State-Action Selection in General Stochastic Games under Regularized Policy Gradient Methods

Omar Abbadi[1,2*] **Rida Laraki**[1] **Panayotis Mertikopoulos**[2,3]

[1] Univ. Mohammed VI Polytechnic, MCGT, College of Computing, 11103 Rabat, Morocco
[2] Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France
[3] Archimedes, Athena Research Center
{omar.abbadi,rida.laraki}@um6p.ma, panayotis.mertikopoulos@imag.fr

Abstract

We study a general class of regularized policy gradient methods for episodic learning in general-sum stochastic games, with the aim of describing the long-run behavior of the induced policy iterates. In this challenging setting, learning even the simplest pointwise solution concepts is known to be intractable. This leads us to switch our focus to less restrictive *setwise* solution concepts, that is, rather than only asking whether the iterates converge to a specific policy, we ask which *sets of policies* are selected by the learning dynamics in the long-run. In one-shot games, this setwise perspective admits a crisp characterization, which is that a product set of pure strategies is asymptotically stable if and only if it is *closed under better replies* (club). We show that the direct statewise analogue of this principle *fails outright* in stochastic games because actions affect not only current rewards but also the future evolution of the game. To restore this principle in a transition-aware form, we introduce Q*-club* sets, namely state-action sets in which every *allowed* action has a strict Q-value advantage over every *excluded* action, thereby providing a setwise generalization of *strict Nash policies*. We then prove that Q-club sets are stochastically stable and attracting under the learning dynamics, even in the presence of noise and uncertainty. Finally, we establish local convergence rates to Q-club sets governed by the regularizer; in particular, entropic regularization yields exponential convergence, while Euclidean regularization yields finite-time convergence.

1 Introduction

Stochastic games, also known as *Markov games*, were introduced by Shapley [49] and have since become the classical model of strategic interaction in dynamic environments [32, 39, 44, 52], as well as the *de facto* framework of multi-agent reinforcement learning [11, 22, 37, 50, 51, 54]. In this setting, the players observe at each stage a common state and choose actions, which jointly determine their rewards and the law of the next state. Each decision thus shapes not only the current payoff, but also the very states in which future play will unfold.

This interplay between strategic and sequential decision making is precisely what makes learning so challenging. Indeed, players may have to forgo immediate rewards to reach more favorable states, as the environment itself moves with the policies of all players. Moreover, the game may be unknown to the players and, even when it is known, computing its equilibria is typically out of reach. In fact, stationary Nash equilibria of general-sum stochastic games with an infinite horizon,

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

although guaranteed to exist [23], are PPAD-hard to compute [29], and even stationary Markov *coarse correlated equilibria*, arguably the most accessible pointwise solution concept in games, are *computationally intractable* [20]. It is thus natural to study how uncoupled, reward-driven learning rules behave in the long run, rather than presuppose equilibrium convergence.

Existing guarantees in this direction, however, come at the cost of strong structural assumptions. For Q-learning and fictitious play methods, convergence is known in zero-sum, identical-interest, and weakly acyclic stochastic games [4, 6, 7, 17, 35, 47, 48]. For policy gradient and regularized schemes, it is known in Markov potential and common-interest games [21, 24, 34, 53, 64], and in zero-sum stochastic games, notably through optimistic and extragradient variants [15, 16, 19, 56, 59, 61, 63]. Beyond these classes, in general-sum stochastic games, many guarantees target weaker solution concepts, such as correlated or coarse correlated equilibria, often through averaged or empirical play and nonstationary policies [13, 20, 28, 38, 63]. Closest to our viewpoint, Giannou et al. [26] follow the sequence of policies, *episode by episode*, and prove that policy gradient converges locally to strict Nash policies.

While we also adopt this last-iterate viewpoint, we do not confine our attention to equilibrium points, and we ask instead a broader question, namely, *which actions remain viable, at each state, along the learning trajectory*. Indeed, learning may eliminate some actions without settling the strategic interaction altogether, as players may abandon persistently unfavorable actions while still adapting among the remaining ones. This brings to the fore the *state-action set* selected by the dynamics—a restriction to a subset of actions at each state—which may stabilize even when the policy itself does not. Such behavior escapes pointwise equilibrium concepts, yet it lies at the heart of the learning process, which leads us to the following question:

Can we guarantee that, at each state, some actions are consistently selected in the long run?
More precisely, which state-action sets are stable and attracting under learning?

In normal-form games, this question admits a crisp answer, which is provided by sets that are *closed under better replies* (club) of Ritzberger and Weibull [45]. These sets capture setwise strategic stability through self-enforcement, in the sense that every pure unilateral deviation leaving the set is costly to the deviator. Remarkably, these are precisely the sets that regularized learning stabilizes, as Boone and Mertikopoulos [12] proved that, in finite games, a product of pure strategies is club if and only if its span is asymptotically stable, a result extended and strengthened in continuous time [1, 14].

In stochastic games, however, this principle does not carry over verbatim. Since actions also steer the distribution of future states, self-enforcement can no longer be judged from stage payoffs alone, and any meaningful notion of stability must account for the transition structure of the game. With this in mind, our contributions can be summarized as follows.

1. We introduce the notion of sets that are *closed under Q-better replies* (Q-club), a *novel setwise solution concept* for stochastic games. These are state-action sets in which every excluded action is strictly worse, in terms of its Q-value, than every allowed action.
2. In the pointwise case, we show that, under full visitation, a Nash policy is variationally stable if and only if it is strict, which settles an open question raised by Giannou et al. [26]. Such policies are, moreover, stochastically asymptotically stable under policy gradient.
3. Through an example with two states, we show that the comparison of Q-values cannot be dispensed with. There, the candidate support is closed under one-shot better replies at each state, yet it is not dynamically stable, as an excluded action profitably redirects play toward a better state.
4. We prove that every Q-club set is stochastically stable and attracting under a broad class of regularized policy gradient methods, even in the presence of noise and uncertainty. This covers exact and stochastic gradients, optimistic and extragradient variants, and feedback from trajectories, thereby demonstrating robustness to the feedback model and to the choice of regularizer.
5. Finally, we derive fast local convergence rates to Q-club sets for a broad range of step size schedules. These rates are governed by the regularizer, with exponential decay in the cumulative step size under entropic regularization and convergence in finite time under Euclidean regularization.

Taken together, these results provide, to our knowledge, the first robust setwise stability concept for learning in stochastic games.

2 Preliminaries

Setup of the game. We work with finite *stochastic games with random stopping*. Formally, such a game is a tuple $\mathcal{G} = (\mathcal{S}, \mathcal{N}, (\mathcal{A}_i, r_i)_{i\in\mathcal{N}}, P, \rho)$, where $\mathcal{N} = \{1, \dots, N\}$ is the finite set of players, $\mathcal{S}$ the finite set of states, and $\mathcal{A}_i$ the finite action set of player $i \in \mathcal{N}$. We also write $\mathcal{A} = \prod_{i\in\mathcal{N}} \mathcal{A}_i$ and $\mathcal{A}_{-i} = \prod_{j\neq i} \mathcal{A}_j$, and we denote action profiles by $\alpha = (\alpha_i, \alpha_{-i}) \in \mathcal{A}$, where $\alpha_i \in \mathcal{A}_i$ is the action of player i and $\alpha_{-i} \in \mathcal{A}_{-i}$ that of their opponents. Moreover, each player has a bounded reward function $r_i \colon \mathcal{S} \times \mathcal{A} \to [-1, 1]$, where $r_i(s, \alpha_i, \alpha_{-i})$ is the reward of player i at state s when they choose α_i and their opponents choose α_{-i}. The map $P(\cdot \mid s, \alpha)$ is a substochastic transition kernel on $\mathcal{S}$, meaning that, after α is played at s, the next state is s' with probability $P(s' \mid s, \alpha)$, while the episode terminates with probability $\zeta_{s,\alpha} := 1 - \sum_{s'\in\mathcal{S}} P(s' \mid s, \alpha) \geq 0$. Throughout, we impose the uniform stopping condition $\zeta_{\min} := \min_{s\in\mathcal{S},\, \alpha\in\mathcal{A}} \zeta_{s,\alpha} > 0$. As a result, every episode ends almost surely, and its length is dominated, uniformly over all policies, by a geometric random variable, with finite moments of all orders. This model encompasses discounted stochastic games, in which the discount factor is read as a continuation probability, as well as stopped games [32, 39, 52].

Episodic play. Given an initial state distribution $\rho \in \Delta(\mathcal{S})$, an episode starts from $s_0 \sim \rho$. At each stage $t = 0, 1, 2, \dots$, all players observe the common state $s_t \in \mathcal{S}$, choose actions $\alpha_{i,t} \in \mathcal{A}_i$, and receive rewards $r_i(s_t, \alpha_t)$, where $\alpha_t = (\alpha_{i,t})_{i\in\mathcal{N}}$. Conditional on $(s_t, \alpha_t) = (s, \alpha)$, the episode stops with probability $\zeta_{s,\alpha}$ and otherwise moves to $s_{t+1} = s'$ with probability $P(s' \mid s, \alpha)$. Writing $r(s_t, \alpha_t) = (r_i(s_t, \alpha_t))_{i\in\mathcal{N}}$ and T for the (random) terminal time, the realized trajectory is $\tau = (s_0, \alpha_0, r(s_0, \alpha_0), \dots, s_T, \alpha_T, r(s_T, \alpha_T))$.

Policy space and score space. Each player i follows a *stationary Markov policy*, that is, a map $\pi_i \colon \mathcal{S} \to \Delta(\mathcal{A}_i)$, where $\pi_i(\alpha_i \mid s)$ is the probability of selecting α_i at state s. We write

$$\Pi_i = \Delta(\mathcal{A}_i)^{\mathcal{S}}, \qquad \Pi = \prod_{i\in\mathcal{N}} \Pi_i, \qquad \mathcal{Y}_i = \mathbb{R}^{\mathcal{S}\times\mathcal{A}_i}, \quad \text{and} \quad \mathcal{Y} = \prod_{i\in\mathcal{N}} \mathcal{Y}_i, \tag{1}$$

for the *policy space*, which hosts the policies of the players, and the *score space*, which hosts the reward signals that drive learning. To fix ideas once and for all, we endow Π with the ℓ_1 norm and $\mathcal{Y}$ with the dual ℓ_∞ norm, and we write $\operatorname{ri}\Pi$ for the relative interior of Π, that is, the set of policies with full action support at all states. Given a policy profile $\pi \in \Pi$, we write (π_i, π_{-i}) for the profile in which player i follows π_i and their opponents follow π_{-i}. Moreover, actions are drawn independently across players conditional on the state, so that $\pi(\alpha \mid s) = \prod_{i\in\mathcal{N}} \pi_i(\alpha_i \mid s)$ and $P^\pi(s' \mid s) = \sum_{\alpha\in\mathcal{A}} \pi(\alpha \mid s) P(s' \mid s, \alpha)$.

Values and policy gradient. We write $\mathbb{P}_s^\pi$ and $\mathbb{E}_s^\pi$ for probability and expectation under π when $s_0 = s$, and $\mathbb{P}_\rho^\pi$ and $\mathbb{E}_\rho^\pi$ when $s_0 \sim \rho$. The *value* of player i under π starting at state s, and its average under ρ, are then given by

$$V_{i,s}(\pi) = \mathbb{E}_s^\pi\left[\sum_{t=0}^{T} r_i(s_t, \alpha_t)\right] \quad \text{and} \quad V_{i,\rho}(\pi) = \sum_{s\in\mathcal{S}} \rho(s) V_{i,s}(\pi), \tag{2}$$

and the (unnormalized) *visitation count* of a state s is

$$d_\rho^\pi(s) = \mathbb{E}_\rho^\pi\left[\sum_{t=0}^{T} \mathbf{1}\{s_t = s\}\right]. \tag{3}$$

In addition, we assume that every state is visited under every stationary policy, that is,

$$d_\rho^\pi(s) > 0 \tag{R}$$

for all $\pi \in \Pi$ and all $s \in \mathcal{S}$. This *reachability* condition is automatic when ρ has full support, and it simply rules out states that would remain invisible to policy gradient learning. Finally, given an action $\alpha_i \in \mathcal{A}_i$, the *state-action value*, also known as *Q-value*, of player i is

$$Q_i^\pi(s, \alpha_i) = \sum_{\alpha_{-i}\in\mathcal{A}_{-i}} \pi_{-i}(\alpha_{-i} \mid s)\left[r_i(s, \alpha_i, \alpha_{-i}) + \sum_{s'\in\mathcal{S}} P(s' \mid s, \alpha_i, \alpha_{-i}) V_{i,s'}(\pi)\right]. \tag{4}$$

Simply put, $Q_i^\pi(s, \alpha_i)$ is the value obtained by forcing player i to play α_i once at state s, while all subsequent play, as well as the current actions of all other players, follow π.

3 Regularized policy gradient learning

The learning framework. Learning proceeds episode by episode as follows.

1. At the beginning of episode n, each player i holds a *score* $Y_{i,n} \in \mathcal{Y}_i$ and a *policy* $\pi_{i,n} \in \Pi_i$.
2. During the episode, which starts from ρ, each player follows a fixed policy based on $\pi_{i,n}$.
3. After termination, each player forms a *policy gradient signal* $\widehat{v}_{i,n} \in \mathcal{Y}_i$, which describes the performance of each action at each state.
4. The scores are updated by a *policy gradient step* $Y_{i,n+1} = Y_{i,n} + \gamma_n \widehat{v}_{i,n}$, where $\gamma_n > 0$ is the *step size*.
5. The scores are mapped back to the policy space through a *mirror map* $\psi_i \colon \mathcal{Y}_i \to \Pi_i$, namely $\pi_{i,n+1} = \psi_i(Y_{i,n+1})$.

In this light, the algorithm may be viewed as constrained gradient ascent, in which the players take a gradient step in score space and project the result back to the policy space through the mirror map.

The gradient signal. Our starting point is the *policy gradient identity* [19, 26]

$$\frac{\partial V_{i,\rho}(\pi)}{\partial \pi_i(\alpha_i \mid s)} = d_\rho^\pi(s) Q_i^\pi(s, \alpha_i)\,. \tag{5}$$

The basic object that player i seeks to track is therefore the *policy gradient field*, given componentwise, for each state-action pair (s, α_i), by

$$v_{i\alpha_i}^s(\pi) = \big(\nabla_i V_{i,\rho}(\pi)\big)_{\alpha_i}^s = d_\rho^\pi(s) Q_i^\pi(s, \alpha_i) \quad \text{and} \quad v_i(\pi) = \nabla_i V_{i,\rho}(\pi)\,, \tag{6}$$

and we write $v(\pi) = (v_i(\pi))_{i \in \mathcal{N}}$, omitting ρ from the notation since it is fixed throughout. Each component of the gradient is thus a Q-value, weighted by the visitation count of its state.

Of course, the players need not observe this field exactly. Depending on the information model, player i may receive the full vector $v_i(\pi_n)$, or only an imperfect estimate of it. To model such imperfections, we decompose the feedback signal, relative to a filtration $(\mathcal{F}_n)$, as

$$\widehat{v}_{i,n} = v_i(\pi_n) + U_{i,n} + b_{i,n}\,, \quad U_{i,n} = \widehat{v}_{i,n} - \mathbb{E}[\widehat{v}_{i,n} \,|\, \mathcal{F}_n]\,, \quad \text{and} \quad b_{i,n} = \mathbb{E}[\widehat{v}_{i,n} \,|\, \mathcal{F}_n] - v_i(\pi_n)\,, \tag{7}$$

where $U_{i,n}$ is the *noise fluctuation* in the feedback of player i and $b_{i,n}$ its *bias*. In what follows, the process is assumed adapted, i.e., Y_n and π_n are $\mathcal{F}_n$-measurable and $\widehat{v}_n$ is $\mathcal{F}_{n+1}$-measurable. Aggregating $b_n = (b_{i,n})_{i \in \mathcal{N}}$ and $U_n = (U_{i,n})_{i \in \mathcal{N}}$, we also assume that there exist deterministic sequences (B_n) and (σ_n) and some $q \geq 2$ such that

$$\|b_n\|_\infty \leq B_n \quad \text{and} \quad \mathbb{E}[\|U_n\|_\infty^q \,|\, \mathcal{F}_n] \leq \sigma_n^q\,. \tag{S}$$

These bounds accommodate both systematic and stochastic errors, whose magnitude will be kept in check by the step size schedules of Section 4. To illustrate, we first describe two classic schemes that fit this framework, before introducing a template that encompasses both.

Example 1 (Projected policy gradient)**.** The most direct instance is obtained by accumulating the gradients in score space $\mathcal{Y}$ and projecting them back to the policy space only after the score update. This yields the *projected policy gradient* recursion

$$Y_{i,n+1} = Y_{i,n} + \gamma_n \widehat{v}_{i,n} \quad \text{and} \quad \pi_{i,n+1} = \mathrm{proj}_{\Pi_i}(Y_{i,n+1})\,, \tag{PPG}$$

where proj_{Π_i} denotes the Euclidean projection onto Π_i. This is the classical policy gradient scheme with "lazy" projections [9, 26, 65]. Since the Euclidean projection can hit the boundary of Π_i, some actions may receive zero probability after finitely many iterations. ✦

Example 2 (KL-regularized policy gradient)**.** Replacing the Euclidean projection with a KL-regularized update yields a second canonical instance. Denoting by D_{KL} the *Kullback–Leibler divergence*, the update reads

$$\pi_{i,n+1} = \operatorname*{arg\,max}_{\pi_i \in \Pi_i} \{\gamma_n \langle \widehat{v}_{i,n}, \pi_i \rangle - D_{\mathrm{KL}}(\pi_i \,\|\, \pi_{i,n})\}\,. \tag{KL-PG}$$

This is the *KL-regularized policy gradient* method [2, 31, 43, 55, 60, 62]. In contrast to the Euclidean case, the KL term keeps the iterates away from the boundary whenever the initial policy has full

support. Moreover, the same recursion admits a dual form in terms of scores, as (KL-PG) is equivalent to

$$Y_{i,n+1} = Y_{i,n} + \gamma_n \widehat{v}_{i,n} \quad \text{and} \quad \pi_{i,n+1}(\alpha_i \mid s) = \frac{\exp(Y^s_{i\alpha_i,n+1})}{\sum_{\beta_i \in \mathcal{A}_i} \exp(Y^s_{i\beta_i,n+1})}\,. \tag{EPG}$$

This entropic scheme thus lends itself to two readings, as a KL-regularized primal update in the policy space or as a score aggregation process followed by a softmax map. ✦

The RPG template. To unify these two examples, we endow the policy space Π_i of each player i with a *decomposable regularizer*

$$h_i(\pi_i) = \sum_{s \in \mathcal{S}} \sum_{\alpha_i \in \mathcal{A}_i} \theta(\pi_i(\alpha_i \mid s))\,, \tag{8}$$

where the *kernel*[2] $\theta\colon [0,1] \to \mathbb{R}$ is continuous, continuously differentiable on $(0,1]$, twice differentiable on $(0,1)$, and satisfies $\inf_{p\in(0,1)} \theta''(p) > 0$. The regularizer induces the mirror map

$$\psi_i(Y_i) = \operatorname*{arg\,max}_{\pi_i \in \Pi_i} \{\langle Y_i, \pi_i \rangle - h_i(\pi_i)\}\,, \tag{9}$$

which may be viewed as a *regularized best-response map*. Absent h_i, it would simply select the policies with maximal score, and the regularizer smooths out this choice according to its own geometry. The *regularized policy gradient* recursion is then defined for each player $i \in \mathcal{N}$ as

$$Y_{i,n+1} = Y_{i,n} + \gamma_n \widehat{v}_{i,n} \quad \text{and} \quad \pi_{i,n+1} = \psi_i(Y_{i,n+1})\,. \tag{RPG}$$

When the regularizer is *steep*, i.e., when the kernel satisfies $\lim_{p\downarrow 0} \theta'(p) = -\infty$, we have $\psi_i(Y_i) \in \operatorname{ri} \Pi_i$ for all scores Y_i, and (RPG) is then equivalent to *mirror policy gradient*

$$\pi_{i,n+1} = \operatorname*{arg\,max}_{\pi_i \in \Pi_i} \{\gamma_n \langle \widehat{v}_{i,n}, \pi_i \rangle - D(\pi_i, \pi_{i,n})\}\,, \tag{MPG}$$

where D is the *Bregman divergence* generated by the regularizer [9, 31, 36, 62]. In particular, (PPG) is induced by the *Euclidean kernel* $\theta(p) = p^2/2$, which is not steep and whose mirror map is the Euclidean projection, whereas (EPG) corresponds to the *entropic kernel* $\theta(p) = p \log p$, which is steep, whose mirror map is the softmax, and whose Bregman divergence is the KL divergence; further examples of regularizers are given in Appendix C.3.

Finally, in Appendix E, we define a deterministic family of algorithms based on *extrapolation*, in the spirit of [5, 33], which we call (RPG+) and which fits the template (RPG). It contains vanilla (RPG) with full information ($\widehat{v} = v$), its *optimistic* variant [3, 16, 38, 56, 59, 63], and its *extragradient* variant [15, 17] as special cases.

Feedback from trajectories. In practice, the players seldom have access to their full policy gradient field v_i; rather, they generate trajectories through their interaction in the game and must recover the gradient from the realized states, actions, transitions, and rewards, a task the following estimator accomplishes from single trajectory feedback. Fix an iteration n, let $\widehat{\pi}_n$ be a policy with full support, and sample one stopped trajectory $\tau_n = (s_0^{(n)}, \alpha_0^{(n)}, \dots, s_{T_n}^{(n)}, \alpha_{T_n}^{(n)})$ from the game started at ρ and run under $\widehat{\pi}_n$. The REINFORCE estimator [19, 26, 54, 58] is defined for each action and state by

$$\widehat{v}^s_{i\alpha_i,n} = \sum_{t=0}^{T_n} \frac{\mathbf{1}\{s_t^{(n)} = s,\, \alpha_{i,t}^{(n)} = \alpha_i\}}{\widehat{\pi}_{i,n}(\alpha_i \mid s)} \sum_{u=t}^{T_n} r_i(s_u^{(n)}, \alpha_u^{(n)})\,. \tag{REINFORCE}$$

Algorithm 1 (REINFORCE-RPG)**.** At iteration n, use the ε_n*-greedy policy*

$$\widehat{\pi}_{i,n}(\cdot \mid s) = (1 - \varepsilon_n)\pi_{i,n}(\cdot \mid s) + \varepsilon_n \operatorname{unif}(\mathcal{A}_i)\,, \tag{ε-greedy}$$

sample one stopped episode under $\widehat{\pi}_n$, form $\widehat{v}_n$ by (REINFORCE), and update

$$Y_{i,n+1} = Y_{i,n} + \gamma_n \widehat{v}_{i,n} \quad \text{and} \quad \pi_{i,n+1} = \psi_i(Y_{i,n+1})\,. \tag{R-RPG}$$

The uniform exploration floor in (ε-greedy) keeps the denominator in (REINFORCE) under control. In Appendix E, we show that this algorithm conforms to the template (RPG).

[2] More generally, we could allow a different kernel θ_i^s for each player i and state s, but we refrain from doing so to keep the notation light.

4 From pointwise to setwise stability

We begin by defining the notion of dynamic stability used throughout.

4.1 Stochastic stability and attraction

Definition 1. Let $S \subseteq \Pi$ be closed. Given a tolerance level $\eta \in (0,1)$, we say that S is

- *stochastically stable* if, for every neighborhood U of S, there exists a neighborhood V of S such that, for every initialization $\pi_1 = \psi(Y_1) \in V$ of (RPG), the iterates satisfy $\pi_n \in U$ for all $n \geq 1$ with probability at least $1 - \eta$,
- *stochastically attracting* if there exists a neighborhood U of S such that, for every initialization $\pi_1 = \psi(Y_1) \in U$ of (RPG), we have $\operatorname{dist}_1(\pi_n, S) \to 0$ with probability at least $1 - \eta$,
- *stochastically asymptotically stable* if it is both stochastically stable and stochastically attracting.

These are the stochastic counterparts of the classical notions of Lyapunov stability and attraction in the theory of dynamical systems [12, 27, 42], and, when the neighborhood in the definition of attraction can be taken to be Π itself, the latter amounts to *global* convergence to S. As a matter of convention, when one of these properties holds at every tolerance level $\eta \in (0,1)$ as soon as the base step size γ of the schedule in Assumption 1 below is small enough—with a threshold that may depend on η and on the neighborhoods at play—we omit the qualifier and simply say that S is stochastically stable, attracting, or asymptotically stable.

In the proofs, stability is established through an *energy function*, that is, a function on score space with local dissipation properties (see Appendix D). To proceed, we work under the following step size, bias and noise schedules, which are rather broad and standard in stochastic approximation [10, 42].

Assumption 1. The sequences (B_n) and (σ_n) of (S) and the step size γ_n satisfy

$$\gamma_n = \gamma n^{-\ell_\gamma}, \qquad B_n = \mathcal{O}(n^{-\ell_b}), \quad \text{and} \quad \sigma_n = \mathcal{O}(n^{\ell_\sigma}), \tag{10}$$

with $\ell_\gamma \in (1/2, 1]$, $\ell_b > 0$, $\ell_\sigma < 1/2$, $\ell_\gamma + \ell_b > 1$, and $\ell_\gamma - \ell_\sigma > 1/2$.

In Appendix E, we verify these schedules for vanilla and stochastic (RPG), the extrapolated family (RPG+), and the algorithm (R-RPG) driven by trajectories, and we summarize them in Table 1. Concretely, these schedules ensure that $\sum_n \gamma_n = \infty$ while $\sum_n \gamma_n^2 < \infty$, that the bias has a finite cumulative effect, i.e., $\sum_n \gamma_n B_n < \infty$, and that the noise, although it may grow, has a finite accumulated variance, i.e., $\sum_n \gamma_n^2 \sigma_n^2 < \infty$.

4.2 Nash policies and pointwise stability

We say that a policy profile $\pi^\star$ is a *Nash policy* if no player can improve their value by a stationary Markov deviation, that is, if

$$V_{i,\rho}(\pi^\star) \geq V_{i,\rho}(\pi_i, \pi^\star_{-i}) \tag{11}$$

for every player $i \in \mathcal{N}$ and every policy $\pi_i \in \Pi_i$. In view of the policy gradient identity (5) and the reachability condition (R), this is equivalent to the variational inequality which requires that, for all $\pi \in \Pi$,

$$\langle v(\pi^\star), \pi - \pi^\star \rangle \leq 0\,. \tag{12}$$

Nash policies are thus the variationally stationary points of the policy gradient field [19, 26], and, although they always exist [23], they may well fail to be dynamically stable [25]. Requiring instead every deviation to be strictly worse leads to the following, more demanding notion.

Definition 2. We say that a policy profile $\pi^\star$ is a *strict Nash policy* if, for every player i and every stationary Markov deviation $\pi_i \neq \pi_i^\star$,

$$V_{i,\rho}(\pi^\star) > V_{i,\rho}(\pi_i, \pi^\star_{-i})\,. \tag{13}$$

Our first result provides equivalent state-action and variational forms of this notion.

Theorem 1. *For a policy profile $\pi^\star$, the following are equivalent:*

(a) *$\pi^\star$ is a strict Nash policy.*

(b) *$\pi^\star$ is deterministic and, writing $\beta_i^\star(s)$ for its prescribed action, we have*

$$Q_i^{\pi^\star}(s, \beta_i^\star(s)) > Q_i^{\pi^\star}(s, \alpha_i) \tag{14}$$

for every player i, state s, and action $\alpha_i \neq \beta_i^\star(s)$.

(c) *$\pi^\star$ is variationally stable, i.e., there exists a neighborhood U of $\pi^\star$ such that, for all $\pi \in U \setminus \{\pi^\star\}$,*

$$\langle v(\pi), \pi - \pi^\star \rangle < 0\,. \tag{15}$$

Strict Nash policies are therefore the deterministic Nash policies at which, state by state, every unilateral deviation strictly lowers the state-action value of the deviator. As we show in Appendix F.1, strictness is in fact equivalent to the seemingly stronger *sharpness* bound

$$\langle v(\pi), \pi - \pi^\star \rangle \leq -c\,\|\pi - \pi^\star\|_1 \tag{16}$$

for some $c > 0$ and all π sufficiently close to $\pi^\star$. Under full visitation, this closes the gap between strictness and stability left open by Giannou et al. [26], whose intermediate notion of second-order stationarity consequently coincides with strictness (see Remark F.1 for the details). Combined with the energy method of Appendix D, the sharpness bound yields the following stability guarantee.

Corollary 1. *Under Assumption 1, every strict Nash policy is stochastically asymptotically stable.*

The sharpness bound has a simple geometric interpretation in terms of the *tangent cone* to Π at $\pi^\star$, which consists of the feasible directions at $\pi^\star$, and of its polar, the *normal cone*, namely

$$\begin{aligned} \mathrm{TC}(\pi^\star) &= \{t(\pi - \pi^\star) : \pi \in \Pi,\, t \geq 0\}\,, \\ \mathrm{NC}(\pi^\star) &= \left\{w \in \mathcal{Y} : \sup\nolimits_{z \in \mathrm{TC}(\pi^\star)} \langle w, z \rangle \leq 0\right\}. \end{aligned} \tag{17}$$

In these terms, the variational inequality (12) states that $v(\pi^\star) \in \mathrm{NC}(\pi^\star)$, while the strict Q-value gaps of Theorem 1(b) place $v(\pi^\star)$ in the interior of $\mathrm{NC}(\pi^\star)$. The field then forms an obtuse angle with every feasible direction, a property which persists near $\pi^\star$ by continuity and is precisely what (16) expresses (see Figure 1). Unlike general Nash policies, strict Nash policies therefore exhibit a strict inward drift, which underlies their stability in Corollary 1. Part of this strength, however, stems from the fact that strict Nash policies sit at vertices of the policy space, which suggests, quite naturally, moving from vertices to *faces*, that is, to *state-action sets*.

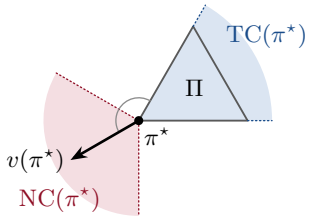


Figure 1: The field at a strict Nash policy $\pi^\star$ lies in the interior of the normal cone.

4.3 One-shot club is not enough

At first sight, it is tempting to impose the normal-form notion of closedness under better replies separately at each state. Recall that, in the one-shot case, a finite normal-form game consists of players $i \in \mathcal{N}$, action sets $\mathcal{A}_i$, and payoffs $u_i \colon \mathcal{A} \to \mathbb{R}$, where $\mathcal{A} = \prod_i \mathcal{A}_i$ (equivalently, it is the stochastic game with a single state that stops immediately after the stage payoff). Given nonempty sets $\mathcal{B}_i \subseteq \mathcal{A}_i$, write $\mathcal{B} = \prod_i \mathcal{B}_i$ and $\mathrm{span}(\mathcal{B}) = \prod_i \Delta(\mathcal{B}_i)$. The support set $\mathcal{B}$ is *closed under better replies* (club) if

$$u_i(\beta_i, \beta_{-i}) > u_i(\alpha_i, \beta_{-i}) \tag{18}$$

for every player i, every $\beta_{-i} \in \mathcal{B}_{-i}$, every $\beta_i \in \mathcal{B}_i$, and every $\alpha_i \in \mathcal{A}_i \setminus \mathcal{B}_i$. That is, as long as the opponents remain in $\mathcal{B}$, leaving $\mathcal{B}_i$ is strictly costly. In finite normal-form games, this condition is sharp, as $\mathrm{span}(\mathcal{B})$ is asymptotically stable under the regularized dynamics if and only if $\mathcal{B}$ is club [1, 12, 14]. As the next example shows, however, this statewise test falls short in stochastic games.

In Figure 2, each state hosts a $2 \times 2 \times 2$ game, in which player 1 chooses top or bottom, player 2 left or right, and player 3 front (F) or back (K). The black edge arrows indicate the preferences of player 3, so that F is one-shot club at state 1 and K is one-shot club at state 2. As such, the statewise support F at state 1 and K at state 2 would be self-enforcing if only stage rewards mattered. The orange arrows indicate the state transitions, in that, with high probability, play remains at state 1, while, at state 2, action F leads to state 1 and action K keeps the game at state 2. The excluded action

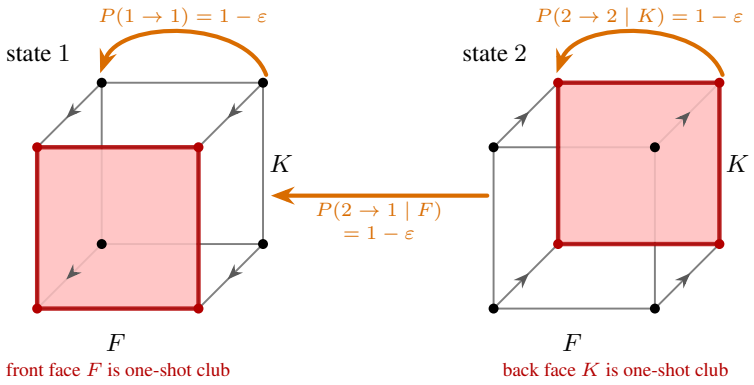


Figure 2: A support which is one-shot club at each state, yet fails to be stable.

F at state 2 therefore leads to the more rewarding state 1, and, in the instance of Appendix F.2, the combined face is neither attracting nor stable under vanilla (EPG), the action K even going *extinct* at state 2. Even more strikingly, modifying the transitions alone may reverse this behavior and alter the selected support (e.g., by setting $P(2 \to 1 \mid K) = 1 - \varepsilon$ instead), in stark contrast with normal-form games, where preferences alone decide the stability of a support. The temporal structure of stochastic games thus changes the picture in a fundamental way.

4.4 Q-club sets

The preceding example also suggests the remedy, namely that the support selected by the dynamics should be justified by *continuation values*, in that, once play is confined to this support, any deviation from it, at any state, lowers the cumulative reward of the deviator. To formalize this idea, for each player $i \in \mathcal{N}$ and state $s \in \mathcal{S}$, let $\mathcal{B}_i(s) \subseteq \mathcal{A}_i$ be a nonempty set of allowed actions, with excluded actions $\bar{\mathcal{B}}_i(s) = \mathcal{A}_i \setminus \mathcal{B}_i(s)$. The resulting *state-action set* and its *policy span* are

$$\mathcal{B} = \prod_{s \in \mathcal{S}} \prod_{i \in \mathcal{N}} \mathcal{B}_i(s) \quad \text{and} \quad \operatorname{span}(\mathcal{B}) = \prod_{s \in \mathcal{S}} \prod_{i \in \mathcal{N}} \Delta(\mathcal{B}_i(s)) \subseteq \Pi\,. \tag{19}$$

Geometrically, $\operatorname{span}(\mathcal{B})$ is the *face* of the policy space in which each player i, at each state s, only plays actions of $\mathcal{B}_i(s)$. We may now state the central definition of this paper.

Definition 3. The state-action set $\mathcal{B}$ is *closed under Q-better replies* (Q-club) if

$$Q_i^{\pi}(s, \beta_i) > Q_i^{\pi}(s, \alpha_i) \tag{20}$$

for every player $i \in \mathcal{N}$, state $s \in \mathcal{S}$, policy profile $\pi \in \operatorname{span}(\mathcal{B})$, allowed action $\beta_i \in \mathcal{B}_i(s)$, and excluded action $\alpha_i \in \bar{\mathcal{B}}_i(s)$.

Put differently, as long as play remains in $\operatorname{span}(\mathcal{B})$, no player can improve their state-action value by leaving their prescribed support, at any state. This is the setwise generalization of strict Nash policies, to which it reduces, by Theorem 1(b), when each $\mathcal{B}_i(s)$ is a singleton. In a normal-form game, where Q-values are simply payoffs, linear in the strategies of the opponents, Definition 3 boils down to (18). In general, however, the comparison must hold on all of $\operatorname{span}(\mathcal{B})$, and not only at its deterministic profiles, since continuation values, unlike stage payoffs, are in general neither linear nor even concave in each player's own policy.

To connect this notion with classical game-theoretic constructions, and in the spirit of Ritzberger and Weibull [45], we next consider the *state-action better-reply correspondence*, which assigns to every policy profile π, player i and state s the set

$$\operatorname{Qbtr}_{i,s}(\pi) = \{\alpha_i \in \mathcal{A}_i : Q_i^{\pi}(s, \alpha_i) \geq V_{i,s}(\pi)\} \tag{21}$$

of pure actions that player i, at state s, has a nonnegative incentive to play instead of the mixed action prescribed by π, and which acts on state-action sets through

$$\operatorname{Qbtr}(\mathcal{B}) = \prod_{s \in \mathcal{S}} \prod_{i \in \mathcal{N}} \bigcup_{\pi \in \operatorname{span}(\mathcal{B})} \operatorname{Qbtr}_{i,s}(\pi)\,. \tag{22}$$

The main game-theoretic properties of Q-club sets, proven in Appendix F.3, are as follows.

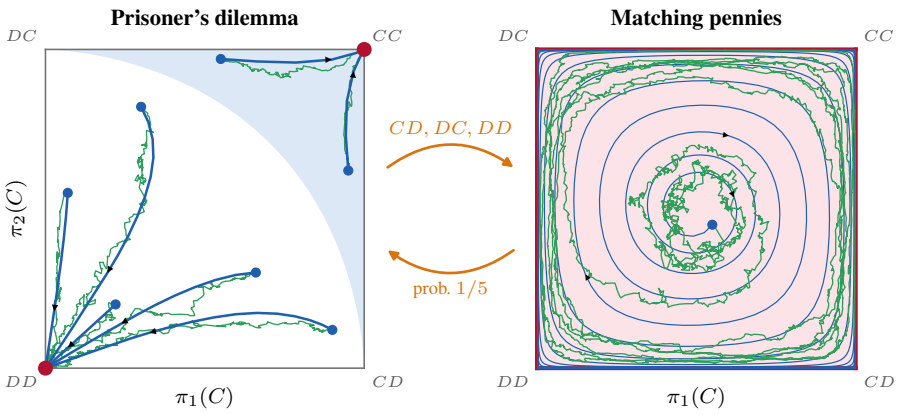


Figure 3: Learning to cooperate in a stochastic prisoner's dilemma. Mutual cooperation keeps the players in a prisoner's dilemma, whereas any defection sends them to a costly matching pennies game, from which play returns with probability $1/5$. Although defection is strictly dominant in the prisoner's dilemma, the state-action set $\mathcal{B}$ prescribing mutual cooperation there and leaving the other state unrestricted is Q-club, as is its counterpart prescribing mutual defection (both spans in red). The squares show the cooperation probabilities along (EPG) (blue) and entropic (R-RPG) (green). From the blue basin, the iterates converge to $\operatorname{span}(\mathcal{B})$, but not to a point, as episodes start in matching pennies with probability $1/2$, where play spirals out toward the boundary (see Appendix F.5).

Proposition 1. *Every Q-club set $\mathcal{B}$ is a fixed set of the state-action better-reply correspondence, i.e., $\operatorname{Qbtr}(\mathcal{B}) = \mathcal{B}$, although the converse fails in general. Moreover, the span of every Q-club set contains a stationary Nash policy.*

Unlike strict Nash policies, Q-club sets always exist. Indeed, the state-action set allowing every action at every state excludes nothing and is thus vacuously Q-club, while minimal ones exist since there are finitely many state-action sets. By Proposition 1, converging to any of them ensures that the iterates at least settle on a face containing a Nash policy, a far more attainable target, as we now show.

Theorem 2. *Under Assumption 1, if a state-action set $\mathcal{B}$ is Q-club, then $\operatorname{span}(\mathcal{B})$ is stochastically asymptotically stable.*

As mutual cooperation and mutual defection illustrate in Figure 3, a game may admit disjoint Q-club sets. This guarantee is therefore necessarily local, and the set which the dynamics select depends on the initialization. The proof, detailed in Appendix F, may be sketched as follows. Let $\mathcal{C} = \operatorname{span}(\mathcal{B})$ and $\pi = \psi(y)$, and let $\pi_{\mathcal{B}} = \arg\max_{p \in \mathcal{C}} \{\langle y, p\rangle - h(p)\}$ be the projection of y on $\mathcal{C}$ adapted to the regularizer. The Fenchel gap $F_{\mathcal{C}}(y) = \min_{p \in \mathcal{C}} \{h(p) + h^*(y) - \langle y, p\rangle\}$, where h^* is the convex conjugate of h, is a "dual distance" to $\mathcal{C}$, insofar as it is always nonnegative and vanishes if and only if $\psi(y) \in \mathcal{C}$, and it plays the role of an *energy function*. Its directional derivative along the policy gradient field is

$$\langle \nabla F_{\mathcal{C}}(y), v(\pi)\rangle = \langle v(\pi), \pi - \pi_{\mathcal{B}}\rangle = \sum_{i,s} d_\rho^\pi(s) \sum_{\alpha_i \in \mathcal{A}_i} (\pi^s_{i\alpha_i} - \pi^s_{\mathcal{B},i\alpha_i}) Q_i^\pi(s, \alpha_i)\,. \tag{23}$$

Now, for decomposable regularizers, the first-order optimality conditions imply that $\pi^s_{i\beta_i} \leq \pi^s_{\mathcal{B},i\beta_i}$ for all $\beta_i \in \mathcal{B}_i(s)$. Moreover, near a Q-club span, every allowed action enjoys a uniform Q-value advantage over every excluded action, and (R) keeps $d_\rho^\pi(s)$ uniformly positive. Putting these facts together yields, for some $c > 0$ and every policy in a sufficiently small neighborhood of $\mathcal{C}$,

$$\langle v(\pi), \pi - \pi_{\mathcal{B}}\rangle \leq -c \operatorname{dist}_1(\pi, \mathcal{C})\,, \tag{24}$$

which is the setwise version of the sharpness bound (16), with c independent of the regularizer. The energy theorem of Appendix D then does the rest, yielding stochastic stability and attraction.

4.5 Fast convergence to Q-club sets

Having established stability, we now address the speed of convergence. To accommodate less restrictive schedules, constant step sizes included, we restrict the initialization to a suitable open domain of score space, and, as before, the proofs are deferred to Appendix F.

Assumption 2. The sequences (B_n) and (σ_n) of (S) and the step size γ_n satisfy

$$\gamma_n = \gamma n^{-\ell_\gamma}, \qquad B_n = \mathcal{O}(n^{-\ell_b}), \quad \text{and} \quad \sigma_n = \mathcal{O}(n^{\ell_\sigma}), \tag{25}$$

where $\ell_\gamma \in [0,1]$ and $\ell_b > 0$, with $\ell_\sigma < 1/2$ if $\ell_\gamma < 1$, and $\ell_\sigma \le 1/2$ if $\ell_\gamma = 1$.

The algorithms covered by these schedules are again listed in Table 1. Next, let $a := \lim_{t \downarrow 0} \theta'(t) \in [-\infty, \infty)$ and $b := \theta'(1)$, and define the *rate function* of the kernel, a nondecreasing function which vanishes at $-\infty$, as

$$\varphi(u) = \begin{cases} 0, & u \le a, \\ (\theta')^{-1}(u), & a < u < b, \\ 1, & u \ge b. \end{cases} \tag{26}$$

Theorem 3. *Under Assumption 2, if a state-action set $\mathcal{B}$ is Q-club, then, for every $\eta \in (0,1)$, there exist an open unbounded domain $\mathcal{D} \subseteq \mathcal{Y}$ and constants $c_0 > 0$, $c_1 \in \mathbb{R}$ and $c_2 > 0$ such that, for every initialization $Y_1 \in \mathcal{D}$, with probability at least $1 - \eta$, we have, for every episode $n \ge 1$,*

$$\operatorname{dist}_1(\pi_n, \operatorname{span}(\mathcal{B})) \le c_0 \, \varphi\left(c_1 - c_2 \sum_{k=1}^{n-1} \gamma_k \right). \tag{27}$$

Corollary 2. *Suppose that the assumptions of Theorem 3 hold. Then, with probability at least $1 - \eta$, the iterates π_n converge to $\operatorname{span}(\mathcal{B})$ at the following rates, for every episode $n \ge 1$,*

$$\operatorname{dist}_1(\pi_n, \operatorname{span}(\mathcal{B})) \le c_0 \cdot \begin{cases} [c_1 - c_2 \sum_{k=1}^{n} \gamma_k]_+, & \text{for (PPG)}, \\ \exp(-c_2 \sum_{k=1}^{n} \gamma_k), & \text{for (EPG)}, \end{cases} \tag{28}$$

for some constants $c_0 > 0$, $c_1 \in \mathbb{R}$ and $c_2 > 0$, and where $[x]_+ := \max\{x, 0\}$.

The entropic scheme (EPG) thus converges geometrically under constant step sizes, whereas the Euclidean scheme (PPG) identifies the span in finite time. Following Boone and Mertikopoulos [12], the proof of Theorem 3 replaces the Fenchel gap by the *linear energies* $E_z(y) = \langle y, z \rangle$, where $z = e^s_{i\alpha} - e^s_{i\beta}$ pairs an excluded action $\alpha \in \bar{\mathcal{B}}_i(s)$ with an allowed action $\beta \in \mathcal{B}_i(s)$, and which measure the lead of β over α in score space. Near $\operatorname{span}(\mathcal{B})$, the Q-club condition gives $\langle v(\pi), z \rangle \le -c$ for some $c > 0$, and these energies decrease linearly in $\Gamma_n = \sum_{k=1}^{n} \gamma_k$, whereas the noise and the bias only accumulate like $G(1 + \Gamma_{n-1})^\alpha$ for some $\alpha < 1$. When the initial energies are negative enough, the linear drift thus prevails along the whole trajectory, with $E_z(Y_n) \le -M - c\Gamma_{n-1}/2$ for some $M > 0$. The optimality conditions of the mirror map then translate this bound into $\pi_{i,n}(\alpha \mid s) \le \varphi(b - M - c\Gamma_{n-1}/2)$, hence the rates. As this argument dispenses with small steps, Assumption 2 allows constant step sizes of any magnitude. On the other hand, since $\psi(\mathcal{D})$ is not a neighborhood of $\operatorname{span}(\mathcal{B})$ in general, Theorem 3 does not establish stability in the sense of Definition 1, but rather a dual, more relaxed form of stability in score space, for which we can quantify the speed of convergence.

5 Concluding remarks

Nash policies, in general, are subject to severe dynamical limitations. To overcome them, we have introduced a setwise stability notion for learning in stochastic games, based on unilateral comparisons of state-action values rather than stage payoffs. The resulting Q-club sets are closed under state-action better replies, and they are stable and attracting under a broad class of regularized policy gradient methods, even under noise and partial feedback. Several questions nonetheless remain open, chief among them whether asymptotically stable state-action sets admit an intrinsic game-theoretic characterization, as they do in normal-form games. Such a characterization is, however, considerably more elusive in stochastic games, where continuation values need not be linear, nor even concave, in the policy of a player. How hard is it to find minimal Q-club sets, and can the local guarantees be complemented by estimates of the size of their basins of attraction? Can stable regions be identified beyond support sets, such as more general spans of deterministic policy profiles? We believe that these questions open a rich avenue for studying collective behavior in stochastic games.

Acknowledgments

This research was supported in part by the French National Research Agency (ANR) in the framework of the PEPR IA FOUNDRY project (ANR-23-PEIA-0003). PM is also a member of the Archimedes Research Unit/Athena RC, and was partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. We would also like to thank the IRP FAIRGAME of the CNRS for supporting this project.

References

[1] O. Abbadi, R. Laraki, and P. Mertikopoulos. What preferences can—and cannot—predict in multi-agent online learning. In *Proc. 43rd Internat. Conf. Machine Learning*, volume 306 of *Proc. Mach. Learn. Res.*, pages 22–68, 2026.

[2] A. Alacaoglu, L. Viano, N. He, and V. Cevher. A natural actor-critic framework for zero-sum Markov games. In *Proc. 39th Internat. Conf. Machine Learning*, volume 162 of *Proc. Mach. Learn. Res.*, pages 307–366, 2022.

[3] I. Anagnostides, I. Panageas, G. Farina, and T. Sandholm. Optimistic policy gradient in multi-player Markov games with a single controller: Convergence beyond the Minty property. In *Proc. 38th AAAI Conf. Artificial Intelligence*, volume 38, pages 9451–9459, 2024.

[4] G. Arslan and S. Yüksel. Decentralized Q-learning for stochastic teams and games. *IEEE Trans. Automat. Control*, 62(4):1545–1558, 2017.

[5] W. Azizian, F. Iutzeler, J. Malick, and P. Mertikopoulos. The rate of convergence of Bregman proximal methods: Local geometry versus regularity versus sharpness. *SIAM J. Optim.*, 34(3):2440–2471, 2024.

[6] L. Baudin and R. Laraki. Fictitious play and best-response dynamics in identical interest and zero-sum stochastic games. In *Proc. 39th Internat. Conf. Machine Learning*, volume 162 of *Proc. Mach. Learn. Res.*, pages 1664–1690, 2022.

[7] L. Baudin and R. Laraki. Smooth fictitious play in stochastic games with perturbed payoffs and unknown transitions. In *Advances in Neural Information Processing Systems 35*, 2022.

[8] A. Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. SIAM, Philadelphia, PA, 2017.

[9] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, 2003.

[10] M. Benaïm. Dynamics of stochastic approximation algorithms. In J. Azéma, M. Émery, M. Ledoux, and M. Yor, editors, *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pages 1–68. Springer, Berlin, 1999.

[11] D. P. Bertsekas. *Rollout, Policy Iteration, and Distributed Reinforcement Learning*. Athena Scientific, Belmont, MA, 2020.

[12] V. Boone and P. Mertikopoulos. The equivalence of dynamic and strategic stability under regularized learning in games. In *Advances in Neural Information Processing Systems 36*, 2023.

[13] W. Brown. Learning in multi-player stochastic games. In *Proc. 37th Conf. Uncertainty in Artificial Intelligence*, volume 161 of *Proc. Mach. Learn. Res.*, pages 1927–1937, 2021.

[14] P.-L. Cauvin, D. Legacci, and P. Mertikopoulos. The impact of uncertainty on regularized learning in games. In *Proc. 42nd Internat. Conf. Machine Learning*, volume 267 of *Proc. Mach. Learn. Res.*, pages 6920–6957, 2025.

[15] S. Cen, Y. Wei, and Y. Chi. Fast policy extragradient methods for competitive games with entropy regularization. In *Advances in Neural Information Processing Systems 34*, 2021.

[16] S. Cen, Y. Chi, S. S. Du, and L. Xiao. Faster last-iterate convergence of policy optimization in zero-sum Markov games. In *Proc. 11th Internat. Conf. Learning Representations*, 2023.

[17] Z. Chen, S. Ma, and Y. Zhou. Sample efficient stochastic policy extragradient algorithm for zero-sum Markov game. In *Proc. 10th Internat. Conf. Learning Representations*, 2022.

[18] Y. S. Chow. A martingale inequality and the law of large numbers. *Proc. Amer. Math. Soc.*, 11(1):107–111, 1960. doi: 10.1090/S0002-9939-1960-0112190-3.

[19] C. Daskalakis, D. J. Foster, and N. Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Advances in Neural Information Processing Systems 33*, 2020.

[20] C. Daskalakis, N. Golowich, and K. Zhang. The complexity of Markov equilibrium in stochastic games. In *Proc. 36th Conf. Learning Theory*, volume 195 of *Proc. Mach. Learn. Res.*, pages 4180–4234, 2023.

[21] D. Ding, C.-Y. Wei, K. Zhang, and M. R. Jovanović. Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *Proc. 39th Internat. Conf. Machine Learning*, volume 162 of *Proc. Mach. Learn. Res.*, pages 5166–5220, 2022.

[22] J. A. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Springer, New York, 1997.

[23] A. M. Fink. Equilibrium in a stochastic n-person game. *J. Sci. Hiroshima Univ. Ser. A-I Math.*, 28(1): 89–93, 1964.

[24] R. Fox, S. McAleer, W. Overman, and I. Panageas. Independent natural policy gradient always converges in Markov potential games. In *Proc. 25th Internat. Conf. Artificial Intelligence and Statistics*, volume 151 of *Proc. Mach. Learn. Res.*, pages 4414–4425, 2022.

[25] A. Giannou, E. V. Vlatakis-Gkaragkounis, and P. Mertikopoulos. Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information. In *Proc. 34th Conf. Learning Theory*, volume 134 of *Proc. Mach. Learn. Res.*, pages 2147–2148, 2021.

[26] A. Giannou, K. Lotidis, P. Mertikopoulos, and E. V. Vlatakis-Gkaragkounis. On the convergence of policy gradient methods to Nash equilibria in general stochastic games. In *Advances in Neural Information Processing Systems 35*, 2022.

[27] M. W. Hirsch, S. Smale, and R. L. Devaney. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic Press, Waltham, MA, 3rd edition, 2012.

[28] C. Jin, Q. Liu, Y. Wang, and T. Yu. V-learning—A simple, efficient, decentralized algorithm for multiagent reinforcement learning. *Math. Oper. Res.*, 49(4):2295–2322, 2024.

[29] Y. Jin, V. Muthukumar, and A. Sidford. The complexity of infinite-horizon general-sum stochastic games. In *Proc. 14th Innovations in Theoretical Computer Science Conf.*, volume 251 of *Leibniz Internat. Proc. Inform.*, pages 76:1–76:20, 2023.

[30] S. M. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *Proc. 19th Internat. Conf. Machine Learning*, pages 267–274, 2002.

[31] G. Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Math. Program.*, 198(1):1059–1106, 2023.

[32] R. Laraki. Jeux stochastiques. In *Théorie des jeux: Introduction à la théorie des jeux répétés, Journées X-UPS 2006*, pages 93–133. Éditions de l'École polytechnique, Palaiseau, 2006.

[33] D. Legacci, P. Mertikopoulos, C. H. Papadimitriou, G. Piliouras, and B. S. R. Pradelski. No-regret learning in harmonic games: Extrapolation in the face of conflicting interests. In *Advances in Neural Information Processing Systems 37*, 2024.

[34] S. Leonardos, W. Overman, I. Panageas, and G. Piliouras. Global convergence of multi-agent policy gradient in Markov potential games. In *Proc. 10th Internat. Conf. Learning Representations*, 2022.

[35] D. S. Leslie, S. Perkins, and Z. Xu. Best-response dynamics in zero-sum stochastic games. *J. Econom. Theory*, 189:105095, 2020.

[36] Y. Li and G. Lan. Policy mirror descent inherently explores action space. *SIAM J. Optim.*, 35(1):116–156, 2025.

[37] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proc. 11th Internat. Conf. Machine Learning*, pages 157–163, 1994.

[38] W. Mao, H. Qiu, C. Wang, H. Franke, Z. Kalbarczyk, and T. Başar. $\widetilde{O}(T^{-1})$ convergence to (coarse) correlated equilibria in full-information general-sum Markov games. In *Proc. 6th Annual Learning for Dynamics and Control Conf.*, volume 242 of *Proc. Mach. Learn. Res.*, pages 361–374, 2024.

[39] J.-F. Mertens, S. Sorin, and S. Zamir. *Repeated Games*. Econometric Society Monographs. Cambridge University Press, Cambridge, 2015.

[40] P. Mertikopoulos, C. H. Papadimitriou, and G. Piliouras. Cycles in adversarial regularized learning. In *Proc. 29th Annual ACM-SIAM Symp. Discrete Algorithms*, 2018.

[41] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *Proc. 7th Internat. Conf. Learning Representations*, 2019.

[42] P. Mertikopoulos, Y.-P. Hsieh, and V. Cevher. A unified stochastic approximation framework for learning in games. *Math. Program.*, 203(1–2):559–609, 2024.

[43] A. Nayak, T. Yang, O. Yağan, G. Joshi, and Y. Chi. Achieving logarithmic regret in KL-regularized zero-sum Markov games. In *Proc. 43rd Internat. Conf. Machine Learning*, 2026.

[44] A. Neyman and S. Sorin, editors. *Stochastic Games and Applications*, volume 570 of *NATO Science Series C*. Kluwer Academic Publishers, Dordrecht, 2003.

[45] K. Ritzberger and J. W. Weibull. Evolutionary selection in normal-form games. *Econometrica*, 63(6): 1371–1399, 1995.

[46] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[47] M. O. Sayin, K. Zhang, D. S. Leslie, T. Başar, and A. Ozdaglar. Decentralized Q-learning in zero-sum Markov games. In *Advances in Neural Information Processing Systems 34*, 2021.

[48] M. O. Sayin, F. Parise, and A. Ozdaglar. Fictitious play in zero-sum stochastic games. *SIAM J. Control Optim.*, 60(4):2095–2114, 2022.

[49] L. S. Shapley. Stochastic games. *Proc. Natl. Acad. Sci. USA*, 39(10):1095–1100, 1953.

[50] Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, Cambridge, 2009.

[51] Y. Shoham, R. Powers, and T. Grenager. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377, 2007.

[52] E. Solan. *A Course in Stochastic Game Theory*, volume 103 of *London Mathematical Society Student Texts*. Cambridge University Press, Cambridge, 2022.

[53] Y. Sun, T. Liu, R. Zhou, P. R. Kumar, and S. Shahrampour. Provably fast convergence of independent natural policy gradient for Markov potential games. In *Advances in Neural Information Processing Systems 36*, 2023.

[54] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018.

[55] M. Tomar, L. Shani, Y. Efroni, and M. Ghavamzadeh. Mirror descent policy optimization. In *Proc. 10th Internat. Conf. Learning Representations*, 2022.

[56] C.-Y. Wei, C.-W. Lee, M. Zhang, and H. Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. In *Proc. 34th Conf. Learning Theory*, volume 134 of *Proc. Mach. Learn. Res.*, pages 4259–4299, 2021.

[57] D. Williams. *Probability with Martingales*. Cambridge University Press, Cambridge, 1991.

[58] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, 1992.

[59] Y. Yang and C. Ma. $O(T^{-1})$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum Markov games. In *Proc. 11th Internat. Conf. Learning Representations*, 2023.

[60] C. Ye, W. Xiong, Y. Zhang, H. Dong, N. Jiang, and T. Zhang. Online iterative reinforcement learning from human feedback with general preference model. In *Advances in Neural Information Processing Systems 37*, 2024.

[61] S. Zeng, T. T. Doan, and J. Romberg. Regularized gradient descent ascent for two-player zero-sum Markov games. In *Advances in Neural Information Processing Systems 35*, 2022.

[62] W. Zhan, S. Cen, B. Huang, Y. Chen, J. D. Lee, and Y. Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM J. Optim.*, 33(2): 1061–1091, 2023.

[63] R. Zhang, Q. Liu, H. Wang, C. Xiong, N. Li, and Y. Bai. Policy optimization for Markov games: Unified framework and faster convergence. In *Advances in Neural Information Processing Systems 35*, 2022.

[64] R. Zhang, J. Mei, B. Dai, D. Schuurmans, and N. Li. On the global convergence rates of decentralized softmax gradient play in Markov potential games. In *Advances in Neural Information Processing Systems 35*, 2022.

[65] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proc. 20th Internat. Conf. Machine Learning*, pages 928–936, 2003.

A General roadmap

This appendix offers a bird's-eye view of the proofs of Theorems 2 and 3. Their common thread is that the policy process is best understood through its score representation, in which the recursion (RPG) becomes additive, namely

$$Y_{n+1} = Y_n + \gamma_n \widehat{v}_n \,, \qquad \pi_n = \psi(Y_n) \,, \quad \text{and} \quad \widehat{v}_n = v(\pi_n) + U_n + b_n \,. \tag{A.1}$$

In this representation, the proof of stochastic stability rests on an energy of Lyapunov type defined on score space, while the proof of the rates refines this picture by tracking the relative scores of allowed and excluded actions.

Energy functions and stochastic stability. Let $S \subseteq \Pi$ be the set of policies whose stability we wish to establish. Our stability results rely on a nonnegative, continuously differentiable function E on score space, which we call an *energy* for S, and whose vanishing is equivalent to approaching S, in that, for every sequence of scores (y_k),

$$E(y_k) \to 0 \quad \Longleftrightarrow \quad \operatorname{dist}_1(\psi(y_k), S) \to 0 \,. \tag{A.2}$$

Moreover, the energy dissipates along the policy gradient field, as its drift is uniformly negative on every region of the form $\{\epsilon < E < \epsilon'\}$ with $0 < \epsilon < \epsilon' \leq \bar{E}$, namely, writing $\pi = \psi(y)$,

$$\dot{E}(y) \coloneqq \langle \nabla E(y), v(\pi) \rangle \leq -c(\epsilon, \epsilon') < 0 \,, \tag{A.3}$$

which makes E the analogue, in score space, of a Lyapunov function. Along the recursion (RPG), if E is κ-smooth, a Taylor expansion yields, at every iteration,

$$E(Y_{n+1}) \leq E(Y_n) + \gamma_n \dot{E}(Y_n) + \gamma_n \xi_n + \gamma_n \chi_n + \frac{\kappa}{2} \gamma_n^2 \|\widehat{v}_n\|_\infty^2 \,, \tag{A.4}$$

where

$$\xi_n = \langle \nabla E(Y_n), U_n \rangle \quad \text{and} \quad \chi_n = \langle \nabla E(Y_n), b_n \rangle \,. \tag{A.5}$$

The term $\dot{E}(Y_n)$ is the drift induced by the policy gradient, ξ_n is a martingale fluctuation, χ_n accounts for the bias, and the last term is the second-order Taylor remainder.

Accordingly, the argument unfolds in two stages. To begin, we stop the process at its first exit from a small sublevel set of the energy. Under Assumption 1, Doob's maximal inequality keeps the stopped martingale small, while the accumulated bias and Taylor remainders have a small expectation as soon as the base step size is small enough. With high probability, these errors are therefore unable to push the energy out of its sublevel set, and the process never leaves the region where the energy dissipates. Next, to establish attraction, we add to the stopped energy the dissipation accumulated so far, together with the expected errors still to come, which makes it a nonnegative supermartingale. As such, the energy converges, and the total dissipation that it undergoes is finite. Should its limit be positive, the drift would eventually be uniformly negative along steps whose sizes sum to infinity, which is absurd, and the energy, together with the distance to S, must therefore vanish in the limit.

Roadmap for the proof of Theorem 2. We now specialize to a Q-club state-action set $\mathcal{B}$ and let $\mathcal{C} \coloneqq \operatorname{span}(\mathcal{B})$. As it happens, the right energy is the Fenchel gap to $\mathcal{C}$, which, for $\pi = \psi(y)$, is defined together with the restricted response $\pi_{\mathcal{B}}$ by

$$\pi_{\mathcal{B}} \coloneqq \arg\max_{p \in \mathcal{C}} \{\langle y, p\rangle - h(p)\} \quad \text{and} \quad F_{\mathcal{C}}(y) \coloneqq \min_{p \in \mathcal{C}} \{h(p) + h^*(y) - \langle y, p\rangle\}\,. \tag{A.6}$$

The Fenchel gap is nonnegative and vanishes if and only if $\pi \in \mathcal{C}$. Its gradient is $\nabla F_{\mathcal{C}}(y) = \pi - \pi_{\mathcal{B}}$, and its derivative along the policy gradient field thus reads

$$\dot{F}_{\mathcal{C}}(y) = \langle v(\pi), \pi - \pi_{\mathcal{B}}\rangle = \sum_{i,s} d_\rho^\pi(s) \sum_{\alpha_i \in \mathcal{A}_i} \big(\pi_i(\alpha_i \mid s) - \pi_{\mathcal{B},i}(\alpha_i \mid s)\big) Q_i^\pi(s, \alpha_i)\,. \tag{A.7}$$

The Q-club condition is precisely what makes this expression negative near $\mathcal{C}$, through the combination of three facts. First, when π is close to $\mathcal{C}$, every allowed action $\beta_i \in \mathcal{B}_i(s)$ enjoys a uniform Q-value advantage $\delta_i^s > 0$ over every excluded action $\alpha_i \in \bar{\mathcal{B}}_i(s)$. Second, the reachability condition (R) keeps $d_\rho^\pi(s)$ bounded away from zero. Third, the decomposable structure of h implies that the restricted response $\pi_{\mathcal{B}}$ reallocates the mass of the excluded actions to the allowed ones, namely

$$\pi_i(\beta_i \mid s) \le \pi_{\mathcal{B},i}(\beta_i \mid s) \quad \text{and} \quad \pi_{\mathcal{B},i}(\alpha_i \mid s) = 0 \tag{A.8}$$

for every allowed action β_i and every excluded action α_i. The pairing (A.7) thus weighs the mass placed by π on excluded actions against the compensating mass placed by $\pi_{\mathcal{B}}$ on allowed actions, and, as the latter have uniformly larger Q-values, we obtain the sharp drift bound

$$\langle v(\pi), \pi - \pi_{\mathcal{B}}\rangle \le -c \operatorname{dist}_1(\pi, \mathcal{C}) \tag{A.9}$$

for all π sufficiently close to $\mathcal{C}$. The Fenchel gap is thereby a local energy for $\mathcal{C}$, and the general energy argument delivers the stochastic asymptotic stability of $\mathcal{C}$.

Roadmap for the proof of Theorem 3. The stability argument shows that the process eventually approaches $\mathcal{C}$, but it says nothing about how fast. For rates, we follow instead the argument of Boone and Mertikopoulos [12], which relies on linear energies. To that end, let $\mathcal{Z}$ be the set of vectors $z = e_{i\alpha}^s - e_{i\beta}^s$, where $\alpha \in \bar{\mathcal{B}}_i(s)$ is excluded and $\beta \in \mathcal{B}_i(s)$ is allowed at the same pair (i, s), and consider, for every $z \in \mathcal{Z}$ and every $M > 0$, the linear energy and the region

$$E_z(y) = \langle y, z\rangle \quad \text{and} \quad \mathcal{D}(M) = \{y \in \mathcal{Y} : \max_{z \in \mathcal{Z}} E_z(y) < -M\}\,. \tag{A.10}$$

Each linear energy is thus the score of an excluded action minus that of an allowed one, and $\mathcal{D}(M)$ is the region in which all of them lie below $-M$. The optimality conditions of the mirror map translate these energies into probabilities, as they imply that, for every such z,

$$\pi_i(\alpha \mid s) \le \varphi\big(b + E_z(y)\big)\,, \tag{A.11}$$

where $\pi = \psi(y)$, $b = \theta'(1)$, and φ is the rate function (26) of the kernel. As a result, for M large enough, the image $\psi(\mathcal{D}(M))$ lies in a neighborhood U of $\mathcal{C}$ on which the Q-club inequalities give $\langle v(\pi), z\rangle \le -c < 0$ uniformly over $z \in \mathcal{Z}$. As long as the process remains in U, the linear energies therefore decrease at a linear pace, and telescoping (A.1) gives

$$E_z(Y_n) \le E_z(Y_1) - c\Gamma_{n-1} + S_{n,z} + \bar{B}_n\,, \tag{A.12}$$

where $S_{n,z} = \sum_{k<n} \gamma_k \langle U_k, z\rangle$ collects the noise, $\bar{B}_n = 2\sum_{k<n} \gamma_k B_k$ bounds the contribution of the bias, and $\Gamma_n = \sum_{k=1}^n \gamma_k$, with $\Gamma_0 = 0$. The crux of the argument is that these errors grow sublinearly in the cumulative step size. Indeed, the Hájek–Rényi inequality and Assumption 2 provide deterministic constants $G > 0$ and $\alpha < 1$ such that, with probability at least $1 - \eta$,

$$S_{n,z} + \bar{B}_n \le G(1 + \Gamma_{n-1})^\alpha \tag{A.13}$$

for every episode $n \ge 1$ and every $z \in \mathcal{Z}$. The excess of this sublinear envelope over half the linear drift, namely

$$\Delta = \sup_{t \ge 0} \{G(1+t)^\alpha - ct/2\}\,, \tag{A.14}$$

is finite since $\alpha < 1$, and an initialization in $\mathcal{D}(M + \Delta)$ leaves enough room to absorb it. By induction on (A.12), all linear energies then satisfy $E_z(Y_n) < -M - c\Gamma_{n-1}/2$, which keeps every iterate in U, and the bound (A.11) on excluded probabilities yields

$$\operatorname{dist}_1(\pi_n, \mathcal{C}) \le c_0\, \varphi\big(c_1 - c_2 \Gamma_{n-1}\big)\,. \tag{A.15}$$

The regularizer only enters this bound through its rate function φ, whose exponential decay in the entropic case yields geometric convergence under constant step sizes, while its vanishing below a finite threshold in the Euclidean case yields the identification of the face in finite time.

B Auxiliary results

This appendix brings together the classical results invoked in the proofs. We begin with the conjugacy and quadratic growth properties of strongly convex functions [8].

Theorem B.1 (Strong convexity and conjugacy)**.** *Let $g\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ be proper, lower semi-continuous, and K-strongly convex with respect to $\|\cdot\|_1$, for some $K>0$. Then its convex conjugate g^* is continuously differentiable, and, for all $y, y' \in \mathbb{R}^d$,*

$$q(y) \coloneqq \nabla g^*(y) = \arg\max_p\{\langle y,p\rangle - g(p)\} \quad \textit{and} \quad \|q(y)-q(y')\|_1 \le K^{-1}\|y-y'\|_\infty\,. \tag{B.1}$$

Moreover, $y \in \partial g(q(y))$, and, for all $p \in \operatorname{dom} g$,

$$g(p) + g^*(y) - \langle y,p\rangle \ge \frac{K}{2}\|p-q(y)\|_1^2\,. \tag{B.2}$$

Next, we recall the descent lemma for smooth functions [8, 42], which we state for the pair of norms used throughout.

Lemma B.1 (Descent lemma)**.** *Let $E\colon \mathcal{Y}\to\mathbb{R}$ be continuously differentiable and κ-smooth, meaning that $\|\nabla E(y)-\nabla E(y')\|_1 \le \kappa\|y-y'\|_\infty$ for all $y, y' \in \mathcal{Y}$. Then, for all $y, z \in \mathcal{Y}$,*

$$E(y+z) \le E(y) + \langle \nabla E(y), z\rangle + \frac{\kappa}{2}\|z\|_\infty^2\,. \tag{B.3}$$

Finally, we recall two classical results from martingale theory [57], together with the martingale form of the Hájek–Rényi inequality [18], which controls the growth of a martingale relative to a deterministic sequence of weights.

Theorem B.2 (Doob's maximal inequality)**.** *Let $(M_k)_{k=0}^n$ be a square-integrable real-valued martingale. Then, for every $a>0$,*

$$\mathbb{P}\left(\max_{0\le k\le n}|M_k| \ge a\right) \le \frac{\mathbb{E}[M_n^2]}{a^2}\,. \tag{B.4}$$

Theorem B.3 (Hájek–Rényi maximal inequality)**.** *Let $(M_k)_{k\ge 0}$ be a square-integrable real-valued martingale with $M_0=0$, and let $(a_k)_{k\ge 1}$ be a deterministic, positive and nondecreasing sequence. Then, for every $\epsilon>0$,*

$$\mathbb{P}\left(\sup_{k\ge 1}\frac{|M_k|}{a_k} > \epsilon\right) \le \frac{1}{\epsilon^2}\sum_{k\ge 1}\frac{\mathbb{E}[(M_k-M_{k-1})^2]}{a_k^2}\,. \tag{B.5}$$

Theorem B.4 (Supermartingale convergence theorem)**.** *Every nonnegative supermartingale converges almost surely to a finite random variable.*

C Regularizers, mirror maps, and Fenchel geometry

Recall that, for each player i, we endow Π_i with a *decomposable regularizer*, which we now allow to vary across players and states, namely

$$h_i(\pi_i) = \sum_{s\in\mathcal{S}}\sum_{\alpha_i\in\mathcal{A}_i}\theta_i^s(\pi_i(\alpha_i \mid s))\,, \tag{C.1}$$

where each kernel $\theta_i^s\colon[0,1]\to\mathbb{R}$ is continuous, continuously differentiable on $(0,1]$, and twice differentiable on $(0,1)$, with $(\theta_i^s)''\ge K_i^s$ on $(0,1)$ for some $K_i^s>0$. We aggregate these regularizers into $h(\pi)=\sum_{i\in\mathcal{N}}h_i(\pi_i)$, which we extend by $+\infty$ outside Π. Let us denote by d the ambient dimension and set

$$d \coloneqq \sum_{i\in\mathcal{N}}|\mathcal{S}||\mathcal{A}_i| \quad\text{and}\quad K \coloneqq \frac{1}{d}\min_{i,s}K_i^s\,. \tag{C.2}$$

As $\|z\|_2^2 \ge \|z\|_1^2/d$ for every vector z of the ambient space, the lower bounds on the second derivatives imply that h is K-strongly convex with respect to $\|\cdot\|_1$. We write $\partial h(\pi)$ for the subdifferential of h at π, that is, the set of vectors g such that $h(p)\ge h(\pi)+\langle g, p-\pi\rangle$ for all $p\in\Pi$, and $\operatorname{dom}\partial h$ for

the set of policies at which it is nonempty. The convex conjugate and the mirror map of h are then given by

$$h^*(y) = \max_{\pi\in\Pi}\{\langle y,\pi\rangle - h(\pi)\} \quad \text{and} \quad \psi(y) = \arg\max_{\pi\in\Pi}\{\langle y,\pi\rangle - h(\pi)\}\,, \tag{C.3}$$

and we say that h is *steep* if $(\theta_i^s)'(p) \to -\infty$ as $p \downarrow 0$ for every player i and state s. Finally, for every pair (i,s), we let

$$a_i^s := \lim_{t\downarrow 0}(\theta_i^s)'(t) \in [-\infty,\infty) \quad \text{and} \quad b_i^s := (\theta_i^s)'(1)\,, \tag{C.4}$$

and we define the rate function φ_i^s of θ_i^s by (26), with a_i^s and b_i^s in place of a and b. As $(\theta_i^s)'$ is strictly increasing on $(0,1]$, the value $\varphi_i^s(u)$ is the unique maximizer of $t \mapsto ut - \theta_i^s(t)$ on $[0,1]$. In particular, φ_i^s is nondecreasing, and $\varphi_i^s(u) \to 0$ as $u \to -\infty$.

C.1 Mirror maps and primal mirror form

Proposition C.1. *The following statements hold.*

(a) *The maximum in (C.3) is attained at a unique point, and $\pi = \psi(y)$ if and only if $y \in \partial h(\pi)$.*

(b) $\operatorname{Im}\psi = \operatorname{dom}\partial h$ *and* $\operatorname{ri}\Pi \subseteq \operatorname{Im}\psi \subseteq \Pi$. *If h is steep, then* $\operatorname{Im}\psi = \operatorname{ri}\Pi$.

(c) h^* *is differentiable,* $\nabla h^* = \psi$, *and ψ is $(1/K)$-Lipschitz from $(\mathcal{Y}, \|\cdot\|_\infty)$ to $(\Pi, \|\cdot\|_1)$.*

(d) *For every pair (i,s), if $q = \psi_i^s(y)$, then there exists $\lambda \in \mathbb{R}$ such that, for every action $\alpha \in \mathcal{A}_i$,*

$$q_\alpha = \varphi_i^s(y_\alpha - \lambda)\,. \tag{C.5}$$

In particular, $y_\alpha - (\theta_i^s)'(q_\alpha) = \lambda$ whenever $q_\alpha > 0$, whereas $q_\alpha = 0$ forces $a_i^s > -\infty$ and $y_\alpha - a_i^s \le \lambda$.

Proof. Statements (a)–(c) are classical [42, 46]. Indeed, as h is proper, lower semicontinuous and K-strongly convex with respect to $\|\cdot\|_1$, Theorem B.1 applies to $g = h$, and it yields (a), (c) and the identity $\operatorname{Im}\psi = \operatorname{dom}\partial h$, while the remaining assertions of (b) follow from the differentiability of the kernels on $(0,1]$ and, under steepness, from their infinite slope at zero. We thus focus on (d).

Fix a pair (i,s) and suppress its indices, so that θ, a, b and φ stand for θ_i^s, a_i^s, b_i^s and φ_i^s. The term $q = \psi_i^s(y)$ maximizes the concave function $f(p) = \sum_\alpha[y_\alpha p_\alpha - \theta(p_\alpha)]$ over the simplex. Fix an action α with $q_\alpha > 0$, and let $\lambda := y_\alpha - \theta'(q_\alpha)$. For every action $\alpha' \neq \alpha$, we have $q_{\alpha'} \le 1 - q_\alpha < 1$, so that $q + t(e_{\alpha'} - e_\alpha)$ lies in the simplex for $0 \le t \le q_\alpha$, and the optimality of q yields

$$0 \ge \lim_{t\downarrow 0}\frac{f(q + t(e_{\alpha'} - e_\alpha)) - f(q)}{t} = \big(y_{\alpha'} - \theta'_+(q_{\alpha'})\big) - \lambda\,, \tag{C.6}$$

where $\theta'_+(q_{\alpha'})$ is the right derivative of θ at $q_{\alpha'}$, which equals $\theta'(q_{\alpha'})$ if $q_{\alpha'} > 0$, and a if $q_{\alpha'} = 0$, by the mean value theorem. If $q_{\alpha'} > 0$, the same argument applies with the roles of α and α' exchanged, and therefore $y_{\alpha'} - \theta'(q_{\alpha'}) = \lambda$. If $q_{\alpha'} = 0$, then (C.6) reads $y_{\alpha'} - a \le \lambda$, which forces $a > -\infty$. This proves the last assertion of (d), and (C.5) follows by distinguishing three cases, for every action α', including α itself. If $0 < q_{\alpha'} < 1$, then $\theta'(q_{\alpha'}) = y_{\alpha'} - \lambda$ lies in (a,b), as θ' is strictly increasing on $(0,1]$, hence $q_{\alpha'} = (\theta')^{-1}(y_{\alpha'} - \lambda) = \varphi(y_{\alpha'} - \lambda)$. If $q_{\alpha'} = 1$, then $y_{\alpha'} - \lambda = \theta'(1) = b$, and $\varphi(b) = 1$. Finally, if $q_{\alpha'} = 0$, then $y_{\alpha'} - \lambda \le a$, and $\varphi(y_{\alpha'} - \lambda) = 0$. ∎

Next, for $p \in \Pi$ and $\pi \in \operatorname{ri}\Pi$, the Bregman divergence generated by h is

$$D(p,\pi) = h(p) - h(\pi) - \langle \nabla h(\pi), p - \pi\rangle\,. \tag{C.7}$$

The next result shows that, under steepness, the score recursion (RPG) can be written in the primal form (MPG).

Proposition C.2. *Assume that h is steep, let $\pi_n = \psi(Y_n)$, and let $\widehat{v}_n \in \mathcal{Y}$. Then the score update*

$$Y_{n+1} = Y_n + \gamma_n \widehat{v}_n \quad \textit{and} \quad \pi_{n+1} = \psi(Y_{n+1}) \tag{C.8}$$

is equivalent to the primal update

$$\pi_{n+1} = \arg\max_{\pi\in\Pi}\{\gamma_n\langle \widehat{v}_n, \pi\rangle - D(\pi,\pi_n)\}\,. \tag{C.9}$$

Proof. By Proposition C.1(b), steepness ensures that $\pi_n \in \operatorname{ri}\Pi$. Proposition C.1(d) then provides multipliers $\lambda_i^s \in \mathbb{R}$ such that $(\theta_i^s)'(\pi_{i,n}(\alpha \mid s)) = Y^s_{i\alpha,n} - \lambda_i^s$ for every triple (i,s,α), that is, $\nabla h(\pi_n) = Y_n - \lambda_n$, where $\lambda_n \in \mathcal{Y}$ has coordinate λ_i^s at every action of the pair (i,s). As the coordinates of every policy sum to one at each pair, every $\pi \in \Pi$ satisfies

$$\langle \lambda_n, \pi - \pi_n \rangle = \sum_{i,s} \lambda_i^s \sum_{\alpha} \big(\pi_i(\alpha \mid s) - \pi_{i,n}(\alpha \mid s)\big) = 0\,, \tag{C.10}$$

so that $D(\pi, \pi_n) = h(\pi) - h(\pi_n) - \langle Y_n, \pi - \pi_n \rangle$, and therefore

$$\gamma_n \langle \widehat{v}_n, \pi \rangle - D(\pi, \pi_n) = \langle Y_{n+1}, \pi \rangle - h(\pi) + h(\pi_n) - \langle Y_n, \pi_n \rangle\,. \tag{C.11}$$

The last two terms do not depend on π, and, by Proposition C.1(a), the unique maximizer over Π of the first two is $\psi(Y_{n+1}) = \pi_{n+1}$. ∎

C.2 Fenchel coupling and Fenchel gap

We may now introduce the *Fenchel coupling* associated with h, which is defined, for all $p \in \Pi$ and $y \in \mathcal{Y}$, as

$$F(p,y) = h(p) + h^*(y) - \langle y, p \rangle\,, \tag{C.12}$$

and, given a nonempty compact convex set $\mathcal{C} \subseteq \Pi$, we define the restricted mirror map and the Fenchel gap to $\mathcal{C}$ as

$$\psi_{\mathcal{C}}(y) = \operatorname*{arg\,max}_{p \in \mathcal{C}} \{\langle y, p \rangle - h(p)\} \quad \text{and} \quad F_{\mathcal{C}}(y) = \min_{p \in \mathcal{C}} F(p,y) = F(\psi_{\mathcal{C}}(y), y)\,. \tag{C.13}$$

The following lemmas contain the only properties of the Fenchel gap that we use in the sequel.

Lemma C.1. *The Fenchel gap enjoys the following properties.*

(a) *For every $y \in \mathcal{Y}$, the point $\psi_{\mathcal{C}}(y)$ exists and is unique.*

(b) *$F_{\mathcal{C}}$ is continuously differentiable on $\mathcal{Y}$, with $\nabla F_{\mathcal{C}}(y) = \psi(y) - \psi_{\mathcal{C}}(y)$.*

(c) *$F_{\mathcal{C}}$ is $(2/K)$-smooth, in the sense that, for all $y, y' \in \mathcal{Y}$,*

$$\|\nabla F_{\mathcal{C}}(y) - \nabla F_{\mathcal{C}}(y')\|_1 \le \frac{2}{K} \|y - y'\|_\infty\,. \tag{C.14}$$

(d) *For every $y \in \mathcal{Y}$,*

$$F_{\mathcal{C}}(y) \ge \frac{K}{2} \|\psi(y) - \psi_{\mathcal{C}}(y)\|_1^2 \ge \frac{K}{2} \operatorname{dist}_1(\psi(y), \mathcal{C})^2\,. \tag{C.15}$$

(e) *For every $y \in \mathcal{Y}$,*

$$\|\nabla F_{\mathcal{C}}(y)\|_1 \le \sqrt{2F_{\mathcal{C}}(y)/K}\,, \tag{C.16}$$

and $\nabla F_{\mathcal{C}}$ is therefore bounded on every bounded sublevel set of $F_{\mathcal{C}}$.

Proof. Let $h_{\mathcal{C}}$ denote the restriction of h to $\mathcal{C}$, extended by $+\infty$ outside $\mathcal{C}$. Both h and $h_{\mathcal{C}}$ are proper, lower semicontinuous and K-strongly convex with respect to $\|\cdot\|_1$, so that Theorem B.1 applies to each of them. In particular, the maximizer $\psi_{\mathcal{C}}(y)$ exists and is unique, which is (a), and $\psi = \nabla h^*$, while $\psi_{\mathcal{C}} = \nabla h_{\mathcal{C}}^*$. Moreover, the definitions (C.12) and (C.13) give, for every $y \in \mathcal{Y}$,

$$F_{\mathcal{C}}(y) = h^*(y) - \max_{p \in \mathcal{C}} \{\langle y, p \rangle - h(p)\} = h^*(y) - h_{\mathcal{C}}^*(y)\,, \tag{C.17}$$

so that $F_{\mathcal{C}}$ is continuously differentiable, with $\nabla F_{\mathcal{C}} = \psi - \psi_{\mathcal{C}}$, which is (b). As both mirror maps are $(1/K)$-Lipschitz from $(\mathcal{Y}, \|\cdot\|_\infty)$ to $(\Pi, \|\cdot\|_1)$, we then have, for all $y, y' \in \mathcal{Y}$,

$$\|\nabla F_{\mathcal{C}}(y) - \nabla F_{\mathcal{C}}(y')\|_1 \le \|\psi(y) - \psi(y')\|_1 + \|\psi_{\mathcal{C}}(y) - \psi_{\mathcal{C}}(y')\|_1 \le \frac{2}{K} \|y - y'\|_\infty\,, \tag{C.18}$$

which is (c). Next, as $F_{\mathcal{C}}(y) = F(\psi_{\mathcal{C}}(y), y)$, the quadratic growth bound (B.2), applied to $g = h$ at the point $p = \psi_{\mathcal{C}}(y) \in \mathcal{C}$, gives

$$F_{\mathcal{C}}(y) \ge \frac{K}{2} \|\psi(y) - \psi_{\mathcal{C}}(y)\|_1^2 \ge \frac{K}{2} \operatorname{dist}_1(\psi(y), \mathcal{C})^2\,, \tag{C.19}$$

which is (d). Finally, (b) and the first inequality of (C.19) yield $\|\nabla F_{\mathcal{C}}(y)\|_1 = \|\psi(y) - \psi_{\mathcal{C}}(y)\|_1 \le \sqrt{2F_{\mathcal{C}}(y)/K}$, which is (e). ∎

Lemma C.2. *Let (y_k) be a sequence in $\mathcal{Y}$. Then $F_{\mathcal{C}}(y_k) \to 0$ if and only if $\operatorname{dist}_1(\psi(y_k), \mathcal{C}) \to 0$.*

Proof. The direct implication is immediate from (C.15), which gives $\operatorname{dist}_1(\psi(y_k), \mathcal{C})^2 \leq 2F_{\mathcal{C}}(y_k)/K$. The converse requires a little more care. Assume that $\operatorname{dist}_1(\pi_k, \mathcal{C}) \to 0$, where $\pi_k \coloneqq \psi(y_k)$, and recall that a real sequence converges to zero as soon as each of its subsequences admits a further subsequence that does. Given a subsequence, the compactness of Π provides a further subsequence, which we still index by k, along which π_k converges to some $p \in \Pi$, and p belongs to $\mathcal{C}$, since $\mathcal{C}$ is closed and $\operatorname{dist}_1(\pi_k, \mathcal{C}) \to 0$. By the definition (C.13) of the Fenchel gap, we then have

$$0 \leq F_{\mathcal{C}}(y_k) \leq F(p, y_k)\,, \tag{C.20}$$

and it is therefore enough to show that $F(p, y_k) \to 0$ along this subsequence.

To that end, we first express the Fenchel coupling through the policies π_k. As $h^*(y_k) = \langle y_k, \pi_k \rangle - h(\pi_k)$, the definition (C.12) gives

$$F(p, y_k) = h(p) - h(\pi_k) - \langle y_k, p - \pi_k \rangle\,, \tag{C.21}$$

which decomposes across the pairs (i, s). We may thus fix one of them and suppress its indices, writing θ for its kernel. By Proposition C.1(d), there exists $\lambda_k \in \mathbb{R}$ such that $y_{k,\alpha} - \lambda_k = \theta'(\pi_{k,\alpha})$ whenever $\pi_{k,\alpha} > 0$. Since $\sum_\alpha (p_\alpha - \pi_{k,\alpha}) = 0$, this multiplier may be inserted at no cost, and the contribution of the pair to (C.21) reads $\sum_\alpha \Phi_{k,\alpha}$, where

$$\Phi_{k,\alpha} \coloneqq \theta(p_\alpha) - \theta(\pi_{k,\alpha}) - (y_{k,\alpha} - \lambda_k)(p_\alpha - \pi_{k,\alpha})\,. \tag{C.22}$$

It remains to show that every term $\Phi_{k,\alpha}$ tends to zero, and two cases arise.

(i) If $p_\alpha > 0$, then $\pi_{k,\alpha} > 0$ for all sufficiently large k, and, for such k,

$$\Phi_{k,\alpha} = \theta(p_\alpha) - \theta(\pi_{k,\alpha}) - \theta'(\pi_{k,\alpha})(p_\alpha - \pi_{k,\alpha})\,, \tag{C.23}$$

which tends to zero by continuity of θ and θ' at p_α.

(ii) If $p_\alpha = 0$, then $\Phi_{k,\alpha} = 0$ whenever $\pi_{k,\alpha} = 0$, whereas, whenever $\pi_{k,\alpha} > 0$,

$$\Phi_{k,\alpha} = \big[\theta(0) - \theta(\pi_{k,\alpha})\big] + \pi_{k,\alpha}\theta'(\pi_{k,\alpha})\,. \tag{C.24}$$

As $\pi_{k,\alpha} \to 0$, the bracket vanishes in the limit by continuity of θ at zero.

The last term of (C.24) calls for a closer look, since $\theta'(t)$ may well diverge to $-\infty$ as $t \downarrow 0$, and we now check that $t\theta'(t) \to 0$ nonetheless. Fix $t \in (0, 1/2]$. As θ is convex, its slopes are nondecreasing, so that the secant over $[t/2, t]$, the derivative at t and the secant over $[t, 2t]$ are ordered as

$$\frac{\theta(t) - \theta(t/2)}{t/2} \leq \theta'(t) \leq \frac{\theta(2t) - \theta(t)}{t}\,. \tag{C.25}$$

Multiplying by $t > 0$, we obtain

$$2\big[\theta(t) - \theta(t/2)\big] \leq t\theta'(t) \leq \theta(2t) - \theta(t)\,, \tag{C.26}$$

and both bounds tend to zero as $t \downarrow 0$ by continuity of θ at zero, which forces $t\theta'(t) \to 0$. Applied along $t = \pi_{k,\alpha}$, this shows that the terms of the second case vanish in the limit as well. Summing over the finitely many actions and pairs, we conclude that $F(p, y_k) \to 0$ along the subsequence, which completes the proof. ∎

C.3 Examples of regularizers

We close this appendix with some standard regularizers [9, 12, 31] which satisfy our assumptions. For each of them, we write $\pi_i = \psi_i(Y_i)$, and the Bregman divergences are computed at $\pi_i' \in \operatorname{ri}\Pi_i$.

Example C.1 (Entropy)**.** Let $\theta(p) = p\log p$, with the convention $0\log 0 = 0$, whose second derivative is $\theta''(p) = 1/p \geq 1$. This kernel is steep, its mirror map is the softmax map

$$\pi_i(\alpha_i \mid s) = \frac{\exp(Y_{i\alpha_i}^s)}{\sum_{\beta_i \in \mathcal{A}_i} \exp(Y_{i\beta_i}^s)}\,, \tag{C.27}$$

and the induced Bregman divergence is the KL divergence

$$D(\pi_i, \pi_i') = \sum_{s \in \mathcal{S}} \sum_{\alpha_i \in \mathcal{A}_i} \pi_i(\alpha_i \mid s) \log \frac{\pi_i(\alpha_i \mid s)}{\pi_i'(\alpha_i \mid s)}\,, \tag{C.28}$$

in which case (MPG) reduces to (KL-PG). ✦

Example C.2 (Euclidean)**.** Let $\theta(p) = p^2/2$, for which $\theta''(p) = 1$. As $\theta'(0) = 0$ is finite, this kernel is not steep, its mirror map is the Euclidean projection $\pi_i(\cdot \mid s) = \mathrm{proj}_{\Delta(\mathcal{A}_i)}(Y_i^s)$ onto the simplex, and the induced Bregman divergence is $D(\pi_i, \pi_i') = \frac{1}{2}\|\pi_i - \pi_i'\|_2^2$. ✦

Example C.3 (Square root)**.** Let $\theta(p) = -4\sqrt{p}$, which satisfies $\theta''(p) = p^{-3/2} \geq 1$. In contrast to the Euclidean case, this kernel is steep, since $\theta'(p) = -2/\sqrt{p} \to -\infty$ as $p \downarrow 0$, and the induced Bregman divergence is

$$D(\pi_i, \pi_i') = 2 \sum_{s \in \mathcal{S}} \sum_{\alpha_i \in \mathcal{A}_i} \frac{\left(\sqrt{\pi_i(\alpha_i \mid s)} - \sqrt{\pi_i'(\alpha_i \mid s)}\right)^2}{\sqrt{\pi_i'(\alpha_i \mid s)}}\,, \tag{C.29}$$

and its mirror map, like the softmax map, takes values in the relative interior of the policy space. ✦

Example C.4 (Tsallis entropy)**.** Fix $q \in (0,1) \cup (1,2]$, and let $\theta_q(p) = p^q/(q(q-1))$, with $\theta_q''(p) = p^{q-2} \geq 1$. Here, the kernel is steep for $q \in (0,1)$ but not for $q \in (1,2]$, and the induced Bregman divergence is

$$D(\pi_i, \pi_i') = \frac{1}{q(q-1)} \sum_{s \in \mathcal{S}} \sum_{\alpha_i \in \mathcal{A}_i} \Big[\pi_i(\alpha_i \mid s)^q - q\,\pi_i'(\alpha_i \mid s)^{q-1}\pi_i(\alpha_i \mid s) + (q-1)\,\pi_i'(\alpha_i \mid s)^q\Big]\,. \tag{C.30}$$

This family contains the Euclidean kernel as the case $q = 2$ and the square root kernel as the case $q = 1/2$, while the entropic kernel is recovered in the limit $q \to 1$, up to affine terms on the simplex, which affect neither the mirror map nor the Bregman divergence. ✦

D Energy functions and stochastic stability

In this appendix, we develop the energy machinery behind our stability results. Mertikopoulos et al. [42] proved that the existence of a local energy function leads to stochastic attraction, and we revisit and strengthen their argument so as to obtain stochastic stability as well. As the statement of Theorem D.1 below makes clear, the neighborhood of initializations depends only on the target neighborhood U, whereas the threshold on the base step size depends on both the tolerance level η and U, in accordance with the convention adopted after Definition 1. We first establish the regularity properties of the policy gradient field that are used throughout.

D.1 Regularity of the policy gradient field

Recall that the stopping probabilities are bounded from below by

$$\zeta_{\min} = \min_{s,\alpha}\Big(1 - \sum_{s' \in \mathcal{S}} P(s' \mid s, \alpha)\Big) > 0\,. \tag{D.1}$$

For a matrix A indexed by $\mathcal{S} \times \mathcal{S}$, we write $\|A\|_\infty = \max_s \sum_{s'} |A(s,s')|$ for its operator norm on $(\mathbb{R}^{\mathcal{S}}, \|\cdot\|_\infty)$, and we recall that $\|u^\top A\|_1 \leq \|u\|_1 \|A\|_\infty$ for every vector $u \in \mathbb{R}^{\mathcal{S}}$. As every row of P^π sums to at most $1 - \zeta_{\min}$, the Neumann series gives, for every $\pi \in \Pi$,

$$\|P^\pi\|_\infty \leq 1 - \zeta_{\min} \quad \text{and} \quad \|(I - P^\pi)^{-1}\|_\infty = \left\|\sum_{t \geq 0} (P^\pi)^t\right\|_\infty \leq \sum_{t \geq 0} (1 - \zeta_{\min})^t = \frac{1}{\zeta_{\min}}\,. \tag{D.2}$$

Lemma D.1. *The policy gradient field is bounded and Lipschitz continuous on* Π*. More precisely, for all* $\pi, \pi' \in \Pi$*,*

$$\|v(\pi)\|_\infty \leq M_v \quad \textit{and} \quad \|v(\pi) - v(\pi')\|_\infty \leq L_v \|\pi - \pi'\|_1\,, \tag{D.3}$$

where $M_v \coloneqq 1/\zeta_{\min}^2$ *and* $L_v \coloneqq 4/\zeta_{\min}^3$*.*

Proof. This estimate is standard [26], and we only recall the computation behind it, which also provides Lipschitz bounds for the values, the visitation counts and the Q-values. Fix $\pi, \pi' \in \Pi$, and let $\delta \coloneqq \|\pi - \pi'\|_1$. As actions are drawn independently across players, telescoping the products $\pi(\alpha \mid s) = \prod_j \pi_j(\alpha_j \mid s)$ over the players gives $\sum_\alpha |\pi(\alpha \mid s) - \pi'(\alpha \mid s)| \leq \sum_j \|\pi_j(\cdot \mid s) - \pi_j'(\cdot \mid$

$s)\|_1 \le \delta$ at every state, and likewise for the profiles π_{-i} and π'_{-i} of the opponents. As $|r_i| \le 1$ and $\sum_{s'} P(s' \mid s, \alpha) \le 1$, the expected stage rewards $r_i^\pi(s) := \sum_\alpha \pi(\alpha \mid s) r_i(s,\alpha)$ and the induced transition matrices thus satisfy $\|r_i^\pi - r_i^{\pi'}\|_\infty \le \delta$ and $\|P^\pi - P^{\pi'}\|_\infty \le \delta$. The values and the visitation counts are then handled by the resolvent identities

$$\begin{aligned} V_i(\pi) - V_i(\pi') &= (I - P^\pi)^{-1}\Big[r_i^\pi - r_i^{\pi'} + (P^\pi - P^{\pi'})V_i(\pi')\Big]\,, \\ d_\rho^\pi - d_\rho^{\pi'} &= \rho^\top (I - P^\pi)^{-1}(P^\pi - P^{\pi'})(I - P^{\pi'})^{-1}\,, \end{aligned} \tag{D.4}$$

which follow from $V_i(\pi) = (I - P^\pi)^{-1} r_i^\pi$ and $d_\rho^\pi = \rho^\top (I - P^\pi)^{-1}$. Together with (D.2), they give $\|V_i(\pi)\|_\infty \le 1/\zeta_{\min}$ and $\|d_\rho^\pi\|_1 \le 1/\zeta_{\min}$, as well as

$$\|V_i(\pi) - V_i(\pi')\|_\infty \le \frac{\delta}{\zeta_{\min}} + \frac{\delta}{\zeta_{\min}^2} \le \frac{2\delta}{\zeta_{\min}^2} \quad \text{and} \quad \|d_\rho^\pi - d_\rho^{\pi'}\|_1 \le \frac{\delta}{\zeta_{\min}^2}\,. \tag{D.5}$$

Next, by (4), the Q-value $Q_i^\pi(s,\alpha_i)$ is the average, under $\pi_{-i}(\cdot \mid s)$, of $g_i^\pi(s,\alpha) := r_i(s,\alpha) + \sum_{s'} P(s' \mid s,\alpha) V_{i,s'}(\pi)$, which satisfies $|g_i^\pi| \le 1 + (1 - \zeta_{\min})/\zeta_{\min} = 1/\zeta_{\min}$ and $|g_i^\pi - g_i^{\pi'}| \le 2\delta/\zeta_{\min}^2$. Hence $|Q_i^\pi| \le 1/\zeta_{\min}$, and

$$\begin{aligned} |Q_i^\pi(s,\alpha_i) - Q_i^{\pi'}(s,\alpha_i)| &\le \sum_{\alpha_{-i}} |\pi_{-i}(\alpha_{-i} \mid s) - \pi'_{-i}(\alpha_{-i} \mid s)||g_i^\pi(s,\alpha)| \\ &\qquad + \sum_{\alpha_{-i}} \pi'_{-i}(\alpha_{-i} \mid s)|g_i^\pi(s,\alpha) - g_i^{\pi'}(s,\alpha)| \\ &\le \frac{\delta}{\zeta_{\min}} + \frac{2\delta}{\zeta_{\min}^2} \le \frac{3\delta}{\zeta_{\min}^2}\,, \end{aligned} \tag{D.6}$$

where $\alpha = (\alpha_i, \alpha_{-i})$. Finally, as $v_{i\alpha_i}^s(\pi) = d_\rho^\pi(s) Q_i^\pi(s,\alpha_i)$, the above bounds give $|v_{i\alpha_i}^s(\pi)| \le 1/\zeta_{\min}^2$ and

$$\begin{aligned} |v_{i\alpha_i}^s(\pi) - v_{i\alpha_i}^s(\pi')| &\le |d_\rho^\pi(s) - d_\rho^{\pi'}(s)||Q_i^\pi(s,\alpha_i)| + d_\rho^{\pi'}(s)|Q_i^\pi(s,\alpha_i) - Q_i^{\pi'}(s,\alpha_i)| \\ &\le \frac{\delta}{\zeta_{\min}^3} + \frac{3\delta}{\zeta_{\min}^3} = \frac{4\delta}{\zeta_{\min}^3}\,, \end{aligned} \tag{D.7}$$

which is (D.3). ■

D.2 Local energy functions

Throughout this subsection, $S \subseteq \Pi$ is nonempty and closed.

Definition D.1. We say that a map $E\colon \mathcal{Y} \to [0,\infty)$ is a *local energy* for S if the following conditions hold.

(a) E is continuously differentiable.

(b) For every sequence (y_k) in $\mathcal{Y}$, we have $E(y_k) \to 0$ if and only if $\operatorname{dist}_1(\psi(y_k), S) \to 0$.

(c) There exists $\bar{E} > 0$ such that, for all $0 < \epsilon < \epsilon' \le \bar{E}$,

$$\sup\{\langle \nabla E(y), v(\psi(y))\rangle : \epsilon < E(y) < \epsilon'\} < 0\,. \tag{D.8}$$

We further say that E is *κ-smooth* if, for all $y, y' \in \mathcal{Y}$,

$$\|\nabla E(y) - \nabla E(y')\|_1 \le \kappa \|y - y'\|_\infty\,, \tag{D.9}$$

and that E has *bounded gradients on bounded sublevels* if, for every $E_+ > 0$,

$$L(E_+) := \sup\{\|\nabla E(y)\|_1 : E(y) \le E_+\} < \infty\,. \tag{D.10}$$

Lemma D.2. *Assume that Definition D.1(b) holds. Then the following two properties hold.*

(a) *For every neighborhood U of S in Π, there exists $\delta_U > 0$ such that $E(y) < \delta_U$ implies $\psi(y) \in U$.*

(b) *For every $\delta > 0$, there exists a neighborhood V_δ of S in Π such that $\psi(y) \in V_\delta$ implies $E(y) < \delta$.*

Proof. For (a), we may assume, upon shrinking it, that U is open and distinct from Π. The set $\Pi \setminus U$ is then compact and disjoint from S, which is itself compact as a closed subset of Π, and it therefore lies at a positive distance r_U from S. If the claim failed, there would exist a sequence (y_k) with $E(y_k) \to 0$ and $\psi(y_k) \notin U$, hence with $\mathrm{dist}_1(\psi(y_k), S) \geq r_U$ for all k, in contradiction with Definition D.1(b). Likewise, if (b) failed for some $\delta > 0$, testing the neighborhoods $\{\pi \in \Pi : \mathrm{dist}_1(\pi, S) < 1/k\}$ would produce a sequence (y_k) with $\mathrm{dist}_1(\psi(y_k), S) < 1/k$ and $E(y_k) \geq \delta$ for all k, which again contradicts Definition D.1(b). ∎

D.3 Stochastic asymptotic stability

We are now ready for the main result of this appendix, in which the descent lemma and the martingale results of Appendix B join forces.

Theorem D.1. *Let $S \subseteq \Pi$ be nonempty and closed, and suppose that S admits a κ-smooth local energy E with bounded gradients on bounded sublevels. Under Assumption 1, for every tolerance level $\eta \in (0,1)$ and every neighborhood U of S, there exist a neighborhood V of S and a threshold $\bar{\gamma} > 0$ with the following property. For every initialization $\pi_1 = \psi(Y_1) \in V \cap \mathrm{Im}\,\psi$ and every base step size $\gamma \in (0, \bar{\gamma}]$, the event that $\pi_n \in U$ for all $n \geq 1$ and $\mathrm{dist}_1(\pi_n, S) \to 0$ has probability at least $1 - \eta$. In particular, S is stochastically asymptotically stable.*

Proof. Fix a tolerance level $\eta \in (0,1)$ and a neighborhood U of S in Π. The proof proceeds in four steps, which lead us from a local picture of the energy to the long-run behavior of the iterates.

Step 1: Localization. By Lemma D.2(a), there exists $\delta_U > 0$ such that $E(y) < \delta_U$ implies $\psi(y) \in U$, and we set

$$E_+ := \frac{1}{2}\min\{\bar{E}, \delta_U\} \quad \text{and} \quad L := L(E_+)\,, \tag{D.11}$$

so that $2E_+ \leq \bar{E}$, while $\|\nabla E(y)\|_1 \leq L$ whenever $E(y) \leq E_+$. By Lemma D.2(b), there also exists a neighborhood V of S, which depends only on U, such that $\psi(y) \in V$ implies $E(y) < E_+/4$. We now fix an initialization Y_1 with $\pi_1 = \psi(Y_1) \in V$, and we write $E_n := E(Y_n)$, with $E_1 < E_+/4$ by construction. Our object of study is the first exit from the sublevel set $\{E \leq E_+\}$, together with the corresponding stopped energy, namely

$$T_+ := \inf\{n \geq 1 : E_n > E_+\}\,, \qquad I_n := \mathbf{1}\{n < T_+\}\,, \quad \text{and} \quad \tilde{E}_n := E_{n \wedge T_+}\,, \tag{D.12}$$

with the convention $\inf \varnothing = \infty$.

Step 2: Energy inequality. Write $\dot{E}_n := \langle \nabla E(Y_n), v(\pi_n)\rangle$ for the drift and $\xi_n := \langle \nabla E(Y_n), U_n\rangle$ for the noise term. The drift is nonpositive whenever $E_n < \bar{E}$, as follows from (D.8) at positive energies, and from the fact that ∇E vanishes at the zeros of the nonnegative differentiable function E. On the event $\{n < T_+\}$, we have $E_n \leq E_+$, hence $\|\nabla E(Y_n)\|_1 \leq L$, and the descent lemma (Lemma B.1), applied with $z = \gamma_n \widehat{v}_n$, together with the decomposition (7), gives

$$\begin{aligned} E_{n+1} &\leq E_n + \gamma_n \langle \nabla E(Y_n), v(\pi_n) + U_n + b_n\rangle + \frac{\kappa}{2}\gamma_n^2 \|\widehat{v}_n\|_\infty^2 \\ &\leq E_n + \gamma_n \dot{E}_n + \gamma_n \xi_n + L\gamma_n B_n + \frac{\kappa}{2}\gamma_n^2 \|\widehat{v}_n\|_\infty^2\,, \end{aligned} \tag{D.13}$$

where the second inequality follows from $|\langle \nabla E(Y_n), b_n\rangle| \leq \|\nabla E(Y_n)\|_1 \|b_n\|_\infty \leq LB_n$. As $\tilde{E}_n = E_n$ and $\tilde{E}_{n+1} = E_{n+1}$ on this event, we obtain

$$\tilde{E}_{n+1} \leq \tilde{E}_n - d_n + (M_{n+1} - M_n) + r_n\,, \tag{D.14}$$

where

$$\begin{aligned} d_n &:= -I_n \gamma_n \dot{E}_n\,, \qquad M_n := \sum_{k<n} I_k \gamma_k \xi_k\,, \\ r_n &:= I_n\Big(L\gamma_n B_n + \frac{\kappa}{2}\gamma_n^2 \|\widehat{v}_n\|_\infty^2\Big)\,, \quad \text{and} \quad R_n := \sum_{k<n} r_k\,. \end{aligned} \tag{D.15}$$

On the event $\{n \geq T_+\}$, the stopped energy no longer moves, while $I_n = 0$ makes all the other terms of (D.14) vanish, so that this inequality holds everywhere. Moreover, $d_n \geq 0$, since $E_n \leq E_+ < \bar{E}$ on $\{n < T_+\}$.

We now bound the error terms. As I_n and $\nabla E(Y_n)$ are $\mathcal{F}_n$-measurable and $\mathbb{E}[U_n \,|\, \mathcal{F}_n] = 0$, the process (M_n) is a martingale. By the conditional Jensen inequality, (S) gives $\mathbb{E}[\|U_n\|_\infty^2 \,|\, \mathcal{F}_n] \leq \mathbb{E}[\|U_n\|_\infty^q \,|\, \mathcal{F}_n]^{2/q} \leq \sigma_n^2$, and the conditional Minkowski inequality then yields

$$\mathbb{E}[\|\widehat{v}_n\|_\infty^2 \,|\, \mathcal{F}_n]^{1/2} \leq \|v(\pi_n)\|_\infty + \|b_n\|_\infty + \mathbb{E}[\|U_n\|_\infty^2 \,|\, \mathcal{F}_n]^{1/2} \leq H_n \,, \tag{D.16}$$

where $H_n \coloneqq M_v + B_n + \sigma_n$. As $|\xi_n| \leq L\|U_n\|_\infty$ on $\{n < T_+\}$, we deduce that

$$\begin{aligned} \mathbb{E}[(M_{n+1} - M_n)^2 \,|\, \mathcal{F}_n] = I_n \gamma_n^2 \, \mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_n] \leq L^2 \gamma_n^2 \sigma_n^2 \,, \\ \mathbb{E}[r_n \,|\, \mathcal{F}_n] \leq q_n \coloneqq L\gamma_n B_n + \frac{\kappa}{2} \gamma_n^2 H_n^2 \,. \end{aligned} \tag{D.17}$$

Finally, as $H_n^2 \leq 3(M_v^2 + B_n^2 + \sigma_n^2)$, Assumption 1 gives

$$\begin{aligned} \gamma_n B_n = \mathcal{O}(n^{-\ell_\gamma - \ell_b}) \,, \qquad \gamma_n^2 \sigma_n^2 = \mathcal{O}(n^{-2\ell_\gamma + 2\ell_\sigma}) \,, \\ \text{and} \quad \gamma_n^2 H_n^2 = \mathcal{O}(n^{-2\ell_\gamma} + n^{-2\ell_\gamma + 2\ell_\sigma}) \,, \end{aligned} \tag{D.18}$$

and the conditions $\ell_\gamma > 1/2$, $\ell_\gamma + \ell_b > 1$ and $\ell_\gamma - \ell_\sigma > 1/2$ bring all these exponents below -1. Both series $\sum_n q_n$ and $\sum_n \gamma_n^2 \sigma_n^2$ are therefore finite.

Step 3: Stability. We first show that $T_+ = \infty$ on the event

$$\Omega \coloneqq \left\{\sup_{n \geq 1} |M_n| \leq E_+/4\right\} \cap \{R_\infty \leq E_+/4\} \,. \tag{D.19}$$

Indeed, if T_+ were finite, then $T_+ \geq 2$, as $E_1 < E_+$, and, since $M_1 = R_1 = 0$, telescoping (D.14) from $n = 1$ to $n = T_+ - 1$ and discarding the nonnegative terms d_n would give, on Ω,

$$E_{T_+} = \tilde{E}_{T_+} \leq E_1 + M_{T_+} + R_{T_+} < \frac{E_+}{4} + \frac{E_+}{4} + \frac{E_+}{4} < E_+ \,, \tag{D.20}$$

in contradiction with the definition of T_+. It thus remains to bound the probability of the complement of Ω. For every horizon $N \geq 1$, Doob's inequality (Theorem B.2) and the orthogonality of martingale increments give

$$\mathbb{P}\left(\max_{n \leq N} |M_n| \geq \frac{E_+}{4}\right) \leq \frac{16}{E_+^2} \, \mathbb{E}[M_N^2] = \frac{16}{E_+^2} \sum_{n<N} \mathbb{E}[(M_{n+1} - M_n)^2] \leq \frac{16L^2}{E_+^2} \sum_{n \geq 1} \gamma_n^2 \sigma_n^2 \,, \tag{D.21}$$

and the same bound holds for the supremum over all $n \geq 1$ by letting $N \to \infty$. For the remainder, Markov's inequality and (D.17) give $\mathbb{P}(R_\infty > E_+/4) \leq 4\,\mathbb{E}[R_\infty]/E_+ \leq (4/E_+) \sum_n q_n$. Altogether, since $\gamma_n = \gamma n^{-\ell_\gamma}$, we obtain

$$\begin{aligned} \mathbb{P}(T_+ < \infty) &\leq \frac{16L^2}{E_+^2} \sum_{n \geq 1} \gamma_n^2 \sigma_n^2 + \frac{4}{E_+} \sum_{n \geq 1} q_n \\ &= \frac{4L\gamma}{E_+} \sum_{n \geq 1} \frac{B_n}{n^{\ell_\gamma}} + \frac{16L^2\gamma^2}{E_+^2} \sum_{n \geq 1} \frac{\sigma_n^2}{n^{2\ell_\gamma}} + \frac{2\kappa\gamma^2}{E_+} \sum_{n \geq 1} \frac{H_n^2}{n^{2\ell_\gamma}} \,. \end{aligned} \tag{D.22}$$

The right-hand side does not depend on the initialization, and, as the bias and noise bounds of Assumption 1 hold uniformly for base step sizes in a bounded interval (as is the case for all the algorithms of Appendix E.3), it vanishes as $\gamma \to 0$. It is therefore at most η as soon as the base step size is small enough, which provides the threshold $\bar{\gamma}$. Finally, on the event $\{T_+ = \infty\}$, we have $E_n \leq E_+ < \delta_U$, and hence $\pi_n \in U$, for all $n \geq 1$.

Step 4: Attraction. Consider the process

$$W_n \coloneqq \tilde{E}_n + \sum_{k<n} d_k + \sum_{k \geq n} q_k \,. \tag{D.23}$$

As d_n is $\mathcal{F}_n$-measurable and (M_n) is a martingale, taking conditional expectations in (D.14) and using (D.17) yields

$$\mathbb{E}[W_{n+1} \,|\, \mathcal{F}_n] \le \tilde{E}_n - d_n + q_n + \sum_{k \le n} d_k + \sum_{k > n} q_k = W_n \,, \tag{D.24}$$

so that (W_n) is a nonnegative supermartingale, which converges almost surely by Theorem B.4. As $0 \le \sum_{k<n} d_k \le W_n$ and $\sum_{k \ge n} q_k \to 0$, the nondecreasing sum of the dissipations converges to a finite limit, and so does $\tilde{E}_n$. Fix now a sample in $\{T_+ = \infty\}$ along which these limits exist, and suppose, toward a contradiction, that the energy converges to some $E_\infty > 0$ along it. As $E_n \le E_+$ for all n, there exists n_0 such that $E_\infty/2 < E_n < 2E_+$ for all $n \ge n_0$, and (D.8), applied with $\epsilon = E_\infty/2$ and $\epsilon' = 2E_+ \le \bar{E}$, provides a constant $c > 0$ such that $\dot{E}_n \le -c$ for all $n \ge n_0$. As $I_n = 1$ for all n on $\{T_+ = \infty\}$, we get

$$\sum_{n \ge n_0} d_n = -\sum_{n \ge n_0} \gamma_n \dot{E}_n \ge c \sum_{n \ge n_0} \gamma_n = \infty \,, \tag{D.25}$$

which contradicts the finiteness of $\sum_n d_n$. Therefore, on $\{T_+ = \infty\}$, the energy vanishes in the limit almost surely, and Definition D.1(b) ensures that $\mathrm{dist}_1(\pi_n, S) \to 0$. Combined with the third step, this concludes the proof. ∎

E Instantiations of regularized policy gradient

In this appendix, we verify that the main feedback models considered in the paper fit the template (RPG), and we derive the corresponding signal bounds.

E.1 Extrapolation

The deterministic extrapolated family mentioned in the main text is defined as follows. Fix deterministic coefficients $\mu_n, \mu'_n \ge 0$ with $\mu_n + \mu'_n \le 1$, and, given the previous intermediate policy $\pi_{n-1/2}$ and the current policy π_n, define the predictor

$$w_n = \mu_n v(\pi_n) + \mu'_n v(\pi_{n-1/2}) \,, \tag{E.1}$$

with the convention $\pi_{1/2} = \pi_1$. The extrapolated update is then

$$\begin{aligned} Y_{n+1/2} &= Y_n + \gamma_n w_n \,, & \pi_{n+1/2} &= \psi(Y_{n+1/2}) \,, \\ Y_{n+1} &= Y_n + \gamma_n v(\pi_{n+1/2}) \,, & \pi_{n+1} &= \psi(Y_{n+1}) \,. \end{aligned} \tag{RPG+}$$

The choice $\mu_n = \mu'_n = 0$ recovers vanilla (RPG), while the choices $(\mu_n, \mu'_n) = (0, 1)$ and $(\mu_n, \mu'_n) = (1, 0)$ respectively yield the *optimistic* and the *extragradient* versions.

Proposition E.1. *The base update of* (RPG+) *can be written as* $Y_{n+1} = Y_n + \gamma_n(v(\pi_n) + b_n)$, *where the deterministic bias* $b_n \coloneqq v(\pi_{n+1/2}) - v(\pi_n)$ *satisfies*

$$\|b_n\|_\infty \le \frac{L_v M_v}{K} \gamma_n \,. \tag{E.2}$$

Proof. The decomposition of the base update is immediate from (RPG+). Moreover, as $\mu_n + \mu'_n \le 1$, Lemma D.1 gives $\|w_n\|_\infty \le M_v$, and Proposition C.1(c) then yields

$$\|\pi_{n+1/2} - \pi_n\|_1 = \|\psi(Y_n + \gamma_n w_n) - \psi(Y_n)\|_1 \le \frac{\gamma_n}{K} \|w_n\|_\infty \le \frac{M_v}{K} \gamma_n \,. \tag{E.3}$$

The Lipschitz continuity of v provided by Lemma D.1 finally implies $\|b_n\|_\infty \le L_v \|\pi_{n+1/2} - \pi_n\|_1$, and (E.2) follows. ∎

Consequently, (RPG+) is an instance of (RPG) with no noise and a deterministic bias of order $\mathcal{O}(\gamma_n)$.

E.2 Feedback from trajectories

Let us now attend to the estimates for (R-RPG), which require a little more bookkeeping. In the estimator (REINFORCE), each visit of the sampled trajectory to a pair (s, α_i) contributes the rewards accrued from that stage onward, reweighted by the inverse of the probability with which the behavior policy selects α_i at s. The uniform stopping condition ensures that, under every behavior policy, the length of an episode has geometric tails, that is, for all $t \geq 0$,

$$\mathbb{P}(T_n \geq t \,|\, \mathcal{F}_n) \leq (1 - \zeta_{\min})^t \,. \tag{E.4}$$

For convenience, we also send the process, after termination, to an absorbing terminal state where a dummy action is played. This harmless convention leaves the dynamics of the game unchanged and ensures that the histories $\mathcal{H}_t^-$ and $\mathcal{H}_t^+$ below are defined for all $t \geq 0$.

Lemma E.1. *Fix a triple (i, s, α_i), let $\widehat{\pi} \in \Pi$ be a behavior policy with $p \coloneqq \widehat{\pi}_i(\alpha_i \mid s) \geq \varepsilon > 0$, let τ be one stopped episode sampled under $\widehat{\pi}$, and let*

$$\Lambda \coloneqq \sum_{t \geq 0} \frac{I_t}{p} \quad \text{and} \quad I_t \coloneqq \mathbf{1}\{t \leq T,\, s_t = s,\, \alpha_{i,t} = \alpha_i\} \,. \tag{E.5}$$

Then, for every integer $m \geq 0$, there exists a constant $C_m < \infty$, which depends only on the finite game data and on m, such that

$$\mathbb{E}_{\widehat{\pi}}[(T+1)^m \Lambda^2] \leq \frac{C_m}{\varepsilon} \,. \tag{E.6}$$

Proof. Throughout, we write $\vartheta \coloneqq 1 - \zeta_{\min} < 1$, and, for every stage $t \geq 0$, we let $\mathcal{H}_t^-$ denote the history of the trajectory up to the observation of s_t, before the actions at stage t are sampled, and $\mathcal{H}_t^+$ the history up to and including these actions. Conditioning on $\mathcal{H}_t^-$, under which player i selects α_i at $s_t = s$ with probability p, the uniform stopping condition gives

$$\mathbb{E}_{\widehat{\pi}}[I_t] = p\, \mathbb{P}_{\widehat{\pi}}(t \leq T,\, s_t = s) \leq p\, \vartheta^t \,. \tag{E.7}$$

Moreover, conditionally on $\mathcal{H}_t^+$ and on the event $\{t \leq T\}$, the episode survives each subsequent stage with probability at most ϑ, so that $\mathbb{P}_{\widehat{\pi}}(T - t \geq u \,|\, \mathcal{H}_t^+) \leq \vartheta^u$ for every integer $u \geq 0$. Hence, for every integer $j \geq 1$, we have, on this event,

$$\mathbb{E}_{\widehat{\pi}}[(T-t)^j \,|\, \mathcal{H}_t^+] = \sum_{u \geq 1} \big(u^j - (u-1)^j\big)\, \mathbb{P}_{\widehat{\pi}}(T - t \geq u \,|\, \mathcal{H}_t^+) \leq \sum_{u \geq 1} u^j \vartheta^u \,, \tag{E.8}$$

and, since $(T+1)^j = \big((t+1) + (T-t)\big)^j \leq 2^{j-1}\big((t+1)^j + (T-t)^j\big)$ by convexity, we infer that

$$\mathbb{E}_{\widehat{\pi}}[(T+1)^j \,|\, \mathcal{H}_t^+] \leq \bar{C}_j (t+1)^j \quad \text{and} \quad \bar{C}_j \coloneqq 2^{j-1}\Big(1 + \sum_{u \geq 1} u^j \vartheta^u\Big) \,. \tag{E.9}$$

With these estimates at hand, we bound the moments of Λ. As the indicators I_t vanish for $t > T$, we have $\sum_{t \geq 0} I_t \leq T + 1$, and hence $\Lambda^2 \leq (T+1) \sum_{t \geq 0} I_t / p^2$ pathwise. In addition, I_t is $\mathcal{H}_t^+$-measurable and vanishes outside $\{t \leq T\}$, so that Tonelli's theorem, (E.9) and (E.7) yield

$$\begin{aligned} \mathbb{E}_{\widehat{\pi}}[(T+1)^m \Lambda^2] &\leq \frac{1}{p^2} \sum_{t \geq 0} \mathbb{E}_{\widehat{\pi}}[I_t\, \mathbb{E}_{\widehat{\pi}}[(T+1)^{m+1} \,|\, \mathcal{H}_t^+]] \\ &\leq \frac{\bar{C}_{m+1}}{p^2} \sum_{t \geq 0} (t+1)^{m+1}\, \mathbb{E}_{\widehat{\pi}}[I_t] \leq \frac{\bar{C}_{m+1}}{p} \sum_{t \geq 0} (t+1)^{m+1} \vartheta^t \,. \end{aligned} \tag{E.10}$$

The last series is finite because $\vartheta < 1$, and, as $p \geq \varepsilon$, the claim follows with $C_m \coloneqq \bar{C}_{m+1} \sum_{t \geq 0} (t+1)^{m+1} \vartheta^t$. ∎

Lemma E.2. *Assume that the behavior policy $\widehat{\pi}_n$ is $\mathcal{F}_n$-measurable and satisfies $\widehat{\pi}_{i,n}(\alpha_i \mid s) \geq \varepsilon_{\min}$ for every triple (i, s, α_i), where $\varepsilon_{\min} \in (0, 1]$. Then the estimator (REINFORCE) satisfies $\mathbb{E}[\widehat{v}_n \,|\, \mathcal{F}_n] = v(\widehat{\pi}_n)$, and there exists a constant $C < \infty$, which depends only on the finite game data, such that*

$$\mathbb{E}[\|\widehat{v}_n - v(\widehat{\pi}_n)\|_\infty^2 \,|\, \mathcal{F}_n] \leq \frac{C}{\varepsilon_{\min}} \,. \tag{E.11}$$

Proof. We condition on $\mathcal{F}_n$ throughout, write $\widehat{\pi} = \widehat{\pi}_n$, drop the superscript (n), and fix a triple (i, s, α_i). Denoting by $\mathcal{R}_{i,t}$ the rewards accrued from stage t onward, the corresponding coordinate of the estimator reads

$$\widehat{v}^{s}_{i\alpha_i} = \sum_{t\geq 0} \frac{I_t}{\widehat{\pi}_i(\alpha_i \mid s)} \mathcal{R}_{i,t} \quad \text{and} \quad \mathcal{R}_{i,t} \coloneqq \sum_{u=t}^{T} r_i(s_u, \alpha_u)\,, \tag{E.12}$$

with I_t as in (E.5). As $|r_i| \leq 1$, we also have $|\mathcal{R}_{i,t}| \leq T+1$, and hence $|\widehat{v}^{s}_{i\alpha_i}| \leq (T+1)\Lambda$, which is square-integrable by Lemma E.1 with $\varepsilon = \varepsilon_{\min}$ and $m = 2$. All the expectations below are therefore absolutely convergent.

Next, on the event $\{t \leq T,\, s_t = s\}$, conditioning on $\mathcal{H}_t^-$ and averaging over the actions of the other players at time t and over the continuation of the trajectory, we obtain, by (4),

$$\begin{aligned}
&\mathbb{E}_{\widehat{\pi}}\Big[\frac{\mathbf{1}\{\alpha_{i,t} = \alpha_i\}}{\widehat{\pi}_i(\alpha_i \mid s)} \mathcal{R}_{i,t} \,\Big|\, \mathcal{H}_t^-\Big] \\
&\quad = \sum_{\alpha_{-i} \in \mathcal{A}_{-i}} \widehat{\pi}_{-i}(\alpha_{-i} \mid s) \Big[r_i(s, \alpha_i, \alpha_{-i}) + \sum_{s' \in \mathcal{S}} P(s' \mid s, \alpha_i, \alpha_{-i}) V_{i,s'}(\widehat{\pi}) \Big] \\
&\quad = Q_i^{\widehat{\pi}}(s, \alpha_i)\,,
\end{aligned} \tag{E.13}$$

and hence

$$\mathbb{E}_{\widehat{\pi}}\Big[\frac{I_t}{\widehat{\pi}_i(\alpha_i \mid s)} \mathcal{R}_{i,t}\Big] = \mathbb{P}_{\widehat{\pi}}(t \leq T,\, s_t = s)\, Q_i^{\widehat{\pi}}(s, \alpha_i)\,. \tag{E.14}$$

Summing (E.14) over $t \geq 0$, which commutes with the expectation by absolute integrability, we get

$$\mathbb{E}_{\widehat{\pi}}[\widehat{v}^{s}_{i\alpha_i}] = \sum_{t\geq 0} \mathbb{P}_{\widehat{\pi}}(t \leq T,\, s_t = s)\, Q_i^{\widehat{\pi}}(s, \alpha_i) = d_\rho^{\widehat{\pi}}(s) Q_i^{\widehat{\pi}}(s, \alpha_i) = v^{s}_{i\alpha_i}(\widehat{\pi})\,, \tag{E.15}$$

and, the triple (i, s, α_i) being arbitrary, $\mathbb{E}[\widehat{v}_n \mid \mathcal{F}_n] = v(\widehat{\pi}_n)$.

As for the second moment, the bound $|\widehat{v}^{s}_{i\alpha_i}| \leq (T+1)\Lambda$ and Lemma E.1 with $m = 2$ give $\mathbb{E}[|\widehat{v}^{s}_{i\alpha_i}|^2 \mid \mathcal{F}_n] \leq C_2/\varepsilon_{\min}$. As $\|v\|_\infty \leq M_v$ by Lemma D.1 and $\varepsilon_{\min} \leq 1$, we infer that

$$\mathbb{E}[|\widehat{v}^{s}_{i\alpha_i} - v^{s}_{i\alpha_i}(\widehat{\pi})|^2 \mid \mathcal{F}_n] \leq 2\,\mathbb{E}[|\widehat{v}^{s}_{i\alpha_i}|^2 \mid \mathcal{F}_n] + 2M_v^2 \leq \frac{2C_2 + 2M_v^2}{\varepsilon_{\min}} \tag{E.16}$$

for every triple (i, s, α_i). Since the squared sup norm of a vector is at most the sum of its d squared coordinates, (E.11) follows with $C \coloneqq 2d(C_2 + M_v^2)$. ∎

Proposition E.2. *Let $\widehat{\pi}_n = (1 - \varepsilon_n)\pi_n + \varepsilon_n \pi^{\mathrm{unif}}$, where $\varepsilon_n \in (0, 1]$ and π^{unif} is the uniform policy, given by $\pi_i^{\mathrm{unif}}(\cdot \mid s) = \mathrm{unif}(\mathcal{A}_i)$ for every player i and state s. Then the estimator* (REINFORCE) *satisfies* (7) *with*

$$b_n = v(\widehat{\pi}_n) - v(\pi_n) \quad \textit{and} \quad U_n = \widehat{v}_n - v(\widehat{\pi}_n)\,, \tag{E.17}$$

where $\|b_n\|_\infty = \mathcal{O}(\varepsilon_n)$ and $\mathbb{E}[\|U_n\|_\infty^2 \mid \mathcal{F}_n] = \mathcal{O}(\varepsilon_n^{-1})$.

Proof. Writing $A_{\max} \coloneqq \max_{i \in \mathcal{N}} |\mathcal{A}_i|$, the uniform component of the behavior policy guarantees that $\widehat{\pi}_{i,n}(\alpha_i \mid s) \geq \varepsilon_n / A_{\max}$ for every triple (i, s, α_i). Lemma E.2, applied with $\varepsilon_{\min} = \varepsilon_n / A_{\max}$, then shows that $\mathbb{E}[\widehat{v}_n \mid \mathcal{F}_n] = v(\widehat{\pi}_n)$, which identifies the bias and the noise as stated, and its second moment estimate provides $\mathbb{E}[\|U_n\|_\infty^2 \mid \mathcal{F}_n] = \mathcal{O}(\varepsilon_n^{-1})$. Finally, by Lemma D.1, the bias satisfies

$$\|b_n\|_\infty \leq L_v \|\widehat{\pi}_n - \pi_n\|_1 = L_v \varepsilon_n \|\pi^{\mathrm{unif}} - \pi_n\|_1\,, \tag{E.18}$$

which is of order $\mathcal{O}(\varepsilon_n)$ because Π is bounded. ∎

E.3 Verification of the signal assumptions

We may now translate these estimates into the assumptions of the main text. Throughout this subsection, $\gamma_n = \gamma n^{-\ell_\gamma}$, and all implicit constants are independent of n and, for γ in a bounded interval, of the base step size γ. The resulting schedules are summarized in Table 1.

Vanilla RPG. With exact feedback, we have $\widehat{v}_n = v(\pi_n)$, hence $U_n = b_n = 0$ and $B_n = \sigma_n = 0$. Assumption 1 thus holds whenever $\ell_\gamma \in (1/2, 1]$, and Assumption 2 holds for every $\ell_\gamma \in [0, 1]$, including constant step sizes.

Stochastic RPG. For the noisy variant of (RPG), we have $\widehat{v}_n = v(\pi_n) + U_n$ with $\mathbb{E}[U_n \,|\, \mathcal{F}_n] = 0$, and we assume that $\mathbb{E}[\|U_n\|_\infty^q \,|\, \mathcal{F}_n] \le \sigma^q$ for some $q \ge 2$ and some constant $\sigma < \infty$. Then $B_n = 0$ and $\sigma_n = \mathcal{O}(1)$, and, as before, Assumption 1 holds whenever $\ell_\gamma \in (1/2, 1]$, while Assumption 2 holds for every $\ell_\gamma \in [0, 1]$.

RPG+. By Proposition E.1, the reduction of (RPG+) to the base template involves no martingale fluctuation and a bias of order $\mathcal{O}(\gamma_n)$, which means that $\sigma_n = 0$ and $B_n = \mathcal{O}(n^{-\ell_\gamma})$. Assumption 1 thus holds whenever $\ell_\gamma \in (1/2, 1]$, and Assumption 2 holds for every $\ell_\gamma \in (0, 1]$. Constant step sizes are excluded here, as the induced bias is then of order $\mathcal{O}(1)$, and not $\mathcal{O}(n^{-\ell_b})$ for any $\ell_b > 0$.

REINFORCE-RPG. Let $\varepsilon_n \asymp n^{-\ell_\varepsilon}$ with $\ell_\varepsilon > 0$. By Proposition E.2, we have $B_n = \mathcal{O}(n^{-\ell_\varepsilon})$ and, with $q = 2$, $\sigma_n = \mathcal{O}(n^{\ell_\varepsilon/2})$. Assumption 1 thus holds whenever $\ell_\gamma + \ell_\varepsilon > 1$ and $\ell_\gamma - \ell_\varepsilon/2 > 1/2$, that is, whenever

$$1 - \ell_\gamma < \ell_\varepsilon < 2\ell_\gamma - 1\,, \tag{E.19}$$

and such an exponent exists if and only if $\ell_\gamma > 2/3$. As for Assumption 2, the same estimates give $\ell_b = \ell_\varepsilon$ and $\ell_\sigma = \ell_\varepsilon/2$, so that this assumption holds, constant step sizes included, for all $\ell_\varepsilon \in (0, 1)$ if $\ell_\gamma < 1$, and for all $\ell_\varepsilon \in (0, 1]$ if $\ell_\gamma = 1$.

Table 1: Summary of signal bounds and admissible schedules.

Algorithm	Signal bounds	Assumption 1	Assumption 2
Vanilla RPG	$B_n = 0, \sigma_n = 0$	$\ell_\gamma \in (1/2, 1]$	$\ell_\gamma \in [0, 1]$
Stochastic RPG	$B_n = 0, \sigma_n = \mathcal{O}(1)$	$\ell_\gamma \in (1/2, 1]$	$\ell_\gamma \in [0, 1]$
RPG+	$B_n = \mathcal{O}(n^{-\ell_\gamma}), \sigma_n = 0$	$\ell_\gamma \in (1/2, 1]$	$\ell_\gamma \in (0, 1]$
R-RPG	$B_n = \mathcal{O}(n^{-\ell_\varepsilon}), \sigma_n = \mathcal{O}(n^{\ell_\varepsilon/2})$	$1 - \ell_\gamma < \ell_\varepsilon < 2\ell_\gamma - 1$	$0 < \ell_\varepsilon < 1$ (≤ 1 if $\ell_\gamma = 1$)

F Omitted proofs from Section 4

In this appendix, we provide the proofs of the results of Section 4 which were omitted from the main text.

F.1 Nash policies and pointwise stability

We begin with Theorem 1, which we restate below for convenience.

Theorem 1. *For a policy profile $\pi^\star$, the following are equivalent:*

(a) *$\pi^\star$ is a strict Nash policy.*

(b) *$\pi^\star$ is deterministic and, writing $\beta_i^\star(s)$ for its prescribed action, we have $Q_i^{\pi^\star}(s, \beta_i^\star(s)) > Q_i^{\pi^\star}(s, \alpha_i)$ for every player i, state s, and action $\alpha_i \neq \beta_i^\star(s)$.*

(c) *$\pi^\star$ is variationally stable, i.e., there exists a neighborhood U of $\pi^\star$ such that $\langle v(\pi), \pi - \pi^\star \rangle < 0$ for all $\pi \in U \setminus \{\pi^\star\}$.*

The whole argument revolves around the performance-difference lemma [30], which expresses every unilateral deviation of a player through the advantages of their actions at $\pi^\star$, and which we recall below for the decision problem faced by a single player against fixed opponents.

Lemma F.1 (Performance-difference lemma)**.** *Fix a policy profile $\pi^\star$ and a player i, and, for every stationary deviation $\pi_i \in \Pi_i$, write*

$$J_i(\pi_i) \coloneqq V_{i,\rho}(\pi_i, \pi^\star_{-i}) \quad \text{and} \quad A_i^\star(s, \alpha_i) \coloneqq Q_i^{\pi^\star}(s, \alpha_i) - V_{i,s}(\pi^\star)\,. \tag{F.1}$$

Then, for every $\pi_i \in \Pi_i$,

$$J_i(\pi_i) - J_i(\pi_i^\star) = \sum_{s\in\mathcal{S}} d_\rho^{(\pi_i,\pi_{-i}^\star)}(s) \sum_{\alpha_i\in\mathcal{A}_i} \pi_i(\alpha_i \mid s) A_i^\star(s,\alpha_i)\,. \tag{F.2}$$

Proof of Theorem 1. Fix a player i, and let J_i and $A_i^\star$ be as in Lemma F.1. By the very definition of $V_{i,s}(\pi^\star)$, we have, for every state s,

$$\sum_{\alpha_i\in\mathcal{A}_i} \pi_i^\star(\alpha_i \mid s) A_i^\star(s,\alpha_i) = 0\,. \tag{F.3}$$

We first show that (a) and (b) are equivalent. Suppose, then, that $\pi^\star$ is strict, and let $\bar\pi_i$ be a deterministic policy with $\bar\pi_i(s) \in \arg\max_{\alpha_i} A_i^\star(s,\alpha_i)$ for every $s \in \mathcal{S}$. As $\max_{\alpha_i} A_i^\star(s,\alpha_i) \geq 0$ by (F.3), Lemma F.1 gives

$$J_i(\bar\pi_i) - J_i(\pi_i^\star) = \sum_{s\in\mathcal{S}} d_\rho^{(\bar\pi_i,\pi_{-i}^\star)}(s) \max_{\alpha_i\in\mathcal{A}_i} A_i^\star(s,\alpha_i) \geq 0\,, \tag{F.4}$$

and strictness forces $\bar\pi_i = \pi_i^\star$. The policy $\pi_i^\star$ is therefore deterministic, and (F.3) yields, for every state s,

$$A_i^\star(s,\beta_i^\star(s)) = \max_{\alpha_i\in\mathcal{A}_i} A_i^\star(s,\alpha_i) = 0\,. \tag{F.5}$$

If, in addition, $A_i^\star(s,\alpha_i) = 0$ for some $\alpha_i \neq \beta_i^\star(s)$, the deterministic policy $\tilde\pi_i \neq \pi_i^\star$ which plays α_i at s and agrees with $\pi_i^\star$ elsewhere satisfies, by (F.2) and (F.5),

$$J_i(\tilde\pi_i) - J_i(\pi_i^\star) = d_\rho^{\tilde\pi}(s) A_i^\star(s,\alpha_i) + \sum_{s'\neq s} d_\rho^{\tilde\pi}(s') A_i^\star(s',\beta_i^\star(s')) = 0\,, \tag{F.6}$$

where $\tilde\pi := (\tilde\pi_i, \pi_{-i}^\star)$, and this flatly contradicts strictness. Hence $A_i^\star(s,\alpha_i) < 0$ for all $\alpha_i \neq \beta_i^\star(s)$, which is (14). Conversely, under (b), any stationary deviation $\pi_i \neq \pi_i^\star$ satisfies $\pi_i(\alpha_i \mid s) > 0$ for some state s and some action $\alpha_i \neq \beta_i^\star(s)$, and (F.2) and (R) give

$$J_i(\pi_i) - J_i(\pi_i^\star) \leq d_\rho^{(\pi_i,\pi_{-i}^\star)}(s)\, \pi_i(\alpha_i \mid s)\, A_i^\star(s,\alpha_i) < 0\,. \tag{F.7}$$

Let us show next that (b) implies (c), together with the sharpness bound (16). By continuity of $\pi \mapsto Q_i^\pi$, there exist a neighborhood U of $\pi^\star$ and $\delta > 0$ such that

$$Q_i^\pi(s,\alpha_i) - Q_i^\pi(s,\beta_i^\star(s)) \leq -\delta \tag{F.8}$$

for all $\pi \in U$, $i \in \mathcal{N}$, $s \in \mathcal{S}$ and $\alpha_i \neq \beta_i^\star(s)$, while (R), the compactness of Π and the continuity of $\pi \mapsto d_\rho^\pi$, established in (D.5), ensure that

$$d_{\min} := \min_{\pi\in\Pi,\, s\in\mathcal{S}} d_\rho^\pi(s) > 0\,. \tag{F.9}$$

Since $\pi^\star$ is deterministic, we have $\pi_i(\beta_i^\star(s) \mid s) - 1 = -\sum_{\alpha_i\neq\beta_i^\star(s)} \pi_i(\alpha_i \mid s)$, and the policy gradient identity (5) yields, for every $\pi \in U$,

$$\begin{aligned}\langle v(\pi), \pi - \pi^\star\rangle &= \sum_{i,s} d_\rho^\pi(s) \sum_{\alpha_i\neq\beta_i^\star(s)} \pi_i(\alpha_i \mid s)\big[Q_i^\pi(s,\alpha_i) - Q_i^\pi(s,\beta_i^\star(s))\big] \\ &\leq -d_{\min}\delta \sum_{i,s} \sum_{\alpha_i\neq\beta_i^\star(s)} \pi_i(\alpha_i \mid s) = -\frac{d_{\min}\delta}{2}\|\pi - \pi^\star\|_1\,,\end{aligned} \tag{F.10}$$

where the last equality follows from the identity $\|\pi_i(\cdot \mid s) - \pi_i^\star(\cdot \mid s)\|_1 = 2\sum_{\alpha_i\neq\beta_i^\star(s)} \pi_i(\alpha_i \mid s)$. This is the sharpness bound (16) with $c = d_{\min}\delta/2$, and (c) follows.

It remains to close the circle by showing that (c) implies (b). Intuitively, variational stability leaves no room for a flat direction at $\pi^\star$, whereas a second action with zero advantage would provide one. Assume (c). For $\pi \neq \pi^\star$ and $\pi_t := (1-t)\pi^\star + t\pi$, we have $t\langle v(\pi_t), \pi - \pi^\star\rangle = \langle v(\pi_t), \pi_t - \pi^\star\rangle < 0$ for all sufficiently small $t > 0$, and letting $t \downarrow 0$ yields (12). The profile $\pi^\star$ is thus a Nash policy, and

none of its advantages can be positive. Indeed, if $A_i^\star(s,\alpha_i) > 0$ for some state s and action α_i, the policy $\hat\pi_i$ which plays α_i at s and agrees with $\pi_i^\star$ elsewhere would satisfy, by (F.2), (F.3) and (R),

$$J_i(\hat\pi_i) - J_i(\pi_i^\star) = d_\rho^{(\hat\pi_i,\pi_{-i}^\star)}(s) A_i^\star(s,\alpha_i) > 0\,, \tag{F.11}$$

in contradiction with (11). Together with (F.3), this shows that $A_i^\star(s,\cdot) \le 0$ at every state s, with equality on the support of $\pi_i^\star(\cdot \mid s)$. Suppose now, toward a contradiction, that two distinct actions have zero advantage at some state s, and pick $q_i \neq \pi_i^\star$ which coincides with $\pi_i^\star$ outside s and is supported at s on $\{\alpha_i : A_i^\star(s,\alpha_i) = 0\}$. Writing $q_i^t := (1-t)\pi_i^\star + t q_i$ and $\pi^t := (q_i^t, \pi_{-i}^\star)$, the policy q_i^t only plays actions with zero advantage, at every state, and Lemma F.1 gives, for all $t \in [0,1]$,

$$J_i(q_i^t) - J_i(\pi_i^\star) = \sum_{s'\in\mathcal{S}} d_\rho^{\pi^t}(s') \sum_{\alpha_i\in\mathcal{A}_i} q_i^t(\alpha_i \mid s') A_i^\star(s',\alpha_i) = 0\,. \tag{F.12}$$

As $\pi^t - \pi^\star = (t(q_i - \pi_i^\star), 0)$, the policy gradient identity (5) then yields, for every $t \in (0,1)$,

$$\langle v(\pi^t), \pi^t - \pi^\star\rangle = t\langle v_i(\pi^t), q_i - \pi_i^\star\rangle = t\,\frac{d}{dt} J_i(q_i^t) = 0\,, \tag{F.13}$$

which contradicts (c), since $\pi^t \neq \pi^\star$ and $\pi^t \to \pi^\star$ as $t \downarrow 0$. Each state thus admits a unique action with zero advantage, which $\pi_i^\star$ plays with probability one, all other actions having a negative advantage. The player i being arbitrary, this is (b), and the three conditions are equivalent. Finally, as (b) implies (16), which itself implies (c), they are also equivalent to (16). ■

Remark F.1. For Nash policies of finite stochastic games with random stopping, Giannou et al. [26] introduce a hierarchy of local notions, in which strict policies are second-order stationary, second-order stationary policies are stable, and stable policies are first-order stationary, i.e., Nash. In their terminology, stability is the local variational condition (15), while strictness is stated as determinism together with a strict first-order stationarity inequality away from $\pi^\star$. They also show that second-order stationary policies satisfy a quadratic drift estimate, and that strict policies satisfy the sharper, linear estimate (16), leaving open whether the gap between these notions can be closed in general. Theorem 1 closes this gap under the reachability condition (R), since the implication (c)⇒(a) shows that every stable Nash policy is strict. Combined with the implications of Giannou et al. [26], this shows that strictness, second-order stationarity and stability are all equivalent for Nash policies, and that they all imply the linear sharpness estimate (16). ✦

The geometric reading of (16) given after Theorem 1 rests on the following observation. At a deterministic profile $\pi^\star$, the tangent cone $\mathrm{TC}(\pi^\star)$ is generated by the edges $e^s_{i\alpha_i} - e^s_{i\beta_i^\star(s)}$ of Π emanating from $\pi^\star$, along which the field has slope

$$\langle v(\pi^\star), e^s_{i\alpha_i} - e^s_{i\beta_i^\star(s)}\rangle = d_\rho^{\pi^\star}(s)\big(Q_i^{\pi^\star}(s,\alpha_i) - Q_i^{\pi^\star}(s,\beta_i^\star(s))\big)\,. \tag{F.14}$$

Under (R), the Q-value gaps of Theorem 1(b) are thus equivalent to the negativity of all these slopes, that is, to the inclusion of $v(\pi^\star)$ in the interior of $\mathrm{NC}(\pi^\star)$.

The proof of Corollary 1 is given at the end of Appendix F.3, as a consequence of Theorem 2.

F.2 One-shot club is not enough

This subsection details the example of Figure 2. There are three players, with action sets $\mathcal{A}_1 = \{T,B\}$, $\mathcal{A}_2 = \{L,R\}$ and $\mathcal{A}_3 = \{F,K\}$, where F denotes the front action and K the back action, and two states, $\mathcal{S} = \{1,2\}$. The candidate state-action set $\mathcal{B}$ is given by $\mathcal{B}_1(s) = \mathcal{A}_1$ and $\mathcal{B}_2(s) = \mathcal{A}_2$ at both states, together with $\mathcal{B}_3(1) = \{F\}$ and $\mathcal{B}_3(2) = \{K\}$, so that

$$\mathrm{span}(\mathcal{B}) = \{\pi \in \Pi : \pi_3(F \mid 1) = 1,\ \pi_3(K \mid 2) = 1\}\,. \tag{F.15}$$

At both states and for both actions of player 3, players 1 and 2 play matching pennies, with $r_1 = -r_2 = 1$ if $(\alpha_1,\alpha_2) \in \{(T,L),(B,R)\}$ and $r_1 = -r_2 = -1$ otherwise. The reward of player 3 does not depend on (α_1,α_2), and it equals 1 under F and 0 under K at state 1, and -1 under F and 0 under K at state 2. All rewards thus lie in $\{-1,0,1\}$, and the stage games, with state 1 in the first row and state 2 in the second, each matrix listing the payoffs (r_1,r_2,r_3) for the action of player 3

indicated in its corner, are given by

F	L	R
T	$(1,-1,1)$	$(-1,1,1)$
B	$(-1,1,1)$	$(1,-1,1)$

K	L	R
T	$(1,-1,0)$	$(-1,1,0)$
B	$(-1,1,0)$	$(1,-1,0)$

F	L	R
T	$(1,-1,-1)$	$(-1,1,-1)$
B	$(-1,1,-1)$	$(1,-1,-1)$

K	L	R
T	$(1,-1,0)$	$(-1,1,0)$
B	$(-1,1,0)$	$(1,-1,0)$

(F.16)

In the one-shot game at state 1, the front action of player 3 thus strictly dominates the back action, and the front face is a strict one-shot club. Symmetrically, the back face is a strict one-shot club at state 2.

The transitions, on the other hand, depend only on the state and on the action of player 3, and they are governed by a single parameter $\varepsilon \in (0, 1/3)$. From state 1, the game returns to state 1 with probability $1-\varepsilon$ under both actions of player 3. From state 2, it stays at state 2 with probability $1-\varepsilon$ under K, and it moves to state 1 with probability $1-\varepsilon$ under F, all other transition probabilities being zero. The stopping probability is thus always equal to ε, which gives $\zeta_{\min} = \varepsilon$, and we take $\rho = (1/2, 1/2)$. For every $\pi \in \operatorname{span}(\mathcal{B})$, player 3 plays F at state 1 and K at state 2, hence

$$P^\pi = (1-\varepsilon)I\,, \qquad (I - P^\pi)^{-1} = \varepsilon^{-1} I\,, \quad \text{and} \quad d_\rho^\pi = \Big(\frac{1}{2\varepsilon}, \frac{1}{2\varepsilon}\Big)\,, \tag{F.17}$$

and the reachability condition holds at both states. Intuitively, a small ε makes the future weigh heavily, and the excluded action F at state 2, which costs one unit of reward now, opens the door to the rewarding state 1 for a long time.

We first show that $\mathcal{B}$ is not Q-club. Fix $\pi \in \operatorname{span}(\mathcal{B})$. The values of player 3 satisfy $V_{3,1}(\pi) = 1 + (1-\varepsilon)V_{3,1}(\pi)$ and $V_{3,2}(\pi) = (1-\varepsilon)V_{3,2}(\pi)$, that is, $V_{3,1}(\pi) = 1/\varepsilon$ and $V_{3,2}(\pi) = 0$. At state 1, we thus have

$$Q_3^\pi(1,F) = 1 + \frac{1-\varepsilon}{\varepsilon} = \frac{1}{\varepsilon} \quad \text{and} \quad Q_3^\pi(1,K) = \frac{1-\varepsilon}{\varepsilon}\,, \tag{F.18}$$

and the prescribed action F is indeed Q-better than K. At state 2, however,

$$Q_3^\pi(2,K) = 0 \quad \text{and} \quad Q_3^\pi(2,F) = -1 + \frac{1-\varepsilon}{\varepsilon} = \frac{1-2\varepsilon}{\varepsilon} > 0\,. \tag{F.19}$$

The excluded action F thus has a strictly larger Q-value than the prescribed action K, and $\mathcal{B}$ is not Q-club.

We next show that $\mathcal{C} \coloneqq \operatorname{span}(\mathcal{B})$ is neither attracting nor stable under the deterministic entropic dynamics, that is, vanilla (EPG) with $Y_{n+1} = Y_n + \gamma_n v(\pi_n)$ and $\pi_n = \psi(Y_n)$, where ψ is the softmax map, $\gamma_n > 0$ and $\sum_n \gamma_n = \infty$. Given an arbitrary policy π, write $q = \pi_3(F \mid 1)$ and $p = \pi_3(F \mid 2)$, two quantities that $\mathcal{C}$ requires to be equal to 1 and 0, respectively. As state 1 returns to itself with probability $1-\varepsilon$ regardless of the action of player 3, we have $V_{3,1}(\pi) = q + (1-\varepsilon)V_{3,1}(\pi)$, that is, $V_{3,1}(\pi) = q/\varepsilon$, which gives

$$Q_3^\pi(2,F) = -1 + \frac{(1-\varepsilon)q}{\varepsilon} \quad \text{and} \quad Q_3^\pi(1,F) - Q_3^\pi(1,K) = 1\,. \tag{F.20}$$

Writing $q_n \coloneqq \pi_{3,n}(F \mid 1)$, $p_n \coloneqq \pi_{3,n}(F \mid 2)$ and $Z_n^1 \coloneqq Y_{3F,n}^1 - Y_{3K,n}^1$ for the score gap at state 1, the policy gradient identity (5) and (F.20) give

$$Z_{n+1}^1 - Z_n^1 = \gamma_n d_\rho^{\pi_n}(1)\big[Q_3^{\pi_n}(1,F) - Q_3^{\pi_n}(1,K)\big] = \gamma_n d_\rho^{\pi_n}(1) > 0\,. \tag{F.21}$$

As the softmax map gives $q_n = e^{Z_n^1}/(1+e^{Z_n^1})$, which increases with Z_n^1, the condition $q_1 > 1/2$ ensures that $q_n \geq q_1 > 1/2$ for all n, and (F.20) yields

$$Q_3^{\pi_n}(2,F) = -1 + \frac{(1-\varepsilon)q_n}{\varepsilon} \geq -1 + \frac{1-\varepsilon}{2\varepsilon} = \omega_\varepsilon \coloneqq \frac{1-3\varepsilon}{2\varepsilon} > 0\,. \tag{F.22}$$

At state 2, on the other hand, $Q_3^{\pi_n}(2,K) = (1-\varepsilon)V_{3,2}(\pi_n)$, and, as the value at state 2 averages the two Q-values, the identity $V_{3,2}(\pi_n) = p_n Q_3^{\pi_n}(2,F) + (1-p_n)(1-\varepsilon)V_{3,2}(\pi_n)$ gives

$$V_{3,2}(\pi_n) = \frac{p_n Q_3^{\pi_n}(2,F)}{\varepsilon + (1-\varepsilon)p_n}\,. \tag{F.23}$$

Since $\varepsilon + (1-\varepsilon)p_n \le 1$ and $Q_3^{\pi_n}(2,F) \ge \omega_\varepsilon > 0$, we conclude that

$$\begin{aligned} Q_3^{\pi_n}(2,F) - Q_3^{\pi_n}(2,K) &= Q_3^{\pi_n}(2,F) - \frac{(1-\varepsilon)p_n Q_3^{\pi_n}(2,F)}{\varepsilon + (1-\varepsilon)p_n} \\ &= \frac{\varepsilon\, Q_3^{\pi_n}(2,F)}{\varepsilon + (1-\varepsilon)p_n} \ge \varepsilon\,\omega_\varepsilon = \frac{1-3\varepsilon}{2}\,. \end{aligned} \tag{F.24}$$

As moreover $d_\rho^{\pi_n}(2) \ge \rho(2) = 1/2$, the policy gradient identity shows that the score gap $Z_n^2 \coloneqq Y_{3F,n}^2 - Y_{3K,n}^2$ at state 2 satisfies

$$Z_{n+1}^2 - Z_n^2 = \gamma_n d_\rho^{\pi_n}(2)\big[Q_3^{\pi_n}(2,F) - Q_3^{\pi_n}(2,K)\big] \ge \frac{1-3\varepsilon}{4}\,\gamma_n\,, \tag{F.25}$$

and summing these increments yields $Z_n^2 \to +\infty$. Hence $p_n = e^{Z_n^2}/(1+e^{Z_n^2}) \to 1$, whereas $\mathcal{C}$ requires $p_n = 0$. As this holds from initializations arbitrarily close to $\mathcal{C}$, the face $\mathcal{C}$ is neither attracting nor stable, whatever the value of $\varepsilon \in (0, 1/3)$, and the action that defines the one-shot club set at state 2 even becomes extinct.

F.3 Q-club sets

We begin with the game-theoretic properties of Q-club sets, which we restate below for convenience.

Proposition 1. *Every Q-club set $\mathcal{B}$ is a fixed set of the state-action better-reply correspondence, i.e., $\mathrm{Qbtr}(\mathcal{B}) = \mathcal{B}$, although the converse fails in general. Moreover, the span of every Q-club set contains a stationary Nash policy.*

The proof relies on the identity

$$V_{i,s}(\pi) = \sum_{\alpha_i \in \mathcal{A}_i} \pi_i(\alpha_i \mid s) Q_i^\pi(s,\alpha_i)\,, \tag{F.26}$$

which expresses the value of continuing with the current mixed action as the average of the Q-values of the pure actions, and which follows directly from the definitions (2) and (4).

Proof of Proposition 1. We first show that every Q-club set is a fixed set of Qbtr. Let $\mathcal{B}$ be any state-action set. Given an allowed action $\beta_i \in \mathcal{B}_i(s)$, every policy $\pi \in \mathrm{span}(\mathcal{B})$ with $\pi_i(\beta_i \mid s) = 1$ satisfies $V_{i,s}(\pi) = Q_i^\pi(s,\beta_i)$ by (F.26), hence $\beta_i \in \mathrm{Qbtr}_{i,s}(\pi)$, which shows that $\mathcal{B} \subseteq \mathrm{Qbtr}(\mathcal{B})$. If, in addition, $\mathcal{B}$ is Q-club and $\alpha_i \in \bar{\mathcal{B}}_i(s)$, then, for every $\pi \in \mathrm{span}(\mathcal{B})$, the mixed action $\pi_i(\cdot \mid s)$ is supported on $\mathcal{B}_i(s)$, and (20) and (F.26) give

$$Q_i^\pi(s,\alpha_i) = \sum_{\beta_i \in \mathcal{B}_i(s)} \pi_i(\beta_i \mid s) Q_i^\pi(s,\alpha_i) < \sum_{\beta_i \in \mathcal{B}_i(s)} \pi_i(\beta_i \mid s) Q_i^\pi(s,\beta_i) = V_{i,s}(\pi)\,. \tag{F.27}$$

The excluded action α_i thus belongs to no set $\mathrm{Qbtr}_{i,s}(\pi)$ with $\pi \in \mathrm{span}(\mathcal{B})$, and $\mathrm{Qbtr}(\mathcal{B}) \subseteq \mathcal{B}$, which proves that $\mathrm{Qbtr}(\mathcal{B}) = \mathcal{B}$.

To see that the converse fails, consider a game with a single player and a single state s, with $\rho = \delta_s$ and three actions a, b and c. The actions a and b yield the rewards 1 and 0, respectively, and terminate the episode, whereas c yields the reward $-1/4$ and returns to s with probability $1/2$, which makes the stopping probability at least $1/2$. Let $\mathcal{B} = \{\mathsf{a}, \mathsf{b}\}$, so that c is the only excluded action, and, given $\pi \in \mathrm{span}(\mathcal{B})$, write $p \coloneqq \pi(\mathsf{a} \mid s)$. As a and b terminate the episode, we have

$$V_s(\pi) = p\,, \qquad Q^\pi(s,\mathsf{b}) = 0\,, \quad \text{and} \quad Q^\pi(s,\mathsf{c}) = -\frac{1}{4} + \frac{p}{2}\,. \tag{F.28}$$

On the one hand, $Q^\pi(s,\mathsf{c}) - V_s(\pi) = -1/4 - p/2 < 0$ for every $p \in [0,1]$, which means that $\mathsf{c} \notin \mathrm{Qbtr}(\mathcal{B})$, and $\mathcal{B}$ is a fixed set of Qbtr by the first part of the argument. On the other hand, the policy which plays a leads to $Q^\pi(s,\mathsf{c}) = 1/4 > 0 = Q^\pi(s,\mathsf{b})$, and $\mathcal{B}$ is therefore not Q-club. Intuitively, the excluded action c, whose value stems from the continuation of the game, is never a better reply to a policy in $\mathrm{span}(\mathcal{B})$, but it beats the allowed action b as soon as the continuation value is high enough. Such a discrepancy cannot arise in normal-form games, where the payoff of a pure action does not depend on the mixed action of the player under consideration.

It remains to show that the span of a Q-club set $\mathcal{B}$ contains a stationary Nash policy. To that end, let $\mathcal{G}_{\mathcal{B}}$ be the stochastic game obtained from $\mathcal{G}$ by restricting player i, at every state s, to the actions of $\mathcal{B}_i(s)$. Being again a finite stochastic game with random stopping, $\mathcal{G}_{\mathcal{B}}$ admits a stationary Nash policy $\pi^\star \in \operatorname{span}(\mathcal{B})$ [23], and we claim that $\pi^\star$ is also a Nash policy of $\mathcal{G}$. Fix a player i, and let J_i and $A_i^\star$ be as in Lemma F.1. If $A_i^\star(s,\beta_i) > 0$ for some state s and some allowed action $\beta_i \in \mathcal{B}_i(s)$, the policy $\hat{\pi}_i$ which plays β_i at s and agrees with $\pi_i^\star$ elsewhere is a deviation within $\mathcal{G}_{\mathcal{B}}$, and, exactly as in (F.11), it satisfies

$$J_i(\hat{\pi}_i) - J_i(\pi_i^\star) = d_\rho^{(\hat{\pi}_i, \pi_{-i}^\star)}(s) A_i^\star(s,\beta_i) > 0\,, \tag{F.29}$$

which is impossible. Hence $Q_i^{\pi^\star}(s,\beta_i) \le V_{i,s}(\pi^\star)$ for every state s and every allowed action $\beta_i \in \mathcal{B}_i(s)$. For every excluded action $\alpha_i \in \bar{\mathcal{B}}_i(s)$, the Q-club condition then ensures that

$$Q_i^{\pi^\star}(s,\alpha_i) < Q_i^{\pi^\star}(s,\beta_i) \le V_{i,s}(\pi^\star)\,. \tag{F.30}$$

All advantages of player i at $\pi^\star$ are thus nonpositive, and Lemma F.1 gives, for every $\pi_i \in \Pi_i$,

$$J_i(\pi_i) - J_i(\pi_i^\star) = \sum_{s\in\mathcal{S}} d_\rho^{(\pi_i, \pi_{-i}^\star)}(s) \sum_{\alpha_i \in \mathcal{A}_i} \pi_i(\alpha_i \mid s) A_i^\star(s,\alpha_i) \le 0\,. \tag{F.31}$$

The player i being arbitrary, $\pi^\star$ is a stationary Nash policy of $\mathcal{G}$. ∎

We now come to the main stability result of the paper, which we restate below for convenience.

Theorem 2. *Under Assumption 1, if a state-action set $\mathcal{B}$ is Q-club, then $\operatorname{span}(\mathcal{B})$ is stochastically asymptotically stable.*

As outlined in Appendix A, our strategy is to show that the Fenchel gap to $\mathcal{C} := \operatorname{span}(\mathcal{B})$ is a local energy in the sense of Definition D.1, and then to invoke Theorem D.1. Intuitively, the restricted mirror map $\psi_{\mathcal{C}}$ moves the mass of the excluded actions onto the allowed ones, and, near a Q-club span, the latter are uniformly better, and the policy gradient field therefore pushes the iterates toward $\mathcal{C}$. To make this intuition rigorous, we fix a Q-club state-action set $\mathcal{B}$, together with the decomposable regularizer (C.1), and we begin with two elementary lemmas, which describe what happens at a single pair (i,s).

Lemma F.2. *Let A be a finite set and $B \subseteq A$ a nonempty subset, let $\theta\colon [0,1] \to \mathbb{R}$ be a kernel as in (C.1), with rate function φ, and, given $y \in \mathbb{R}^A$, let*

$$q := \operatorname*{arg\,max}_{p\in\Delta(A)} \sum_{\alpha\in A} \big[y_\alpha p_\alpha - \theta(p_\alpha)\big] \quad \text{and} \quad q^B := \operatorname*{arg\,max}_{p\in\Delta(B)} \sum_{\beta\in B} \big[y_\beta p_\beta - \theta(p_\beta)\big]^{3}\,. \tag{F.32}$$

Then $w := q - q^B$ satisfies $w_\alpha \ge 0$ for every $\alpha \notin B$ and $w_\beta \le 0$ for every $\beta \in B$, and the excluded mass $m := \sum_{\alpha\notin B} q_\alpha$ satisfies

$$m = \sum_{\alpha\notin B} w_\alpha = -\sum_{\beta\in B} w_\beta\,. \tag{F.33}$$

Proof. By Proposition C.1(d), applied on $\Delta(A)$ and on $\Delta(B)$, there exist $\lambda, \lambda^B \in \mathbb{R}$ such that

$$q_\alpha = \varphi(y_\alpha - \lambda) \quad \text{and} \quad q_\beta^B = \varphi(y_\beta - \lambda^B) \tag{F.34}$$

for every $\alpha \in A$ and every $\beta \in B$. Suppose, toward a contradiction, that $q_\beta > q_\beta^B$ for some $\beta \in B$. As φ is nondecreasing, the inequality $\varphi(y_\beta - \lambda) > \varphi(y_\beta - \lambda^B)$ forces $\lambda < \lambda^B$, and the same monotonicity gives $q_{\beta'} = \varphi(y_{\beta'} - \lambda) \ge \varphi(y_{\beta'} - \lambda^B) = q_{\beta'}^B$ for every $\beta' \in B$. Summing over B, we get

$$1 \ge \sum_{\beta'\in B} q_{\beta'} > \sum_{\beta'\in B} q_{\beta'}^B = 1\,, \tag{F.35}$$

which is absurd. Hence $w \le 0$ on B, while $w = q \ge 0$ outside B, and, as q and q^B both sum to one,

$$m = \sum_{\alpha\notin B} q_\alpha = \sum_{\alpha\notin B} w_\alpha = -\sum_{\beta\in B} w_\beta\,, \tag{F.36}$$

which is (F.33). ∎

[3] Here, q^B is extended to an element of $\Delta(A)$ by setting $q_\alpha^B := 0$ for $\alpha \notin B$.

Lemma F.3. *Let A be a finite set and $B \subseteq A$ a nonempty subset. Then, for every $p \in \Delta(A)$, we have $\mathrm{dist}_1(p, \Delta(B)) = 2m(p)$, where $m(p) \coloneqq \sum_{\alpha \notin B} p_\alpha$.*

Proof. Every $q \in \Delta(B)$ vanishes outside B, and therefore satisfies

$$\|p - q\|_1 = m(p) + \sum_{\beta \in B} |p_\beta - q_\beta| \geq m(p) + \left| \sum_{\beta \in B} (p_\beta - q_\beta) \right| = 2m(p)\,. \tag{F.37}$$

Conversely, keeping the coordinates of p on B, moving all its excluded mass to a single action of B, and setting the excluded coordinates to zero produces a point $q \in \Delta(B)$ with $\|p - q\|_1 = 2m(p)$, which shows that the bound is attained. ∎

With these lemmas at hand, the proof of Theorem 2 reduces to two further lemmas. The first one shows that the Q-club condition survives, with a uniform margin, in a neighborhood of $\mathcal{C}$, and the second one deduces from it the setwise sharpness bound (24) announced in the main text. If no action is excluded, then $\mathcal{C} = \Pi$, every iterate lies in $\mathcal{C}$, and there is nothing to prove. We may thus assume that $\bar{\mathcal{B}}_i(s) \neq \varnothing$ for at least one pair (i, s), and we call such pairs *active*.

Lemma F.4. *There exist $r > 0$ and constants $c_i^s > 0$ such that, on the neighborhood $U \coloneqq \{\pi \in \Pi : \mathrm{dist}_1(\pi, \mathcal{C}) < r\}$ of $\mathcal{C}$,*

$$v_{i\beta}^s(\pi) - v_{i\alpha}^s(\pi) \geq c_i^s \tag{F.38}$$

for every active pair (i, s), every allowed action $\beta \in \mathcal{B}_i(s)$, and every excluded action $\alpha \in \bar{\mathcal{B}}_i(s)$.

Proof. For each active pair (i, s), the map

$$g_i^s(\pi) \coloneqq \min_{\beta \in \mathcal{B}_i(s)} Q_i^\pi(s, \beta) - \max_{\alpha \in \bar{\mathcal{B}}_i(s)} Q_i^\pi(s, \alpha) \tag{F.39}$$

is positive on $\mathcal{C}$ by the Q-club property. By (D.6), in the proof of Lemma D.1, each map $\pi \mapsto Q_i^\pi(s, \alpha)$ is $(3/\zeta_{\min}^2)$-Lipschitz with respect to $\|\cdot\|_1$, and so are the minimum and the maximum in (F.39), which makes g_i^s a $(6/\zeta_{\min}^2)$-Lipschitz function. In particular, g_i^s is continuous, and the compactness of $\mathcal{C}$ gives

$$\delta_i^s \coloneqq \frac{1}{2} \min_{\pi \in \mathcal{C}} g_i^s(\pi) > 0\,. \tag{F.40}$$

Let $r \coloneqq \zeta_{\min}^2 \min_{i,s} \delta_i^s / 6$, the minimum being taken over the finitely many active pairs. For every $\pi \in U$, the compactness of $\mathcal{C}$ provides $p \in \mathcal{C}$ with $\|\pi - p\|_1 < r$, hence, for every active pair,

$$g_i^s(\pi) \geq g_i^s(p) - \frac{6}{\zeta_{\min}^2} \|\pi - p\|_1 \geq 2\delta_i^s - \delta_i^s = \delta_i^s\,. \tag{F.41}$$

Combining this bound with the policy gradient identity (5) and the lower bound (F.9) on the visitation counts, we obtain, for every allowed action $\beta \in \mathcal{B}_i(s)$ and every excluded action $\alpha \in \bar{\mathcal{B}}_i(s)$,

$$v_{i\beta}^s(\pi) - v_{i\alpha}^s(\pi) = d_\rho^\pi(s) \big[Q_i^\pi(s, \beta) - Q_i^\pi(s, \alpha) \big] \geq d_{\min}\, g_i^s(\pi) \geq d_{\min} \delta_i^s\,, \tag{F.42}$$

which is (F.38) with $c_i^s \coloneqq d_{\min} \delta_i^s$. ∎

Before stating the second lemma, we take a closer look at the restricted mirror map. Fix $y \in \mathcal{Y}$, and write $\pi \coloneqq \psi(y)$ and $\pi_{\mathcal{B}} \coloneqq \psi_{\mathcal{C}}(y)$. The maps ψ and $\psi_{\mathcal{C}}$ inherit the decomposability of h across players and states, and, for every pair (i, s), we let

$$w_i^s \coloneqq \pi_i^s - \pi_{\mathcal{B},i}^s \quad \text{and} \quad m_i^s \coloneqq \sum_{\alpha \in \bar{\mathcal{B}}_i(s)} \pi_i(\alpha \mid s)\,. \tag{F.43}$$

By Lemma F.2, applied to each pair, we have $w_{i\alpha}^s \geq 0$ for every $\alpha \in \bar{\mathcal{B}}_i(s)$, $w_{i\beta}^s \leq 0$ for every $\beta \in \mathcal{B}_i(s)$, and

$$m_i^s = \sum_{\alpha \in \bar{\mathcal{B}}_i(s)} w_{i\alpha}^s = - \sum_{\beta \in \mathcal{B}_i(s)} w_{i\beta}^s\,. \tag{F.44}$$

In addition, since both the face $\mathcal{C}$ and the ℓ_1 norm decompose across players and states, Lemma F.3 yields

$$\mathrm{dist}_1(\pi, \mathcal{C}) = \sum_{i \in \mathcal{N}} \sum_{s \in \mathcal{S}} \mathrm{dist}_1\big(\pi_i^s, \Delta(\mathcal{B}_i(s))\big) = 2 \sum_{i \in \mathcal{N}} \sum_{s \in \mathcal{S}} m_i^s\,. \tag{F.45}$$

Lemma F.5. *Let $c_* > 0$ be the minimum of the constants c_i^s of Lemma F.4 over all active pairs. Then, for every $y \in \mathcal{Y}$ such that $\pi = \psi(y) \in U$,*

$$\langle v(\pi), \pi - \psi_{\mathcal{C}}(y)\rangle \le -\frac{c_*}{2}\,\mathrm{dist}_1(\pi, \mathcal{C})\,. \tag{F.46}$$

Proof. Fix an active pair (i, s), and let $v^s_{i,\min}(\pi)$ denote the minimum of $v^s_{i\beta}(\pi)$ over all allowed actions $\beta \in \mathcal{B}_i(s)$. By (F.38), every excluded action α and every allowed action β satisfy

$$v^s_{i\alpha}(\pi) \le v^s_{i,\min}(\pi) - c_i^s \quad \text{and} \quad v^s_{i\beta}(\pi) \ge v^s_{i,\min}(\pi)\,. \tag{F.47}$$

The pairing of $v_i^s(\pi)$ with w_i^s thus weighs the mass that π places on excluded actions, which is evaluated at a field lower than $v^s_{i,\min}(\pi) - c_i^s$, against the compensating mass that $\pi_{\mathcal{B}}$ places on allowed actions, which is evaluated at a field higher than $v^s_{i,\min}(\pi)$. More precisely, we have

$$\begin{aligned}\langle v_i^s(\pi), w_i^s\rangle &= \sum_{\alpha\in\bar{\mathcal{B}}_i(s)} v^s_{i\alpha}(\pi) w^s_{i\alpha} + \sum_{\beta\in\mathcal{B}_i(s)} v^s_{i\beta}(\pi) w^s_{i\beta} \\ &\le \big(v^s_{i,\min}(\pi) - c_i^s\big) m_i^s - v^s_{i,\min}(\pi) m_i^s = -c_i^s m_i^s\,,\end{aligned} \tag{F.48}$$

where the inequality combines (F.47) with the sign conditions $w^s_{i\alpha} \ge 0$ on excluded actions and $w^s_{i\beta} \le 0$ on allowed ones, and then evaluates both sums of weights by (F.44). For inactive pairs, nothing is excluded, so that $\pi_{\mathcal{B}}$ and π coincide at the pair, and $w_i^s = 0$. Every pair thus contributes at most $-c_* m_i^s$ to the drift, and, summing over all pairs and invoking (F.45), we conclude that

$$\langle v(\pi), \pi - \pi_{\mathcal{B}}\rangle = \sum_{i,s}\langle v_i^s(\pi), w_i^s\rangle \le -c_* \sum_{i,s} m_i^s = -\frac{c_*}{2}\,\mathrm{dist}_1(\pi, \mathcal{C})\,, \tag{F.49}$$

which is (F.46). ∎

Armed with these two lemmas, we are now ready to prove Theorem 2.

Proof of Theorem 2. By Lemma C.1(b), (c), (e) and Lemma C.2, the Fenchel gap $F_{\mathcal{C}}$ is continuously differentiable and $(2/K)$-smooth, it has bounded gradients on bounded sublevels, and it satisfies Definition D.1(b). It thus remains to verify the dissipation condition Definition D.1(c), and, to that end, we set

$$\bar{E} := \frac{Kr^2}{8}\,, \tag{F.50}$$

in such a way that, by Lemma C.1(d), every score y with $F_{\mathcal{C}}(y) < \bar{E}$ satisfies

$$\frac{K}{2}\,\mathrm{dist}_1(\psi(y), \mathcal{C})^2 \le F_{\mathcal{C}}(y) < \frac{Kr^2}{8}\,, \tag{F.51}$$

that is, $\mathrm{dist}_1(\psi(y), \mathcal{C}) < r/2$, and hence $\psi(y) \in U$. Fix now $0 < \epsilon < \epsilon' \le \bar{E}$. By Lemma C.2 and Lemma D.2(b), there is a neighborhood of $\mathcal{C}$ on which the Fenchel gap is smaller than ϵ, and, as $\mathcal{C}$ is compact, this neighborhood contains all policies within some distance $\delta > 0$ of $\mathcal{C}$. Accordingly, every score y with $\epsilon < F_{\mathcal{C}}(y) < \epsilon'$ satisfies

$$\mathrm{dist}_1(\psi(y), \mathcal{C}) \ge \delta\,. \tag{F.52}$$

For every such y, the policy $\pi = \psi(y)$ lies in U, and, since $\nabla F_{\mathcal{C}}(y) = \psi(y) - \psi_{\mathcal{C}}(y)$, Lemma F.5 and (F.52) yield

$$\langle \nabla F_{\mathcal{C}}(y), v(\pi)\rangle = \langle v(\pi), \pi - \psi_{\mathcal{C}}(y)\rangle \le -\frac{c_*\delta}{2} < 0\,. \tag{F.53}$$

Hence $F_{\mathcal{C}}$ is a local energy for $\mathcal{C}$, and Theorem D.1 completes the proof. ∎

Finally, Corollary 1 follows from Theorem 2 by viewing a strict Nash policy as a Q-club set with singleton supports.

Proof of Corollary 1. Let $\pi^\star$ be a strict Nash policy. By Theorem 1, $\pi^\star$ is deterministic, with prescribed actions $\beta_i^\star(s)$, and the state-action set $\mathcal{B}$ defined by $\mathcal{B}_i(s) = \{\beta_i^\star(s)\}$ has $\mathrm{span}(\mathcal{B}) = \{\pi^\star\}$. Its Q-club inequalities (20) are then exactly the inequalities (14) of Theorem 1(b), which makes $\mathcal{B}$ a Q-club set, and Theorem 2 shows that $\{\pi^\star\}$ is stochastically asymptotically stable. ∎

F.4 Fast convergence to Q-club sets

Our aim in this subsection is to prove Theorem 3, which we restate below for convenience.

Theorem 3. *Under Assumption 2, if a state-action set $\mathcal{B}$ is Q-club, then, for every $\eta \in (0,1)$, there exist an open unbounded domain $\mathcal{D} \subseteq \mathcal{Y}$ and constants $c_0 > 0$, $c_1 \in \mathbb{R}$ and $c_2 > 0$ such that, for every initialization $Y_1 \in \mathcal{D}$, with probability at least $1-\eta$, we have, for every episode $n \geq 1$,*

$$\operatorname{dist}_1(\pi_n, \operatorname{span}(\mathcal{B})) \leq c_0\, \varphi\left(c_1 - c_2 \sum_{k=1}^{n-1} \gamma_k \right).$$

Throughout this subsection, we work under Assumption 2, we let $\mathcal{B}$ be a Q-club state-action set, and we write $\mathcal{C} \coloneqq \operatorname{span}(\mathcal{B})$ and $\Gamma_n \coloneqq \sum_{k=1}^{n} \gamma_k$, with $\Gamma_0 = 0$. In the spirit of Boone and Mertikopoulos [12], we track, instead of a single energy, a finite family of linear energies, which compare the score of every excluded action with that of every allowed action at the same pair (i,s), namely

$$\mathcal{Z} \coloneqq \left\{ e^s_{i\alpha} - e^s_{i\beta} : \alpha \in \bar{\mathcal{B}}_i(s),\ \beta \in \mathcal{B}_i(s) \right\} \quad \text{and} \quad E_z(y) \coloneqq \langle y, z \rangle . \tag{F.54}$$

Intuitively, a very negative value of E_z means that the excluded action of z is heavily outscored by an allowed one, and should thus be played with a small probability. The following lemma makes this intuition quantitative, as it converts the linear energies into bounds on the probabilities of excluded actions. It also covers kernels which depend on the player and the state.

Lemma F.6. *Let $y \in \mathcal{Y}$, let $\pi = \psi(y)$, and let $z = e^s_{i\alpha} - e^s_{i\beta} \in \mathcal{Z}$. Then*

$$\pi_i(\alpha \mid s) \leq \varphi^s_i\big(b^s_i + E_z(y)\big) . \tag{F.55}$$

Proof. We suppress the player and state indices, and write q for the block of $\psi(y)$ at the pair under consideration. By Proposition C.1(d), there exists $\lambda \in \mathbb{R}$ such that $q_{\alpha'} = \varphi(y_{\alpha'} - \lambda)$ for every action α' of the pair. We first bound the multiplier from below. If $q_\beta > 0$, then $y_\beta - \lambda = \theta'(q_\beta) \leq \theta'(1) = b$, while, if $q_\beta = 0$, then $y_\beta - \lambda \leq a \leq b$. In both cases, $y_\beta - \lambda \leq b$, and therefore

$$y_\alpha - \lambda = (y_\alpha - y_\beta) + (y_\beta - \lambda) \leq E_z(y) + b . \tag{F.56}$$

As φ is nondecreasing, we conclude that $\pi_i(\alpha \mid s) = q_\alpha = \varphi(y_\alpha - \lambda) \leq \varphi(b + E_z(y))$. ∎

If no action is excluded, then $\mathcal{C} = \Pi$ and there is nothing to prove, so we assume in what follows that $\mathcal{Z}$ is nonempty. For $M > 0$, we also consider the region of score space in which all linear energies lie below $-M$, namely

$$\mathcal{D}(M) \coloneqq \{ y \in \mathcal{Y} : \max_{z \in \mathcal{Z}} E_z(y) < -M \} . \tag{F.57}$$

The proof of Theorem 3 then rests on three lemmas. The first one identifies a region of score space in which every linear energy decreases at a linear pace, the second one shows that the accumulated noise and bias grow sublinearly in the cumulative step size, and the third one combines both facts to keep the iterates in the region of linear decrease forever.

Lemma F.7. *There exist a neighborhood U of $\mathcal{C}$ and constants $c > 0$ and $M > 0$ such that*

$$\langle v(\pi), z \rangle \leq -c \tag{F.58}$$

for every $\pi \in U$ and every $z \in \mathcal{Z}$, and such that

$$\psi(\mathcal{D}(M)) \subseteq U . \tag{F.59}$$

Proof. By Lemma F.4, there exists a neighborhood U of $\mathcal{C}$ on which (F.38) holds, and, for every $z = e^s_{i\alpha} - e^s_{i\beta} \in \mathcal{Z}$, this gives $\langle v(\pi), z \rangle = v^s_{i\alpha}(\pi) - v^s_{i\beta}(\pi) \leq -c^s_i \leq -c_*$, which is (F.58) with $c \coloneqq c_*$. Furthermore, let $y \in \mathcal{D}(M)$ and $\pi = \psi(y)$. For every active pair (i,s) and every excluded action $\alpha \in \bar{\mathcal{B}}_i(s)$, any allowed action $\beta \in \mathcal{B}_i(s)$ provides some $z \in \mathcal{Z}$ with $E_z(y) < -M$, and Lemma F.6 gives $\pi_i(\alpha \mid s) \leq \varphi^s_i(b^s_i - M)$. By (F.45), we thus have

$$\operatorname{dist}_1(\pi, \mathcal{C}) = 2 \sum_{i,s} \sum_{\alpha \in \bar{\mathcal{B}}_i(s)} \pi_i(\alpha \mid s) \leq 2 \sum_{i,s} |\bar{\mathcal{B}}_i(s)|\, \varphi^s_i\big(b^s_i - M\big) , \tag{F.60}$$

and the right-hand side vanishes as $M \to \infty$, since every rate function vanishes at $-\infty$. Choosing M large enough that the right-hand side is smaller than the radius r of U yields (F.59). ∎

We now confine the noise and the bias below a sublinear envelope. To this end, we consider, for every $z \in \mathcal{Z}$, the processes

$$S_{n,z} \coloneqq \sum_{k<n} \gamma_k \langle U_k, z \rangle \quad \text{and} \quad \bar{B}_n \coloneqq 2 \sum_{k<n} \gamma_k B_k \,, \tag{F.61}$$

which respectively collect the noise along z and bound the contribution of the bias, since $|\langle b_k, z \rangle| \le \|z\|_1 \|b_k\|_\infty \le 2B_k$.

Lemma F.8. *There exist deterministic constants $G > 0$ and $\alpha \in (0,1)$, which do not depend on the initialization, and an event Ω of probability at least $1-\eta$, on which*

$$S_{n,z} + \bar{B}_n \le G(1+\Gamma_{n-1})^\alpha \tag{F.62}$$

for every episode $n \ge 1$ and every $z \in \mathcal{Z}$.

Proof. We first choose $\alpha \in (0,1)$ such that

$$V_\alpha \coloneqq \sum_{n \ge 1} \frac{\gamma_n^2 \sigma_n^2}{(1+\Gamma_n)^{2\alpha}} < \infty \quad \text{and} \quad H_\alpha \coloneqq \sum_{n \ge 1} \frac{\gamma_n B_n}{(1+\Gamma_n)^\alpha} < \infty \,. \tag{F.63}$$

Such a choice is possible under Assumption 2. Indeed, if $\ell_\gamma < 1$, then $\Gamma_n \asymp n^{1-\ell_\gamma}$, and any $\alpha < 1$ with

$$\alpha > \max\left\{0, \frac{1-2\ell_\gamma+2\ell_\sigma}{2(1-\ell_\gamma)}, 1 - \frac{\ell_b}{1-\ell_\gamma}\right\} \tag{F.64}$$

makes the summands of (F.63) of order $\mathcal{O}(n^{-2\ell_\gamma+2\ell_\sigma-2\alpha(1-\ell_\gamma)})$ and $\mathcal{O}(n^{-\ell_\gamma-\ell_b-\alpha(1-\ell_\gamma)})$, respectively, with both exponents below -1. Note that the lower bound in (F.64) is indeed smaller than 1, as $\ell_\sigma < 1/2$ and $\ell_b > 0$. If $\ell_\gamma = 1$, then $\Gamma_n \asymp \log n$, and any $\alpha \in (1/2, 1)$ makes the summands of order $\mathcal{O}(n^{-2+2\ell_\sigma}/\log^{2\alpha} n)$ and $\mathcal{O}(n^{-1-\ell_b}/\log^\alpha n)$ for $n \ge 2$, which are summable, the case $\ell_\sigma = 1/2$ included.

Next, fix $z \in \mathcal{Z}$, and let $N_k \coloneqq S_{k+1,z}$ for $k \ge 0$, so that $N_0 = 0$ and $N_k - N_{k-1} = \gamma_k \langle U_k, z \rangle$ for $k \ge 1$. As $\|z\|_1 = 2$, the conditional Jensen inequality and (S) give

$$\mathbb{E}[(N_k - N_{k-1})^2 \,|\, \mathcal{F}_k] = \gamma_k^2 \,\mathbb{E}[\langle U_k, z \rangle^2 \,|\, \mathcal{F}_k] \le 4\gamma_k^2 \,\mathbb{E}[\|U_k\|_\infty^2 \,|\, \mathcal{F}_k] \le 4\gamma_k^2 \sigma_k^2 \,, \tag{F.65}$$

and (N_k) is a square-integrable martingale with respect to $(\mathcal{F}_{k+1})$. The weights $a_k \coloneqq (1+\Gamma_k)^\alpha$ being deterministic, positive and nondecreasing, the Hájek–Rényi inequality (Theorem B.3) yields, for every $R > 0$,

$$\mathbb{P}\left(\sup_{n \ge 1} \frac{|S_{n,z}|}{(1+\Gamma_{n-1})^\alpha} > R\right) = \mathbb{P}\left(\sup_{k \ge 1} \frac{|N_k|}{a_k} > R\right) \le \frac{1}{R^2} \sum_{k \ge 1} \frac{4\gamma_k^2 \sigma_k^2}{(1+\Gamma_k)^{2\alpha}} = \frac{4V_\alpha}{R^2} \,, \tag{F.66}$$

where the first equality uses $S_{1,z} = 0$. Taking a union bound over the finitely many $z \in \mathcal{Z}$, we then get

$$\mathbb{P}\left(\sup_{n \ge 1} \max_{z \in \mathcal{Z}} \frac{|S_{n,z}|}{(1+\Gamma_{n-1})^\alpha} > R\right) \le \frac{4|\mathcal{Z}|V_\alpha}{R^2} \,. \tag{F.67}$$

As for the bias, we have $\Gamma_k \le \Gamma_{n-1}$ for every $k < n$, and therefore

$$\bar{B}_n = 2 \sum_{k<n} (1+\Gamma_k)^\alpha \frac{\gamma_k B_k}{(1+\Gamma_k)^\alpha} \le 2(1+\Gamma_{n-1})^\alpha H_\alpha \,. \tag{F.68}$$

Choosing R large enough that the right-hand side of (F.67) is at most η, we let Ω be the complement of the event in (F.67), and $G \coloneqq R + 2H_\alpha$. On Ω, (F.68) then gives $S_{n,z} + \bar{B}_n \le (R+2H_\alpha)(1+\Gamma_{n-1})^\alpha = G(1+\Gamma_{n-1})^\alpha$ for every episode $n \ge 1$ and every $z \in \mathcal{Z}$, which is (F.62). As V_α and H_α are deterministic, so are R and G, which therefore do not depend on the initialization. ∎

As $\alpha < 1$, the sublinear envelope of Lemma F.8 exceeds half the linear drift of Lemma F.7 by a bounded amount only, namely

$$\Delta \coloneqq \sup_{t \ge 0} \left\{ G(1+t)^\alpha - \frac{c\,t}{2} \right\} < \infty \,, \tag{F.69}$$

and we set $\mathcal{D} \coloneqq \mathcal{D}(M+\Delta)$. This set is an intersection of finitely many open half-spaces, hence open. It is nonempty, as it contains the score which assigns 0 to every allowed action and $-(M+\Delta+1)$ to every excluded one, and it is unbounded, since lowering the excluded scores further keeps the score in $\mathcal{D}$. The initialization thus leaves enough room to absorb the excess Δ, and the following lemma makes this precise.

Lemma F.9. *For every initialization $Y_1 \in \mathcal{D}$, we have, on the event Ω of Lemma F.8,*

$$E_z(Y_n) < -M - \frac{c}{2}\Gamma_{n-1} \tag{F.70}$$

for every episode $n \geq 1$ and every $z \in \mathcal{Z}$. In particular, $\pi_n \in U$ for every $n \geq 1$.

Proof. We argue by induction on n. As $\Gamma_0 = 0$ and $\Delta \geq G > 0$, the bound holds for $n = 1$. Assume now that it holds up to episode $n-1$. Then $Y_k \in \mathcal{D}(M)$ for every $k \leq n-1$, and all the corresponding policies lie in U by (F.59). Telescoping (A.1), and using (F.58), the bound $|\langle b_k, z\rangle| \leq 2B_k$ and (F.62), we then obtain

$$\begin{aligned} E_z(Y_n) &\leq E_z(Y_1) - c\Gamma_{n-1} + S_{n,z} + \bar{B}_n \\ &< -M - \Delta - c\Gamma_{n-1} + G(1+\Gamma_{n-1})^\alpha \leq -M - \frac{c}{2}\Gamma_{n-1}\,, \end{aligned} \tag{F.71}$$

where the last inequality follows from the definition (F.69) of Δ. This completes the induction, and the last assertion follows from (F.59). ∎

We may now reap the fruits of our labor.

Proof of Theorem 3. Let $b_{\max}$ denote the maximum of b_i^s over all active pairs, and $\varphi_{\mathcal{B}}(u)$ the maximum of the rate functions $\varphi_i^s(u)$ over these pairs, which is again nondecreasing and vanishes as $u \to -\infty$. Fix an initialization $Y_1 \in \mathcal{D}$. On the event Ω, Lemma F.6 and Lemma F.9 give, for every episode $n \geq 1$, every active pair (i,s) and every excluded action $\alpha \in \bar{\mathcal{B}}_i(s)$,

$$\pi_{i,n}(\alpha \mid s) \leq \varphi_i^s\Big(b_i^s - M - \frac{c}{2}\Gamma_{n-1}\Big) \leq \varphi_{\mathcal{B}}\Big(b_{\max} - M - \frac{c}{2}\Gamma_{n-1}\Big)\,. \tag{F.72}$$

Summing over all excluded actions and invoking (F.45), we conclude that, on Ω and for every episode $n \geq 1$,

$$\begin{aligned} \operatorname{dist}_1(\pi_n, \mathcal{C}) = 2\sum_{i,s}\sum_{\alpha\in\bar{\mathcal{B}}_i(s)} \pi_{i,n}(\alpha \mid s) &\leq 2\sum_{i,s}|\bar{\mathcal{B}}_i(s)|\,\varphi_{\mathcal{B}}\Big(b_{\max} - M - \frac{c}{2}\Gamma_{n-1}\Big) \\ &= c_0\,\varphi_{\mathcal{B}}\big(c_1 - c_2\Gamma_{n-1}\big)\,, \end{aligned} \tag{F.73}$$

where $c_0 \coloneqq 2\sum_{i,s}|\bar{\mathcal{B}}_i(s)|$, $c_1 \coloneqq b_{\max} - M$ and $c_2 \coloneqq c/2$. As $\mathbb{P}(\Omega) \geq 1-\eta$ for every initialization in $\mathcal{D}$, and as $\varphi_{\mathcal{B}} = \varphi$ when the kernel does not depend on the player and the state, this is precisely the statement of Theorem 3. ∎

Proof of Corollary 2. It suffices to evaluate the rate function in Theorem 3, after shifting the cumulative step size by one iteration. Since $\Gamma_n = \Gamma_{n-1} + \gamma_n$ with $\gamma_n \leq \gamma$, we have

$$c_1 - c_2\Gamma_{n-1} \leq (c_1 + c_2\gamma) - c_2\Gamma_n\,. \tag{F.74}$$

For (PPG), the kernel is $\theta(p) = p^2/2$, with $\theta'(p) = p$, which gives $a = 0$, $b = 1$ and $\varphi(u) = \min\{[u]_+, 1\} \leq [u]_+$. Theorem 3 combined with the monotonicity of φ then yields

$$\operatorname{dist}_1(\pi_n, \operatorname{span}(\mathcal{B})) \leq c_0[(c_1 + c_2\gamma) - c_2\Gamma_n]_+\,, \tag{F.75}$$

which is the first bound of (28) with $c_1 + c_2\gamma$ in place of c_1. For (EPG), the kernel is $\theta(p) = p\log p$, with $\theta'(p) = 1 + \log p$, so that $a = -\infty$, $b = 1$ and $\varphi(u) = \min\{e^{u-1}, 1\} \leq e^{u-1}$, and Theorem 3 delivers

$$\operatorname{dist}_1(\pi_n, \operatorname{span}(\mathcal{B})) \leq c_0 e^{c_1 + c_2\gamma - 1}\exp(-c_2\Gamma_n)\,, \tag{F.76}$$

which is the second bound of (28) with $c_0 e^{c_1+c_2\gamma-1}$ in place of c_0. ∎

F.5 Numerical illustration

We close this appendix with the details of the numerical illustration of Figure 3. The game has two players and two states, a prisoner's dilemma (state 1) and a matching pennies game (state 2), and both players choose between C and D at both states. The stage games, with player 1 choosing the row and player 2 the column, each matrix listing the payoffs (r_1, r_2) at the state indicated in its corner, are given by

$$\begin{array}{c|cc} 1 & C & D \\ \hline C & (1/2, 1/2) & (-1, 1) \\ D & (1, -1) & (0, 0) \end{array} \qquad \begin{array}{c|cc} 2 & C & D \\ \hline C & (0, -1) & (-1, 0) \\ D & (-1, 0) & (0, -1) \end{array} \tag{F.77}$$

At state 1, defection thus strictly dominates cooperation for both players. At state 2, player 1 wins by matching the action of player 2, player 2 wins by mismatching it, and the winner receives 0 while the loser receives -1. After every stage, the game terminates with probability $1/10$, which gives $\zeta_{\min} = 1/10$. Conditionally on continuation, the profile CC keeps the game at state 1, while every other profile sends it to state 2, from which it returns to state 1 with probability $1/5$, regardless of the actions. Finally, the initial distribution is $\rho = (1/2, 1/2)$, so that the reachability condition (R) holds.

Let $\mathcal{B}_C$ and $\mathcal{B}_D$ be the state-action sets which prescribe, respectively, C and D to both players at state 1 and leave state 2 unrestricted. Both are Q-club, and, as nothing is excluded at state 2, it suffices to compare the Q-values at state 1. For $\pi \in \operatorname{span}(\mathcal{B}_C)$, the game remains at state 1 with probability $9/10$ under CC, which gives $V_{i,1}(\pi) = 5$ for both players, while, at state 2,

$$V_{i,2}(\pi) = \bar{r}_i + \frac{9}{50} V_{i,1}(\pi) + \frac{18}{25} V_{i,2}(\pi) = \bar{r}_i + \frac{9}{10} + \frac{18}{25} V_{i,2}(\pi)\,, \tag{F.78}$$

where $\bar{r}_i \leq 0$ is the expected reward of player i at state 2 under π, so that $V_{i,2}(\pi) = 25(\bar{r}_i + 9/10)/7 \leq 45/14$. Accordingly, for player 1, and symmetrically for player 2,

$$Q_1^\pi(1, C) = \frac{1}{2} + \frac{9}{10} \cdot 5 = 5 \quad \text{and} \quad Q_1^\pi(1, D) = 1 + \frac{9}{10} V_{1,2}(\pi) \leq \frac{109}{28} < 5\,. \tag{F.79}$$

For $\pi \in \operatorname{span}(\mathcal{B}_D)$, the opponent defects at state 1. Both actions of player 1 then lead to state 2 with probability $9/10$, so that the continuation terms cancel out in

$$Q_1^\pi(1, D) - Q_1^\pi(1, C) = \Big[0 + \frac{9}{10} V_{1,2}(\pi)\Big] - \Big[-1 + \frac{9}{10} V_{1,2}(\pi)\Big] = 1\,. \tag{F.80}$$

The same computation applies verbatim to player 2.

At state 2, on the other hand, the continuation values do not depend on the actions, and the players thus face a constant-sum version of matching pennies, in which the continuous-time entropic dynamics cycle around the interior equilibrium [40], while their discrete-time counterparts spiral away from it toward the boundary [41]. This is why, in Figure 3, the iterates converge to the span of $\mathcal{B}_C$ without settling at a single policy.

Both dynamics of Figure 3 are entropic, and both are run on the full stochastic game, for the same number of episodes. The deterministic trajectories follow (EPG) with exact feedback and $\gamma = 0.15$, whereas the stochastic ones follow (R-RPG) with the softmax map, $\gamma = 5 \cdot 10^{-4}$ and $\varepsilon_n = n^{-0.15}/2$, with one sampled episode per iteration.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: Our abstract and introduction follow the same logic and order as the technical results of the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: We discuss exact assumptions throughout. Moreover, in the concluding remarks we focus on limitations that are also natural avenues for future research.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: All the results state their assumptions explicitly, and full rigorous proofs are included in the appendix.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: The paper is theoretical, and its only experiment is an easily reproducible illustration, described in Appendix F.5.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: The paper is theoretical and uses no data, and its only experiment is an easily reproducible illustration (see Appendix F.5).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The illustration of Figure 3 is described in Appendix F.5.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A].

   Justification: The illustration of Figure 3 makes no statistical claim.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

Justification: The illustration of Figure 3 runs in a few minutes on a standard laptop.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes].

   Justification: The paper is theoretical and does not raise ethical concerns.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A].

    Justification: The paper is theoretical and has no clear direct societal impact.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper is theoretical and releases no data or models.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A].

    Justification: The paper is theoretical and uses no existing assets.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: The paper is theoretical and introduces no new assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: The core methodology does not involve LLMs as an important, original, or non-standard component.