LEARNING IN GAMES FROM A STOCHASTIC APPROXIMATION VIEWPOINT

PANAYOTIS MERTIKOPOULOS[*,⋆], YA-PING HSIEH[§], AND VOLKAN CEVHER[♯]

Abstract. We develop a unified stochastic approximation framework for analyzing the long-run behavior of multi-agent online learning in games. Our framework is based on a "primal-dual", mirrored Robbins–Monro (MRM) template which encompasses a wide array of popular game-theoretic learning algorithms (gradient methods, their optimistic variants, the EXP3 algorithm for learning with payoff-based feedback in finite games, etc.). In addition to providing an integrated view of these algorithms, the proposed MRM blueprint allows us to obtain a broad range of new convergence results, both asymptotic and in finite time, in both continuous and finite games.

1. Introduction

The prototypical setting of online learning in games can be summarized as follows:

(1) At each stage of the decision-making process, every player selects an action.

(2) The players receive a reward determined by their chosen actions and their individual payoff functions – assumed a priori unknown.

(3) Based on these payoffs and any other observed information, the players update their actions and the process repeats.

A key question that arises in this general setting is whether the players eventually settle down to a stable profile from which no player has an incentive to deviate. Put differently:

Does the players' learning process converge to a Nash equilibrium?

This question has occupied the forefront of game-theoretic research ever since the field's earliest steps in the 1950's. On the positive side, fictitious play [12, 57] and its variants [29, 40] provided the first equilibrium convergence results for certain classes of finite games – zero-sum, potential and 2×2 games. On the negative side, the well-known impossibility results of Hart and Mas-Colell [26, 27] showed that an unconditional positive answer to this question is out of reach: it is not possible to devise an uncoupled learning rule – deterministic or stochastic – that converges to Nash equilibrium in all games. As a result, contemporary research on learning in games has focused on extending the classes of games in which positive results can be obtained, relaxing the feedback requirements of the players' learning process, and understanding the convergence failures of popular learning algorithms.

[*] Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France.

[⋆] Criteo AI Lab.

[§] Institute for Machine Learning, CAB G69.3, Universitaetstrasse 6, 8092 Zurich, Switzerland.

[♯] Laboratory for Information and Inference Systems, IEL STI EPFL, 1015 Lausanne, Switzerland.

E-mail addresses: panayotis.mertikopoulos@imag.fr, yaping.hsieh@inf.ethz.ch, volkan.cevher@epfl.ch.

2020 *Mathematics Subject Classification.* Primary 91A10, 91A26; secondary 68Q32, 68T02.

Key words and phrases. Nash equilibrium; continuous games; finite games; stochastic approximation; variational stability; primal attractors; coherence.

The first author is grateful to Angeliki Giannou, Kyriakos Lotidis, Sylvain Sorin, and Manolis Vlatakis for many fruitful discussions during the preparation and revision of this manuscript.

In turn, this has led to a vast, interdisciplinary corpus of literature that is impossible to survey here. Historically, much of this literature has focused on games with a *finite* number of actions, which are prevalent in economic models of rationality. More recently however, the surge of breakthroughs in data science and machine learning (robust reinforcement learning, generative adversarial networks, etc.), has generated an intense interest in games with *continuous* action spaces. These two threads of the literature have evolved essentially in isolation, with little communication between them, and sometimes with overlooked connections.

Our contributions. Our paper aims to partially bridge this disconnect by providing a unified framework for the analysis of a wide range of game-theoretic learning algorithms, for both continuous and finite games. The basic ingredients of our approach are twofold: First, we introduce an abstract *mirrored Robbins–Monro* (MRM) "parent scheme" which includes as special cases many popular methods for learning in games (gradient schemes, their optimistic variants, the exponential weights and EXP3 algorithms for finite games, etc.). We then couple the proposed template with a suitable "primal-dual" dynamical system in continuous time which is sufficiently flexible to accommodate different types of feedback and update structures, and whose asymptotic behavior reflects that of the algorithms under study.

In this general context, the principal axes of our analysis can be summarized as follows:

(1) First, regarding solutions and behaviors contained in the interior of the game's action space, we introduce the notion of *subcoercivity*, a structural condition which ensures that all MRM algorithms converge to the internally chain transitive (ICT) sets of the underlying continuous-time dynamics (more specifically, that the algorithm's limit sets are internally chain transitive). This allows us to deduce a wide range of almost sure equilibrium convergence results for min-max and potential games, both constrained and unconstrained.

(2) To deal with outcomes at the boundary of the game's action space, we introduce the concept of a *primal attractor*, and we show that the induced trajectory of play converges locally to such sets with arbitrarily high probability (or globally with probability 1, depending on the attractor). As an immediate corollary of this result, our analysis directly implies convergence to Nash equilibrium in all strictly monotone games, and we are likewise able to infer a series of local convergence results to equilibria that satisfy a variational stability or a second-order sufficient condition in continuous games – and this, even with bandit, payoff-based feedback.

(3) We further introduce the notion of *coherence* (which covers strict Nash equilibria in finite games, sharp equilibria in continuous games, linear programs, etc.), and we show that MRM methods converge to such sets under significantly weaker conditions for their runtime parameters (step-size, sampling radius, etc.). In addition, we are also able to show in this case that the induced trajectory of play achieves convergence in a *finite* number of iterations if the players' mirror map is surjective (e.g., as in projection-based schemes).

An appealing feature of our analysis is that it applies to both *first-order* ("oracle-based") and *zeroth-order* ("payoff-based") methods. In this regard, our results provide an integrated proof technique that can be easily adapted to many other learning algorithms in the literature, reducing in this way the number of ad hoc elements required to analyze a given method.

A note on related work. The set of example algorithms that we use to illustrate our analysis includes (stochastic) gradient methods in the spirit of [1], extra-gradient [34, 37, 50] and optimistic gradient schemes [53, 54], the Hedge and EXP3 algorithms for learning in finite games [3, 42, 67], as well as the single-point stochastic approximation (SPSA) [11, 63] and dampened gradient approximation (DGA) [9] methods for payoff-based learning in continuous games. Of course, given the breadth and depth of the relevant literature, it is impossible to survey here all methods covered by the proposed MRM template – or that *could* be covered modulo minor modifications. Our selection is only meant to highlight different trends in the literature, and to show how some algorithms that initially seem unrelated – like DGA – can be included in our framework.

Paper outline. In Section 2, we introduce the game-theoretic background of our work, including the various solution concepts that we use throughout our paper (critical points, Nash equilibria, variationally stable states, etc.). Subsequently, in Sections 3 and 4, we introduce a range of well-known algorithms for learning in games, and we show how they can be seen as special instances of the MRM blueprint. Our analysis proper begins in Section 5, where we introduce the notion of subcoercivity and present our ICT convergence results. Subsequently, in Sections 6 and 7, we state and prove our main convergence results for primal attractors and coherent sets respectively.

2. Preliminaries

Notation. In what follows, $\mathcal{V}$ will denote a d-dimensional real space with norm $\|\cdot\|$. We will also write $\mathcal{Y} \coloneqq \mathcal{V}^*$ for the dual space of $\mathcal{V}$, $\langle y, x\rangle$ for the canonical pairing between $y \in \mathcal{Y}$ and $x \in \mathcal{V}$, and $\|y\|_* \coloneqq \max\{\langle y, x\rangle : \|x\| \leq 1\}$ for the induced dual norm on $\mathcal{Y}$. As is customary, if $\mathcal{V}$ is Euclidean, we will not distinguish between primal and dual vectors.

2.1. Games in normal form.

Throughout the sequel, we will focus on games with a finite number of players $i \in \mathcal{N} = \{1, \dots, N\}$, each selecting an *action* x_i from some closed convex subset $\mathcal{X}_i$ of a d_i-dimensional normed space $\mathcal{V}_i$. Gathering all players together, we will write $\mathcal{X} = \prod_i \mathcal{X}_i$ for the space of all *action profiles* $x = (x_i)_{i\in\mathcal{N}}$ and $d = \sum_i d_i$ for the dimension of the ambient space $\mathcal{V} = \prod_i \mathcal{V}_i$. Finally, when we want to distinguish between the action of the i-th player and that of all other players, we will employ the shorthand $(x_i; x_{-i})$.

Given an action profile $x \in \mathcal{X}$, each player $i \in \mathcal{N}$ is assumed to receive a reward $u_i(x) \equiv u_i(x_i; x_{-i})$ based on an associated *payoff function* $u_i \colon \mathcal{X} \to \mathbb{R}$. In terms of regularity, we will tacitly assume that u_i is differentiable and we will write

$$v_i(x) = \nabla_{x_i} u_i(x_i; x_{-i}) \quad \text{and} \quad v(x) = (v_i(x))_{i\in\mathcal{N}} \tag{1}$$

for the players' *individual payoff gradients* and the ensemble thereof. Finally, unless explicitly mentioned otherwise, we will treat each $v_i(x)$ as an element of the corresponding dual space $\mathcal{Y}_i = \mathcal{V}_i^*$ of $\mathcal{V}_i$, and we will make the following blanket assumption throughout the rest of our paper:

Assumption 1. The players' payoff functions are *Lipschitz continuous and smooth*, i.e., there exist constants $G_i, L_i \geq 0$, $i \in \mathcal{N}$, such that

$$|u_i(x') - u_i(x)| \leq G_i\|x' - x\| \quad \text{and} \quad \|\nabla u_i(x') - \nabla u_i(x)\|_* \leq L_i\|x' - x\|. \tag{2}$$

for all $x, x' \in \mathcal{X}$, $i \in \mathcal{N}$. For concision, we will also write $G \coloneqq \max_i G_i$ and $L \coloneqq \max_i L_i$.

With all this in hand, a *continuous game in normal form* is a tuple $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{X}, u)$ with players, actions and payoff functions as above. For concreteness, we provide some examples below:

Example 2.1 (Min-max games)**.** Consider two players $i \in \{1, 2\}$ with action spaces $\mathcal{X}_1$ and $\mathcal{X}_2$, and payoff functions $u_1 = -\ell = -u_2$ for some smooth function $\ell \colon \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$. Player 1 (the "min" player) seeks to minimize $\ell = -u_1$ whereas Player 2 (the "max" player) seeks to maximize $\ell = u_2$. In many applications, ℓ is (strictly) convex-concave, in which case von Neumann's theorem asserts that the game $\min_{x_1 \in \mathcal{X}_1} \max_{x_2 \in \mathcal{X}_2} \ell(x_1, x_2)$ always admits a solution if $\mathcal{X}_1 \times \mathcal{X}_2$ is compact. ¶

Example 2.2 (Cournot oligopolies)**.** Consider N firms supplying the market with a quantity $x_i \in [0, C_i]$ of some good up to each firm's capacity C_i. The good is priced as a function $P(x) = a - b \sum_{i=1}^{N} x_i$ of the total quantity of the good in the market, so the net utility of the i-th firm is $u_i(x) = x_i P(x) - c_i x_i$ where a, b and c_i are market-related positive constants. The resulting game $\mathcal{G}(\mathcal{N}, \mathcal{X}, u)$ is known as a Cournot competition game and it plays a central role in economic theory. ¶

Example 2.3 (Power control)**.** As another example, consider N users transmitting a stream of packets to a common receiver over K shared wireless channels [66]. If the channel gain for the i-th user over the k-th channel is g_{ik} and each user transmits with total power $P_{\max}$ split over each channel as p_{ik}, $\sum_k p_{ik} = P_{\max}$, the user's transmission rate will be given by the Shannon formula

$$R_i(p_i; p_{-i}) = \sum_{k=1}^{K} \log\left(1 + \frac{g_{ik} p_{ik}}{\sigma + \sum_{j \neq i} g_{jk} p_{jk}}\right), \qquad p_i \in \mathcal{P}_i \coloneqq P_{\max} \, \Delta(K), \tag{3}$$

where $\sigma > 0$ denotes the ambient noise in the channel. The resulting game $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{P}, r)$ is known as the power control problem and it is a core aspect of wireless network design [66]. ¶

Example 2.4 (Finite games)**.** In a finite game $\Gamma \equiv \Gamma(\mathcal{N}, \mathcal{A}, u)$, each player $i \in \mathcal{N}$ chooses an action α_i from some finite set $\mathcal{A}_i$; the players' payoffs are then determined by the action profile $\alpha = (\alpha_1, \dots, \alpha_N) \in \mathcal{A} \coloneqq \prod_i \mathcal{A}_i$ and an ensemble of payoff functions $u_i \colon \mathcal{A} \to \mathbb{R}$, $i = 1, \dots, N$. In the *mixed extension* of Γ, a player may pick an action according to a probability distribution $x_i \in \Delta(\mathcal{A}_i)$: this is known as a *mixed strategy*, and the corresponding mixed payoff to the i-th player is $u_i(x) = \sum_{\alpha \in \mathcal{A}} x_\alpha u_i(\alpha)$ where $x_\alpha = \prod_i x_{i\alpha_i}$ is the probability of the action profile $\alpha = (\alpha_1, \dots, \alpha_N)$.

Letting $\mathcal{X}_i = \Delta(\mathcal{A}_i)$, the mixed extension of Γ is defined as the continuous game $\Delta(\Gamma) = \mathcal{G}(\mathcal{N}, \mathcal{X}, u)$. For posterity, we note here that the "payoff gradient" of each player $i \in \mathcal{N}$ is simply their mixed payoff vector, i.e., $v_i(x) = \nabla_{x_i} u_i(x) = (u_i(\alpha_i; x_{-i}))_{\alpha_i \in \mathcal{A}_i}$. ¶

2.2. **Solution concepts.** The standard solution concept in game theory is that of a Nash equilibrium, i.e., an action profile that is resilient to unilateral deviations. Formally, we say that $x^* \in \mathcal{X}$ is a *Nash equilibrium* of the game $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{X}, u)$ if

$$u_i(x^*) \geq u_i(x_i; x^*_{-i}) \qquad \text{for all } x_i \in \mathcal{X}_i,\ i \in \mathcal{N}. \tag{NE}$$

Nash equilibria always exist if $\mathcal{X}$ is compact and each u_i is individually concave in x_i [19]. Otherwise, Nash equilibria may fail to exist, in which case we will consider the following relaxations:

(1) *Local Nash equilibria*, i.e., profiles $x^* \in \mathcal{X}$ for which (NE) holds locally:

$$u_i(x^*) \geq u_i(x_i; x^*_{-i}) \qquad \text{for all } x \text{ in a neighborhood } \mathcal{U} \text{ of } x^* \text{ in } \mathcal{X}. \tag{LNE}$$

(2) *Critical points*, i.e., profiles $x^* \in \mathcal{X}$ that satisfy the first-order stationarity condition:

$$\tfrac{d}{dt}\Big|_{t=0^+} u_i(x^*_i + t(x_i - x^*_i); x^*_{-i}) \leq 0 \qquad \text{for all } x_i \in \mathcal{X}_i,\ i \in \mathcal{N}. \tag{FOS}$$

Equivalently, (FOS) can be reformulated as a Stampacchia variational inequality of the form

$$\langle v(x^*), x - x^* \rangle \leq 0 \quad \text{for all } x \in \mathcal{X}. \tag{SVI}$$

The solutions of (SVI) are precisely the fixed points of the "linearized" best-response correspondence $x \mapsto \arg\max_{x' \in \mathcal{X}} \langle v(x), x' \rangle$ so, by standard fixed point arguments, the set of critical points of $\mathcal{G}$ is always nonempty if $\mathcal{X}$ is compact (independently of concavity or other considerations).

Remark. In operator theory and optimization, the direction of (SVI) is reversed because optimization problems are typically formulated as cost minimization problems. The utility maximization viewpoint is more common in game theory, so we will maintain the above sign convention throughout. ¶

Dually to the above, the *Minty variational inequality* associated to $\mathcal{G}$ is

$$\langle v(x), x - x^* \rangle \leq 0 \quad \text{for all } x \in \mathcal{X}. \tag{MVI}$$

It is straightforward to verify that the solutions of (MVI) comprise a convex set of Nash equilibria of $\mathcal{G}$, so (MVI) can be seen as an equilibrium refinement criterion for $\mathcal{G}$. Taking this a step further, a state $x^* \in \mathcal{X}$ is said to be *variationally stable* if

$$\langle v(x), x - x^* \rangle < 0 \qquad \text{for all } x \neq x^* \text{ in a neighborhood } \mathcal{U} \text{ of } x^* \text{ in } \mathcal{X} \tag{VS}$$

and x^* is called *neutrally stable* if the strict inequality "$<$" in (VS) is relaxed to "$\leq$", i.e., if

$$\langle v(x), x - x^* \rangle \leq 0 \qquad \text{for all } x \text{ in a neighborhood } \mathcal{U} \text{ of } x^* \text{ in } \mathcal{X}. \tag{NS}$$

Finally, we say that x^* is *globally variationally stable* [resp. *globally neutrally stable*] if (VS) [resp. (NS)] holds with $\mathcal{U} = \mathcal{X}$ (i.e., for all $x \in \mathcal{X}$).

In general, the solution concepts discussed above are related as follows:

$$\begin{tikzcd}
\mathrm{GVS} \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \mathrm{GNS} \equiv \mathrm{MVI} \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \mathrm{NE} \arrow[d, Rightarrow] & \\
\mathrm{VS} \arrow[r, Rightarrow] & \mathrm{NS} \arrow[r, Rightarrow] & \mathrm{LNE} \arrow[r, Rightarrow] & \mathrm{FOS} \equiv \mathrm{SVI}
\end{tikzcd} \tag{4}$$

Without further assumptions, the implications in (4) are all one-way; in the next paragraph, we discuss a number of cases where some (or all) of these implications become equivalences.

Remark. The definition of variational stability echoes the seminal notion of *evolutionary stability* as introduced by Maynard Smith and Price [43] in the context of population games. To make this connection precise, consider a unit mass population of players with a finite set of pure strategies $\mathcal{A}$, and let $v_\alpha(x)$ denote the payoff to α-strategists when the state of the population is $x \in \Delta(\mathcal{A})$. Then a state $x^* \in \Delta(\mathcal{A})$ is *evolutionarily stable* if $\langle v(\delta x + (1-\delta)x^*), x - x^* \rangle < 0$ for all sufficiently small $\delta > 0$ and all $x \neq x^*$. As was shown by Taylor [65] and Hofbauer et al [31], a state x^* is evolutionarily stable if and only if it satisfies (VS), an equivalence which justifies our choice of terminology. ¶

2.3. **Special cases and classes of games.** We close this section with a discussion of some special cases of the above definitions that will play a major role in the sequel.

Monotone games. A game is said to be *monotone* if it satisfies the monotonicity condition

$$\langle v(x') - v(x), x' - x \rangle \leq 0 \quad \text{for all } x, x' \in \mathcal{X}. \tag{Mon}$$

The strict version of this requirement (i.e., that equality holds if and only if $x = x'$) is sometimes referred to as *diagonal strict concavity* (DSC), a terminology due to Rosen [58]. In monotone games, the solutions of (MVI) and (SVI) coincide, leading to the string of equivalences MVI $\iff$ NE $\iff$ LNE $\iff$ FOS $\iff$ SVI. By comparison, if a game is strictly monotone, every implication in (4) becomes an equivalence, so the game admits a unique, globally variationally stable Nash equilibrium.

Examples 2.1–2.3 are all strictly monotone (assuming ℓ is strictly convex-concave in Example 2.1); other examples include Kelly auctions [36], Fisher markets [51], resource allocation problems in communication networks [62], and many other classes of problems in economics and control. ¶

Potential games. First formalized by Monderer and Shapley [49], potential games are games that admit a *potential function* $\Phi\colon \mathcal{X} \to \mathbb{R}$ such that

$$u_i(x_i; x_{-i}) - u_i(x_i'; x_{-i}) = \Phi(x_i; x_{-i}) - \Phi(x_i'; x_{-i}) \qquad \text{for all } x, x' \in \mathcal{X} \text{ and all } i \in \mathcal{N}. \tag{Pot}$$

If $\mathcal{G}$ is a potential game, we have $v(x) = \nabla\Phi(x)$ so any local maximum of Φ is a local Nash equilibrium of $\mathcal{G}$ and any *strict* local maximum of Φ is variationally stable.

In terms of examples, all finite congestion games are potential games [59]; Example 2.3 likewise admits the potential $R(p) = \sum_{k=1}^{K} \log(1 + \sum_j g_{jk} p_{jk})$ [47]. ¶

Finite games. Let $\mathcal{G} = \Delta(\Gamma)$ be the mixed extension of a finite game $\Gamma \equiv \Gamma(\mathcal{N}, \mathcal{A}, u)$. Since each player's payoff function $u_i(x_i; x_{-i})$ is linear in x_i, we immediately get the equivalences NE $\iff$ LNE $\iff$ FOS $\iff$ SVI. In addition, we have the following important characterization of variationally stable states in finite games:

Proposition 1. *A mixed strategy profile x^* is variationally stable if and only if it is a strict Nash equilibrium of Γ, i.e., if and only if* (NE) *holds as a strict inequality for all $x \neq x^*$.*

The analogue of Proposition 1 for evolutionary stability is a folk result in evolutionary game theory, cf. Sandholm [61, Chap. 8]; the two notions coincide for finite games, so we omit the proof. ¶

Second-order stationarity. Our last example concerns critical points that satisfy a condition similar to second-order sufficient conditions in optimization:

Proposition 2. *Let x^* be a critical point of $\mathcal{G}$ satisfying the second-order stationarity condition*

$$z^\top \operatorname{Jac}_v(x^*) z < 0 \quad \textit{for all nonzero tangent vectors } z \textit{ to } \mathcal{X} \textit{ at } x^*, \tag{SOS}$$

where $\operatorname{Jac}_v(x^*)$ denotes the Jacobian of v at x^*. Then x^* is variationally stable.

For the proof of Proposition 2, see Hsieh et al [32, Lemma A.4]. In the context of saddle-point problems and continuous games, the second-order condition (SOS) has been widely studied in the machine learning and control literatures, cf. [4, 5, 32, 41, 46, 55, 56] and references therein. ¶

3. Online learning policies and algorithms

We now proceed to describe a representative range of methods for learning in games. Depending on the information available to the players, we classify the algorithms under study as *oracle-based* or *payoff-based*; our presentation scheme in the rest of this section reflects this taxonomy.

3.1. Oracle-based methods.

The commmon denominator of the algorithms we discuss below is that players have access to a "black-box" feedback mechanism – a *stochastic first-order oracle* (SFO) – that returns an estimate of their individual payoff gradients at their chosen action profile.[1] Formally, when queried at $x \in \mathcal{X}$, an SFO outputs a random vector of the form

$$V(x; \theta) = v(x) + \operatorname{Err}(x; \theta), \tag{SFO}$$

where θ is a random variable taking values in some measurable space Θ and $\operatorname{Err}(x; \theta)$ is an umbrella error term capturing all sources of uncertainty in the model.

[1] Methods that require full, "white-box" knowledge of the game's payoff functions are not treated here.

In practice, (SFO) is queried repeatedly at a sequence of action profiles $X_n \in \mathcal{X}$, $n = 1, 2, \dots,$ possibly with a different random seed θ_n at each time.[2] To keep track of this sequence of events, we will view X_n as a stochastic process on some complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and we will write $\mathcal{F}_n \coloneqq \mathcal{F}(X_1, \dots, X_n) \subseteq \mathcal{F}$ for the *history of play* up to – and including – stage n. Accordingly, since the randomness entering the oracle is triggered only *after* each player has selected an action, we will posit throughout that θ_n is $\mathcal{F}_{n+1}$-measurable – though not necessarily $\mathcal{F}_n$-measurable. For concreteness, we will also assume that the noise in (SFO) is zero-mean and bounded in L^q for some $q \geq 2$, i.e.,

$$\mathbb{E}[\operatorname{Err}(x;\theta_n) \,|\, \mathcal{F}_n] = 0 \quad \text{and} \quad \mathbb{E}[\|\operatorname{Err}(x;\theta_n)\|_*^q \,|\, \mathcal{F}_n] \leq \sigma^q \tag{5}$$

for some $\sigma \geq 0$ and all $x \in \mathcal{X}$. In particular (and in a slight abuse of notation), the case $q = \infty$ above refers to the case when the noise is bounded w.p.1, that is, $\|\operatorname{Err}(x;\theta_n)\|_* \leq \sigma$ (a.s.).

We are now in a position to introduce the array of oracle-based methods under study. For simplicity, we present some of these policies in an unconstrained setting; this is only done to lighten the notation.

Algorithm 1 (Stochastic gradient ascent)**.** Perhaps the most basic iterative policy for multi-agent online learning is the standard (individual) gradient ascent method

$$X_{i,n+1} = X_{i,n} + \gamma_n V_i(X_n;\theta_n). \tag{SGA}$$

Informally, (SGA) implies that all players are simultaneously taking a stochastic gradient step with step-size $\gamma_n > 0$. From a loss minimization viewpoint, (SGA) is a multi-agent analogue of the standard stochastic gradient descent algorithm; analogously, in min-max games, (SGA) boils down to the Arrow–Hurwicz method [1]. ¶

Algorithm 2 (Sequential gradient ascent)**.** A common variant of (SGA) is when players update their actions turn-by-turn instead of simultaneously. This results in the *sequential gradient ascent* policy

$$X_{i,n+1} = X_{i,n} + \gamma_n V_i(\dots, X_{i-1,n+1}, X_{i,n}, X_{i+1,n}, \dots;\theta_n). \tag{seqGA}$$

Formally, the only difference between (SGA) and (seqGA) lies in the oracle query at the update step: instead of computing all gradients at X_n, each individual gradient in (seqGA) is computed sequentially after each player has locked in an action. This variant is widely used in generative adversarial networks (GANs) [18, 24, 33]. ¶

Algorithm 3 (Extra-gradient)**.** Going a step further from (SGA), the *extra-gradient* (EG) algorithm of Korpelevich [37] is based on the following principle: starting at some "base" state X_n, the players first take a gradient step to an interim, "leading" state $X_{n+1/2}$; subsequently, to anticipate their payoff landscape, they update the base state X_n with gradient information from $X_{n+1/2}$ instead of X_n. Formally, this leads to the policy

$$\begin{aligned} X_{i,n+1/2} &= X_{i,n} + \gamma_n V_i(X_n;\theta_n), \\ X_{i,n+1} &= X_{i,n} + \gamma_n V_i(X_{n+1/2};\theta_{n+1/2}). \end{aligned} \tag{EG}$$

For applications of this method to GANs and robust reinforcement learning, see [35, 48]. ¶

Algorithm 4 (Optimistic gradient)**.** A computational drawback of (EG) is that it requires two oracle queries per update – and hence, more overhead per iteration. One way to overcome

[2] In some cases, the index set may be enlarged to include all positive half-integers ($n = 1/2, 1, 3/2, \dots$).

this hurdle is to reuse past gradient information in the hope that it provides a good enough approximation of the present; this leads to the *optimistic gradient* policy

$$\begin{aligned} X_{i,n+1/2} &= X_{i,n} + \gamma_n V_i(X_{n-1/2}; \theta_{n-1}), \\ X_{i,n+1} &= X_{i,n} + \gamma_n V_i(X_{n+1/2}; \theta_n). \end{aligned} \tag{OG}$$

The "gradient reuse" idea in (OG) dates back at least to Popov [53], and it has resurfaced several times in the literature since then, cf. [18, 32, 54] and references therein. To simplify our presentation, we will assume in the sequel that the method is run with an SFO satisfying (5) with $q = \infty$. ¶

The next method concerns learning in mixed extensions of finite games.

Algorithm 5 (Exponential weights)**.** Let $\mathcal{G} = \Delta(\Gamma)$ be the mixed extension of a finite game $\Gamma(\mathcal{N}, \mathcal{A}, u)$ as per Example 2.4. In this setting, the players' learning process typically unfolds as follows: at each stage $n = 1, 2, \dots$, every player selects a mixed strategy $X_{i,n} \in \Delta(\mathcal{A}_i)$ and draws a pure strategy $\alpha_{i,n} \in \mathcal{A}_i$ according to $X_{i,n}$. Then, depending on the amount of information available to the players, we have the following oracle models:

(1) *Full information feedback:* in this case, players observe their *mixed* payoff vectors, i.e.,

$$V_i(X_n; \alpha_n) = v_i(X_{i,n}; X_{-i,n}). \tag{6a}$$

(2) *Realization-based feedback:* here, players instead observe their *pure* payoff vectors, i.e.,

$$V_i(X_n; \alpha_n) = v_i(\alpha_{i,n}; \alpha_{-i,n}). \tag{6b}$$

In both models, the seed of the oracle is the action profile α_n chosen by the players at stage n: the oracle (6a) is deterministic, while the oracle (6b) is stochastic and satisfies (5) with $q = \infty$.

In this context, one of the most widely used methods is the *exponential weights* algorithm

$$\begin{aligned} Y_{i,n+1} &= Y_{i,n} + \gamma_n V_i(X_n; \alpha_n) \\ X_{i,n+1} &= \Lambda_i(Y_{i,n+1}) \end{aligned} \tag{EW}$$

where Λ_i denotes the "logit choice" map

$$\Lambda_i(y_i) = \frac{(\exp(y_{i\alpha_i}))_{\alpha_i \in \mathcal{A}_i}}{\sum_{\alpha_i \in \mathcal{A}_i} \exp(y_{i\alpha_i})} \tag{7}$$

and V_i is given by (6a) or (6b) depending on the information available to the players. This method has a very long history in online learning and game theory; for an appetizer, see Littlestone and Warmuth [42], Vovk [67] and Auer et al [3]. ¶

The last oracle-based method we present concerns games with general action spaces.

Algorithm 6 (Mirror-prox)**.** An important generalization of the extra-gradient method (Algorithm 3) is the *mirror-prox* algorithm [34, 50], which we define here as follows:

$$\begin{aligned} Y_{n+1/2} &= Y_n + \gamma_n V(X_n; \theta_n) \qquad & Y_{n+1} &= Y_n + \gamma_n V(X_{n+1/2}; \theta_{n+1/2}) \\ X_{n+1/2} &= Q(Y_{n+1/2}) & X_{n+1} &= Q(Y_{n+1}). \end{aligned} \tag{MP}$$

In the above, $Q\colon \mathcal{Y} \to \mathcal{X}$ denotes the namesake *mirror map* of the method, as it is used to "mirror" the extra-gradient chain $Y_n \to Y_{n+1/2} \to Y_{n+1} \in \mathcal{Y}$ into feasible action profiles $X_n \to X_{n+1/2} \to X_{n+1} \in \mathcal{X}$. As an example, in the Euclidean unconstrained case ($\mathcal{X} = \mathcal{V}$), the identity map $Q(y) = y$ yields the extra-gradient algorithm (Algorithm 3); as another example, in finite games, letting $Q = \Lambda$ gives the so-called "simplex setup" of (MP). We provide more details on this construction in Section 4 below. ¶

3.2. **Payoff-based methods.** In many applications to machine learning and data science, unbiased gradient estimates can be obtained by making a partial pass over the problem's training data; in such cases, players can query (SFO) directly. By contrast, in applications to auctions, online advertising and networks, players may only be able to observe their realized payoffs. In this case, gradients must be somehow reconstructed from payoff-based observations, as players cannot perform direct queries to (SFO). In view of this distinction, we describe below a range of *payoff-based* policies where players do not have access to an oracle, but must infer gradient information indirectly.

Algorithm 7 (Single-point stochastic approximation)**.** A straightforward way of reconstructing gradients from zeroth-order feedback is via the *single-point stochastic approximation* framework of Spall [63]. In the unconstrained case ($\mathcal{X} = \mathcal{V}$), the relevant update step is:

$$\begin{aligned} \hat{X}_{i,n} &= X_{i,n} + \delta_n W_{i,n}, \\ X_{i,n+1} &= X_{i,n} + \gamma_n (u_i(\hat{X}_n)/\delta_n)\, W_{i,n}. \end{aligned} \tag{SPSA}$$

In (SPSA), each player's "query state" $\hat{X}_{i,n}$, $i = 1, \dots, N$, is a perturbation of the "base state" $X_{i,n}$ by a step of magnitude $\delta_n > 0$ along a random direction $W_{i,n}$ drawn from the ensemble of signed basis vectors $\mathcal{E}_i \coloneqq \{(\pm e_1, \dots, \pm e_{d_i})\}$. Importantly, (SPSA) can be seen as a special case of (SGA) with the "virtual" oracle $V_i(x; w) = (d_i/\delta)\, u_i(x + \delta w)\, w_i$ where each w_i is drawn uniformly at random from $\mathcal{E}_i$. As we discuss in Section 4.3, this estimator is biased, so it does not satisfy (5); see also Table 1.

We should stress here that the above formulation of (SPSA) is tailored to unconstrained problems. In this case, to ensure that the resulting gradient estimator remains bounded, it is customary to include an indicator of the form $\mathbb{1}(\|\hat{X}_n\| \leq R_n)$ for some suitably chosen sequence $R_n \to \infty$ [63]. This would lead to the same analysis but at the cost of heavier notation; thus, to avoid overloading the presentation, we will assume instead that the players' payoff functions are bounded when discussing (SPSA).

Of course, in games with compact action spaces (as per Section 2) this last point is moot. In that case however, certain book-keeping adjustments are required to ensure that $\hat{X}_n \in \mathcal{X}$ for all n. For a detailed discussion of how to adapt (SPSA) in the presence of constraints, we refer the reader to Bravo et al [11] who show that the relevant entries of Table 1 apply verbatim when $\mathcal{X}$ is compact. ¶

Algorithm 8 (Dampened gradient approximation)**.** An alternative approach to (SPSA) is the "explore-then-update", two-point sampling approach of Bervoets et al [9]. Specifically, Bervoets et al [9] focused on games with $\mathcal{X}_i = [0, \infty)$ for all $i \in \mathcal{N}$ and introduced the *dampened gradient approximation* policy

$$\begin{aligned} X_{i,n+1/2} &= X_{i,n} + (1/n) W_{i,n}, \\ X_{i,n+1} &= X_{i,n}[1 + (u_i(X_{n+1/2}) - u_i(X_n)) W_{i,n}], \end{aligned} \tag{DGA}$$

where, at each $n = 1, 2, \dots$, the "exploration direction" $W_{i,n}$ is sampled uniformly at random from $\{\pm 1\}$. In words, (DGA) is a two-stage process in which players first "explore" their individual payoff functions at a nearby state, and then use this information to estimate their individual payoff gradients and update their base state. ¶

Algorithm 9 (The EXP3 algorithm)**.** In our final example, we return to finite games, and we focus on the "bandit" case where players can only observe the payoffs of the pure strategies that they actually played. In this setting, it is common to employ the *importance-weighted* *estimator*

$$V_{i\alpha_i}(\hat{X}_n;\hat{\alpha}_n) = \frac{\mathbb{1}(\hat{\alpha}_{i,n}=\alpha_i)}{\hat{X}_{i\alpha_i,n}} u_i(\hat{\alpha}_{i,n};\hat{\alpha}_{-i,n}) \qquad \text{for all } \alpha_i \in \mathcal{A}_i,\, i \in \mathcal{N}, \tag{IWE}$$

where each player $i \in \mathcal{N}$ draws an action $\hat{\alpha}_{i,n}$ from $\mathcal{A}_i$ according to a mixed strategy $\hat{X}_{i,n} \in \Delta(\mathcal{A}_i)$, cf. [15, 23, 39] and references therein. Then, plugging (IWE) into (EW), we obtain the method known as *exponential weights for exploration and exploitation* (EXP3), viz.

$$\begin{aligned} Y_{i,n+1} &= Y_{i,n} + \gamma_n V_i(\hat{X}_n;\hat{\alpha}_n), \\ X_{i,n+1} &= \Lambda_i(Y_{i,n+1}), \end{aligned} \tag{EXP3}$$

where the sampling strategy of the i-th player at stage n is given by

$$\hat{X}_{i,n} = (1-\delta_n) X_{i,n} + \delta_n \operatorname{unif}(\mathcal{A}_i). \tag{8}$$

In the above, $\delta_n \geq 0$ is an "explicit exploration" parameter that determines the mixing between $X_{i,n}$ and the uniform distribution $\operatorname{unif}(\mathcal{A}_i)$ on $\mathcal{A}_i$. ¶

In closing this section, we note that we can also "mix'n'match" the above algorithms to form new ones: for instance, incorporating the gradient reuse step of (OG) in the simplex setup for (MP) yields the optimistic multiplicative weights (OMW) method of Daskalakis and Panageas [17]. In the sections to come, we provide a synthetic view and analysis of all these policies.

4. Stochastic approximation framework

4.1. Algorithmic template.

The basic blueprint that we will use to analyze the algorithms encountered so far (and more), will be based on the *mirrored Robbins–Monro* template

$$Y_{n+1} = Y_n + \gamma_n \hat{v}_n \qquad X_{n+1} = Q(Y_{n+1}), \tag{MRM}$$

where:

1. $X_n = (X_{i,n})_{i\in\mathcal{N}} \in \mathcal{X}$ denotes the algorithm's state at each stage $n = 1, 2, \dots$
2. $\hat{v}_n = (\hat{v}_{i,n})_{i\in\mathcal{N}} \in \mathcal{Y}$ is a "gradient signal" related to the players' inidividual payoffs.
3. $Y_n = (Y_{i,n})_{i\in\mathcal{N}} \in \mathcal{Y}$ is an auxiliary state that aggregates individual gradient steps.
4. $\gamma_n > 0$ is the method's step-size, for which we will assume throughout that $\sum_n \gamma_n = \infty$ (typically the method is run with $\gamma_n \propto 1/n^p$ for some $p \geq 0$).
5. $Q\colon \mathcal{Y} \to \mathcal{X}$ is the players' so-called "mirror map" (and namesake of the method).

We detail each of these elements below; to streamline our presentation, we also defer to Section 4.3 a systematic account of how Algorithms 1–9 can be recast in the framework of (MRM).

► **The gradient signal.** In the spirit of standard Robbins–Monro schemes, we will decompose the gradient signal $\hat{v}_n$ in (MRM) as

$$\hat{v}_n = v(X_n) + U_n + b_n \tag{9}$$

where

$$U_n = \hat{v}_n - \mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n] \quad \text{and} \quad b_n = \mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n] - v(X_n). \tag{10}$$

By definition, $\mathbb{E}[U_n \,|\, \mathcal{F}_n] = 0$ and b_n is $\mathcal{F}_n$-measurable, so U_n can be intepreted as a random, zero-mean error relative to $v(X_n)$, whereas b_n captures all systematic (non-zero-mean) errors. To make this precise, we will further assume that b_n, U_n and $\hat{v}_n$ are bounded for some $q \geq 2$ as

$$\mathbb{E}[\|b_n\|_* \,|\, \mathcal{F}_n] \leq B_n \qquad \mathbb{E}[\|U_n\|_*^q \,|\, \mathcal{F}_n] \leq \sigma_n^q \qquad \text{and} \qquad \mathbb{E}[\|\hat{v}_n\|_*^q \,|\, \mathcal{F}_n] \leq M_n^q \tag{11}$$

where the sequences B_n, σ_n and M_n, $n = 1, 2, \dots$, are to be construed as deterministic upper bounds on the bias, fluctuations, and magnitude of the gradient signal $\hat{v}_n$ (with $q = \infty$ interpreted as in (5) by convention). Depending on these bounds, a gradient signal with $B_n = 0$ will be called *unbiased*, and an unbiased signal with $\sigma_n = 0$ will be called *perfect*.

We should stress here that the gradient signal $\hat{v}_n$ *does not* play the same role as the gradient oracle (SFO). To see this, consider the unconstrained setting $\mathcal{X} = \mathcal{V}$ with the identity map $Q(y) = y$. In this case, (SGA) can be encoded as an instance of (MRM) by taking $\hat{v}_n = V(X_n; \theta_n)$. Likewise, despite its different update structure, (EG) can *also* be cast as an instance of (MRM) by letting $\hat{v}_n = V(X_{n+1/2}; \theta_{n+1/2})$. In both cases, the oracle V is unbiased as per the discussion surrounding (5) in the previous section; however, even though the gradient signal $\hat{v}_n$ is unbiased in (SGA), it has a non-zero bias in (EG), namely $b_n = \mathbb{E}[v(X_{n+1/2}) \,|\, \mathcal{F}_n] - v(X_n)$.

In view of the above, $\hat{v}_n$ should not be interpreted as an oracle query, but as a "model-agnostic" proxy for $v(X_n)$. In Section 4.3, we show how the methods discussed in Section 3 (including payoff-based ones) can be covered by (MRM) and we record the relevant values of B_n and σ_n in Table 1.

Remark. By Assumption 1 and the inequality $(\sum_{i=1}^m a_i)^q \leq m^{q-1} \sum_{i=1}^m a_i^q$, the decomposition (9) of $\hat{v}_n$ shows that we can always pick $M_n^q = 3^{q-1}(G^q + B_n^q + \sigma_n^q)$ in (11). This makes the last part of (11) redundant, but we will maintain the explicit bound for M_n to simplify the presentation. ¶

► **The players' mirror map.** The second defining element of (MRM) is the "mirror map" $Q_i \colon \mathcal{Y}_i \to \mathcal{X}_i$ of each player – or, in aggregate form, the product map $Q = (Q_i)_{i \in \mathcal{N}} \colon \mathcal{Y} \to \mathcal{X}$. This is defined by means of a "*regularizer*" on $\mathcal{X}$ as follows:[3]

Definition 1. We say that $h_i \colon \mathcal{V}_i \to \mathbb{R} \cup \{\infty\}$ is a *regularizer* on $\mathcal{X}_i$ if:

(1) h_i is *supported* on $\mathcal{X}_i$, i.e., $\operatorname{dom} h_i = \{x_i \in \mathcal{V}_i : h_i(x_i) < \infty\} = \mathcal{X}_i$.

(2) h_i is continuous and *strongly convex* on $\mathcal{X}_i$, i.e., there exists a constant $K_i > 0$ such that

$$h_i(\lambda x_i + (1-\lambda)x_i') \leq \lambda h_i(x_i) + (1-\lambda) h_i(x_i') - \tfrac{1}{2} K_i \lambda (1-\lambda) \|x_i' - x_i\|^2 \tag{12}$$

for all $x_i, x_i' \in \mathcal{X}_i$ and all $\lambda \in [0, 1]$.

The *mirror map* associated to h_i is defined for all $y_i \in \mathcal{Y}_i$ as

$$Q_i(y_i) = \arg\max\nolimits_{x_i \in \mathcal{X}_i} \{\langle y_i, x_i \rangle - h_i(x_i)\} \tag{13}$$

and the image $\mathcal{X}_{h_i} = \operatorname{im} Q_i$ of Q_i is called the *prox-domain* of h_i.

For concision, we will also write $h(x) = \sum_i h_i(x_i)$ for the players' aggregate regularizer and $Q = (Q_i)_{i \in \mathcal{N}}$ for the induced mirror map. We provide three examples of this construction below:

Example 4.1 (Euclidean projection)**.** Consider the quadratic regularizer $h(x) = \|x\|_2^2/2$, $x \in \mathcal{X}$. Then the induced mirror map is the Euclidean projector $Q(y) = \arg\min_{x \in \mathcal{X}} \|y - x\|_2$; as a special case, in unconstrained settings ($\mathcal{X} = \mathcal{V}$), we have $Q(y) = y$ as per Algorithms 1–4 and 7.

Example 4.2 (Entropic regularization and exponential weights)**.** Let $\mathcal{X}_i = \Delta(\mathcal{A}_i)$ for an ensemble of pure strategies $\mathcal{A}_i$, $i \in \mathcal{N}$, and let $h_i(x_i) = \sum_{\alpha_i \in \mathcal{A}_i} x_{i\alpha_i} \log x_{i\alpha_i}$ be the (negative) Gibbs–Shannon entropy on $\mathcal{X}_i$. By standard arguments, the associated mirror map is the logit choice map (7), as per Algorithms 5 and 9 in the previous section.

[3] The authors thank S. Sorin for proposing this definition.

Algorithm	**Actions** ($\mathcal{X}_i$)	**Mirror Map** (Q)	**Feedback**	**Bias** (B_n)	**Magnitude** (M_n)
(SGA)	$\mathbb{R}^{d_i}$	y	oracle	0	$\mathcal{O}(1)$
(seqGA)	$\mathbb{R}^{d_i}$	y	oracle	$\mathcal{O}(1/n^p)$	$\mathcal{O}(1)$
(EG) / (OG)	$\mathbb{R}^{d_i}$	y	oracle	$\mathcal{O}(1/n^p)$	$\mathcal{O}(1)$
(EW)	$\Delta(\mathcal{A}_i)$	$\Lambda(y)$	oracle	0	$\mathcal{O}(1)$
(MP)	general	general	oracle	$\mathcal{O}(1/n^p)$	$\mathcal{O}(1)$
(SPSA)	$\mathbb{R}^{d_i}$	y	payoff	$\mathcal{O}(1/n^r)$	$\mathcal{O}(n^r)$
(DGA)	$[0,\infty)$	$\exp(y)$	payoff	$\mathcal{O}(1/n)$	$\mathcal{O}(1)$
(EXP3)	$\Delta(\mathcal{A}_i)$	$\Lambda(y)$	payoff	$\mathcal{O}(1/n^r)$	$\mathcal{O}(n^r)$

Table 1: The algorithms of Section 3 as instances of (MRM). Where applicable, the methods' step-size and sampling parameters are assumed to be of the form $\gamma_n = \gamma/n^p$ and $\delta_n \propto 1/n^r$ for some $p \in [0,1]$ and $r \in (0,1/2)$ respectively. All oracle-based methods are further assumed to employ an SFO satisfying (5) (with $q=\infty$ for Algorithms 4 and 5).

Example 4.3 (Regularization on the orthant)**.** Let $\mathcal{X}_i = [0,\infty)$ and set $h_i(x_i) = x_i \log x_i - x_i$ for all $x_i \in \mathcal{X}_i$, $i \in \mathcal{N}$. By a straightforward calculation, the induced mirror map is $Q_i(y_i) = \exp(y_i)$. As we discuss in Section 4.3, this provides the setup for the DGA method of [9] (Algorithm 8).[4]

For convenience, we collect the relevant regularizer setups for Algorithms 1–9 in Table 1.

4.2. **Mean dynamics and stochastic approximation.** A key feature of (MRM) is that it can be seen as a "noisy" discretization of the *mean dynamics*

$$\dot{y} = v(x) \qquad x = Q(y). \tag{MD}$$

In this interpretation, $\dot{y}$ represents the continuous-time limit of the finite difference quotient $(Y_{n+1} - Y_n)/\gamma_n$. In particular, if γ_n is "sufficiently small" and the gradient signal $\hat{v}_n$ is a "good enough" approximation of $v(X_n)$, it is plausible to expect that the iterates of (MRM) and the solutions of (MD) will eventually come together.

Following [7, 8], this heuristic can be made precise as follows: First, we define the *(semi)flow* associated to (MD) as the map $\Psi\colon \mathbb{R}\times\mathcal{Y}\to\mathcal{Y}$ which sends an initial condition $y\in\mathcal{Y}$ to the point $\Psi_t(y)\in\mathcal{Y}$ obtained by following the orbit of (MD) starting at y for time $t\in\mathbb{R}$. Then, to compare the sequence of iterates Y_n generated by (MRM) with the solution orbits of (MD), we introduce the *effective time* $\tau_n = \sum_{k=1}^{n} \gamma_k$ and we define the continuous-time affine interpolation $Y(t)$ of Y_n as

$$Y(t) = Y_n + \frac{t-\tau_n}{\tau_{n+1}-\tau_n}(Y_{n+1}-Y_n) \quad \text{for all } t\in[\tau_n,\tau_{n+1}],\ n = 1,2,\dots \tag{14}$$

We then have the following notion of "asymptotic closeness" between (MRM) and (MD):

Definition 2 (Benaïm, 1999)**.** The sequence Y_n is an *asymptotic pseudotrajectory* (APT) of (MD) if

$$\lim\nolimits_{t\to\infty} \sup\nolimits_{0\leq s\leq T} \|Y(t+s) - \Psi_s(Y(t))\| = 0 \quad \text{for all } T>0. \tag{APT}$$

In a slight abuse of terminology, we will also say that the sequence $X_n = Q(Y_n)$ is an APT of (MD).

[4] Strictly speaking, the regularizer $x\log x - x$ is not strongly convex over $\mathbb{R}_+$ but it *is* strongly convex over any bounded subset of $\mathbb{R}_+$ – and it can be made strongly convex over all of $\mathbb{R}_+$ by adding a small quadratic penalty of the form $\varepsilon x^2/2$. This issue does not change the essence of our results, so we sidestep the details.

In words, Definition 2 posits that X_n asymptotically tracks the orbits $x(t) = Q(y(t))$ of (MD) with arbitrary precision over windows of arbitrary length. Of course, this property is quite difficult to verify in general, but the following proposition can be used as an explicit criterion to that effect:

Proposition 3. *Suppose that* (MRM) *is run with a step-size sequence* γ_n *such that*

a) $\lim_{n\to\infty} \gamma_n = 0.$ (15a)

b) $\lim_{n\to\infty} B_n = 0$ *and* $\sum_n \gamma_n^{1+q/2} M_n^q < \infty.$ (15b)

Then the sequence $X_n = Q(Y_n)$ *is an APT of* (MD) *with probability* 1*.*

Proposition 3 is a variant of a basic result by Benaïm [7, Props. 4.1 and 4.2] for processes with $\sup_n M_n < \infty$. The proof proceeds as in Benaïm [7, cf. Eq. (13) and onwards] so we omit it.

4.3. **Examples and applications.** We now return to the online policies presented in Section 3 and illustrate how they can be recast in the framework of (MRM). We begin by presenting the continuous-time dynamics associated to Examples 4.1–4.3:

Example 4.4. Take $\mathcal{X} = \mathcal{V}$ and $Q(y) = y$ as in Example 4.1. Then (MD) gives rise to the (Euclidean) *gradient dynamics*

$$\dot{x} = v(x) \tag{GD}$$

Example 4.5. Let $\mathcal{X}_i = \Delta(\mathcal{A}_i)$ and take $Q_i = \Lambda_i$ as in Example 4.2. Then, by a standard calculation, (MD) boils down to the *replicator dynamics* [61]

$$\dot{x}_{i\alpha_i} = x_{i\alpha_i}[u_i(\alpha_i; x_{-i}) - u_i(x_i; x_{-i})]. \tag{RD}$$

Example 4.6. Let $\mathcal{X}_i = [0, \infty)$ and take $Q_i(y_i) = \exp(y_i)$ as in Example 4.3. In this way, by differentiating $x_i = e^{y_i}$ we obtain the *dampened gradient dynamics* [9]

$$\dot{x}_i = x_i v_i(x). \tag{DGD}$$

We now proceed to illustrate how Algorithms 1–9 can be viewed as instances of (MRM).

► **Algorithm 1: Stochastic gradient ascent.** To recover (SGA), it suffices to take $Q(y) = y$ (cf. Example 4.1) and run (MRM) with gradient signals $\hat{v}_n = V(X_n; \theta_n)$.

► **Algorithm 2: Sequential gradient ascent.** As per Algorithm 1, but with player-specific gradient signals $\hat{v}_{i,n} = V_i(\dots, X_{i-1,n+1}, X_{i,n}, X_{i+1,n}, \dots; \theta_n)$.

► **Algorithm 3: Extra-gradient.** As per Algorithm 1, but with $\hat{v}_n = V(X_{n+1/2}; \theta_{n+1/2})$.

► **Algorithm 4: Optimistic gradient.** As per Algorithm 1, but with $\hat{v}_n = V(X_{n+1/2}; \theta_n)$.

► **Algorithm 5: Exponential weights.** Here, the relevant mirror map is the logit choice map of Example 4.2. The corresponding sequence of oracle signals is then given by (6a) or (6b), depending on the information available to the players.

► **Algorithm 6: Mirror-prox.** The mirror map here is general but, otherwise, the gradient signal is as in Algorithm 3, i.e., $\hat{v}_n = V(X_{n+1/2}; \theta_{n+1/2})$.

► **Algorithm 7: Single-point stochastic approximation.** As per Algorithm 1, but with gradient signals $\hat{v}_{i,n} = (u_i(\hat{X}_n)/\delta_n)\, W_{i,n}$.

► **Algorithm 8: Dampened gradient approximation.** To include (DGA) in the framework of (MRM), take $Q_i(y_i) = \exp(y_i)$ as in Example 4.3. Then, letting $Y_n = \log X_n$, we get

$$Y_{n+1} = Y_n + \log(1 + (u_i(X_{n+1/2}) - u_i(X_n))W_{i,n}). \tag{16}$$

We may therefore view (DGA) as an instance of (MRM) with $\gamma_n = 1/n$ and gradient signals $\hat{v}_{i,n} = n \cdot \log(1 + (u_i(X_{n+1/2}) - u_i(X_n))W_{i,n})$.

► **Algorithm 9: Exponential weights for exploration and exploitation.** As per Algorithm 5, but with $\hat{v}_n$ given by (IWE) and pure strategies $\hat{\alpha}_n$ chosen according to $\hat{X}_n$.

The above justifies the characterization of (MRM) as a common ancestor of the policies discussed in Section 3. On the other hand, it is not clear if the sequence of play generated by each of these methods is, indeed, an APT of the associated dynamics (MD). In this regard, the following proposition provides a range of easily verifiable requirements for Algorithms 1–9:

Proposition 4. *Suppose that Algorithms 1–9 are run with step-size $\gamma_n \propto 1/n^p$, $p \in [0, 1]$, and if applicable, a sampling parameter $\delta_n \propto 1/n^r$, $r \in (0, 1/2)$. Then the corresponding sequence of gradient signals $\hat{v}_n$ in (MRM) enjoys the bounds:*

- *For Algorithms 1 and 5: $B_n = 0$ and $M_n = \mathcal{O}(1)$.*
- *For Algorithms 2–4 and 6: $B_n = \mathcal{O}(1/n^p)$ and $M_n = \mathcal{O}(1)$.*
- *For Algorithms 7 and 9: $B_n = \mathcal{O}(1/n^r)$ and $M_n = \mathcal{O}(n^r)$.*
- *For Algorithm 8: $B_n = \mathcal{O}(1/n)$ and $M_n = \mathcal{O}(1)$.*

Thus, combining Propositions 3 and 4, we obtain the following APT criterion for Algorithms 1–9:

Corollary 1. *Suppose that Algorithms 1–9 are run with parameters as in Proposition 4. Then the sequence $X_n = Q(Y_n)$ comprises an APT of the corresponding instance of (MD) provided that:*

- *For Algorithms 1–4 and 6: $p > 2/(2 + q)$.*
- *For Algorithm 5: $p > 0$.*
- *For Algorithms 7 and 9: $p > 2r > 0$.*

Corollary 1 provides a minimal set of hypotheses under which (MD) is a faithful representation of Algorithms 1–9. For some of the algorithms discussed above, this property is already known in the literature, cf. [7] for (SGA), [46] for (EW), and [9] for (DGA). For others however, the link with (MD) seems to be new: especially in the case of (EG) / (OG), Corollary 1 settles a standing question in the online learning literature concerning the mean dynamics of optimistic gradient methods.

The key element in the proof of Proposition 4 is to bound the bias and moments of the corresponding gradient signal sequence. To streamline the flow of our paper, we defer the relevant calculations to Appendix B and we present an overview of these bounds in Table 1.

5. General convergence analysis

5.1. **The primal-dual interplay.**

In this section, we proceed to derive a series of general convergence results for (MRM) by linking the algorithm's long-run behavior to that of (MD). However, before digging into this analysis, it is important to highlight an important subtlety concerning the interface between the dynamics' driving, dual state $y(t) \in \mathcal{Y}$, and the players' primal, action profile $x(t) = Q(y(t)) \in \mathcal{X}$.

To explain the issue, take a single player with payoff function $u(x) = 1 - x$ for $x \in \mathcal{X} = [0, \infty)$, and consider two different instances of (MD):

(1) *The Euclidean projection dynamics:* This case corresponds to the Euclidean projector $Q(y) = [y]_+$ on $[0, \infty)$, as induced by the quadratic regularizer $h(x) = x^2/2$ (cf. Example 4.1). Since $v(x) = u'(x) = -1$, (MD) gives $\dot{y} = -1$, i.e., $y(t) = y(0) - t$ and $x(t) = [y(0) - t]_+$ for all $t \geq 0$. In particular, this means that $x(t)$ reaches 0 after time $\tau = y(0)$, and subsequently remains there for all time. Since different primal trajectories coalesce at 0 in finite time, the dynamics on $x(t)$ cannot be described by a system of ordinary differential equations on $\mathcal{X}$.

(2) *The dampened gradient dynamics:* This case corresponds to the setup of Algorithm 8, namely the mirror map $Q(y) = \exp(y)$ induced by the entropic regularizer $h(x) = x \log x - x$. As above, the dynamics on $\mathcal{Y}$ are given by $\dot{y} = -1$, so $y(t) = y(0) - t$ for all $t \geq 0$. However, we now have $x(t) = \exp(y(0) - t)$, i.e., $x(t)$ only converges to 0 *asymptotically* as $t \to \infty$.

This qualitative difference in behavior (asymptotic vs. finite-time convergence) is deeply rooted in the boundary behavior of the chosen regularizer. Relegating the detailed statements (and proofs) to Appendix A, we can sum up this dichotomy along the following principal axes:

(1) *Surjective mirror maps:* in this case, $\mathcal{X}_h = \operatorname{im} Q = \mathcal{X}$, cf. Example 4.1. The dynamics' primal orbits $x(t) = Q(y(t))$ may split or coalesce in finite time, so they do not fully capture the state of the system. In particular, the moments at which $x(t)$ enters or exits the boundary of $\mathcal{X}$ cannot always be anticipated by a dynamical system on $\mathcal{X}$.

(2) *Interior-valued mirror maps:* in this case, $\mathcal{X}_h = \operatorname{im} Q = \operatorname{ri} \mathcal{X}$, so Q is not surjective, cf. Examples 4.2 and 4.3. The dynamics' primal orbits remain in $\operatorname{ri} \mathcal{X}$ for all time, and they can be fully described by a well-posed dynamical system on $\mathcal{X}$.

What makes this duality important is that game-theoretic solution concepts reside by necessity in the primal space $\mathcal{X} \subseteq \mathcal{V}$; however, the dynamics (MD) actually evolve on the *dual* space $\mathcal{Y}$. This creates a relatively awkward situation in which *dynamical* notions of equilibrium and stability must be defined on $\mathcal{Y}$, whereas the corresponding *game-theoretic* notions are defined on $\mathcal{X} \subseteq \mathcal{V}$. This difficulty permeates our analysis, and we will return to it several times below.

5.2. **Convergence to internally chain transitive sets.** To analyze the convergence properties of (MD), we will first require some basic definitions from the theory of dynamical systems.

Definition 3. Let $\mathcal{D}$ be a nonempty compact subset of $\mathcal{Y}$, and let $\Psi\colon \mathbb{R} \times \mathcal{Y} \to \mathcal{Y}$ denote the flow map of (MD). Then:

(1) $\mathcal{D}$ is *invariant* under (MD) if $\Psi_t(\mathcal{D}) = \mathcal{D}$ for all $t \in \mathbb{R}$.

(2) $\mathcal{D}$ is an *attractor* for (MD) if it admits a neighborhood $\mathcal{D} \subseteq \mathcal{Y}$ such that $\operatorname{dist}(\Psi_t(y), \mathcal{D}) \to 0$ uniformly in $y \in \mathcal{D}$ as $t \to \infty$.

(3) $\mathcal{D}$ is *internally chain transitive* (ICT) if it is invariant and $\Psi|_{\mathcal{D}}$ has no attractors other than $\mathcal{D}$.

The general theory of Benaïm and Hirsch [8] then yields the following convergence result:

Theorem 1 (Benaïm and Hirsch, 1996)**.** *Let Y_n, $n = 1, 2, \dots,$ be an APT of* (MD) *with* $\sup_n \|Y_n\|_* < \infty$*. Then Y_n converges to an ICT set of* (MD)*.*

Proof. Lemma A.1 in Appendix A shows that Q is Lipschitz continuous. Since v is also Lipschitz continuous by Assumption 1, our assertion follows from Benaïm and Hirsch [8, Theorem 0.1]. ■

Taken together, Corollary 1 and Theorem 1 assert that the behavior of the various algorithms presented in Section 3 (and many more) can be understood by looking at the ICT sets of the *same* mean dynamics. However, from a practical point of view, this comes with two important caveats: First, the boundedness assumption for Y_n cannot be easily validated from the game's primitives, so it is not clear when Theorem 1 applies – and, in much of the literature, this assumption has persisted as a condition that needs to be enforced "by hand" [7, 38]. Second, in certain cases, it is actually *desirable* to have Y_n escape to infinity. As an example, consider the toy problem $u(x) = 1 - x$ with $x \in [0,1]$ and entropic regularization, i.e., $Q(y) = 1/(1+e^{-y})$: in this case, the only way that $X_n = Q(Y_n)$ can converge to the unique solution at 0 is if $\lim_n Y_n = -\infty$, i.e., if Y_n is *unbounded*.

Because of the above, Theorem 1 and the relevant boundedness requirement for Y_n are more relevant for games whose solutions are contained in the *interior* of $\mathcal{X}$. To account for this, we will employ the following "subcoercivity" requirement:

Definition 4. We say that $\mathcal{G}$ is *subcoercive* if there exists a compact set $\mathcal{K} \in \operatorname{ri}\mathcal{X}$ and a reference point $p \in \mathcal{K}$ such that

$$\langle v(x), x - p\rangle \leq 0 \quad \text{for all } x \in \mathcal{X} \setminus \mathcal{K}. \tag{SC}$$

Remark. The term "subcoercivity" alludes to the standard notion of coercivity as defined in convex analysis, viz. $\lim_{\|x\|\to\infty} \langle v(x), x\rangle/\|x\| = \infty$. That said, we should stress that (SC) *does not* impose a superlinear growth requirement for v (hence the "sub"), and also applies to bounded spaces. ¶

Geometrically, subcoercivity simply posits that the Nash field $v(x)$ of the game points weakly towards p outside $\mathcal{K}$, so any "attracting" behavior in $\mathcal{G}$ must be contained in $\mathcal{K}$: for example, it is straightforward to verify that any variationally stable state of $\mathcal{G}$ must lie within $\mathcal{K}$ if (SC) holds. Beyond this, it is important to note that $\mathcal{K}$ can be arbitrarily large (relative to $\mathcal{X}$) and there are no limitations on what type of solutions – or behaviors – may occur *within* $\mathcal{K}$. For concreteness, we provide some important examples of classes of games that satisfy (SC) below.

Example 5.1 (Potential games)**.** Suppose that $\mathcal{G}$ admits a quasiconcave potential function Φ with $\arg\max \Phi \subseteq \operatorname{ri}\mathcal{X}$. If we fix a maximizer p of Φ, we have $\langle v(x), x-p\rangle = \langle \nabla\Phi(x), x-p\rangle \leq 0$ for all $x \in \mathcal{X}$, so $\mathcal{G}$ is subcoercive. More generally, $\mathcal{G}$ is subcoercive if Φ is "eventually quasiconcave", i.e., the upper level sets $L_c^+(\Phi) = \{x \in \mathcal{X} : \Phi(x) \geq c\}$ of Φ are convex for sufficiently small $c > \inf\Phi$ and at least one such set is contained in $\operatorname{ri}\mathcal{X}$.[5] ¶

Example 5.2 (Min-max games)**.** Consider the toy game $\min_{x_1\in[-1,1]} \max_{x_2\in[-1,1]} x_1 x_2$. Since $\langle v(x), x\rangle = -x_2 x_1 + x_1 x_2 = 0$ for all $x \in [-1,1]\times[-1,1]$, the game is trivially subcoercive. More generally, it is easy to check that any two-player, quasi-convex / quasi-concave game with an interior equilibrium is subcoercive. ¶

By itself, subcoercivity ensures that there is no consistent drift pointing away from $\mathcal{K}$, so it is reasonable to expect that Y_n is not repelled to infinity either. To control the inherent stochasticity in Y_n and make this intuition precise, we will require the following summability conditions regarding the bias, variance, and magnitude of the gradient signal process $\hat{v}_n$:

$$\textstyle\sum_n \gamma_n B_n < \infty \qquad \sum_n \gamma_n^2 \sigma_n^2 < \infty \qquad \text{and} \qquad \sum_n \gamma_n^2 M_n^2 < \infty. \tag{Sum}$$

Under these conditions, we have the following stability result:

[5] To see this, let $\mathcal{K} = L_{c_0}^+(\Phi)$ be a convex upper level set of Φ in $\operatorname{ri}\mathcal{X}$. Then, for all $c \leq c_0$ and all x with $\Phi(x) = c$, the segment $x + \tau(p-x)$, $\tau \in [0,1]$, is contained in $L_c^+(\Phi) \supseteq L_{c_0}^+(\Phi)$, so the function $\phi(\tau) = \Phi(x + \tau(p-x))$ cannot have $\phi'(0) < 0$. This implies that $0 \leq \langle \nabla\Phi(x), p-x\rangle = \langle v(x), p-x\rangle$ for all $x \in \mathcal{X}\setminus\mathcal{K}$, i.e., $\mathcal{G}$ is subcoercive.

Proposition 5. *Suppose that* (MRM) *is run with step-size and gradient signal sequences satisfying* (Sum)*. If $\mathcal{G}$ is subcoercive, the sequence of iterates Y_n generated by* (MRM) *is bounded w.p.*1*.*

Before proving Proposition 5, some remarks and corollaries are in order. First, it is important to note that subcoercivity only concerns the primitives of the game under study, and it is otherwise "algorithm-agnostic". In this regard, given the primal-dual nature of the underlying dynamics (MD), Proposition 5 plays a major role in enabling the use of stochastic approximation tools and techniques (otherwise, the boundedness of $\mathcal{X}$ by itself does not suffice).

Second, under Proposition 4, the verification of (Sum) becomes a trivial affair for the example algorithms under study. In particular, a joint application of Corollary 1, Propositions 4 and 5, and Theorem 1 readily yields the general convergence result below:

Theorem 2. *Suppose that Algorithms 1–9 are run with step-size $\gamma_n \propto 1/n^p$, $p \in (1/2, 1]$, and where applicable, a sampling parameter $\delta_n \propto 1/n^r$ such that $1 - p < r < p - 1/2$. If $\mathcal{G}$ is subcoercive, the sequence of play $X_n = Q(Y_n)$ converges to an ICT set of* (MD) *w.p.*1*.*

Corollary 2. *If $\mathcal{G}$ admits a subcoercive potential, X_n converges to a component of critical points of $\mathcal{G}$ w.p.*1*. In particular, if the potential is concave, X_n converges to the set of Nash equilibria of $\mathcal{G}$.*

Corollary 3. *Suppose that $\mathcal{G}$ is a strictly convex-concave min-max game with an interior equilibrium $x^* \in \operatorname{ri} \mathcal{X}$. Then X_n converges to x^* w.p.*1*.*

Corollaries 2 and 3 follow respectively from the fact that the only ICT sets of potential games and strictly convex-concave games are their sets of critical points, see e.g., [7, 46] and references therein. On the other hand, this is not the end of the story, and one should not presume that Theorem 2 can only be used to derive equilibrium convergence results. In Fig. 1, we present an example of an "almost bilinear" 2-player game whose only ICT sets are an unstable critical point at the origin and a "spurious" (but otherwise stable) limit cycle that contains no critical points. In this example, Theorem 2 shows that the range of more sophisticated algorithms (extra-gradient, optimistic gradient, etc.) that have been proposed to overcome the convergence deficiencies of (SGA) in the class of convex-concave games all fail to converge as soon as a modicum of non-convexity is present in the game. By this token, Theorem 2 should not be viewed as a narrow equilibrium convergence criterion, but as a characterization of what types of behaviors may arise in the limit of a game-theoretic learning process – equilibrium and non-equilibrium alike.

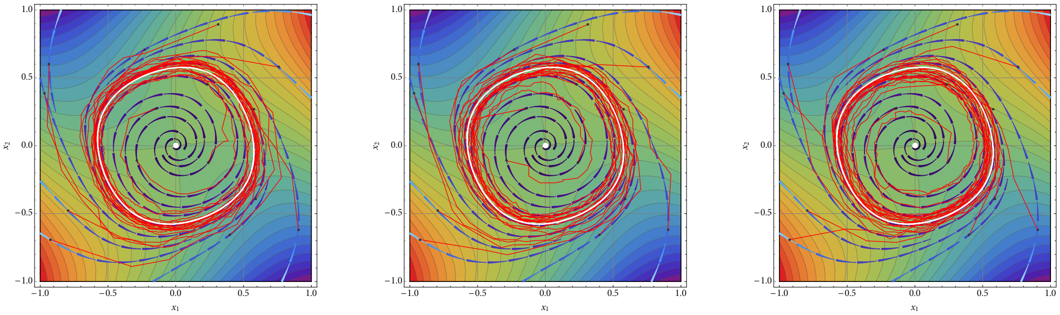
We conclude this section with the proof of our iterate boundedness result.

Proof of Proposition 5. Our proof hinges on the construction of a suitable "energy function" $E\colon \mathcal{Y} \to \mathbb{R}_+$ for (MRM). To define it, we will assume for simplicity – and without loss of generality – that $\mathcal{X}$ has nonempty topological interior in $\mathcal{V}$ (which can be achieved by redefining $\mathcal{V}$ to be the affine hull of $\mathcal{X}$), that the reference point p in Definition 4 is the origin $0 \in \mathcal{V}$, and that $h(p) = 0$ (which can be achieved by a simple translation).

With this in mind, let $h^*(y) = \max_{x \in \mathcal{X}} \{\langle y, x\rangle - h(x)\}$ denote the convex conjugate of h. Then, by Lemma A.2 in Appendix A, we have

$$(K/2)\|Q(y)\|^2 \leq h^*(y) \leq -\min h + \langle y, Q(y)\rangle + (2/K)\|y\|_*^2 \quad \text{for all } y \in \mathcal{Y} \tag{17}$$

where we note that $\min h \leq h(p) = 0$ by assumption. Since h is lower-semicontinuous, we have $h = h^{**}$ by the Fenchel–Moreau theorem. In addition, the Moreau–Rockafellar theorem [6, Theorem 4.17] implies that h^* is coercive because it can be written as $h^*(y) = h^*(y) - \langle y, p\rangle$ and $0 = p \in \operatorname{ri} \mathcal{X} = \operatorname{ri} \operatorname{dom} h^{**}$ by subcoercivity. Finally, since $\mathcal{X}$ has nonempty interior, it



(a) Gradient ascent **(b)** Extra-gradient **(c)** Optimistic gradient

Figure 1: Non-equilibrium behavior in an "almost bilinear" min-max game. The plots show the trajectory of play under (SGA), (EG) and (OG) in a min-max game with loss function $\ell(x_1, x_2) = x_1x_2 + \varepsilon[\phi(x_2) - \phi(x_1)]$ where $\phi(z) = 2z^2 - 4z^4$, $\varepsilon = 2^{-6}$ and $x_1, x_2 \in [-1, 1]$. All methods converge to a spurious limit cycle that contains no critical points. A direct check gives $\langle v(x), x\rangle \leq 0$ outside the quartic $4x_1^4 + 4x_2^4 = x_1^2 + x_2^2$ so ℓ is subcoercive and Theorem 2 applies.

follows that the polar cone $\mathrm{PC}(x)$ is trivially 0 for all $x \in \operatorname{ri}\mathcal{X}$, so the subdifferential ∂h of h is compact-valued on $\mathcal{K} \subseteq \operatorname{ri}\mathcal{X}$. Thus, by the upper hemicontinuity of the subdifferential and the compactness of $\mathcal{K}$, we deduce that the image $\mathcal{D} = \partial h(\mathcal{K})$ of $\mathcal{K}$ under ∂h is compact, cf. [28, p. 201]. Hence, by the coercivity of h^* and the fact that $Q(y) = x$ if and only if $\partial h(x) \ni y$ (cf. Lemma A.1 in Appendix A), there exists some $c > 0$ such that $h^*(y) \leq c$ whenever $Q(y) \in \mathcal{K}$, i.e., $Q^{-1}(\mathcal{K})$ is contained in the c-sublevel set $L_c^-(h^*)$ of h^*.

With all this said and done, fix some $c' > c$ and let

$$E(y) = \varphi(h^*(y)) \quad \text{for all } y \in \mathcal{Y} \tag{18}$$

where $\varphi\colon \mathbb{R}_+ \to \mathbb{R}_+$ is a C^2-smooth "gauge function" with the following properties: *i)* $\varphi(u) = 0$ for $u \leq c$; *ii)* $\varphi(u) = \sqrt{u}$ for $u \geq c'$; *iii)* $\varphi'(u) \geq 0$ and $\varphi''(u) \leq 1$ for all $u \in \mathbb{R}_+$.[6] Then, setting $x = Q(y)$ and differentiating, we readily obtain

$$\nabla E(y) = \varphi'(h^*(y)) \cdot \nabla h^*(y) = \varphi'(h^*(y)) \cdot x \quad \text{for all } y \in \mathcal{Y} \tag{19}$$

and hence, by the smoothness properties of φ and h^*, there exists some constant $C_2 \geq 0$ such that

$$E(y + w) = E(y) + \varphi'(h^*(y)) \cdot \langle w, x\rangle + C_2\|w\|_*^2 \quad \text{for all } y, w \in \mathcal{Y}. \tag{20}$$

Therefore, combining Eqs. (19) and (20) and letting $E_n = E(Y_n)$, we obtain

$$E_{n+1} \leq E_n + \varphi'(h^*(Y_n)) \cdot \langle \hat{v}_n, X_n\rangle + C_2\|\hat{v}_n\|_*^2 \leq E_n + \varphi_n\langle b_n + U_n, X_n\rangle + C_2\|\hat{v}_n\|_*^2 \tag{21}$$

where we set $\varphi_n = \varphi'(h^*(Y_n))$ and we used the fact that $\varphi(h^*(y)) \cdot \langle v(x), x\rangle \leq 0$ for all $y \in \mathcal{Y}$ (the latter being a consequence of subcoercivity and the defining properties of φ). Accordingly, conditioning on $\mathcal{F}_n$ and taking expectations, we finally get

$$\mathbb{E}[E_{n+1} \,|\, \mathcal{F}_n] \leq E_n + \gamma_n\varphi_n B_n\|X_n\| + \gamma_n^2 M_n^2, \tag{22}$$

where we used the Cauchy-Schwarz inequality to bound $\langle b_n, X_n\rangle$ from above by $B_n\|X_n\|$ (recall also that $\mathbb{E}[U_n \,|\, \mathcal{F}_n] = 0$ by definition).

Now, let $\varepsilon_n = \gamma_n\varphi_n B_n\|X_n\| + M_n^2$ denote the "residual" term in (22), and consider the auxiliary process $E_n = E_{n+1} + \sum_{k=n+1}^{\infty} \varepsilon_k$. By (22), we have $\mathbb{E}[E_n \,|\, \mathcal{F}_n] \leq E_n + \sum_{k=n}^{\infty} \varepsilon_n =$

[6] That such a function exists is a straightforward exercise in the construction of aproximate identities, which we omit.

E_{n-1}, i.e., E_n is a supermartingale relative to $\mathcal{F}_n$. Moreover, by (17) and the definition of φ, we further have

$$\varphi_n = \frac{1}{2\sqrt{h^*(Y_n)}} \leq \frac{1}{\sqrt{2K}\|X_n\|} \quad \text{whenever } h^*(Y_n) \geq c' \tag{23}$$

so there exists some (deterministic) positive constant C_1 such that $\sup_n \varphi_n \|X_n\| \leq C_1$. We thus get

$$\sum_{n=1}^{\infty} \varepsilon_n \leq C_1 \sum_{n=1}^{\infty} \gamma_n B_n + C_2 \sum_{n=1}^{\infty} \gamma_n^2 M_n^2 < \infty \tag{24}$$

by the summability condition (Sum). This shows that $\mathbb{E}[\sum_n \varepsilon_n] < \infty$ and, in turn, that $\mathbb{E}[E_n] \leq \mathbb{E}[E_1] < \infty$, i.e., E_n is uniformly bounded in L^1. Accordingly, by Doob's submartingale convergence theorem [25, Theorem 2.5], it follows that E_n converges with probability 1 to some finite random limit E_∞. Since $\sum_n \varepsilon_n < \infty$, this implies that $E_n = E_{n-1} - \sum_{k=n}^{\infty} \varepsilon_n$ also converges to some (random) finite limit (a.s.). Thus, by the coercivity of E, we deduce that $\limsup_n \|X_n\| < \infty$ w.p.1, as claimed. ■

6. Convergence to primal attractors

Theorems 1 and 2 provide a generalist view of the convergence properties of the class of algorithms under study. At the same time however, they do not suffice to answer sharper convergence questions such as what type of sets could be attracting under (MRM) and/or how to reconcile the fact that a boundary equilibrium of a game could attract all primal orbits $x(t) = Q(y(t))$ of (MD) even when the actual solution orbit $y(t)$ of (MD) escapes to infinity. A crucial tool to address these questions will be the notion of a *primal attractor*, which we present and discuss below.

6.1. **Primal attractors.**

By Conley's theorem [16], attractors are characterized by the existence of a local *Lyapunov function*, i.e., a smooth function Φ which is (i) zero on the attractor; (ii) positive everywhere else; and (iii) strictly decreasing along every nearby trajectory that does not belong to the attractor ($\dot{\Phi} < 0$). In our case however, the situation is more complicated because of the primal-dual interplay that we highlighted in Section 5.1: we are interested in the attracting properties of subsets of the *primal space* $\mathcal{X} \subseteq \mathcal{V}$, but the dynamics evolve in the *dual space* $\mathcal{Y} = \mathcal{V}^*$ of $\mathcal{V}$ – and as we already noted, convergence in $\mathcal{X}$ may require escape to infinity in $\mathcal{Y}$.

In view of all this, we introduce the following notions:

Definition 5. A Lipschitz continuous and smooth function $E\colon \mathcal{Y} \to \mathbb{R}$ is a *local energy function* for (MD) if $\sup\{\dot{E}(y) : E_- < E(y) < E_+\} < 0$ for all sufficiently small $E_+ > E_- > \inf E$. Accordingly, a nonempty compact subset $\mathcal{S}$ of $\mathcal{X}$ is said to be a *primal attractor* of (MD) if it admits an energy function E with $\inf E > -\infty$ and such that $Q(y) \to \mathcal{S}$ whenever $E(y) \to \inf E$. In particular, if the above requirements hold for all $E_+ \leq \sup E$, we will refer to E and $\mathcal{S}$ as *global*.

Informally, Definition 5 simply posits that E is a "primal-dual" Lyapunov function for $\mathcal{S}$: it is smooth, positive-definite, and strictly decreasing along any nearby primal orbit $x(t) = Q(y(t))$ that does not lie in $\mathcal{S}$. In this regard, a natural question that arises is whether a primal attractor could be defined instead as the mirror image $Q(\mathcal{D})$ of an ordinary attractor $\mathcal{D} \subseteq \mathcal{Y}$ of (MD). Indeed, if $\mathcal{D}$ is an attractor for (MD), then $Q(\mathcal{D})$ is trivially a primal attractor for (MD); in the converse direction however, primal orbits could converge to a point x^* *outside* the prox-domain $\mathcal{X}_h$ of Q, in which case $Q^{-1}(x^*)$ could be empty – so it wouldn't make sense to talk about the attracting properties of $Q^{-1}(x^*)$ under (MD). To see this, consider again the toy problem $u(x) = 1 - x$ with $x \in [0, 1]$ and the mirror map $Q(y) = 1/(1 + e^{-y})$. In this instance, all primal orbits $x(t) = Q(y(t))$ of (MD) converge to 0, but $Q^{-1}(0)$ is empty so 0 cannot be the mirror image of *any* attractor of (MD).

This simple example highlights several interesting aspects of the problem. First, as we already saw before, having trajectories that escape to infinity may be a desirable property as long as the orbits escape to infinity along the "right" direction. The second is that the notion of an ICT set may not suffice to capture the long-run behavior of (MD) in constrained problems: in the above example, (MD) has no ICT sets but, nonetheless, all primal orbits converge to the game's unique Nash equilibrium. Because of this, stochastic approximation results based on ICT sets become more difficult to apply in games with constrained action spaces (such as mixed extensions of finite games).

6.2. **First steps.** Our convergence analysis will hinge on deriving a suitable "energy inequality" for (MRM). To that end, note first that if E is an energy function for (MD), there exists some $E_* > \inf E$ (possibly equal to ∞) such that the sublevel set $\mathcal{W} = \{y \in \mathcal{Y} : E(y) \leq E_*\}$ is forward invariant under (MD) and $\sup\{\dot{E}(y) : E_* \geq E(y) > E_-\} < 0$ for all $E_- \in (\inf E, E_*)$. Then, if $\mathcal{S}$ is a primal attractor with energy function E, we will refer to $\mathcal{B} = Q(\mathcal{W})$ as a *fundamental neighborhood* of $\mathcal{S}$. Finally, we note that there exist positive constants $\beta, H > 0$ such that $\|\nabla E(y)\| \leq H$ and

$$E(y') \leq E(y) + \langle \nabla E(y), y' - y \rangle + \tfrac{1}{2}\beta \|y' - y\|_*^2 \tag{25}$$

for all $y, y' \in \mathcal{Y}$. We then have the following template inequality:

Lemma 1. *Let $E_n \coloneqq E(Y_n)$. Then, for all $n = 1, 2, \dots,$ we have*

$$E_{n+1} \leq E_n + \gamma_n \langle v(X_n), \nabla E(Y_n) \rangle + \gamma_n \xi_n + \gamma_n \chi_n + \gamma_n^2 \psi_n^2, \tag{26}$$

where the error terms ξ_n, χ_n, and ψ_n are given by

$$\xi_n = \langle U_n, \nabla E(Y_n) \rangle, \quad \chi_n = H B_n \quad \textit{and} \quad \psi_n^2 = \tfrac{1}{2}\beta \|\hat{v}_n\|_*^2. \tag{27}$$

Proof. Simply unroll (25) after substituting $y \leftarrow Y_n$ and $y' \leftarrow Y_{n+1} = Y_n + \gamma_n \hat{v}_n$ with $\hat{v}_n$ as in (9). ■

Now, by the definition of E, we have $\dot{E}(y) = \langle v(Q(y)), \nabla E(y) \rangle < 0$ whenever $y \in \mathcal{W} \setminus Q^{-1}(\mathcal{S})$. Hence, for $X_n \in \mathcal{B}$, (26) becomes

$$E_{n+1} \leq E_n + \gamma_n \xi_n + \gamma_n \chi_n + \gamma_n^2 \psi_n^2. \tag{28}$$

Of course, each of these error terms can be positive, so E_n may fail to be decreasing, even when $X_n \in \mathcal{B}$. On that account, it will be convenient to introduce the error processes

$$\mathrm{I}_n = \textstyle\sum_{k=1}^n \gamma_k \xi_k \qquad \mathrm{II}_n = \sum_{k=1}^n \gamma_k \chi_k \qquad \text{and} \qquad \mathrm{III}_n = \sum_{k=1}^n \gamma_k^2 \psi_k^2 \tag{29}$$

which measure directly the aggregate effect of each error term in (26). As it turns out, under (Sum), these errors can all be compensated by the negative drift of (26), leading to the following global convergence result:

Proposition 6. *Let $\mathcal{S}$ be a global attractor of* (MD)*, and let $X_n = Q(Y_n)$ be the sequence of play generated by* (MRM)*. If* (Sum) *holds, then X_n converges to $\mathcal{S}$ with probability* 1.

To streamline our discussion, we defer the proof of Proposition 6 to the end of this section, and we focus instead on deriving a similar convergence result for non-global attractors. In this case, even if the algorithm begins play very close to $\mathcal{S}$, a single "bad" realization of the noise could force the process to exit the basin of attraction of $\mathcal{S}$, possibly never to return.

Thus, to control the probability with which this event occurs, we will introduce the random variables

$$\mathrm{I}_\infty = \sup_n \mathrm{I}_n \qquad \mathrm{II}_\infty = \sup_n \mathrm{II}_n \qquad \text{and} \qquad \mathrm{III}_\infty = \sup_n \mathrm{III}_n \tag{30}$$

which can be seen as a "worst-case" measure of the aggregate error entering (26). Then, given an error tolerance $\varepsilon > 0$ and a confidence level $\rho > 0$, we will consider the stability requirement

$$\mathbb{P}(\mathrm{I}_\infty > \varepsilon) \leq \rho \quad (\text{Stb.I}) \qquad \mathbb{P}(\mathrm{II}_\infty > \varepsilon) \leq \rho \quad (\text{Stb.II}) \qquad \mathbb{P}(\mathrm{III}_\infty > \varepsilon) \leq \rho \quad (\text{Stb.III})$$

which, in turn, leads to the following local analogue of Proposition 6:

Proposition 7. *Let $\mathcal{S}$ be a primal attractor of* (MD)*, fix some confidence index $\rho > 0$, and let $X_n = Q(Y_n)$ be the sequence of play generated by* (MRM)*. Assume further that the algorithm begins play at a neighborhood $\mathcal{U}$ of $\mathcal{S}$ such that $E(Y_1) \leq E_*/4 =: \varepsilon$. If* (Sum) *and* (Stb) *hold, then*

$$\mathbb{P}(X_n \text{ converges to } \mathcal{S} \mid X_1 \in \mathcal{U}) \geq 1 - 3\rho. \tag{32}$$

As with its global counterpart, we defer the proof of Proposition 7 to the end of this section. For now, we simply note that Propositions 6 and 7 can be difficult to employ in practice because of their reliance on the conditions (Sum) and (Stb). In the next section, we provide a range of explicit parameter schedules that can be used to verify these implicit requirements directly.

6.3. **Main results, implications, and applications.** To obtain an explicit version of Propositions 6 and 7, we will assume in the sequel that (MRM) is run with step-size and gradient signal sequences such that

$$\gamma_n = \gamma/n^p \qquad B_n = \mathcal{O}(1/n^b) \qquad \text{and} \qquad M_n = \mathcal{O}(n^s) \tag{33}$$

for some $p \in [0,1]$, $b > 0$ and $s < 1/2$. Since the schedule (33) involves B_n and M_n (which, depending on the algorithm, may be beyond the players' control), this requirement may also seem unverifiable at first glance. However, in view of Proposition 4, the exponents b and s can be directly expressed in terms of the parameters of the specific algorithm under study, so this is not an issue.

With all this in mind, our main result in this section is as follows:

Theorem 3. *Let $\mathcal{S}$ be a primal attractor of* (MD)*, and let X_n be the sequence of play of* (MRM) *with step-size and gradient signal sequences such that $p + b > 1$ and $p - s > 1/2$ in* (33)*. Then:*

Case 1: *If $\mathcal{S}$ is global, X_n converges to $\mathcal{S}$ with probability 1.*

Case 2: *If $\mathcal{S}$ is local, there exists a neighborhood $\mathcal{U}$ of $\mathcal{S}$ such that, for any given $\rho > 0$, we have*

$$\mathbb{P}(X_n \text{ converges to } \mathcal{S} \mid X_1 \in \mathcal{U}) \geq 1 - \rho \tag{34}$$

provided that $\gamma > 0$ is small enough relative to ρ.

Corollary 4. *Suppose that Algorithms 1–9 are run with step-size $\gamma_n = \gamma/n^p$, $p \in (1/2, 1]$, and where applicable, a sampling parameter $\delta_n = \delta/n^r$ such that $1 - p < r < p - 1/2$. Then the conclusions of Theorem 3 hold.*

Theorem 3 and Corollary 4 are our primary results concerning primal attractors so, before proving them, we proceed to show how they apply to the range of solution concepts – and classes of games – discussed in Section 2.

► **Variational stability, global.** A key property of globally variationally stable states is that they are global attractors of (MD). To state this fact formally, we will need an additional regularity assumption for h, namely that

$$h(x_n) + \langle y_n, x - x_n\rangle \to h(x) \quad \text{whenever} \quad x_n \to x \tag{R}$$

for all $x \in \mathcal{X}$ and every sequence of subgradients $y_n \in \partial h(x_n)$. This condition simply posits that the first-order approximation of $h(x)$ from $h(x_n)$ is always accurate in the limit $x_n \to x$, a property which is satisfied by all examples of regularizers that we have considered so far; for a more detailed discussion, cf. Censor and Lent [13], Chen and Teboulle [14] and references therein.

With this caveat in mind, we have:

Proposition 8. *Suppose that $x^* \in \mathcal{X}$ is globally variationally stable. If the players' regularizers satisfy* (R)*, x^* is a global attractor.*

Proof. Let $F(y) = h(x^*) + h^*(y) - \langle y, x^*\rangle$ where $h^*(y) = \max_{x\in\mathcal{X}}\{\langle y, x\rangle - h(x)\}$ denotes the convex conjugate of h. By the Fenchel–Young inequality, we have $F(y) \geq 0$ for all $y \in \mathcal{Y}$. Moreover, by the definition (13) of Q, we have $F(y_n) = h(x^*) + h^*(y_n) - \langle y_n, x^*\rangle = h(x^*) - h(x_n) + \langle y_n, x_n - x^*\rangle$ for every sequence $y_n \in \mathcal{Y}$, $x_n = Q(y_n)$, so (R) yields $F(y_n) \to 0$ if and only if $x_n = Q(y_n) \to x^*$. Finally, by Lemma A.1 in Appendix A, we have $\nabla h^*(y) = Q(y)$ so (MD) gives

$$\dot{F}(y) = \langle \dot{y}, \nabla F(y)\rangle = \langle v(Q(y)), \nabla h^*(y) - x^*\rangle = \langle v(x), x - x^*\rangle < 0 \tag{35}$$

whenever $x = Q(y) \neq x^*$. Thus, putting everything together, we conclude that $\dot{F}(y) \to 0$ if and only if $F(y) \to 0$, which implies in turn that $\sup\{\dot{F}(y) : F(y) > F_-\} < 0$ for all $F_- > 0$.

Consider now the gauge function $\varphi\colon \mathbb{R}_+ \to \mathbb{R}_+$ with $\varphi(u) = u$ if $u \in [0,1]$ and $\varphi(u) = 2\sqrt{u} - 1$ for $u \geq 1$, so $\varphi'(u) \leq 1/\sqrt{u}$ for all $u > 0$. Then, setting $E(y) = \varphi(F(y))$, we readily get $\nabla E(y) = \varphi'(F(y))\nabla F(y)$ so (35) yields $\dot{E}(y) = \langle \dot{y}, \nabla E(y)\rangle = \varphi'(F(y))\langle v(x), x - x^*\rangle < 0$ whenever $x = Q(y) \neq x^*$. Furthermore, by Lemma A.2 we have $F(y) \geq (K/2)\|Q(y) - x^*\|^2$, and hence

$$\|\nabla E(y)\|_* = \varphi'(F(y))\|\nabla F(y)\| \leq \frac{\|\nabla F(y)\|}{\sqrt{F(y)}} \leq \frac{\sqrt{2/K}\|Q(y) - x^*\|}{\sqrt{\|Q(y) - x^*\|^2}} \leq \sqrt{2/K} \tag{36}$$

so we can take $H = \sqrt{2/K}$. Finally, again by Lemma A.2 in Appendix A, $F(y)$ is $(1/K)$-Lipschitz smooth, so Eq. (25) follows immediately from the concavity of φ. This shows that E is a global energy function for x^*, i.e., x^* is a global attractor of (MD). ■

Armed with this observation, we immediately obtain the following corollaries of Theorem 3:

Corollary 5. *Suppose that Algorithms 1–9 are run with parameters as in Corollary 4 – or, more generally, that* (MRM) *is run with parameters as in Theorem 3. If x^* is globally variationally stable, then X_n converges to x^* with probability* 1.

Corollary 6. *Suppose that Algorithms 1–9 are run with parameters as in Corollary 4 – or, more generally, that* (MRM) *is run with parameters as in Theorem 3. If $\mathcal{G}$ is strictly monotone, then X_n converges to the game's unique Nash equilibrium with probability* 1.

We should stress here that neither of the above results can be inferred by the ICT convergence results of Section 5. In particular, if x^* lies at the boundary of $\mathcal{X}$, it might fail to be accessible unless the dual process Y_n escapes to infinity, in which case Theorems 1 and 2 no longer apply. This illustrates the flexibility of the concept of a primal attractor, as it allows us to tackle at the same time both boundary and interior solutions.

To the best of our knowledge, the only comparable result in the literature for oracle-based methods concerns the convergence of the standard mirror descent algorithm ($B_n = 0$, $\sup_n \sigma_n^2 < \infty$) in strictly monotone games with compact domains [46]. For payoff-based learning, the closest results we are aware of are by Bravo et al [11] and Tatarenko and Kamgarpour [64] for a constrained variant of (SPSA) in strictly monotone games with compact domains (the latter actually showing convergence in probability, but without requiring strict monotonicity).

► **Variational stability, local.** Now, if $x^* \in \mathcal{X}$ is variationally stable but not globally so, we have the following local analogue of Proposition 8:

Proposition 9. *Suppose that $x^* \in \mathcal{X}$ is variationally stable. If the players' regularizers satisfy* (R)*, x^* is a primal attractor of* (MD)*.*

Proof. Let $F(y)$ and $E(y)$ be defined as in the proof of Proposition 8, and let $\mathcal{K}$ be a compact neighborhood of x^* in $\mathcal{X}$ such that $\langle v(x), x - x^* \rangle < 0$ for all $x \in \mathcal{K}$. Then, under (R), Lemma A.2 in Appendix A shows that there exists some $E_* > 0 = \inf E$ such that $Q(y) \in \mathcal{K}$ whenever $E(y) \leq E_*$. In turn, this shows that E is a local energy function for x^*, so our assertion follows. ■

As a consequence of the above, Theorem 3 readily yields the following local convergence results (which, again, cannot be inferred by the ICT convergence analysis of Section 5):

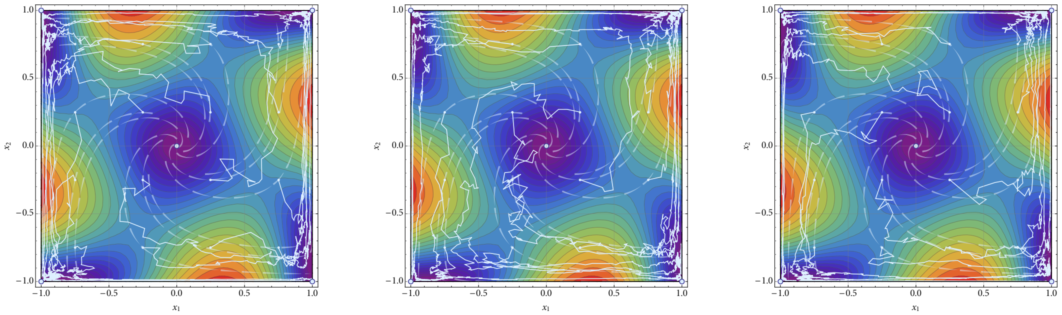
Corollary 7. *Suppose that Algorithms 1–9 are initialized and run as per Corollary 4 – or, more generally, that* (MRM) *is initialized and run as per Theorem 3. If x^* satisfies the second-order sufficient condition* (SOS)*, then X_n converges locally to x^* with arbitrarily high probability.*

Corollary 8. *Let x^* be a strict Nash equilibrium of a finite game. If Algorithms 5 and 9 are initialized and run as per Corollary 4, X_n converges locally to x^* with arbitrarily high probability.*

In a certain precise sense, local Nash equilibria satisfying (SOS) are the game-theoretic analogue of minimizers with a positive-definite Hessian in non-convex minimization problems [55, 56]. In this regard, Corollary 7 is particularly important as it shows that such equilibria are stable and attracting under the entire class of algorithms under study. Likewise, Corollary 8 is a key special case of this implication because, generically – i.e., except on a set of games which is meager in the sense of the Baire category theorem – pure Nash equilibria in finite games are always strict. Thus, coupled with the inherent instability of mixed equilibria in finite games [21], Corollary 8 goes a long way towards establishing a learning analogue of the "folk theorem" of evolutionary game theory [30] which states that a mixed strategy profile is stable and attracting if and only if it is a strict Nash equilibrium.

► **Discoordination games.** As a last example, consider a two-player discoordination game with payoff functions $u_1(x_1, x_2) = (x_1 - x_2)^2/2$ and $u_2(x_1, x_2) = (x_1 + x_2)^2/2$ for $x_1, x_2 \in [-1, 1]$. This game admits five critical points, the origin $(0, 0)$ and the four vertices $\{\pm 1, \pm 1\}$ of $\mathcal{X} = [-1, 1]^2$. None of these critical points is an equilibrium: the origin is unstable to deviations by both players, whereas the vertices are unstable to deviations by one of the players (but not the other). Given the lack of an equilibrium in pure strategies (a standard feature of discoordination games), the players' limiting behavior is quite difficult to predict; however, since the critical point at $(0, 0)$ is unstable for both players, it is reasonable to expect that it should be selected against.

To examine this issue in the context of (MRM), consider for concreteness the mirror map $Q_i(y_i) = \tanh(y_i/2)$ that is induced by the entropic regularizer $h_i(x_i) = (1 -$



(a) Gradient ascent **(b)** Extra-gradient **(c)** Optimistic gradient

Figure 2: Learning in a 2-player discoordination game. All algorithms under study converge to the boundary of the game's domain, which contains all critical points that are resilient to deviations by one of the players (but not the other). The interior critical point at $(0, 0)$ is unstable to deviations by *both* players, and no trajectories converge there, even though it is the only ICT of (MD).

$x_i) \log x_i + (1 + x_i) \log(1 + x_i)$. In this case, it is straightforward to check that $E(y_1, y_2) = 2 \operatorname{sech}(y_1/2) \operatorname{sech}(y_2/2)$ is an (almost global) energy function for the boundary $\mathcal{S} = \operatorname{bd}(\mathcal{X})$ of the game's domain; hence, by Theorem 3, the induced sequence of play will converge to $\mathcal{S}$ (cf. Fig. 2). On the other hand, since all solution orbits $y(t)$ of (MD) escape to infinity unless they start at the origin, the only ICT set of (MD) is the origin itself. On that account, a casual reading of the ICT convergence results of Section 5 (viz. one that ignores the boundedness caveat) would lead to a prediction that is diametrically opposed to what actually happens in the long run – i.e., that X_n converges to the boundary of $\mathcal{X}$ and not to $(0, 0)$.

6.4. **Technical proofs.** We conclude this section with the proofs of Propositions 6 and 7 and Theorem 3. We begin with a technical lemma showing that the aggregate error processes $\mathrm{I}_n, \mathrm{II}_n$ and III_n of (29) are subleading relative to the long-run drift of (26).

Lemma 2. *Under* (Sum)*, the aggregate error processes of* (29) *are sublinear in* τ_n*, i.e.,*

$$\mathrm{I}_n/\tau_n \to 0 \quad \text{(Sub.I)} \qquad \mathrm{II}_n/\tau_n \to 0 \quad \text{(Sub.II)} \qquad \mathrm{III}_n/\tau_n \to 0 \quad \text{(Sub.III)}$$

with all limits interpreted in the almost sure sense.

Proof. We treat each term separately. For (Sub.I), we have

$$\sum_{n=1}^{\infty} \mathbb{E}[\gamma_n^2 \xi_n^2 \,|\, \mathcal{F}_n] \leq \sum_{n=1}^{\infty} \gamma_n^2 \,\mathbb{E}[\|\nabla E(Y_n)\|^2 \|U_n\|_*^2 \,|\, \mathcal{F}_n] \leq H^2 \sum_{n=1}^{\infty} \gamma_n^2 \sigma_n^2 < \infty \tag{38}$$

by (Sum). Thus, by the strong law of large numbers for martingale difference sequences [25, Theorem 2.18], we conclude that I_n/τ_n converges to 0 with probability 1. For (Sub.II), the conclusion is immediate by the fact that $\sum_n \gamma_n B_n < \infty$ under (Sum). Finally, for the submartingale term of (Sub.III), we have

$$\mathbb{E}[\mathrm{III}_n] = \sum_{k=1}^{n} \gamma_k^2 \,\mathbb{E}[\psi_k^2] \leq \tfrac{1}{2}\beta \sum_{k=1}^{n} \gamma_k^2 \,\mathbb{E}[\|\hat{v}_k\|_*^2], \leq \tfrac{1}{2}\beta \sum_{k=1}^{n} \gamma_k^2 M_k^2, \tag{39}$$

so, by (Sum), it follows that III_n is bounded in L^1. Therefore, by Doob's submartingale convergence theorem [25, Theorem 2.5], we further deduce that III_n converges (a.s.) to some (finite) random variable III_∞, implying in turn that $\mathrm{III}_n/\tau_n \to 0$ with probability 1. ∎

Moving forward, we present two lemmas that will allow us to deduce the convergence of the energy iterates $E_n \coloneqq E(Y_n)$ if a certain favorable event occurs. To state them, recall by the discussion in the beginning of Section 6.2 that there exists some $E_* > \inf E$ such that the sublevel set $\mathcal{W} = \{y \in \mathcal{Y} : E(y) \leq E_*\}$ is forward invariant under (MD) and $\sup\{\dot{E}(y) : E_* \geq E(y) > E_-\} < 0$ for all $E_- \in (\inf E, E_*)$. Then, letting

$$\mathcal{E} = \{Y_n \in \mathcal{W} \text{ for all } n\} \tag{40}$$

denote the favorable event in question, we have the following series of implications.

Lemma 3. *Suppose that* $\mathbb{P}(\mathcal{E}) > 0$. *If* (Sub) *holds, then* $\mathbb{P}(\liminf_{n\to\infty} E_n = \inf E \mid \mathcal{E}) = 1$.

Lemma 4. *Suppose that* $\mathbb{P}(\mathcal{E}) > 0$. *If* (Sum) *holds and* $\inf E > -\infty$, *there exists some finite random variable* E_∞ *such that* $\mathbb{P}(\lim_{n\to\infty} E_n = E_\infty \mid \mathcal{E}) = 1$.

Proposition 10. *Let* $\mathcal{S}$ *be a primal attractor of* (MD) *with energy function* E. *If* $\mathbb{P}(\mathcal{E}) > 0$ *and* (Sum) *holds, then* $\mathbb{P}(X_n$ *converges to* $\mathcal{S} \mid \mathcal{E}) = 1$.

Proof of Lemma 3. Since $\mathbb{P}(\mathcal{E}) > 0$, it suffices to show that the hitting time $N_a = \inf\{n \in \mathbb{N} : E_n \leq a\}$ is finite with probability 1 on $\mathcal{E}$ for all sufficiently small $a > \inf E$. More precisely, building on a technique of Duvocelle et al [20], we will show that the event $\mathcal{N}_a = \mathcal{E} \cap \{N_a = \infty\}$ has $\mathbb{P}(\mathcal{N}_a) = 0$ whenever $\inf E < a \leq E_*$: indeed, if this is the case and $a_k \in (\inf E, E_*)$, $k = 1, 2, \dots$, is a sequence converging monotonically to $\inf E$, we will have $\mathbb{P}(\mathcal{N}_{a_k}) = 0$ for all $k \in \mathbb{N}$. Thus, with only a countable number of $\mathcal{N}_{a_k}$ in play, we will have

$$\begin{aligned} \mathbb{P}(\liminf\nolimits_{n\to\infty} E_n = \inf E \mid \mathcal{E}) &= \mathbb{P}(N_{a_k} < \infty \text{ for all } k \mid \mathcal{E}) \\ &= \mathbb{P}(\textstyle\bigcap_{k=1}^\infty \{N_{a_k} < \infty\} \mid \mathcal{E}) = 1 - \mathbb{P}(\textstyle\bigcup_{k=1}^\infty \{N_{a_k} < \infty\} \mid \mathcal{E}) \\ &= 1 - \frac{\mathbb{P}(\mathcal{E} \cap (\bigcup_{k=1}^\infty \{N_{a_k} < \infty\}))}{\mathbb{P}(\mathcal{E})} = 1 - \frac{\mathbb{P}(\bigcup_{k=1}^\infty \mathcal{N}_{a_k})}{\mathbb{P}(\mathcal{E})} = 1, \end{aligned} \tag{41}$$

as per our original assertion.

Now, to establish our claim for $\mathcal{N}_a$, assume to the contrary that $\mathbb{P}(\mathcal{N}_a) > 0$ for some sufficiently small $a > \inf E$, and let $c_a = -\sup\{\dot{E}(y) : a \leq E(y) \leq E_*\}$, so $c_a > 0$ by Definition 5. Then, by telescoping (26), we get

$$\begin{aligned} E_{n+1} &\leq E_1 + \sum_{k=1}^n \gamma_k \dot{E}(Y_k) + \sum_{k=1}^n \gamma_k \xi_k + \sum_{k=1}^n \gamma_k \chi_k + \sum_{k=1}^n \gamma_k \psi_k^2 \\ &\leq E_1 - \left[c_a - \frac{\mathrm{I}_n + \mathrm{II}_n + \mathrm{III}_n}{\tau_n}\right] \cdot \tau_n \qquad \text{for all } n = 1, 2, \dots \end{aligned} \tag{42}$$

with probability 1 on $\mathcal{N}_a$. Since $\mathbb{P}(\mathcal{N}_a) > 0$ by assumption and $(\mathrm{I}_n + \mathrm{II}_n + \mathrm{III}_n)/\tau_n \to 0$ with probability 1 by (Sub), the above gives $\mathbb{P}(\lim_{n\to\infty} E_n = -\infty \mid \mathcal{N}_a) = 1$. However, with $\inf_n E_n \geq a > -\infty$ on $\mathcal{N}_a$ by construction, we get a contradiction, and our proof is complete. ■

Proof of Lemma 4. Consider the nested sequence of events

$$\mathcal{E}_n = \{\dot{E}(Y_k) \leq 0 \text{ for all } k = 1, 2, \dots, n\} \tag{43}$$

so $\mathcal{E} = \bigcap_{n=1}^\infty \mathcal{E}_n$. Then, letting $\tilde{E}_n = \mathbb{1}_{\mathcal{E}_n}(E_n - \inf E)$, Eq. (26) readily gives

$$\begin{aligned} \tilde{E}_{n+1} = \mathbb{1}_{\mathcal{E}_{n+1}}(E_{n+1} - \inf E) &\leq \mathbb{1}_{\mathcal{E}_n}(E_{n+1} - \inf E) \\ &\leq \mathbb{1}_{\mathcal{E}_n}(E_n - \inf E) + \left(\gamma_n \dot{E}(Y_n) + \gamma_n \xi_n + \gamma_n \chi_n + \gamma_n^2 \psi_n^2\right) \mathbb{1}_{\mathcal{E}_n} \\ &\leq \tilde{E}_n + \gamma_n \mathbb{1}_{\mathcal{E}_n} \xi_n + \left(\gamma_n \chi_n + \gamma_n^2 \psi_n^2\right) \mathbb{1}_{\mathcal{E}_n}, \end{aligned} \tag{44}$$

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

$$\begin{aligned}\mathbb{E}[\tilde{E}_{n+1} \,|\, \mathcal{F}_n] &\leq \tilde{E}_n + \gamma_n \mathbb{1}_{\mathcal{E}_n} \mathbb{E}[\xi_n \,|\, \mathcal{F}_n] + \mathbb{1}_{\mathcal{E}_n} \mathbb{E}[\gamma_n \chi_n + \gamma_n^2 \psi_n^2 \,|\, \mathcal{F}_n] \\ &\leq \tilde{E}_n + \mathbb{E}[\gamma_n \chi_n + \gamma_n^2 \psi_n^2 \,|\, \mathcal{F}_n] \\ &\leq \tilde{E}_n + \gamma_n H B_n + \tfrac{1}{2}\beta \gamma_n^2 M_n^2. \end{aligned} \tag{45}$$

Now, given that $\sum_n \gamma_n B_n$ and $\sum_n \gamma_n^2 M_n^2$ are both finite by (Sum), $\tilde{E}_n$ is an almost supermartingale with summable increments, i.e., $\sum_n \Big[\mathbb{E}[\tilde{E}_{n+1} \,|\, \mathcal{F}_n] - \tilde{E}_n\Big] < \infty$ w.p.1. Therefore, by Gladyshev's lemma [52, p. 49], we conclude that $\tilde{E}_n$ converges almost surely to some (finite) random variable. Since $\mathbb{P}(\mathcal{E}) > 0$ and $\mathbb{1}_{\mathcal{E}_n} = 1$ for all n if and only if $\mathcal{E}$ occurs, we further deduce that $\mathbb{P}(E_n \text{ converges} \,|\, \mathcal{E}) = \mathbb{P}(\tilde{E}_n \text{ converges} \,|\, \mathcal{E}) = 1$, and our claim follows. ■

Proof of Proposition 10. By Lemma 2, (Sub) is satisfied whenever (Sum) is. Thus, by a tandem application of Lemmas 3 and 4, we conclude that $\lim_{n\to\infty} E_n = \inf E$ with probability 1 on $\mathcal{E}$. Since $Q(y) \to \mathcal{S}$ whenever $E(y) \to \inf E$, our claim follows. ■

We are now in a position to prove our main results, beginning with Propositions 6 and 7.

Proof of Proposition 6. By the definition of a global attractor, we have $E_* = \sup E$, so $\mathbb{P}(\mathcal{E}) = 1$. Our claim is then an immediate consequence of Proposition 10. ■

Proof of Proposition 7. Suppose that $E(Y_1) \leq E_*/4$. We then claim that the event $\mathcal{E}$ always occurs on the intersection of the events $\mathcal{E}_{\mathrm{I}}$, $\mathcal{E}_{\mathrm{II}}$, and $\mathcal{E}_{\mathrm{III}}$, where $\mathcal{E}_Z = \{Z_n \leq E_*/4 \text{ for all } n\}$. Indeed, this being trivially the case for $n = 1$, assume that $Y_k \in \mathcal{W}$ for all $k = 1, 2, \dots, n$ for some $n \geq 1$. Then, telescoping (26) yields

$$E_{n+1} \leq E_1 + \sum_{k=1}^n \gamma_k \dot{E}(Y_k) + \mathrm{I}_n + \mathrm{II}_n + \mathrm{III}_n \leq E_*/4 + 0 + E_*/4 + E_*/4 + E_*/4 = E_* \tag{46}$$

by the inductive hypothesis and our other assumptions. This shows that $E_{n+1} \in \mathcal{W}$, so the induction argument is complete, and we conclude that $\mathcal{E} \supseteq \mathcal{E}_{\mathrm{I}} \cap \mathcal{E}_{\mathrm{II}} \cap \mathcal{E}_{\mathrm{III}}$. Now, by (Stb), we have $\mathbb{P}(\mathcal{E}_{\mathrm{I}}) \leq \rho$ and likewise for the rest, so we get

$$\mathbb{P}(\mathcal{E}) \geq \mathbb{P}(\mathcal{E}_{\mathrm{I}} \cap \mathcal{E}_{\mathrm{II}} \cap \mathcal{E}_{\mathrm{III}}) = 1 - \mathbb{P}(\mathcal{E}_{\mathrm{I}} \cup \mathcal{E}_{\mathrm{II}} \cup \mathcal{E}_{\mathrm{III}}) \geq 1 - \mathbb{P}(\mathcal{E}_{\mathrm{I}}) - \mathbb{P}(\mathcal{E}_{\mathrm{II}}) - \mathbb{P}(\mathcal{E}_{\mathrm{III}}) \geq 1 - 3\rho. \tag{47}$$

Our claim then follows directly from Proposition 10. ■

With all this in hand, the proof of Theorem 3 proceeds as follows.

Proof of Theorem 3. We begin by noting that (Sum) holds trivially under the stated conditions for $\gamma_n \propto 1/n^p$, $B_n = \mathcal{O}(1/n^b)$ and $M_n = \mathcal{O}(n^s)$. As a result, the first part of the theorem follows immediately from Proposition 6.

Likewise, for the second part, it will suffice to establish the stability condition (Stb). To that end, consider the "maximal" processes

$$\mathrm{I}_n^* = \max_{1\leq k\leq n} \mathrm{I}_k \qquad \mathrm{II}_n^* = \max_{1\leq k\leq n} \mathrm{II}_k \qquad \text{and} \qquad \mathrm{III}_n^* = \max_{1\leq k\leq n} \mathrm{III}_k, \tag{48}$$

so $\mathrm{I}_\infty = \lim_{n\to\infty} \mathrm{I}^*$ (and likewise for the rest). Then, proceeding term-by-term, we have:

(1) Since I_n is a martingale, Kolmogorov's inequality [25, Corollary 2.1] gives

$$\mathbb{P}(\mathrm{I}_n^* \geq \varepsilon) \leq \mathbb{P}\left(\max_{1\leq k\leq n} |\mathrm{I}_k| \geq \varepsilon\right) \leq \frac{\mathbb{E}[\mathrm{I}_n^2]}{\varepsilon^2} = \frac{\mathbb{E}\big[\big(\sum_{k=1}^n \gamma_k \xi_k\big)^2\big]}{\varepsilon^2} \leq \frac{H^2 \sum_{k=1}^n \gamma_k^2 \sigma_k^2}{\varepsilon^2} \tag{49}$$

where we used the variance bound

$$\mathbb{E}[\xi_k^2] = \mathbb{E}[\mathbb{E}[|\langle U_k, \nabla E(Y_k)\rangle|^2 \,|\, \mathcal{F}_k]] \leq H^2 \sigma_k^2 \tag{50}$$

and the fact that $\mathbb{E}[\xi_k \xi_m] = \mathbb{E}[\xi_k \xi_m \,|\, \mathcal{F}_{k\vee m}] = 0$ whenever $k \neq m$. Since $\{\mathrm{I}_\infty \geq \varepsilon\} = \bigcup_n \{\mathrm{I}_n^* \geq \varepsilon\}$ is a union of nested events, we conclude that (Stb.I) holds whenever $\rho \geq C_{\mathrm{I}} \coloneqq (H/\varepsilon)^2 \sum_n \gamma_n^2 \sigma_n^2$.

(2) For the second term, we have $\mathrm{II}_\infty \leq \sum_n \gamma_n B_n$, so (Stb.II) holds for all $\rho \geq 0$ provided that $C_{\mathrm{II}} \coloneqq \sum_n \gamma_n B_n/\varepsilon \leq 1$.

(3) Finally, for the last term, Kolmogorov's inequality again yields

$$\mathbb{P}(\mathrm{III}_n^* \geq \varepsilon) \leq \frac{\mathbb{E}[\mathrm{III}_n^*]}{\varepsilon} = \frac{\beta \sum_{k=1}^n \gamma_k^2 M_k^2}{2\varepsilon}. \tag{51}$$

Consequently, the event $\{\mathrm{III}_\infty \geq \varepsilon\} = \bigcup_n \{\mathrm{III}_n^* \geq \varepsilon\}$ occurs with probability no more than $C_{\mathrm{III}} \coloneqq (\beta/2\varepsilon) \sum_n \gamma_n^2 M_n^2$, i.e., (Stb.III) holds whenever $\rho \geq C_{\mathrm{III}}$.

Assume now that ρ has been fixed. Since $\gamma_n = \gamma/n^p$ for some $\gamma > 0$ and $p \in (0,1]$, we can choose γ sufficiently small so that $C_{\mathrm{I}} \leq \rho/3$, $C_{\mathrm{II}} \leq 1$ and $C_{\mathrm{III}} \leq \rho/3$. In this case, (Stb) holds for $\rho \leftarrow \rho/3$, and our claim follows from Proposition 7. ■

7. Sharper convergence guarantees: the role of coherence

7.1. **Coherence: definition and examples.** In this section, we will show that our results on primal attractors can be strengthened considerably under the notion of *coherence:*

Definition 6. A nonempty compact subset $\mathcal{S}$ of $\mathcal{X}$ will be called *coherent* if it admits a (finite) set of *deviation directions* $\mathcal{Z} = \{z_1, \dots, z_m\} \subseteq \mathcal{V}$ such that

a) $\langle v(x), z\rangle < 0 \quad$ for all $x \in \mathcal{S}$ and all $z \in \mathcal{Z}$. (52a)

b) $Q(y) \to \mathcal{S} \quad$ whenever $\max_{z\in\mathcal{Z}} \langle y, z\rangle \to -\infty$. (52b)

In particular, if (52a) holds for all $x \in \mathcal{X}$, we will say that $\mathcal{S}$ is *globally coherent*; and if we want to stress that $\mathcal{S}$ is coherent but not globally so, we will say that $\mathcal{S}$ is *locally coherent.*

The motivation behind Definition 6 is as follows. First, condition (52a) posits that any deviation from $\mathcal{S}$ along a vector $z \in \mathcal{Z}$ is actively disincentivized by the players' individual gradient field v so, in a certain sense, v points locally "towards" $\mathcal{S}$. The second condition is game-independent and asks that the elements of $\mathcal{Z}$ are sufficient to identify $\mathcal{S}$ by acting as primal-dual "support vectors" for $\mathcal{S}$ under Q. The terminology "coherence" has been chosen precisely to indicate that these two properties dovetail to create a favorable convergence landscape under (MRM).

To illustrate the notion of coherence, we proceed below with a diverse range of examples. The first two concern finite games; the last two concern continuous ones.

Example 7.1 (Strict equilibria in finite games)**.** Recall that a strict Nash equilibrium of a finite game $\Gamma = \Gamma(\mathcal{N}, \mathcal{A}, u)$ is a strategy profile x^* such that (NE) holds as a strict inequality for all $x \neq x^*$. An immediate consequence of this definition is that *a)* x^* is *pure*, i.e., it is supported on a single pure strategy profile $\alpha^* \in \mathcal{A}$; and that *b)* unilateral deviations from α^* lead to *strictly* inferior payoffs, i.e., $u_\alpha(\alpha_i^*; \alpha_{-i}^*) > u_i(\alpha_i; \alpha_{-i}^*)$ for all $\alpha_i \in \mathcal{A}_i \backslash \{\alpha_i^*\}$, $i \in \mathcal{N}$.

With this in mind, consider the set of unilateral deviations

$$\mathcal{Z} = \{e_{i\alpha_i} - e_{i\alpha_i^*} : \alpha_i \in \mathcal{A}_i \backslash \{\alpha_i^*\}, i \in \mathcal{N}\}. \tag{53}$$

Since $\langle v(x^*), e_{i\alpha_i} - e_{i\alpha_i^*}\rangle = u_i(\alpha_i; \alpha_{-i}^*) - u_i(\alpha_i^*; \alpha_{-i}^*) < 0$ for all $\alpha_i \in \mathcal{A}_i \backslash \{\alpha_i^*\}$, $i \in \mathcal{N}$, condition (52a) is satisfied. Lemma A.3 further shows that $Q_{i\alpha_i}(y) \to 0$ whenever $y_{i\alpha_i} -$

$y_{i\alpha_i^*} \to -\infty$, so the requirement $Q(y) \to x^*$ of (52b) is also satisfied. In other words, *strict equilibria are coherent.* ¶

Example 7.2 (Extinction of dominated strategies)**.** Recall that a pure strategy $\alpha_i \in \mathcal{A}_i$ is *dominated* by $\beta_i \in \mathcal{A}_i$ if $u_i(\alpha_i; x_{-i}) < u_i(\beta_i; x_{-i})$ for all $x \in \mathcal{X}$. We then say that α_i is *eliminated* in a mixed strategy profile $x \in \mathcal{X}$ if α_i is not supported in x_i, i.e., if $x_{i\alpha_i} = 0$. A fundamental requirement for game-theoretic learning is that dominated strategies become extinct over time, i.e., that the trajectory of play converges to the set $\mathcal{X}^*$ of action profiles that eliminate all dominated strategies.[7]

This set is globally coherent. To see this, consider the set of dominant deviations

$$\mathcal{Z} = \{e_{i\alpha_i} - e_{i\beta_i} : \alpha_i \text{ is dominated by } \beta_i\}. \tag{54}$$

By definition, $\langle v(x), e_{i\alpha_i} - e_{i\beta_i}\rangle = u_i(\alpha_i; x_{-i}) - u_i(\beta_i; x_{-i}) < 0$ for all $x \in \mathcal{X}$, so (52a) holds globally. Moreover, for any finite game, $\mathcal{X}^*$ is a face of $\mathcal{X}$ [60] and hence compact. Finally, Lemma A.3 shows that $Q_{i\alpha_i}(y) \to 0$ if $y_{i\alpha_i} - y_{i\beta_i} \to -\infty$, so the requirement $Q(y) \to \mathcal{X}^*$ of (52b) is also satisfied, and we conclude that *the set of undominated strategies is globally coherent.* ¶

Example 7.3 (Sharp equilibria in concave games)**.** Following Polyak [52], a Nash equilibrium of a concave game is *sharp* if the stationarity condition (FOS) holds as a strict inequality for all $x \neq x^*$, i.e.,

$$\langle v(x^*), x - x^*\rangle < 0 \quad \text{for all } x \neq x^*. \tag{Sharp}$$

Examples of sharp equilibria include deterministic Nash policies in generic stochastic games [68], the Nash equilibria of the power control game of Example 2.3 [47], etc.

Geometrically, sharp equilibria can be characterized by the condition that $v(x^*)$ lies in the (topological) interior of the polar cone $\mathrm{PC}(x^*)$ to $\mathcal{X}$ at x^*. This means in particular that there exists a polyhedral cone $\mathcal{C}$ that is spanned by a finite set of vectors $\mathcal{Z} = \{z_1, \dots, z_m\} \subseteq \mathcal{V}$ such that *a*) the tangent cone $\mathrm{TC}(x^*)$ to $\mathcal{X}$ at x^* is contained in the interior of $\mathcal{C}$; and *b*) $\langle v(x^*), z\rangle < 0$ for all $z \in \mathcal{Z}$. Lemma A.4 in Appendix A shows that $Q(y) \to \mathcal{S}$ if $\max_{z\in\mathcal{Z}}\langle y, z\rangle \to -\infty$, so we conclude that *sharp equilibria are coherent.* ¶

Example 7.4 (Stochastic linear programming)**.** To borrow an example from optimization (viewed here as a single-player game), let $\mathcal{X}$ be a convex polytope and consider the stochastic linear program

$$\begin{aligned} \text{maximize} \quad & u(x) = \mathbb{E}_\theta[\langle V(\theta), x\rangle] \\ \text{subject to} \quad & x \in \mathcal{X} \end{aligned} \tag{SLP}$$

where $V(\theta)$ is a random payoff vector drawn from some complete probability space $(\Theta, \mathbb{P}_\theta)$. By linearity, the set of solutions $\mathcal{X}^* = \arg\max u$ of (SLP) is a face of $\mathcal{X}$; moreover, if we let $v = \mathbb{E}_\theta[V(\theta)] = \nabla u(x)$, we have $\langle v, x - x^*\rangle < 0$ whenever $x^* \in \mathcal{X}^*$ and $x \in \mathcal{X} \setminus \mathcal{X}^*$. Finally, since $\mathcal{X}$ is a convex polytope, there exists a finite set of vectors $\mathcal{Z} = \{z_1, \dots, z_m\}$ such that *a*) $x^* + z \in \mathcal{X} \setminus \mathcal{X}^*$ for all $x^* \in \mathcal{X}^*$, $z \in \mathcal{Z}$; and *b*) every point $x \in \mathcal{X} \setminus \mathcal{X}^*$ can be decomposed as $x = x^* + \lambda z$ for some $x^* \in \mathcal{X}^*$, $z \in \mathcal{Z}$ and $\lambda > 0$. Lemma A.4 in Appendix A shows that $Q(y) \to \mathcal{X}^*$ whenever $\langle y, z\rangle \to -\infty$ for all $z \in \mathcal{Z}$, so (52b) is satisfied and we conclude that *the solution set $\mathcal{X}^*$ of* (SLP) *is globally coherent.* ¶

The above examples illustrate that the notion of coherence underlies a diverse range of game-theoretic settings and problems. In light of this, we devote the rest of this section to analyzing the convergence properties of coherent sets under (MRM).

[7] The case of mixed strategies dominated by mixed strategies requires heavier notation, so we do not treat it here.

7.2. **Convergence analysis.** The first thing to note is that coherent sets are primal attractors. Indeed, if $\mathcal{S}$ is coherent, it is straightforward to check that the function

$$E(y) = \log\Big(1 + \sum\nolimits_{z\in\mathcal{Z}} \exp\langle y, z\rangle\Big) \tag{55}$$

is a local energy function for E. First, if $E(y) \to \inf E = 0$, we must have $\langle y, z\rangle \to -\infty$ for all $z \in \mathcal{Z}$, and hence $Q(y) \to \mathcal{S}$ by Definition 6. Moreover, for all y such that $x = Q(y)$ is sufficiently close to $\mathcal{S}$, we have $\nabla E(y) = \sum_{z\in\mathcal{Z}} \langle v(x), z\rangle e^{\langle y,z\rangle} / (1 + \sum_{z\in\mathcal{Z}} e^{\langle y,z\rangle}) < 0$ by the continuity of v. This shows that the requirements of Definition 5 are all satisfied, leading to the following corollary of Theorem 3:

Corollary 9. *Suppose that $\mathcal{S}$ is coherent, and let X_n be the sequence of play of* (MRM) *with step-size and gradient signal assumptions as in Theorem 3. Then the conclusions of Theorem 3 hold, namely (i) if $\mathcal{S}$ is globally coherent, X_n converges to $\mathcal{S}$ with probability 1; and (ii) if $\mathcal{S}$ is locally coherent, X_n converges locally to $\mathcal{S}$ with probability at least $1-\rho$ if γ is small enough relative to ρ.*

Corollary 9 is a strong convergence guarantee in itself, but it does not exploit the sharper structural properties of coherent sets. As we show below, the assumptions of Theorem 3 on the method's step-size and gradient signals can be relaxed considerably, allowing in many cases the use of *constant* step-sizes. The key step to achieve this is the following refinement of Lemma 1 for coherent sets.

Lemma 5. *Suppose that $\mathcal{S} \subseteq \mathcal{X}$ is coherent, and let $E_z(y) = \langle y, z\rangle$ for $y \in \mathcal{Y}$, $z \in \mathcal{Z}$. Then the iterates $E_n = E_z(Y_n)$ of E_z satisfy the template inequality*

$$E_{n+1} \leq E_n + \gamma_n \langle v(X_n), z\rangle + \gamma_n \xi_n + \gamma_n \chi_n. \tag{56}$$

where the error terms ξ_n and χ_n are now given by

$$\xi_n = \langle U_n, z\rangle \qquad \textit{and} \qquad \chi_n = \max\nolimits_{z\in\mathcal{Z}} \|z\| \cdot B_n. \tag{57}$$

Proof. Simply set $y \leftarrow Y_{n+1}$ in $E_z(y)$ and invoke the definition of (MRM). ■

Compared to Lemma 1, the template inequality (56) *does not* have a second-order term, so the second moment of $\hat{v}_n$ plays a much more minor role when dealing with coherent sets. This can be seen very clearly in the following coherent analogue of Proposition 6:

Proposition 11. *Suppose that $\mathcal{S}$ is globally coherent, and let $X_n = Q(Y_n)$ be the sequence of play generated by* (MRM)*. If* (Sub) *holds, then X_n converges to $\mathcal{S}$ with probability 1.*

The crucial difference between Propositions 6 and 11 is that the former requires the summability condition (Sum), while the latter requires *only* the subleading growth requirement (Sub). The latter assumption grants much more flexibility to the players because they can employ practically *any* step-size of the form $\gamma_n \propto 1/n^p$ for some $p \in [0,1]$. A similar situation arises for locally coherent sets, in which case the stability requirement (Stb) can be replaced by the "dominance" condition

$$\mathbb{P}(\mathrm{I}_n \leq C\tau_n^{\mu}/2 \ \text{ for all } n) \geq 1 - \rho \tag{Dom.I}$$

$$\mathbb{P}(\mathrm{II}_n \leq C\tau_n^{\mu}/2 \ \text{ for all } n) \geq 1 - \rho \tag{Dom.II}$$

for some $C > 0$ and $\mu \in [0,1)$. Under this milder condition, we have:

Proposition 12. *Suppose that $\mathcal{S}$ is locally coherent, fix some confidence level $\rho > 0$, and let $X_n = Q(Y_n)$ be the sequence of play generated by* (MRM)*. If* (Sub) *and* (Dom) *hold, there exists an unbounded open set $\mathcal{D} \subseteq \mathcal{Y}$ of initializations such that*

$$\mathbb{P}(X_n \textit{ converges to } \mathcal{S} \mid Y_1 \in \mathcal{D}) \geq 1 - (m+1)\rho. \tag{59}$$

To streamline our presentation, we defer the proof of Propositions 11 and 12 to the end of this section, and we proceed to state an explicit version of these results when (MRM) adheres to the general schedule (33). Our main result in this regard is as follows:

Theorem 4. *Let $X_n = Q(Y_n)$ be the sequence of play generated by* (MRM) *with step-size and gradient signal sequences as per* (33)*. Then:*

Case 1: *If $\mathcal{S}$ is globally coherent, X_n converges to $\mathcal{S}$ with probability* 1.

Case 2: *If $\mathcal{S}$ is locally coherent and, in addition, (i) $p - s > 1/2$; or (ii) $0 \leq p < q/(2+q)$ and $s < 1/2 - 1/q$, there exists an open set $\mathcal{D} \subseteq \mathcal{Y}$ of initializations such that, for any $\rho > 0$*

$$\mathbb{P}(X_n \text{ converges to } \mathcal{S} \mid Y_1 \in \mathcal{D}) \geq 1 - \rho \tag{60}$$

provided that $\gamma > 0$ is small enough relative to ρ.

In particular, by Proposition 4, we readily obtain the following explicit guarantees:

Corollary 10. *Suppose that Algorithms 1–9 are run with step-size $\gamma_n \propto 1/n^p$, $p \in [0,1]$, and where applicable, a sampling parameter $\delta_n \propto 1/n^r$, $r \in (0, 1/2)$. If $\mathcal{S}$ is globally coherent, X_n converges to $\mathcal{S}$ with probability* 1 *provided the following conditions are met:*

- *For Algorithms 1, 5* and *7–9: no additional requirements needed.*
- *For Algorithms 2–4* and *6: $p > 0$.*

Corollary 11. *Suppose that Algorithms 1–9 are run with step-size $\gamma_n \propto 1/n^p$, $p \in [0,1]$, and where applicable, a sampling parameter $\delta_n \propto 1/n^r$, $r \in (0, 1/2)$. Then the conclusions of Theorem 4 for locally coherent sets continue to hold provided the following conditions are met:*

- *For Algorithm 1: $p > 1/2$ if $q = 2$; no such requirement needed if $q > 2$.*
- *For Algorithms 2–4* and *6: $p > 1/2$ if $q = 2$; $p > 0$ otherwise.*
- *For Algorithms 5* and *7–9: no other requirements needed.*

We should stress here that, depending on the statistical properties of the players' feedback mechanism, the above results imply convergence even with a *constant* step-size, a feature which is quite unique in the context of stochastic approximation. To the best of our knowledge, the only comparable result in the literature in terms of step-size assumptions is the recent work of Giannou et al [22] for local convergence to strict Nash equilibria in finite games under a *"follow-the-generalized-leader"* (FTGL) scheme: since strict equilibria are locally coherent, the analysis of Giannou et al [22] corresponds to the last item of Corollary 11.

Perhaps surprisingly, the principal reason for this relaxation in terms of step-size requirements is *not* the boundedness of the q-th moments of the players' oracle: the step-size requirements of Section 6 cannot be relaxed for non-coherent attractors even if $q = \infty$; at the same time, the convergence guarantees of Theorem 4 for globally coherent sets yield convergence with a constant step-size even when $q = 2$. Instead, as we hinted at before, these sharper convergence properties are due to the fact that the quadratic error term $\mathrm{III}_n = \sum_{k=1}^n \gamma_k^2 \psi_k^2$ is not present in the case of coherent sets: it is precisely this simplification that leads to convergence with significantly faster step-size schedules.

Our last result builds on this observation to show that convergence occurs at a finite number of iterations if the mirror map of the process is surjective (e.g., if it is a Euclidean projection):

Theorem 5. *Suppose that the mirror map $Q\colon \mathcal{Y} \to \mathcal{X}$ of* (MRM) *is surjective. If $\mathcal{S}$ is coherent, then, with probability* 1*, every trajectory $X_n = Q(Y_n)$ that converges to $\mathcal{S}$ does so in a finite number of iterations, i.e., there exists some n_0 such that $X_n \in \mathcal{S}$ for all $n \geq n_0$.*

Corollary 12. *Suppose that* (MRM) *is run with Euclidean projections and step-size and gradient signal sequences as per* (33). *If $\mathcal{S}$ is globally coherent and $\mathcal{X}$ is compact, the induced sequence of play $X_n = Q(Y_n)$ converges to $\mathcal{S}$ in a finite number of iterations (a.s.).*

In view of the above, coherent sets comprise perhaps the most well-behaved class of rational outcomes under (MRM): the agents' sequence of play converges to such sets in a finite number of iterations, even with bandit, payoff-based feedback. We find this aspect particularly intriguing because it shows that the algorithms' long-run behavior remains robust in the face of uncertainty, a property with important implications for the theory of learning in games.

7.3. **Technical proofs.** We conclude this section with the proofs of Propositions 11 and 12 and Theorems 4 and 5. To set the stage for our analysis, it will be convenient to introduce the family of sets

$$\mathcal{D}(a) = \{y \in \mathcal{Y} : \max\nolimits_{z\in\mathcal{Z}} \langle y, z\rangle < -a\}. \tag{61}$$

By Definition 6, these sets are mapped to neighborhoods of $\mathcal{S}$ under Q, so they are particularly well-suited to serve as initialization domains for (MRM). In particular, by the requirements of Definition 6 and the continuity of v, there exists some a such that $c \coloneqq -\sup\{\langle v(Q(y)), z\rangle : y \in \mathcal{D}(a), z \in \mathcal{Z}\} < 0$. With all this in hand, the proofs of Propositions 11 and 12 are fairly straightforward.

Proof of Proposition 11. Since $\mathcal{S}$ is globally coherent, we can take $a = -\infty$ in the definition of $\mathcal{D}(a)$ above. Then, telescoping (56) readily yields

$$E_z(Y_{n+1}) \leq E_z(Y_1) - c\tau_n + \mathrm{I}_n + \mathrm{II}_n \quad \text{for all } z \in \mathcal{Z}. \tag{62}$$

Thus, if (Sub) holds, we get $E_z(Y_n) \to -\infty$ for all $z \in \mathcal{Z}$, i.e., $X_n = Q(Y_n) \to \mathcal{S}$. ■

Proof of Proposition 12. Let $\mu \in [0, 1)$ be such that (Dom) holds for every $z \in \mathcal{Z}$ (recall that ξ_n depends on z), and let $\Delta a = \max_n\{C\tau_n^\mu - c\tau_n\}$. Then, if Y_1 is initialized in $\mathcal{D} \coloneqq \mathcal{D}(a+\Delta a)$, we claim that $Y_n \in \mathcal{D}(a)$ for all n. Indeed, this being trivially true for $n = 1$, assume it to be the case for all $k = 1, 2, \dots, n$. Then, by (56) and our inductive hypothesis, we get

$$\begin{aligned} E_z(Y_{n+1}) &\leq E_z(Y_1) - \sum_{k=1}^{n} \gamma_k \langle v(X_k), z\rangle + \mathrm{I}_n + \mathrm{II}_n \\ &\leq E_z(Y_1) - c\tau_n + C\tau_n^\mu/2 + C\tau_n^\mu/2 \leq -a - \Delta a + \Delta a \leq -a \end{aligned} \tag{63}$$

i.e., $Y_{n+1} \in \mathcal{D}(a)$, as claimed. Since $Y_n \in \mathcal{D}(a)$ for all n, we conclude that (62) holds with probability 1 on the event that (Dom.I) and (Dom.II) both hold for all $z \in \mathcal{Z}$. Since (Dom.I) involves $|\mathcal{Z}| = m$ separate events (one for each $z \in \mathcal{Z}$) and II_n does not depend on z, it follows that $E_z(Y_n) \to -\infty$ for all $z \in \mathcal{Z}$ with probability at least $1 - (m + 1)\rho$. Our claim then follows from Definition 6. ■

We are now in a position to prove Theorem 4.

Proof of Theorem 4. As in the case of Theorem 3, our proof will hinge on showing that (Sub) and (Dom) hold under the stated step-size and sampling parameter schedules. Our claim will then follow by a direct application of Propositions 11 and 12.

First, regarding (Sub), the law of large numbers for martingale difference sequences [25, Theorem 2.18] shows that $\mathrm{I}_n/\tau_n \to 0$ w.p.1 on the event $\left\{\sum_n \gamma_n^2\, \mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_n]/\tau_n^2 < \infty\right\}$. However

$$\mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_n] \leq \|z\|^2\, \mathbb{E}[\|U_n\|_*^2 \,|\, \mathcal{F}_n] \leq \|z\|^2 \sigma_n^2 = \mathcal{O}(n^{2s}) \tag{64}$$

so, in turn, given that $s < 1/2$, we get

$$\sum_n \frac{\gamma_n^2\,\mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_n]}{\tau_n^2} = \mathcal{O}\left(\sum_n \frac{\gamma_n^2 \sigma_n^2}{\tau_n^2}\right) = \mathcal{O}\left(\sum_n \frac{n^{-2p} n^{2s}}{n^{2(1-p)}}\right) = \mathcal{O}\left(\sum_n \frac{1}{n^{2-2s}}\right) < \infty. \tag{65}$$

This establishes (Sub.I); as for the requirement (Sub.II), this follows by noting that $\sum_{k=1}^n \gamma_k B_k \big/ \sum_{k=1}^n \gamma_k \to 0$ if and only if $B_n \to 0$, which is immediate from (33). This shows that (Sub) holds, so the first case of the theorem follows from Proposition 11.

Now, for the second case of the theorem, since B_n is deterministic and $B_n = \mathcal{O}(1/n^b)$ for some $b > 0$, it is always possible to find $C > 0$ and $\mu \in (0,1)$ so that (Dom.II) holds. We are thus left to establish (Dom.I). To that end, let $\mathrm{I}_n^* = \sup_{1\leq k\leq n} |\mathrm{I}_n|$ and set $P_n \coloneqq \mathbb{P}(\mathrm{I}_n^* > C\tau_n^\mu/2)$ so

$$P_n \leq \frac{\mathbb{E}[|\mathrm{I}_n|^q]}{(C/2)^q \tau_n^{\mu q}} \leq c_q \frac{\mathbb{E}\big[\big(\sum_{k=1}^n \gamma_k^2 \|U_k\|_*^2\big)^{q/2}\big]}{\tau_n^{\mu q}} \tag{66}$$

where c_q is a positive constant depending only on C and q, and we used Kolmogorov's inequality [25, Corollary 2.1] in the first step and the Burkholder–Davis–Gundy inequality [25, Theorem 2.10] in the second. To proceed, we will require the following variant of Hölder's inequality [7, p. 15]:

$$\left(\sum_{k=1}^n a_k b_k\right)^\rho \leq \left(\sum_{k=1}^n a_k^{\frac{\lambda\rho}{\rho-1}}\right)^{\rho-1} \sum_{k=1}^n a_k^{(1-\lambda)\rho} b_k^\rho \tag{67}$$

valid for all $a_k, b_k \geq 0$ and all $\rho > 1$, $\lambda \in [0,1)$. Then, substituting $a_k \leftarrow \gamma_k^2$, $b_k \leftarrow \|U_k\|_*^2$, $\rho \leftarrow q/2$ and $\lambda \leftarrow 1/2 - 1/q$, (66) gives

$$P_n \leq c_q \frac{\big(\sum_{k=1}^n \gamma_k\big)^{q/2-1} \sum_{k=1}^n \gamma_k^{1+q/2}\, \mathbb{E}[\|U_k\|_*^q]}{\tau_n^{\mu q}} \leq c_q \frac{\sum_{k=1}^n \gamma_k^{1+q/2} \sigma_k^q}{\tau_n^{1+(\mu-1/2)q}} \tag{68}$$

We now consider two cases, depending on whether the numerator of (68) is summable or not.

Case 1: $p(1+q/2) \geq 1 + qs$. In this case, the numerator of (68) is summable under (33), so the fraction in (68) behaves as $\mathcal{O}(1/n^{(1-p)(1+(\mu-1/2)q)})$.

Case 2: $p(1+q/2) < 1 + qs$. In this case, the numerator of (68) is not summable under (33), so the fraction in (68) behaves as $\mathcal{O}\big(n^{1-p(1+q/2)+qs} \big/ n^{(1-p)(1+(\mu-1/2)q)}\big)$.

Thus, working out the various exponents, a straightforward – if tedious – calculation shows that there exists some $\mu \in (0,1)$ such that P_n is summable as long as $s < 1/2 - 1/q$ and $0 \leq p < q/(2+q)$. Hence, if γ is sufficiently small relative to ρ, we conclude that

$$\mathbb{P}(\mathrm{I}_n \leq C\tau_n^\mu/2 \text{ for all } n) \geq 1 - \textstyle\sum_n P_n \geq 1 - \rho/2. \tag{69}$$

Finally, if $p > 1/2 + s$, (Dom.I) is a straightforward consequence of (Stb.I). Our assertion then follows by putting everything together and invoking Proposition 12. ■

We conclude this section with the proof of our finite-time convergence result.

Proof of Theorem 5. Since Q is surjective, Lemma A.1 shows that $Q^{-1}(\mathcal{S})$ contains a shifted copy of $\bigcup_{x\in\mathcal{S}} \mathrm{PC}(x)$. Thus, given that $\max_{z\in\mathcal{Z}} \langle Y_n, z\rangle \to -\infty$ by the proof of Theorem 4, it follows that, for every $a \in \mathbb{R}$, there exists some (possibly random) $n_0 \equiv n_0(a)$ such that $\max_{z\in\mathcal{Z}} \langle Y_n, z\rangle < -a$ for all $n \geq n_0$. This shows that Y_n converges to $Q^{-1}(\mathcal{S})$ within a finite number of iterations, as claimed. ■

8. Concluding remarks

The proposed mirrored Robbins–Monro (MRM) stochastic approximation framework captures a wide range of existing algorithms, both first- and zeroth-order, and it allows us to derive a series of convergence results in a unified way. Conceptually speaking, an appealing feature of this framework lies in the fact that it provides a scaffolding that can be used in several other settings and algorithms of interest. The associated workflow is as follows:

(1) Estimate the bounds B_n and M_n for the signal sequence $\hat{v}_n$ of the method under study.

(2) Find suitable exponents b and s such that $B_n = 1/n^b$ and $M_n = 1/n^s$, as per Proposition 4.

(3) Determine the allowable range of step-size and/or other parameters by backsolving the requirements of Theorems 2–4 for b and s.

In this way, we can immediately derive the properties of several other algorithmic schemes in the literature, such as extra-gradient algorithms with zeroth-order feedback, optimistic multiplicative weights updates with payoff-based information, learning with implicitly normalized forecasters in the spirit of Audibert and Bubeck [2], etc. We leave the inclusion of even more general frameworks – such as algorithms with adaptive step-sizes, learning with asyncrhonous and/or delayed feedback, etc. – to future work.

Acknowledgments

Part of this work was done while the first author was visiting the Simons Institute for the Theory of Computing. P. Mertikopoulos is grateful for financial support by the French National Research Agency (ANR) in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02), the LabEx PERSYVAL (ANR-11-LABX-0025-01), MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and the grant ALIAS (ANR-19-CE48-0018-01).

Appendix A. Regularizers and mirror maps

In this appendix we present some basic properties of the mirror map Q. To state them, recall first that the subdifferential of a h at $x \in \mathcal{X}$ is defined as $\partial h(x) \coloneqq \{y \in \mathcal{Y} : h(x') \geq h(x) + \langle y, x' - x\rangle \text{ for all } x' \in \mathcal{V}\}$, the *domain of subdifferentiability* of h is $\operatorname{dom}\partial h \coloneqq \{x \in \operatorname{dom} h : \partial h \neq \varnothing\}$, and the convex conjugate of h is defined as $h^*(y) = \max_{x\in\mathcal{X}}\{\langle y, x\rangle - h(x)\}$ for all $y \in \mathcal{Y}$. We then have the following basic results.

Lemma A.1. *Let h be a regularizer on $\mathcal{X}$, and let $Q\colon \mathcal{Y} \to \mathcal{X}$ be its induced mirror map. Then:*

(1) Q is single-valued on $\mathcal{Y}$: in particular, for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, we have $x = Q(y) \iff y \in \partial h(x)$.

(2) The prox-domain $\mathcal{X}_h \coloneqq \operatorname{im} Q$ of h satisfies $\operatorname{ri}\mathcal{X} \subseteq \mathcal{X}_h \subseteq \mathcal{X}$.

(3) Q is $(1/K)$-Lipschitz continuous and $Q = \nabla h^*$.

Lemma A.2. *Let h be a regularizer on $\mathcal{X}$ with induced mirror map $Q\colon \mathcal{Y} \to \mathcal{X}$, and let $F(p, y) = h(p) + h^*(y) - \langle y, p\rangle$ for $p \in \mathcal{X}$, $y \in \mathcal{Y}$. Then, for all $y' \in \mathcal{Y}$, we have:*

$$a)\quad F(p, y) \geq \tfrac{1}{2}K\,\|Q(y) - p\|^2. \tag{A.1a}$$

$$b)\quad F(p, y') \leq F(p, y) + \langle y' - y, Q(y) - p\rangle + \tfrac{1}{2K}\|y' - y\|_*^2. \tag{A.1b}$$

In particular, if $h(0) = 0$, we have

$$(K/2)\|Q(y)\|^2 \leq h^*(y) \leq -\min h + \langle y, Q(y)\rangle + (2/K)\|y\|_*^2 \quad \textit{for all } y \in \mathcal{Y} \tag{A.2}$$

Variants of these lemmas can be found in [10, 45, 46], so we omit their proof. The next properties we discuss concern the way that different regions of $\mathcal{Y}$ are mapped to $\mathcal{X}$ under Q.

Lemma A.3 (Mertikopoulos and Sandholm, 2016, Prop. A.1)**.** *Let h be a regularizer on the simplex $\Delta(\mathcal{A}) \subseteq \mathbb{R}^{\mathcal{A}}$. If $y_\alpha - y_\beta \to -\infty$, then $Q_\alpha(y) \to 0$.*

Lemma A.4. *Let h be a regularizer on $\mathcal{X}$, let y_n, $n = 1, 2, \dots$ be a sequence in $\mathcal{Y}$, and fix some $x \in \mathcal{X}$. If $\langle y_n, z\rangle \to -\infty$ for every nonzero $z \in \mathrm{TC}(x)$, we have $Q(y_n) \to x$.*

Proof. Assume that $\limsup_n \|x_n - x\| > 0$. Then, given that $y_n \in \partial h(x_n)$, we get $h(x) \geq h(x_n) + \langle y_n, x - x_n\rangle \geq h(x_n) - \langle y_n, z_n\rangle \|x_n - x\|$, where we set $z_n = (x_n - x)/\|x_n - x\|$. If we further assume (by descending to a subsequence if needed) that z_n converges in the unit sphere of $\|\cdot\|$, there exists some $z \in \mathrm{TC}(x)$ with $\|z\| = 1$ and such that $\langle y_n, z_n\rangle \leq (1+\varepsilon)\langle y_n, z\rangle$ for some $\varepsilon > 0$. Thus, taking the lim sup of the above estimate gives $h(x) \geq \infty$, a contradiction which proves our claim. ∎

Lemma A.5. *Let h be a regularizer on a convex polytope $\mathcal{P}$ of $\mathcal{V}$, let $\mathcal{S}$ be a face of $\mathcal{P}$, and let $\mathcal{Z} = \{z_1, \dots, z_m\}$ be a set of unit vectors of $\mathcal{V}$ such that every point $x \in \mathcal{P} \setminus \mathcal{S}$ can be written as $x = p + \lambda z$ for some $p \in \mathcal{S}$, $z \in \mathcal{Z}$ and $\lambda > 0$. If $\max_{z \in \mathcal{Z}} \langle y, z\rangle \to -\infty$, then $Q(y) \to \mathcal{S}$.*

Proof. By the compactness of $\mathcal{P}$ (and descending to a subsequence if necessary), we may assume that $x_n = Q(y_n)$ converges to some $x \in \mathcal{P}$. If $x \notin \mathcal{S}$, there exist $p \in \mathcal{S}$, $z \in \mathcal{Z}$ and $\lambda > 0$ such that $x = p + \lambda z$. In turn, this gives $h(p) \geq h(x_n) + \langle y_n, p - x_n\rangle = h(x_n) - \langle y_n, z_n\rangle \|x_n - p\|$ where we set $z_n = (x_n - p)/\|x_n - p\|$. Since $z_n \to z$, taking $n \to \infty$ yields $h(p) \geq \infty$, a contradiction which shows that $x = \lim x_n \in \mathcal{S}$, as claimed. ∎

Appendix B. Error estimates

Our aim in this appendix is to prove the bounds on the bias and magnitude of $\hat{v}_n$ reported in Table 1. We proceed to do so on a method-by-method basis.

Proof of Proposition 4. We begin with the oracle-based methods of Section 3.1, namely Algorithms 1–6. For this, we will make free use of the fact that we can take $M_n^q = 3^{q-1}(G^q + B_n^q + \sigma_n^q)$ in (11), cf. the discussion after (9).

Algorithm 1: Stochastic gradient ascent. For (SGA), we have $U_n = \mathrm{Err}(X_n; \theta_n)$ and $b_n = 0$, so our claim follows immediately from the stated assumptions for (SFO).

Algorithm 2: Sequential gradient ascent. For (seqGA), we have

$$\hat{v}_{i,n} = V_i(\hat{X}_n^i; \theta_n) = v_i(\hat{X}_n^i) + \mathrm{Err}_i(\hat{X}_n^i; \theta_n). \tag{B.1}$$

where $\hat{X}_n^i = (\dots, X_{i-1,n+1}, X_{i,n}, X_{i+1,n}, \dots)$. We thus have $\mathbb{E}[\hat{v}_{i,n} \,|\, \mathcal{F}_n] = \mathbb{E}[v_i(\hat{X}_n^i) \,|\, \mathcal{F}_n]$ and hence

$$\begin{aligned}
\|b_{i,n}\|_* &\leq \mathbb{E}[\|v_i(\hat{X}_n^i) - v_i(X_n)\|_* \,|\, \mathcal{F}_n] \\
&\leq L_i\, \mathbb{E}[\|\hat{X}_n^i - X_n\| \,|\, \mathcal{F}_n] \\
&\leq \gamma_n L_i \max\nolimits_{j<i} \mathbb{E}[\|v_j(\hat{X}_n^j) + \mathrm{Err}(\hat{X}_n^j; \theta_n)\|_* \,|\, \mathcal{F}_n] \\
&\leq \gamma_n L_i (G + \sigma) = \mathcal{O}(\gamma_n) = \mathcal{O}(1/n^p).
\end{aligned} \tag{B.2}$$

Likewise, the noise term $U_{i,n}$ in (9) can be bounded as

$$\begin{aligned}
\|U_{i,n}\|_* &= \|\hat{v}_{i,n} - \mathbb{E}[\hat{v}_{i,n} \,|\, \mathcal{F}_n]\|_* \\
&= \|v_i(\hat{X}_n^i) - \mathbb{E}[v_i(\hat{X}_n^i) \,|\, \mathcal{F}_n] + \mathrm{Err}_i(\hat{X}_n^i; \theta_n)\|_* \leq 2G + \|\mathrm{Err}_i(\hat{X}_n^i; \theta_n)\|_*
\end{aligned} \tag{B.3}$$

which yields $\mathbb{E}[\|U_n\|_*^q \,|\, \mathcal{F}_n] = \mathcal{O}(G^q + \sigma^q) = \mathcal{O}(1)$ under the requirements (5) for (SFO).

Algorithm 3: Extra-gradient. For (EG), we have $\hat{v}_n = V(X_{n+1/2}; \theta_{n+1/2})$ so $\mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n] = \mathbb{E}[v(X_{n+1/2}) \,|\, \mathcal{F}_n]$. We thus get

$$\begin{aligned} \|b_n\|_* = \|\mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n] - v(X_n)\|_* &\leq \mathbb{E}[\|v(X_{n+1/2}) - v(X_n)\|_* \,|\, \mathcal{F}_n] \\ &\leq L\,\mathbb{E}[\|X_{n+1/2} - X_n\| \,|\, \mathcal{F}_n] \\ &\leq \gamma_n L\,\mathbb{E}[\|V(X_n;\theta_n)\|_* \,|\, \mathcal{F}_n] \\ &= \gamma_n L\,\mathbb{E}[\|v(X_n) + \mathrm{Err}(X_n;\theta_n)\|_* \,|\, \mathcal{F}_n] \\ &\leq \gamma_n L(G+\sigma) = \mathcal{O}(\gamma_n) = \mathcal{O}(1/n^p) \end{aligned} \tag{B.4}$$

and, analogously

$$\|U_n\|_* = \|\hat{v}_n - \mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n]\|_* = \|v(X_{n+1/2}) - \mathbb{E}[v(X_{n+1/2}) \,|\, \mathcal{F}_n] + \mathrm{Err}(X_{n+1/2}; \theta_{n+1/2})\|_* \tag{B.5}$$

so $\mathbb{E}[\|U_n\|_*^q \,|\, \mathcal{F}_n] = \mathcal{O}(G^q + \sigma^q) = \mathcal{O}(1)$ under (5), as claimed.

Algorithm 4: Optimistic gradient. For (OG), we have again $\mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n] = \mathbb{E}[v(X_{n+1/2}) \,|\, \mathcal{F}_n]$, so the same series of arguments as above gives

$$\begin{aligned} \|b_n\|_* &= \|\mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n] - v(X_n)\|_* \\ &\leq L\,\mathbb{E}[\|X_{n+1/2} - X_n\| \,|\, \mathcal{F}_n] \\ &\leq \gamma_n L\,\mathbb{E}[\|V(X_{n-1/2};\theta_{n-1})\|_* \,|\, \mathcal{F}_n] \\ &= \gamma_n L\,\mathbb{E}[\|v(X_{n-1/2}) + \mathrm{Err}(X_{n-1/2};\theta_{n-1})\|_* \,|\, \mathcal{F}_n] \\ &\leq \gamma_n L(G+\sigma) = \mathcal{O}(\gamma_n) = \mathcal{O}(1/n^p) \end{aligned} \tag{B.6}$$

under (5) with $q = \infty$. The noise term U_n can be bounded in exactly the same way, so we omit the calculations.

Algorithm 5: Exponential weights. We consider two cases, based on the information available to the players. For the full information oracle (6a), we have $\hat{v}_n = v(X_n)$ so $b_n = U_n = 0$ by definition (i.e., the oracle is perfect). Otherwise, under the realization-based oracle (6b), we have $\mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n] = \mathbb{E}[v(\alpha_n) \,|\, \mathcal{F}_n] = v(X_n)$ because α_n is sampled according to X_n. We thus get $b_n = 0$ and $U_n = \mathcal{O}(1)$, which proves our assertion.

Algorithm 6: Mirror-prox. Mirroring the analysis for (EG), we have

$$\begin{aligned} \|b_n\|_* &\leq \mathbb{E}[\|v(X_{n+1/2}) - v(X_n)\|_* \,|\, \mathcal{F}_n] \\ &\leq L\,\mathbb{E}[\|X_{n+1/2} - X_n\| \,|\, \mathcal{F}_n] \\ &\leq (L/K)\,\mathbb{E}[\|Y_{n+1/2} - Y_n\|_* \,|\, \mathcal{F}_n] \\ &\leq \gamma_n (L/K)\,\mathbb{E}[\|V(X_n;\theta_n)\|_* \,|\, \mathcal{F}_n] \\ &\leq \gamma_n L(G+\sigma)/K = \mathcal{O}(\gamma_n) = \mathcal{O}(1/n^p) \end{aligned} \tag{B.7}$$

where the estimate in the second line follows from Lemma A.1. The rest now follows as in the case of Algorithm 3.

We now proceed with the payoff-based methods of Section 3.2, namely Algorithms 7–9.

Algorithm 7: Single-point stochastic approximation. Since u_i is assumed bounded in the context of (SPSA), the bound for M_n follows trivially. As for the bias of (SPSA), it will be convenient to set $V_i^\delta(x;w) = (d_i/\delta)\, u_i(x+\delta w)\, w_i$ so, in obvious notation, $\hat{v}_{i,n} = V_i^{\delta_n}(X_n; W_n)$. Thus, if we fix a pivot point $x \in \mathcal{X}$ and a query point $\hat{x} = x + \delta w$ for some $w \in \mathcal{E} = \prod_i \mathcal{E}_i$, a first-order Taylor expansion of u_i with integral remainder gives

$$V_i^\delta(x;w) = \frac{d_i}{\delta} u_i(\hat{x}) \cdot w_i = \frac{d_i}{\delta} u_i(x) \cdot w_i + \frac{d_i}{\delta} \langle \nabla u_i(x), z \rangle \cdot w_i \tag{B.8a}$$

$$+ \int_0^1 \langle \nabla u_i(x+\tau z) - \nabla u_i(x), z\rangle \, d\tau \cdot w_i \tag{B.8b}$$

where we set $z = \hat{x} - x = \delta w$. Hence, if w is drawn uniformly at random from $\mathcal{E}$, taking expectations yields

$$\begin{aligned}\mathbb{E}[(\text{B.8a})] &= \frac{d_i}{\delta}\,\mathbb{E}[\langle v_i(x), z_i\rangle\, w_i] + \frac{d_i}{\delta}\sum_{j\neq i}\langle \nabla_{x_j} u_i(x), \mathbb{E}[z_j]\rangle\; \mathbb{E}[w_i] \\ &= d_i\,\mathbb{E}[\langle v_i(x), w_i\rangle\, w_i] = d_i \cdot \frac{1}{2d_i}\sum_{\ell=1}^{d_i}[v_{i\ell}(x)e_{i\ell} - v_{i\ell}(x)(-e_{i\ell})] = v_i(x)\end{aligned} \tag{B.9}$$

where we used the fact that $\mathbb{E}[w_i] = 0$ for all $i \in \mathcal{N}$ and that w_i and w_j are independent for all $i, j \in \mathcal{N}$, $i \neq j$. As for the second term, Assumption 1 readily yields

$$\|\mathbb{E}[(\text{B.8b})]\| \leq \frac{d_i}{\delta}\int_0^1 L_i \delta^2 \|w\|^2 \tau \, d\tau = \mathcal{O}(L\delta). \tag{B.10}$$

Thus, by combining (B.9) and (B.10), we conclude that $b_{i,n} = \mathbb{E}[V_i^{\delta_n}(X_n; W_n) \,|\, \mathcal{F}_n] - v_i(X_n) = \mathcal{O}(\delta_n)$, which immediately yields the desired bound $B_n = \mathcal{O}(\delta_n) = \mathcal{O}(1/n^r)$ for (SPSA).

Algorithm 8: Dampened gradient approximation. Recall that $\hat{v}_{i,n} = n \cdot \log(1 + (u_i(X_{n+1/2}) - u_i(X_n))W_{i,n})$. Since $u_i(X_{n+1/2}) - u_i(X_n) = (1/n)v_i(X_n)W_{i,n} + \mathcal{O}(1/n^2)$ by the definition of $X_{n+1/2}$, expanding the logairthm readily yiels $B_n = \mathcal{O}(1/n)$ and $M_n = \mathcal{O}(1)$. Our claim then follows as above.

Algorithm 9: Exponential weights for exploration and exploitation. Since $\hat{\alpha}_n$ is sampled according to $\hat{X}_n$, we readily get $\mathbb{E}[\hat{v}_{i,n} \,|\, \mathcal{F}_n] = v_i(\hat{X}_n)$, so $B_n = \mathcal{O}(\|\hat{X}_n - X_n\|) = \mathcal{O}(\delta_n) = \mathcal{O}(1/n^r)$. Moreover, since $\hat{X}_{i\alpha_i,n} \geq \delta_n/A_i$, it follows that $\|\hat{v}_n\|_* = \mathcal{O}(1/\delta_n) = \mathcal{O}(n^r)$, and our proof is complete. ■

References

[1] Arrow KJ, Hurwicz L, Uzawa H (1958) Studies in linear and non-linear programming. Stanford University Press

[2] Audibert JY, Bubeck S (2010) Regret bounds and minimax policies under partial monitoring. Journal of Machine Learning Research 11:2635–2686

[3] Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (1995) Gambling in a rigged casino: The adversarial multi-armed bandit problem. In: Proceedings of the 36th Annual Symposium on Foundations of Computer Science

[4] Azizian W, Iutzeler F, Malick J, Mertikopoulos P (2021) The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities. In: COLT '21: Proceedings of the 34th Annual Conference on Learning Theory

[5] Balduzzi D, Racaniere S, Martens J, Foerster J, Tuyls K, Graepel T (2018) The mechanics of n-player differentiable games. In: ICML '18: Proceedings of the 35th International Conference on Machine Learning

[6] Bauschke HH, Combettes PL (2017) Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2nd edn. Springer, New York, NY, USA

[7] Benaïm M (1999) Dynamics of stochastic approximation algorithms. In: Azéma J, Émery M, Ledoux M, Yor M (eds) Séminaire de Probabilités XXXIII, Lecture Notes in Mathematics, vol 1709, Springer Berlin Heidelberg, pp 1–68

[8] Benaïm M, Hirsch MW (1996) Asymptotic pseudotrajectories and chain recurrent flows, with applications. Journal of Dynamics and Differential Equations 8(1):141–176

[9] Bervoets S, Bravo M, Faure M (2020) Learning with minimal information in continuous games. Theoretical Economics 15:1471–1508

[10] Bravo M, Mertikopoulos P (2017) On the robustness of learning in games with stochastically perturbed payoff observations. Games and Economic Behavior 103, John Nash Memorial issue:41–66

[11] Bravo M, Leslie DS, Mertikopoulos P (2018) Bandit learning in concave N-person games. In: NeurIPS '18: Proceedings of the 32nd International Conference of Neural Information Processing Systems

[12] Brown GW (1951) Iterative solutions of games by fictitious play. In: Coopmans TC (ed) Activity Analysis of Productions and Allocation, 374-376, Wiley

[13] Censor Y, Lent A (1981) An iterative row action method for internal convex programming. Journal of Optimization Theory and Applications 34:321–353

[14] Chen G, Teboulle M (1993) Convergence analysis of a proximal-like minimization algorithm using Bregman functions. SIAM Journal on Optimization 3(3):538–543

[15] Cohen J, Héliou A, Mertikopoulos P (2017) Learning with bandit feedback in potential games. In: NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems

[16] Conley CC (1978) Isolated Invariant Set and the Morse Index. American Mathematical Society, Providence, RI

[17] Daskalakis C, Panageas I (2019) Last-iterate convergence: Zero-sum games and constrained min-max optimization. In: ITCS '19: Proceedings of the 10th Conference on Innovations in Theoretical Computer Science

[18] Daskalakis C, Ilyas A, Syrgkanis V, Zeng H (2018) Training GANs with optimism. In: ICLR '18: Proceedings of the 2018 International Conference on Learning Representations

[19] Debreu G (1952) A social equilibrium existence theorem. Proceedings of the National Academy of Sciences of the USA 38(10):886–893

[20] Duvocelle B, Mertikopoulos P, Staudigl M, Vermeulen D (to appear) Multi-agent online learning in time-varying games. Mathematics of Operations Research

[21] Flokas L, Vlatakis-Gkaragkounis EV, Lianeas T, Mertikopoulos P, Piliouras G (2020) No-regret learning and mixed Nash equilibria: They do not mix. In: NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems

[22] Giannou A, Vlatakis-Gkaragkounis EV, Mertikopoulos P (2021) The convergence rate of regularized learning in games: From bandits and uncertainty to optimism and beyond. In: NeurIPS '21: Proceedings of the 35th International Conference on Neural Information Processing Systems

[23] Giannou A, Vlatakis-Gkaragkounis EV, Mertikopoulos P (2021) Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information. In: COLT '21: Proceedings of the 34th Annual Conference on Learning Theory

[24] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: NIPS '14: Proceedings of the 28th International Conference on Neural Information Processing Systems

[25] Hall P, Heyde CC (1980) Martingale Limit Theory and Its Application. Probability and Mathematical Statistics, Academic Press, New York

[26] Hart S, Mas-Colell A (2003) Uncoupled dynamics do not lead to Nash equilibrium. American Economic Review 93(5):1830–1836

[27] Hart S, Mas-Colell A (2006) Stochastic uncoupled dynamics and Nash equilibrium. Games and Economic Behavior 57:286–303

[28] Hiriart-Urruty JB, Lemaréchal C (2001) Fundamentals of Convex Analysis. Springer, Berlin

[29] Hofbauer J, Sandholm WH (2002) On the global convergence of stochastic fictitious play. Econometrica 70(6):2265–2294

[30] Hofbauer J, Sigmund K (2003) Evolutionary game dynamics. Bulletin of the American Mathematical Society 40(4):479–519

[31] Hofbauer J, Schuster P, Sigmund K (1979) A note on evolutionarily stable strategies and game dynamics. Journal of Theoretical Biology 81(3):609–612

[32] Hsieh YG, Iutzeler F, Malick J, Mertikopoulos P (2019) On the convergence of single-call stochastic extra-gradient methods. In: NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp 6936–6946

[33] Hsieh YP, Mertikopoulos P, Cevher V (2021) The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In: ICML '21: Proceedings of the 38th International Conference on Machine Learning

[34] Juditsky A, Nemirovski AS, Tauvel C (2011) Solving variational inequalities with stochastic mirror-prox algorithm. Stochastic Systems 1(1):17–58

[35] Kamalaruban P, Huang YT, Hsieh YP, Rolland P, Shi C, Cevher V (2020) Robust reinforcement learning via adversarial training with langevin dynamics. arXiv preprint arXiv:200206063

[36] Kelly FP, Maulloo AK, Tan DKH (1998) Rate control for communication networks: shadow prices, proportional fairness and stability. Journal of the Operational Research Society 49(3):237–252

[37] Korpelevich GM (1976) The extragradient method for finding saddle points and other problems. Èkonom i Mat Metody 12:747–756

[38] Kushner HJ, Yin GG (1997) Stochastic approximation algorithms and applications. Springer-Verlag, New York, NY

[39] Lattimore T, Szepesvári C (2020) Bandit Algorithms. Cambridge University Press, Cambridge, UK

[40] Leslie DS, Collins EJ (2006) Generalised weakened fictitious play. Games and Economic Behavior 56(2):285–298

[41] Liang T, Stokes J (2019) Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In: AISTATS '19: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics

[42] Littlestone N, Warmuth MK (1994) The weighted majority algorithm. Information and Computation 108(2):212–261

[43] Maynard Smith J, Price GR (1973) The logic of animal conflict. Nature 246:15–18

[44] Mertikopoulos P, Sandholm WH (2016) Learning in games via reinforcement and regularization. Mathematics of Operations Research 41(4):1297–1324

[45] Mertikopoulos P, Staudigl M (2018) On the convergence of gradient-like flows with noisy gradient input. SIAM Journal on Optimization 28(1):163–197

[46] Mertikopoulos P, Zhou Z (2019) Learning in games with continuous action sets and unknown payoff functions. Mathematical Programming 173(1-2):465–507

[47] Mertikopoulos P, Belmega EV, Moustakas AL, Lasaulce S (2011) Dynamic power allocation games in parallel multiple access channels. In: ValueTools '11: Proceedings of the 5th International Conference on Performance Evaluation Methodologies and Tools

[48] Mertikopoulos P, Lecouat B, Zenati H, Foo CS, Chandrasekhar V, Piliouras G (2019) Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In: ICLR '19: Proceedings of the 2019 International Conference on Learning Representations

[49] Monderer D, Shapley LS (1996) Potential games. Games and Economic Behavior 14(1):124 – 143

[50] Nemirovski AS (2004) Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM Journal on Optimization 15(1):229–251

[51] Nisan N, Roughgarden T, Tardos É, Vazirani VV (eds) (2007) Algorithmic Game Theory. Cambridge University Press

[52] Polyak BT (1987) Introduction to Optimization. Optimization Software, New York, NY, USA

[53] Popov LD (1980) A modification of the Arrow–Hurwicz method for search of saddle points. Mathematical Notes of the Academy of Sciences of the USSR 28(5):845–848

[54] Rakhlin A, Sridharan K (2013) Optimization, learning, and games with predictable sequences. In: NIPS '13: Proceedings of the 27th International Conference on Neural Information Processing Systems

[55] Ratliff LJ, Burden SA, Sastry SS (2014) Genericity and structural stability of non-degenerate differential nash equilibria. In: ACC '14: Proceedings of the 2014 American Control Conference

[56] Ratliff LJ, Burden SA, Sastry SS (2016) On the characterization of local Nash equilibria in continuous games. IEEE Trans Autom Control 61(8):2301–2307

[57] Robinson J (1951) An iterative method for solving a game. Annals of Mathematics 54:296–301

[58] Rosen JB (1965) Existence and uniqueness of equilibrium points for concave N-person games. Econometrica 33(3):520–534

[59] Rosenthal RW (1973) A class of games possessing pure-strategy Nash equilibria. International Journal of Game Theory 2:65–67

[60] Samuelson L, Zhang J (1992) Evolutionary stability in asymmetric games. Journal of Economic Theory 57:363–391

[61] Sandholm WH (2010) Population Games and Evolutionary Dynamics. MIT Press, Cambridge, MA

[62] Scutari G, Facchinei F, Palomar DP, Pang JS (2010) Convex optimization, game theory, and variational inequality theory in multiuser communication systems. IEEE Signal Process Mag 27(3):35–49

[63] Spall JC (1992) Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Trans Autom Control 37(3):332–341

[64] Tatarenko T, Kamgarpour M (2019) Learning generalized Nash equilibria in a class of convex games. IEEE Trans Autom Control 64(4):1426–1439

[65] Taylor PD (1979) Evolutionarily stable strategies with two types of player. Journal of Applied Probability 16(1):76–83

[66] Tse D, Viswanath P (2005) Fundamentals of Wireless Communication. Cambridge University Press, Cambridge, UK

[67] Vovk VG (1990) Aggregating strategies. In: COLT '90: Proceedings of the 3rd Workshop on Computational Learning Theory, pp 371–383

[68] Zhang R, Ren Z, Li N (2021) Gradient play in multi-agent Markov stochastic games: Stationary points and convergence. https://arxiv.org/abs/2106.00198