# On the convergence of policy gradient methods to Nash equilibria in general stochastic games

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Multi-agent learning in stochastic $N$-player games is a notoriously difficult problem because, in addition to their changing strategic decisions, the players of the game must also contend with the fact that the game itself evolves over time, possibly in a very complicated manner. Because of this, the equilibrium convergence properties of popular learning algorithms – like policy gradient and its variants – are poorly understood, except in specific classes of games (such as potential or two-player, zero-sum games). In view of all this, we examine the long-run behavior of policy gradient methods with respect to Nash equilibrium policies that are second-order stationary (SOS) in a sense similar to the type of KKT sufficiency conditions used in optimization. Our analysis shows that SOS policies are locally attracting with high probability, and we show that policy gradient trajectories with gradient estimates provided by the REINFORCE algorithm achieve an $\mathcal{O}(1/\sqrt{n})$ convergence rate to such equilibria if the method's step-size is chosen appropriately. On the other hand, when the equilibrium in question is *deterministic*, we show that this rate can be improved dramatically and, in fact, policy gradient methods converge within a *finite* number of iterations in that case.

## 1 Introduction

Ever since they were introduced by Shapley [51] in the 1950's, stochastic games have comprised one of the staples of non-cooperative game theory, with a range of pioneering applications to multi-agent reinforcement learning [8, 28, 65], unmanned vehicles [11, 35, 48, 50, 62], general game-playing [6, 7, 38, 52, 58], etc. Informally, a stochastic game evolves in discrete time as follows: At each point in time, the players are at a given state which determines the rules of the game for that stage. The actions of the players in this state determine not only their instantaneous payoffs (as defined by the stage game), but also the transition probabilities towards the next state of the process. In this way, each player has to balance two distinct – and often competing – objectives: optimizing the payoffs of *today* versus picking a possibly suboptimal action which could yield significant benefits *tomorrow* (i.e., by influencing the transitions of the process towards a more favorable state for the player).

Since all players in the game are involved in a similar dilemma, the decision-making problem for each player is a very complicated affair. In particular, in addition to their changing strategic decisions, the players of the game must also contend with the fact that the game itself evolves over time. Because of this, even the existence of a Nash equilibrium policy – viz. a stationary Markovian policy that is stable to unilateral deviations [20] – is far more difficult to prove compared to standard, stateless normal form games; for a comprehensive survey, see [42, 53, 67] and references therein.

The question we seek to address in this paper is whether an ensemble of boundedly rational players can reach an equilibrium policy in a stochastic game. Specifically, if players do not have sufficient information – or the computational resources required – to solve a Bellman equation in very high

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

dimensions [55, 59], it is not at all clear if they would somehow end up playing a Nash policy in the long run. After all, the complexity of most games increases exponentially with the number of players, so the identification of a game's equilibria quickly becomes prohibitively difficult [17, 29, 34, 36].

**Our contributions in the context of related work.** This issue has sparked a vigorous literature with important implications for the series of applications mentioned above [3, 54, 64]. On the downside, these efforts also have to grapple with a series of strong lower bounds for computing weaker solution concepts like coarse correlated equilibria in turn-based stochastic games [16, 29]. On that account, a recent line of work has instead focused on understanding specific sub-classes of stochastic games, like *min-max* [12, 15, 49, 60] and common interest *potential* games [18, 33, 68], or computing relaxed solution concepts where either the stationarity or the Markov property has been dropped [16].

Our paper focuses on episodic playing in random stopping games – in lieu of learning in ergodic stochastic games with an infinite horizon [34, 44] – and considers the general class of policy gradient methods, first introduced by [30, 31, 56, 61] and subsequently popularized in single-agent reinforcement learning by [2, 10, 27, 63]. Concretely, this means that the sequence of play evolves episode-by-episode: within each episode, the players commit a policy and play the game, and from one episode to the next, they use an iterative gradient step to update their policy and continue playing.

Our main contributions in this general context may then be summarized as follows:

1. We introduce a flexible algorithmic template for the analysis of policy gradient methods which accounts for different information and update frameworks – from perfect policy gradients to value-based estimates obtained per episode, e.g., via the Reinforce algorithm [4, 56, 61].
2. Within this framework, we show that Nash policies that satisfy a certain strategic stability condition are locally attracting with arbitrarily high probability. Moreover, to estimate the method's rate of convergence, we focus on Nash policies that satisfy a second-order sufficiency condition similar to the type of KKT conditions used in optimization, and we show that such policies enjoy an $\mathcal{O}(1/\sqrt{n})$ convergence rate in terms of squared distance.
3. Finally, we also consider the method's convergence to *deterministic* Nash policies and we show that, generically, the above rate can be improved dramatically. By a simple tweak to the method's projection step, we are able to show that the induced sequence of play converges to equilibrium in a *finite* number of iterations, despite all the noise and uncertainty in the process.

It is worth mentioning that our results focus squarely on the convergence of the actual, inter-episode trajectory of play – as opposed to "best-iterate" or ergodic convergence results. In addition, obtaining guarantees using stochastic estimators (cf. Reinforce) greatly alleviate the burden of exact gradient computations that are otherwise beyond reach in low-compute / low-memory practical environments. This aspect of our results is especially relevant for multi-agent reinforcement learning scenarios where agents learn "on the fly", and is a property with important ramifications for many of the practical applications of stochastic games.

## 2 Preliminaries

**2.1. Game formulation.** Throughout this work we consider $N$-player generic stochastic games, where players repeatedly select actions in a shared Markov decision process (MDP) with the goal of maximizing their individual value functions. Formally, we study the tabular version with random stopping of general stochastic games, which is specified by a tuple $\mathcal{G} = (\mathcal{S}, \mathcal{N}, \{\mathcal{A}_i, R_i\}_{i\in\mathcal{N}}, P, \zeta, \rho)$ with the following primitives:

- A finite set of *agents* $i \in \mathcal{N} = \{1, 2, \dots, N\}$ and a finite set of *states* $\mathcal{S} = \{1, \dots, S\}$.
- For each $i \in \mathcal{N}$, a finite space of *actions* (or *pure strategies*) $\mathcal{A}_i$ indexed by $\alpha_i = 1, \dots, A_i = |\mathcal{A}_i|$. We will write $\mathcal{A} = \prod_{i\in\mathcal{N}} \mathcal{A}_i$ and $\mathcal{A}_{-i} = \prod_{j\neq i} \mathcal{A}_j$ for the action space of all agents and that of all agents other than $i$ respectively. In a similar vein, we will also write $\alpha = (\alpha_i, \alpha_{-i})$ when we want to highlight the action $\alpha_i$ of player $i$ against the action profile $\alpha_{-i}$ of $i$'s opponents.

- For each $i \in \mathcal{N}$, we will write $R_i \colon \mathcal{S} \times \mathcal{A} \to [-1, 1]$ for the *reward function* of agent $i \in \mathcal{N}$, i.e., $R_i(s, \alpha_i, \alpha_{-i})$ will denote the value of the reward of agent $i$ when the game is at state $s \in \mathcal{S}$, the focal agent $i \in \mathcal{N}$ plays $\alpha_i \in \mathcal{A}_i$, and all other agents take actions $\alpha_{-i} \in \mathcal{A}_{-i}$.
- The game transits from one state to another according to a Markov transition process, so that $P(s' \mid s, \alpha)$ denotes the probability of transitioning from $s$ to $s'$ when $\alpha \in \mathcal{A}$ is the action profile chosen by the agents.
- Given an action profile $\alpha$ at state $s$, the process terminates with probability $\zeta_{s,a} > 0$, i.e., $\zeta_{s,a} = 1 - \sum_{s' \in \mathcal{S}} P(s' \mid s, \alpha)$; for convenience, we will write $\zeta := \min_{s,a}\{\zeta_{s,a}\}$.
- $\rho \in \Delta(\mathcal{S})$ is the distribution for the initial state of the game.

**Episodic Setting.** We consider an episodic setting, where in each episode a realization of the game is completed. At every time step $t \geq 0$ of each episode, all agents observe the common state $s_t \in \mathcal{S}$, select actions $\alpha_t$ and receive rewards $\{R_i(s_t, \alpha_t)\}_{i \in \mathcal{N}}$. Then, with probability $\zeta_{s_t,\alpha_t}$ the game terminates, and with probability $1 - \zeta_{s_t,\alpha_t}$, it moves to the state $s_{t+1}$, which is drawn according to $P(\cdot|s_t, \alpha_t)$. Denoting the realized reward of player $i$ at time $t$ as $r_{i,t} := R_i(s_t, \alpha_t)$, we will write $\tau = (s_t, \alpha_t, r_t)_{t \leq T(\tau)}$ to denote the trajectory of the episode, where $r_t := (r_{i,t})_{i \in \mathcal{N}}$, and $T(\tau)$ the time the episode terminates.

**Policies and value functions.** We consider *stationary Markovian* policies, i.e., policies that do not depend on the time-step and the history, given the current state of the game. More specifically, for each agent $i \in \mathcal{N}$, a *policy* $\pi_i \colon \mathcal{S} \to \Delta(\mathcal{A}_i)$ specifies a probability distribution over the actions of agent $i$ in state $s \in \mathcal{S}$, i.e., $\alpha_i \sim \pi_i(\cdot|s)$ denotes the (random) action drawn by agent $i$ at state $s \in \mathcal{S}$ according to $\pi_i$, viewed here as an element of $\Pi_i := \Delta(\mathcal{A}_i)^{\mathcal{S}}$. In addition, we will also write $\pi = (\pi_i)_{i \in \mathcal{N}} \in \Pi := \prod_i \Pi_i$ and $\pi_{-i} = (\pi_j)_{j \neq i} \in \Pi_{-i} := \prod_{j \neq i} \Pi_j$ for the policy profile of all agents and all agents other than $i$, respectively.

The expected reward of agent $i \in \mathcal{N}$ if agents follow policy $\pi$, starting from initial state $s \in \mathcal{S}$, defines the *value function* of agent $i$, denoted as $V_{i,s}(\pi)$, and is equal to

$$V_{i,s}(\pi) := \mathbb{E}_{\tau \sim \mathrm{MDP}}\Big[\sum\nolimits_{t=0}^{T(\tau)} R_i(s_t, a_t)\Big| s_0 = s\Big] \tag{1}$$

where $\tau \sim$ MDP denotes the randomness induced by the policy profile $\pi$, and the state-transition probabilities of the MDP. Overloading the notation, we set $V_{i,\rho}(\pi) := \mathbb{E}_{s \sim \rho}[V_{i,s}(\pi)]$. Although value functions are, in general, non-convex, they share similar smoothness properties with the payoff functions of normal form games, namely bounded and Lipschitz gradients. For precise statements, we defer to the paper's supplement.

**Visitation distribution and the mismatch coefficient.** For a policy profile $\pi \in \Pi$ and an arbitrary initial state distribution $s_0 \sim \rho$, we define the discounted state visitation measure/distribution as

$$\tilde{d}^{\pi}_{\rho}(s) = \mathbb{E}_{\tau \sim \mathrm{MDP}}\Big[\sum\nolimits_{t=0}^{T(\tau)} \mathbb{1}\{s_t = s\}\Big| s_0 \sim \rho\Big], \quad d^{\pi}_{\rho}(s) := \tilde{d}^{\pi}_{\rho}(s)/Z^{\pi}_{\rho}$$

In the appendix, we prove formally that the above definition is well-posed for the random stopping episodic framework described above, i.e., $\tilde{d}^{\pi}_{\rho}(s) < \infty$, so $Z^{\pi}_{\rho} := \sum_{s \in \mathcal{S}} \tilde{d}^{\pi}_{\rho}(s)$ is well-defined. In our proofs, we will leverage a standard property of visitation distributions, namely the equivalence of the expected value of state-action function and the expected cumulative value over a random trajectory. More precisely, we have:

**Lemma 1.** [Conversion Lemma] *For an arbitrary state-action function $f \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a policy profile $\pi$ and an initial state distribution $s_0 \sim \rho$, we have*

$$\mathbb{E}_{\tau \sim \mathrm{MDP}}\Big[\sum\nolimits_{t=0}^{T(\tau)} f(s_t, \alpha_t)\Big] = Z^{\pi}_{\rho}\, \mathbb{E}_{s \sim d^{\pi}_{\rho}}\, \mathbb{E}_{\alpha \sim \pi(\cdot|s)}[f(s, \alpha)] \tag{2}$$

Finally, to quantify the difficulty of hard-to-reach states via a policy gradient method, we will follow the standard approach of [13, 19, 39, 40, 68] and use an appropriately-defined distribution "mismatch coefficient", generalizing the single-agent counterpart of Agarwal et al. [1]. More precisely, for a stochastic game $\mathcal{G}$, we define the *minimax mismatch coefficient* as $\mathcal{C}_{\mathcal{G}} := \max_{\pi,\pi' \in \Pi}\{\|\tilde{d}^{\pi}_{\rho}/\tilde{d}^{\pi'}_{\rho}\|_{\infty}\}$. Similar to prior work in this direction [1, 5, 15], we will assume $\mathcal{C}_{\mathcal{G}}$ is finite, which, equivalently, means that $d^{\pi}_{\rho}(s) > 0$ for any policy $\pi$ and state $s$.

**2.2. Solution Concepts.** The most widely used solution concept in game theory is that of a Nash equilibrium i.e., a strategy profile $\pi^* \in \Pi$ that discourages unilateral deviations. However, in stochastic games, the definition of a Nash policy is much more involved because of the existence of multiple states and steps, cf. [20, 51, 53, 57]. Formally, we have the following definition:

**Definition 1** (Nash Policy)**.** A policy $\pi^* = (\pi_i^*)_{i\in\mathcal{N}} \in \Pi$ is said to be a *Nash policy* for a given distribution of initial states $\rho \in \Delta(\mathcal{S})$ if, for every player $i \in \mathcal{N}$, we have

$$V_{i,\rho}(\pi_i^*; \pi_{-i}^*) \geq V_{i,\rho}(\pi_i; \pi_{-i}^*) \qquad \forall i \in \mathcal{N}, \forall \pi_i \in \Delta(\mathcal{A}_i)^{\mathcal{S}} \tag{NE}$$

In contrast to general non-convex continuous games, stochastic games satsify a version of the well-known Polyak-Łojasiewicz condition [46] but with linear gradient growth, also known as a *gradient dominance property* (GDP) [1, 5]. For the multi-agent case, [15, 68] showed that a similar property holds even in an episodic setting:

**Lemma 2.** [Gradient dominance property] *For any policy profile $\pi = (\pi_i)_{i\in\mathcal{N}} \in \Pi$, we have that*

$$V_{i,\rho}(\pi_i'; \pi_{-i}) - V_{i,\rho}(\pi_i; \pi_{-i}) \leq C_{\mathcal{G}} \max_{\bar{\pi}_i \in \Pi_i} \langle \nabla_i V_{i,\rho}(\pi), \bar{\pi}_i - \pi_i \rangle \tag{GDP}$$

*for any unilateral deviation $\pi_i' \in \Pi_i$ of each player $i \in \mathcal{N}$.*

*Remark.* In the above and throughout our paper, we will write $\nabla_i$ to denote the gradient of the quantity in question with respect to $\pi_i$, i.e., when $\pi_{-i}$ is kept fixed and only $\pi_i$ is varied. For concision, we will write $v_i(\pi) = \nabla_i V_{i,\rho}(\pi)$ for the individual gradient of player $i$'s value function, and $v(\pi) = (v_i(\pi))_{i\in\mathcal{N}}$ for the ensemble thereof. ¶

Thanks to (GDP), it is straightforward to check that first-order stationary (FOS) points of $V$ are Nash policies. Formally, as in [15, 33, 68], we have the following characterization:

**Lemma 3.** [First-order stationary policies are Nash] *A profile $\pi^* = (\pi_i^*)_{i\in\mathcal{N}} \in \Pi$ is a Nash policy profile if and only if it satisfies the first-order stationary condition*

$$\langle v(\pi^*), \pi - \pi^* \rangle \leq 0 \quad \textit{for all } \pi \in \Pi. \tag{FOS}$$

Leonardos et al. [33] and Zhang et al. [68] proved a relaxation of the above lemma to the effect that policies that satisfy (FOS) up to $\varepsilon$ (i.e., in lieu of 0 in the RHS) are $\mathcal{O}(\varepsilon)$-Nash. Going in the other direction, we will consider the following series of refinements of Nash policies which are particularly important from a learning standpoint [32, 37, 53]:

**Definition 2.** Let $\pi^* = (\pi_i^*)_{i\in\mathcal{N}} \in \Pi$ be a Nash policy. Then:

- $\pi^*$ is *stable* if $\langle v(\pi), \pi - \pi^* \rangle < 0$ for all $\pi \neq \pi^*$ close to $\pi^*$.
- $\pi^*$ is *second-order stationary* if it satisfies the sufficiency condition

$$(\pi - \pi^*)^\top \operatorname{Jac}_v(\pi^*)(\pi - \pi^*) < 0 \quad \text{for al } \pi \in \Pi \backslash \{\pi^*\}, \tag{SOS}$$

  where $\operatorname{Jac}_v(\pi^*) = (\nabla_j v_i(\pi^*))_{i,j\in\mathcal{N}} = (\nabla_j \nabla_i V_i(\pi^*))_{i,j\in\mathcal{N}}$ denotes the Jacobian of $v$ at $\pi^*$.
- $\pi^*$ is *deterministic* if it induces a deterministic selection rule $\pi_i^* : \mathcal{S} \to \mathcal{A}_i$ for all $i \in \mathcal{N}$.
- $\pi^*$ is *strict* if it is deterministic and (FOS) holds as a strict inequality whenever $\pi \neq \pi^*$.

Intuitively, the condition for equilibrium stability is the game-theoretic analogue of a first-order KKT sufficiency condition, while the condition for second-order stationarity is the second-order version thereof. In this regard, the distinction between first-order stationary, stable and second-order stationary points is formally analogous to the distinction between critical points, minimizer, and second-order minimum points in optimization. As for deterministic policies, we should mention that, generically – i.e., except on a set which is meager in the sense of Baire [22, 32] – deterministic policies are also strict, so we will use the two terms interchangeably.

Importantly, as we show in the appendix, these refinements admit the following characterizations:

**Proposition 1.** *Let $\pi^* = (\pi_i^*)_{i \in \mathcal{N}} \in \Pi$ be a Nash policy. Then:*

*a) If $\pi^*$ is second-order stationary, there exists some $\mu > 0$ such that*

$$\langle v(\pi), \pi - \pi^* \rangle \leq -\mu \, \|\pi - \pi^*\|^2 \qquad \text{for all } \pi \text{ sufficiently close to } \pi^*. \tag{3a}$$

*b) If $\pi^*$ is strict, there exists some $\mu > 0$ such that*

$$\langle v(\pi), \pi - \pi^* \rangle \leq -\mu \, \|\pi - \pi^*\| \qquad \text{for all } \pi \text{ sufficiently close to } \pi^*. \tag{3b}$$

In view of all the above, we get the following string of implications for equilibria in generic games:

$$\text{strict/deterministic} \implies \text{SOS} \implies \text{stable} \implies \text{FOS} = \text{Nash} \tag{4}$$

For posterity, we only note here that it is plausible to except that more refined solution concepts should enjoy stronger convergence properties; we will confirm this intuition in the sequel.

## 3 Policy gradient methods

We now proceed to describe our general model for learning in stochastic games. In tune with the episodic framework described in the previous section, we will likewise consider a learning framework where agents follow a specific policy profile $\pi_n$ within each episode, and update it from one episode to the next with the objective of increasing their individual rewards.

Formally, our approach will adhere to the following inter-episode sequence of events:

1. At the beginning of each episode $n = 1, 2, \dots$, every agent $i \in \mathcal{N}$ chooses a policy $\pi_{i,n} \in \Pi_i$.
2. Within the $n$-th episode, each player executes their chosen policy $\pi_{i,n}$, inducing in this way an intra-episode trajectory of play $\tau_n = (s_t^{(n)}, \alpha_t^{(n)}, r_t^{(n)})_{t \leq T(\tau_n)}$.
3. Once the episode terminates, agents update their policies, and the process repeats.

In terms of feedback, we will treat several models, depending on what type of information is available to the agents during play. To that end, we will focus on the generic policy gradient (PG) template

$$\pi_{n+1} = \operatorname{proj}_\Pi(\pi_n + \gamma_n \hat{v}_n) \tag{PG}$$

where:

1. $\pi_n = (\pi_{i,n})_{i \in \mathcal{N}} \in \Pi$ denotes the player's policy profile at each episode $n = 1, 2, \dots$
2. $\hat{v}_n = (\hat{v}_{i,n})_{i \in \mathcal{N}} \in \prod_i \left(\mathbb{R}^{\mathcal{A}_i}\right)^{\mathcal{S}}$ is an estimate for the agents' inidividual policy gradients.
3. $\gamma_n > 0$ is the method's step-size, for which we will assume throughout that $\sum_n \gamma_n = \infty$; typically, (PG) is run with a step-size of the form $\gamma_n = \gamma/(n+m)^p$ for some $\gamma > 0$, $m \geq 0$ and $p \geq 0$.
4. $\operatorname{proj}_\Pi \colon \prod_i \left(\mathbb{R}^{\mathcal{A}_i}\right)^{\mathcal{S}} \to \Pi$ denotes the Euclidean projection to the agents' policy space $\Pi$.

Regarding the gradient signal $\hat{v}_n$, we will decompose it as

$$\hat{v}_n = v(\pi_n) + U_n + b_n \tag{5}$$

where

$$U_n = \hat{v}_n - \mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n] \quad \text{and} \quad b_n = \mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n] - v(\pi_n). \tag{6}$$

In the above, we treat $\pi_n$ as a stochastic process on some complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and we write $\mathcal{F}_n \coloneqq \mathcal{F}(\pi_1, \dots, \pi_n) \subseteq \mathcal{F}$ for the history (adapted filtration) of $\pi_n$ up to – and including – stage $n$.

By definition, $\mathbb{E}[U_n \,|\, \mathcal{F}_n] = 0$ and $b_n$ is $\mathcal{F}_n$-measurable, so $U_n$ can be intepreted as a random, zero-mean error relative to $v(\pi_n)$, whereas $b_n$ captures all systematic (non-zero-mean) errors. To make this precise, we will further assume that $b_n$ and $U_n$ are bounded as

$$\mathbb{E}[\|b_n\| \,|\, \mathcal{F}_n] \leq B_n \qquad \text{and} \qquad \mathbb{E}[\|U_n\|^2 \,|\, \mathcal{F}_n] \leq \sigma_n^2 \tag{7}$$

**Algorithm 1:** Reinforce

**Input:** $\hat{\pi} \in \Pi, \tau = (s_t, \alpha_t, r_t)_{t \leq T(\tau)} \in \mathcal{T}$
**for** $i = 1, \dots, N$ **do**
  $R_i(\tau) \leftarrow \sum_{t=0}^{T(\tau)} r_{i,t}$
  $\Lambda_i(\tau) \leftarrow \sum_{t=0}^{T(\tau)} \nabla_i(\log \hat{\pi}_i(a_{i,t}|s_t))$
  $\hat{v}_i \leftarrow R_i(\tau) \cdot \Lambda_i(\tau)$
**end for**
**return** $\{\hat{v}_i\}_{i \in \mathcal{N}}$

**Algorithm 2:** $\varepsilon$-Greedy Policy Gradient

**Input:** $\pi_1, \{\gamma_n\}_{n \in \mathbb{N}}, \{\varepsilon_n\}_{n \in \mathbb{N}}$
**for** $n = 1, 2, \dots$ **do**
  $\hat{\pi}_n \leftarrow (1 - \varepsilon_n)\pi_n + \frac{\varepsilon_n}{|\mathcal{A}|}$
  Sample $\tau_n \sim \mathrm{MDP}(\hat{\pi}_n|s_0)$
  $\hat{v}_n \leftarrow \textsc{Reinforce}(\hat{\pi}_n, \tau_n)$
  $\pi_{n+1} \leftarrow \mathrm{proj}_\Pi(\pi_n + \gamma_n \hat{v}_n)$
**end for**

where the sequences $B_n$ and $\sigma_n$ , $n = 1, 2, \dots$, are to be construed as deterministic upper bounds on the bias, fluctuations, and magnitude of the gradient signal $\hat{v}_n$. Depending on these bounds, a gradient signal with $B_n = 0$ will be called *unbiased*, and an unbiased signal with $\sigma_n = 0$ will be called *perfect*. More generally, we will assume that the above statistics are bounded as

$$B_n = \mathcal{O}(1/n^{\ell_b}) \qquad \text{and} \qquad \sigma_n = \mathcal{O}(n^{\ell_\sigma}) \tag{8}$$

for some $\ell_b, \ell_\sigma > 0$ which depend on the specific model under consideration. For concreteness, we describe below three basic models that adhere to the above template for $\hat{v}_n$ in order of decreasing information requirements:

**Model 1** (Full gradient information)**.** The first model we will consider assumes that agents observe their *full policy gradients*, i.e.,

$$\hat{v}_n = v(\pi_n) \tag{9}$$

implying in particular that $U_n = b_n = 0$. This model is fully deterministic across episodes (though intra-episode play remains stochastic). In particular, it tacitly assumes that agents know the game (and can observe their opponents' policies) sufficiently well so as to calculate the full gradients of their individual value functions $V_{i,\rho}$, cf. [2, 33, 68] and references therein. ¶

**Model 2** (Learning with stochastic gradients)**.** A relaxation of the above model which is particularly relevant when the game involves training over datasets concerns the case where the player have access to stochastic policy gradients, i.e., unbiased gradient estimates of the form

$$\hat{v}_n = v(\pi_n) + U_n \tag{10}$$

with $\mathbb{E}[U_n \,|\, \mathcal{F}_n] = 0$ (so we can formally take $\ell_b = \infty$ and $\ell_\sigma = 0$ in Eq. (8) above). This case is considered in [66] and [43]. ¶

**Model 3** (Value-based learning)**.** The last model we will consider concerns the case where agents only have access to their realized values and need to reconstruct their individual gradients based on this information. A widely used method to achieve this is via the Reinforce subroutine, which we describe in pseudocode form in Algorithm 1. In words, when employing Reinforce, each agent $i \in i$ commits to a sampling policy $\hat{\pi}_i \in \Pi_i$ and executes it in an episode of the stochastic game in play. Then, at the end of the episode, players gather the total reward $R_i(\tau) \leftarrow \sum_{t=0}^{T(\tau)} r_{i,t}$ associated to the intra-episode trajectory of play $\tau$, and they estimate their policy gradients via the so-called "log-trick" [61] as

$$\hat{v}_i = R_i(\tau) \cdot \sum_{t=0}^{T(\tau)} \nabla_i(\log \hat{\pi}_i(a_{i,t}|s_t)). \tag{11}$$

Lemma 4 below provides the vital statistics of the Reinforce estimator:

**Lemma 4.** *Suppose that each agents $i \in \mathcal{N}$ follows a stationary policy $\pi_i \in \Pi_i$. Then, letting $\kappa_i = \min_{s \in \mathcal{S}, \alpha_i \in \mathcal{A}_i} \pi_i(\alpha_i|s)$ for each $i \in \mathcal{N}$, we have*

*a)* $$\mathbb{E}_{\tau \sim \mathrm{MDP}}[\textsc{Reinforce}(\pi)] = v(\pi). \tag{12a}$$

*b)* $$\mathbb{E}_{\tau \sim \mathrm{MDP}}\Big[\|\textsc{Reinforce}_i(\pi) - v_i(\pi)\|^2\Big] \leq \frac{24|\mathcal{A}_i|}{\kappa_i \zeta^4}. \tag{12b}$$

Thus, if Reinforce is executed at $\hat{\pi} \leftarrow \pi_n$ at each episode $n = 1, 2, \dots$, we will have

$$\mathbb{E}[\hat{v}_{i,n}] = v_i(\pi_n) \qquad \text{and} \qquad \mathbb{E}[\|U_{i,n}\|^2 \mid \mathcal{F}_n] \le \frac{24|\mathcal{A}_i|}{\zeta^4 \min_{s\in\mathcal{S},\alpha_i\in\mathcal{A}_i} \pi_{i,n}(\alpha_i|s)}. \tag{13}$$

This means that we will always have $B_n = 0$ for the bias of the estimator, but its variance could be unbounded if $\pi_n$ gets close to the boundary of $\Pi$. For this reason, Reinforce is typically paired with an explicit exploration step that modifies the sampling policy of the $n$-th episode to

$$\hat{\pi}_{i,n} = (1-\varepsilon_n)\pi_{i,n} + \varepsilon_n \operatorname{Unif}_{\mathcal{A}_i}. \tag{14}$$

i.e., $\hat{\pi}_{i,n}$ is the mixture between $\pi_{i,n}$ and the uniform distribution $\operatorname{Unif}_{\mathcal{A}_i}$ over $\mathcal{A}_i$. The resulting algorithm is known as $\varepsilon$-Greedy Policy Gradient; for a pseudocode, see Algorithm 2.

Importantly, by calling Reinforce at $\hat{\pi}_n$, $\hat{v}_n$ becomes biased (because of the difference between $\hat{\pi}_n$ and $\pi_n$), but its variance is bounded; in particular, by invoking Lemma 4, we have

$$\mathbb{E}[\|b_{i,n}\| \mid \mathcal{F}_n] \le G\varepsilon_n \qquad \text{and} \qquad \mathbb{E}[\|U_{i,n}\|^2 \mid \mathcal{F}_n] \le \frac{24|\mathcal{A}_i|^2}{\varepsilon_n \zeta^4} \tag{15}$$

where $G$ is a constant that depends on the smoothness of $V$ and the cardinalities of $\mathcal{A}$ and $\mathcal{S}$. In this way, Algorithm 2 can be seen as a special case of (PG) with $B_n = \mathcal{O}(\varepsilon_n)$ and $\sigma_n = \mathcal{O}(1/\sqrt{\varepsilon_n})$. ¶

# 4 Convergence analysis and results

We are now in a position to state and discuss our main results. For convenience, we will present our results in order of increasing structure, starting with stable policies, and then moving on to second-order stationary and deterministic Nash policies. All proofs are deferred to the appendix.

**4.1. Asymptotic convergence to stable Nash policies.** Our first convergence result concerns Nash policies that satisfy the stability requirement $\langle v(\pi), \pi - \pi^* \rangle < 0$ of Definition 2. In this case, we have the following guarantee:

**Theorem 1.** *Let $\pi^*$ be a stable Nash policy, and let $\pi_n$ be the sequence of play generated by (PG) with step-size $\gamma_n = \gamma/(n+m)^p$, $p \in (1/2, 1]$, and policy gradient estimates such that $p + \ell_b > 1$ and $p - \ell_\sigma > 1/2$ as per (8). Then there exists a neighborhood $\mathcal{U}$ of $\pi^*$ in $\Pi$ such that, for any given $\delta > 0$, we have*

$$\mathbb{P}(\pi_n \text{ converges to } \pi^* \mid \pi_1 \in \mathcal{U}) \ge 1 - \delta \tag{16}$$

*provided that $\gamma$ is small enough (or $m$ large enough) relative to $\delta$.*

**Corollary 1.** *Suppose that Models 1–3 are run with a step-size of the form $\gamma_n = \gamma/(n+m)^p$, $p > 1/2$, and if applicable, an exploration parameter $\varepsilon_n = \varepsilon/(n+m)^r$ such that $1 - p < r < 2p - 1$. Then:*

- *For Models 1 and 2: the conclusions of Theorem 1 hold as stated.*
- *For Model 3: the conclusions of Theorem 1 hold as long as $p > 2/3$.*

We note here that Theorem 1 provides a trajectory convergence guarantee which is otherwise quite difficult to obtain even in structured stochastic games. For example, if we zoom in on the class of stochastic potential (or min-max) games, the existing guarantees in the literature concern the "best iterate" of the algorithm, cf. [33, 68] and references therein. Because of this, said guarantees do not apply to the actual trajectory of play generated by (PG); this makes them less suitable for agent-based learning where the players involved are learning "as they go", as opposed to *simulating* the game in order to approximately compute an equilibrium policy offline.

We should also note that the convergence guarantees of Theorem 1 hold locally with arbitrarily high probability. Without further assumptions, it is not possible to obtain global trajectory convergence guarantees that hold with probability 1, even in the simple case where the game only has a single state – that is, the case of learning in finite normal form games. In this (much simpler) setting, the well-known impossibility result of Hart and Mas-Colell [24, 25] shows that it is not possible to expect convergence to Nash equilibrium in all games – not even locally. In this regard, the local convergence caveat in Theorem 1 cannot be lifted without further structural properties in place – such as the existence of a potential function in the spirit of [33].

**4.2. Convergence to second-order stationary policies.** Albeit valuable as an asymptotic convergence guarantee, Theorem 1 does not provide an indication of how long it will take players to actually converge to a Nash policy. Of course, in full generality, it is not plausible to expect to be able to derive such a convergence rate because the stability requirement provides no indication on how fast the players' policy gradients stabilize near a solution. This kind of estimate is provided by the second-order sufficient condition (SOS), which allows us to establish sufficient control over the sequence of play as indicated by the following theorem.

**Theorem 2.** *Let $\pi^*$ be a Nash policy such that (SOS) holds on some open set $\mathcal{B}$ containing $\pi^*$, and let $\pi_n$ be the sequence of play generated by (PG) with step-size $\gamma_n = \gamma/(n+m)^p$, $p \in (1/2, 1]$, and policy gradient estimates such that $p + \ell_b > 1$ and $p - \ell_\sigma > 1/2$ as per (8). Then:*

1. *There exists a neighborhood $\mathcal{U}$ of $\pi^*$ in $\Pi$ such that, for any confidence level $\delta > 0$, the event*
$$\mathcal{E} = \{\pi_n \in \mathcal{B} \text{ for all } n = 1, 2, \dots\} \tag{17}$$
*occurs with probability $\mathbb{P}(\mathcal{E} \mid \pi_1 \in \mathcal{U}) \geq 1 - \delta$ if $m$ is large enough relative to $\delta$.*

2. *The sequence $\pi_n$ converges to $\pi^*$ with probability 1 on $\mathcal{E}$; in particular, we have*
$$\mathbb{P}(\pi_n \text{ converges to } \pi^* \mid \pi_1 \in \mathcal{U}) \geq 1 - \delta \tag{18}$$
*if $m$ is large relative to $\delta$. Moreover, conditioned on $\mathcal{E}$ and taking $q = \min\{\ell_b, p - 2\ell_\sigma\}$, we have*
$$\mathbb{E}[\|\pi_n - \pi^*\|^2 \mid \mathcal{E}] = \begin{cases} \mathcal{O}(1/n^{2\mu\gamma}) & \text{if } p = 1 \text{ and } 2\mu\gamma < q, \\ \mathcal{O}(1/n^q) & \text{otherwise}. \end{cases} \tag{19}$$

**Corollary 2.** *Suppose that Models 1–3 are run with a step-size of the form $\gamma_n = \gamma/(n+m)^p$, $p > 1/2$, and if applicable, an exploration parameter $\varepsilon_n = \varepsilon/(n+m)^{p/2}$. Then:*

- *For Models 1 and 2: the conclusions of Theorem 2 hold with $q = p$; in particular, (19) gives an $\mathcal{O}(1/n)$ rate of convergence if $p = 1$ and $2\mu\gamma > q$.*
- *For Model 3: the conclusions of Theorem 2 hold for $p > 2/3$ with $q = p/2$; in particular, (19) gives an $\mathcal{O}(1/\sqrt{n})$ rate of convergence if $p = 1$ and $2\mu\gamma > q$.*

Besides providing a general framework for achieving trajectory convergence, Theorem 2 gives the rates of convergence of the sequence of play to the Nash policy in question. In particular, with this result in hand, one can confidently argue about the distance of the iterates of (PG) from equilibrium in a series of different environments. More to the point, this convergence guarantee allows the algorithm designer to adapt the parameters of the learning process according to the complexity and limitations of the environment, a feature which further highlights the significance of this result.

We should also note the delicate interplay between the method's step-size and the achieved convergence rate. In the case of Model 1, Corollary 2 suggests a step-size of the form $\gamma_n = \Theta(1/n)$, leading to a $\mathcal{O}(1/n)$ convergence rate. As we show in the appendix, this rate can be improved: in the deterministic case with perfect gradient information, (PG) with a suitably chosen constant step-size achieves a *geometric* convergence rate, i.e., $\|\pi_n - \pi^*\| = \mathcal{O}(\exp(-\rho n))$ for some $\rho > 0$. By contrast, in the case of Model 2, the $\mathcal{O}(1/n)$ rate we provide cannot be improved, even if the quadratic minorant (3a) that characterizes SOS policies holds *globally* – and this because the learning process is running against standard lower bounds from convex optimization [9, 41].

Perhaps the most significant guarantee from a practical point of view is the $\mathcal{O}(1/\sqrt{n})$ convergence rate attained in Model 3 (cf. Algorithms 1 and 2). This guarantee amounts to a $\mathcal{O}(1/n^{1/4})$ convergence rate in terms of the (non-squared) distance to equilibrium which, mutatis mutandis, represents a notable improvement over the $\mathcal{O}(1/n^{1/6})$ guarantee of Leonardos et al. [33] (expressed in norm values). Of course, the latter guarantee is global – because the focus of [33] is stochastic *potential* games – but it also concerns the "best iterate" of the process (not its "last iterate"), so the two results are not immediately comparable. However, a useful "best-of-both-worlds" heuristic that can be inferred by the combination of these works is that, given a budget of training episodes, Algorithm 2 can be run with a constant step-size as per [33] for a sufficient fraction of this budget, and then with a $\mathcal{O}(1/n)$ "cooldown" schedule for the rest. In this way, after an aggressive "exploration" phase, the algorithm's $\mathcal{O}(1/n^{1/4})$ rate would kick in and supply faster stabilization to an SOS policy.

**4.3. Convergence to deterministic Nash policies.** Our last series of results concerns the rate of convergence to deterministic Nash policies in generic stochastic games. As we discussed in Section 2, deterministic Nash policies also satisfy (SOS), so the rate of convergence of (PG) to such policies can be harvested directly from Theorem 2. However, as we show below, a simple projection tweak in (SOS) can improve this rate dramatically.

The tweak in question is inspired by the geometry of $\Pi$ around a deterministic policy: by definition, such policies are corner points of $\Pi$, so any consistent drift towards them will cause $\pi_n$ to hit the boundary of $\Pi$ in finite time. Of course, under (PG), the process may rebound from the boundary and return to the interior of $\Pi$ if the policy gradient estimate is not particularly good at a given iteration of the algorithm. However, if we replace the projection step of (PG) with a "lazy projection" in the spirit of Zinkevich [69], the aggregation of gradient steps will eventually push the process far inside the normal cone of $\Pi$ at $\pi^*$, so rebounds of this type can no longer occur.

Formally, we will consider the following *lazy policy gradient* (LPG) scheme:

$$y_{n+1} = y_n + \gamma_n \hat{v}_n \qquad \pi_{n+1} = \operatorname{proj}_{\Pi}(y_{n+1}) \tag{LPG}$$

where $y_n = (y_{i,n})_{i\in\mathcal{N}} \in \prod_i \left(\mathbb{R}^{\mathcal{A}_i}\right)^{\mathcal{S}}$ is an auxiliary variable that maintains an aggregate of gradient steps *before* projecting them back to $\Pi$. We then have the following convergence result:

**Theorem 3.** *Let $\pi_n$ be the sequence of play under (LPG) with step-size and policy gradient estimates such that $p + \ell_b > 1$ and $p - \ell_\sigma > 1/2$ as per (8). If $\pi^*$ is a deterministic Nash policy, there exists an unbounded open set $\mathcal{W} \subseteq \prod_i \left(\mathbb{R}^{\mathcal{A}_i}\right)^{\mathcal{S}}$ of initializations such that, for any $\delta > 0$, we have*

$$\mathbb{P}(\pi_n \text{ converges to } \pi^* \mid y_1 \in \mathcal{W}) \geq 1 - \delta, \tag{20}$$

*provided that $\gamma > 0$ is small enough. Moreover, conditioned on this event, $\pi_n$ converges to $\pi^*$ at a finite number of iterations, i.e., there exists some $n_0$ such that $\pi_n = \pi^*$ for all $n \geq n_0$.*

**Corollary 3.** *Suppose that Models 1–3 are run with parameters $\gamma_n = \gamma/n^p$, $p \in (1/2, 1]$, and if applicable, $\varepsilon_n = \varepsilon/n^r$ with $1 - p < r < 2p - 1$. Then the conclusions of Theorem 3 hold.*

Theorem 3 – and, by extension, Corollary 3 – are fairly unique because they provide a guarantee for convergence to an *exact* Nash equilibrium in a *finite* number of iterations. To the best of our knowledge, the only comparable result in the literature is that of [68], where the authors provide a finite-time convergence guarantee to strict equilibria with *perfect* policy gradients (as per Model 1). The result of Zhang et al. [68] echoes the convergence properties of deterministic first-order algorithms around sharp minima of convex functions [45], but the fact that Theorem 3 applies to models with *stochastic* gradient feedback of *unbounded* variance (Models 2 and 3 respectively) is a major difference. As far as we are aware, this is the first guarantee of its kind in the literature on learning in stochastic games.

**Concluding remarks.** A key roadblock encountered by practical applications of multi-agent reinforcement learning is the lack of universal equilibrium convergence guarantees. While the impossibility results of [24, 25] imply that unconditional convergence is not a reasonable aspiration without further assumptions on the game, the existence of local convergence results mitigates this deficiency as it provides a range of theoretically grounded stability and runtime guarantees. In this regard, second-order stationary and deterministic policies acquire particular importance, as the convergence of policy gradient methods is especially rapid and robust and this case. Of course, this leaves open the question of non-tabular settings and parametrically encoded policies, e.g., as in the case of deep reinforcement learning; we defer these investigations to future work.

## References


[1] Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 64–66. PMLR, 2020. URL http://proceedings.mlr.press/v125/agarwal20a.html.

[2] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22 (98):1–76, 2021.

[3] Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pages 551–560. PMLR, 2020.

[4] Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

[5] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

[6] Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018. doi: 10.1126/science.aao1733.

[7] Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019. doi: 10.1126/science.aay2400.

[8] Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining deep reinforcement learning and search for imperfect-information games. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

[9] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–358, 2015.

[10] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1283–1294. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/cai20d.html.

[11] Zhangjie Cao, Erdem Bıyık, Woodrow Z Wang, Allan Raventos, Adrien Gaidon, Guy Rosman, and Dorsa Sadigh. Reinforcement learning based control of imitative policies for near-accident driving. *arXiv preprint arXiv:2007.00178*, 2020.

[12] Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 27952–27964, 2021.

[13] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

[14] Kuo-Liang Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3): 463–483, 1954.

[15] Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540, 2020.

[16] Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of markov equilibrium in stochastic games. *arXiv preprint arXiv:2204.03991*, 2022.

[17] Xiaotie Deng, Yuhao Li, David Henry Mguni, Jun Wang, and Yaodong Yang. On the complexity of computing markov perfect equilibrium in general-sum stochastic games. *arXiv preprint arXiv:2109.01795*, 2021.

[18] Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo R Jovanović. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. *arXiv preprint arXiv:2202.04129*, 2022.

[19] Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR, 2020.

[20] Arlington M Fink. Equilibrium in a stochastic $n$-person game. *Journal of science of the hiroshima university, series ai (mathematics)*, 28(1):89–93, 1964.

[21] Gerald B. Folland. *Real Analysis*. Wiley-Interscience, 2 edition, 1999.

[22] Drew Fudenberg and Jean Tirole. *Game Theory*. The MIT Press, 1991.

[23] P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Probability and Mathematical Statistics. Academic Press, New York, 1980.

[24] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, September 2000.

[25] Sergiu Hart and Andreu Mas-Colell. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review*, 93(5):1830–1836, 2003.

[26] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 6936–6946, 2019.

[27] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

[28] Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

[29] Yujia Jin, Vidya Muthukumar, and Aaron Sidford. The complexity of infinite-horizon general-sum stochastic games, 2022.

[30] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[31] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

[32] Rida Laraki, Jérôme Renault, and Sylvain Sorin. *Mathematical Foundations of Game Theory*. Universitext. Springer, 2019.

[33] Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gfwON7rAm4.

[34] David S. Leslie, Steven Perkins, and Zibo Xu. Best-response dynamics in zero-sum stochastic games. *Journal of Economic Theory*, 189:105095, 2020.

[35] Konstantinos Makantasis, Maria Kontorinaki, and Ioannis Nikolos. A deep reinforcement learning driving policy for autonomous road vehicles, 2019. URL https://arxiv.org/abs/1905.09046.

[36] Eric Mazumdar, Lillian J Ratliff, Michael I Jordan, and S Shankar Sastry. Policy-gradient algorithms have no guarantees of convergence in linear quadratic games. *arXiv preprint arXiv:1907.03712*, 2019.

[37] Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1-2):465–507, January 2019.

[38] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017. doi: 10.1126/science.aam6960.

[39] Rémi Munos. Error bounds for approximate policy iteration. In *ICML*, volume 3, pages 560–567, 2003.

[40] Rémi Munos. Error bounds for approximate value iteration. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 1006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[41] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Number 87 in Applied Optimization. Kluwer Academic Publishers, 2004.

[42] Abraham Neyman and Sylvain Sorin, editors. *Stochastic Games and Applications*. NATO ASI. Kluwer Academic Publishers, 2003.

[43] Santiago Paternain. *Stochastic control foundations of autonomous behavior*. PhD thesis, University of Pennsylvania, 2018.

[44] Steven Perkins. *Advanced stochastic approximation frameworks and their applications*. PhD thesis, University of Bristol, 2013.

[45] Boris Teodorovich Polyak. *Introduction to Optimization*. Optimization Software, New York, NY, USA, 1987.

[46] B.T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. ISSN 0041-5553.

[47] Ralph Tyrrell Rockafellar and Roger J. B. Wets. *Variational Analysis*, volume 317 of *A Series of Comprehensive Studies in Mathematics*. Springer-Verlag, Berlin, 1998.

[48] Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 29(19):70–76, jan 2017. doi: 10.2352/issn.2470-1173.2017.19.avm-023. URL https://doi.org/10.2352%2Fissn.2470-1173.2017.19.avm-023.

[49] Muhammed O Sayin, Francesca Parise, and Asuman Ozdaglar. Fictitious play in zero-sum stochastic games. *arXiv preprint arXiv:2010.04223*, 2020.

[50] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

[51] Lloyd S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences of the USA*, 39: 1095–1100, 1953.

[52] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nat.*, 550(7676):354–359, 2017. doi: 10.1038/nature24270.

[53] Eilon Solan and Nicolas Vieille. Stochastic games. *Proceedings of the National Academy of Sciences*, 112 (45):13743–13746, 2015.

[54] Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

[55] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[56] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[57] Masayuki Takahashi. Stochastic games with infinitely many strategies. *Journal of Science of the Hiroshima University, Series AI (Mathematics)*, 26(2):123–134, 1962.

[58] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom Le Paine, Çaglar Gülçehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in starcraft II using multi-agent reinforcement learning. *Nat.*, 575(7782):350–354, 2019. doi: 10.1038/s41586-019-1724-z.

[59] Okko Jan Vrieze. Stochastic games with finite state and action spaces. *CWI tracts*, 1987.

[60] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4259–4299. PMLR, 15–19 Aug 2021.

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

[62] Markus Wulfmeier, Dushyant Rao, Dominic Zeng Wang, Peter Ondruska, and Ingmar Posner. Large-scale cost function learning for path planning using deep inverse reinforcement learning. *The International Journal of Robotics Research*, 36(10):1073–1087, 2017.

[63] Lin Xiao. On the convergence rates of policy gradient methods. *arXiv preprint arXiv:2201.07443*, 2022.

[64] Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pages 3674–3682. PMLR, 2020.

[65] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5872–5881. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/zhang18n.html.

[66] Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020.

[67] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.

[68] Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in multi-agent markov stochastic games: Stationary points and convergence. *arXiv e-prints*, pages arXiv–2106, 2021.

[69] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML '03: Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.

# Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

# Organization of the appendix



| Notation | Description |
|---|---|
| $s \in \mathcal{S}$ | States of the game |
| $\alpha_i \in \mathcal{A}_i$ | Actions of agent $i \in \mathcal{N}$ |
| $\tau$ | Episode trajectory |
| $T(\tau)$ | Episode stopping time |
| $\zeta$ | Minimum stopping probability |
| $r_{i,t}$ | Realized reward of $i$-th player at time $t$ |
| $\gamma_n$ | Step size at episode $n$ |
| $\varepsilon_n$ | Explicit exploration parameter at episode $n$ |
| $v(\pi_n)$ | Policy gradients at policy $\pi_n$ of episode $n$ |
| $\hat{v}_n$ | Policy gradient proxy at episode $n$. |

**Table 1:** Index of the most common notations used in our paper.

# A Errata and omissions

When preparing the supplementary material of our paper, we noticed a number of typographic errors and omissions in the main paper that could possibly cause confusion. We clarify those below:

- L48: The reference pointers should point to Perkins [44] and Leslie et al. [34].
- L157: (NE) should read (FOS)
- L166: Only the one-way implication is relevant; Proposition 1 was amended accordingly.
- L188: The text should read $\gamma_n = \gamma/(n+m)^p$ for some $\gamma > 0$, $m \geq 0$ and $p \geq 0$.
- L246: The text of Theorem 1 was amended to explicitly include the above clarification.
- L251–L252: the relation "$1 - p < r/2 < p - 1/2$" should read "$1 - p < r < 2p - 1$".
- L250, L283: "$\varepsilon_n = \varepsilon/n^r$" should read "$\varepsilon_n = \varepsilon/(n+m)^r$" and "$\varepsilon_n = \varepsilon/(n+m)^{p/2}$" respectively.
- L331, Eq. (20): "$\mathcal{U}$" should read "$\mathcal{W}$"
- L125, the ~~minimax~~ *mismatch coefficient* can be defined either as $\mathcal{C}_\mathcal{G} \coloneqq \max_{\pi,\pi'\in\Pi}\{\|\tilde{d}_\rho^\pi/\tilde{d}_\rho^{\pi'}\|_\infty\}$ or simpler. $\mathcal{C}_\mathcal{G} \coloneqq \max_{\pi,\in\Pi}\{\frac{1}{\zeta}\|d_\rho^\pi/\rho\|_\infty\}$.

The errata and omissions identified above have all been corrected in the file at hand.

# B Asymptotic convergence to stable Nash policies

Our goal in this appendix is to prove Theorem 1 and Corollary 1, which we restate below for convenience:

**Theorem 1.** *Let $\pi^*$ be a stable Nash policy, and let $\pi_n$ be the sequence of play generated by* (PG) *with step-size $\gamma_n = \gamma/(n+m)^p$, $p \in (1/2, 1]$, and policy gradient estimates such that $p + \ell_b > 1$ and $p - \ell_\sigma > 1/2$ as per* (8)*. Then there exists a neighborhood $\mathcal{U}$ of $\pi^*$ in $\Pi$ such that, for any given $\delta > 0$, we have*

$$\mathbb{P}(\pi_n \text{ converges to } \pi^* \mid \pi_1 \in \mathcal{U}) \geq 1 - \delta \tag{16}$$

*provided that $\gamma$ is small enough* (*or $m$ large enough*) *relative to $\delta$.*

**Corollary 1.** *Suppose that Models 1–3 are run with a step-size of the form $\gamma_n = \gamma/(n+m)^p$, $p > 1/2$, and if applicable, an exploration parameter $\varepsilon_n = \varepsilon/(n+m)^r$ such that $1 - p < r < 2p - 1$. Then:*

- *For Models 1 and 2: the conclusions of Theorem 1 hold as stated.*
- *For Model 3: the conclusions of Theorem 1 hold as long as $p > 2/3$.*

Our proof strategy will comprise the following basic steps:

1. To begin with, we will show that the squared distance
$$D(\pi) = \frac{1}{2}\|\pi - \pi^*\|^2 \tag{B.1}$$
can be seen as a "local Lyapunov function" for (PG) in the sense that it is locally decreasing near $\pi^*$, up to a series of error terms – both zero-mean and non-zero-mean.
2. Due to these errors, the evolution of the iterates $D_n \coloneqq D(\pi_n)$ of $D$ over time may exhibit *significant* jumps: in particular, a single "bad" realization of the noise could carry $\pi_n$ out of the basin of attraction of $\pi^*$, possibly never to return. To exclude this event, our second step will be to show that the aggregation of these errors can be controlled with probability at least $1 - \delta$.
3. Conditioned on the above, we will show that, with probability at least $1 - \delta$, the iterates $D_n$ cannot grow more than a token value. As a result, if (PG) is initialized close to $\pi^*$, it will remain in a neighborhood thereof for all $n$ (again, with probability at least $1 - \delta$).
4. Thanks to this "stochastic Lyapunov stability" result, we employ a series of martingale limit theory arguments to extract a subsequence converging to $\pi^*$.

5. Finally, we show that the increments of $D_n$ are summable; hence, by invoking the Gladyshev's lemma [45, p. 49], we conclude that $D_n$ converges to some (finite) random variable $D_\infty$. Combining this fact with the existence of a convergent subsequence, we obtain the desired conclusion that $\pi_n$ converges to $\pi^*$ with probability at least $1-\delta$.

In the sequel, we make the above precise in a series of intermediate results.

**B.1. Energy inequality.** We begin by establishing a "quasi-Lyapunov" inequality for the iterates $D_n = \|\pi_n - \pi^*\|^2/2$ of (B.1).

**Lemma B.1.** *Let $D_n \coloneqq D(\pi_n)$. Then, for all $n = 1, 2, \dots$, we have*

$$D_{n+1} \leq D_n + \gamma_n \langle v(\pi_n), \pi_n - \pi^* \rangle + \gamma_n \xi_n + \gamma_n \chi_n + \gamma_n^2 \psi_n^2, \tag{B.2}$$

*where the error terms $\xi_n$, $\chi_n$, and $\psi_n$ are given by*

$$\xi_n = \langle U_n, \pi_n - \pi^* \rangle, \quad \chi_n = \|\Pi\| B_n \quad \textit{and} \quad \psi_n^2 = \tfrac{1}{2}\|\hat{v}_n\|^2. \tag{B.3}$$

*with $\|\Pi\| \coloneqq \max_{\pi,\pi' \in \Pi} \|\pi - \pi'\|$.*

*Proof.* By the definition of the iterates of (PG), we have

$$\begin{aligned} D_{n+1} = \frac{1}{2}\|\pi_{n+1} - \pi^*\|^2 &= \frac{1}{2}\|\operatorname{proj}_\Pi(\pi_n + \gamma_n \hat{v}_n) - \operatorname{proj}_\Pi(\pi^*)\|^2 \\ &\leq \frac{1}{2}\|\pi_n + \gamma_n \hat{v}_n - \pi^*\|^2 \\ &= \frac{1}{2}\|\pi_n - \pi^*\|^2 + \gamma_n \langle \hat{v}_n, \pi_n - \pi^* \rangle + \frac{1}{2}\gamma_n^2 \|\hat{v}_n\|^2 \\ &= D_n + \gamma_n \langle v(\pi_n) + U_n + b_n, \pi_n - \pi^* \rangle + \frac{1}{2}\gamma_n^2 \|\hat{v}_n\|^2 \\ &\leq D_n + \gamma_n \langle v(\pi_n), \pi_n - \pi^* \rangle + \gamma_n \xi_n + \gamma_n \chi_n + \gamma_n^2 \psi_n^2 \end{aligned} \tag{B.4}$$

where we used the Cauchy-Schwarz inequality to bound the bias term as $\langle b_n, \pi_n - \pi^* \rangle \leq \|b_n\| \cdot \|\pi_n - \pi^*\| \leq \|\Pi\| B_n = \chi_n$. ∎

**B.2. Error control and stability.** The second major step in our proof (and the most challenging one from a technical standpoint) is to establish a suitable measure of control over the error increments in (B.1), with the aim of showing that the process $\pi_n$ never leaves a neighborhood of $\pi^*$.

To make this idea precise, let $\mathcal{B} = \{\pi \in \Pi : \|\pi - \pi^*\| \leq r\}$ be a ball of radius $r$ based on $\pi^*$ in $\Pi$ so that $\langle v(\pi), \pi - \pi^* \rangle < 0$ for all $\pi \in \mathcal{B} \setminus \{\pi^*\}$ (without loss of generality, we can assume that $\mathcal{B}$ is maximal in that regard). We will then examine the event that the aggregation of the error terms in (B.1) is not sufficient to drive $\pi_n$ to escape from $\mathcal{B}$.

To that end, we will begin by aggregating the errors in (B.1) as

$$M_n = \sum_{k=1}^n \gamma_k \xi_k \quad \text{and} \quad S_n = \sum_{k=1}^n [\gamma_k \chi_k + \gamma_k^2 \psi_k^2]. \tag{B.5}$$

Since $\mathbb{E}[\xi_n \,|\, \mathcal{F}_n] = 0$, we have $\mathbb{E}[M_n \,|\, \mathcal{F}_n] = M_{n-1}$, so $M_n$ is a martingale; likewise, $\mathbb{E}[S_n \,|\, \mathcal{F}_n] \geq S_{n-1}$, so $S_n$ is a submartingale. Then, using a technique of Hsieh et al. [26] that builds on an earlier idea by Mertikopoulos and Zhou [37], we will also consider the "mean square" error process

$$R_n = M_n^2 + S_n, \tag{B.6}$$

and the associated indicator events

$$\mathcal{E}_n = \{\pi_k \in \mathcal{B} \text{ for all } k = 1, 2, \dots, n\} \quad \text{and} \quad H_n = \{R_k \leq a \text{ for all } k = 1, 2, \dots, n\}, \tag{B.7a}$$

where, with a fair amount of hindsight, the error tolerance level $a > 0$ is such that $2a + \sqrt{a} < r$, and we are employing the convention $\mathcal{E}_0 = H_0 = \Omega$ (since every statement is true for the elements of the empty set). We will then assume that $\pi_1$ is initialized in a ball of radius $\sqrt{2a}$ centered at $\pi^*$, viz.

$$\mathcal{U} = \{\pi \in \Pi : D(\pi) \leq a\} = \{\pi \in \Pi : \|\pi - \pi^*\|^2/2 \leq a\}. \tag{B.8}$$

With all this in hand, the key to showing that $\pi_n$ remains close to $\pi^*$ with high probability is the following conditional estimate:

**Lemma B.2.** *Let $\pi_n$ be the sequence of play generated by* (PG) *initialized at $\pi_1 \in \mathcal{U}$. We then have:*

1. *$\mathcal{E}_{n+1} \subseteq \mathcal{E}_n$ and $H_{n+1} \subseteq H_n$ for all $n = 1, 2, \ldots$*
2. *$H_{n-1} \subseteq \mathcal{E}_n$ for all $n = 1, 2, \ldots$*
3. *Consider the "bad realization" event*

$$\tilde{H}_n \coloneqq H_{n-1} \setminus H_n = \{R_k \leq a \text{ for } k = 1, 2, \ldots, n-1 \text{ and } R_n > a\}, \tag{B.9}$$

*and let $\tilde{R}_n = R_n \mathbb{1}_{H_{n-1}}$ be the cumulative error subject to the noise being "small". Then we have:*

$$\mathbb{E}[\tilde{R}_n] \leq \mathbb{E}[\tilde{R}_{n-1}] + \gamma_n \|\Pi\| B_n + \gamma_n^2 \|\Pi\|^2 \sigma_n^2 + \tfrac{3}{2}\gamma_n^2 (G^2 + B_n^2 + \sigma_n^2) - a\,\mathbb{P}(\tilde{H}_{n-1}), \tag{B.10}$$

*where, by convention, $\tilde{H}_0 = \varnothing$ and $\tilde{R}_0 = 0$.*

*Remark.* In the above (and what follows), the notation $\mathbb{1}_A$ is used to indicate the logical indicator of an event $A \subseteq \Omega$, i.e., $\mathbb{1}_A(\omega) = 1$ if $\omega \in A$ and $\mathbb{1}_A(\omega) = 0$ otherwise.

The proof of Lemma B.2 is quite technical, so we first proceed to derive an important stability result based on this estimate.

**Proposition B.1.** *Fix some confidence threshold $\delta > 0$ and let $\pi_n$ be the sequence of play generated by* (PG) *with step-size and policy gradient estimates as per Theorem 1. We then have:*

$$\mathbb{P}(H_n \mid \pi_1 \in \mathcal{U}) \geq 1 - \delta \quad \text{for all } n = 1, 2, \ldots \tag{B.11}$$

*provided that $\gamma$ is small enough (or $m$ large enough) relative to $\delta$.*

*Proof.* We begin by bounding the probability of the "bad realization" event $\tilde{H}_n = H_{n-1} \setminus H_n$. Indeed, if $\pi_1 \in \mathcal{U}$, we have:

$$\mathbb{P}(\tilde{H}_n) = \mathbb{P}(H_{n-1} \setminus H_n) = \mathbb{E}[\mathbb{1}_{H_{n-1}} \times \mathbb{1}\{R_n > a\}] \leq \mathbb{E}[\mathbb{1}_{H_{n-1}} \times (R_n/a)] = \mathbb{E}[\tilde{R}_n]/a \tag{B.12}$$

where, in the penultimate step, we used the fact that $R_n \geq 0$ (so $\mathbb{1}\{R_n > a\} \leq R_n/a$). Telescoping (B.10) then yields

$$\mathbb{E}[\tilde{R}_n] \leq \mathbb{E}[\tilde{R}_0] + \|\Pi\| \sum_{k=1}^{n} \gamma_k B_k + \sum_{k=1}^{n} \gamma_k^2 \varrho_k^2 - a \sum_{k=1}^{n} \mathbb{P}(\tilde{H}_{k-1}) \tag{B.13}$$

where we set

$$\varrho_n^2 = \|\Pi\|^2 \sigma_n^2 + \tfrac{3}{2}(G^2 + B_n^2 + \sigma_n^2). \tag{B.14}$$

Hence, combining (B.12) and (B.13) and invoking our stated assumptions for $\gamma_n$, $B_n$ and $\sigma_n$, we get

$$\sum_{k=1}^{n} \mathbb{P}(\tilde{H}_k) \leq \frac{1}{a} \sum_{k=1}^{n} [\gamma_k B_k \|\Pi\| + \gamma_k^2 \varrho_k^2] \leq \frac{C}{a} \tag{B.15}$$

for some $C \equiv C(\gamma, m) > 0$ with $\lim_{\gamma \to 0^+} C(\gamma, m) = \lim_{m \to \infty} C(\gamma, m) = 0$.

Now, by choosing $\gamma$ sufficiently small (or $m$ sufficiently large), we can ensure that $C/a < \delta$; thus, given that the events $\tilde{H}_k$ are disjoint for all $k = 1, 2, \ldots$, we get $\mathbb{P}(\bigcup_{k=1}^n \tilde{H}_k) = \sum_{k=1}^n \mathbb{P}(\tilde{H}_k) \leq \delta$. In turn, this implies that $\mathbb{P}(H_n) = \mathbb{P}(\tilde{H}_1^c \cap \cdots \cap \tilde{H}_n^c) \geq 1 - \delta$, and our assertion follows. ■

We conclude this appendix with the proof of our technical result on the events $\mathcal{E}_n$ and $H_n$:

*Proof of Lemma B.2.* The first claim of the lemma is obvious. For the second, we proceed inductively:

1. For the base case $n = 1$, we have $\mathcal{E}_1 = \{\pi_1 \in \mathcal{B}\} \supseteq \{\pi_1 \in \mathcal{U}\} = \Omega$ (recall that $\pi_1$ is initialized in $\mathcal{U} \subseteq \mathcal{B}$). Since $H_0 = \Omega$, our claim follows.

2. Inductively, assume that $H_{n-1} \subseteq \mathcal{E}_n$ for some $n \geq 1$. To show that $H_n \subseteq \mathcal{E}_{n+1}$, suppose that $R_k \leq a$ for all $k = 1, 2, \dots, n$. Since $H_n \subseteq H_{n-1}$, this implies that $\mathcal{E}_n$ also occurs, i.e., $\pi_k \in \mathcal{B}$ for all $k = 1, 2, \dots, n$; as such, it suffices to show that $\pi_{n+1} \in \mathcal{B}$. To do so, given that $\pi_k \in \mathcal{U} \subseteq \mathcal{B}$ for all $k = 1, 2, \dots n$, telescoping the bound (B.2) over $k = 1, 2, \dots, n$ gives

$$D_{k+1} \leq D_k + \gamma_k \xi_k + \gamma_k \chi_k + \gamma_k^2 \psi_k^2, \quad \text{for all } k = 1, 2, \dots n, \tag{B.16}$$

and hence, after telescoping over $k = 1, 2, \dots, n$, we get

$$D_{n+1} \leq D_1 + M_n + S_n \leq D_1 + \sqrt{R_n} + R_n \leq a + \sqrt{a} + a = 2a + \sqrt{a}. \tag{B.17}$$

We conclude that $D(\pi_{n+1}) \leq 2a + \sqrt{a}$, i.e., $\pi_{n+1} \in \mathcal{B}$, as required for the induction.

For our third claim, note first that

$$\begin{aligned} R_n &= (M_{n-1} + \gamma_n \xi_n)^2 + S_{n-1} + \gamma_n \chi_n + \gamma_n^2 \psi_n^2 \\ &= R_{n-1} + 2\gamma_n \xi_n M_{n-1} + \gamma_n^2 \xi_n^2 + \gamma_n \chi_n + \gamma_n^2 \psi_n^2, \end{aligned} \tag{B.18}$$

so, after taking expectations, we get

$$\mathbb{E}[R_n \,|\, \mathcal{F}_n] = R_{n-1} + 2M_{n-1}\gamma_n \,\mathbb{E}[\xi_n \,|\, \mathcal{F}_n] + \mathbb{E}[\gamma_n^2 \xi_n^2 + \gamma_n \chi_n + \gamma_n^2 \psi_n^2 \,|\, \mathcal{F}_n] \geq R_{n-1}, \tag{B.19}$$

i.e., $R_n$ is a submartingale. To proceed, let $\tilde{R}_n = R_n \mathbb{1}_{H_{n-1}}$ so

$$\begin{aligned} \tilde{R}_n = R_n \mathbb{1}_{H_{n-1}} &= R_{n-1} \mathbb{1}_{H_{n-1}} + (R_n - R_{n-1}) \mathbb{1}_{H_{n-1}} \\ &= R_{n-1} \mathbb{1}_{H_{n-2}} - R_{n-1} \mathbb{1}_{\tilde{H}_{n-1}} + (R_n - R_{n-1}) \mathbb{1}_{H_{n-1}}, \\ &= \tilde{R}_{n-1} + (R_n - R_{n-1}) \mathbb{1}_{H_{n-1}} - R_{n-1} \mathbb{1}_{\tilde{H}_{n-1}}, \end{aligned} \tag{B.20}$$

where we used the fact that $H_{n-1} = H_{n-2} \setminus \tilde{H}_{n-1}$ so $\mathbb{1}_{H_{n-1}} = \mathbb{1}_{H_{n-2}} - \mathbb{1}_{\tilde{H}_{n-1}}$ (since $H_{n-1} \subseteq H_{n-2}$). Then, (B.18) yields

$$R_n - R_{n-1} = 2M_{n-1}\gamma_n \xi_n + \gamma_n^2 \xi_n^2 + \gamma_n \chi_n + \gamma_n^2 \psi_n^2 \tag{B.21}$$

and hence, given that $H_{n-1}$ is $\mathcal{F}_n$-measurable, we get:

$$\begin{aligned} \mathbb{E}[(R_n - R_{n-1}) \mathbb{1}_{H_{n-1}}] &= 2\,\mathbb{E}[\gamma_n M_{n-1} \xi_n \mathbb{1}_{H_{n-1}}] && \text{(B.22a)} \\ &\quad + \mathbb{E}[\gamma_n^2 \xi_n^2 \mathbb{1}_{H_{n-1}}] && \text{(B.22b)} \\ &\quad + \mathbb{E}[(\gamma_n \chi_n + \gamma_n^2 \psi_n^2) \mathbb{1}_{H_{n-1}}]. && \text{(B.22c)} \end{aligned}$$

However, since $H_{n-1}$ and $M_{n-1}$ are both $\mathcal{F}_n$-measurable, we have the following estimates:

1. For the noise term in (B.22a), we have:

$$\mathbb{E}[M_{n-1}\xi_n \mathbb{1}_{H_{n-1}}] = \mathbb{E}[M_{n-1} \mathbb{1}_{H_{n-1}} \mathbb{E}[\xi_n \,|\, \mathcal{F}_n]] = 0. \tag{B.23}$$

2. The term (B.22b) is where the reduction to $H_{n-1}$ kicks in; indeed, we have:

$$\begin{aligned} \mathbb{E}[\xi_n^2 \mathbb{1}_{H_{n-1}}] &= \mathbb{E}[\mathbb{1}_{H_{n-1}} \mathbb{E}[|\langle \pi_n - \pi^*, U_n \rangle|^2 \,|\, \mathcal{F}_n]] \\ &\leq \mathbb{E}[\mathbb{1}_{H_{n-1}} \|\pi_n - \pi^*\|^2 \,\mathbb{E}[\|U_n\|^2 \,|\, \mathcal{F}_n]] && \textit{\# by Cauchy–Schwarz} \\ &\leq \mathbb{E}[\mathbb{1}_{\mathcal{E}_n} \|\pi_n - \pi^*\|^2 \,\mathbb{E}[\|U_n\|^2 \,|\, \mathcal{F}_n]] && \textit{\# because } H_{n-1} \subseteq \mathcal{E}_n \\ &\leq \|\Pi\|^2 \sigma_n^2. \end{aligned} \tag{B.24}$$

3. Finally, for the term (B.22c), we have:

$$\mathbb{E}[\psi_n^2 \mathbb{1}_{H_{n-1}}] \leq \tfrac{3}{2}[G^2 + B_n^2 + \sigma_n^2] \tag{B.25}$$

where we used the bound $\|v(\pi)\| \leq G$. Likewise, $\chi_n \mathbb{1}_{H_{n-1}} \leq \|\Pi\| B_n$, so

$$\text{(B.22c)} \leq \gamma_n \|\Pi\| B_n + \tfrac{3}{2}\gamma_n^2 (G^2 + B_n^2 + \sigma_n^2) \tag{B.26}$$

Thus, putting together all of the above, we obtain:

$$\mathbb{E}[(R_n - R_{n-1}) \mathbb{1}_{H_{n-1}}] \leq \gamma_n \|\Pi\| B_n + \gamma_n^2 \|\Pi\|^2 \sigma_n^2 + \tfrac{3}{2}\gamma_n^2 (G^2 + B_n^2 + \sigma_n^2) \tag{B.27}$$

Going back to (B.20), we have $R_{n-1} > a$ if $\tilde{H}_{n-1}$ occurs, so the last term becomes

$$\mathbb{E}[R_{n-1} \mathbb{1}_{\tilde{H}_{n-1}}] \geq a\,\mathbb{E}[\mathbb{1}_{\tilde{H}_{n-1}}] = a\,\mathbb{P}(\tilde{H}_{n-1}). \tag{B.28}$$

Our claim then follows by combining Eqs. (B.20), (B.25), (B.26) and (B.28). ■

**B.3. Extraction of a convergent subsequence.** Our next step is to show that any realization $\pi_n$ of (PG) that is contained in $\mathcal{B}$ admits a subsequence $\pi_{n_k}$ converging to $\pi^*$.

**Proposition B.2.** *Let $\pi^*$ be a stable Nash policy, and let $\pi_n$ be the sequence of play generated by* (PG) *with step-size and policy gradient estimates such that $p + \ell_b > 1$ and $p - \ell_\sigma > 1/2$ as per* (8)*. Then $\pi_n$ admits a subsequence $\pi_{n_k}$ that converges to $\pi^*$ with probability 1 on the event $\mathcal{E} = \bigcap_n \mathcal{E}_n = \{\pi_n \in \mathcal{B}$ for all $n = 1, 2, \dots\}$.*

*Proof.* Let $\mathcal{Q} = \{\pi_n \in \mathcal{B} \text{ for all } n\} \cap \{\liminf_n \|\pi_n - \pi^*\| > 0\}$ denote the event that $\pi_n$ is contained in $\mathcal{B}$ but the sequence $\pi_n$ does not admit a subsequence converging to $\pi^*$. We will show that $\mathbb{P}(\mathcal{Q}) = 0$.

Indeed, assume ad absurdum that $\mathbb{P}(\mathcal{Q}) > 0$. Hence, with probability 1 on $\mathcal{Q}$, there exists some positive constant $c > 0$ (again, possibly random) such that $\langle v(\pi_n), \pi_n - \pi^* \rangle \leq -c < 0$ for all $n$. Thus, going back to (B.1), we get

$$D_{n+1} \leq D_n - \gamma_n c + \gamma_n \xi_n + \gamma_n \chi_n + \gamma_n^2 \psi_n^2, \tag{B.29}$$

so if we let $\tau_n = \sum_{k=1}^n \gamma_k$ and telescope the above, we obtain the bound

$$D_{n+1} \leq D_1 - \tau_n \left[ c - \frac{M_n}{\tau_n} - \frac{S_n}{\tau_n} \right] \tag{B.30}$$

with $\xi_n, \chi_n$ and $\psi_n$ given by (B.3), and $M_n = \sum_{k=1}^n \gamma_k \xi_k$, $S_n = \sum_{k=1}^n [\gamma_k \chi_k + \gamma_k^2 \psi_k^2]$ defined as in (D.10). Also, (7) readily gives

$$\sum_{n=1}^\infty \mathbb{E}[\gamma_n^2 \xi_n^2 \,|\, \mathcal{F}_n] \leq \sum_{n=1}^\infty \gamma_n^2 \, \mathbb{E}[\|\pi_n - \pi^*\|^2 \|U_n\|^2 \,|\, \mathcal{F}_n] \leq \|\Pi\|^2 \sum_{n=1}^\infty \gamma_n^2 \sigma_n^2 < \infty \tag{B.31}$$

so, by the strong law of large numbers for martingale difference sequences [23, Theorem 2.18], we conclude that $M_n/\tau_n$ converges to 0 with probability 1. In a similar vein, for the submartingale $S_n$ we have

$$\mathbb{E}[S_n] = \sum_{k=1}^n \gamma_k \chi_k \sum_{k=1}^n \gamma_k^2 \, \mathbb{E}[\psi_k^2] \leq \|\Pi\| \sum_{k=1}^n \gamma_k B_k + \frac{3}{2} \sum_{k=1}^n \gamma_k^2 [G^2 + B_k^2 + \sigma_k^2], \tag{B.32}$$

so, by (7) and the stated conditions for the method's step-size and bias/noise parameters, it follows that $S_n$ is bounded in $L^1$. Therefore, by Doob's submartingale convergence theorem [23, Theorem 2.5], we further deduce that $S_n$ converges with probability 1 to some (finite) random variable $S_\infty$.

Going back to (B.30) and letting $n \to \infty$, the above shows that $D_n \to -\infty$ with probability 1 on $\mathcal{Q}$. Since $D$ is nonnegative by construction and $\mathbb{P}(\mathcal{Q}) > 0$ by assumption, we obtain a contradiction and our proof is complete. ∎

**B.4. Convergence of the energy values.** Our last auxiliary result concerns the convergence of the values of the dual energy function $D$. We encode this as follows.

**Proposition B.3.** *If* (PG) *is run with assumptions as in Proposition B.1, there exists a finite random variable $D_\infty$ such that*

$$\mathbb{P}(D_n \to D_\infty \text{ as } n \to \infty \,|\, \pi_n \in \mathcal{B} \text{ for all } n) = 1. \tag{B.33}$$

*Proof.* Let $\mathcal{E}_n = \{\pi_k \in \mathcal{B} \text{ for all } k = 1, 2, \dots, n\}$ be defined as in (B.7), and let $\tilde{D}_n = \mathbb{1}_{\mathcal{E}_n} D_n$. Then, by the energy inequality (B.2) and the fact that $\mathcal{E}_{n+1} \subseteq \mathcal{E}_n$, we get

$$\begin{aligned} \tilde{D}_{n+1} = \mathbb{1}_{\mathcal{E}_{n+1}} D_{n+1} &\leq \mathbb{1}_{\mathcal{E}_n} D_{n+1} \\ &\leq \mathbb{1}_{\mathcal{E}_n} D_n + \mathbb{1}_{\mathcal{E}_n} \gamma_n \langle v(\pi_n), \pi_n - \pi^* \rangle + (\gamma_n \xi_n + \gamma_n \chi_n + \gamma_n^2 \psi_n^2) \mathbb{1}_{\mathcal{E}_n} \\ &\leq \tilde{D}_n + \gamma_n \mathbb{1}_{\mathcal{E}_n} \xi_n + (\gamma_n \chi_n + \gamma_n^2 \psi_n^2) \mathbb{1}_{\mathcal{E}_n}, \end{aligned} \tag{B.34}$$

where we used the fact that $\langle v(\pi_k), \pi_k - \pi^* \rangle \leq 0$ for all $k = 1, 2, \dots, n$ if $\mathcal{E}_n$ occurs. Since $\mathcal{E}_n$ is $\mathcal{F}_n$-measurable, conditioning on $\mathcal{F}_n$ and taking expectations yields

$$\mathbb{E}[\tilde{D}_{n+1} \,|\, \mathcal{F}_n] \leq \tilde{D}_n + \gamma_n \mathbb{1}_{\mathcal{E}_n} \mathbb{E}[\xi_n \,|\, \mathcal{F}_n] + \mathbb{1}_{\mathcal{E}_n} \gamma_n \chi_n + \mathbb{1}_{\mathcal{E}_n} \mathbb{E}[\gamma_n^2 \psi_n^2 \,|\, \mathcal{F}_n]$$

$$
\begin{aligned}
&\leq \tilde{D}_n + \gamma_n \|\Pi\| B_n + \gamma_n \chi_n + \mathbb{E}[\gamma_n^2 \psi_n^2 \,|\, \mathcal{F}_n] \\
&\leq \tilde{D}_n + \gamma_n \|\Pi\| B_n + \frac{3}{2}[G^2 + B_n^2 + \sigma_n^2]. \qquad \text{(B.35)}
\end{aligned}
$$

By our step-size assumptions, we have $\sum_n \gamma_n^2(1 + B_n^2 + \sigma_n^2) < \infty$ and $\sum_n \gamma_n B_n < \infty$, which means that $\tilde{D}_n$ is an almost supermartingale with almost surely summable increments, i.e.,

$$
\sum_{n=1}^{\infty} \Big[\mathbb{E}[\tilde{D}_{n+1} \,|\, \mathcal{F}_n] - \tilde{D}_n\Big] < \infty \quad \text{with probability 1} \tag{B.36}
$$

Therefore, by Gladyshev's lemma [45, p. 49], we conclude that $\tilde{D}_n$ converges almost surely to some (finite) random variable $D_\infty$. Since $\mathbb{1}_{\mathcal{E}_n} = 1$ for all $n$ if and only if $\pi_n \in \mathcal{B}$ for all $n$, we conclude that $\mathbb{P}(D_n \text{ converges} \,\big|\, \pi_n \in \mathcal{B} \text{ for all } n) = \mathbb{P}(\tilde{D}_n \text{ converges}) = 1$, and our claim follows. ∎

**B.5. Putting everything together.** We are now in a position to prove Theorem 1 and Corollary 1.

*Proof of Theorem 1.* Let $\mathcal{E} = \bigcap_n \mathcal{E}_n = \{\pi_n \in \mathcal{B} \text{ for all } n\}$ denote the event that $\pi_n$ lies in $\mathcal{B}$ for all $n$. By Proposition B.1, if $\pi_1$ is initialized within the neighborhood $\mathcal{U}$ defined in (B.8), we have $\mathbb{P}(\mathcal{E} \,|\, \pi_1 \in \mathcal{U}) \geq 1 - a$, noting also that the neighborhood $\mathcal{U}$ is independent of the required confidence level $a$. Then, by Propositions B.2 and B.3, it follows that *a)* $\liminf_n \|\pi_n - \pi^*\| = 0$; and *b)* $D_n$ converges, both events occurring with probability 1 on the set $\mathcal{E} \cap \{\pi_1 \in \mathcal{U}\}$. We thus conclude that $\lim_{n\to\infty} D_n = 0$ and hence

$$
\begin{aligned}
\mathbb{P}(\pi_n \to \pi^* \,|\, \pi_1 \in \mathcal{U}) &\geq \mathbb{P}(\mathcal{E} \cap \{\pi_n \to \pi^*\} \,|\, \pi_1 \in \mathcal{U}) \\
&= \mathbb{P}(\pi_n \to \pi^* \,|\, \pi_1 \in \mathcal{U}, \mathcal{E}) \times \mathbb{P}(\mathcal{E} \,|\, \pi_1 \in \mathcal{U}) \geq 1 - \delta,
\end{aligned}
$$

and our proof is complete. ∎

*Proof of Corollary 1.* For Models 1 and 2, taking $\ell_b = \infty, \ell_\sigma = 0$, we obtain $p > 1/2$. Since we have that $\sum_{n=1}^{\infty} \gamma_n = \infty$, we get that $p \leq 1$, i.e., $p \in (1/2, 1]$.

For Model 3, we have that $B_n = \mathcal{O}(\varepsilon_n)$ and $\sigma_n = \mathcal{O}(1/\sqrt{\varepsilon_n})$, i.e., $\ell_b = r$ and $\ell_\sigma = r/2$. Now, since $p \leq 1$, $p + \ell_b > 1$ and $p - \ell_\sigma > 1/2$, we obtain that $p \in (2/3, 1]$ and $(1 - p)/2 < r/2 < p - 1/2$. ∎

# C Rate of convergence to second-order stationary policies

We now proceed with the proof of Theorem 2, which we again restate below for convenience:

**Theorem 2.** *Let $\pi^*$ be a Nash policy such that (SOS) holds on some open set $\mathcal{B}$ containing $\pi^*$, and let $\pi_n$ be the sequence of play generated by (PG) with step-size $\gamma_n = \gamma/(n + m)^p$, $p \in (1/2, 1]$, and policy gradient estimates such that $p + \ell_b > 1$ and $p - \ell_\sigma > 1/2$ as per (8). Then:*

1. *There exists a neighborhood $\mathcal{U}$ of $\pi^*$ in $\Pi$ such that, for any confidence level $\delta > 0$, the event*

$$
\mathcal{E} = \{\pi_n \in \mathcal{B} \text{ for all } n = 1, 2, \dots\} \tag{17}
$$

*occurs with probability $\mathbb{P}(\mathcal{E} \,|\, \pi_1 \in \mathcal{U}) \geq 1 - \delta$ if $m$ is large enough relative to $\delta$.*

2. *The sequence $\pi_n$ converges to $\pi^*$ with probability 1 on $\mathcal{E}$; in particular, we have*

$$
\mathbb{P}(\pi_n \text{ converges to } \pi^* \,|\, \pi_1 \in \mathcal{U}) \geq 1 - \delta \tag{18}
$$

*if $m$ is large relative to $\delta$. Moreover, conditioned on $\mathcal{E}$ and taking $q = \min\{\ell_b, p - 2\ell_\sigma\}$, we have*

$$
\mathbb{E}[\|\pi_n - \pi^*\|^2 \,|\, \mathcal{E}] = \begin{cases} \mathcal{O}(1/n^{2\mu\gamma}) & \text{if } p = 1 \text{ and } 2\mu\gamma < q, \\ \mathcal{O}(1/n^q) & \text{otherwise.} \end{cases} \tag{19}
$$

*Proof.* We will follow an approach similar to Theorem 1 for the first part of the theorem. More precisely, let $\mathcal{B} = \{\pi \in \Pi : \|\pi - \pi^*\| \leq r\}$ be a ball of radius $r$ centered at $\pi^*$ in $\Pi$ such that (SOS) holds for all $\pi \in \mathcal{B}$. Then, for all $\pi \in \mathcal{B}\backslash\{\pi^*\}$, we have $\langle v(\pi), \pi - \pi^*\rangle \leq -\mu\|\pi - \pi^*\| < 0$ by Proposition 1. Hence, defining the events $\mathcal{E}_n$ and $H_n$ as in Eq. (B.7), and assuming that $\pi_1$ is initialized in a ball of radius $\sqrt{2a}$ centered at $\pi^*$, viz.

$$\mathcal{U} = \{\pi \in \Pi : D(\pi) \leq a\} = \{\pi \in \Pi : \|\pi - \pi^*\|^2/2 \leq a\}. \tag{C.1}$$

then, by Lemma B.2 and Proposition B.1, we readily obtain that

$$\mathbb{P}(H_n \mid \pi_1 \in \mathcal{U}) \geq 1 - \delta \quad \text{for all } n = 1, 2, \ldots \tag{C.2}$$

which implies that

$$\mathbb{P}(\mathcal{E} \mid \pi_1 \in \mathcal{U}) \geq 1 - \delta \tag{C.3}$$

if $m$ is large enough relative to $\delta$.

For the second part, constraining Eq. (B.2) on the event $\mathcal{E}_n$, we get:

$$\begin{aligned} D_{n+1} \mathbb{1}_{\mathcal{E}_n} &\leq D_n \mathbb{1}_{\mathcal{E}_n} + \gamma_n \langle v(\pi_n), \pi_n - \pi^*\rangle \mathbb{1}_{\mathcal{E}_n} + \mathbb{1}_{\mathcal{E}_n}\Big(\gamma_n\xi_n + \gamma_n\chi_n + \gamma_n^2\psi_n^2\Big) \\ &\leq (1 - 2\mu\gamma_n) D_n \mathbb{1}_{\mathcal{E}_n} + \mathbb{1}_{\mathcal{E}_n}\Big(\gamma_n\xi_n + \gamma_n\chi_n + \gamma_n^2\psi_n^2\Big) \end{aligned} \tag{C.4}$$

where the last inequality comes from (SOS). Therefore, taking expectations, we obtain:

$$\begin{aligned} \mathbb{E}[D_{n+1} \mathbb{1}_{\mathcal{E}_n}] &\leq (1 - 2\mu\gamma_n)\,\mathbb{E}[D_n \mathbb{1}_{\mathcal{E}_n}] + \mathbb{E}\Big[\mathbb{1}_{\mathcal{E}_n}\Big(\gamma_n\xi_n + \gamma_n\chi_n + \gamma_n^2\psi_n^2\Big)\Big] \\ &\leq (1 - 2\mu\gamma_n)\,\mathbb{E}[D_n \mathbb{1}_{\mathcal{E}_n}] + \gamma_n\,\mathbb{E}[\mathbb{1}_{\mathcal{E}_n}\xi_n] + \gamma_n\,\mathbb{E}[\mathbb{1}_{\mathcal{E}_n}\chi_n] + \gamma_n^2\,\mathbb{E}[\mathbb{1}_{\mathcal{E}_n}\psi_n^2] \\ &= (1 - 2\mu\gamma_n)\,\mathbb{E}[D_n \mathbb{1}_{\mathcal{E}_n}] + \gamma_n\,\mathbb{E}[\mathbb{1}_{\mathcal{E}_n}\chi_n] + \gamma_n^2\,\mathbb{E}[\mathbb{1}_{\mathcal{E}_n}\psi_n^2] \\ &\leq (1 - 2\mu\gamma_n)\,\mathbb{E}[D_n \mathbb{1}_{\mathcal{E}_n}] + \|\Pi\|\,\mathbb{P}(\mathcal{E}_n)\gamma_n B_n + \mathbb{P}(\mathcal{E}_n)\Big(G\gamma_n^2 + 3\gamma_n^2\sigma_n^2 + 3\gamma_n^2 B_n^2\Big) \end{aligned} \tag{C.5}$$

where the equality in the third line comes from the fact that

$$\mathbb{E}[\mathbb{1}_{\mathcal{E}_n}\xi_n] = \mathbb{E}[\mathbb{E}[\xi_n \mathbb{1}_{\mathcal{E}_n} \mid \mathcal{F}_n]] = \mathbb{E}[\mathbb{1}_{\mathcal{E}_n}\,\mathbb{E}[\xi_n \mid \mathcal{F}_n]] = 0. \tag{C.6}$$

Now, since $\mathbb{1}_{\mathcal{E}_{n+1}} \leq \mathbb{1}_{\mathcal{E}_n}$, we further have

$$\mathbb{E}[D_{n+1} \mathbb{1}_{\mathcal{E}_{n+1}}] \leq \mathbb{E}[D_{n+1} \mathbb{1}_{\mathcal{E}_n}] \tag{C.7}$$

and hence, setting $\bar{D}_n := \mathbb{E}[D_n \mathbb{1}_{\mathcal{E}_n}]$, we get

$$\begin{aligned} \bar{D}_{n+1} &\leq (1 - 2\mu\gamma_n)\bar{D}_n + \|\Pi\|\,\mathbb{P}(\mathcal{E}_n)\gamma_n B_n + \mathbb{P}(\mathcal{E}_n)\Big(G\gamma_n^2 + 3\gamma_n^2\sigma_n^2 + 3\gamma_n^2 B_n^2\Big) \\ &\leq (1 - 2\mu\gamma_n)\bar{D}_n + \|\Pi\|\gamma_n B_n + G\gamma_n^2 + 3\gamma_n^2\sigma_n^2 + 3\gamma_n^2 B_n^2. \end{aligned} \tag{C.8}$$

Therefore, taking $\gamma_n, B_n, \sigma_n$ as per the statement of the theorem and noting that the terms $\gamma_n^2$ and $\gamma_n^2 B_n^2$ are respectively dominated by the terms $\gamma_n^2\sigma_n^2$ and $\gamma_n B_n$, we obtain

$$\begin{aligned} \bar{D}_{n+1} &\leq \left(1 - \frac{2\mu\gamma}{(n+m)^p}\right)\bar{D}_n + \frac{C_1}{(n+m)^{p+\ell_b}} + \frac{C_2}{(n+m)^{2p-2\ell_\sigma}} \\ &\leq \left(1 - \frac{2\mu\gamma}{(n+m)^p}\right)\bar{D}_n + \frac{C_1 + C_2}{(n+m)^{p+q}} \end{aligned} \tag{C.9}$$

for some $C_1, C_2 > 0$, where $q = \min\{\ell_b, p - 2\ell_\sigma\}$, as per the theorem's statement. Therefore, by a straightforward modification of Chung's lemma [14, Lemmas 2&3], [45, p. 45], we get

$$\bar{D}_n = \begin{cases} \mathcal{O}(1/n^{2\mu\gamma}) & \text{if } p = 1 \text{ and } 2\mu\gamma < q, \\ \mathcal{O}(1/n^q) & \text{otherwise.} \end{cases} \tag{C.10}$$

Accordingly, letting $n \to \infty$ and recalling that $\mathbb{E}[D_n \mathbb{1}_{\mathcal{E}}] \leq \mathbb{E}[D_n \mathbb{1}_{\mathcal{E}_n}] = \bar{D}_n$

$$\lim_{n\to\infty} \mathbb{E}[D_n \mathbb{1}_{\mathcal{E}}] = 0. \tag{C.11}$$

Then, by Fatou's lemma [21], we obtain

$$0 \leq \mathbb{E}[\liminf_{n\to\infty} D_n \mathbb{1}_{\mathcal{E}}] \leq \liminf_{n\to\infty} \mathbb{E}[D_n \mathbb{1}_{\mathcal{E}}] = 0, \tag{C.12}$$

which readily shows that $\mathbb{E}[\liminf_{n\to\infty} D_n \mathbb{1}_{\mathcal{E}}] = 0$. Finally, since $\liminf_{n\to\infty} D_n \mathbb{1}_{\mathcal{E}} \geq 0$ (a.s.) and $\mathbb{E}[\liminf_{n\to\infty} D_n \mathbb{1}_{\mathcal{E}}] = 0$, we get that

$$\liminf_{n\to\infty} D_n \mathbb{1}_{\mathcal{E}} = 0 \quad \text{with probability 1.} \tag{C.13}$$

Therefore, there exists a subsequence $D_{n_k}$ that converges to 0 with probability 1 on the event $\mathcal{E}$, i.e., $\pi_{n_k}$ converges to $\pi^*$. Hence, invoking Proposition B.3, we further deduce that $D_n$ converges to some $D_\infty$ with probability 1 on $\mathcal{E}$, and thus, we obtain that $\lim_{n\to\infty} D_n = 0$ on $\mathcal{E}$. We thus get

$$\begin{aligned}\mathbb{P}(\pi_n \to \pi^* \mid \pi_1 \in \mathcal{U}) &\geq \mathbb{P}(\mathcal{E} \cap \{\pi_n \to \pi^*\} \mid \pi_1 \in \mathcal{U}) \\ &= \mathbb{P}(\pi_n \to \pi^* \mid \pi_1 \in \mathcal{U}, \mathcal{E}) \times \mathbb{P}(\mathcal{E} \mid \pi_1 \in \mathcal{U}) \geq 1 - \delta,\end{aligned} \tag{C.14}$$

as claimed.

For the last part of the theorem, note that

$$\begin{aligned}\bar{D}_n = \mathbb{E}[D_n \mathbb{1}_{\mathcal{E}_n}] \geq \mathbb{E}[D_n \mathbb{1}_{\mathcal{E}}] &= \mathbb{E}[\mathbb{E}[D_n \mid \sigma(\mathcal{E})] \mathbb{1}_{\mathcal{E}}] \\ &= \mathbb{E}[\mathbb{E}[D_n \mid \mathcal{E}] \mathbb{1}_{\mathcal{E}}] \\ &= \mathbb{E}[D_n \mid \mathcal{E}] \, \mathbb{E}[\mathbb{1}_{\mathcal{E}}] \\ &= \mathbb{E}[D_n \mid \mathcal{E}] \, \mathbb{P}(\mathcal{E})\end{aligned} \tag{C.15}$$

where we used the fact that $\mathbb{E}[D_n \mid \sigma(\mathcal{E})] \mathbb{1}_{\mathcal{E}} = \mathbb{E}[D_n \mid \mathcal{E}] \mathbb{1}_{\mathcal{E}}$. We thus conclude that

$$\mathbb{E}\Big[\|\pi_n - \pi^*\|^2 \,\Big|\, \mathcal{E}\Big] = 2\,\mathbb{E}[D_n \mid \mathcal{E}] \leq \frac{2}{\mathbb{P}(\mathcal{E})} \bar{D}_n \leq \frac{2}{1-\delta} \bar{D}_n \tag{C.16}$$

and hence

$$\mathbb{E}\Big[\|\pi_n - \pi^*\|^2 \,\Big|\, \mathcal{E}\Big] = \begin{cases} \mathcal{O}(1/n^{2\mu\gamma}) & \text{if } p = 1 \text{ and } 2\mu\gamma < q, \\ \mathcal{O}(1/n^q) & \text{otherwise.}\end{cases}$$

■

*Proof of Corollary 2.* For Models 1 and 2, taking $\ell_b = \infty, \ell_\sigma = 0$ we readily get that $q = p$ and $p > 1/2$. Since we require that $\sum_{n=1}^\infty \gamma_n = \infty$, we obtain that $p \in (1/2, 1]$. Hence, for $p = 1$ and $2\mu\gamma > 1$ we obtain $\mathcal{O}(1/n)$ rate of convergence.

For Model 3, we have that $B_n = \mathcal{O}(\varepsilon_n)$ and $\sigma_n = \mathcal{O}(1/\sqrt{\varepsilon_n})$, i.e., $\ell_b = p/2$ and $\ell_\sigma = p/4$, and, hence, we readily get that $q = p/2$. Now, since $p \leq 1$, $p + \ell_b > 1$ and $p - \ell_\sigma > 1/2$, we obtain that $p \in (2/3, 1]$. Hence, for $p = 1$ and $\mu\gamma > 1$, we obtain $\mathcal{O}(1/\sqrt{n})$ rate of convergence. ■

# D Rate of convergence to strict Nash policies

**D.1. Structural preliminaries.** To prove Theorem 3, we will first require some notions describing the geometry of $\Pi$ near $\pi^*$. Referring to [47] for a full treatment, we have:

**Definition 3.** Let $\mathcal{C}$ be a convex set and let $x \in \mathcal{C}$. Then the tangent cone $\mathrm{TC}_{\mathcal{C}}(x)$ is defined as the set of all rays emanating from $x$ and intersecting $\mathcal{C}$ to at least one other point different from $x$. The *polar cone* $\mathrm{PC}_{\mathcal{C}}(x)$ to $\mathcal{C}$ at $x$ is then defined $\mathrm{PC}_{\mathcal{C}}(x) = \{y : \langle y, z\rangle \leq 0 \text{ for all } z \in \mathrm{TC}_{\mathcal{C}}(x)\}$, where $y$ belong in the dual space of the vector space in which $\mathcal{C}$ is defined.

With these general definitions in hand, we proceed to characterize some further projections of Euclidean projections on $\Pi$ that will play an important role in the sequel. For notational simplicity, we suppress the player and state indices in the statement and proof of the next lemma.

**Lemma D.1.** $x = \mathrm{proj}(y)$ *if and only if there exist* $\mu \in \mathbb{R}$ *and* $\nu_\alpha \in \mathbb{R}_+$ *such that, for all* $\alpha \in \mathcal{A}$*, we have* $y_\alpha = x_\alpha + \mu - \nu_\alpha$ *with* $\nu_\alpha \geq 0$ *and* $x_\alpha \nu_\alpha = 0$*.*

*Proof.* Recall that $\mathrm{proj}(y) = \arg\min_{x\in\Delta(\mathcal{A})} \|y - x\|^2$. Our result then follows by applying the KKT conditions to this optimization problem and noting that, since the constraints are affine, the KKT conditions are sufficient for optimality. Our Langragian is

$$\mathcal{L}(x, \mu, \nu) = \sum_{\alpha\in\mathcal{A}} \frac{1}{2}(y_\alpha - x_\alpha)^2 - \mu\Big(\sum_{\alpha\in\mathcal{A}} x_\alpha - 1\Big) + \sum_{\alpha\in\mathcal{A}} \nu_\alpha x_\alpha$$

where the set of constraints (i) of the statement of the lemma are the stationarity constraints, which in our case are $\nabla\mathcal{L}(x,\mu,\nu)=0 \Leftrightarrow \nabla(\sum_{\alpha\in\mathcal{A}}\frac{1}{2}(y_\alpha-x_\alpha)^2)=\mu\nabla(\sum_{\alpha\in\mathcal{A}}x_\alpha-1)-\sum_{\alpha\in\mathcal{A}}\nu_\alpha\nabla x_\alpha$ , while the set of constraints (ii) of the statement of the lemmas are the complementary slackness constraints. Note that complementary slackness implies $\nu_\alpha>0$ whenever $\alpha\notin\operatorname{supp}(x)$, so our proof is complete. ■

Our next result is a concrete consequence of Proposition 1 which will be very useful in establishing the stability estimates required for the proof of Theorem 3.

**Lemma D.2.** *Let $\pi^*=(\alpha^*_{i,s})_{i\in\mathcal{N},s\in\mathcal{S}}$ be a strict Nash policy. Then there exists a neighborhood $\mathcal{U}$ of $\pi^*$ and constants $c_{i,s}$ such that for each player $i\in\mathcal{N}$ and state $s\in\mathcal{S}$, we have:*

$$v_{i\alpha^*_{i,s}}(\pi)-v_{i\alpha_{i,s}}(\pi)\geq c_{i,s} \text{ for all } \pi\in\mathcal{U} \text{ and } \alpha_i\neq\alpha^*_i,\ \alpha_i\in\mathcal{A}_i. \tag{D.1}$$

*Proof.* Our claim is a consequence of the definition of strict Nash policies. Specifically, from Proposition 1 we have

$$\langle v(\pi^*),z\rangle<0 \quad \text{for all} \quad z\in\mathrm{TC}(\pi^*), z\neq 0 \tag{D.2}$$

Let $z=e_{i,\alpha_{i,s}}-e_{i,\alpha^*_{i,s}}$, then we get that

$$v_{i\alpha^*_{i,s}}(\pi^*)-v_{i\alpha_{i,s}}(\pi^*)>0 \tag{D.3}$$

where $e_{i,\alpha_{i,s}}$ is the vector that has one only in the index and zero anywhere else. By continuity there exists a neighborhood $\mathcal{U}\subseteq\mathcal{X}$ and $c_{i,s}>0$ for each player $i\in\mathcal{N}$ such that

$$v_{i\alpha^*_{i,s}}(\pi)-v_{i\alpha_{i,s}}(\pi)\geq c_{i,s} \quad \text{for all} \quad \pi\in\mathcal{U}$$

■

Our final result is intimately tied to the lazy projection step in (LPG), and quantifies the relation between initializations in $\prod_i\left(\mathbb{R}^{\mathcal{A}_i}\right)^{\mathcal{S}}$ and $\Pi$.

**Lemma D.3.** *Let $\pi^*=(\alpha^*_{i,s})_{i\in\mathcal{N},s\in\mathcal{S}}$, be a deterministic policy. For each agent $i\in\mathcal{N}$ and each state $s\in\mathcal{S}$, let $y_{i,\alpha_{i,s}}-y_{i,\alpha^*_{i,s}}$ be the difference of the aggregated gradients between the strategy of the equilibrium and any other strategy $\alpha^*_i\neq\alpha_i\in\mathcal{A}_i$. Then for any $\varepsilon>0$ such that $\mathcal{U}_\varepsilon=\{\pi:\pi_{i,\alpha^*_{i,s}}\geq 1-\varepsilon$ for all $i\in\mathcal{N}$ and $s\in\mathcal{S}\}$, there exist $M_{i,\varepsilon,s}$ such that if $\mathcal{W}_{i,s}=\{y\in\mathbb{R}^{\mathcal{A}_i}:y_{i,\alpha_{i,s}}-y_{i,\alpha^*_{i,s}}<-M_{i,\varepsilon,s}\}$ then $\prod_{i\in\mathcal{N},s\in\mathcal{S}}\operatorname{proj}_{\Pi_i}(\mathcal{W}_{i,s})\subseteq\mathcal{U}_\varepsilon$.*

*Proof.* Consider an arbitrary player $i\in\mathcal{N}$, a state $s\in\mathcal{S}$, and let $\mathcal{W}_i(M_{i,\varepsilon,s})$ be an open set as defined in the statement of the lemma. For notational simplicity, we will drop the index $s$. We will show that any $M_{i,\varepsilon}>1-\frac{\varepsilon}{|\mathcal{A}_i|}>0$ satisfies our claim. By using Lemma D.1 for a $y_i\in\mathcal{W}_i(M_{i,\varepsilon})$ with $\pi_i=\operatorname{proj}(y_i)$ we have that

$$y_{i\alpha^*_i}-y_{i\alpha_i}>M_{i,\varepsilon} \tag{D.4}$$

$$\pi_{i\alpha^*_i}-\pi_{i\alpha_i}-(\nu_{\alpha^*_i}-\nu_{\alpha_i})>M_{i,\varepsilon} \tag{D.5}$$

with $\nu_{\alpha_i}\geq 0$ and $\pi_{i\alpha_i}=0$ whenever $\nu_{\alpha_i}>0$. Notice that since $M_{i,\varepsilon}>1-\frac{\varepsilon}{|\mathcal{A}_i|}$ we have that $\pi_{i\alpha^*_i}>\pi_{i\alpha_i}+1-\frac{\varepsilon}{|\mathcal{A}_i|}+(\nu_{\alpha^*_i}-\nu_{\alpha_i})$ or

$$\pi_{i\alpha_i}<\pi_{i\alpha^*_i}-1+\frac{\varepsilon}{|\mathcal{A}_i|}-(\nu_{\alpha^*_i}-\nu_{\alpha_i})<\frac{\varepsilon}{|\mathcal{A}_i|} \tag{D.6}$$

Hence, by summing over all strategies of player $i$ we get the desired result. ■

**D.2. Proof of the main theorem.** We are now in a position to prove our main result on the rate of convergence towards strict Nash policies. For ease of reference, we restate Theorem 3 below.

**Theorem 3.** *Let $\pi_n$ be the sequence of play under (LPG) with step-size and policy gradient estimates such that $p+\ell_b>1$ and $p-\ell_\sigma>1/2$ as per (8). If $\pi^*$ is a deterministic Nash policy, there exists an unbounded open set $\mathcal{W}\subseteq\prod_i\left(\mathbb{R}^{\mathcal{A}_i}\right)^{\mathcal{S}}$ of initializations such that, for any $\delta>0$, we have*

$$\mathbb{P}(\pi_n \text{ converges to } \pi^* \mid y_1\in\mathcal{W})\geq 1-\delta, \tag{20}$$

*provided that $\gamma>0$ is small enough. Moreover, conditioned on this event, $\pi_n$ converges to $\pi^*$ at a finite number of iterations, i.e., there exists some $n_0$ such that $\pi_n=\pi^*$ for all $n\geq n_0$.*

*Proof of Theorem 3.* We start by fixing a confidence level $\delta > 0$ and all the parameters of the algorithm, such that all the assumptions stated in the theorem are satisfied and. We will prove that for each agent $i \in \mathcal{N}$, $s \in \mathcal{S}$ there exist $M_{1,i,s} > 0$, $\mathcal{W}_{1,i,s} = \{y \in \mathbb{R}^{\mathcal{A}_i} : y_{i,\alpha_i} - y_{i,\alpha_i^*} < -M_{1,i,s}$ for all $\alpha_i \in \mathcal{A}_i, \alpha_i \neq \alpha_i^*\}$, such that if $y_1 \in \mathcal{W}_1 \coloneqq \prod_{i\in\mathcal{N},s\in\mathcal{S}} \mathcal{W}_{1,i,s}$ then the agents' sequence of play, converge to the deterministic Nash policy, in finite number of iterations.

To simplify the notation, we will drop the indices $s$ and $i$ referring to the states and agents, accordingly, and we will focus on a specific agent and a specific state. From Lemma D.3, Lemma D.2 we have that there exist constants $c, M$, neighborhood $\mathcal{U}_c = \{\pi \in \Pi : \|\pi - \pi^*\| \leq \beta\}$ and open set $\mathcal{W}_M$ such that

$$v_{\alpha^*}(\pi) - v_\alpha(\pi) \geq c \quad \text{for all } \alpha \neq \alpha^*, \alpha \in \mathcal{A} \text{ and } \pi \in \mathcal{U}_c \tag{D.7}$$

$$y_{\alpha^*} - y_\alpha > M_c \quad \text{for all } \alpha \neq \alpha^*, \alpha \in \mathcal{A} \text{ and } \pi = \operatorname{proj}(y) \in \mathcal{U}_c \tag{D.8}$$

The first step is to prove that for an appropriate initialization for $y_1$, we have $y_n \in \mathcal{W}(M_c)$ for all $n = 1, 2, \dots$, with probability at least $1 - \delta$. Assume that $y_k \in \mathcal{W}(M_c)$ for all $k = 1, \dots, n$; then for the differences of the scores at a round $n+1$ between any $\alpha \in \mathcal{A}$ and the equilibrium strategy $\alpha^*$, we have

$$\begin{aligned}
y_{\alpha,n+1} - y_{\alpha^*,n+1} &= y_{\alpha,n} - y_{\alpha^*,n} + (\hat{v}_{\alpha,n} - \hat{v}_{\alpha^*,n}) \\
&= y_{\alpha,1} - y_{\alpha^*,1} + \sum_{k=1}^n \gamma_k[(v_{\alpha,k} - v_{\alpha^*,k}) + (U_{\alpha,k} - U_{\alpha^*,k}) + (b_{\alpha,k} - b_{\alpha^*,k})] \\
&\leq -M_1 + \sum_{k=1}^n \gamma_k[(v_{\alpha,k} - v_{\alpha^*,k}) + (U_{\alpha,k} - U_{\alpha^*,k}) + (b_{\alpha,k} - b_{\alpha^*,k})] \\
&\leq -M_1 - c\sum_{k=1}^n \gamma_k + \sum_{k=1}^n \gamma_k[(U_{\alpha,k} - U_{\alpha^*,k}) + (b_{\alpha,k} - b_{\alpha^*,k})] \\
&\leq -M_1 - c\sum_{k=1}^n \gamma_k + \sum_{k=1}^n \gamma_k[\xi_k + \chi_k]
\end{aligned} \tag{D.9}$$

where $\xi_k = (U_{\alpha,k} - U_{\alpha^*,k})$ and $\chi_k = 2\|b_k\|$. Now, similarly to the proofs of Theorems 1 and 2 we will proceed to control the aggregate error terms

$$R_n = \sum_{k=1}^n \gamma_k \xi_k \quad \text{and} \quad S_n = \sum_{k=1}^n \gamma_k \chi_k. \tag{D.10}$$

Since $\mathbb{E}[\xi_n \,|\, \mathcal{F}_n] = 0$, we have $\mathbb{E}[R_n \,|\, \mathcal{F}_n] = R_{n-1}$, so $R_n$ is a martingale; likewise, $\mathbb{E}[S_n \,|\, \mathcal{F}_n] \geq S_{n-1}$, so $S_n$ is a sub-martingale. Furthermore from (7) we have:

I. $\mathbb{E}[\xi_n^2] \leq \mathbb{E}[\|U_n\|^2] \leq \mathbb{E}[\mathbb{E}[\|U_n\|^2 \,|\, \mathcal{F}_n]] \leq \sigma_n^2$

II. $\mathbb{E}[\chi_n] = 2\,\mathbb{E}[\|b_n\|] \leq \mathbb{E}[\mathbb{E}[\|b_n\| \,|\, \mathcal{F}_n]] \leq B_n$

Moreover, for any $\eta_1 > 0$, we get by Doob's Maximal Inequality:

$$\mathbb{P}\left(\sup_{1\leq k\leq n} R_k \geq \eta_1\right) \leq \frac{\mathbb{E}[R_n^2]}{\eta_1^2} \overset{(a)}{=} \frac{\sum_{k=1}^n \gamma_k^2\, \mathbb{E}[\xi_k^2]}{\eta_1^2} \overset{(I.)}{\leq} \frac{\sum_{k=1}^n \gamma_k^2 \sigma_k^2}{\eta_1^2} \tag{D.11}$$

where $(a)$ comes from the fact that $\mathbb{E}[\xi_i \xi_j] = 0$ for $i \neq j$. Since $\gamma_n = \gamma/n^p$, $\sigma_n = \mathcal{O}(n^{\ell_\sigma})$ and $p - \ell_\sigma > 1/2$, there exists $\gamma_1$ sufficiently small such that if $\gamma \leq \gamma_1$ then

$$\sum_{k=1}^\infty \gamma_k^2 \sigma_k^2 < \frac{\delta\eta_1^2}{2} \tag{D.12}$$

and so we automatically get that

$$\mathbb{P}\left(\sup_{1\leq k\leq n} R_k \geq \eta_1\right) \leq \frac{\delta}{2} \tag{D.13}$$

Furthermore, notice that the term $\{S_n\}_{n\in\mathbb{N}}$ is a sub-martingale, since $\mathbb{E}[|S_n| \,|\, \mathcal{F}_n] < \infty$ and $\mathbb{E}[S_{n+1} \,|\, \mathcal{F}_n] > S_n$, for all $n$. As before, using Doob's Maximal Inequality, we get for any $\eta_2 > 0$:

$$\mathbb{P}\left(\sup_{1\leq k\leq n} S_k \geq \eta_2\right) \leq \frac{\mathbb{E}[S_n]}{\eta_2} = \frac{\sum_{k=1}^{n} \gamma_k \,\mathbb{E}[\chi_k]}{\eta_2} \leq \frac{2\sum_{k=1}^{n} \gamma_k B_k}{\eta_2} \tag{D.14}$$

So, since $p + \ell_b > 1$ there exists $\gamma_2$ sufficiently small such that if $\gamma \leq \gamma_2$ then

$$\sum_{k=1}^{n} \gamma_k B_k \leq \frac{\eta_2 \delta}{4} \tag{D.15}$$

which immidiately implies that

$$\mathbb{P}\left(\sup_{1\leq k\leq n} S_k \geq \eta_2\right) \leq \frac{\delta}{2} \tag{D.16}$$

By choosing $\gamma \leq \min\{\gamma_1, \gamma_2\}$ we get that

$$\mathbb{P}\left(\sup_{1\leq k\leq n} R_n + S_n \leq M_c\right) \geq 1 - \delta. \tag{D.17}$$

Notice now that by choosing $M_1 > M_c + \eta_1 + \eta_2$, from (D.9) we have that with probability at least $1 - \delta$, $y_{\alpha,n+1} - y_{\alpha^*,n+1} < -M_c$, which implies that $\pi_{n+1} \in \mathcal{U}_c$.

Defining the sequences of "good" events $\{\mathcal{E}_n\}_{n\in\mathbb{N}}$ and $\{\mathcal{E}'_n\}_{n\in\mathbb{N}}$ as $\mathcal{E}_n := \{\pi_k \in \mathcal{U}_c, \forall k = 1, \dots, n\}$ and $\mathcal{E}'_n := \left\{\sup_{1\leq k\leq n} R_k + S_k \leq \eta_1 + \eta_2\right\}$, accordingly, we get that $\mathcal{E}'_n \subseteq \mathcal{E}_n$ for all $n$. Because $\mathbb{P}(\mathcal{E}'_n) \geq 1 - \delta$, we get that

$$\mathbb{P}(\mathcal{E}_n) \geq 1 - \delta \tag{D.18}$$

and since $\{\mathcal{E}_n\}_{n\in\mathbb{N}}$ is a decreasing sequence converging to $\mathcal{E} := \{\pi_n \in \mathcal{U}_c, \forall n \in \mathbb{N}\}$, we obtain

$$\mathbb{P}(\mathcal{E}) \geq 1 - \delta. \tag{D.19}$$

i.e.,

$$\mathbb{P}(\pi_n \in \mathcal{U}_c,\ \forall n \mid y_1 \in \mathcal{W}_1) \geq 1 - \delta \tag{D.20}$$

Notice that the above conclusions immediately imply convergence in finite time. More specifically, constrained to the event $\mathcal{E}$ with probability at least $1 - \delta$, from Eq. (D.9) we have

$$y_{\alpha,n+1} - y_{\alpha^*,n+1} \leq -M_c - c\sum_{k=1}^{n} \gamma_k \tag{D.21}$$

for all $n = 1, 2, \dots$. Assume ad absurdum that there exists at least one strategy $\alpha \neq \alpha^*, \alpha \in \mathcal{A}$ such that $\limsup_{n\to\infty} \pi_{\alpha,n} \geq \varepsilon > 0$. for all sufficiently large $n$. Recall also that for $\pi \in \mathcal{U}_c$, it holds that $\pi_{\alpha^*} > 0$ by construction. Using Lemma D.1 we get

$$y_{\alpha,n+1} - y_{\alpha^*,n+1} = \pi_{\alpha,n+1} - \pi_{\alpha^*,n+1} \leq -M_c - c\sum_{k=1}^{n} \gamma_k \tag{D.22}$$

Notice that the L.H.S. of this inequality is bounded, while the R.H.S. goes to $-\infty$, which is a contradiction. Thus, with probability at least $1 - \delta$, $\pi_n \to \pi^*$ as $n \to \infty$.

We can rewrite the previous inequality as

$$\pi_{\alpha,n+1} \leq 1 - M_c - c\sum_{k=1}^{n} \gamma_k \quad \text{for all} \quad \alpha^* \neq \alpha \in \mathcal{A} \tag{D.23}$$

Now aggregating over all strategies, on the previous inequality, we get that

$$\|\pi_{n+1} - \pi^*\|_1 = 2(1 - \pi_{\alpha^*,n+1}) \leq 2\sum_{\alpha^*\neq\alpha\in\mathcal{A}} (1 - M_c - c\sum_{k=1}^{n} \gamma_k) \tag{D.24}$$

Thus, once $\sum_{k=1}^{n} \gamma_k$ becomes at least $(1 - M_c)/c$, which occurs in finite time, the convergence is implied. ■

*Proof of Corollary 3.* For Models 1 and 2, taking $\ell_b = \infty, \ell_\sigma = 0$ we readily get that $p > 1/2$. Since we require that $\sum_{n=1}^{\infty} \gamma_n = \infty$, we obtain that $p \in (1/2, 1]$.

For Model 3, we have that $B_n = \mathcal{O}(\varepsilon_n)$ and $\sigma_n = \mathcal{O}(1/\sqrt{\varepsilon_n})$, i.e., $\ell_b = r$ and $\ell_\sigma = r/2$. Now, since $p \leq 1$, $p + \ell_b > 1$ and $p - \ell_\sigma > 1/2$, we obtain that $p \in (2/3, 1]$. ■

# E Structural properties of policy gradient methods

In this part of the appendix we will establish the necessary properties about the value function, its gradient. More precisely,

- In Lemma E.1 we prove that in the random stopping episodic framework visitation the notion of discounted state visitation distribution is well-defined.
- In Lemma 1, we prove the conversion lemma, a standard lemma that connects a sample by visitation distribution and a random trajectory.
- In Lemma E.4, we establish different versions of Policy Gradient theorem via $Q$-value function for the random stopping episodic framework.
- In Lemma E.5 and E.7, we establish the boundedness and the Lipschitz smoothness of policy gradient vector field, i.e., $v(\pi) = (v_i(\pi))_{i\in\mathcal{N}}$ where $v_i(\pi) = \nabla_{\pi_i} V_{i,s}(\pi)$

For a policy profile $\pi \in \Pi$ and an arbitrary initial state distribution $s_0 \sim \rho$, let's recall the definition of discounted state visitation measure/distribution as

$$\tilde{d}_\rho^\pi(s) = \mathbb{E}_{\tau\sim\text{MDP}}\Big[\sum\nolimits_{t=0}^{T(\tau)} \mathbb{1}\{s_t = s\}\Big| s_0 \sim \rho\Big], \quad d_\rho^\pi(s) := \tilde{d}_\rho^\pi(s)/Z_\rho^\pi$$

To begin with, we prove formally that the above definition is well-posed for the random stopping episodic framework described above, i.e., $\tilde{d}_\rho^\pi(s) < \infty$, so $Z_\rho^\pi := \sum_{s\in\mathcal{S}} \tilde{d}_\rho^\pi(s)$ is well-defined.

**Lemma E.1.** *For any $s \in \mathcal{S}$, $\tilde{d}_\rho^\pi(s) < \infty$ and $Z_\rho^\pi \leq \frac{1}{\zeta}$.*

*Proof.* For the sake of the proof, we define a new state $s_f$, indicating that the game has stopped. In other words, we have that $P(s_f \mid s, \alpha) = \zeta_{s,\alpha} \geq \zeta > 0$ for all $\alpha \in \mathcal{A}, s \in \mathcal{S}$. Hence, for $s \in \mathcal{S}$ we obtain:

$$\begin{aligned}
\tilde{d}_\rho^\pi(s) &= \mathbb{E}_{\tau\sim\text{MDP}}\Big[\sum\nolimits_{t=0}^{T(\tau)} \mathbb{1}\{s_t = s\}\Big| s_0 \sim \rho\Big] && \text{(E.1)}\\
&= \mathbb{E}_{\tau\sim\text{MDP}}\left[\sum_{t=0}^{\infty} \mathbb{1}\{s_t = s, s_i \neq s_f, 1 \leq i \leq t\} \mid s_0 \sim \rho\right] && \text{(E.2)}\\
&\leq \sum_{s\in\mathcal{S}} \tilde{d}_\rho^\pi(s) && \text{(E.3)}\\
&= \mathbb{E}_{\tau\sim\text{MDP}}\left[\sum_{t=0}^{\infty} \mathbb{1}\{s_i \neq s_f, 1 \leq i \leq t\} \mid s_0 \sim \rho\right] && \text{(E.4)}\\
&= \sum_{t=0}^{\infty} \mathbb{P}(s_i \neq s_f, 1 \leq i \leq t \mid s_0 \sim \rho) && \text{(E.5)}\\
&= \sum_{t=0}^{\infty} \prod_{i=1}^{t} \mathbb{P}(s_i \neq s_f \mid s_0 \sim \rho, s_j \neq s_f, 1 \leq j \leq i-1) && \text{(E.6)}\\
&\leq \sum_{t=0}^{\infty} (1-\zeta)^t \leq \frac{1}{\zeta} && \text{(E.7)}\\
&< \infty. && \text{(E.8)}
\end{aligned}$$

■

**Lemma 1.** [Conversion Lemma] *For an arbitrary state-action function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a policy profile $\pi$ and an initial state distribution $s_0 \sim \rho$, we have*

$$\mathbb{E}_{\tau\sim\text{MDP}}\Big[\sum\nolimits_{t=0}^{T(\tau)} f(s_t, \alpha_t)\Big] = Z_\rho^\pi \, \mathbb{E}_{s\sim d_\rho^\pi} \mathbb{E}_{\alpha\sim\pi(\cdot|s)}[f(s, \alpha)] \tag{2}$$

*Proof.*

$$\mathbb{E}_{\tau\sim\text{MDP}}\Big[\sum\nolimits_{t=0}^{T(\tau)} f(s_t, \alpha_t)\Big] = \sum_{t=0}^{\infty} \sum_{s\in\mathcal{S}} \sum_{\alpha\in\mathcal{A}} \mathbb{E}_{\tau\sim\text{MDP}} \left[\mathbb{1}\{t \leq T(\tau), s_t = s, \alpha_t = \alpha\} f(s, \alpha)\right]$$

$$
\begin{aligned}
&= \sum_{s\in\mathcal{S}} \sum_{t=0}^{\infty} \sum_{\alpha\in\mathcal{A}} \mathbb{P}^{\pi}(s = s_t \mid s_0 \sim \rho)\pi(\alpha \mid s) f(s,\alpha) \\
&= \sum_{s\in\mathcal{S}} \sum_{t=0}^{\infty} \mathbb{P}^{\pi}(s = s_t \mid s_0 \sim \rho) \sum_{\alpha\in\mathcal{A}} \pi(\alpha \mid s) f(s,\alpha) \\
&= \sum_{s\in\mathcal{S}} \tilde{d}^{\pi}_{\rho}(s)\, \mathbb{E}_{\alpha\sim\pi(\cdot|s)} [f(s,\alpha)] \\
&= Z^{\pi}_{\rho}\, \mathbb{E}_{s\sim d^{\pi}_{\rho}}\, \mathbb{E}_{\alpha\sim\pi(\cdot|s)}[f(s,\alpha)] \qquad \text{(E.9)}
\end{aligned}
$$

where $Z^{\pi}_{\rho} := \mathbb{E}_{s\sim\mathrm{Unif}(\mathcal{S})}[\tilde{d}^{\pi}_{\rho}(s)] \cdot |\mathcal{S}|$ is well-defined by E.1. ■

An equivalent but very useful way to describe compactly the aforementioned lemma is via the matrix representation of the discounted visitation distribution:

**Lemma E.2** ( Conversion Lemma (Matrix form) ). *For an arbitrary state-action function $f\colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ and a policy profile $\pi$, we have*

$$
\mathbb{E}_{\tau\sim\mathrm{MDP}}\Big[\sum\nolimits_{t=0}^{T(\tau)} f(s_t,\alpha_t) \mid \alpha_0 = \alpha, s_0 = s\Big] = e^{\top}_{s,\alpha}\mathcal{T}(\pi) f \qquad \text{(E.10)}
$$

*where $\mathcal{T}$ is a discounted visitation distribution (action-state)-matrix under poliy profile $\pi$ i.e., $[\mathcal{T}(\pi)]_{\underbrace{(\alpha,\, s)}_{\text{Row Index}}\to\underbrace{(\alpha',\, s')}_{\text{Column Index}}} = \sum_{t=0}^{\infty} \mathbb{P}^{\pi}(s_t = s', \alpha_t = \alpha' \mid s_0 = s, \alpha_0 = \alpha)$*

*Proof.* By definition we have

$$
\begin{aligned}
e^{\top}_{s,\alpha}\mathcal{T}(\pi) f &= \langle e^{\top}_{s,\alpha}\mathcal{T}(\pi), f\rangle && \text{(E.11)} \\
&= \sum_{s'\in\mathcal{S}} \sum_{\alpha'\in\mathcal{A}} \left(e^{\top}_{s,\alpha}\mathcal{T}(\pi)\right)_{(s',\alpha')} \cdot f(s',\alpha') && \text{(E.12)} \\
&= \sum_{s'\in\mathcal{S}} \sum_{\alpha'\in\mathcal{A}} e^{\top}_{s,\alpha}\mathcal{T}(\pi) e_{s',\alpha'} \cdot f(s',\alpha') && \text{(E.13)} \\
&= \sum_{s'\in\mathcal{S}} \sum_{\alpha'\in\mathcal{A}} \sum_{t=0}^{\infty} \mathbb{P}^{\pi}(s_t = s', \alpha_t = \alpha' \mid s_0 = s, \alpha_0 = \alpha) \cdot f(s',\alpha') && \text{(E.14)} \\
&= \sum_{t=0}^{\infty} \sum_{s'\in\mathcal{S}} \sum_{\alpha'\in\mathcal{A}} \mathbb{E}_{\tau\sim\mathrm{MDP}} \left[\mathbb{1}\{t \le T(\tau), s'_t = s, \alpha'_t = \alpha, \} f(s,\alpha) \mid s_0 = s, \alpha_0 = \alpha\right] && \text{(E.15)} \\
&= \mathbb{E}_{\tau\sim\mathrm{MDP}}\Big[\sum\nolimits_{t=0}^{T(\tau)} f(s_t,\alpha_t) \mid \alpha_0 = \alpha, s_0 = s\Big] && \text{(E.16)}
\end{aligned}
$$

■

*Remark* 1. Notice that $\mathcal{T}$ is a well-defined matrix. Indeed, let's us define $\mathcal{P}(\pi)$ as the state-action one step transition matrix:

$$
[\mathcal{P}(\pi)]_{\underbrace{(\alpha,\, s)}_{\text{Row Index}}\to\underbrace{(\alpha',\, s')}_{\text{Column Index}}} = \mathbb{P}^{\pi}(s_1 = s', \alpha_1 = \alpha' \mid s_0 = s, \alpha_0 = \alpha) = \pi(\alpha' \mid s') P(s'|s,\alpha).
$$

Notice that $\mathcal{P}(\pi)$ is a substochastic matrix and therefore spectral$(\mathcal{P}(\pi)) < 1$ or equivalently $(I - \mathcal{P}(\pi))^{-1}$ is invertible. Thus using Neumann series we have that $(I - \mathcal{P}(\pi))^{-1} = \sum_{t=0}^{\infty} \mathcal{P}(\pi)^t$. By induction, a folklore probabilistic-graph theoretic fact, we can show that $\sum_{t=0}^{\infty} \mathcal{P}(\pi)^t = \mathcal{T}(\pi)$.

In order to analyze the gradient of MARL policy gradient methods, we will introduce the notions $Q, A$ and their per-player averages that are useful in the MDP analysis.

**Definition 4.** For a state $s \in \mathcal{S}$, a policy $\pi$ and $\alpha = (\alpha_1, \ldots, \alpha_N) \in \mathcal{A}$, we define:

(i) The $Q$-value function of player $i$ as:

$$
Q^{\pi}_i(s,\alpha) := \mathbb{E}_{\tau\sim\mathrm{MDP}(\pi|s)}\left[\sum_{t=0}^{T(\tau)} R_i(s_t(\tau), \alpha_t(\tau)) \mid s_0 = s, \alpha_0 = \alpha\right] \qquad \text{(E.17)}
$$

(ii) The *Advantage*-function of player $i$ as:

$$A_i^\pi(s,\alpha) := Q_i^\pi(s,\alpha) - V_{i,s}(\pi) \tag{E.18}$$

We also define $\overline{Q_i^\pi}, \overline{A_i^\pi}$ to be the averaged for $i$-th player single MDP $Q$-value and advantage functions:

(i) The averaged $\overline{Q_i^\pi}$-value function of player $i$ as:

$$\overline{Q_i^\pi}(s,\alpha_i) := \mathbb{E}_{\alpha_{-i}\sim\pi_{-i}(\cdot|s)}\left[Q_i^\pi\left(s,(\alpha_i;\alpha_{-i})\right)\right] \tag{E.19}$$

(ii) The averaged *Advantage* ,$\overline{A_i^\pi}$-function of player $i$ as:

$$\overline{A_i^\pi}(s,\alpha_i) := \mathbb{E}_{\alpha_{-i}\sim\pi_{-i}(\cdot|s)}\left[A_i^\pi\left(s,(\alpha_i;\alpha_{-i})\right)\right], \tag{E.20}$$

Using Remark 1, we can rewrite the above notations using $\mathcal{T},\mathcal{P}$.

**Lemma E.3.** *For a policy profile $\pi$, we have that*

*1.* $Q_i^\pi(s,\alpha) = e_{s,\alpha}^\top \mathcal{T}(\pi) r_i$

*2.* $\tilde{d}_\rho^\pi(s) = \left[\sum_{s'\in\mathcal{S}}\rho(s')\sum_{\alpha'\in\mathcal{A}}\pi(\alpha'\mid s')e_{s',\alpha'}\right]^\top \mathcal{T}(\pi)\sum_{\alpha\in\mathcal{A}}e_{s,\alpha}$

*Proof.* We separately have using Lemma E.3 and Remark 1.

1. $Q_i^\pi(s,\alpha) = \mathbb{E}_{\tau\sim\mathrm{MDP}(\pi|s)}\left[\sum_{t=0}^{T(\tau)}R_i(s_t(\tau),\alpha_t(\tau))\mid s_0=s,\alpha_0=\alpha\right] = e_{s,\alpha}^\top\mathcal{T}(\pi)R_i$

2.

$$\tilde{d}_\rho^\pi(s) = \mathbb{E}_{\tau\sim\mathrm{MDP}}\left[\sum\nolimits_{t=0}^{T(\tau)}\mathbb{1}\{s_t=s\}\middle|s_0\sim\rho\right] \tag{E.21}$$

$$= \mathbb{E}_{s'\sim\rho}\,\mathbb{E}_{\tau\sim\mathrm{MDP}}\left[\sum_{t=0}^{T(\tau)}\sum_{\alpha\in\mathcal{A}}\mathbb{1}\{s_t=s,\alpha_t=\alpha\}\middle|s_0=s'\right] \tag{E.22}$$

$$= \mathbb{E}_{s'\sim\rho}\,\mathbb{E}_{\alpha'\sim\pi(\cdot|s)}\,\mathbb{E}_{\tau\sim\mathrm{MDP}}\left[\sum_{t=0}^{T(\tau)}\sum_{\alpha\in\mathcal{A}}\mathbb{1}\{s_t=s,\alpha_t=\alpha\}\middle|s_0=s',\alpha_0=\alpha'\right] \tag{E.23}$$

$$= \mathbb{E}_{s'\sim\rho}\,\mathbb{E}_{\alpha'\sim\pi(\cdot|s)}\left[e_{s',\alpha'}^\top\mathcal{T}(\pi)\sum_{\alpha\in\mathcal{A}}e_{s,\alpha}\right] \tag{E.24}$$

$$= \left[\sum_{s'\in\mathcal{S}}\rho(s')\sum_{\alpha'\in\mathcal{A}}\pi(\alpha'\mid s')e_{s',\alpha'}\right]^\top\mathcal{T}(\pi)\sum_{\alpha\in\mathcal{A}}e_{s,\alpha} \tag{E.25}$$

■

Having defined the above notions, we are ready to provide equivalent forms of the $v(\pi)$ operator that will permit us to prove its boundedness and smoothness. We start with the following versions of Policy gradient theorem for random stopping setting:

**Lemma E.4.** *For the independent gradient operator $v(\pi)$ per player the following expressions are equal to $v_i(\pi)$:*

*1.* $v_i(\pi) = \mathbb{E}_{\tau\sim\mathrm{MDP}}\left[\sum_{t=0}^{T(\tau)}\nabla_i\left(\log\pi_i(\alpha_{i,t}(\tau)\mid s_t(\tau))\right)\overline{Q_i^\pi}(s_t(\tau),\alpha_{i,t}(\tau))\right]$

*2.* $v_i(\pi) = Z_\rho^\pi\,\mathbb{E}_{s\sim d_\rho^\pi}\,\mathbb{E}_{\alpha_i\sim\pi_i(\cdot|s)}\left[\nabla_i\left(\log\pi_i(\alpha_i\mid s)\right)\overline{Q_i^\pi}(s,\alpha_i)\right]$

*3.* $(v_i(\pi))_{\alpha_i^\circ,s^\circ} = \frac{\partial V_{i,\rho}(\pi)}{\partial\pi_i(\alpha_i^\circ|s^\circ)} = \tilde{d}_\rho^\pi(s^\circ)\overline{Q_i^\pi}(s^\circ,\alpha_i^\circ) = Z_\rho^\pi d_\rho^\pi(s^\circ)\overline{Q_i^\pi}(s^\circ,\alpha_i^\circ)$

*Proof.* Let as recall again the definition of our independent gradient operator $v(\pi)$:

$$v_i(\pi) = \nabla_i V_{i,\rho}(\pi)$$

First, we will show that:

$$\nabla_i \left(V_{i,\rho}(\pi)\right) = \mathbb{E}_{\tau\sim\text{MDP}}\left[\sum_{t=0}^{T(\tau)} \nabla_i \left(\log \pi_i(\alpha_{i,t}(\tau) \mid s_t(\tau))\right) \overline{Q_i^\pi}(s_t(\tau), \alpha_{i,t}(\tau))\right] \tag{E.26}$$

We will start with an arbitrary $s_0$, and by linearity of $\nabla_{\pi_i}(\cdot)$ and $\mathbb{E}_{s_0\sim\rho}[\cdot]$, we will obtain the result.

$$\begin{aligned}
\nabla_i \left(V_{i,s_0}(\pi)\right) &= \nabla_i \left(\mathbb{E}_\tau \left[R_i(\tau)\right]\right) \\
&= \nabla_i \left(\mathbb{E}_{\alpha_i\sim\pi_i(\cdot|s_0)}\left[\overline{Q_i^\pi}(s_0, \alpha_i)\right]\right) \\
&= \nabla_i \left(\sum_{\alpha_i\in\mathcal{A}_i} \pi_i(\alpha_i \mid s_0)\overline{Q_i^\pi}(s_0, \alpha_i)\right) \\
&= \sum_{\alpha_i\in\mathcal{A}_i} \nabla_i \left(\pi_i(\alpha_i \mid s_0)\right) \overline{Q_i^\pi}(s_0, \alpha_i) + \pi_i(\alpha_i \mid s_0)\nabla_i \left(\overline{Q_i^\pi}(s_0, \alpha_i)\right) \\
&= \sum_{\alpha_i\in\mathcal{A}_i} \nabla_i \left(\log \pi_i(\alpha_i \mid s_0)\right) \pi_i(\alpha_i \mid s_0)\overline{Q_i^\pi}(s_0, \alpha_i) + \pi_i(\alpha_i \mid s_0)\nabla_i \left(\overline{Q_i^\pi}(s_0, \alpha_i)\right) \\
&= \mathbb{E}_{\alpha_i\sim\pi_i(\cdot|s_0)}\left[\nabla_i \left(\log \pi_i(\alpha_i \mid s_0)\right) \overline{Q_i^\pi}(s_0, \alpha_i)\right] \\
&\quad + \sum_{\alpha_i\in\mathcal{A}_i} \pi_i(\alpha_i \mid s_0)\nabla_i \left(\mathbb{E}_{\alpha_{-i}\sim\pi_{-i}(\cdot|s_0)}\left[R_i(s_0, \alpha) + \sum_{s_1\in\mathcal{S}} P(s_1 \mid s_0, \alpha)V_{i,s_1}(\pi)\right]\right) \\
&= \mathbb{E}_{\alpha_i\sim\pi_i(\cdot|s_0)}\left[\nabla_i \left(\log \pi_i(\alpha_i \mid s_0)\right) \overline{Q_i^\pi}(s_0, \alpha_i)\right] \\
&\quad + \sum_{\alpha_i\in\mathcal{A}_i} \pi_i(\alpha_i \mid s_0)\, \mathbb{E}_{\alpha_{-i}\sim\pi_{-i}(\cdot|s_0)}\left[\sum_{s_1\in\mathcal{S}} P(s_1 \mid s_0, \alpha)\nabla_i \left(V_{i,s_1}(\pi)\right)\right] \\
&= \mathbb{E}_{\alpha_i\sim\pi_i(\cdot|s_0)}\left[\nabla_i \left(\log \pi_i(\alpha_i \mid s_0)\right) \overline{Q_i^\pi}(s_0, \alpha_i)\right] \\
&\quad + \mathbb{E}_{\alpha\sim\pi(\cdot|s_0)}\left[\sum_{s_1\in\mathcal{S}} P(s_1 \mid s_0, \alpha)\nabla_i \left(V_{i,s_1}(\pi)\right)\right]
\end{aligned} \tag{E.27}$$

Thus, we can rewrite it as:

$$\begin{aligned}
\nabla_i \left(V_{i,s_0}(\pi)\right) &= \mathbb{E}_{\alpha_i\sim\pi_i(\cdot|s_0)}\left[\nabla_i \left(\log \pi_i(\alpha_i \mid s_0)\right) \overline{Q_i^\pi}(s_0, \alpha_i)\right] \\
&\quad + \mathbb{E}_{\alpha\sim\pi(\cdot|s_0)}\left[\sum_{s_1\in\mathcal{S}} P(s_1 \mid s_0, \alpha)\nabla_i \left(V_{i,s_1}(\pi)\right)\right] \\
&= \mathbb{E}_{\tau\sim\text{MDP}(\pi|s_0)}\left[\nabla_i \left(\log \pi_i(\alpha_{i,0}(\tau) \mid s_0)\right) \overline{Q_i^\pi}(s_0, \alpha_{i,0}(\tau))\right] \\
&\quad + \mathbb{E}_{\tau\sim\text{MDP}(\pi|s_0)}\left[\mathbb{1}\left\{T(\tau) \geq 1\right\} \nabla_i \left(V_{i,s_1(\tau)}(\pi)\right)\right] \\
&= \sum_{t=0}^{\infty} \mathbb{E}_{\tau\sim\text{MDP}(\pi|s_0)}\left[\mathbb{1}\{t \leq T(\tau)\}\nabla_i \left(\log \pi_i(\alpha_{i,t}(\tau) \mid s_t(\tau))\right) \overline{Q_i^\pi}(s_t(\tau), \alpha_{i,t}(\tau))\right] \\
&\quad + \mathbb{E}_{\tau\sim\text{MDP}(\pi|s_0)}\left[\mathbb{1}\{T(\tau) = \infty\}A_\infty\right] \\
&\overset{(a)}{=} \mathbb{E}_{\tau\sim\text{MDP}(\pi|s_0)}\left[\sum_{t=0}^{T(\tau)} \nabla_i \left(\log \pi_i(\alpha_{i,t}(\tau) \mid s_t(\tau))\right) \overline{Q_i^\pi}(s_t(\tau), \alpha_{i,t}(\tau))\right]
\end{aligned} \tag{E.28}$$

where $(a)$ holds because $\mathbb{P}(T(\tau) = \infty) = 0$, and $A_\infty$ is some limiting quantity.

Hence,we readily obtain:

$$\nabla_i \left(V_{i,\rho}(\pi)\right) = \mathbb{E}_{s_0\sim\rho}\left[\nabla_i \left(V_{i,s_0}(\pi)\right)\right] \tag{E.29}$$

Now we are ready to utilize our Lemma 1:

$$\nabla_i \left(V_{i,\rho}(\pi)\right) = Z_\rho^\pi\, \mathbb{E}_{s\sim d_\rho^\pi}\, \mathbb{E}_{\alpha\sim\pi(\cdot|s)}\left[\nabla_i \left(\log \pi_i(\alpha_i \mid s)\right) \overline{Q_i^\pi}(s, \alpha_i)\right] \tag{E.30}$$

$$= Z_\rho^\pi \, \mathbb{E}_{s\sim d_\rho^\pi} \, \mathbb{E}_{\alpha_i\sim\pi_i(\cdot|s)}\Big[\nabla_i \left(\log \pi_i(\alpha_i \mid s)\right) \overline{Q_i^\pi}(s,\alpha_i)\Big] \tag{E.31}$$

Decoupling $\nabla_i$ per a state $s^\circ$ and action $\alpha_i^\circ$, we get

$$\frac{\partial V_{i,\rho}(\pi)}{\partial \pi_i(\alpha_i^\circ \mid s^\circ)} = Z_\rho^\pi \, \mathbb{E}_{s\sim d_\rho^\pi} \, \mathbb{E}_{\alpha_i\sim\pi_i(\cdot|s)}\left[\frac{\partial \left(\log \pi_i(\alpha_i \mid s)\right)}{\partial \pi_i(\alpha_i^\circ \mid s^\circ)} \overline{Q_i^\pi}(s,\alpha_i)\right] \tag{E.32}$$

$$= Z_\rho^\pi \, \mathbb{E}_{s\sim d_\rho^\pi} \, \mathbb{E}_{\alpha_i\sim\pi_i(\cdot|s)}\left[\mathbb{1}\{\alpha_i^\circ = \alpha_i, s^\circ = s\}\frac{1}{\pi_i(\alpha_i^\circ \mid s^\circ)} \overline{Q_i^\pi}(s^\circ,\alpha_i^\circ)\right] \tag{E.33}$$

$$= \sum_{s\in\mathcal{S}} \tilde{d}_\rho^\pi(s) \sum_{\alpha_i\in\mathcal{A}_i} \pi_i(\alpha_i \mid s)\mathbb{1}\{\alpha_i^\circ = \alpha_i, s^\circ = s\}\frac{1}{\pi_i(\alpha_i^\circ \mid s^\circ)} \overline{Q_i^\pi}(s^\circ,\alpha_i^\circ) \tag{E.34}$$

$$= \tilde{d}_\rho^\pi(s^\circ)\overline{Q_i^\pi}(s^\circ,\alpha_i^\circ) = Z_\rho^\pi d_\rho^\pi(s^\circ)\overline{Q_i^\pi}(s^\circ,\alpha_i^\circ) \tag{E.35}$$

■

We are ready to bound the amplitude of the independent player gradient operator:

**Lemma E.5.** *For a given initial state distribution $\rho$, the independent player policy gradient operator $v(\pi)$ is bounded. More precisely,*

$$\|v_i(\pi)\| \leq \frac{\sqrt{|\mathcal{A}_i|}}{\zeta^2} \quad \& \quad \|v(\pi)\| \leq \frac{\sum_{i\in\mathcal{N}} \sqrt{|\mathcal{A}_i|}}{\zeta^2}$$

*Proof.* We start by analyzing $\|v_i(\pi)\|^2$ using the aforementioned Lemma E.4.

$$\begin{aligned}
\|v_i(\pi)\|^2 &= \sum_{\alpha_i^\circ, s^\circ, \in \mathcal{A}_i, \mathcal{S}} (v_i(\pi)_{\alpha_i^\circ, s^\circ})^2 \\
&= \sum_{s^\circ\in\mathcal{S}} \sum_{\alpha_i^\circ\in\mathcal{A}_i} \Big(\frac{\partial V_{i,\rho}(\pi)}{\partial \pi_i(\alpha_i^\circ \mid s^\circ)}\Big)^2 \\
&= \sum_{s^\circ\in\mathcal{S}} \sum_{\alpha_i^\circ\in\mathcal{A}_i} (Z_\rho^\pi d_\rho^\pi(s^\circ)\overline{Q_i^\pi}(s^\circ,\alpha_i^\circ))^2 \\
&\leq (Z_\rho^\pi)^2 \max_{\alpha_i^\circ, s^\circ, \in \mathcal{A}_i, \mathcal{S}} (\overline{Q_i^\pi}(s^\circ,\alpha_i^\circ))^2 \sum_{s^\circ\in\mathcal{S}} \sum_{\alpha_i^\circ\in\mathcal{A}_i} d_\rho^\pi(s^\circ)^2 \\
&\leq \frac{1}{\zeta^2} \max_{\alpha_i^\circ, s^\circ, \in \mathcal{A}_i, \mathcal{S}} (\mathbb{E}_{\alpha_{-i}\sim\pi_{-i}(\cdot|s)}\left[Q_i^\pi\left(s^\circ, (\alpha_i^\circ; \alpha_{-i})\right)\right])^2 \sum_{s^\circ\in\mathcal{S}} \sum_{\alpha_i^\circ\in\mathcal{A}_i} d_\rho^\pi(s^\circ) \\
&\leq \frac{1}{\zeta^2} \max_{\alpha^\circ, s^\circ, \in \mathcal{A}, \mathcal{S}} (Q_i^\pi(s^\circ,\alpha^\circ))^2 \sum_{\alpha_i^\circ\in\mathcal{A}_i} \sum_{s^\circ\in\mathcal{S}} d_\rho^\pi(s^\circ) \\
&\leq \frac{1}{\zeta^2} \max_{\alpha^\circ, s^\circ, \in \mathcal{A}, \mathcal{S}} \left(\mathbb{E}_{\tau\sim\mathrm{MDP}(\pi|s)}\left[\sum_{t=0}^{T(\tau)} R_i(s_t(\tau),\alpha_t(\tau)) \mid s_0 = s^\circ, \alpha_0 = \alpha^\circ\right]\right)^2 \sum_{\alpha_i^\circ\in\mathcal{A}_i} 1 \\
&\leq \frac{|\mathcal{A}_i|}{\zeta^2} \left(\mathbb{E}_{\tau\sim\mathrm{MDP}(\pi|s)}\left[\sum_{t=0}^{T(\tau)} 1 \mid s_0 = s^\circ, \alpha_0 = \alpha^\circ\right]\right)^2 \\
&\leq \frac{|\mathcal{A}_i|}{\zeta^4}
\end{aligned}$$

Thus we conclude that

$$\|v_i(\pi)\| \leq \frac{\sqrt{|\mathcal{A}_i|}}{\zeta^2} \quad \& \quad \|v(\pi)\| \leq \frac{\sum_{i\in\mathcal{N}} \sqrt{|\mathcal{A}_i|}}{\zeta^2}$$

■

To prove the smoothness of the policy gradient operator, we have first to establish the performance lemma for our setting. Respectively, we get

**Lemma E.6** (Performance lemma)**.** *For any pair of policy profiles $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi'_i, \pi'_{-i})$, it holds*

$$V_{i,\rho}(\pi_i, \pi_{-i}) - V_{i,\rho}(\pi'_i, \pi'_{-i}) = \mathbb{E}_{\tau\sim \mathrm{MDP}(\pi|\rho)}\left[\sum_{t=0}^{T(\tau)} A_i^{\pi'_i,\pi'_{-i}}(s_t, \alpha_t)\right] \tag{E.36}$$

*where* $\mathrm{MDP}(\pi|\rho)$ *signifies that players follow $\pi$ as policy profile with $\rho$ as the initial state distribution.*

*Proof.* We will initial prove the aforementioned result for an arbitrary deterministic initial state $s_0 = s$:

$$V_{i,s}(\pi) - V_{i,s}(\pi') = \mathbb{E}_{\tau\sim \mathrm{MDP}(\pi|\rho)}\left[\sum_{t=0}^{T(\tau)} R_i(s_t, \alpha_t)\right] - V_{i,s}(\pi') \tag{E.37}$$

$$= \mathbb{E}_{\tau\sim \mathrm{MDP}(\pi|s)}\left[\sum_{t=0}^{T(\tau)} (R_i(s_t, \alpha_t) + V_{i,s_t}(\pi') - V_{i,s_t}(\pi'))\right] - V_{i,s}(\pi') \tag{E.38}$$

$$= \mathbb{E}_{\tau\sim \mathrm{MDP}(\pi|s)}\left[\sum_{t=0}^{T(\tau)} R_i(s_t, \alpha_t) + \sum_{t=0}^{T(\tau)} (V_{i,s_t}(\pi') - V_{i,s}(\pi') - V_{i,s_t}(\pi'))\right] \tag{E.39}$$

$$= \mathbb{E}_{\tau\sim \mathrm{MDP}(\pi|s)}\left[\sum_{t=0}^{T(\tau)} (R_i(s_t, \alpha_t) + \mathbb{1}\{T(\tau) \geq t+1\} V_{i,s_{t+1}}(\pi')) - V_{i,s_t}(\pi')\right] \tag{E.40}$$

$$= \mathbb{E}_{\tau\sim \mathrm{MDP}(\pi|s)}\left[\sum_{t=0}^{T(\tau)} \left(Q_i^{\pi'}(s_t, \alpha_t) - V_{i,s_t}(\pi')\right)\right] \tag{E.41}$$

$$= \mathbb{E}_{\tau\sim \mathrm{MDP}(\pi|s)}\left[\sum_{t=0}^{T(\tau)} A_i^{\pi'}(s_t, \alpha_t)\right] \tag{E.42}$$

where in the last equation we recall the definition of the Advantage function and in the pre-last the equivalent definitions of $Q_i^{\pi}(s, \alpha)$

$$\begin{aligned} Q_i^{\pi}(s, \alpha) &= \mathbb{E}_{\tau\sim \mathrm{MDP}(\pi|s)}\left[\sum_{t=0}^{T(\tau)} R_i(s_t(\tau), \alpha_t(\tau)) \mid s_0 = s, \alpha_0 = \alpha\right] \\ &= R_i(s, \alpha) + \mathbb{E}_{\tau\sim \mathrm{MDP}(\pi|s)}\left[\mathbb{1}\{T(\tau) \geq 1\} V_{i,s_1}(\pi) \mid s_0 = s, \alpha_0 = \alpha\right] \end{aligned} \tag{E.43}$$

Applying the linearity of $\mathbb{E}_{s\sim\rho}[\cdot]$, we get the desired result:

$$V_{i,\rho}(\pi) - V_{i,\rho}(\pi') = \mathbb{E}_{\tau\sim \mathrm{MDP}(\pi|\rho)}\left[\sum_{t=0}^{T(\tau)} A_i^{\pi'}(s_t, \alpha_t)\right] = Z_\rho^{\pi}\, \mathbb{E}_{s\sim d_\rho^{\pi}}\, \mathbb{E}_{\alpha\sim\pi(\cdot|s)}\left[A_i^{\pi'}(s, \alpha)\right] \tag{E.44}$$

where the last expression comes from Lemma 1. ■

Before closing this section by proving the Lipschitz-smoothness of our operator, we describe a useful observation that would be helpful in the smoothness bounds.

**Proposition E.1.** *For any pair of policy profiles $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi'_i, \pi'_{-i})$ and an arbitrary initial state distribution $\rho$ and a subset $\mathcal{M} \subseteq \mathcal{N}$, it holds that:*

$$\sum_s d_\rho^{\pi}(s) \sum_{\alpha_{\mathcal{M}}} |(\pi_{\mathcal{M}} - \pi'_{\mathcal{M}})(\alpha_{\mathcal{M}} \mid s)| \leq \sum_{i\in\mathcal{M}} \sqrt{|\mathcal{A}_i|}\|\pi_i - \pi'_i\|$$

*where $\pi_{\mathcal{M}} = (\pi_i)_{i\in\mathcal{M}}$ and $\alpha_{\mathcal{M}} = (\alpha_i)_{i\in\mathcal{M}}$, correspondingly.*

*Proof.*

$$\sum_s d_\rho^{\pi}(s) \sum_{\alpha_{\mathcal{M}}} |(\pi_{\mathcal{M}} - \pi'_{\mathcal{M}})(\alpha_{\mathcal{M}} \mid s)| = 2\sum_s d_\rho^{\pi}(s)\frac{1}{2}\|(\pi_{\mathcal{M}} - \pi'_{\mathcal{M}})\|_1 \tag{E.45}$$

$$= 2\sum_s d_\rho^\pi(s)\frac{1}{2}\mathrm{d_{TV}}(\pi_{\mathcal{M}}(\cdot|s),\pi'_{\mathcal{M}}(\cdot|s)) \tag{E.46}$$

$$\leq 2\sum_s d_\rho^\pi(s)\sum_{i\in\mathcal{M}}\frac{1}{2}\mathrm{d_{TV}}(\pi_i(\cdot|s),\pi'_i(\cdot|s)) \tag{E.47}$$

$$= \sum_s d_\rho^\pi(s)\sum_{i\in\mathcal{M}}\|(\pi_i(\cdot|s)-\pi'_i(\cdot|s))\|_1 \tag{E.48}$$

$$= \sum_s d_\rho^\pi(s)\sum_{i\in\mathcal{M}}\sqrt{|\mathcal{A}_i|}\|\pi_i-\pi'_i\|_2 \tag{E.49}$$

$$= \sum_{i\in\mathcal{M}}\sqrt{|\mathcal{A}_i|}\|\pi_i-\pi'_i\|_2(\sum_s d_\rho^\pi(s)) \tag{E.50}$$

$$= \sum_{i\in\mathcal{M}}\sqrt{|\mathcal{A}_i|}\|\pi_i-\pi'_i\|_2 \tag{E.51}$$

where $\mathrm{d_{TV}}$ corresponds to the total variation distance. Indeed notice that $\mathrm{d_{TV}}$ actually equals to the normalized difference of the histograms between two distributions. Additionally, the first inequality is derived by the "triangle inequality" that holds for $\mathrm{d_{TV}}$ in product-measure distributions. ∎

**Lemma E.7.** *For a given initial state distribution $\rho$, the independent player policy gradient operator $v(\pi)$ is lipschitz-smooth. More precisely, for any pair of policy profiles $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi'_i, \pi'_{-i})$, it holds*

$$\|v_i(\pi) - v_i(\pi')\| = \|\nabla_i(V_{i,\rho}(\pi) - \nabla_i(V_{i,\rho}(\pi')\| \leq \frac{3\sqrt{|\mathcal{A}_i|}}{\zeta^3}\sum_{j=1}^{N}\sqrt{|\mathcal{A}_i|}\|\pi_j - \pi'_j\| \quad \forall i \in \mathcal{N}$$

*and consequently,*

$$\|v(\pi) - v(\pi')\| \leq \frac{3|\mathcal{A}|}{\zeta^3}\|\pi - \pi'\|$$

*Proof.* For the proof, we will follow the approach of Zhang et al. [68] and Agarwal et al. [1]. Our first task is to bound the directional derivative of the $i$-th player's value function. We start by setting some notation. Let $\pi, \pi' \in \Pi$ and pert $\in \mathcal{S}\times\mathcal{A}$ such that $\|\mathrm{pert}\| = 1$. Then, we define the following $\lambda$-almost perturbed policies:

$$\begin{cases}\pi_\lambda^{\mathbb{A}}(\alpha \mid s) = (\pi_i + \lambda\mathrm{pert}, \pi_{-i}) \\ \pi_\lambda^{\mathbb{B}}(\alpha \mid s) = (\pi'_i + \lambda\mathrm{pert}, \pi'_{-i})\end{cases}$$

$$\left|\frac{\partial V_{i,\rho}(\pi_\lambda^{\mathbb{A}})}{\partial\lambda} - \frac{\partial V_{i,\rho}(\pi_\lambda^{\mathbb{B}})}{\partial\lambda}\right| = \left|\frac{\partial V_{i,\rho}(\pi_\lambda^{\mathbb{A}}) - V_{i,\rho}(\pi_\lambda^{\mathbb{B}})}{\partial\lambda}\right| \tag{E.52}$$

$$= \left|\frac{\partial\left(V_{i,\rho}(\pi_\lambda^{\mathbb{A}}) - V_{i,\rho}(\pi_\lambda^{\mathbb{B}})\right)}{\partial\lambda}\right| \tag{E.53}$$

$$= \left|\frac{\partial\left(Z_\rho^{\pi_\lambda^{\mathbb{A}}}\,\mathbb{E}_{s\sim d_\rho^{\pi_\lambda^{\mathbb{A}}}}\mathbb{E}_{\alpha\sim\pi_\lambda^{\mathbb{A}}(\cdot|s)}\left[A_i^{\pi_\lambda^{\mathbb{B}}}(s,\alpha)\right]\right)}{\partial\lambda}\right| \tag{E.54}$$

$$= \left|\frac{\partial\left(Z_\rho^{\pi_\lambda^{\mathbb{A}}}\,\mathbb{E}_{s\sim d_\rho^{\pi_\lambda^{\mathbb{A}}}}\mathbb{E}_{\alpha\sim\pi_\lambda^{\mathbb{A}}(\cdot|s)}\left[A_i^{\pi_\lambda^{\mathbb{B}}}(s,\alpha)\right]\right)}{\partial\lambda}\right| \tag{E.55}$$

$$= \left|\frac{\partial\left(Z_\rho^{\pi_\lambda^{\mathbb{A}}}\sum_{s,\alpha}d_\rho^{\pi_\lambda^{\mathbb{A}}}(s)(\pi_\lambda^{\mathbb{A}} - \pi_\lambda^{\mathbb{B}})(\alpha \mid s)A_i^{\pi_\lambda^{\mathbb{B}}}(s,\alpha)\right)}{\partial\lambda}\right| \tag{E.56}$$

$$
= \left| \frac{\partial \left( Z_\rho^{\pi_\lambda^{\mathbb{A}}} \sum_{s,\alpha} d_\rho^{\pi_\lambda^{\mathbb{A}}}(s)(\pi_\lambda^{\mathbb{A}} - \pi_\lambda^{\mathbb{B}})(\alpha \mid s) Q_i^{\pi_\lambda^{\mathbb{B}}}(s,\alpha) \right)}{\partial \lambda} \right| \tag{E.57}
$$

$$
= \left| \frac{\partial \left( \sum_{s,\alpha} \widetilde{d}_\rho^{\pi_\lambda^{\mathbb{A}}}(s)(\pi_\lambda^{\mathbb{A}} - \pi_\lambda^{\mathbb{B}})(\alpha \mid s) Q_i^{\pi_\lambda^{\mathbb{B}}}(s,\alpha) \right)}{\partial \lambda} \right| \tag{E.58}
$$

where (E.54) leverages the Performance Lemma E.6 and (E.56) uses the fact $\sum_{\alpha \in \mathcal{A}} \pi(\alpha \mid s) A_i^\pi(s,\alpha) =$, for all $s \in \mathcal{S}$ and the last one is derived by the definition $d_\rho^\pi(s) := \widetilde{d}_\rho^\pi(s)/Z_\rho^\pi$.

By triangular inequality, the linearity of $\partial$ operator and Lemma E.1, we have:

$$
\begin{aligned}
\left| \frac{\partial (V_{i,\rho}(\pi_\lambda^{\mathbb{A}}) - V_{i,\rho}(\pi_\lambda^{\mathbb{B}}))}{\partial \lambda} \Big|_{\lambda=0} \right| \leq & \left| \sum_{s,\alpha} \frac{\partial \widetilde{d}_\rho^{\pi_\lambda^{\mathbb{A}}}(s)}{\partial \lambda} \Big|_{\lambda=0} (\pi - \pi^{'})(\alpha \mid s) Q_i^{\pi^{'}}(s,\alpha) \right| \\
& + Z_\rho^{\pi^{\mathbb{A}}} \left| \sum_{s,\alpha} d_\rho^\pi(s) \frac{\partial (\pi_\lambda^{\mathbb{A}} - \pi_\lambda^{\mathbb{B}})(\alpha \mid s)}{\partial \lambda} \Big|_{\lambda=0} Q_i^{\pi^{'}}(s,\alpha) \right| \\
& + Z_\rho^{\pi^{\mathbb{A}}} \left| \sum_{s,\alpha} d_\rho^\pi(s) (\pi - \pi^{'})(\alpha \mid s) \frac{\partial Q_i^{\pi_\lambda^{\mathbb{B}}}(s,\alpha)}{\partial \lambda} \Big|_{\lambda=0} \right|
\end{aligned} \tag{E.59}
$$

We will bound the following three terms separately:

$$
\begin{cases}
\text{Term}_A = \left| \sum_{s,\alpha} \frac{\partial \widetilde{d}_\rho^{\pi_\lambda^{\mathbb{A}}}(s)}{\partial \lambda} \Big|_{\lambda=0} (\pi - \pi^{'})(\alpha \mid s) Q_i^{\pi^{'}}(s,\alpha) \right| \\
\text{Term}_B = \left| \sum_{s,\alpha} d_\rho^\pi(s) \frac{\partial (\pi_\lambda^{\mathbb{A}} - \pi_\lambda^{\mathbb{B}})(\alpha \mid s)}{\partial \lambda} \Big|_{\lambda=0} Q_i^{\pi^{'}}(s,\alpha) \right| \\
\text{Term}_C = \left| \sum_{s,\alpha} d_\rho^\pi(s) (\pi - \pi^{'})(\alpha \mid s) \frac{\partial Q_i^{\pi_\lambda^{\mathbb{B}}}(s,\alpha)}{\partial \lambda} \Big|_{\lambda=0} \right|
\end{cases}
$$

For $\text{Term}_A$, we will use Lemma E.3 in order to compute compactly the derivative:

$$
\frac{\partial \widetilde{d}_\rho^{\pi_\lambda^{\mathbb{A}}}(s)}{\partial \lambda} = \frac{\partial \left( \left[ \sum_{s' \in \mathcal{S}} \rho(s') \sum_{\alpha' \in \mathcal{A}} \pi_\lambda^{\mathbb{A}}(\alpha' \mid s') e_{s',\alpha'} \right]^\top \mathcal{T}(\pi_\lambda^{\mathbb{A}}) \sum_{\alpha \in \mathcal{A}} e_{s,\alpha} \right)}{\partial \lambda} \tag{E.60}
$$

$$
\begin{aligned}
= & \left( \left[ \sum_{s' \in \mathcal{S}} \rho(s') \sum_{\alpha' \in \mathcal{A}} \frac{\partial \pi_\lambda^{\mathbb{A}}(\alpha' \mid s')}{\partial \lambda} e_{s',\alpha'} \right]^\top \mathcal{T}(\pi_\lambda^{\mathbb{A}}) \sum_{\alpha \in \mathcal{A}} e_{s,\alpha} \right) \\
& + \left( \left[ \sum_{s' \in \mathcal{S}} \rho(s') \sum_{\alpha' \in \mathcal{A}} \pi_\lambda^{\mathbb{A}}(\alpha' \mid s') e_{s',\alpha'} \right]^\top \frac{\partial \mathcal{T}(\pi_\lambda^{\mathbb{A}})}{\partial \lambda} \sum_{\alpha \in \mathcal{A}} e_{s,\alpha} \right)
\end{aligned} \tag{E.61}
$$

$$
\begin{aligned}
= & \left( \left[ \sum_{s' \in \mathcal{S}} \rho(s') \sum_{\alpha' \in \mathcal{A}} \text{pert}(\alpha_i' \mid s') \cdot \pi_{-i}(\alpha_{-i}' \mid s') e_{s',\alpha'} \right]^\top \mathcal{T}(\pi_\lambda^{\mathbb{A}}) \sum_{\alpha \in \mathcal{A}} e_{s,\alpha} \right) \\
& + \left( \left[ \sum_{s' \in \mathcal{S}} \rho(s') \sum_{\alpha' \in \mathcal{A}} \pi_\lambda^{\mathbb{A}}(\alpha' \mid s') e_{s',\alpha'} \right]^\top \frac{\partial (I - \mathcal{P}(\pi_\lambda^{\mathbb{A}}))^{-1}}{\partial \lambda} \sum_{\alpha \in \mathcal{A}} e_{s,\alpha} \right)
\end{aligned} \tag{E.62}
$$

$$
\begin{aligned}
= & \left( \left[ \sum_{s' \in \mathcal{S}} \rho(s') \sum_{\alpha' \in \mathcal{A}} \text{pert}(\alpha_i' \mid s') \cdot \pi_{-i}(\alpha_{-i}' \mid s') e_{s',\alpha'} \right]^\top \mathcal{T}(\pi_\lambda^{\mathbb{A}}) \sum_{\alpha \in \mathcal{A}} e_{s,\alpha} \right) \\
& + \left( \left[ \sum_{s' \in \mathcal{S}} \rho(s') \sum_{\alpha' \in \mathcal{A}} \pi_\lambda^{\mathbb{A}}(\alpha' \mid s') e_{s',\alpha'} \right]^\top (\mathcal{T}(\pi_\lambda^{\mathbb{A}}) \frac{\partial \mathcal{P}(\pi_\lambda^{\mathbb{A}})}{\partial \lambda} \mathcal{T}(\pi_\lambda^{\mathbb{A}})) \sum_{\alpha \in \mathcal{A}} e_{s,\alpha} \right)
\end{aligned} \tag{E.63}
$$

Thus for $\lambda = 0$, we get

$$
\frac{\partial \widetilde{d}_\rho^{\pi_\lambda^{\mathbb{A}}}(s)}{\partial \lambda} \Big|_{\lambda=0} = \left( \left[ \sum_{s' \in \mathcal{S}} \rho(s') \sum_{\alpha' \in \mathcal{A}} \text{pert}(\alpha_i' \mid s') \cdot \pi_{-i}(\alpha_{-i}' \mid s') e_{s',\alpha'} \right]^\top \mathcal{T}(\pi) \sum_{\alpha \in \mathcal{A}} e_{s,\alpha} \right)
$$

$$+\left(\left[\sum_{s'\in\mathcal{S}}\rho(s')\sum_{\alpha'\in\mathcal{A}}\pi(\alpha'\mid s')e_{s',\alpha'}\right]^{\top}(\mathcal{T}(\pi)\frac{\partial\mathcal{P}(\pi_\lambda^{\mathbb{A}})}{\partial\lambda}\Big|_{\lambda=0}\mathcal{T}(\pi))\sum_{\alpha\in\mathcal{A}}e_{s,\alpha}\right) \tag{E.64}$$

Notice that $\left[\frac{\partial\mathcal{P}(\pi_\lambda^{\mathbb{A}})}{\partial\lambda}\Big|_{\lambda=0}\right]_{(s^\circ,\alpha^\circ)\to(s^\star,\alpha^\star)} = \mathrm{pert}(\alpha_i^\star\mid s^\star)\cdot\pi_{-i}(\alpha_{-i}^\star\mid s^\star)P(s^\star|s^\circ,\alpha^\circ)$.

To compactify the notation let us call $\mathrm{aux}_A := \left[\sum_{s'\in\mathcal{S}}\rho(s')\sum_{\alpha'\in\mathcal{A}}\mathrm{pert}(\alpha_i'\mid s')\cdot\pi_{-i}(\alpha_{-i}'\mid s')e_{s',\alpha'}\right]$, $\mathrm{aux}_B := [\sum_{s'\in\mathcal{S}}\rho(s')\sum_{\alpha'\in\mathcal{A}}\pi(\alpha'\mid s')e_{s',\alpha'}]$ and $\mathrm{aux}_C(s) := \sum_{\alpha\in\mathcal{A}}e_{s,\alpha}$.

Then, we get that:

$$\mathrm{Term}_A = \left|\sum_{s,\alpha}\frac{\partial\widetilde{d}_\rho^{\pi_\lambda^{\mathbb{A}}}(s)}{\partial\lambda}\Big|_{\lambda=0}(\pi_{\prime}-\pi^{'})(\alpha\mid s)Q_i^{\pi_{s'}^{'}}(s,\alpha)\right| \tag{E.65}$$

$$= \left|\sum_{s,\alpha}\left(\mathrm{aux}_A^\top\mathcal{T}(\pi)\mathrm{aux}_C(s)+\mathrm{aux}_B^\top(\mathcal{T}(\pi)\frac{\partial\mathcal{P}(\pi_\lambda^{\mathbb{A}})}{\partial\lambda}\Big|_{\lambda=0}\mathcal{T}(\pi))\mathrm{aux}_C(s)\right)(\pi-\pi^{'})(\alpha\mid s)Q_i^{\pi^{'}}(s,\alpha)\right| \tag{E.66}$$

$$= \left|\left(\mathrm{aux}_A^\top\mathcal{T}(\pi)+\mathrm{aux}_B^\top(\mathcal{T}(\pi)\frac{\partial\mathcal{P}(\pi_\lambda^{\mathbb{A}})}{\partial\lambda}\Big|_{\lambda=0}\mathcal{T}(\pi))\right)\underbrace{\sum_{s,\alpha}(\pi-\pi^{'})(\alpha\mid s)Q_i^{\pi^{'}}(s,\alpha)\mathrm{aux}_C(s)}_{\mathrm{aux}_D}\right| \tag{E.67}$$

$$\leq \|\mathrm{aux}_A\|_1\|\mathcal{T}(\pi)\mathrm{aux}_D\|_\infty+\|\mathrm{aux}_B\|_1\|(\mathcal{T}(\pi)\frac{\partial\mathcal{P}(\pi_\lambda^{\mathbb{A}})}{\partial\lambda}\Big|_{\lambda=0}\mathcal{T}(\pi))\mathrm{aux}_D\|_\infty \tag{E.68}$$

It is easy to see that $\|\mathrm{aux}_A\|_1 \leq \sqrt{|\mathcal{A}_i|}$, $\|\mathrm{aux}_B\|_1 = 1$. Indeed,

$$\|\mathrm{aux}_A\|_1 = \sum_{s'\in\mathcal{S}}\rho(s')\sum_{\alpha'\in\mathcal{A}}|\mathrm{pert}(\alpha_i'\mid s')|\cdot\pi_{-i}(\alpha_{-i}'\mid s') = \sum_{s'\in\mathcal{S}}\rho(s')\sum_{\alpha_i'\in\mathcal{A}_i}|\mathrm{pert}(\alpha_i'\mid s')|$$

$$= \sum_{s'\in\mathcal{S}}\rho(s')\|\mathrm{pert}_{i|s'}\|_1 \leq \sum_{s'\in\mathcal{S}}\rho(s')\sqrt{|\mathcal{A}_i|}\|\mathrm{pert}_{i|s'}\|_2 \leq \sqrt{|\mathcal{A}_i|} \tag{E.69}$$

$$\|\mathrm{aux}_B\|_1 = \sum_{s'\in\mathcal{S}}\rho(s')\sum_{\alpha'\in\mathcal{A}}\pi(\alpha'\mid s') = 1 \tag{E.70}$$

Additionally by Conversion Lemma in Matrix form (See Lemma E.2), we have that:

$$\|\mathcal{T}(\pi)x\|_\infty = \max_{s,\alpha}|e_{s,\alpha}^\top\mathcal{T}(\pi)x| = \max_{s,\alpha}|\mathbb{E}_{\tau\sim\mathrm{MDP}}\left[\sum_{t=0}^{T(\tau)}x(s_t,\alpha_t)\mid\alpha_0=\alpha,s_0=s\right]| \leq \frac{1}{\zeta}\|x\|_\infty \tag{E.71}$$

Similarly, for the matrix $\frac{\partial\mathcal{P}(\pi_\lambda^{\mathbb{A}})}{\partial\lambda}\Big|_{\lambda=0}$, we have that

$$\|\frac{\partial\mathcal{P}(\pi_\lambda^{\mathbb{A}})}{\partial\lambda}\Big|_{\lambda=0}x\|_\infty = \max_{s,\alpha}|e_{s,\alpha}^\top\frac{\partial\mathcal{P}(\pi_\lambda^{\mathbb{A}})}{\partial\lambda}\Big|_{\lambda=0}x| = \max_{s,\alpha}|\sum_{s',\alpha'}\mathrm{pert}(\alpha_i'\mid s')\cdot\pi_{-i}(\alpha_{-i}'\mid s')P(s'|s,\alpha)x_{s',\alpha'}|$$

$$\leq \sum_{s',\alpha'}|\mathrm{pert}(\alpha_i'\mid s')|\cdot\pi_{-i}(\alpha_{-i}'\mid s')P(s'|s,\alpha) \leq \sqrt{|\mathcal{A}_i|}\|\mathrm{pert}_{i|s'}\|_2\|x\|_\infty \leq \sqrt{|\mathcal{A}_i|}\|x\|_\infty \tag{E.72}$$

since $\|\mathrm{pert}\|_2 = 1$. Then, using (E.72) and (E.71) in (E.68) we get that :

$$\mathrm{Term}_A \leq \frac{\sqrt{|\mathcal{A}_i|}}{\zeta}\|\mathrm{aux}_D\|_\infty+\frac{\sqrt{|\mathcal{A}_i|}}{\zeta^2}\|\mathrm{aux}_D\|_\infty \tag{E.73}$$

$$\leq \frac{\sqrt{|\mathcal{A}_i|}}{\zeta}(1+\frac{1}{\zeta})\left\|\sum_{s,\alpha}(\pi-\pi^{'})(\alpha\mid s)Q_i^{\pi^{'}}(s,\alpha)\mathrm{aux}_C(s)\right\|_\infty \tag{E.74}$$

$$\leq \frac{\sqrt{|\mathcal{A}_i|}}{\zeta^2}(1+\frac{1}{\zeta})\max_s\left|\sum_\alpha(\pi-\pi^{'})(\alpha\mid s)\right|\|\mathrm{aux}_C(s)\|_\infty \tag{E.75}$$

$$\leq \frac{\sqrt{|\mathcal{A}_i|}}{\zeta^2}(1+\frac{1}{\zeta})\sum_{j=1}^{N}\sqrt{|\mathcal{A}_i|}\|\pi_j-\pi'_j\| \leq \frac{\sqrt{|\mathcal{A}_i|}}{\zeta^3}\sum_{j=1}^{N}\sqrt{|\mathcal{A}_i|}\|\pi_j-\pi'_j\| \tag{E.76}$$

where we used above the fact that $Q$ function is bounded by $1/\zeta$, $\|\text{pert}\|=1$ and the proposition E.1 to bound the difference of the policy profiles.

For the $\text{Term}_B$, we have that:

$$\text{Term}_B = \left|\sum_{s,\alpha} d_\rho^\pi(s)\frac{\partial(\pi_\lambda^{\mathbb{A}}-\pi_\lambda^{\mathbb{B}})(\alpha\mid s)}{\partial\lambda}\Big|_{\lambda=0} Q_i^{\pi'}(s,\alpha)\right| \tag{E.77}$$

$$= \left|\sum_{s,\alpha} d_\rho^\pi(s)\text{pert}(\alpha_i\mid s)(\pi_{-i}-\pi'_{-i})(\alpha\mid s)Q_i^{\pi'}(s,\alpha)\right| \tag{E.78}$$

$$\leq \frac{1}{\zeta}\left|\sum_{s} d_\rho^\pi(s)\sum_{\alpha_i}\text{pert}(\alpha_i\mid s)\sum_{\alpha_{-i}}(\pi_{-i}-\pi'_{-i})(\alpha\mid s)\right| \tag{E.79}$$

$$\leq \frac{1}{\zeta}\sum_{s}\left|d_\rho^\pi(s)\max_s\sum_{\alpha_i}|\text{pert}(\alpha_i\mid s)|\sum_{\alpha_{-i}}(\pi_{-i}-\pi'_{-i})(\alpha\mid s)\right| \tag{E.80}$$

$$\leq \frac{1}{\zeta}\max_s\|\text{pert}_{i|s}\|_1\sum_s d_\rho^\pi(s)\sum_{\alpha_{-i}}|(\pi_{-i}-\pi'_{-i})(\alpha\mid s)| \tag{E.81}$$

$$\leq \frac{\sqrt{|\mathcal{A}_i|}}{\zeta}\max_s\|\text{pert}_{i|s}\|_2(\sum_s d_\rho^\pi(s)\sum_{\alpha_{-i}}|(\pi_{-i}-\pi'_{-i})(\alpha\mid s)|) \tag{E.82}$$

$$\leq \frac{\sqrt{|\mathcal{A}_i|}}{\zeta}\sum_{j\in\mathcal{N}\setminus\{i\}}\sqrt{|\mathcal{A}_i|}\|\pi_j-\pi'_j\| \leq \frac{\sqrt{|\mathcal{A}_i|}}{\zeta}\sum_{j=1}^{N}\sqrt{|\mathcal{A}_i|}\|\pi_j-\pi'_j\| \tag{E.83}$$

where we used again the fact that $Q$ function is bounded by $1/\zeta$ and the proposition E.1 to bound the difference of the policy profiles.

For the $\text{Term}_C$, we get that:

$$\text{Term}_C = \left|\sum_{s,\alpha} d_\rho^\pi(s)(\pi-\pi')(\alpha\mid s)\frac{\partial Q_i^{\pi_\lambda^{\mathbb{B}}}(s,\alpha)}{\partial\lambda}\Big|_{\lambda=0}\right| \tag{E.84}$$

$$\leq \max_{s,\alpha}\left|\frac{\partial Q_i^{\pi_\lambda^{\mathbb{B}}}(s,\alpha)}{\partial\lambda}\Big|_{\lambda=0}\right|\sum_{s,\alpha} d_\rho^\pi(s)\big|(\pi-\pi')(\alpha\mid s)\big| \tag{E.85}$$

$$\leq \max_{s,\alpha}\left|\frac{\partial Q_i^{\pi_\lambda^{\mathbb{B}}}(s,\alpha)}{\partial\lambda}\Big|_{\lambda=0}\right|\sum_{j=1}^{N}\sqrt{|\mathcal{A}_i|}\|\pi_j-\pi'_j\| \tag{E.86}$$

$$\leq \max_{s,\alpha}\left|e_{s,\alpha}^\top\frac{\partial\mathcal{T}(\pi_\lambda^{\mathbb{B}})}{\partial\lambda}\Big|_{\lambda=0}r_i\right|\sum_{j=1}^{N}\sqrt{|\mathcal{A}_i|}\|\pi_j-\pi'_j\| \tag{E.87}$$

$$\leq \max_{s,\alpha}\left|e_{s,\alpha}^\top\frac{\partial(I-\mathcal{P}(\pi_\lambda^{\mathbb{A}}))^{-1}}{\partial\lambda}\Big|_{\lambda=0}r_i\right|\sum_{j=1}^{N}\sqrt{|\mathcal{A}_i|}\|\pi_j-\pi'_j\| \tag{E.88}$$

$$\leq \max_{s,\alpha}\left|e_{s,\alpha}^\top(\mathcal{T}(\pi)\frac{\partial\mathcal{P}(\pi_\lambda^{\mathbb{A}})}{\partial\lambda}\Big|_{\lambda=0}\mathcal{T}(\pi))r_i\right|\sum_{j=1}^{N}\sqrt{|\mathcal{A}_i|}\|\pi_j-\pi'_j\| \tag{E.89}$$

$$\leq \frac{\sqrt{|\mathcal{A}_i|}}{\zeta^2}\sum_{j=1}^{N}\sqrt{|\mathcal{A}_i|}\|\pi_j-\pi'_j\| \tag{E.90}$$

using again (E.72) and (E.71) and proposition E.1. Thus, we are ready now to bound the gradient per player:

$$\left|\frac{\partial(V_{i,\rho}(\pi_\lambda^{\mathbb{A}})-V_{i,\rho}(\pi_\lambda^{\mathbb{B}}))}{\partial\lambda}\Big|_{\lambda=0}\right| \leq \text{Term}_A + Z_\rho^{\pi^{\mathbb{A}}}(\text{Term}_B+\text{Term}_C) \leq \frac{3\sqrt{|\mathcal{A}_i|}}{\zeta^3}\sum_{j=1}^{N}\sqrt{|\mathcal{A}_i|}\|\pi_j-\pi'_j\|$$

where we recall that $Z_\rho^{\pi^{\mathrm{A}}} \leq \frac{1}{\zeta}$ Since we prove it for an arbitrary perturbation vector pert for the directional derivative, for the independent player's policy gradient it holds also that:

$$\|v_i(\pi) - v_i(\pi')\| = \|\nabla_i(V_{i,\rho}(\pi) - \nabla_i(V_{i,\rho}(\pi^{'})\| \leq \frac{3\sqrt{|\mathcal{A}_i|}}{\zeta^3} \sum_{j=1}^{N} \sqrt{|\mathcal{A}_i|}\|\pi_j - \pi_j'\| \quad \forall i \in \mathcal{N}$$

Finally for the concatenated gradient operator we get:

$$\|v(\pi) - v(\pi')\| = \sqrt{\sum_{i\in\mathcal{N}} \|v_i(\pi) - v_i(\pi')\|^2} = \sqrt{\sum_{i\in\mathcal{N}} \|\nabla_i(V_{i,\rho}(\pi) - \nabla_i(V_{i,\rho}(\pi')\|^2} \tag{E.91}$$

$$\leq \sqrt{\sum_{i\in\mathcal{N}} \frac{9|\mathcal{A}_i|}{\zeta^6}(\sum_{j\in\mathcal{N}} \sqrt{|\mathcal{A}_i|}\|\pi_j - \pi_j'\|)^2} \leq \sqrt{\sum_{i\in\mathcal{N}} \frac{9|\mathcal{A}_i|}{\zeta^6} \sum_{j\in\mathcal{N}} |\mathcal{A}_i| \sum_{j\in\mathcal{N}} \|\pi_j - \pi_j'\|^2} \tag{E.92}$$

$$\leq \frac{3}{\zeta^3} \sqrt{(\sum_{i\in\mathcal{N}} |\mathcal{A}_i|)^2 \|\pi - \pi'\|^2} \leq \frac{3|\mathcal{A}|}{\zeta^3}\|\pi - \pi'\| \tag{E.93}$$

■

# F Statistics of Reinforce

> Let's first recall our notation: We will write $\nabla_i$ to denote the gradient of the quantity in question with respect to $\pi_i$, i.e., when $\pi_{-i}$ is kept fixed and only $\pi_i$ is varied. For concision, we will write $v_i(\pi) = \nabla_i V_{i,\rho}(\pi)$ for the individual gradient of player $i$'s value function, and $v(\pi) = (v_i(\pi))_{i\in\mathcal{N}}$ for the ensemble thereof. Below we present two fundamental properties of Reinforce Policy Gradient estimator that we will utilize later in the our analysis.
>
> - Reinforce is an unbiased estimator of $v(\pi)$.
> - Reinforce's variance is bounded by $\mathcal{O}(1/\min_{s\in\mathcal{S},\alpha_i\in\mathcal{A}_i} \pi_i(\alpha_i|s))$ for each $i \in \mathcal{N}$.

**Lemma 4.** *Suppose that each agents $i \in \mathcal{N}$ follows a stationary policy $\pi_i \in \Pi_i$. Then, letting $\kappa_i = \min_{s\in\mathcal{S},\alpha_i\in\mathcal{A}_i} \pi_i(\alpha_i|s)$ for each $i \in \mathcal{N}$, we have*

*a)* $\mathbb{E}_{\tau\sim\mathrm{MDP}}[\textsc{Reinforce}(\pi)] = v(\pi).$ (12a)

*b)* $\mathbb{E}_{\tau\sim\mathrm{MDP}}\Big[\|\textsc{Reinforce}_i(\pi) - v_i(\pi)\|^2\Big] \leq \frac{24|\mathcal{A}_i|}{\kappa_i\zeta^4}.$ (12b)

*Proof.* In order to prove $\mathbb{E}_{\tau\sim\mathrm{MDP}}[\textsc{Reinforce}(\pi)] = v(\pi)$, it is equivalent to prove that

$$\mathbb{E}_{\tau\sim\mathrm{MDP}}[\textsc{Reinforce}_i(\pi)] = v_i(\pi) \text{ for each } i \in \mathcal{N}.$$

Without loss of generality let's assume that MDP ≡ MDP($\pi \mid \rho$) for some initial state distribution $\rho$. Additionally, we denote $\mathbb{P}^\pi(\tau)$ the induced probability of a random trajectory $\tau = (s_t, \alpha_t, r_t)_{t\leq T(\tau)}$.

$$\mathbb{E}_{\tau\sim\mathrm{MDP}}[\hat{v}_i] = \mathbb{E}_{\tau\sim\mathrm{MDP}}[R_i(\tau)\cdot\Lambda_i(\tau)] = \sum_{\tau\in\mathcal{T}} \mathbb{P}^\pi(\tau)R_i(\tau)\cdot\Lambda_i(\tau) \tag{F.1}$$

$$= \sum_{\tau\in\mathcal{T}} \mathbb{P}^\pi(\tau)R_i(\tau)\cdot[\sum_{t=0}^{T(\tau)} \nabla_i(\log\pi_i(a_{i,t}|s_t))] \tag{F.2}$$

$$= \sum_{\tau\in\mathcal{T}} \mathbb{P}^\pi(\tau)R_i(\tau)\cdot\nabla_i\left[\sum_{t=0}^{T(\tau)} \log\pi_i(a_{i,t}|s_t)\right] \tag{F.3}$$

$$\begin{aligned} = &\sum_{\tau\in\mathcal{T}} \mathbb{P}^\pi(\tau)R_i(\tau)\nabla_i \sum_{t=0}^{T(\tau)} \log\pi_i(a_{i,t}|s_t) \\ &+ \sum_{\tau\in\mathcal{T}} \mathbb{P}^\pi(\tau)R_i(\tau)\left(\nabla_i \sum_{j\in\mathcal{N}\setminus\{i\}} \sum_{t=0}^{T(\tau)} \log\pi_j(\alpha_{j,t}|s_t) + \nabla_i \sum_{t=0}^{T(\tau)} \log\mathbb{P}(s_t \mid s_{t-1}, a_{t-1})\right) \\ &+ \sum_{\tau\in\mathcal{T}} \mathbb{P}^\pi(\tau)R_i(\tau)\nabla_i \log\rho(s_0) \end{aligned} \tag{F.4}$$

$$= \sum_{\tau\in\mathcal{T}} \mathbb{P}^\pi(\tau) R_i(\tau)\nabla_i(\log \mathbb{P}^\pi(\tau)) = \sum_{\tau\in\mathcal{T}} (\nabla_i \mathbb{P}^\pi(\tau)) R_i(\tau) = \nabla_i(\sum_{\tau\in\mathcal{T}} \mathbb{P}^\pi(\tau) R_i(\tau)) \tag{F.5}$$

$$= \nabla_i V_{i,\rho}(\pi) \tag{F.6}$$

where in the second to last inequality we used the definition for the derivative of the logarithm. We also note here that

$$\mathbb{E}_{\tau\sim\text{MDP}}[\hat{v}_i] = \mathbb{E}_{\tau\sim\text{MDP}}[R_i(\tau)\nabla_i(\log \mathbb{P}^\pi(\tau))] \tag{F.7}$$

For the variance of REINFORCE estimator we have that

$$\begin{aligned}\mathbb{E}_{\tau\sim\text{MDP}}\Big[\|\text{REINFORCE}_i(\pi) - v_i(\pi)\|^2\Big] = &\ \mathbb{E}_{\tau\sim\text{MDP}}\Big[\|\text{REINFORCE}_i(\pi)\|^2\Big] \\ &- 2\,\mathbb{E}_{\tau\sim\text{MDP}}[\langle \text{REINFORCE}_i(\pi), v_i(\pi)\rangle] \\ &+ \mathbb{E}_{\tau\sim\text{MDP}}\Big[\|v_i(\pi)\|^2\Big]\end{aligned}$$

or equivalently $\mathbb{E}_{\tau\sim\text{MDP}}\Big[\|\text{REINFORCE}_i(\pi) - v_i(\pi)\|^2\Big] = \mathbb{E}_{\tau\sim\text{MDP}}\Big[\|\text{REINFORCE}_i(\pi)\|^2\Big] - \mathbb{E}_{\tau\sim\text{MDP}}\Big[\|v_i(\pi)\|^2\Big]$. Therefore, we have that

$$\mathbb{E}_{\tau\sim\text{MDP}}\Big[\|\text{REINFORCE}_i(\pi) - v_i(\pi)\|^2\Big] \le \mathbb{E}_{\tau\sim\text{MDP}}\Big[\|\text{REINFORCE}_i(\pi)\|^2\Big] = \mathbb{E}[\|\hat{v}_i\|^2] \tag{F.8}$$

$$\mathbb{E}[\|\hat{v}_i\|^2] = \mathbb{E}_{\tau\sim\text{MDP}}[\|R_i(\tau)\Lambda_i(\tau)\|^2] \le \mathbb{E}_{\tau\sim\text{MDP}}[\|R_i(\tau)\|^2\|\Lambda_i(\tau)\|^2] \tag{F.9}$$

$$\le \mathbb{E}_{\tau\sim\text{MDP}}[(T(\tau)+1)^2\|\sum_{t=0}^{T(\tau)} \nabla_i \log \pi_i(a_{i,t}, s_t)\|^2] \tag{F.10}$$

$$\le \mathbb{E}_{\tau\sim\text{MDP}}[(T(\tau)+1)^3 \sum_{t=0}^{\infty} \sum_{s,a\in\mathcal{S}\times\mathcal{A}_i} \mathbb{1}\{t\le T\}\mathbb{1}\{s_t = s, a_{i,t} = a\}\|\nabla_i \log \pi_i(a,s)\|^2] \tag{F.11}$$

$$= \sum_{t=0}^{\infty} \sum_{s,a\in\mathcal{S}\times\mathcal{A}_i} \mathbb{E}_{\tau\sim\text{MDP}}[(T(\tau)+1)^3 \mathbb{1}\{t\le T\}\mathbb{1}\{s_t = s, a_{i,t} = a\}\frac{1}{(\pi_i(a,s))^2}] \tag{F.12}$$

$$\le \sum_{t=0}^{\infty} \sum_{s,a\in\mathcal{S}\times\mathcal{A}_i} \frac{1}{(\pi_i(a,s))^2}\, \mathbb{E}_{\tau\sim\text{MDP}}[(T(\tau)+1)^3 \mathbb{1}\{t\le T\}\mathbb{1}\{s_t = s, a_{i,t} = a\}] \tag{F.13}$$

$$\le \sum_{t=0}^{\infty} \sum_{s,a\in\mathcal{S}\times\mathcal{A}_i} \frac{1}{\pi_i(a,s)}\, \mathbb{E}_{\tau\sim\text{MDP}}[(T(\tau)+1)^3 \mathbb{1}\{t\le T\}\mathbb{1}\{s_t = s\}] \tag{F.14}$$

$$\le \sum_{t=0}^{\infty} \sum_{s,a\in\mathcal{S}\times\mathcal{A}_i} \frac{1}{\kappa_i}\{(T(\tau)+1)^3 \mathbb{1}\{t\le T\}\mathbb{1}\{s_t = s\}\} \tag{F.15}$$

$$= \sum_{t=0}^{\infty} \sum_{s\in\mathcal{S}} \frac{|A_i|}{\kappa_i}\, \mathbb{E}_{\tau\sim\text{MDP}}[(T(\tau)+1)^3 \mathbb{1}\{t\le T\}\mathbb{1}\{s_t = s\}] \tag{F.16}$$

$$= \frac{|A_i|}{\kappa_i}\, \mathbb{E}_{\tau\sim\text{MDP}}[(T(\tau)+1)^3 \sum_{t=0}^{T} \mathbb{1}\{t\le T\}] \tag{F.17}$$

$$\le \frac{|A_i|}{\kappa_i}\, \mathbb{E}_{\tau\sim\text{MDP}}[(T(\tau)+1)^4] \tag{F.18}$$

$$\le \frac{|A_i|}{\kappa_i} \sum_{t=0}^{\infty} (1-\zeta)^t \zeta (t+1)^4 \le \frac{24}{\zeta^4}\frac{|A_i|}{\kappa_i} \tag{F.19}$$

we note that to go from the first to the second inequality we used the boundeness by one of the rewards, while from the second to the third using Jensen's inequality. ■

# G Solution concepts

In this part, we will establish three important facts that certifies the leitmotif of our focus to variational optima. More precisely,

- In Lemma 2, we prove the crucial property of Gradient Dominance for the multi-agent random stopping setting.
- In Lemma 3, we establish that any stationary point corresponds to Nash Equilibria.
- In Proposition 1, we prove the "drift" inequalities for all the different types of stationary points.

Proposition 1 will be crucial to prove the corresponding rate of convergence at the following sections of the supplement

**Lemma 2.** [Gradient dominance property] *For any policy profile $\pi = (\pi_i)_{i\in\mathcal{N}} \in \Pi$, we have that*

$$V_{i,\rho}(\pi_i'; \pi_{-i}) - V_{i,\rho}(\pi_i; \pi_{-i}) \le \mathcal{C}_{\mathcal{G}} \max_{\bar{\pi}_i\in\Pi_i} \langle \nabla_i V_{i,\rho}(\pi), \bar{\pi}_i - \pi_i\rangle \tag{GDP}$$

*for any unilateral deviation $\pi_i' \in \Pi_i$ of each player $i \in \mathcal{N}$.*

*Proof.* We start by rewriting the LHS of the demanded expression using Performance Lemma E.6 and Conversion Lemma 1 for $\pi^{\mathbb{A}} = (\pi_i'; \pi_{-i})$ and $\pi^{\mathbb{B}} = (\pi_i; \pi_{-i})$:

$$\begin{aligned}
V_{i,\rho}(\pi^{\mathbb{A}}) - V_{i,\rho}(\pi^{\mathbb{B}}) &= \sum_{s\in\mathcal{S}} \tilde{d}_\rho^{\pi^{\mathbb{A}}}(s)\, \mathbb{E}_{\alpha\sim\pi^{\mathbb{A}}(\cdot|s)}\left[A_i^{\pi_i^{\mathbb{B}}}(s,\alpha)\right] && \text{(G.1)}\\
&= \sum_{s\in\mathcal{S}} \tilde{d}_\rho^{\pi^{\mathbb{A}}}(s) \sum_{a_i\in\mathcal{A}_i} \pi_i'(a_i|s) \sum_{a_{-i}\in\mathcal{A}_{-i}} \pi_{-i}(a_{-i}|s) A_i^{\pi^{\mathbb{B}}}(s,\alpha) && \text{(G.2)}\\
&= \sum_{s\in\mathcal{S}} \tilde{d}_\rho^{\pi^{\mathbb{A}}}(s) \sum_{a_i\in\mathcal{A}_i} \pi_i'(a_i|s) \overline{A}_i^{\pi^{\mathbb{B}}}(s,a_i) && \text{(G.3)}\\
&\le \sum_{s\in\mathcal{S}} \tilde{d}_\rho^{\pi^{\mathbb{A}}}(s) \sum_{a_i\in\mathcal{A}_i} \pi_i'(a_i|s) \max_{a_i\in\mathcal{A}_i} \overline{A}_i^{\pi^{\mathbb{B}}}(s,a_i) && \text{(G.4)}
\end{aligned}$$

$$\begin{aligned}
V_{i,\rho}(\pi^{\mathbb{A}}) - V_{i,\rho}(\pi^{\mathbb{B}}) &\le \max_{\tilde{\pi}_i\in\Delta(\mathcal{A})^{\mathcal{S}}} \sum_{s\in\mathcal{S}} \tilde{d}_\rho^{\pi^{\mathbb{A}}}(s) \sum_{a_i\in\mathcal{A}_i} \tilde{\pi}_i(a_i|s) \overline{A}_i^{\pi^{\mathbb{B}}}(s,a_i) && \text{(G.5)}\\
&\le \max_{\tilde{\pi}_i\in\Delta(\mathcal{A})^{\mathcal{S}}} \sum_{s\in\mathcal{S}} \tilde{d}_\rho^{\pi^{\mathbb{A}}}(s) \sum_{a_i\in\mathcal{A}_i} (\tilde{\pi}_i(a_i|s) - \pi_i(a_i|s)) \overline{A}_i^{\pi^{\mathbb{B}}}(s,a_i) && \text{(G.6)}\\
&\le \max_{\tilde{\pi}_i\in\Delta(\mathcal{A})^{\mathcal{S}}} \sum_{s\in\mathcal{S}} \frac{\tilde{d}_\rho^{\pi^{\mathbb{A}}}(s)}{\tilde{d}_\rho^{\pi^{\mathbb{B}}}(s)} \tilde{d}_\rho^{\pi^{\mathbb{B}}}(s) \sum_{a_i\in\mathcal{A}_i} (\tilde{\pi}_i(a_i|s) - \pi_i(a_i|s)) \overline{A}_i^{\pi^{\mathbb{B}}}(s,a_i) && \text{(G.7)}\\
&\le \left\| \frac{\tilde{d}_\rho^{\pi^{\mathbb{A}}}(s)}{\tilde{d}_\rho^{\pi^{\mathbb{B}}}(s)} \right\|_\infty \max_{\tilde{\pi}_i\in\Delta(\mathcal{A})^{\mathcal{S}}} \sum_{s\in\mathcal{S}} \sum_{a_i\in\mathcal{A}_i} \tilde{d}_\rho^{\pi^{\mathbb{B}}}(s) (\tilde{\pi}_i(a_i|s) - \pi_i(a_i|s)) \overline{Q}_i^{\pi^{\mathbb{B}}}(s,a_i) && \text{(G.8)}\\
&\le \left\| \frac{\tilde{d}_\rho^{\pi^{\mathbb{A}}}(s)}{\tilde{d}_\rho^{\pi^{\mathbb{B}}}(s)} \right\|_\infty \max_{\tilde{\pi}_i\in\Delta(\mathcal{A})^{\mathcal{S}}} \sum_{s\in\mathcal{S}, a_i\in\mathcal{A}_i} (\tilde{\pi}_i(a_i|s) - \pi_i(a_i|s)) \tilde{d}_\rho^{\pi^{\mathbb{B}}}(s) \overline{Q}_i^{\pi^{\mathbb{B}}}(s,a_i) && \text{(G.9)}\\
&\le \left\| \frac{\tilde{d}_\rho^{\pi^{\mathbb{A}}}(s)}{\tilde{d}_\rho^{\pi^{\mathbb{B}}}(s)} \right\|_\infty \max_{\tilde{\pi}_i\in\Delta(\mathcal{A})^{\mathcal{S}}} \sum_{s\in\mathcal{S}, a_i\in\mathcal{A}_i} (\tilde{\pi}_i(a_i|s) - \pi_i(a_i|s)) \frac{\partial V_{i,\rho}(\pi)}{\partial \pi_i(\alpha_i \mid s)} && \text{(G.10)}\\
V_{i,\rho}(\pi_i'; \pi_{-i}) - V_{i,\rho}(\pi_i; \pi_{-i}) &\le \mathcal{C}_{\mathcal{G}} \max_{\bar{\pi}_i\in\Pi_i} \langle \nabla_i V_{i,\rho}(\pi), \bar{\pi}_i - \pi_i\rangle && \text{(G.11)}
\end{aligned}$$

Notice that we have assumed that $\tilde{d}_\rho^{\pi^{\mathbb{B}}} > 0$. If this wasn't the case we could take a trivial bound of $\infty$.

∎

**Lemma 3.** [First-order stationary policies are Nash] *A profile $\pi^* = (\pi_i^*)_{i\in\mathcal{N}} \in \Pi$ is a Nash policy profile if and only if it satisfies the first-order stationary condition*

$$\langle v(\pi^*), \pi - \pi^*\rangle \le 0 \quad \textit{for all } \pi \in \Pi. \tag{FOS}$$

*Proof.* Let's apply the definition of first-order stationary point for the pair of policy profiles $\{\pi^*, \pi\}$: $\pi^* = (\pi_i^*, \pi_{-i}^*)$ and $\pi = (\pi_i, \pi_{-i}^*)$:

$$\begin{aligned}
\langle v(\pi^*), \pi^* - \pi\rangle \geq 0 \quad &\Leftrightarrow && \text{(G.12)}\\
\langle v(\pi^*), (\pi_i^*, \pi_{-i}^*) - (\pi_i, \pi_{-i}^*)\rangle \geq 0 \quad &\Leftrightarrow && \text{(G.13)}\\
\langle v(\pi^*), (\pi_i^* - \pi_i, 0)\rangle \geq 0 \quad &\Leftrightarrow && \text{(G.14)}\\
\langle v_i(\pi^*), \pi_i^* - \pi_i\rangle \geq 0 \quad &\Leftrightarrow && \text{(G.15)}\\
\langle \nabla_i V_{i,\rho}(\pi^*), \pi_i^* - \pi_i\rangle \geq 0 \quad &\Leftrightarrow && \text{(G.16)}\\
\min_{\bar{\pi}_i \in \Pi_i} \langle \nabla_i V_{i,\rho}(\pi^*), \pi_i^* - \bar{\pi}_i\rangle \geq 0 \quad &\Leftrightarrow && \text{(G.17)}\\
\max_{\bar{\pi}_i \in \Pi_i} \langle \nabla_i V_{i,\rho}(\pi^*), \pi_i - \bar{\pi}_i^*\rangle \leq 0 \quad &\Leftrightarrow && \text{(G.18)}\\
& && \text{(G.19)}
\end{aligned}$$

By Gradient Dominance Property and Lemma 2, we have that

$$V_{i,\rho}(\pi_i; \pi_{-i}^*) - V_{i,\rho}(\pi_i^*; \pi_{-i}^*) \leq \mathcal{C}_{\mathcal{G}} \max_{\bar{\pi}_i \in \Pi_i} \langle \nabla_i V_{i,\rho}(\pi^*), \bar{\pi}_i - \pi_i^*\rangle \leq 0 \Rightarrow \tag{G.20}$$

$$V_{i,\rho}(\pi_i; \pi_{-i}^*) \leq V_{i,\rho}(\pi_i^*; \pi_{-i}^*) \quad \forall \pi_i \in \Pi_i. \tag{G.21}$$

■

With all this in place, we are finally in a position to prove the characterization of second-order stationary and strict Nash policies that of Proposition 1. For ease of reference, we restate the relevant claims below.

**Proposition 1.** *Let $\pi^* = (\pi_i^*)_{i \in \mathcal{N}} \in \Pi$ be a Nash policy. Then:*

*a) If $\pi^*$ is second-order stationary, there exists some $\mu > 0$ such that*

$$\langle v(\pi), \pi - \pi^*\rangle \leq -\mu \, \|\pi - \pi^*\|^2 \qquad \textit{for all } \pi \textit{ sufficiently close to } \pi^*. \tag{3a}$$

*b) If $\pi^*$ is strict, there exists some $\mu > 0$ such that*

$$\langle v(\pi), \pi - \pi^*\rangle \leq -\mu \, \|\pi - \pi^*\| \qquad \textit{for all } \pi \textit{ sufficiently close to } \pi^*. \tag{3b}$$

*Proof.* We begin with the characterization of second-order stationary policies. To that end, let $d = |\mathcal{S}| \sum_i |\mathcal{A}_i|$ denote the ambient dimension of $\prod_i \big(\mathbb{R}^{\mathcal{A}_i}\big)^{\mathcal{S}}$ and consider the mapping $\varphi \colon \mathbb{R}^{d \times d} \to \mathbb{R}$ mapping $H \mapsto \max\{z^\top H z : z \in \mathrm{TC}(\pi^*), \|z\| = 1\}$. Clearly, $\varphi$ is convex as the pointwise maximum of a set of linear – and hence convex – functions. This in turn implies the continuity of $\varphi$ as every convex function is continuous on the interior of its effective domain. Since $\pi^*$ satisfies (SOS) by assumption, we have $\varphi(\mathrm{Jac}_v(\pi^*)) < 0$, so, by continuity and the convexity of $\Pi$, there exists some $\mu > 0$ and a convex neighborhood $\mathcal{U}$ of $\pi^*$ in $\Pi$ such that $\varphi(\mathrm{Jac}_v(\pi)) \leq -\mu$ for all $\pi \in \mathcal{U}$.

With this in mind, letting $z = \pi - \pi^* \in \mathrm{TC}(\pi^*)$ for some $\pi \in \mathcal{U}$, a straightforward Taylor expansion with integral remainder yields

$$v(\pi) - v(\pi^*) = \int_0^1 \mathrm{Jac}_v(\pi^* + \tau z) z \, d\tau \tag{G.22}$$

and hence, setting $\pi_\tau = \pi^* + \tau z$, we get

$$\begin{aligned}
\langle v(\pi) - v(\pi^*), \pi - \pi^*\rangle &= \int_0^1 z^\top \, \mathrm{Jac}_v(\pi_\tau) z \, d\tau \\
&\leq \|z\|^2 \int_0^1 \varphi(\mathrm{Jac}_v(\pi_\tau)) \, d\tau \leq -\mu \|z\|^2 = -\mu \|\pi - \pi^*\|^2
\end{aligned} \tag{G.23}$$

However, by (FOS), we have $\langle v(\pi^*), \pi - \pi^*\rangle \leq 0$ which, combined with the above, yields $\langle v(\pi), \pi - \pi^*\rangle \leq -\mu \|\pi - \pi^*\|^2$, as claimed.

For the second part of our lemma, pick some $\pi \neq \pi^*$ and let $z = (\pi - \pi^*)/\|\pi - \pi^*\|$, so $z \in \mathrm{TC}(\pi^*)$ and $\|z\| = 1$. Then, given that (FOS) is satisfied as a strict inequality for all $\pi \neq \pi^*$, we readily get $\langle v(\pi^*), z\rangle < 0$ for all $z \in \mathrm{TC}(\pi^*)$ with $\|z\| = 1$. Thus, by the joint continuity of the function $\langle v(\pi), z\rangle$ in $\pi$ and $z$, there exists a compact convex neighborhood $\mathcal{K}$ of $\pi^*$ in $\Pi$ such that $\mu \coloneqq \min\{\langle v(\pi), z\rangle : \pi \in \mathcal{K}, z \in \mathrm{TC}(\pi^*), \|z\| = 1\} < 0$. Thus, letting $z = (\pi - \pi^*)/\|\pi - \pi^*\|$ as above, we conclude that $\langle v(\pi), \pi - \pi^*\rangle \leq -\mu\|\pi - \pi^*\|$, as claimed. ■