Equilibrium tracking and convergence in dynamic games

Equilibrium tracking and convergence in dynamic games

Panayotis Mertikopoulos, Mathias Staudigl

Equilibrium tracking and convergence in dynamic games

Panayotis Mertikopoulos[⋆,‡] Mathias Staudigl[◦]

***Abstract*— In this paper, we examine the equilibrium tracking and convergence properties of no-regret learning algorithms in continuous games that evolve over time. Specifically, we focus on learning via "mirror descent", a widely used class of no-regret learning schemes where players take small steps along their individual payoff gradients and then "mirror" the output back to their action sets. In this general context, we show that the induced sequence of play stays asymptotically close to the evolving equilibrium of the sequence of stage games (assuming they are strongly monotone), and converges to it if the game stabilizes to a strictly monotone limit. Our results apply to both gradient- and payoff-based feedback, i.e., the "bandit" case where players only observe the payoffs of their chosen actions.**

I. Introduction

Consider the following multi-agent learning framework:

1) At each stage $t = 1, 2, \dots$ of a repeated decision process, every participating agent selects an action from some continuous set.
2) Each agent receives a reward based on their chosen action and the actions of all other players. These rewards are determined by a normal form game $\mathcal{G}_t$ that evolves over time and is a priori unknown to the players.
3) Based on the reward that they received and any other observed information, the players update their actions and the process repeats.

In this general setting, the main questions that we seek to address are as follows: *Are there online learning policies that allow players to track a Nash equilibrium over time* (*or converge to one if the stage games stabilize*)*? What is the impact of the information available to the players and the variability of the stage game sequence?*

In this regard, one of the most widely used policies for online learning is the *mirror descent* (MD) family of algorithms, cf. [1], [2], and references therein. This first-order scheme has a long history in optimization and contains as special cases the online gradient descent (OGD) policy of [3], the entropic/exponentiated gradient descent method of [4], and the *"Hedge"* (or exponential/multiplicative weights) algorithm for mixed-strategy learning in multi-armed bandits and finite games [5]–[7]. Importantly, when the payoff functions encountered by the learner are concave, MD methods guarantee an $\mathcal{O}(\sqrt{T})$ static regret bound which is well known to be order-optimal [8]; moreover, if the problem has a favorable geometry (e.g., when the learner's action set is a simplex or a spectrahedron), these bounds are "almost" dimension-free, a fact which is of crucial importance in practical applications.

In view of these desirable guarantees, methods based on mirror descent are also natural candidates for learning in *multi-agent*, game-theoretic environments [9]–[11]. However, the multi-agent case is considerably more involved because, in addition to the *exogenous* variability of the stage game $\mathcal{G}_t$, the individual payoff function of any given player also varies *endogenously* as a function of the actions chosen by all other players at time t. Moreover, in addition to this extra dimension of the problem, the standard figure of merit in game theory is that of a *Nash equilibrium* – not the players' regret. Thus, even though the learning algorithms under study remain essentially unchanged in single- and multi-agent settings, the type of results obtained in the literature are quite different.

Related work: Most of the literature on game-theoretic learning has focused on the case where the players encounter the same game at each stage – i.e., when there are no exogenous variations in the players' individual payoff functions. Starting with mixed-strategy learning in finite games, a "folk" result in the field states that the empirical frequency of play under concurrent no-regret play converges to the game's *Hannan set* (also known as the set of coarse correlated equilibria). However, as was shown by [12], the Hannan set of a game may contain strategies that assign positive weight *only* to dominated strategies – which, of course, cannot be supported in a Nash equilibrium. More to the point, the impossibility result of [13] shows that there are no uncoupled dynamics leading to Nash equilibrium in all games: since no-regret dynamics are uncoupled by construction, it is not possible to establish a blanket causal link between regret minimization and convergence to Nash equilibrium.

The work that is most relevant for our purposes is the recent paper [9] that focused on a class of continuous games satisfying the so-called *diagonal strict concavity* (DSC) condition of [14]. Using similar stochastic approximation techniques as above, [9] showed that the sequence of play generated by a class of mirror-based policies converges to Nash equilibrium with probability 1, even with imperfect gradient information on the players' side. Finally, in a very recent paper, [15] established the convergence of a mirror-like, *dampened gradient approximation* (DGA) scheme in two classes of one-dimensional concave games: games with strategic complements, and ordinal potential games with isolated equilibria. Interestingly, the continuous-time limit of both methods is the same; however, the method of [15] has the significant advantage that it requires only payoff-based feedback (which is in turn used to reconstruct individual payoff gradients by means of a two-

⋆ Univ. Grenoble Alpes, CNRS, Inria, LIG, 38000, Grenoble, France.
‡ Criteo AI Lab.
◦ Maastricht University, Department of Data Science and Knowledge Engineering, P.O. Box 616, NL–6200 MD Maastricht, The Netherlands

point estimation process).

***Our contributions**:* In contrast to the works discussed above, our paper seeks to tackle problems where the sequence of games encountered by the players also evolves *exogenously* over time – i.e., players encounter a *time-varying game*. Given their popularity, we focus throughout on a class of *generalized mirror descent* (GMD) policies and we consider two distinct regimes: (*a*) when the sequence of stage games converges to some well-defined limit (in our case, a strictly monotone game); and (*b*) when $\mathcal{G}_t$ evolves over time without converging.

In terms of feedback, we consider an agnostic oracle model which provides noisy payoff gradient estimates to the players based on the actions that they chose at each stage of the process. We then show that, if the sequence of stage games stabilizes to some well-defined limit, the induced sequence of play converges to a Nash equilibrium of the limit game with probability 1, irrespective of the magnitude of the noise entering the players' gradient signals. On the other hand, if the stage games do not stabilize, there is no equilibrium state to converge to (either static or in the mean); in this case, we focus instead on the players' ability to track the equilibrium of $\mathcal{G}_t$ as it evolves over time. More precisely, we show that the average distance from equilibrium at each stage vanishes over time, and we provide an explicit estimate for this "tracking error" in terms of the variation of the sequence of stage games (assuming they are strongly monotone).

Finally, to account for environments where gradient information is not available to the players, we also consider the case of learning with *payoff-based* feedback. By considering a one-shot gradient estimation process based on single-point stochastic approximation techniques [16]–[18], we map the problem of payoff-based learning to our generic gradient oracle model, and we show that our convergence and equilibrium tracking results still apply in this case (though the corresponding rates are reduced as a consequence of the players' having even less information at their disposal).

II. Preliminaries

A. Notation

Let $\mathcal{X}$ be an n-dimensional real space with norm $\|\cdot\|$, and let $\mathcal{K}$ be a compact convex subset of $\mathcal{X}$. Throughout the sequel, we will write $\mathcal{Y} = \mathcal{X}^*$ for the dual of $\mathcal{X}$, $\langle y, x\rangle$ for the duality pairing between $y \in \mathcal{Y}$ and $x \in \mathcal{X}$, and $\|y\|_* = \sup\{\langle y, x\rangle : \|x\| \leq 1\}$ for the dual norm of $y \in \mathcal{Y}$. We will also write $\operatorname{ri}(\mathcal{K})$ for the relative interior of $\mathcal{K}$, $\operatorname{bd}(\mathcal{K})$ for its boundary, and $\operatorname{diam}(\mathcal{K}) = \sup\{\|x' - x\| : x, x' \in \mathcal{K}\}$ for its diameter. Finally, for concision, we write $[a\,..\,b] = \{a, a+1, \dots, b\}$ for the interval of positive integers spanned by $a, b \in \mathbb{N}$.

B. Concave games

Throughout this paper, we will focus on games with a finite number of players and continuous action sets. Specifically, every player $i \in \mathcal{N} = \{1, \dots, N\}$ is assumed to select an *action* x_i from a compact convex subset $\mathcal{K}_i$ of a finite-dimensional normed space $\mathcal{X}_i$. Subsequently, based on each player's individual objective and the *action profile* $x = (x_i; x_{-i}) \equiv (x_1, \dots, x_N)$ of all players' actions, every player receives a *reward*, and the process repeats.

In more detail, writing $\mathcal{K} \coloneqq \prod_{i\in\mathcal{N}} \mathcal{K}_i$ for the game's *action space* and $\mathcal{X} \coloneqq \prod_{i\in\mathcal{N}} \mathcal{X}_i$ for its corresponding ambient space, we assume that each player's reward is determined by an associated *payoff* (or *utility*) *function* $u_i \colon \mathcal{K} \to \mathbb{R}$. We will denote this tuple as $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{K}, u)$.

Since players are not assumed to "know the game" (or even that they are involved in one) these payoff functions may be unknown to the players, especially with respect to the dependence on the actions of other players. Throughout the sequel, we will only make the following blanket assumption:

Assumption 1. The players' payoff functions are continuously differentiable and *individually concave*, i.e.,

$$u_i(x_i; x_{-i}) \text{ is concave in } x_i \tag{1}$$

for all $x_{-i} \in \mathcal{K}_{-i}$ and all $i \in \mathcal{N}$.

C. Solution concepts and Nash equilibrium

The most prevalent solution concept in game theory is that of a *Nash equilibrium* (NE), defined here as an action profile $x^* \in \mathcal{K}$ that is resilient to unilateral deviations, i.e.,

$$u_i(x_i^*; x_{-i}^*) \geq u_i(x_i; x_{-i}^*) \quad \text{for all } x_i \in \mathcal{K}_i \text{ and all } i \in \mathcal{N}. \tag{NE}$$

The set of Nash equilibria of $\mathcal{G}$ will be denoted throughout as $\mathcal{K}^* \coloneqq \operatorname{NE}(\mathcal{G})$.

By the individual concavity of the game's payoff functions, Nash equilibria can also be characterized via the first-order optimality condition

$$\langle v_i(x^*), x_i - x_i^*\rangle \leq 0 \quad \text{for all } x_i \in \mathcal{K}_i,\ i \in \mathcal{N}, \tag{2}$$

where $v_i(x)$ denotes the individual payoff gradient of the i-th player, i.e.,

$$v_i(x) = \nabla_{x_i} u_i(x_i; x_{-i}). \tag{3}$$

Geometrically, this variational characterization of Nash equilibria simply means that $v_i(x^*)$ forms an obtuse angle with any displacement vector of the form $z_i = x_i - x_i^*$, $x_i \in \mathcal{K}_i$. We will use this geometric intuition freely in what follows.

Starting with the seminal work of [14], much of the literature on continuous games has focused on problems where the vector field $v(x)$ of individual payoff gradients satisfies the monotonicity condition

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

Owing to the link between (MC) and the theory of monotone operators in optimization, games that satisfy (MC) are commonly referred to as *monotone games*. In particular, mirroring the corresponding terminology from operator theory, we will say that a game is:

1) *Strictly monotone* if (MC) holds as a strict inequality when $x' \neq x$.
2) *Strongly monotone* if there exists a positive constant $\alpha > 0$ such that

$$\langle v(x') - v(x), x' - x\rangle \leq -\alpha\|x' - x\|^2 \quad \text{for all } x, x' \in \mathcal{K}. \tag{4}$$

The set of Nash equilibria of a monotone game is itself convex and compact; in particular, if the game is strictly or strongly monotone, its Nash set is a singleton. Moreover, Nash equilibria of monotone games can also be characterized via the *Minty* variational inequality [9], [19]

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

This property of Nash equilibria of monotone games will play a crucial role in our analysis and we will use it freely in the sequel; for a detailed discussion, see [9], [19], [20].

III. Problem setup

We now turn to a detailed description of our model for time-varying games. From the viewpoint of a single agent, this can be captured by the following sequence of events:

Time-varying games: sequence of events

Require: players $i \in \mathcal{N}$, action spaces $\mathcal{K}_i \subseteq \mathbb{R}^{n_i}$

for $t = 1, 2, \dots,\ i = 1, \dots, N$ **do**
  choose $X_{i,t} \in \mathcal{K}_i$ # select pivot point
  each player receives $u_{i,t}(X_{i,t})$ # collect reward
  each player gets signal $V_{i,t}$ # receive feedback
end for

The core ingredients of the above framework are (*a*) the way that the players' payoff functions are determined by a sequence of games $\mathcal{G}_t$, $t = 1, 2, \dots$; and (*b*) the sequence of feedback signals V_t received by the players. We discuss these elements in detail below.

First, in terms of regularity, we will be assuming throughout that the players' individual payoff gradients satisfy the following regularity conditions:

Assumption 2. Let $v_t(x) = (v_{i,t}(x))_{i \in \mathcal{N}}$ denote the players' individual gradient profile for the game $\mathcal{G}_t$. Then there exist constants $G, L > 0$ such that

$$\|v_t(x)\|_* \leq G \tag{5a}$$

$$\|v_t(x') - v_t(x)\|_* \leq L\|x' - x\| \tag{5b}$$

for all $t = 1, 2, \dots$, and all $x, x' \in \mathcal{K}$.

Second, the feedback model that we will employ is that of a *stochastic first-order oracle* (SFO), i.e., a "black-box" mechanism that outputs a (possibly imperfect) measurement of each player's individual payoff gradient at the point where it was queried. More precisely, when called at $X_t = (X_{1,t}, \dots, X_{N,t}) \in \mathcal{K}$, an SFO returns a gradient signal $V_t = (V_{1,t}, \dots, V_{N,t})$ of the form

$$V_t = v_t(X_t) + Z_t \tag{SFO}$$

where the "observational error" Z_t captures all sources of uncertainty in the oracle.

In more detail, to differentiate between "random" (zero-mean) and "systematic" (non-zero-mean) errors in V_t, it will be convenient to decompose the error process Z_t as

$$Z_t = U_t + b_t \tag{6}$$

where U_t is zero-mean and b_t denotes the mean of Z_t. To define all this formally, we will subsume all sources of randomness in V_t in an abstract probability law $\mathbb{P}$. Since this randomness is generated *after* players select their actions, the process Z_t is, in general, not adapted to the history of X_t. More explicitly, writing $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ for the natural filtration of X_t, we set

$$b_t = \mathbb{E}[Z_t \mid \mathcal{F}_t] \qquad \text{and} \qquad U_t = Z_t - b_t \tag{7}$$

so, by definition, $\mathbb{E}[U_t \mid \mathcal{F}_t] = 0$.

In view of all this, the oracle feedback received by the players can be classified according to the following statistics:

1) *Bias:*
$$\|b_t\|_* \leq B_t. \tag{8a}$$

2) *Variance:*
$$\mathbb{E}[\|U_t\|_*^2 \mid \mathcal{F}_t] \leq \sigma_t^2. \tag{8b}$$

3) *Second moment:*
$$\mathbb{E}[\|V_t\|_*^2 \mid \mathcal{F}_t] \leq M_t^2. \tag{8c}$$

Finally, to simplify notation later on, we will also consider the "signal plus noise" error bound

$$S_t^2 = M_t^2 + \sigma_t^2. \tag{8d}$$

To streamline our presentation, we will first present our results in an abstract, model-agnostic manner, i.e., without specifying the origins of the oracle model (SFO); subsequently, in Section VI, we provide an explicit construction of such an oracle from payoff-based observations, and we discuss in detail what this entails for our analysis and results.

IV. Learning via mirror descent

The most widely used method for no-regret learning is the family of algorithms known as *online mirror descent* (OMD). Viewed abstractly, the basic idea of the method is as follows: each player $i \in \mathcal{N}$ plays an action $x_i \in \mathcal{K}_i$ and receives a gradient signal $v_i \in \mathcal{Y}_i$; subsequently, each player takes an "approximate gradient" step from x_i along v_i to generate a new action x_i^+, and the process repeats. Formally, this can be written in recursive form as

$$x_i^+ = \mathcal{P}_i(x_i, \gamma v_i) \tag{9}$$

where:

1) $\mathcal{P}_i$ denotes the "prox-mapping" of player i (discussed in detail below).
2) γ is a step-size parameter controlling the weight attributed to the signal v_i.

The "prox-scheme" (9) will be our main focus in the sequel so some remarks are in order:

To get some intuition about the method, the archetypal example of a prox-mapping is the Euclidean projector

$$\mathcal{P}(x, y) = \Pi(x + y) = \arg\min_{x' \in \mathcal{C}} \left\{ \langle y, x - x' \rangle + \tfrac{1}{2}\|x' - x\|_2^2 \right\} \tag{10}$$

i.e., the closest-point projection of $x + y$ onto a given convex $\mathcal{C}$. Going beyond this familiar example, the key novelty of mirror

descent is to replace the quadratic term in (10) by the so-called *Bregman divergence*

$$D(x', x) = h(x') - h(x) - \langle \nabla h(x), x' - x \rangle, \tag{11}$$

induced by a "*distance-generating function*" h on $\mathcal{C}$. More precisely, in the spirit of [21], we have the following definition:

Definition 1. Let $\mathcal{C}$ be a compact convex subset of $\mathcal{X} \cong \mathbb{R}^n$. We say that $h\colon \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is a *distance-generating function* (DGF) on $\mathcal{C}$ if

1) h is proper, lower semi-continuous (l.s.c.) and convex.
2) The effective domain of h is $\operatorname{dom} h := \{x \in \mathcal{X} : h(x) < \infty\} = \mathcal{C}$.
3) The subdifferential of h admits a *continuous selection*: specifically, writing
$$\mathcal{C}^\circ := \operatorname{dom} \partial h = \{x \in \mathcal{C} : \partial h(x) \neq \varnothing\} \tag{12}$$
for the domain of ∂h, there exists a continuous mapping $\nabla h\colon \mathcal{C}^\circ \to \mathcal{Y}$ such that $\nabla h(x) \in \partial h(x)$ for all $x \in \mathcal{C}^\circ$.
4) h is K-strongly convex relative to $\|\cdot\|$, i.e.,
$$h(x') \geq h(x) + \langle \nabla h(x), x' - x \rangle + \tfrac{K}{2}\|x' - x\|^2 \tag{13}$$
for all $x \in \mathcal{C}^\circ$ and all $x' \in \mathcal{C}$.

Given a DGF h on $\mathcal{C}$, the *Bregman divergence* $D\colon \mathcal{C}^\circ \times \mathcal{C} \to \mathbb{R}$ induced by h is given by (11), and the associated *prox-mapping* $\mathcal{P}\colon \mathcal{C}^\circ \times \mathcal{Y} \to \mathcal{C}^\circ$ is defined as

$$\mathcal{P}(x, y) = \operatorname*{arg\,min}_{x' \in \mathcal{C}} \{\langle y, x - x' \rangle + D(x', x)\} \tag{14}$$

Finally, we say that h is *Lipschitz* if $\sup_{x \in \mathcal{C}^\circ} \|\nabla h(x)\|_* < \infty$.

Going back to our multi-agent setting, let $\mathcal{G}_t \equiv \mathcal{G}_t(\mathcal{N}, \mathcal{K}, u_t)$ be a sequence of stage games, and assume each player $i \in \mathcal{N}$ is endowed with an individual DGF $h_i\colon \mathcal{K}_i \to \mathbb{R}$ and corresponding prox-mapping $\mathcal{P}_i\colon \mathcal{K}_i^\circ \times \mathcal{Y}_i \to \mathcal{K}_i^\circ$, where $\mathcal{K}_i^\circ := \operatorname{dom} \partial h_i$ and $\mathcal{Y}_i := \mathcal{Y}_i$. We then obtain the general class of *prox-learning* methods

$$X_{i,t+1} = \mathcal{P}_i(X_{i,t}, \gamma_t V_{i,t}), \tag{PL$_i$}$$

or, in more compact notation

$$X_{t+1} = \mathcal{P}(X_t, \gamma_t V_t) \tag{PL}$$

where:

1. $t = 1, 2, \dots$ denotes the stage of the process.
2. $X_t = (X_{i,t})_{i \in \mathcal{N}}$ denotes the players' action profile at time t.
3. $V_t = (V_{i,t})_{i \in \mathcal{N}}$ denotes the signals received by the players at stage t, assumed throughout to be provided by an oracle of the general form (SFO).
4. $\gamma_t > 0$ is a (nonincreasing) step-size sequence.
5. The *collective* (or aggregate) prox-mapping $\mathcal{P} := \prod_i \mathcal{P}_i\colon \mathcal{K}^\circ \times \mathcal{Y} \to \mathcal{K}^\circ$ is defined as $\mathcal{P}(x, y) = (\mathcal{P}_i(x_i, y_i))_{i \in \mathcal{N}}$ for all $x \in \mathcal{K}^\circ := \prod_i \mathcal{K}_i^\circ$ and all $y \in \mathcal{Y} := \prod_i \mathcal{Y}_i$.

Unless explicitly mentioned otherwise, all learning policies described in the sequel will be of the form above.

Algorithm 1: Prox-learning [player indices suppressed]

```
Require: sequence of stage games G_t, prox-mapping P,
    step-size γ_t > 0
initialize X_1 ← arg min h                  # initialization
for t = 1, 2, ... do
    play X_t ∈ K                            # select action
    get gradient signal V_t                 # oracle feedback
    set X_{t+1} ← P(X_t, γ_t V_t)           # update action
end for
```

V. Equilibrium tracking and convergence analysis

In this section, we return to the multi-agent viewpoint and we examine the players' long-run behavior in two distinct regimes: *a*) when the sequence of stage games $\mathcal{G}_t$ converges to some limit game $\mathcal{G} \equiv \mathcal{G}_\infty$; and *b*) when $\mathcal{G}_t$ evolves over time without converging. In both cases, we will treat the process defining the time-varying game as a "black box" and we will not scrutinize its origins in detail; we do so in order to focus on the interplay between the fluctuations of the sequence of stage games and the induced sequence of play.

A. Stabilization and convergence

We begin with the case where the sequence of games encountered stabilizes to some limit game $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{K}, u)$. Formally, it will be convenient to characterize this convergence in terms of the quantity

$$R_{i,t} = \max_{x \in \mathcal{K}} \|v_{i,t}(x) - v_i(x)\|_*, \tag{15}$$

i.e., via the maximum difference in the (unilateral) gradient field of the stage game $\mathcal{G}_t$ and the limit game $\mathcal{G}$. We then say that the sequence of games $\mathcal{G}_t$ converges to $\mathcal{G}$ if

$$R_t := \sum_{i \in \mathcal{N}} R_{i,t} \to 0 \quad \text{as } t \to 0. \tag{16}$$

The reason for defining the convergence of a sequence of games in terms of payoff gradients instead of payoff functions is twofold: First, if the payoff functions of a game are perturbed by arbitrary, player-specific constants, the game's equilibrium points will remain unchanged, but the corresponding payoff differences may be large (so $\|u_{i,t} - u_i\|$ may fail to converge to 0 as $t \to \infty$). Second, the first-order optimality condition (2) shows that a Nash equilibrium of a (concave) game can be seen as a solution of a variational inequality that only involves the players' individual payoff gradients – not their payoff functions per se. As such, characterizing the convergence of a sequence of stage games in terms of payoff gradients is closer to the true primitives that define the players' equilibrium behavior.

As in the previous section, we will focus on learning processes adhering to the basic template of Algorithm 1. However, since we are now interested in the convergence of the generated sequence of play to a *specific* point in $\mathcal{K}$, we will assume in what follows that the Bregman divergence (11) satisfies the *Bregman reciprocity* condition

$$x_k \to p \quad \text{whenever} \quad D(p, x_k) \to 0, \tag{BR}$$

for every sequence of actions $x_k \in \mathcal{K}^\circ$. This requirement is fairly standard in the "last iterate" analysis of mirror descent

algorithms – see e.g., [22], [9], and references therein. In particular, if h is Lipschitz, we have

$$D(p, x_k) \leq h(p) - h(x_k) + \|\nabla h(x_k)\|_* \|x_k - p\| = \mathcal{O}(\|x_k - p\|) \quad (17)$$

so (BR) always holds in that case. The converse to this condition holds automatically by the strong convexity of h; thus, taken together, strong convexity and Bregman reciprocity guarantee that $x_k \to p$ if and only if $D(p, x_k) \to 0$.

With all this in hand, we have the following equilibrium convergence result:

Theorem 1. *Let $\mathcal{G}_t$ be a time-varying game converging to a strictly monotone game $\mathcal{G}$. Assume further that each player runs Algorithm 1 with a distance-generating function satisfying (BR), feedback of the form (SFO), and a step-size sequence γ_t such that*

$$\sum_{t=1}^{\infty} \gamma_t = \infty, \quad \sum_{t=1}^{\infty} \gamma_t (R_t + B_t) < \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \gamma_t^2 \sigma_t^2 < \infty \quad (a.s.). \quad (18)$$

Then, with probability 1, the sequence of realized actions X_t converges to the (necessarily unique) Nash equilibrium x^* of $\mathcal{G}$.

Corollary 1. *Suppose that $\mathcal{G}_t$ stabilizes at a rate $R_t = \mathcal{O}(1/t^r)$ and the feedback received by the players has bias $B_t = \mathcal{O}(1/t^b)$ and variance $\sigma_t^2 = \mathcal{O}(t^{2s})$ for some $r, b, s \geq 0$. If Algorithm 1 is run with $\gamma_t \propto t^{-p}$, $1 \geq p > \max\{1 - r, 1 - b, 1/2 + s\}$, the induced sequence of play X_t converges to Nash equilibrium (a.s.).*

Corollary 2. *If Algorithm 1 is run with perfect oracle feedback and assumptions as above, taking $p \in (1 - r, 1]$ guarantees that X_t converges to Nash equilibrium with probability 1.*

Before discussing the proof of Theorem 1, some remarks are in order. First, the players of the game are not required to know the rate of convergence of $\mathcal{G}_t$ to $\mathcal{G}$:

focusing for simplicity on the case of an unbiased oracle with bounded variance, Corollary 1 shows that the Robbins-Monro step-size policy $\gamma_t = 1/t$ guarantees convergence to equilibrium as long as the game stabilizes at a sub-polynomial rate (i.e., $R_t = \mathcal{O}(1/t^r)$ for some $r > 0$). In fact, by including a logarithmic "failsafe" and running the algorithm with the slightly more conservative step-size policy $\gamma_t = 1/(t \log t)$, convergence is guaranteed even if the game stabilizes at a much slower, sub-logarithmic rate $R_t = \mathcal{O}(1/(\log t)^\varepsilon)$ for some $\varepsilon > 0$.

This observation highlights an important difference between regret minimization and convergence to equilibrium. On the one hand, a rapidly-decaying step-size policy is more robust in terms of convergence, as it guarantees convergence under the slowest possible stabilization rate of $\mathcal{G}_t$. On the other hand, a rapidly vanishing step-size may be suboptimal from the point of view of regret minimization, because it may incur higher regret. This disparity is due to the fact that a sequence of games that converges fast to a limit game is very different relative to a sequence of games that oscillates without converging at the same time-scale; this can be seen more clearly in Theorem 2 below.

Our proof strategy for Theorem 1 will be based on a two-pronged approach in the spirit of [18]: First, we show that the sequence of generated actions converges (a.s.) to a level set of the Bregman divergence $D(x^*, \cdot)$ relative to x^*; subsequently, we show that X_t cannot remain at uniformly positive distance away from x^* for all sufficiently large t. Combining these results will show that X_t can only converge to the zero-level set of the Bregman divergence, i.e., $\lim_{t\to\infty} X_t = x^*$.

We begin by establishing the convergence of the Bregman divergence of X_t:

Proposition 1. *With probability 1, the Bregman divergence $D(x^*, X_t)$ converges (a.s.) to a random variable D_∞ with $\mathbb{E}[D_\infty] < \infty$.*

Proof. Let $D_t := D(x^*, X_t)$ and decompose the oracle signal received by the players as

$$V_t = v_t(X_t) + b_t + U_t = v(X_t) + r_t + b_t + U_t, \quad (19)$$

where $r_t = v_t(X_t) - v(X_t)$. Then, by a straightforward calculation, we get:

$$\begin{aligned} D_{t+1} &\leq D_t + \gamma_t \langle V_t, X_t - x^* \rangle + \frac{\gamma_t^2}{2K} \|V_t\|_*^2 \\ &\leq D_t + \gamma_t \rho_t + \gamma_t \beta_t + \gamma_t \psi_t + \frac{\gamma_t^2}{2K} \|V_t\|_*^2 \end{aligned} \quad (20)$$

where we used the fact that $\langle v(X_t), X_t - x^* \rangle \leq 0$ (since x^* is a Nash equilibrium of the limit game $\mathcal{G}$), and we set respectively

$$\rho_t = \langle r_t, X_t - x^* \rangle, \quad (21a)$$
$$\beta_t = \langle b_t, X_t - x^* \rangle, \quad (21b)$$
$$\psi_t = \langle U_t, X_t - x^* \rangle. \quad (21c)$$

Now, by the definition (15) of R_t, we have

$$\rho_t = \langle r_t, X_t - x^* \rangle \leq \|X_t - x^*\| \cdot \|r_t\|_* \leq \operatorname{diam}(\mathcal{K}) R_t \quad (22)$$

and, similarly, $\beta_t \leq \operatorname{diam}(\mathcal{K}) B_t$. Therefore, conditioning on the history $\mathcal{F}_t$ of X_t up to stage t (inclusive) and taking expectations, we get:

$$\begin{aligned} \mathbb{E}[D_{t+1} \mid \mathcal{F}_t] &\leq \mathbb{E}\left[D_t + \gamma_t \rho_t + \gamma_t \beta_t + \gamma_t \psi_t + \frac{\gamma_t^2}{2K} \|V_t\|_*^2 \,\middle|\, \mathcal{F}_t \right] \\ &\leq D_t + \gamma_t \operatorname{diam}(\mathcal{K}) \cdot (R_t + B_t) + \frac{2\gamma_t^2}{K} A_t^2 \end{aligned} \quad (23)$$

where $A_t^2 = \|v(X_t)\|_*^2 + R_t^2 + B_t^2 + \sigma_t^2$. To proceed, rewrite (23) in more compact form as

$$\mathbb{E}[D_{t+1} \mid \mathcal{F}_t] \leq D_t + \varepsilon_t \quad (24)$$

where ε_t collects all terms other than D_t in the RHS of (23). Since $\mathcal{K}$ is compact, v is bounded, so $\sup_t \|v(X_t)\|_* < \infty$ surely. Hence, by the stated assumptions for γ_t, R_t, B_t and σ_t, we get

$$\sum_{t=1}^{\infty} \varepsilon_t = \sum_{t=1}^{\infty} \mathcal{O}(\gamma_t (R_t + B_t) + \gamma_t^2 + \gamma_t^2 \sigma_t^2) < \infty. \quad (25)$$

Consider now the auxiliary process $\zeta_t = D_{t+1} + \sum_{s=t+1}^{\infty} \varepsilon_s$. By Doob's (sub)martingale convergence theorem [23], it follows that ζ_t converges almost surely to some random variable ζ

that is itself finite (almost surely and in L^1). Since $D_t = \zeta_{t-1} - \sum_{s=t}^{\infty} \varepsilon_s$ and $\lim_{t\to\infty} \sum_{s=t}^{\infty} \varepsilon_s = 0$, we conclude that D_t converges (a.s.) to ζ, as was to be shown. ∎

Moving on, our next result shows that we can extract a subsequence of X_t that converges to a Nash equilibrium of $\mathcal{G}$:

Proposition 2. *With probability* 1*, there exists a* (*random*) *subsequence* X_{t_k} *of* X_t *which converges to* x^*.

The proof of Proposition 2 follows the same line of reasoning as a similar result in [9], so we omit the detailed proof. Instead, we proceed to state and prove our main Nash equilibrium convergence result:

Proof of Theorem 1. With probability 1, Proposition 2 shows that there exists a (possibly random) subsequence t_k such that $X_{t_k} \to x^*$. By the reciprocity condition (BR), this implies that $\liminf_{t\to\infty} D(x^*, X_t) = 0$ (a.s.). However, since $\lim_{t\to\infty} D(x^*, X_t)$ exists by Proposition 1 (also with probability 1), it follows that

$$\lim_{t\to\infty} D(x^*, X_t) = \liminf_{t\to\infty} D(x^*, X_t) = 0 \tag{26}$$

i.e., X_t converges to x^*. ∎

B. Nash equilibrium tracking

We now turn to time-varying games that evolve *without* converging. In this case, any notion of convergence for X_t is meaningless: there is no equilibrium state to converge to, either static or in the mean. As a result, we will focus instead on whether X_t is capable of "tracking" the game's equilibria as they evolve over time.

To that end, consider the *equilibrium tracking error*

$$\operatorname{err}(\mathcal{T}) := \sum_{t\in\mathcal{T}} \|X_t - x_t^*\|^2 = \sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{N}} \|X_{i,t} - x_{i,t}^*\|^2 \tag{27}$$

where $x_t^* \in \mathcal{K}_t^* := \operatorname{NE}(\mathcal{G}_t)$ denotes a Nash equilibrium of $\mathcal{G}_t$ and $\mathcal{T} = [\tau \,..\, T]$, $\tau, T \in \mathbb{N}$, denotes the window of play.[1] As before, if $\mathcal{T}$ is of the form $\mathcal{T} = [1 \,..\, T]$, we will simply write $\operatorname{err}(T)$ instead of $\operatorname{err}(\mathcal{T})$; by construction, if $\operatorname{err}(T)$ is small relative to T, the sequence of play X_t will be close to equilibrium for most of the horizon of play.

Clearly, if the variability of the stage games (and, in particular, of x_t^*) is too high, it is not possible to achieve a sublinear tracking error, even in the single-player case. To quantify this, we define below the game's *equilibrium variation* as

$$\mathrm{V}(T) := \sum_{t=1}^{T} \|x_{t+1}^* - x_t^*\|, \tag{28}$$

where $x_t^* \in \mathcal{K}_t^* := \operatorname{NE}(\mathcal{G}_t)$ denotes a Nash equilibrium of $\mathcal{G}_t$ and, as before, T denotes the horizon of play.[1] We will then say that the equilibrium variation of $\mathcal{G}_t$ is *tame* if

$$\mathrm{V}(T) = o(T) \quad \text{as } T \to \infty. \tag{29}$$

[1] There is a certain ambiguity here involved in the choice of $x_t^* \in \mathcal{K}_t^*$; this will not play a role in the sequel because we will focus on games with a unique equilibrium.

Remarkably, under this tame variability assumption, the prox-learning methods under study enjoy the following equilibrium tracking guarantee:

Theorem 2. *Let* $\mathcal{G}_t$ *be a sequence of strongly monotone games satisfying Assumptions 1 and 2. Assume further that each player runs Algorithm 1 with step-size* $\gamma_t \propto t^{-p}$, $p \in (0,1)$, *a Lipschitz distance-generating function, and feedback of the form* (SFO) *with* $B_t = \mathcal{O}(1/t^b)$ *and* $S_t^2 = \mathcal{O}(t^{2s})$ *for some* $b, s \geq 0$. *Then the players' tracking error is bounded as*

$$\mathbb{E}[\operatorname{err}(T)] = \mathcal{O}\Big(T^{1+2s-p} + T^{1-b} + T^{2p-2s}\,\mathrm{V}(T)\Big). \tag{30}$$

Corollary 3. *Suppose that the players' oracle feedback is unbiased and bounded in mean square* (*formally,* $b = \infty$, $s = 0$). *If the equilibrium variation of the game is* $\mathrm{V}(T) = \mathcal{O}(T^r)$ *for some* $r > 0$, *we have*

$$\mathbb{E}[\operatorname{err}(T)] = \mathcal{O}(T^{1-p} + T^{2p+r}). \tag{31}$$

In particular, if Algorithm 1 is run with $\gamma_t \propto t^{-(1-r)/3}$, *the players enjoy the bound*

$$\mathbb{E}[\operatorname{err}(T)] = \mathcal{O}(T^{\frac{2+r}{3}}). \tag{32}$$

Proof. Our proof strategy will be to leverage the gap minimization guarantees of Algorithm 1 together with a batch comparison idea due to [24]. Specifically, for the sake of the analysis (and only the analysis), we will first partition the horizon of play $\mathcal{T} = [1 \,..\, T]$ in m contiguous batches $\mathcal{T}_k$, $k = 1, \dots, m$, each of length Δ (except possibly the m-th one, which might be smaller). Then, we will proceed to establish the error bound (30) by linking $\operatorname{err}(\mathcal{T}_k)$ to $\operatorname{Gap}(\mathcal{T}_k) := \sum_{i\in\mathcal{N}} \operatorname{Gap}_i(\mathcal{T}_k)$ for all $k = 1, \dots, m = \lceil T/\Delta \rceil$.

More explicitly, take the batch length to be of the form $\Delta = \lceil T^q \rceil$ for some constant $q \in [0,1]$ to be determined later. In this way, the number of batches is $m = \lceil T/\Delta \rceil = \Theta(T^{1-q})$ and the k-th batch will be of the form $\mathcal{T}_k = [(k-1)\Delta + 1 \,..\, k\Delta]$ for all $k = 1, \dots, m-1$ (the value $k = m$ is excluded as the m-th batch might be smaller). Then, to bound the players' equilibrium tracking error within $\mathcal{T}_k$, note that, by the strong monotonicity property (4) for $\mathcal{G}_t$, we have

$$\begin{aligned}\alpha\|X_t - x_t^*\|^2 &\leq \langle v_t(X_t), x_t^* - X_t\rangle \\ &= \langle v_t(X_t), \hat{x} - X_t\rangle + \langle v_t(X_t), x_t^* - \hat{x}\rangle \end{aligned} \tag{33}$$

for every reference action profile $\hat{x} \in \mathcal{K}$ and all $t \in \mathcal{T}$. We thus obtain the batch bound

$$\begin{aligned}\alpha \operatorname{err}(\mathcal{T}_k) = \alpha \sum_{t\in\mathcal{T}_k} \|X_t - x_t^*\|^2 &\leq \sum_{t\in\mathcal{T}_k} \langle v_t(X_t), x_t^* - X_t\rangle \\ &= \sum_{t\in\mathcal{T}_k} \langle v_t(X_t), \hat{x} - X_t\rangle + \sum_{t\in\mathcal{T}_k} \langle v_t(X_t), x_t^* - \hat{x}\rangle \\ &\leq \operatorname{Gap}(\mathcal{T}_k) + \sum_{t\in\mathcal{T}_k} \langle v_t(X_t), x_t^* - \hat{x}\rangle, \end{aligned} \tag{34}$$

where, as a reminder, we set $\operatorname{Gap}(\mathcal{T}_k) := \sum_{i\in\mathcal{N}} \operatorname{Gap}_i(\mathcal{T}_k)$.

To proceed, pick a batch-specific reference action $\hat{x}_k \in \mathcal{K}$ for each $k = 1, \dots, m$, and replace $\hat{x}$ by $\hat{x}_k$ to get

$$C_k = \sum_{t\in\mathcal{T}_k} \langle v_t(X_t), x_t^* - \hat{x}_k\rangle, \tag{35}$$

for the last term of (34). A meaningful bound for C_k can then be obtained by taking $\hat{x}_k$ to be the (unique) Nash equilibrium of the first game encountered in the batch $\mathcal{T}_k$, i.e., setting $\hat{x}_k = x^*_{\min \mathcal{T}_k}$. Doing this, we obtain the series of estimates:

$$\begin{aligned} C_k &\leq \sum_{t\in\mathcal{T}_k} \|v_t(X_t)\|_* \cdot \|x_t^* - \hat{x}_k\| && \{\text{by Cauchy-Schwarz}\} \\ &\leq \sum_{t\in\mathcal{T}_k} G\|x_t^* - \hat{x}_k\| && \{\text{by Assumption 2}\} \\ &\leq G\Delta \max_{t\in\mathcal{T}_k} \|x_t^* - \hat{x}_k\| && \{\text{term-by-term bound}\} \\ &\leq G\Delta \sum_{t\in\mathcal{T}_k} \|x_{t+1}^* - x_t^*\| && \{\text{by definition of } \hat{x}_k\} \\ &= G\Delta\, \mathrm{V}(\mathcal{T}_k). && (36) \end{aligned}$$

Thus, plugging everything back in (34) and summing over all batches $k = 1, \dots, m$, we get the total bound

$$\mathbb{E}[\mathrm{err}(\mathcal{T})] \leq \frac{1}{\alpha}\,\mathbb{E}[\mathrm{Gap}(\mathcal{T})] + \frac{G\Delta}{\alpha}\,\mathrm{V}(\mathcal{T}). \tag{37}$$

With this estimate in hand, let $D_{\mathcal{K}} := \sup_{x,x'} D(x, x')$. Then, with γ_t decreasing, a straightforward regret calculation (that we omit for reasons of space) yields

$$\begin{aligned} \sum_{k=1}^m \mathbb{E}[\mathrm{Gap}(\mathcal{T}_k)] &\leq \sum_{k=1}^m \frac{2D_{\mathcal{K}}}{\gamma_{k\Delta}} + 2\,\mathrm{diam}(\mathcal{K}) \sum_{t=1}^T B_t + \frac{1}{2K}\sum_{t=1}^T \gamma_t S_t^2 \\ &= \mathcal{O}\left(\Delta^p \sum_{k=1}^m k^p + \sum_{t=1}^T t^{-b} + \sum_{t=1}^T t^{2s-p}\right) \\ &= \mathcal{O}\left(\Delta^p m^{1+p} + T^{1-b} + T^{1+2s-p}\right). \end{aligned} \tag{38}$$

Since $\Delta = \mathcal{O}(T^q)$ and $m = \mathcal{O}(T/\Delta) = \mathcal{O}(T^{1-q})$, we get

$$\Delta^p m^{1+p} = \mathcal{O}((m\Delta)^p\, m) = \mathcal{O}(T^{qp} T^{(1-q)(1+p)}) = \mathcal{O}(T^{1+p-q}). \tag{39}$$

In turn, this yields the error bound

$$\mathbb{E}[\mathrm{err}(T)] = \mathcal{O}\left(T^{1+p-q} + T^{1-b} + T^{1+2s-p} + T^{2p-2s}\,\mathrm{V}(T)\right). \tag{40}$$

The guarantee (30) then follows by tuning the batch size exponent q so as to balance the first and third terms in the above expression, i.e., by taking $q = 2p - 2s$. ∎

VI. Learning with payoff-based information

In this section, we proceed to examine a "payoff-based" learning scheme, i.e., a method that relies only on observations of the players' realized, in-game payoffs – the so-called "bandit setting" [1], [2]. The first step will be to introduce a payoff-based stochastic first-order oracle in the spirit of [18]; in our game-theoretic framework, this process can be implemented as follows:

1) At each stage $t = 1, 2, \dots$, every player decides on a *candidate action* $X_{i,t} \in \mathcal{K}_i$; this action is *not* played, but it is momentarily kept in memory.
2) Instead of playing $X_{i,t}$, each player draws a random perturbation direction $E_{i,t} \in \mathcal{E}_i \equiv \{\pm e_1, \dots, \pm e_{n_i}\}$ and plays the nearby action $\hat{X}_{i,t}$ defined as

$$\hat{X}_{i,t} = (1 - \delta_t/r_i)X_{i,t} + (\delta_t/r_i)(p_i + r_i\, E_{i,t}). \tag{41}$$

Algorithm 2: Payoff-based prox-learning

Require: step-size $\gamma_t > 0$; sampling radius $\delta_t > 0$; safety parameters $r_i > 0$, $p_i \in \mathrm{int}(\mathcal{K}_i)$

initial candidate $X_1 \leftarrow \arg\min h$ # initialization
for $t = 1, 2, \dots$ **do simultaneously** for all $i = 1, \dots, N$
  draw $E_{i,t}$ uniformly from $\{\pm e_1, \dots, \pm e_{n_i}\}$ # query direction
  play $\hat{X}_{i,t} = (1 - \delta_t/r_i)X_{i,t} + (\delta_t/r_i)(p_i + r_i E_{i,t})$ # select action
  receive $\hat{u}_{i,t} \equiv u_{i,t}(\hat{X}_{i,t}; \hat{X}_{-i,t})$ # get payoff
  set $V_{i,t} = (n_i/\delta_t)\,\hat{u}_{i,t} E_{i,t}$ # estimate gradient
  set $X_{t+1} \leftarrow \mathcal{P}(X_t, \gamma_t V_t)$ # update candidate action
end for

3) Players receive their corresponding payoffs $\hat{u}_{i,t} = u_{i,t}(\hat{X}_{i,t}; \hat{X}_{-i,t})$.
4) Each player estimates their individual payoff gradient as

$$V_{i,t} = \frac{n_i}{\delta_t}\hat{u}_{i,t}\, E_{i,t}, \tag{42}$$

where $\delta_t > 0$ is a variable "sampling radius" parameter and n_i is the dimensionality of the i-th player's action space.

In this way, the estimate $V_t = (V_{i,t})_{i=1,\dots,N}$ can be seen as a payoff-generated stochastic first-order oracle which can be coupled with Algorithm 1 to generate a new candidate action X_{t+1}. For a pseudocode implementation of the resulting policy, see Algorithm 2 above.

Remark VI.1. Throughout this section, we tacitly assume that the players' action spaces are convex bodies, i.e., they have nonempty topological interior. This assumption is only made for convenience: if this is not the case, it suffices to replace the basis vectors $\{\pm e_k\}$ with a basis of the affine hull of each player's action space and proceed in the same way. §

The first step in our analysis of Algorithm 2 consists of quantifying the statistics of the players' gradient estimation process:

Lemma 1. *The single-point stochastic approximation (SPSA) estimator* (42) *satisfies:*

$$\|\mathbb{E}[V_t \mid \mathcal{F}_t] - v_t(X_t)\|_* = \mathcal{O}(\delta_t), \tag{43a}$$

and

$$\mathbb{E}[\|V_t\|_*^2 \mid \mathcal{F}_t] = \mathcal{O}(1/\delta_t^2). \tag{43b}$$

The estimation arguments used in the proof of Lemma 1 are relatively straightforward, so we omit them here. Instead, we proceed to state our main result for the payoff-based learning policy outlined in Algorithm 2:

Theorem 3. *Let $\mathcal{G}_t$ be a time-varying game satisfying Assumptions 1 and 2. Suppose further that each player follows Algorithm 2 with a Lipschitz distance-generating function, variable step-size $\gamma_t \propto t^{-p}$, and sampling radius $\delta_t \propto t^{-q}$ for some $p, q \in (0, 1]$. Then:*

1) If $\mathcal{G}_t$ stabilizes to a strictly monotone game $\mathcal{G}$ at a rate $R_t = \mathcal{O}(1/t^r)$ and $p > \max\{1-r, 1-q, 1/2+q\}$, the induced

sequence of play $\hat{X}_t$ converges to the Nash equilibrium of $\mathcal{G}$ with probability 1.

2) *If $\mathcal{G}_t$ is strongly monotone and its drift is bounded as $\mathrm{V}(T) = \mathcal{O}(T^r)$ for some $r < 1$, the induced sequence of play $\hat{X}_t$ tracks the game's evolving equilibrium x_t^* as*

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}} \|\hat{X}_t - x_t^*\|^2\right] = \mathcal{O}\Big(T^{1+2q-p} + T^{1-q} + T^{2p-2q+r}\Big). \tag{44}$$

In particular, for $p = 3(1-r)/5$, $q = (1-r)/5$, we get the optimized bound:

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}} \|\hat{X}_t - x_t^*\|^2\right] = \mathcal{O}\Big(T^{\frac{4+r}{5}}\Big). \tag{45}$$

Theorem 3 combines two regimes: Part 1 treats time-varying games that stabilize to a well-defined limit, while Part 2 concerns the case where the game evolves without converging. This is in direct analogy to Theorems 1 and 2 for the case of generic SFO feedback.

Due to space constraints, we omit the detailed proof here. We only note that Part 1 of Theorem 3 implies that the sequence of play induced by Algorithm 2 in a *fixed* strictly monotone game $\mathcal{G}_t \equiv \mathcal{G}$ converges to Nash equilibrium with probability 1 as long as $p > \max\{1-q, 1/2+q\}$. In this way, we recover a recent result by [18] who used a different form of the SPSA estimator (42) to establish the convergence of payoff-based no-regret learning in *constant*, monotone games. It is also possible to undertake a finer analysis for the method's rate of convergence in the case where the limit game $\mathcal{G}$ is strongly monotone, but this lies beyond the scope of this work.

VII. Concluding remarks

Our equilibrium tracking and convergence results comprise a first step towards understanding the behavior of utility-maximizing agents in unknown, online environments where the top-down, "rationalistic" viewpoint of dynamic/stochastic games does not apply. Specifically, even though the standard rationality postulates do not hold in our setting (knowledge of the game being played, common knowledge of rationality, etc.), our analysis shows that learning based on mirror descent can still lead to equilibrium in dynamic environments. We find this property particularly appealing, as it provides an important link between online learning and the emergence of rational behavior in strategic environments that evolve over time.

There are many interesting points for future research. A particularly promising one is to bridge the gap between the step-size policies that guarantee an optimal equilibrium tracking error and the policies that guarantee convergence to a Nash equilibrium in the case where $\mathcal{G}_t$ stabilizes to a well-defined limit. Heuristically, these considerations are incompatible: when the rules of the game fluctuate constantly, players are better off using a slowly-varying step-size in order to adapt to the changing landscape; by contrast, when the game stabilizes, a rapidly decaying step-size is best suited to guarantee convergence to Nash equilibrium. Balancing these two objectives in an adaptive, context-agnostic manner is a rich and promising direction for future research.

References

[1] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations and Trends in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2011.

[2] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[3] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *ICML '03: Proceedings of the 20th International Conference on Machine Learning*, 2003, pp. 928–936.

[4] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[5] V. G. Vovk, "Aggregating strategies," in *COLT '90: Proceedings of the 3rd Workshop on Computational Learning Theory*, 1990, pp. 371–383.

[6] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Information and Computation*, vol. 108, no. 2, pp. 212–261, 1994.

[7] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," in *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, 1995.

[8] J. Abernethy, P. L. Bartlett, A. Rakhlin, and A. Tewari, "Optimal strategies and minimax lower bounds for online convex games," in *COLT '08: Proceedings of the 21st Annual Conference on Learning Theory*, 2008.

[9] P. Mertikopoulos and Z. Zhou, "Learning in games with continuous action sets and unknown payoff functions," *Mathematical Programming*, vol. 173, no. 1-2, pp. 465–507, January 2019.

[10] P. Mertikopoulos and M. Staudigl, "Convergence to Nash equilibrium in continuous games with noisy first-order feedback," in *CDC '17: Proceedings of the 56th IEEE Annual Conference on Decision and Control*, 2017.

[11] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[12] Y. Viossat and A. Zapechelnyuk, "No-regret dynamics and fictitious play," *Journal of Economic Theory*, vol. 148, no. 2, pp. 825–842, March 2013.

[13] S. Hart and A. Mas-Colell, "Uncoupled dynamics do not lead to Nash equilibrium," *American Economic Review*, vol. 93, no. 5, pp. 1830–1836, 2003.

[14] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave N-person games," *Econometrica*, vol. 33, no. 3, pp. 520–534, 1965.

[15] S. Bervoets, M. Bravo, and M. Faure, "Learning with minimal information in continuous games," *Theoretical Economics*, vol. 15, pp. 1471–1508, 2020.

[16] J. C. Spall, "A one-measurement form of simultaneous perturbation stochastic approximation," *Automatica*, vol. 33, no. 1, pp. 109–112, 1997.

[17] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: gradient descent without a gradient," in *SODA '05: Proceedings of the 16th annual ACM-SIAM Symposium on Discrete Algorithms*, 2005, pp. 385–394.

[18] M. Bravo, D. S. Leslie, and P. Mertikopoulos, "Bandit learning in concave N-person games," in *NeurIPS '18: Proceedings of the 32nd International Conference of Neural Information Processing Systems*, 2018.

[19] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, ser. Springer Series in Operations Research. Springer, 2003.

[20] R. Laraki, J. Renault, and S. Sorin, *Mathematical Foundations of Game Theory*, ser. Universitext. Springer, 2019.

[21] A. Juditsky, A. S. Nemirovski, and C. Tauvel, "Solving variational inequalities with stochastic mirror-prox algorithm," *Stochastic Systems*, vol. 1, no. 1, pp. 17–58, 2011.

[22] G. Chen and M. Teboulle, "Convergence analysis of a proximal-like minimization algorithm using Bregman functions," *SIAM Journal on Optimization*, vol. 3, no. 3, pp. 538–543, August 1993.

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

[24] O. Besbes, Y. Gur, and A. Zeevi, "Non-stationary stochastic optimization," *Operations Research*, vol. 63, no. 5, pp. 1227–1244, October 2015.