $a_t = (a_{j,t}, a_{-j,t})$. Given the above notation, we can now apply the standard policy gradient theorem for MDPs (single-agent RL, see Proposition 12) to write

$$\nabla_\pi \tilde{v}^\pi_{\mu,j}(\tilde{r}_j) = \mathbb{E}_{\mu,\pi}\left[\sum_{t=0}^{+\infty} \gamma^t \tilde{r}_j(s_t, a_t) \sum_{t'=0}^{t} \nabla_\pi \log \pi(a_{t'}|s_{t'})\right]\,. \tag{42}$$

Now using our product policy form $\left(\pi(a|s) = \prod_{i=1}^N \pi_i(a_i|s)\right)$ and the definition of $\tilde{r}_j$ (which implies that for any policy $\pi \in \Pi, v^\pi_{\mu,j}(r_j) = \tilde{v}^\pi_{\mu,j}(\tilde{r}_j)$, see (38) and (41)), it follows that

$$\nabla_{\pi_i} v^\pi_{\mu,j}(r_j) = \nabla_{\pi_i} \tilde{v}^\pi_{\mu,j}(\tilde{r}_j) = \mathbb{E}_{\mu,\pi}\left[\sum_{t=0}^{+\infty} \gamma^t r_j(s_t, a_{j,t}) \sum_{t'=0}^{t} \nabla_{\pi_i} \log \pi_i(a_{i,t'}|s_{t'})\right]\,, \tag{43}$$

which gives the desired result when combined with (40).

Proof of (ii). Notice first from (40) that for every state-action pair $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$,

$$\frac{\partial u_{\mu,i}(\pi)}{\partial \pi_i(a_i|s)} = \sum_{j=1}^{N} \frac{\partial v^{\pi}_{\mu,j}(r)}{\partial \pi_i(a_i|s)}_{|r=\bar{r}^{\pi}_{i,j}}, \tag{44}$$

By the classical policy gradient theorem (see Proposition 12), we have for every $j \in [N]$ and every $r \in \mathbb{R}^{|S|\cdot|A|}$,

$$\nabla_{\pi} v^{\pi}_{\mu,j}(r) = \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \lambda^{\pi}_{\mu}(s', a')\, Q^{\pi}_{s',a'}(r)\, \nabla \log \pi(a'|s')\,. \tag{45}$$

Hence, it follows that for every $r \in \mathbb{R}^{|S|\cdot|A|}$ and every state-action pair $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$,

$$\frac{\partial v^{\pi}_{\mu,j}(r)}{\partial \pi_i(a_i|s)} = \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} d^{\pi}_{\mu}(s')\pi(a'|s')\, Q^{\pi}_{s',a'}(r)\, \frac{\partial \log \pi(a'|s')}{\partial \pi_i(a_i|s)}\,. \tag{46}$$

Since $\pi(a'|s') = \prod_{i=1}^{N} \pi_i(a_i|s)$ and π_i is a direct parameterized policy, we have

$$\frac{\partial \log \pi(a'|s')}{\partial \pi_i(a_i|s)} = \frac{\partial \log \pi_i(a'|s')}{\partial \pi_i(a_i|s)} = \frac{1}{\pi_i(a_i|s)} \mathbf{1}_{\{s'=s, a'_i=a_i\}}\,, \tag{47}$$

where we use the notation $\mathbf{1}_{\mathcal{X}}$ for the indicator function of the set $\mathcal{X}$. Therefore, we have:

$$\begin{aligned}
\frac{\partial v^{\pi}_{\mu,j}(r)}{\partial \pi_i(a_i|s)} &= \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} d^{\pi}_{\mu}(s')\pi(a'|s')\, Q^{\pi}_{s',a'}(r^{\pi}_i) \frac{1}{\pi_i(a_i|s)} \mathbf{1}_{\{s'=s, a'_i=a_i\}} \\
&= \frac{1}{1-\gamma} d^{\pi}_{\mu}(s) \sum_{a'_{-i} \in \mathcal{A}_{-i}} \pi_{-i}(a'_{-i}|s) Q^{\pi}_{s,(a_i,a'_{-i})}(r) \\
&= \frac{1}{1-\gamma} d^{\pi}_{\mu}(s) \bar{Q}^{\pi}_{s,a_i}(r)\,,
\end{aligned} \tag{48}$$

where the average Q-function $\bar{Q}^{\pi}(r)$ is defined for any joint policy $\pi \in \Pi$, any reward $r \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ and for any state-action pair $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$ as follows:

$$\bar{Q}^{\pi}_{s,a_i}(r) := \sum_{a_{-i} \in \mathcal{A}_{-i}} \pi_{-i}(a_{-i}|s)\, Q^{\pi}_{s,(a_i,a_{-i})}(r)\,. \tag{49}$$

Using (44) together with (48), we obtain

$$\frac{\partial u_{\mu,i}(\pi)}{\partial \pi_i(a_i|s)} = \sum_{j=1}^{N} \frac{\partial v^{\pi}_{\mu,j}(r)}{\partial \pi_i(a_i|s)}_{|r=\bar{r}^{\pi}_{i,j}} = \frac{1}{1-\gamma} d^{\pi}_{\mu}(s) \sum_{j=1}^{N} \bar{Q}^{\pi}_{s,a_i}(\bar{r}^{\pi}_{i,j})\,, \tag{50}$$

which concludes the proof. □

11 PROOFS FOR SECTION 5

11.1 Proof of Proposition 4

Proof. Let $i \in \mathcal{N}, \pi, \pi' \in \Pi$. We show smoothness by controlling each one of the partial gradients w.r.t. policy π_i of agent i since:

$$\|\nabla_{\pi} u_{\mu,i}(\pi) - \nabla_{\pi} u_{\mu,i}(\pi')\|_2^2 = \sum_{i=1}^{N} \|\nabla_{\pi_i} u_{\mu,i}(\pi) - \nabla_{\pi_i} u_{\mu,i}(\pi')\|_2^2\,. \tag{51}$$

Using Prop. 3 (see eq. (40)), we can write the following decomposition:

$$\begin{aligned}\nabla_{\pi_i} u_{\mu,i}(\pi) - \nabla_{\pi_i} u_{\mu,i}(\pi') &= \sum_{j=1}^{N} \nabla_{\pi_i} v_{\mu,j}^{\pi}(\bar{r}_{i,j}^{\pi}) - \nabla_{\pi_i} v_{\mu,j}^{\pi'}(\bar{r}_{i,j}^{\pi'}) \\ &= \sum_{j=1}^{N} \nabla_{\pi_i} v_{\mu,j}^{\pi}(\bar{r}_{i,j}^{\pi}) - \nabla_{\pi_i} v_{\mu,j}^{\pi'}(\bar{r}_{i,j}^{\pi}) + \nabla_{\pi_i} v_{\mu,j}^{\pi'}(\bar{r}_{i,j}^{\pi}) - \nabla_{\pi_i} v_{\mu,j}^{\pi'}(\bar{r}_{i,j}^{\pi'}) \,. \end{aligned} \tag{52}$$

Using Lemma D.7 in Giannou et al. (2022) and boundedness of the pseudo-rewards (Assumption 3), we obtain

$$\|\nabla_{\pi_i} v_{\mu,j}^{\pi}(\bar{r}_{i,j}^{\pi}) - \nabla_{\pi_i} v_{\mu,j}^{\pi'}(\bar{r}_{i,j}^{\pi})\| \leq \frac{3\sqrt{|\mathcal{A}_i|} l_\infty}{(1-\gamma)^3} \sum_{k=1}^{N} \sqrt{|\mathcal{A}_k|} \cdot \|\pi_k - \pi_k'\|_2 \,. \tag{53}$$

Using Prop. 13 and then the smoothness assumption 3 together with the definition of pseudo-rewards, we get

$$\|\nabla_{\pi_i} v_{\mu,j}^{\pi'}(\bar{r}_{i,j}^{\pi}) - \nabla_{\pi_i} v_{\mu,j}^{\pi'}(\bar{r}_{i,j}^{\pi'})\|_2 \leq \frac{\sqrt{|\mathcal{A}_i|}}{(1-\gamma)^2} \|\bar{r}_{i,j}^{\pi} - \bar{r}_{i,j}^{\pi'}\|_\infty \leq \frac{\sqrt{|\mathcal{A}_i|} L}{(1-\gamma)^2} (\|\lambda_{\mu,1:N}^{\pi} - \lambda_{\mu,1:N}^{\pi'}\|_1 + \|\pi_{-i} - \pi_{-i}'\|_1) \,. \tag{54}$$

The next lemma shows that policy deviations only scale in norm with the sum of the sizes of the action spaces rather than their product.

Lemma 2. *The following statements hold when using product policies:*

(i) $\forall s \in \mathcal{S}$, $\|\pi_{-i}(\cdot|s) - \pi_{-i}'(\cdot|s)\|_1 \leq \sum_{k=1, k\neq i} \|\pi_k(\cdot|s) - \pi_k'(\cdot|s)\|_1$,

(ii) $\forall s \in \mathcal{S}$, $\|\pi_{-i}(\cdot|s) - \pi_{-i}'(\cdot|s)\|_1 \leq \sqrt{\sum_{k=1}^{N} |\mathcal{A}_k|} \cdot \|\pi - \pi'\|_2$,

(iii) $\sum_{k=1}^{N} \|\pi_k - \tilde{\pi}_k\|_1 \leq \sqrt{|\mathcal{S}| \cdot \sum_{k=1}^{N} |\mathcal{A}_k|} \cdot \|\pi - \tilde{\pi}\|_2$.

(iv) $\|\pi - \tilde{\pi}\|_1 \leq \sqrt{|\mathcal{S}| \cdot \sum_{k=1}^{N} |\mathcal{A}_k|} \cdot \|\pi - \tilde{\pi}\|_2$.

Proof. See complete proof in section 11.2 below. □

We now upper bound each one of the terms in (54). For the first term, we have:

$$\begin{aligned}\|\lambda_{\mu,1:N}^{\pi} - \lambda_{\mu,1:N}^{\pi'}\|_1 &= \sum_{i=1}^{N} \sum_{s\in\mathcal{S}, a_i\in\mathcal{A}_i} |d_\mu^{\pi}(s)\pi_i(a_i|s) - d_\mu^{\pi'}(s)\pi_i'(a_i|s)| \\ &\leq \sum_{i=1}^{N} \sum_{s\in\mathcal{S}, a_i\in\mathcal{A}_i} |d_\mu^{\pi}(s) - d_\mu^{\pi'}(s)|\pi_i(a_i|s) + d_\mu^{\pi'}(s)|\pi_i(a_i|s) - \pi_i'(a_i|s)| \\ &\leq N\|d_\mu^{\pi} - d_\mu^{\pi'}\|_1 + \sum_{i=1}^{N} \sum_{s\in\mathcal{S}, a_i\in\mathcal{A}_i} d_\mu^{\pi'}(s)|\pi_i(a_i|s) - \pi_i'(a_i|s)| \\ &\leq N\|d_\mu^{\pi} - d_\mu^{\pi'}\|_1 + \sum_{i=1}^{N} \|\pi_i - \pi_i'\|_1 \\ &\leq \frac{N\gamma}{1-\gamma}\|\pi - \tilde{\pi}\|_1 + \sqrt{|\mathcal{S}| \cdot \sum_{k=1}^{N} |\mathcal{A}_k|}\|\pi - \pi'\|_2 \\ &\leq \left(1 + \frac{N\gamma}{1-\gamma}\right) \sqrt{|\mathcal{S}| \cdot \sum_{k=1}^{N} |\mathcal{A}_k|}\|\pi - \pi'\|_2 \,, \end{aligned} \tag{55}$$

where we used Proposition 15 and Lemma 2.

Using again Lemma 2 yields $\|\pi_{-i} - \pi'_{-i}\|_1 \leq |\mathcal{S}| \cdot \sqrt{\sum_{k=1}^N |\mathcal{A}_k|} \cdot \|\pi - \pi'\|_2$ and we obtain then using (55):

$$\|\lambda^{\pi}_{\mu,1:N} - \lambda^{\pi'}_{\mu,1:N}\|_1 + \|\pi_{-i} - \pi'_{-i}\|_1 \leq \sqrt{\sum_{k=1}^N |\mathcal{A}_k|} \left[\left(1 + \frac{N\gamma}{1-\gamma}\right) \sqrt{|\mathcal{S}|} + |\mathcal{S}| \right] \|\pi - \pi'\|_2 \,. \tag{56}$$

Plugging the above inequality into (54) and then using (53) and (52) yields:

$$\begin{aligned} \|\nabla_{\pi_i} u_{\mu,i}(\pi) - \nabla_{\pi_i} u_{\mu,i}(\pi')\|_2 \leq \frac{3N\sqrt{|\mathcal{A}_i|} l_\infty}{(1-\gamma)^3} \sum_{k=1}^N \sqrt{|\mathcal{A}_k|} \cdot \|\pi_k - \pi'_k\|_2 \\ + \frac{N\sqrt{|\mathcal{A}_i|} L}{(1-\gamma)^2} \sqrt{\sum_{k=1}^N |\mathcal{A}_k|} \left[\left(1 + \frac{N\gamma}{1-\gamma}\right) \sqrt{|\mathcal{S}||} + |\mathcal{S}| \right] \|\pi - \pi'\|_2 \,. \end{aligned} \tag{57}$$

Using Cauchy-Schwarz's inequality gives:

$$\sum_{k=1}^N \sqrt{|\mathcal{A}_k|} \cdot \|\pi_k - \pi'_k\|_2 \leq \sqrt{\sum_{k=1}^N |\mathcal{A}_k|} \cdot \sqrt{\sum_{i=1}^N \|\pi_i - \pi'_i\|_2^2} \leq \sqrt{\sum_{k=1}^N |\mathcal{A}_k|} \cdot \|\pi - \pi'\|_2 \,. \tag{58}$$

Using the above inequality in (57), we obtain:

$$\|\nabla_{\pi_i} u_{\mu,i}(\pi) - \nabla_{\pi_i} u_{\mu,i}(\pi')\|_2 \leq \beta_i \|\pi - \pi'\|_2 \,, \tag{59}$$

where the constant β_i is defined as follows:

$$\beta_i := \frac{3N\sqrt{|\mathcal{A}_i|} l_\infty}{(1-\gamma)^3} \sqrt{\sum_{k=1}^N |\mathcal{A}_k|} + \frac{N\sqrt{|\mathcal{A}_i|} L}{(1-\gamma)^2} \sqrt{\sum_{k=1}^N |\mathcal{A}_k|} \left[\left(1 + \frac{N\gamma}{1-\gamma}\right) \sqrt{|\mathcal{S}|} + |\mathcal{S}| \right] . \tag{60}$$

It follows from this that:

$$\|\nabla_{\pi} u_{\mu,i}(\pi) - \nabla_{\pi} u_{\mu,i}(\pi')\|_2 \leq \sum_{i=1}^N \|\nabla_{\pi_i} u_{\mu,i}(\pi) - \nabla_{\pi_i} u_{\mu,i}(\pi')\|_2 \leq \left(\sum_{i=1}^N \beta_i \right) \cdot \|\pi - \pi'\|_2 \,. \tag{61}$$

Using again Cauchy-Schwarz's inequality gives $\sum_{i=1}^N \sqrt{|\mathcal{A}_i|} \leq \sqrt{N \sum_{i=1}^N |\mathcal{A}_i|}$, we have the following bound:

$$\sum_{i=1}^N \beta_i = \frac{3N l_\infty \sum_{i=1}^N \sqrt{|\mathcal{A}_i|}}{(1-\gamma)^3} \sqrt{\sum_{k=1}^N |\mathcal{A}_k|} + \frac{N\sqrt{|\mathcal{A}_i|} L}{(1-\gamma)^2} \sqrt{\sum_{k=1}^N |\mathcal{A}_k|} \left[\left(1 + \frac{N\gamma}{1-\gamma}\right) \sqrt{|\mathcal{S}||} + |\mathcal{S}| \right] \tag{62}$$

$$\leq \beta := \frac{N^{\frac{3}{2}} \sum_{k=1}^N |\mathcal{A}_k|}{(1-\gamma)^2} (3l_\infty + L \left[\left(1 + \frac{N\gamma}{1-\gamma}\right) \sqrt{|\mathcal{S}|} + |\mathcal{S}| \right]) \,. \tag{63}$$

Finally, we obtain our desired β-smoothness bound:

$$\|\nabla_{\pi} u_i(\pi) - \nabla_{\pi} u_i(\pi')\|_2 \leq \beta \|\pi - \pi'\|_2 \,, \tag{64}$$

where the smoothness constant β is defined as follows:

$$\beta := \frac{N^{\frac{3}{2}} \sum_{k=1}^N |\mathcal{A}_k|}{(1-\gamma)^2} \left(3l_\infty + L \left[\left(1 + \frac{N\gamma}{1-\gamma}\right) \sqrt{|\mathcal{S}|} + |\mathcal{S}| \right] \right) \,, \tag{65}$$

and this concludes the proof. □

11.2 Proof of Lemma 2

Proof. We prove each one of the points in what follows.

Item (i). First we use the product form of the joint policy to write:

$$\begin{aligned}|\pi_{-i}(a_{-i}|s) - \pi'_{-i}(a_{-i}|s)| &= \left|\prod_{j=1,j\neq i}^{N} \pi_j(a_j|s) - \prod_{j=1,j\neq i}^{N} \pi'_j(a_j|s)\right| \\ &= \left|\sum_{k=1,k\neq i}^{N} (\pi_k - \pi'_k)(a_k|s) \prod_{l=1,l\neq i}^{k-1} \pi_l(a_l|s) \prod_{l=k+1,l\neq i}^{N} \pi'_l(a_l|s)\right| . \end{aligned} \tag{66}$$

As a consequence, we obtain by summing up the previous inequality over $a_{-i} \in \mathcal{A}_{-i}$:

$$\begin{aligned}\sum_{a_{-i}\in\mathcal{A}_{-i}} |\pi_{-i}(a_{-i}|s) - \pi'_{-i}(a_{-i}|s)| &\leq \sum_{k=1,k\neq i}^{N} \sum_{a_k\in\mathcal{A}_k} |(\pi_k - \pi'_k)(a_k|s)| \sum_{a_{-k}\in\mathcal{A}_{-k}} \prod_{l=1,l\neq i}^{k-1} \pi_l(a_l|s) \prod_{l=k+1,l\neq i}^{N} \pi'_l(a_l|s) \\ &\leq \sum_{k=1,k\neq i} \|\pi_k(\cdot|s) - \pi'_k(\cdot|s)\|_1 \\ &\leq \sum_{k=1,k\neq i} \sqrt{|\mathcal{A}_k|}\|\pi_k(\cdot|s) - \pi'_k(\cdot|s)\|_2 , \end{aligned} \tag{67}$$

where the second inequality follows from noting that $\sum_{a_{-k}\in\mathcal{A}_{-k}} \prod_{l=1,l\neq i}^{k-1} \pi_l(a_l|s) \prod_{l=k+1,l\neq i}^{N} \pi'_l(a_l|s) = 1$ by the definition of policies.

Item (ii). Using the first item, we can write for any $s \in \mathcal{S}$,

$$\begin{aligned}\|\pi_{-i}(\cdot|s) - \pi'_{-i}(\cdot|s)\|_1 &\leq \sum_{k=1,k\neq i} \|\pi_k(\cdot|s) - \pi'_k(\cdot|s)\|_1 \\ &\leq \sum_{k=1,k\neq i} \sqrt{|\mathcal{A}_k|}\|\pi_k(\cdot|s) - \pi'_k(\cdot|s)\|_2 , \\ &\leq \sqrt{\sum_{k=1}^{N} |\mathcal{A}_k|} \cdot \|\pi - \pi'\|_2 , \end{aligned} \tag{68}$$

where the last inequality follows from the Cauchy-Schwarz inequality.

Item (iii). We upper bound the norm as follows:

$$\begin{aligned}\sum_{k=1}^{N} \|\pi_k - \tilde{\pi}_k\|_1 &\leq \sum_{k=1}^{N} \sqrt{|\mathcal{S}| \cdot |\mathcal{A}_k|} \cdot \|\pi_k - \tilde{\pi}_k\|_2 \\ &\leq \sqrt{|\mathcal{S}| \cdot \sum_{k=1}^{N} |\mathcal{A}_k|} \sqrt{\sum_{k=1}^{N} \|\pi_k - \tilde{\pi}_k\|_2^2} \\ &= \sqrt{|\mathcal{S}| \cdot \sum_{k=1}^{N} |\mathcal{A}_k|} \cdot \|\pi - \tilde{\pi}\|_2 , \end{aligned} \tag{69}$$

where the first inequality follows similarly to the derivations in (66) and (67).

Item (iv). Observe using similar derivations as in the proof of the first two items that:

$$\|\pi - \tilde{\pi}\|_1 = \sum_{s\in\mathcal{S},a\in\mathcal{A}} |\pi(a|s) - \tilde{\pi}(a|s)| \leq \sum_{s\in\mathcal{S}} \sum_{k=1}^{N} \|\pi_k(\cdot|s) - \tilde{\pi}_k(\cdot|s)\|_1 = \sum_{k=1}^{N} \|\pi_k - \tilde{\pi}_k\|_1 , \tag{70}$$

and use the previous item. □

11.3 Proof of Thm. 2

Proof sketch. The proof follows a similar strategy to prior work (e.g. Leonardos et al. (2022); Zhang et al. (2024)) and consists of three main steps:

1. Relate the Nash gap to first-order stationarity using gradient domination (Proposition 1): see end of the proof in this section.
2. Relate the first-order stationarity quantity to the gradient mapping (using Proposition 6).
3. Establish a first-order stationarity iteration complexity using the gradient mapping and the smoothness of the potential function (see Proposition 5). This part of the proof is standard in non-convex smooth optimization.

Proof. In the remainder of this section, we provide a complete proof.

We start by presenting a standard result for smooth non-convex optimization reported and proved in e.g. (Zhang et al., 2024, Lemma 12).

To state the result, we first define the standard gradient mapping to quantify first-order stationarity in our constrained setting: For any policy $\pi \in \Pi$ and any greediness policy parameter $\alpha > 0$, we define:

$$G^{\eta,\alpha}(\pi) := \frac{1}{\eta}\left(\mathrm{Proj}(\pi + \eta\nabla\Phi(\pi)) - \pi\right) , \tag{71}$$

where $\Phi : \Pi \to \mathbb{R}$ is a given smooth function. Note that we will use the simpler notation

$$G^{\eta}(\pi) := G^{\eta,\alpha}(\pi) \quad \text{when} \quad \alpha = 0 , \tag{72}$$

as it will be the case in the rest of this section since α-greedy policies are not needed in the exact gradient setting.

Proposition 5 (e.g. Lemma 12, Zhang et al. (2024))**.** *Let $\Phi : \Pi \to \mathbb{R}$ be β-smooth and bounded, i.e. there exist $\Phi_{\min}, \Phi_{\max} > 0$ s.t. for all $\pi \in \Pi, \Phi_{\min} \leq \Phi(\pi) \leq \Phi_{\max}$. Then running projected gradient descent:*

$$\pi^{t+1} = Proj_{\Pi}(\pi^t - \eta\nabla_{\pi}\Phi(\pi^t)) , \tag{73}$$

with stepsize $\eta \leq 1/\beta$ guarantees that $\lim_{t\to+\infty} \|G^{\eta}(\pi^t)\|_2 = 0$. Moreover, we also have:

$$\frac{1}{T}\sum_{t=1}^{T} \|G^{\eta}(\pi^t)\|_2^2 \leq \frac{2\beta(\Phi_{\max} - \Phi_{\min})}{T} . \tag{74}$$

The following result relates FOS with the gradient mapping norm in the deterministic setting.

Proposition 6 (Lemma 11, Zhang et al. (2024))**.** *Let $\Phi : \Pi \to \mathbb{R}$ be β-smooth for some constant $\beta > 0$ and let $\pi^+ = Proj_{\Pi^\alpha}(\pi + \eta\nabla\Phi(\pi))$ for $\eta > 0$. Then for all $\pi' \in \Pi^\alpha$,*

$$\langle \nabla\Phi(\pi^+), \pi' - \pi^+\rangle \leq (1 + \eta\beta)\|G^{\eta,\alpha}(\pi)\| \cdot \|\pi' - \pi^+\| . \tag{75}$$

Moreover, we have

$$\max_{\bar{\pi}_i \in \Pi_i} \langle \nabla_{\pi_i}\Phi(\pi^+), \bar{\pi}_i - \pi_i^+\rangle \leq 2(1 + \eta\beta)\|G^{\eta,\alpha}(\pi)\| + \frac{2Nl_\infty}{(1-\gamma)^2}\alpha . \tag{76}$$

The proof consists in connecting the Nash gap to FOS via agent-wise gradient domination and then using the link between FOS and the gradient mapping as well as the gradient mapping rate which follows from smoothness of the potential function.

Using Propositions 5 and 6 together with our agent-wise gradient domination (Proposition 1), we are now ready to complete the proof.

End of Proof of Thm. 2. By definition of the Nash equilibrium gap and using agent-wise gradient domination (Proposition 1), we have for every $i \in \mathcal{N}, t \geq 0$:

$$\text{NE-Gap}_i(\pi^{t+1}) \leq \frac{C_{\mathcal{G}}}{1-\gamma} \max_{\pi_i' \in \Pi_i} \langle \pi_i' - \pi_i^{t+1}, \nabla_{\pi_i} u_i(\pi^{t+1})\rangle . \tag{77}$$

Using the potential structure, we have $\nabla_{\pi_i} u_i(\pi^{t+1}) = \nabla_{\pi_i}\phi(\pi^{t+1})$ and we obtain:

$$\text{NE-Gap}_i(\pi^{t+1}) \leq \frac{C_{\mathcal{G}}}{1-\gamma} \max_{\pi_i' \in \Pi_i} \langle \pi_i' - \pi_i^{t+1}, \nabla_{\pi_i}\phi(\pi^{t+1})\rangle \,. \tag{78}$$

Using now Proposition 6 with $\alpha = 0$ together with the above inequality and taking the maximum over $i \in \mathcal{N}$ yields:

$$\text{NE-Gap}(\pi^{t+1}) \leq \frac{2C_{\mathcal{G}}(1+\eta\beta)}{1-\gamma} \|G^{\eta}(\pi^{t+1})\|_2 \,. \tag{79}$$

To conclude, it remains to sum up the above inequalities over $t \in [T-1]$, use Jensen's inequality with the square root function and Proposition 5 to obtain:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \text{NE-Gap}(\pi^{t+1}) &\leq \frac{2C_{\mathcal{G}}(1+\eta\beta)}{1-\gamma} \frac{1}{T}\sum_{t=0}^{T-1} \|G^{\eta}(\pi^{t+1})\|_2 \\
&\leq \frac{2C_{\mathcal{G}}(1+\eta\beta)}{1-\gamma} \sqrt{\frac{1}{T}\sum_{t=1}^{T} \|G^{\eta}(\pi^{t})\|_2^2} \\
&\leq \frac{4C_{\mathcal{G}}}{1-\gamma} \sqrt{\frac{2\beta(\Phi_{\max} - \Phi_{\min})}{T}} \,,
\end{aligned} \tag{80}$$

where the last step also uses the condition $\eta \leq 1/\beta$. It follows that for any $\varepsilon > 0$, if

$$T \geq \frac{32 C_{\mathcal{G}}^2 \beta(\Phi_{\max} - \Phi_{\min})}{(1-\gamma)^2 \varepsilon^2} \,, \tag{81}$$

then there exists $t \in [T]$ s.t. $\text{NE-Gap}(\pi^t) \leq \varepsilon$.

11.4 Proof of Thm. 3

Proof sketch and technical comparison to prior work. The proof follows similar lines as the proof of Theorem 2 but we now need to account for the errors due to inexactness of gradients used in the policy update rule. In particular the following steps deviate from the previous deterministic setting:

- As we prove a high-probability result, the first-order stationarity guarantee now holds pathwise (i.e. without expectation) and has an additional error term due to inexact gradients (see Proposition 7).
- Our policy gradient estimator in the generative model setting differs from both Leonardos et al. (2022) (which does not consider the generative model setting) and Zhang et al. (2024). Compared to the latter we rather use trajectory sampling with minibatches and we do not resort to model estimation (estimating the transition kernel). Our resulting sample complexity of $\tilde{O}(\varepsilon^{-4})$ in the generative model setting improves over both the results of Leonardos et al. (2022) and Zhang et al. (2024).
- To control inexactness of gradients (see Proposition 9), we decompose the errors into a bias term of the order γ^H and a sampling error depending on the minibatch size M (using concentration inequalities). Then we select the minibatch size M to guarantee that the sampling error is of the same order as the bias term. This leads to a minibatch size of order $M = \tilde{O}(\gamma^{-2H})$. It remains to choose the horizon H logarithmically in the inverse desired precision to control the bias and hence choose the corresponding minibatch size which is as a consequence of order $\tilde{O}(\varepsilon^{-2})$.

Proof. We now develop the proof in details.

We introduce the stochastic gradient mapping similarly to its deterministic counterpart introduced in (71). For any policy $\pi \in \Pi$ and any greediness policy parameter $\alpha \geq 0$, we define for all $t \geq 0$,

$$\hat{G}^{\eta,\alpha}(\pi^t) := \frac{1}{\eta}\left(\text{Proj}_{\Pi^\alpha}(\pi + \eta g^t) - \pi\right) \,, \tag{82}$$

where for every $i \in [N], g_i^t := \hat{\nabla}_{\pi_i} u_{\mu,i}(\pi^t)$.

Proposition 7. *Let $\Phi : \Pi \to \mathbb{R}$ be β-smooth and bounded, i.e. there exist $\Phi_{\min}, \Phi_{\max} > 0$ s.t. for all $\pi \in \Pi, \Phi_{\min} \leq \Phi(\pi) \leq \Phi_{\max}$. Let (π^t) be the sequence of policies generated by running stochastic projected gradient descent:*[3]

$$\pi^{t+1} = Proj_{\Pi}(\pi^t - \eta g^t)\,, \quad g_i^t = \hat{\nabla}_{\pi_i} u_{\mu,i}(\pi^t)\,, \forall i \in [N]\,, \tag{83}$$

with stepsize $\eta \leq 1/(1+\beta)$. If there exists $\tau > 0$ s.t. for all $t, \|g^t - \nabla\Phi(\pi^t)\| \leq \tau$, then we have for all $t \geq 1$:

$$\frac{1}{T}\sum_{t=1}^{T} \|\hat{G}^{\eta,\alpha}(\pi^t)\|_2^2 \leq \frac{2(\Phi_{\max} - \Phi_{\min})}{\eta T} + \frac{1}{\eta}\tau^2\,. \tag{84}$$

Proof. By β-smoothness of Φ, we have for every time step $t \geq 0$:

$$\begin{aligned}
\Phi(\pi^{t+1}) &\leq \Phi(\pi^t) + \langle \nabla\Phi(\pi^t), \pi^{t+1} - \pi^t\rangle + \frac{\beta}{2}\|\pi^{t+1} - \pi^t\|_2^2\,, \\
&= \Phi(\pi^t) + \langle g^t, \pi^{t+1} - \pi^t\rangle + \langle \nabla\Phi(\pi^t) - g^t, \pi^{t+1} - \pi^t\rangle + \frac{\beta}{2}\|\pi^{t+1} - \pi^t\|_2^2\,, \\
&\leq \Phi(\pi^t) + \langle g^t, \pi^{t+1} - \pi^t\rangle + \frac{1}{2}\|\nabla\Phi(\pi^t) - g^t\|_2^2 + \frac{1+\beta}{2}\|\pi^{t+1} - \pi^t\|_2^2\,.
\end{aligned} \tag{85}$$

By characterization of the projection, the update rule of biased stochastic projected gradient descent yields:

$$\forall \pi \in \Pi, \langle \pi - \pi^{t+1}, (\pi^t - \eta g^t) - \pi^{t+1}\rangle \leq 0\,. \tag{86}$$

Setting $\pi = \pi^t$ in the above inequality and rearranging, we obtain:

$$\langle g^t, \pi^{t+1} - \pi^t\rangle \leq -\frac{1}{\eta}\|\pi^{t+1} - \pi^t\|_2^2 = -\eta\|\hat{G}^{\eta,\alpha}(\pi^t)\|_2^2\,. \tag{87}$$

Moreover, using our inexactness bound assumption, we also have:

$$\frac{1}{2}\|\nabla\Phi(\pi^t) - g^t\|_2^2 \leq \frac{1}{2}\tau^2\,. \tag{88}$$

Using the definition of the (stochastic) gradient mapping together with (88) and (87) in (85), we obtain:

$$\Phi(\pi^{t+1}) \leq \Phi(\pi^t) - \eta\left(1 - \frac{\eta(1+\beta)}{2}\right)\|\hat{G}^{\eta,\alpha}(\pi^t)\|_2^2 + \frac{1}{2}\tau^2\,. \tag{89}$$

Choosing $\eta \leq \frac{1}{1+\beta}$ gives $1 - \frac{\eta(1+\beta)}{2} \geq \frac{1}{2}$ and we obtain by rearranging the above inequality:

$$\|\hat{G}^{\eta,\alpha}(\pi^t)\|_2^2 \leq \frac{2}{\eta}(\Phi(\pi^t) - \Phi(\pi^{t+1})) + \frac{1}{\eta}\tau^2\,. \tag{90}$$

Summing the above inequality for $t \in [T]$ yields:

$$\frac{1}{T}\sum_{t=1}^{T} \|\hat{G}^{\eta,\alpha}(\pi^t)\|_2^2 \leq \frac{2(\Phi_{\max} - \Phi_{\min})}{\eta T} + \frac{1}{\eta}\tau^2\,. \tag{91}$$

□

Proposition 8 (Lemma 11 in Zhang et al. (2024))**.** *Let $\Phi : \Pi \to \mathbb{R}$ be β-smooth and bounded, i.e. there exist $\Phi_{\min}, \Phi_{\max} > 0$ s.t. for all $\pi \in \Pi, \Phi_{\min} \leq \Phi(\pi) \leq \Phi_{\max}$. Let (π^t) be the sequence of policies generated by running stochastic projected gradient descent:*

$$\pi^{t+1} = Proj_{\Pi}(\pi^t - \eta g^t)\,, \quad g_i^t = \hat{\nabla}_{\pi_i} u_{\mu,i}(\pi^t)\,, \forall i \in [N]\,, \tag{92}$$

with stepsize $\eta \leq 1/(1+\beta)$. If there exists $\tau > 0$ s.t. for all $t, \|g^t - \nabla\Phi(\pi^t)\|_2 \leq \tau$, then we have for all $t \geq 1$:

$$\max_{i\in\mathcal{N}} \max_{\pi_i' \in \Pi_i} \langle \pi_i' - \pi_i^{t+1}, \nabla_{\pi_i}\Phi(\pi^{t+1})\rangle \leq 2\sqrt{|\mathcal{S}|}(1+\eta\beta)\|\hat{G}^{\eta,\alpha}(\pi^t)\| + 2\sqrt{|\mathcal{S}|}\tau\,. \tag{93}$$

[3] In this section, we will only use $\alpha = 0$ given a generative model for sampling.

We now quantify the magnitude τ of the gradient errors using concentration results for depending on the minibatch size M used for sampled trajectories and the truncation horizon H.

Proposition 9. *For any $\delta \in (0,1)$, if $M \geq 2\gamma^{-2H}\log(\frac{2|\mathcal{S}|\cdot\max_{k\in[N]}|\mathcal{A}_k|}{\delta})$, then with probability at least $1-3\delta$,*

$$\|\hat{\nabla}_{\pi_i} u_{\mu,i}(\pi^t) - \nabla_{\pi_i} u_{\mu,i}(\pi^t)\|_\infty \leq \frac{2N}{1-\gamma}\left[l_\infty\left(1+\frac{1}{1-\gamma}\right)+NL\right]\gamma^H . \tag{94}$$

Proof. See section 11.5 below. □

End of Proof of Thm. 3. Similarly to the proof of the deterministic setting (Thm. 2, see (78)), we have:

$$\text{NE-Gap}_i(\pi^{t+1}) \leq \frac{C_{\mathcal{G}}}{1-\gamma}\max_{\pi_i'\in\Pi_i}\langle \pi_i' - \pi_i^{t+1}, \nabla_{\pi_i}\phi(\pi^{t+1})\rangle . \tag{95}$$

Using Proposition 11 with $\alpha = 0$ together with the above inequality and taking the maximum over $i \in \mathcal{N}$ yields:

$$\text{NE-Gap}(\pi^{t+1}) \leq \frac{2C_{\mathcal{G}}\sqrt{|\mathcal{S}|}}{1-\gamma}\left((1+\eta\beta)\|\hat{G}^{\eta,\alpha}(\pi^t)\|_2 + \tau\right) . \tag{96}$$

Summing up the above inequalities over $t \in \{1,\dots,T\}$, using Jensen's inequality with the square root function and Proposition 7, we obtain:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\text{NE-Gap}(\pi^{t+1}) &\leq \frac{2C_{\mathcal{G}}\sqrt{|\mathcal{S}|}}{1-\gamma}\frac{1}{T}\sum_{t=1}^{T}\left((1+\eta\beta)\|\hat{G}^{\eta,\alpha}(\pi^t)\|_2+\tau\right)\\
&\leq \frac{2C_{\mathcal{G}}\sqrt{|\mathcal{S}|}}{1-\gamma}\left(\tau + \frac{1+\eta\beta}{T}\sum_{t=1}^{T}\|\hat{G}^{\eta,\alpha}(\pi^t)\|_2\right)\\
&\leq \frac{2C_{\mathcal{G}}\sqrt{|\mathcal{S}|}}{1-\gamma}\left(\tau + (1+\eta\beta)\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\hat{G}^{\eta,\alpha}(\pi^t)\|_2^2}\right)\\
&\leq \frac{2C_{\mathcal{G}}\sqrt{|\mathcal{S}|}}{1-\gamma}\left(\tau + (1+\eta\beta)\sqrt{\frac{2(\Phi_{\max}-\Phi_{\min})}{\eta T}+\frac{1}{\eta}\tau^2}\right)\\
&\leq \frac{2C_{\mathcal{G}}\sqrt{|\mathcal{S}|}}{1-\gamma}\left(\left(1+\frac{2}{\sqrt{\eta}}\right)\tau + \sqrt{\frac{8(\Phi_{\max}-\Phi_{\min})}{\eta T}}\right) ,
\end{aligned} \tag{97}$$

where the last step also uses the condition $\eta \leq 1/(1+\beta)$. It follows that for any $\varepsilon > 0$, if

$$T \geq \frac{128(1+\beta)C_{\mathcal{G}}^2|\mathcal{S}|(\Phi_{\max}-\Phi_{\min})}{(1-\gamma)^2}\cdot\frac{1}{\varepsilon^2}, \quad \tau \leq \frac{1-\gamma}{4(1+2\sqrt{1+\beta})C_{\mathcal{G}}\sqrt{|\mathcal{S}|}}\cdot\varepsilon , \tag{98}$$

then, we have $\frac{1}{T}\sum_{t=1}^{T}\text{NE-Gap}(\pi^{t+1}) \leq \varepsilon$. Hence, there exists $t \in [T]$ s.t. $\text{NE-Gap}(\pi^t) \leq \varepsilon$.

It remains to select the minibatch size M and the truncation horizon H appropriately using Proposition 9 and the condition on τ above which gives:

$$C_1\gamma^H \leq C_2\varepsilon , \tag{99}$$

where the constants C_1, C_2 are independent of H and defined as follows:

$$C_1 := \frac{2N}{1-\gamma}\left[l_\infty\left(1+\frac{1}{1-\gamma}\right)+NL\right], \quad C_2 := \frac{1-\gamma}{4(1+2\sqrt{1+\beta})C_{\mathcal{G}}\sqrt{|\mathcal{S}|}} . \tag{100}$$

It suffices to choose H such that $\exp(-\frac{H}{2}(1-\gamma)) \leq \frac{C_2}{C_1}\varepsilon$, i.e. select H such that:

$$H \geq \frac{2}{1-\gamma}\log\left(\frac{C_2}{C_1}\cdot\frac{1}{\varepsilon}\right) , \tag{101}$$

where C_2, C_1 are defined in (100). Then, using Proposition 9, we obtain the minibatch size condition:

$$M \geq 2\gamma^{-2H} \log\left(\frac{2|\mathcal{S}| \cdot \max_{k\in[N]} |\mathcal{A}_k|}{\delta}\right) \geq 2\left(\frac{C_1}{C_2}\right)^2 \frac{1}{\varepsilon^2} \log\left(\frac{2|\mathcal{S}| \cdot \max_{k\in[N]} |\mathcal{A}_k|}{\delta}\right), \tag{102}$$

since $\gamma^H \leq \frac{C_2}{C_1}\varepsilon$. Overall the total sample complexity is given by:

$$T \times M \times H = \tilde{O}\left(\varepsilon^{-2} \times \varepsilon^{-2} \times \log\left(\frac{1}{\varepsilon}\right)\right) = \tilde{O}(\varepsilon^{-4}), \tag{103}$$

where all the constants are specified above in the choice of T, M, H respectively in (98), (102) and (101). This concludes the proof.

11.5 Proof of Proposition 9

We use the policy gradient theorem (Theorem 3) to write:

$$\begin{aligned}
&\|\hat{\nabla}_{\pi_i} u_{\mu,i}(\pi^t) - \nabla_{\pi_i} u_{\mu,i}(\pi^t)\|_\infty = \max_{s\in\mathcal{S}, a_i\in\mathcal{A}_i} \left|\frac{\hat{\partial} u_{\mu,i}(\pi^t)}{\partial \pi_i(a_i|s)} - \frac{\partial u_{\mu,i}(\pi^t)}{\partial \pi_i(a_i|s)}\right| \\
&= \frac{1}{1-\gamma} \max_{s\in\mathcal{S}, a_i\in\mathcal{A}_i} \left|\hat{d}_\mu^{\pi^t}(s) \sum_{j=1}^N \hat{q}_{i,j,(s,a_i)}^{\pi^t} - d_\mu^{\pi^t}(s) \sum_{j=1}^N \bar{Q}_{s,a_i}^{\pi^t}(\bar{r}_{i,j}^{\pi^t})\right| \\
&\leq \frac{1}{1-\gamma} \max_{s\in\mathcal{S}, a_i\in\mathcal{A}_i} \left(\hat{d}_\mu^{\pi^t}(s) \cdot \sum_{j=1}^N \|\hat{q}_j^{\pi^t} - \bar{Q}^{\pi^t}(\bar{r}_{i,j}^{\pi^t})\|_\infty + |\sum_{j=1}^N \bar{Q}_{s,a_i}^{\pi^t}(\bar{r}_{i,j}^{\pi^t})| \cdot |\hat{d}_\mu^{\pi^t}(s) - d_\mu^{\pi^t}(s)|\right) \\
&\leq \frac{1}{1-\gamma}\left(\sum_{j=1}^N \|\hat{q}_j^{\pi^t} - \bar{Q}^{\pi^t}(\bar{r}_{i,j}^{\pi^t})\|_\infty + \frac{N l_\infty}{1-\gamma}\|\hat{d}_\mu^{\pi^t} - d_\mu^{\pi^t}\|_\infty\right).
\end{aligned} \tag{104}$$

In what follows, we establish high probability error bounds for each one of the error terms in the above inequality.

Lemma 3. *For any $\delta \in (0,1)$, if $M \geq \frac{\gamma^{-2H}}{2}\log(\frac{2|\mathcal{S}|}{\delta})$, then with probability at least $1-\delta$,*

$$\|\hat{d}_\mu^{\pi^t} - d_\mu^{\pi^t}\|_\infty \leq 2\gamma^H, \tag{105}$$

for all iterations $t \geq 1$.

Proof. First, we upper bound the error into two terms: a bias error due to horizon truncation and a concentration error which vanishes when increasing the number of sampled trajectories:

$$\|\hat{d}_\rho^{\pi^t} - d_\rho^{\pi^t}\|_\infty \leq \|\hat{d}_\rho^{\pi^t} - \mathbb{E}[\hat{d}_\rho^{\pi^t}]\| + \|\mathbb{E}[\hat{d}_\rho^{\pi^t}] - d_\rho^{\pi^t}\|_\infty. \tag{106}$$

For the second error, we observe that:

$$\|\mathbb{E}[\hat{d}_\rho^{\pi^t}] - d_\rho^{\pi^t}\|_\infty = \max_{s\in\mathcal{S}} \left|\sum_{t=H}^{+\infty}(1-\gamma)\gamma^t \mathbb{P}(s_t^{(j)} = s)\right| \leq (1-\gamma)\sum_{t=H}^{+\infty}\gamma^t = \gamma^H. \tag{107}$$

As for the first error term, we use Hoeffding's inequality. For any state $s \in \mathcal{S}$ and any $\sigma > 0$, since $\hat{d}_\rho^{\pi^t}(s)$ is bounded (in $[0,1]$), we have:

$$\mathbb{P}(|\hat{d}_\rho^{\pi^t}(s) - \mathbb{E}[\hat{d}_\rho^{\pi^t}(s)]| \geq \sigma) \leq 2\exp(-2M\sigma^2), \tag{108}$$

where M is the minibatch size used in the Monte Carlo estimate of $\hat{d}_\rho^{\pi^t}(s)$. Using a union bound, we obtain:

$$\mathbb{P}(\|\hat{d}_\rho^{\pi^t}(s) - \mathbb{E}[\hat{d}_\rho^{\pi^t}(s)]\|_\infty \geq \sigma) \leq \mathbb{P}\left(\cup_{s\in\mathcal{S}}\{|\hat{d}_\rho^{\pi^t}(s) - \mathbb{E}[\hat{d}_\rho^{\pi^t}(s)]| \geq \sigma\}\right) \leq 2|\mathcal{S}|\exp(-2M\sigma^2). \tag{109}$$

Hence, for any $\delta \in (0,1)$, if $M \geq \frac{1}{2\sigma^2}\log(\frac{2|\mathcal{S}|}{\delta})$, then with probability at least $1-\delta$, we have with probability at least $1-\delta$,

$$\|\hat{d}_\rho^{\pi^t}(s) - \mathbb{E}[\hat{d}_\rho^{\pi^t}(s)]\|_\infty \leq \sigma\,. \tag{110}$$

Combining (107) and (110) yields:

$$\|\hat{d}_\rho^{\pi^t} - d_\rho^{\pi^t}\|_\infty \leq \gamma^H + \sigma\,, \tag{111}$$

with probability at least $1-\delta$. It remainss to set $\sigma = \gamma^H$ to obtain the desired result. □

Lemma 4. *For any $\delta \in (0,1)$, if $M \geq 2\gamma^{-2H}\log(\frac{2|\mathcal{S}|\cdot\max_{k\in[N]}|\mathcal{A}_k|}{\delta})$, then with probability at least $1-2\delta$,*

$$\|\hat{q}_j^{\pi^t} - \bar{Q}^{\pi^t}(\bar{r}_{i,j}^{\pi^t})\|_\infty \leq \left(\frac{2l_\infty}{1-\gamma} + 2NL\right)\gamma^H\,, \tag{112}$$

for all iterations $t \geq 1$ and for all $i,j \in [N]$.

Proof. We decompose the Q-value estimation errors as follows:

$$\begin{aligned}\|\hat{q}_j^{\pi^t} - \bar{Q}^{\pi^t}(\bar{r}_{i,j}^{\pi^t})\|_\infty &= \|\hat{q}_j^{\pi^t}(\hat{r}_{i,j}^{\pi^t}) - \bar{Q}^{\pi^t}(\bar{r}_{i,j}^{\pi^t})\|_\infty \\ &= \|\hat{q}_j^{\pi^t}(\hat{r}_{i,j}^{\pi^t}) - \hat{q}_j^{\pi^t}(\bar{r}_{i,j}^{\pi^t}) + \hat{q}_j^{\pi^t}(\bar{r}_{i,j}^{\pi^t}) - \bar{Q}^{\pi^t}(\bar{r}_{i,j}^{\pi^t})\|_\infty \\ &\leq \|\hat{q}_j^{\pi^t}(\hat{r}_{i,j}^{\pi^t}) - \hat{q}_j^{\pi^t}(\bar{r}_{i,j}^{\pi^t})\|_\infty + \|\hat{q}_j^{\pi^t}(\bar{r}_{i,j}^{\pi^t}) - \bar{Q}^{\pi^t}(\bar{r}_{i,j}^{\pi^t})\|_\infty\,.\end{aligned} \tag{113}$$

For the first term, we observe that:

$$\|\hat{q}_j^{\pi^t}(\hat{r}_{i,j}^{\pi^t}) - \hat{q}_j^{\pi^t}(\bar{r}_{i,j}^{\pi^t})\|_\infty \leq \frac{1}{1-\gamma}\|\hat{r}_{i,j}^{\pi^t} - \bar{r}_{i,j}^{\pi^t}\|_{\infty,\tau_{1:M}} \leq L\|\hat{\lambda}_{\mu,1:N}^{\pi^t} - \lambda_{\mu,1:N}^{\pi^t}\|_1 \leq L\sum_{k=1}^{N}\|\hat{\lambda}_{\mu,k}^{\pi^t} - \lambda_{\mu,k}^{\pi^t}\|_1 = NL\|\hat{d}_\mu^{\pi^t} - d_\mu^{\pi^t}\|_1\,. \tag{114}$$

where the second inequality follows from using Assumption 3 together with the fact that there is no dependence on π_{-i}^t in the pseudo-rewards in the potential GUMG setting.

We can now apply Lemma 3 to obtain that, if $M \geq \frac{\gamma^{-2H}}{2}\log(\frac{2|\mathcal{S}|}{\delta})$, then with probability at least $1-\delta$,

$$\|\hat{q}_j^{\pi^t}(\hat{r}_{i,j}^{\pi^t}) - \hat{q}_j^{\pi^t}(\bar{r}_{i,j}^{\pi^t})\|_\infty \leq 2NL\gamma^H\,. \tag{115}$$

As for the second term in (113), we use the same proof technique as the proof of Lemma 3, combining a bias term bound and an application of Hoeffding's concentration inequality. Observe for this that the random variables $\hat{q}_j^{\pi^t}(\hat{r}_{i,j}^{\pi^t})$ are bounded (in $[\frac{-l_\infty}{1-\gamma}, \frac{l_\infty}{1-\gamma}]$) by boundedness of the pseudo-rewards and use $\sigma = \frac{l_\infty}{1-\gamma}\gamma^H$. Therefore, following the same lines we obtain that for any $\delta \in (0,1)$, if $M \geq 2\gamma^{-2H}\log(\frac{2|\mathcal{S}|\cdot\max_{k\in[N]}|\mathcal{A}_k|}{\delta})$, then with probability at least $1-\delta$,

$$\|\hat{q}_j^{\pi^t}(\bar{r}_{i,j}^{\pi^t}) - \bar{Q}^{\pi^t}(\bar{r}_{i,j}^{\pi^t})\|_\infty \leq \frac{2l_\infty}{1-\gamma}\gamma^H\,. \tag{116}$$

□

Combining the results of Lemma 3 and Lemma 4 in (104) gives the desired inequality and concludes the proof of Proposition 9.

11.6 Proof of Thm. 4

Proof sketch and comparison to prior work. Our proof technique here differs from the proof of a similar result of Leonardos et al. (2022) (Theorem 4.4) for the particular case of Markov potential games. Note that we do not use the Moreau-envelope in our analysis for this stochastic setting. We rather provide an analysis of projected stochastic gradient descent with minibatch, controlling the bias and variance of the policy gradient estimates. The following decomposition gives more intuition on the different errors involved in the analysis:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\text{NE-Gap}(\pi^{t+1})] = \underbrace{\tilde{O}\left(\frac{1}{\sqrt{T}}\right)}_{\text{Optimization error}} + \underbrace{\tilde{O}\left(\gamma^H\right)}_{\text{bias due to truncation}} + \underbrace{\tilde{O}\left(\frac{1}{\sqrt{M\alpha}}\right)}_{\text{variance term}} + \underbrace{\tilde{O}\left(\alpha\right)}_{\text{bias due to }\alpha\text{-greediness}}\,. \tag{117}$$

This suggests to choose $T = \tilde{O}(\varepsilon^{-2})$ (like in the deterministic setting), a greediness parameter $\alpha = \tilde{O}(\varepsilon)$, a minibatch size $M = \tilde{O}(\varepsilon^{-3})$ and a horizon length $H = O\left(\frac{1}{1-\gamma}\log\left(\frac{1}{\varepsilon}\right)\right)$. The resulting total sample complexity is: $TMH = \tilde{O}(\varepsilon^{-5})$.

Proof. We provide below a complete proof. We start with a first-order stationarity guarantee for smooth non-convex optimization with biased stochastic gradients.

Proposition 10. *Let $\Phi : \Pi \to \mathbb{R}$ be β-smooth and bounded, i.e. there exist $\Phi_{\min}, \Phi_{\max} > 0$ s.t. for all $\pi \in \Pi, \Phi_{\min} \leq \Phi(\pi) \leq \Phi_{\max}$. Let (π^t) be the sequence of policies generated by running stochastic projected gradient descent:*

$$\pi^{t+1} = Proj_{\Pi^\alpha}(\pi^t - \eta g^t)\,, \quad g_i^t = \hat{\nabla}_{\pi_i} u_{\mu,i}(\pi^t)\,, \forall i \in [N]\,, \tag{118}$$

with stepsize $\eta \leq 1/(1+\beta)$. Suppose that the bias and variance of the policy gradient estimator g^t are bounded by positive constants δ, σ^2 as follows:

$$\|\mathbb{E}_t[g^t] - \nabla\Phi(\pi^t)\|_2 \leq \delta\,, \quad \mathbb{E}\left[\|g^t - \mathbb{E}_t[g^t]\|_2^2\right] \leq \sigma^2\,, \tag{119}$$

where $\mathbb{E}_t$ is the conditional expectation w.r.t. randomness up to time t. Then we have for any $\alpha \geq 0, t \geq 1$:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\hat{G}^{\eta,\alpha}(\pi^t)\|_2^2] \leq \frac{2(\Phi_{\max} - \Phi_{\min})}{\eta T} + \delta^2 + \sigma^2\,, \tag{120}$$

where we recall that $\hat{G}^{\eta,\alpha}(\pi^t) = \frac{1}{\eta}\left(Proj_{\Pi^\alpha}(\pi^t + \eta g^t) - \pi^t\right)$ for any $\alpha \geq 0, t \geq 1$.

Proof. See section 11.6.1 □

Lemma 5 (Bias bound)**.** *Under Assumption 3, we have for any $i \in \mathcal{N}$:*

$$\mathbb{E}\left[\|\nabla_\pi u_{\mu,i,H}(\pi) - \nabla_\pi u_{\mu,i}(\pi)\|_2\right] \leq \left(\frac{1+l_\infty}{1-\gamma} + NL\right)|\mathcal{S}|\left(\sum_{k=1}^{N}|\mathcal{A}_k|\right)\gamma^H\,, \tag{121}$$

where $\nabla_\pi u_{\mu,i,H}(\pi)$ is a notation for the (conditional) expectation of the truncated policy gradient estimator (as in Proposition 10).

Proof. The proof follows similar lines to the proof of Proposition 9 where the bias is controlled in the proof of Lemma 3 (see eq. (107)) and Lemma 4. □

Lemma 6 (Variance bound)**.** *For any $i \in \mathcal{N}$ and any policy $\pi \in \Pi$,*

$$\mathbb{E}\left[\|\hat{\nabla}_{\pi_i} u_{\mu,i}(\pi) - \nabla_{\pi_i} u_{\mu,i,H}(\pi)\|_2^2\right] \leq \frac{N^2 l_\infty^2}{(1-\gamma)^4 M\alpha}\,. \tag{122}$$

Proof. The proof follows similar lines to the proof of (Daskalakis et al., 2020, Lemma 2). It is enough to observe that the sum over j of pseudo-rewards are bounded by Nl_∞ and our minibatch size M scales the variance. Then the variance for each trajectory estimator term is bounded using (Daskalakis et al., 2020, Lemma 2). □

Proposition 11. *Let $\Phi : \Pi \to \mathbb{R}$ be β-smooth and bounded, i.e. there exist $\Phi_{\min}, \Phi_{\max} > 0$ s.t. for all $\pi \in \Pi, \Phi_{\min} \leq \Phi(\pi) \leq \Phi_{\max}$. Let (π^t) be the sequence of policies generated by running stochastic projected gradient descent:*

$$\pi^{t+1} = Proj_{\Pi^\alpha}(\pi^t - \eta g^t)\,, \quad g_i^t = \hat{\nabla}_{\pi_i} u_{\mu,i}(\pi^t)\,, \forall i \in [N]\,, \tag{123}$$

with stepsize $\eta \leq 1/(1+\beta)$. Suppose that the bias and variance of the policy gradient estimator g^t are bounded by positive constants δ, σ^2 as follows:

$$\|\mathbb{E}_t[g^t] - \nabla\Phi(\pi^t)\|_2 \leq \delta\,, \quad \mathbb{E}\left[\|g^t - \mathbb{E}_t[g^t]\|_2^2\right] \leq \sigma^2\,, \tag{124}$$

where $\mathbb{E}_t$ is the conditional expectation w.r.t. randomness up to time t. Then we have for any $\alpha \geq 0, t \geq 1$:

$$\mathbb{E}\left[\max_{i\in\mathcal{N}}\max_{\pi_i'\in\Pi_i}\langle \pi_i' - \pi_i^{t+1}, \nabla_{\pi_i}\Phi(\pi^{t+1})\rangle\right] \leq 2\sqrt{|\mathcal{S}|}(1+\eta\beta)\mathbb{E}[\|\hat{G}^{\eta,\alpha}(\pi^t)\|_2] + 2\sqrt{|\mathcal{S}|}(\delta+\sigma) + \frac{2Nl_\infty}{(1-\gamma)^2}\alpha\,. \tag{125}$$

Proof. See section 11.6.2. □

The proof follows similar lines as the proof of (Zhang et al., 2024, Lemma 11) with the following differences:

- Our setting involves general utilities beyond standard RL. This implies a few modifications needed including bounding the pseudo-rewards and using a suitable policy gradient theorem for our setting.
- We control the bias and variance of our estimator differently. Indeed, in this section, we are interested in deriving (on-policy) guarantees in expectation rather than with high probability.

End of Proof of Thm. 4. Similarly to the proof of Thm. 2 with exact gradients, we start by bounding the Nash equilibrium gap using agent-wise gradient domination. We have for every $i \in \mathcal{N}$ and every time step $t \geq 0$:

$$\text{NE-Gap}_i(\pi^{t+1}) \leq \frac{C_{\mathcal{G}}}{1-\gamma} \max_{\pi_i' \in \Pi_i} \langle \pi_i' - \pi_i^{t+1}, \nabla_{\pi_i} \Phi(\pi^{t+1}) \rangle \,. \tag{126}$$

Taking the maximum over $i \in \mathcal{N}$ on both sides of the inequality, taking expectation and using Proposition 11, we obtain

$$\mathbb{E}[\text{NE-Gap}(\pi^{t+1})] \leq \frac{C_{\mathcal{G}}}{1-\gamma} \left[2\sqrt{|\mathcal{S}|}(1+\eta\beta)\mathbb{E}[\|\hat{G}^{\eta,\alpha}(\pi^{t+1})\|_2] + 2\sqrt{|\mathcal{S}|}(\delta+\sigma) + \frac{2Nl_\infty}{(1-\gamma)^2}\alpha \right] . \tag{127}$$

Taking the average of the above inequality over $t \in [T-1]$ and using a stepsize $\eta \leq 1/(1+\beta)$, it follows that:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\text{NE-Gap}(\pi^{t+1})] &\leq \frac{C_{\mathcal{G}}}{1-\gamma} \left[4\sqrt{|\mathcal{S}|}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\hat{G}^{\eta,\alpha}(\pi^{t+1})\|_2] + 2\sqrt{|\mathcal{S}|}(\delta+\sigma) + \frac{2Nl_\infty}{(1-\gamma)^2}\alpha \right] \\
&\leq \frac{C_{\mathcal{G}}}{1-\gamma} \left[4\sqrt{|\mathcal{S}|}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\hat{G}^{\eta,\alpha}(\pi^{t+1})\|_2^2]} + 2\sqrt{|\mathcal{S}|}(\delta+\sigma) + \frac{2Nl_\infty}{(1-\gamma)^2}\alpha \right] \\
&\leq \frac{C_{\mathcal{G}}}{1-\gamma} \left[4\sqrt{|\mathcal{S}|}\sqrt{\frac{2(\Phi_{\max}-\Phi_{\min})}{\eta T} + \delta^2 + \sigma^2} + 2\sqrt{|\mathcal{S}|}(\delta+\sigma) + \frac{2Nl_\infty}{(1-\gamma)^2}\alpha \right] \\
&\leq \frac{C_{\mathcal{G}}}{1-\gamma} \left[4\sqrt{|\mathcal{S}|}\sqrt{\frac{2(\Phi_{\max}-\Phi_{\min})}{\eta T}} + 6\sqrt{|\mathcal{S}|}(\delta+\sigma) + \frac{2Nl_\infty}{(1-\gamma)^2}\alpha \right]
\end{aligned} \tag{128}$$

where the second inequality follows from Jensen's inequality and the third one from using Proposition 10.

It remains to plug in the bias δ and variance σ^2 obtained in Lemma 5 and Lemma 6 respectively, and select the minibatch size M, the horizon length H and the greediness parameter α appropriately to obtain the desired sample complexity.

Using Lemma 5 and Lemma 6 in (128), we obtain:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\text{NE-Gap}(\pi^{t+1})] \leq \underbrace{\frac{\tilde{C}_1}{\sqrt{T}}}_{\text{Optimization error}} + \underbrace{\tilde{C}_2\gamma^H}_{\text{bias due to truncation}} + \underbrace{\frac{\tilde{C}_3}{\sqrt{M\alpha}}}_{\text{variance term}} + \underbrace{\tilde{C}_4\alpha}_{\text{bias due to } \alpha\text{-greediness}} , \tag{129}$$

where the constants $\tilde{C}_1, \tilde{C}_2, \tilde{C}_3, \tilde{C}_4$ are independent of T, M, H, α and defined as follows:

$$\begin{aligned}
&\tilde{C}_1 := \frac{4\sqrt{|\mathcal{S}|}C_{\mathcal{G}}}{1-\gamma}\sqrt{\frac{2(\Phi_{\max}-\Phi_{\min})}{\eta}} , \quad \tilde{C}_2 := \frac{6C_{\mathcal{G}}\sqrt{|\mathcal{S}|}}{1-\gamma}\left(\frac{1+l_\infty}{1-\gamma} + NL\right)|\mathcal{S}|\left(\sum_{k=1}^{N}|\mathcal{A}_k|\right) , \\
&\tilde{C}_3 := \frac{6C_{\mathcal{G}}\sqrt{|\mathcal{S}|}N^{\frac{3}{2}}l_\infty}{(1-\gamma)^3} , \quad \tilde{C}_4 := \frac{2C_{\mathcal{G}}Nl_\infty}{(1-\gamma)^3} .
\end{aligned} \tag{130}$$

Selecting then T, H, M, α as follows:

$$T \geq \frac{\tilde{C}_1^2}{\varepsilon^2}\,, \quad \alpha \leq \frac{\varepsilon}{\tilde{C}_4}\,, \quad M \geq \left(\frac{\tilde{C}_3}{\varepsilon\sqrt{\alpha}}\right)^2 = \frac{\tilde{C}_3^2\tilde{C}_4}{\varepsilon^3}\,, \quad H \geq \frac{1}{1-\gamma}\log\left(\frac{\tilde{C}_2}{\varepsilon}\right)\,, \tag{131}$$

guarantees that $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\text{NE-Gap}(\pi^{t+1})] \leq \varepsilon$. In particular the total sample complexity is given by:

$$T \times M \times H \leq \frac{\tilde{C}_1^2}{\varepsilon^2} \times \frac{\tilde{C}_3^2\tilde{C}_4}{\varepsilon^3} \times \frac{1}{1-\gamma}\log\left(\frac{\tilde{C}_2}{\varepsilon}\right) = \frac{\tilde{C}_1^2\tilde{C}_3^2\tilde{C}_4}{1-\gamma}\frac{1}{\varepsilon^5}\log\left(\frac{\tilde{C}_2}{\varepsilon}\right)\,, \tag{132}$$

where $\tilde{C}_1, \tilde{C}_2, \tilde{C}_3, \tilde{C}_4$ are defined in (130). This concludes the proof.

11.6.1 Proof of Proposition 10

The proof follows standard optimization arguments, we provide a full proof for completeness. Throughout this proof, we use the shorthand notation $\bar{g}^t := \mathbb{E}_t[g^t]$.

By β-smoothness of Φ, we have for every time step $t \geq 0$:

$$\begin{aligned}
\Phi(\pi^{t+1}) &\leq \Phi(\pi^t) + \langle \nabla\Phi(\pi^t), \pi^{t+1} - \pi^t\rangle + \frac{\beta}{2}\|\pi^{t+1} - \pi^t\|_2^2\,,\\
&= \Phi(\pi^t) + \langle g^t, \pi^{t+1} - \pi^t\rangle + \langle \nabla\Phi(\pi^t) - g^t, \pi^{t+1} - \pi^t\rangle + \frac{\beta}{2}\|\pi^{t+1} - \pi^t\|_2^2\,,\\
&\leq \Phi(\pi^t) + \langle g^t, \pi^{t+1} - \pi^t\rangle + \frac{1}{2}\|\nabla\Phi(\pi^t) - g^t\|_2^2 + \frac{1+\beta}{2}\|\pi^{t+1} - \pi^t\|_2^2\,.
\end{aligned} \tag{133}$$

By characterization of the projection, the update rule of biased stochastic projected gradient descent yields:

$$\forall \pi \in \Pi, \langle \pi - \pi^{t+1}, (\pi^t - \eta g^t) - \pi^{t+1}\rangle \leq 0\,. \tag{134}$$

Setting $\pi = \pi^t$ in the above inequality and rearranging, we obtain:

$$\langle g^t, \pi^{t+1} - \pi^t\rangle \leq -\frac{1}{\eta}\|\pi^{t+1} - \pi^t\|_2^2 = -\eta\|\hat{G}^{\eta,\alpha}(\pi^t)\|_2^2\,. \tag{135}$$

Moreover, using our bias bound assumption, we also have:

$$\frac{1}{2}\|\nabla\Phi(\pi^t) - g^t\|_2^2 \leq \|\nabla\Phi(\pi^t) - \bar{g}^t\|_2^2 + \|g^t - \bar{g}^t\|_2^2 \leq \delta^2 + \|g^t - \bar{g}^t\|_2^2\,. \tag{136}$$

Using the definition of the (stochastic) gradient mapping together with (136) and (135) in (133), we obtain:

$$\Phi(\pi^{t+1}) \leq \Phi(\pi^t) - \eta\left(1 - \frac{\eta(1+\beta)}{2}\right)\|\hat{G}^{\eta,\alpha}(\pi^t)\|_2^2 + \delta^2 + \|g^t - \bar{g}^t\|_2^2\,. \tag{137}$$

Choosing $\eta \leq \frac{1}{1+\beta}$ gives $1 - \frac{\eta(1+\beta)}{2} \geq \frac{1}{2}$ and we obtain by rearranging the above inequality:

$$\|\hat{G}^{\eta,\alpha}(\pi^t)\|_2^2 \leq \frac{2}{\eta}(\Phi(\pi^t) - \Phi(\pi^{t+1})) + \delta^2 + \|g^t - \bar{g}^t\|_2^2\,. \tag{138}$$

Taking total expectation and using the variance bound assumption gives:

$$\mathbb{E}[\|\hat{G}^{\eta,\alpha}(\pi^t)\|_2^2] \leq \frac{2}{\eta}\mathbb{E}[\Phi(\pi^t) - \Phi(\pi^{t+1})] + \delta^2 + \sigma^2\,. \tag{139}$$

It remains to sum up the above inequality for $t \in [T-1]$ to obtain:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\hat{G}^{\eta,\alpha}(\pi^t)\|_2^2] \leq \frac{2}{\eta T}\mathbb{E}[\Phi(\pi^0) - \Phi(\pi^T)] + \delta^2 + \sigma^2 \leq \frac{2(\Phi_{\max} - \Phi_{\min})}{\eta T} + \delta^2 + \sigma^2\,, \tag{140}$$

and this concludes the proof.

11.6.2 Proof of Proposition 11

In this proof, we use the shorthand notation $g := g^t, \pi = \pi^t, \pi^+ = \pi^{t+1}$.

We use the following decomposition of our term of interest for any policy $\pi' \in \Pi_\alpha$,

$$\langle \pi' - \pi^+, \nabla\Phi(\pi^+)\rangle = \langle \pi' - \pi^+, g\rangle + \langle \pi' - \pi^+, g - \nabla\Phi(\pi)\rangle + \langle \pi' - \pi^+, \nabla\Phi(\pi) - \nabla\Phi(\pi^+)\rangle\,. \tag{141}$$

We now control each one of the terms in the above decomposition.

For the first term we use the characterization of the projection that for any policy $\pi' \in \Pi^\alpha$,

$$\langle \pi' - \pi^+, (\pi + \eta g) - \pi^+\rangle \leq 0\,, \tag{142}$$

which yields the following after rearranging the inequality and using the Cauchy-Schwarz inequality:

$$\langle \pi' - \pi^+, g\rangle \leq -\frac{1}{\eta}\langle \pi' - \pi^+, \pi - \pi^+\rangle = \langle \pi' - \pi^+, \hat{G}^{\eta,\alpha}(\pi)\rangle \leq \|\pi - \pi'\|_2 \cdot \|\hat{G}^{\eta,\alpha}(\pi)\|_2 \leq 2\sqrt{|\mathcal{S}|}\cdot\|\hat{G}^{\eta,\alpha}(\pi)\|_2\,. \tag{143}$$

For the second term involving the bias of the gradient estimator, we immediately have

$$\langle \pi' - \pi^+, g - \nabla\Phi(\pi)\rangle \leq \|\pi' - \pi^+\|_2 \cdot \|g - \nabla\Phi(\pi)\|_2 \leq \|\pi' - \pi^+\|_1 \cdot \|g - \nabla\Phi(\pi)\|_2 \leq 2\sqrt{|\mathcal{S}|}\cdot\|g - \nabla\Phi(\pi)\|_2\,. \tag{144}$$

For the last term, we use smoothness of the function Φ as follows:

$$\langle \pi' - \pi^+, \nabla\Phi(\pi) - \nabla\Phi(\pi^+)\rangle \leq \|\pi' - \pi^+\|_2 \cdot \|\nabla\Phi(\pi) - \nabla\Phi(\pi^+)\|_2 \leq 2\sqrt{|\mathcal{S}|}\beta\|\pi - \pi^+\|_2 = 2\eta\beta\sqrt{|\mathcal{S}|}\|\hat{G}^{\eta,\alpha}(\pi)\|_2\,. \tag{145}$$

Combining (143), (144) and (145) in (141), we obtain for any $\pi' \in \Pi^\alpha$,

$$\langle \pi' - \pi^+, \nabla\Phi(\pi^+)\rangle \leq 2(1+\eta\beta)\sqrt{|\mathcal{S}|}\cdot\|\hat{G}^{\eta,\alpha}(\pi)\|_2 + 2\sqrt{|\mathcal{S}|}\cdot\|g - \nabla\Phi(\pi)\|_2\,. \tag{146}$$

Now for any $\pi'_i \in \Pi_i$, set $\pi' = (\pi'_i, \pi^+_{-i})$ in the previous inequality. Define $\pi^\alpha_i := (1-\alpha)\pi'_i + \frac{\alpha}{|\mathcal{A}_i|} \in \Pi^\alpha_i$. Hence, we get for any $\pi'_i \in \Pi_i$,

$$\langle \pi'_i - \pi^+_i, \nabla_{\pi_i}\Phi(\pi^+)\rangle = \langle \pi^\alpha_i - \pi^+_i, \nabla_{\pi_i}\Phi(\pi^+)\rangle + \langle \pi'_i - \pi^\alpha_i, \nabla_{\pi_i}\Phi(\pi^+)\rangle \tag{147}$$

$$\leq 2(1+\eta\beta)\sqrt{|\mathcal{S}|}\cdot\|\hat{G}^{\eta,\alpha}(\pi)\|_2 + 2\sqrt{|\mathcal{S}|}\cdot\|g - \nabla\Phi(\pi)\|_2 + \alpha\left\langle \pi'_i - \frac{1}{|\mathcal{A}_i|}, \nabla_{\pi_i}\Phi(\pi^+)\right\rangle\,. \tag{148}$$

We bound the last term in the above inequality as follows using our policy gradient theorem (see Proposition 3):

$$\begin{aligned}
\left\langle \pi'_i - \frac{1}{|\mathcal{A}_i|}, \nabla_{\pi_i}\Phi(\pi^+)\right\rangle &= \sum_{s\in\mathcal{S}, a_i\in\mathcal{A}_i}\left(\pi'_i(a_i|s) - \frac{1}{|\mathcal{A}_i|}\right)\cdot[\nabla_{\pi_i}\Phi(\pi^+)]_{s,a_i} \\
&= \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}, a_i\in\mathcal{A}_i}\left(\pi'_i(a_i|s) - \frac{1}{|\mathcal{A}_i|}\right) d^{\pi^+}_\rho(s)\sum_{j=1}^N \bar{Q}^{\pi^+}_{s,a_i}(\bar{r}^{\pi^+}_{i,j}) \\
&\leq \frac{\sum_{j=1}^N \|\bar{r}^{\pi^+}_{i,j}\|_\infty}{(1-\gamma)^2}\sum_{s\in\mathcal{S}, a_i\in\mathcal{A}_i}\left|\pi'_i(a_i|s) - \frac{1}{|\mathcal{A}_i|}\right|\cdot d^{\pi^+}_\rho(s) \\
&\leq \frac{2Nl_\infty}{(1-\gamma)^2}\,.
\end{aligned} \tag{149}$$

Using (149) in (147) yields:

$$\max_{i\in\mathcal{N}}\max_{\pi'_i\in\Pi_i}\langle \pi'_i - \pi^+_i, \nabla_{\pi_i}\Phi(\pi^+)\rangle \leq 2(1+\eta\beta)\sqrt{|\mathcal{S}|}\cdot\|\hat{G}^{\eta,\alpha}(\pi)\|_2 + 2\sqrt{|\mathcal{S}|}\cdot\|g - \nabla\Phi(\pi)\|_2 + \frac{2Nl_\infty}{(1-\gamma)^2}\alpha\,. \tag{150}$$

We conclude the proof by taking total expectation in the above inequality and using our assumptions in (124) (and Jensen's inequality) to bound the term $\|g - \nabla\Phi(\pi)\|_2$ in expectation and obtain the desired inequality:

$$\mathbb{E}\left[\max_{i\in\mathcal{N}}\max_{\pi'_i\in\Pi_i}\langle \pi'_i - \pi^+_i, \nabla_{\pi_i}\Phi(\pi^+)\rangle\right] \leq 2\sqrt{|\mathcal{S}|}(1+\eta\beta)\mathbb{E}[\|\hat{G}^{\eta,\alpha}(\pi)\|_2] + 2\sqrt{|\mathcal{S}|}(\delta+\sigma) + \frac{2Nl_\infty}{(1-\gamma)^2}\alpha\,. \tag{151}$$

12 AUXILIARY RESULTS: SINGLE-AGENT RL

Let $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, r, \rho, \gamma)$ where $r \in \mathcal{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ be a Markov Decision Process (MDP). Define the value and action-value functions as follows for every state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$V_\rho^\pi(r) := \mathbb{E}_{\rho,\pi}\left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, a_t)\right], \quad Q_{s,a}^\pi(r) := \mathbb{E}_{\rho,\pi}\left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a\right] = \langle \lambda_{s,a}^\pi, r\rangle, \tag{152}$$

where $\lambda_{s,a}^\pi$ is the occupancy measure corresponding to the Dirac initial distribution putting all the weight on the state action pair (s, a).

Proposition 12 (Policy Gradient Theorem)**.** *Let $\mathcal{M}(r) := (\mathcal{S}, \mathcal{A}, P, r, \rho, \gamma)$ where $r \in \mathcal{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ be an MDP parameterized by a reward function r. Then, we have for all $\pi \in \Pi$ and all $r \in \mathcal{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$,*

$$\nabla_\pi V_\rho^\pi(r) = \frac{1}{1-\gamma} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \lambda_\rho^\pi(s, a)\, Q_{s,a}^\pi(r)\, \nabla \log \pi(a|s). \tag{153}$$

Moreover, for the direct policy parameterization, we have for any $s \in \mathcal{S}, a \in \mathcal{A}$,

$$\frac{\partial V_\rho^\pi(r)}{\partial \pi(a|s)} = \frac{1}{1-\gamma} d_\rho^\pi(s) Q_{s,a}^\pi(r). \tag{154}$$

Proposition 13. *For any policy $\pi \in \Pi$, for any reward functions $r, r' \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, we have*

$$\|\nabla_\pi V_\rho^\pi(r) - \nabla_\pi V_\rho^\pi(r')\|_2 \le \frac{\sqrt{|\mathcal{A}|}}{(1-\gamma)^2} \|r - r'\|_\infty. \tag{155}$$

Proof. Using (154), we have

$$\begin{aligned}
\left|\frac{\partial V_\rho^\pi(r)}{\partial \pi(a|s)} - \frac{\partial V_\rho^\pi(r')}{\partial \pi(a|s)}\right| &= \frac{1}{1-\gamma} d_\rho^\pi(s) |Q_{s,a}^\pi(r) - Q_{s,a}^\pi(r')| \\
&= \frac{1}{(1-\gamma)^2} d_\rho^\pi(s) |\langle \lambda_{s,a}^\pi, r - r'\rangle| \\
&\le \frac{1}{(1-\gamma)^2} d_\rho^\pi(s) \|\lambda_{s,a}^\pi\|_1 \cdot \|r - r'\|_\infty \\
&= \frac{1}{(1-\gamma)^2} d_\rho^\pi(s) \|r - r'\|_\infty,
\end{aligned} \tag{156}$$

where for all $s' \in \mathcal{S}, a' \in \mathcal{A}$, $\lambda_{s,a}^\pi(s', a') = (1-\gamma)\sum_{t=0}^{+\infty} \gamma^t \mathbb{P}_{\delta_{s,a},\pi}(s_t = s', a_t = a')$ and $\delta_{s,a}$ is the Dirac measure at the state-action pair (s, a). It follows that

$$\|\nabla_\pi V_\rho^\pi(r) - \nabla_\pi V_\rho^\pi(r')\|_2^2 = \sum_{s,a} \left|\frac{\partial V_\rho^\pi(r)}{\partial \pi(a|s)} - \frac{\partial V_\rho^\pi(r')}{\partial \pi(a|s)}\right|^2 \tag{157}$$

$$\le \frac{1}{(1-\gamma)^4} \sum_{s,a} d_\rho^\pi(s)^2 \cdot \|r - r'\|_\infty^2 \tag{158}$$

$$\le \frac{|\mathcal{A}|}{(1-\gamma)^4} \|r - r'\|_\infty^2, \tag{159}$$

where the first inequality uses (156) and the last inequality stems from recalling that $0 \le d_\rho^\pi(s) \le 1$ which leads to $\sum_{s,a} d_\rho^\pi(s)^2 \le \sum_{s,a} d_\rho^\pi(s) = |\mathcal{A}|$. □

Proposition 14 (Variational Gradient Domination, Lemma 4 in Agarwal et al. (2021))**.** *Let $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, r, \rho, \gamma)$ be an MDP and let $\pi^\star$ be an optimal policy. Then for any policy $\pi \in \Pi$ and any state distribution $\mu \in \Delta(\mathcal{S})$ we have*

$$V_\rho^{\pi^\star}(r) - V_\rho^\pi(r) \le \left\|\frac{d_\rho^{\pi^\star}}{d_\mu^\pi}\right\|_\infty \max_{\pi'\in\Pi} \langle \nabla_{\pi_i} V_\rho^\pi(r), \pi' - \pi\rangle, \tag{160}$$

where we suppose that $d_\mu^\pi(s) > 0$ for all $s \in \mathcal{S}$.

Proposition 15. *For any policies $\pi, \tilde{\pi} \in \Pi$, we have*

$$\|d_\rho^\pi - d_\rho^{\tilde{\pi}}\|_1 \leq \frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d_\rho^\pi}\left[\|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1\right] \leq \frac{\gamma}{1-\gamma}\|\pi - \tilde{\pi}\|_1 \,. \tag{161}$$

Proof. By definition, the state occupancy measures d_ρ^π and $d_\rho^{\pi'}$ satisfy the Bellman consistency equations, i.e. for every state $s \in \mathcal{S}$,

$$d_\rho^\pi(s) = \rho(s) + \gamma \sum_{s'\in\mathcal{S}, a'\in\mathcal{A}} P(s|s',a')\pi(a'|s')d_\rho^\pi(s') \,, \tag{162}$$

$$d_\rho^{\tilde{\pi}}(s) = \rho(s) + \gamma \sum_{s'\in\mathcal{S}, a'\in\mathcal{A}} P(s|s',a')\tilde{\pi}(a'|s')d_\rho^{\tilde{\pi}}(s') \,. \tag{163}$$

As a consequence, we have the following upper bound:

$$|d_\rho^\pi(s) - d_\rho^{\tilde{\pi}}(s)| \leq \gamma \sum_{s'\in\mathcal{S}, a'\in\mathcal{A}} P(s|s',a')|\pi(a'|s')d_\rho^\pi(s') - \tilde{\pi}(a'|s')d_\rho^{\tilde{\pi}}(s')| \,. \tag{164}$$

Therefore we obtain by summing the above inequality over states and actions:

$$\begin{aligned}
\|d_\rho^\pi - d_\rho^{\tilde{\pi}}\|_1 = \sum_{s\in\mathcal{S}} |d_\rho^\pi(s) - d_\rho^{\tilde{\pi}}(s)| &\leq \gamma \sum_{s'\in\mathcal{S}, a'\in\mathcal{A}} \sum_{s\in\mathcal{S}} P(s|s',a')|\pi(a'|s') - \tilde{\pi}(a'|s')|d_\rho^\pi(s') \\
&= \gamma \sum_{s'\in\mathcal{S}, a'\in\mathcal{A}} |\pi(a'|s')d_\rho^\pi(s') - \tilde{\pi}(a'|s')d_\rho^{\tilde{\pi}}(s')| \\
&\leq \gamma \sum_{s'\in\mathcal{S}, a'\in\mathcal{A}} d_\rho^\pi(s') \cdot |\pi(a'|s') - \tilde{\pi}(a'|s')| + \gamma \sum_{s'\in\mathcal{S}, a'\in\mathcal{A}} \tilde{\pi}(a'|s') \cdot |d_\rho^\pi(s') - d_\rho^{\tilde{\pi}}(s')| \\
&= \gamma\mathbb{E}_{s\sim d_\rho^\pi}\left[\|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1\right] + \gamma\|d_\rho^\pi - d_\rho^{\tilde{\pi}}\|_1 \,.
\end{aligned} \tag{165}$$

Rearranging the above inequality gives the desired result. □

13 NUMERICAL SIMULATIONS

While our contributions are mainly theoretical, we performed simple simulations to illustrate the performance of our PG algorithm (Algorithm 1) in the stochastic setting. We consider three different examples from section 2.4 (a, d and e) in the potential setting with $N = 3$ agents. We set $\gamma = 0.95$, $\eta = 0.01$, $H = 20$, $A_i = 5$ (number of different actions) on a 5×5 grid environment with 512 episodes per iteration. In all the different simulations, we observe that the potential is increasing and the occupancy/Nash gaps are reducing as expected along the iterations of the algorithm.

Table 1: Multi-agent Imitation Learning (example a in section 2.4)

Iter	Potential	Occupancy Gap	KL Occupancy Gap
1	-1.0712	1.2263	1.0712
20	-0.9606	1.1851	0.9606
40	-0.8469	1.1373	0.8469
60	-0.6965	1.0277	0.6965
100	-0.4607	0.8135	0.4607
120	-0.2687	0.6321	0.2687
160	-0.1932	0.5467	0.1932
180	-0.2173	0.5782	0.2173
200	-0.1591	0.4882	0.1591

Table 2: Team Coverage and Distribution Matching (example d in section 2.4)

Iter	Potential	Nash Gap	Occupancy Gap	KL Occupancy Gap
1	-1.8612	95.0888	1.3509	1.8612
10	-1.7698	92.1758	1.3128	1.7698
20	-1.6501	88.4389	1.2765	1.6501
50	-1.2508	74.4521	1.1924	1.2508
80	-0.8294	55.9668	1.0197	0.8294
100	-0.4804	42.0560	0.7538	0.4804
130	-0.1305	22.7316	0.3542	0.1305
160	-0.0148	7.8765	0.1297	0.0148
190	-0.0040	5.3988	0.0629	0.0040

Table 3: Collective Exploration (example e in section 2.4)

Iter	Potential	Nash Gap	Occupancy Gap
1	2.9909	95.0888	1.3434
10	3.0684	92.1724	1.3079
30	3.3050	84.3933	1.2522
50	3.5616	74.7144	1.1992
80	3.9804	56.6424	1.0325
100	4.3174	43.2140	0.7834
150	4.7797	15.6444	0.2363
200	4.8165	7.7096	0.1010
210	4.8198	7.6516	0.0872