LEARNING IN TIME-VARYING GAMES

BENOIT DUVOCELLE$^\sharp$, PANAYOTIS MERTIKOPOULOS*,
MATHIAS STAUDIGL$^\sharp$, AND DRIES VERMEULEN$^\sharp$

Abstract. In this paper, we examine the long-term behavior of regret-minimizing agents in time-varying games with continuous action spaces. In its most basic form, (external) regret minimization guarantees that an agent's cumulative payoff is no worse in the long run than that of the agent's best fixed action in hindsight. Going beyond this worst-case guarantee, we consider a dynamic regret variant that compares the agent's accrued rewards to those of *any* sequence of play. Specializing to a wide class of no-regret strategies based on mirror descent, we derive explicit rates of regret minimization relying only on imperfect gradient obvservations. We then leverage these results to show that players are able to stay close to Nash equilibrium in time-varying monotone games – and even converge to Nash equilibrium if the sequence of stage games admits a limit.

1. Introduction

A key requirement for decision-making in unknown, non-stationary environments is the minimization of *regret:* no rational agent would want to realize in hindsight that the decision policy they employed was strictly inferior to a crude policy prescribing the same action at each stage. Depending on the context, this minimal worst-case guarantee admits several refinements. For starters, agents could tighten their baseline and, instead of comparing their accrued rewards to those of the best *fixed* action, they could employ more general "comparator sequences" that evolve over time. Moreover, if agents interact with one another and their rewards are determined by a fixed underlying mechanism – a *non-cooperative game* – there are much finer criteria that apply, chief among them being that of convergence to a Nash equilibrium.

Since real-world scenarios are rarely stationary and typically involve several interacting agents, both issues are of high practical relevance and should be treated in tandem. With this in mind, the central question that we seek to address in this paper can be stated as follows: *What is the long-run behavior of strategic agents*

$^\sharp$ Maastricht University, Department of Quantitative Economics, P.O. Box 616, NL–6200 MD Maastricht, The Netherlands.

* Univ. Grenoble Alpes, CNRS, Inria, LIG, F-38000, Grenoble, France.

E-mail addresses: b.duvocelle@maastrichtuniversity.nl, panayotis.mertikopoulos@imag.fr, m.staudigl@maastrichtuniversity.nl, d.vermeulen@maastrichtuniversity.nl.

2010 *Mathematics Subject Classification.* Primary: 90C25, 60H10; secondary: 90C33, 90C47.

Key words and phrases. Dynamic regret; Nash equilibrium; mirror descent; time-varying games.

The research of P. Mertikopoulos and M. Staudigl was partially supported by the COST Action CA16228 "European Network for Game Theory" (GAMENET). P. Mertikopoulos is also grateful for financial support by the French National Research Agency (ANR) under grant no. ANR–16–CE33–0004–01 (ORACLESS).

that adhere to a no-regret policy when the underlying game evolves over time in an unknown, unpredictable manner?

Our contributions and prior work. Our analysis revolves around two main axes, as outlined below:

Dynamic regret minimization. We begin with the so-called "unilateral setting", i.e., when an agent seeks to minimize their regret against a fixed (but otherwise arbitrary) stream of payoff functions. As a benchmark, we posit that the agent compares the rewards accrued by their chosen sequence of play to any other test sequence (as opposed to a fixed action). In particular, as a special case, this definition of regret also includes the agent's best *dynamic* policy in hindsight, i.e., the sequence of actions that maximizes the payoff function encountered at each stage of the process.

This measure of regret is considerably more ambitious than the standard definition of external regret (which only considers constant sequences as performance benchmarks). One of its antecedents is the notion of *shifting regret* which considers piecewise constant benchmark sequences and keeps track of the number of "shifts" relative to the horizon of play – see e.g., Cesa-Bianchi et al. [7]. Much closer in spirit is the dynamic regret definition of Besbes et al. [4] which takes as a benchmark the sequence of instantaneous payoff maximizers (individual best responses) of each stage; unfortunately however, it is *not* possible to achieve sublinear dynamic regret if this sequence varies arbitrarily over time. To counter this, Besbes et al. [4] introduced a restart procedure that amortizes a policy with no static regret over a sequence of time windows of increasing length. In so doing, they showed that if the "variation budget" of the stream of payoff functions encountered is sublinear,[1] the agent can indeed achieve no dynamic regret.

In view of this, our first step is to examine the applicability of this restart heuristic against arbitrary test sequences. To that end, we show in Section 4 that a carefully crafted restart procedure in the spirit of Besbes et al. [4] allows agents to achieve no dynamic regret relative to any slowly-varying test sequence (i.e., any test sequence whose total variation grows sublinearly with the horizon of play). Then, to obtain concrete bounds, we focus on a family of learning policies known as *online mirror descent* (OMD), a flexible meta-algorithm that includes as special cases the online gradient descent (OGD) method of Zinkevich [51], the multiplicative weights (MW) algorithm of Auer et al. [2], the matrix exponentiation schemes of Tsuda et al. [49] and Mertikopoulos et al. [25], and many others.[2] Specifically, building on the work of Besbes et al. [4], we show in Section 6 that the class of policies under consideration attains a regret minimization rate of $\mathcal{O}(T^{2/3}V_T^{1/3})$ relative to test sequences with variation at most V_T over T stages, even when the agent only has access to imperfect gradient observations.

This result essentially coincides with the bound obtained by Besbes et al. [4] for slowly-varying streams of payoff functions and should be contrasted to recent work by Hall and Willett [18], Jadbabaie et al. [21] and Shahrampour and Jadbabaie [44] who established dynamic regret bounds without a restart procedure. In particular, the very recent analysis of Shahrampour and Jadbabaie [44] provides a better

[1] Specifically, the "variation budget" of a sequence of payoff functions u^t, $t = 1, 2, \dots T$, is defined as $\mathrm{VB}_T = \sum_{t=1}^{T-1} \|u^{t+1} - u^t\|_\infty$.

[2] In the above, "descent" should really be "ascent", because players are typically maximizers in game theory. To avoid this clash in terminology, we use the more neutral term "proximal method".

dependence on the horizon of play T ($T^{1/2}$ instead of $T^{2/3}$), but a worse dependence on the variation of the comparator sequence ($V_T^{1/2}$ instead of $V_T^{1/3}$). This suggests that *restarting is more advantageous in environments with higher variability:* this is an important observation that we encounter again (in a different guise) in the game-theoretic analysis of Section 7.

Game-theoretic learning. The second element of our analysis concerns the underlying assumption that the sequence of payoff functions encountered by the player is *oblivious*. Concretely, this means that, when calculating the payoffs that the agent would have obtained by employing a different sequence of actions, the stream of payoff functions encountered by the agent remains the same. This assumption is well-grounded in the literature of (adversarial) online optimization as a minimal requirement; however, in a game-theoretic setting, it is considerably more difficult to justify. For instance, if two regret-minimizing players are involved in a game, the payoff functions encountered by one player will be influenced by the action choices of the other. Thus, if one player were to employ a different sequence of actions, the other player would most likely respond differently, altering in this way the sequence of payoff functions encountered by the first player (and vice versa). As such, in a game-theoretic setting, even *dynamic* regret minimization against a given stream of payoff functions does not provide any equilibration guarantees.

To address this issue, we consider a general multi-agent framework where, at every stage $t = 1, 2, \dots,$ each player's payoff function is determined by the action choices of all other players via a non-cooperative game $\mathcal{G}^t$. The stage game $\mathcal{G}^t$ may vary with time, but the rules governing its evolution are not a priori known to the players (so the rationalistic viewpoint of the literature on repeated/dynamic games does not apply). In this context, the main question we seek to address is as follows: *If all players adhere a dynamic regret minimization policy, do their actions eventually track a Nash equilibrium of the stage game?*

Without further assumptions, the answer to this question is “no”, even when players face the same stage game throughout. Indeed, (external) regret minimization in finite games guarantees that the players' empirical frequencies of play converge to the game's *Hannan set* (also known as the set of coarse correlated equilibria). However, as was recently shown by Viossat and Zapechelnyuk [50], this set may contain strategies that assign positive weight *only* to dominated strategies (and these cannot be supported at a Nash equilibrium). In fact, in two-player zero-sum games, the analysis of Mertikopoulos et al. [26] shows that no-regret learning may cycle indefinitely without converging, always remaining a bounded distance away from the game's Nash set. On the other hand, if the game satisfies a monotonicity condition known as *diagonal strict concavity* (DSC), Mertikopoulos and Zhou [30] showed that no-regret policies based on mirror descent converge to Nash equilibrium with probability 1, even with imperfect gradient information on the players' side.

Building on all this, we first prove that if *a*) the stage games are monotone; and *b*) they admit a slowly-varying sequence of Nash equilibria, no-regret learning with a judiciously chosen restart schedule allows players to remain close to the game's evolving equilibrium (at least on average). More to the point, as a refinement of this result, we show that if the sequence of stage games converges to a strictly monotone game, the induced sequence of play converges to a Nash equilibrium thereof. Importantly, this last result holds globally (i.e., independently of the algorithm's initialization) and with probability 1, irrespective of the magnitude of the noise entering the players' gradient signals.

In our view, these results constitute a first step towards understanding the behavior of utility-maximizing agents in unknown, online environments where the top-down, "rationalistic" viewpoint of the theory of repeated and dynamic 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 results show that no-regret learning can still lead to equilibrium in dynamic environments. We find this property of regret minimization particularly appealing, as it provides an important link between online learning and the emergence of rational behavior in strategic environments that evolve over time.

Notation. Given a finite-dimensional vector space $\mathcal{V}$ with norm $\|\cdot\|$, we will write $\mathcal{V}^*$ for its (algebraic) dual, $\langle y, x\rangle$ for the duality pairing between $y \in \mathcal{V}^*$ and $x \in \mathcal{V}$, and $\|y\|_* = \sup\{\langle y, x\rangle : \|x\| \leq 1\}$ for the dual norm of $y \in \mathcal{V}^*$. If $\mathcal{X}$ is a closed convex subset of $\mathcal{V}$, we write $\operatorname{ri}(\mathcal{X})$ for its relative interior and $\operatorname{diam}(\mathcal{X}) = \sup\{\|x' - x\| : x, x' \in \mathcal{X}\}$ for its diameter. Finally, if x^t, $t = 1, 2, \dots$, is a sequence of elements of $\mathcal{X}$, we will write $x^{\mathcal{T}} \equiv (x^t)_{t\in\mathcal{T}}$ for the subfamily of elements indexed by a subset $\mathcal{T}$ of $\mathbb{N}$.

2. Preliminaries

2.1. Concave games.

The focal point of our analysis will be games with a finite number of players and continuous action sets. Specifically, every player $i \in \mathcal{N} \equiv \{1, \dots, N\}$ is assumed to select an *action* x_i from a compact convex subset $\mathcal{X}_i$ of a finite-dimensional normed space $\mathcal{V}_i$. 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 $\mathcal{X} \equiv \prod_{i\in\mathcal{N}} \mathcal{X}_i$ for the game's *action space* and $\mathcal{V} \equiv \prod_{i\in\mathcal{N}} \mathcal{V}_i$ for its corresponding ambient space,[3] we assume that each player's reward is determined by an associated *payoff* (or *utility*) *function* $u_i \colon \mathcal{X} \to \mathbb{R}$.[4] Since players are not assumed to "know the game" (or even that they are involved in one) these payoff functions might be a priori unknown, especially with respect to the dependence on the actions of other players. Following Rosen [40], our only blanket assumption for u_i will be that

$$u_i(x_i; x_{-i}) \textit{ is concave in } x_i \textit{ for all } x_{-i} \in \mathcal{X}_{-i}, \tag{2.1}$$

where, in obvious notation, $\mathcal{X}_{-i} = \prod_{j\neq i} \mathcal{X}_j$ denotes the action space of all players other than the i-th one. For regularity purposes, it will also be convenient (albeit not necessary) to assume that each u_i is C^1-smooth in x; to streamline our presentation, these will be our standing assumptions in what follows.

With all this in hand, a *concave game* will be a tuple $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{X}, u)$ with players, action spaces and payoffs defined as above. Below, we briefly discuss some recurring examples of such games:

[3] Unless explicitly mentioned otherwise, we will assume that $\mathcal{V}$ is endowed with the norm $\|x\|^2 = \sum_i \|x_i\|^2$. Also, to streamline our presentation, we will use the same notation for the norm of each factor space $\mathcal{V}_i$ and rely on the context to resolve any ambiguities.

[4] For book-keeping reasons, it will be convenient to assume that u_i is actually defined on an open neighborhood of $\mathcal{X}$ in $\mathcal{V}$. However, none of our calculations depend on this device, so we do not make this assumption explicit.

Example 2.1 (Mixed extension of finite games). In *finite games*, each player $i \in \mathcal{N}$ chooses an action (or *pure strategy*) α_i from a finite set $\mathcal{A}_i$. The players' payoffs are then determined by the pure strategy profile $\alpha = (\alpha_i)_{i\in\mathcal{N}}$ of all players' actions via a collection of payoff functions $u_i \colon \mathcal{A} \equiv \prod_j \mathcal{A}_j \to \mathbb{R}$.

In the *mixed extension* of a finite game, players are allowed to randomize their decisions by playing *mixed strategies*, i.e., probability distributions $x_i \in \Delta(\mathcal{A}_i)$ with the interpretation that $x_{i\alpha_i}$ represents the probability of choosing action $\alpha_i \in \mathcal{A}_i$. In this case (and in a slight abuse of notation), the expected payoff to player i under the mixed strategy profile $x = (x_i)_{i\in\mathcal{N}}$ is

$$u_i(x_i; x_{-i}) = \sum_{\alpha_1\in\mathcal{A}_1} \cdots \sum_{\alpha_N\in\mathcal{A}_N} u_i(\alpha_1, \dots, \alpha_N)\, x_{1,\alpha_1} \cdots x_{N,\alpha_N}.$$

Since each player's mixed strategy space $\mathcal{X}_i = \Delta(\mathcal{A}_i)$ is convex and u_i is individually linear in x_i, we see that mixed extensions of finite games are concave in the sense of (2.1).

Example 2.2 (Saddle-point problems). Consider a saddle-point problem of the general form

$$\min_{x_1\in\mathcal{X}_1} \max_{x_2\in\mathcal{X}_2} f(x_1, x_2) \tag{SP}$$

where each feasible region $\mathcal{X}_i$, $i = 1, 2$, is a compact convex subset of $\mathcal{V}_i \equiv \mathbb{R}^{d_i}$ and $f \colon \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$ is assumed to be convex in x_1 and concave in x_2. Letting $u_1 = -f$ and $u_2 = f$, the saddle-point problem (SP) can be seen as a zero-sum game with player set $\mathcal{N} = \{1, 2\}$ and payoff functions u_i, $i = 1, 2$. Since f is convex-concave, the resulting game $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{X}, u)$ is itself concave in the sense of (2.1).

Example 2.3 (Resource allocation auctions). Consider a service provider with a splittable *resource* (bandwidth, computing cores, ad display time, etc.). Fractions of this resource can be leased to a set of N bidders (players) who can place monetary bids $x_i \geq 0$ for the utilization of said resource up to each player's total budget b_i. Once all bids are in, resources are allocated proportionally to each player's bid, i.e., the i-th player gets $\rho_i = (qx_i)\big/(c + \sum_{j\in\mathcal{N}} x_j)$ units of the auctioned resource, with q denoting the total amount of the resource and $c \geq 0$ representing an "entry barrier" for bidding on it. A simple model for the utility of player i is then given by

$$u_i(x_i; x_{-i}) = g_i \rho_i - x_i,$$

where g_i denotes the marginal gain of player i from acquiring a unit slice of resources. Writing $\mathcal{X}_i = [0, b_i]$ for the action space of player i, it is easy to see that the resulting game $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{X}, u)$ is concave in the sense of (2.1).

Many other important scenarios can be formulated as concave games; for an incomplete list, see Kannan and Shanbhag [23], Facchinei et al. [14], Orda et al. [37], Sorin and Wan [46], Mertikopoulos et al. [25], and references therein.

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

$$u_i(\hat{x}_i; \hat{x}_{-i}) \geq u_i(x_i; \hat{x}_{-i}) \quad \text{for all } x_i \in \mathcal{X}_i \text{ and all } i \in \mathcal{N}. \tag{NE}$$

By the classical existence theorem of Debreu [12], concave games with compact action spaces always admit a Nash equilibrium. Moreover, thanks to the individual concavity of the game's payoff functions, Nash equilibria can also be characterized via the first-order optimality condition

$$\langle v_i(\hat{x}), x_i - \hat{x}_i \rangle \leq 0 \quad \text{for all } x_i \in \mathcal{X}_i,\ i \in \mathcal{N}, \tag{2.2}$$

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

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

and ∇_{x_i} denotes differentiation with respect to the variable x_i.[5]

Geometrically, this characterization of Nash equilibria simply means that $v_i(\hat{x})$ belongs to the polar cone

$$\mathrm{PC}_{\mathcal{X}_i}(\hat{x}_i) = \{y_i \in \mathcal{V}_i^* : \langle y_i, x_i - \hat{x}_i \rangle \leq 0 \text{ for all } x_i \in \mathcal{X}_i\}.$$

of $\mathcal{X}_i$ at $\hat{x}_i$, i.e., $v_i(\hat{x})$ forms an obtuse angle with any displacement vector of the form $z_i = x_i - \hat{x}_i$, $x_i \in \mathcal{X}_i$. By concavity, this means that $u_i(\hat{x}_i + tz_i; \hat{x}_{-i})$ is nonincreasing in t, so (NE) holds for all $x_i \in \mathcal{X}_i$. We will use this geometric intuition freely in what follows.

2.3. Variational inequalities and monotonicity.

The first-order characterization (2.2) of Nash equilibria can be written more concisely (but otherwise equivalently) as a variational inequality of the form

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

where

$$v(x) = (v_1(x), \dots, v_N(x))$$

denotes the players' individual gradient profile at $x \in \mathcal{X}$. As a result, finding a Nash equilibrium of a concave game boils down to solving the (Stampacchia) variational inequality problem (VI). This important observation has been the starting point of an extensive literature at the interface of game theory and optimization; for an overview, we refer the reader to Nikaido and Isoda [36], Facchinei and Pang [15], Scutari et al. [43], Mertikopoulos and Zhou [30], and references therein.

Most of this literature has focused on problems where the vector field $v(x)$ of individual payoff gradients satisfies the monotonicity condition

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

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* – see e.g., Scutari et al. [43], Mertikopoulos and Zhou [30], and references therein.[6] In particular, mirroring the corresponding terminology from operator theory, we will say that a game is:

a) *Strictly monotone* if (MC) holds as a strict inequality when $x' \neq x$.

[5] We adopt here the established convention of treating $v_i(x)$ as an element of the dual space $\mathcal{V}_i^*$ of $\mathcal{V}_i$. We do so in order to emphasize the fact that $v_i(x)$ acts naturally on vectors $z_i \in \mathcal{V}_i$ via the (linear) directional derivative mapping $z_i \mapsto u_i'(x; z_i) = d/dt|_{t=0}\, u_i(x_i + tz_i; x_{-i})$.

[6] Rosen [40] uses the name *diagonal strict concavity* (DSC) for a weighted variant of (MC) which holds as a strict inequality when $x' \neq x$. Hofbauer and Sandholm [20] use the term "stable" to refer to a class of population games that satisfy a condition similar to (MC), while Sandholm [42] and Sorin and Wan [46] respectively call such games "contractive" and "dissipative". We use the term "monotone" throughout to underline the connection of (MC) with operator theory and variational inequalities.

b) *Strongly monotone* if there exists a positive constant $\beta > 0$ such that

$$\langle v(x') - v(x), x' - x \rangle \leq -\beta \|x' - x\|^2 \quad \text{for all } x, x' \in \mathcal{X}. \tag{2.3}$$

Obviously, we have the inclusions "strongly monotone" $\subsetneq$ "strictly monotone" $\subsetneq$ "monotone", mirroring the corresponding chain of inclusions "strongly convex" $\subsetneq$ "strictly convex" $\subsetneq$ "convex" for convex functions.

The set of Nash equilibria of a monotone game – which coincides with the solution set of (VI) – is itself convex and compact; in particular, if the game is strictly or strongly monotone, its Nash set is a singleton. Moreover, on account of (MC), Nash equilibria of monotone games can also be characterized in this case as solutions of the Minty variational inequality

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

This property of Nash equilibria of monotone games will play a crucial role in our analysis and we will make free use of it in the rest of our paper; for a more detailed discussion, we refer the reader to Mertikopoulos and Zhou [30].

In terms of applications, monotone games constitute a very rich and diverse class. For instance, Examples 2.2 and 2.3 are both monotone (see Nemirovski et al. [32] for a proof), as are Cournot oligopoly models (Monderer and Shapley [31]), atomic splittable congestion games in networks with parallel links (Sorin and Wan [46]), signal covariance optimization problems in wireless communications (Mertikopoulos et al. [25]), and many other problems where online decision-making is the norm. In particular, the class of monotone games contains all games that admit a (strictly) concave *potential*, i.e., a function $f\colon \mathcal{X} \to \mathbb{R}$ such that

$$v_i(x) = \nabla_{x_i} f(x) \quad \text{for all } x \in \mathcal{X},\ i \in \mathcal{N}.$$

In view of all this, monotone games will comprise a key part of our analysis, especially in the context of convergence to Nash equilibrium.

3. Problem setup

We now turn to a detailed description of our model for time-varying games. In its most general form, this can be captured by the following sequence of events:

Time-varying games: sequence of events

```
Require: set of players $i \in \mathcal{N}$, action spaces $\mathcal{X}_i \subseteq \mathbb{R}^{d_i}$
for $t = 1, 2, \dots$ do
    set $\mathcal{G} \leftarrow \mathcal{G}^t(\mathcal{N}, \mathcal{X}, u^t)$          # stage game
    for all $i \in \mathcal{N}$ do
        play $X_i^t \in \mathcal{X}_i$                            # choose action
        receive $u_i^t(X_i^t; X_{-i}^t)$                   # collect reward
        get signal $Y_i^t$                             # receive feedback
    end for
end for
```

The core ingredients of the above framework are *a*) the sequence of games $\mathcal{G}^t$, $t = 1, 2, \dots$, encountered by the players at each stage of the process; and *b*) the sequence of feedback signals Y^t. We discuss both in detail below.

3.1. **The stage game sequence.** Our standing assumptions for the sequence of stage games $\mathcal{G}^t \equiv \mathcal{G}^t(\mathcal{N}, \mathcal{X}, u^t)$ will be that *a*) they are concave (in the sense of Section 2.1); and *b*) only the players' payoff functions evolve over time. Two important special cases of this framework are when:

(1) There is a single player and the sequence of stage games is fixed in advance (but is otherwise arbitrary). This unilateral framework is the gold standard in *online learning* (cf. the various definitions of regret in the next section) and is a priori *oblivious*, i.e., a different sequence of play would yield the same sequence of payoff functions. This is because, in contrast to the literature on stochastic games, the sequence $\mathcal{G}^t$ is typically assumed given – albeit unknown – at the outset of the game.

(2) The sequence of stage games is *constant*, i.e., $\mathcal{G}^t = \mathcal{G}$ for some fixed game $\mathcal{G}$. This case is the norm in *game-theoretic learning* and, in addition to comprising several players, its main difference with the online learning framework is that, from a unilateral standpoint, the sequence of generated payoff functions is *not* oblivious. In general, given the dependence of the payoff function of player i on the actions of all other players, a different sequence of actions $X^t \equiv (X_i^t; X_{-i}^t)$ would yield a different sequence of payoff functions $u_i(\cdot, X_{-i}^t)$.

In view of the above, a time-varying game can be seen as an amalgamation of these two classical frameworks: in particular, the dependence of a player's payoff function on the stage index t is both *explicit* (via the sequence of stage games $\mathcal{G}^t$ and *implicit* (via the sequence of actions chosen by all other players). This "dual" dependence on t will play a key role in what follows, especially in the equilibrium analysis of Section 7.

Remark 1. We should note here that the set of players $\mathcal{N}$ and their action spaces $\mathcal{X}_i$, $i \in \mathcal{N}$, are formally assumed to remain unchanged for all t. However, this is not necessarily so: for instance, if the payoff function u_i^t of some player $i \in \mathcal{N}$ is identically equal to zero at stage t and the actions of player i have no impact on the payoff function u_j^t of any other player $j \in \mathcal{N}$, the i-th player is effectively removed from the game at stage t. As a result, the proposed time-varying game model is flexible enough to account for games with a variable number of players, a case which has significant interest for practical applications of game theory (e.g., in networks and data science).[7]

In terms of regularity, we will be assuming throughout that the players' individual payoff gradients are uniformly bounded, i.e., there exists some finite $G_i \geq 0$ such that

$$\|v_i^t(x)\|_{i,*} \leq G_i \quad \text{for all } t = 1, 2, \dots, \text{ and all } x \in \mathcal{X}, \tag{3.1}$$

where, in obvious notation, we have set

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

Other than that, we make no prior assumptions about the process that defines each stage game. For instance, this evolution could be random (i.e., $\mathcal{G}^t$ could be determined by some randomly drawn parameter θ^t), it could depend on the players' actions (e.g., as in the literature on dynamic/repeated games), or any other

[7] Similar devices can also account for action spaces that vary with time (at least, as long as they are contained in some compact set).

mechanism. Moreover, we do not assume that such information is available to the players: from their individual viewpoint, each player is involved in a repeated decision process where the choice of an action returns a reward, and they have no knowledge of the mechanism generating this reward. The reason for this "agnostic" approach is that, in many cases of practical interest, the standard rationality postulates (full rationality, common knowledge of rationality, etc.) are not realistic: for instance, a commuter choosing a route to work has no way of knowing how many commuters will be making the same choice, let alone how these choices might influence their thinking for the next day.

3.2. **Signals and feedback.** The other basic ingredient of our model is the feedback available to each player after choosing an action. In tune with the "bounded rationality" framework outlined above, we do not assume that players can observe the actions of other players, their payoffs, or any other such information. Instead, we take a "partial monitoring" approach as in Rustichini [41], Cesa-Bianchi et al. [9] and Lugosi et al. [24], and we only posit that, at each stage $t = 1, 2, \dots$, every player $i \in \mathcal{N}$ receives a (random) signal Y_i^t from some space containing payoff-relevant information. In particular, we will be assuming that the random signal received by player i at stage t is of the general form

$$Y_i^t = v_i^t(X^t) + U_i^t, \tag{3.2}$$

where $X^t = (X_i^t; X_{-i}^t) \in \mathcal{X}$ is the profile of actions at stage t (possibly random), and U_i^t is a stochastic perturbation of the realized payoff gradient, modeling observational noise in the feedback signal. As such, under this model, the signals of player $i \in \mathcal{N}$ are drawn from the dual space $\mathcal{Y}_i \equiv \mathcal{V}_i^*$ of the ambient space $\mathcal{V}_i$ of $\mathcal{X}_i$.

Remark 2. In optimization-theoretic terms, the signal model (3.2) means that each player has access to a (stochastic) *first-order oracle*, i.e., a black-box feedback mechanism providing (possibly noisy) gradient information at each stage. From a game-theoretic standpoint, the motivation for this signal model is rooted in the case where players can only observe their realized, in-game payoffs (the so-called *bandit* setting). In this extremely low-information environment, it is still possible to construct an oracle of the form (3.2) by means of a simultaneous perturbation stochastic approximation (SPSA) procedure as in Spall [47], Flaxman et al. [16] and Bravo et al. [5]; we defer the details of this analysis to a future paper.

In terms of measurability, it is tacitly assumed that both $X^t = (X_i^t)_{i \in \mathcal{N}}$ and $Y^t = (Y_i^t)_{i \in \mathcal{N}}$ are defined over a common (complete) probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and all expectations or probabilities will be taken with reference to this space. The *private history* of player i is then defined as the filtration $\mathbb{F}_i = (\mathcal{F}_i^t)_{t=0}^\infty$, where

$$\mathcal{F}_i^t = \sigma(X_i^1, Y_i^1, \dots, X_i^t, Y_i^t)$$

is the σ-algebra generated by the player's chosen actions and signals received up to stage t (inclusive), while $\mathcal{F}_i^0$ is chosen so as to complete the filtration (not necessarily in a trivial way). Aggregating over all players, the *history of play* is likewise defined as the filtration $\mathbb{F} = (\mathcal{F}^t)_{t=0}^\infty$, where

$$\mathcal{F}^t = \sigma(X^1, Y^1, \dots, X^t, Y^t).$$

Given all this, we posit that a player's action at stage $t + 1$ is determined by the player's private history up to stage t, i.e.,

$$X_i^{t+1} = \mathsf{Alg}_i(X_i^1, Y_i^1, \dots, X_i^t, Y_i^t) \tag{3.3}$$

for some measurable deterministic function Alg_i, which will be referrred to as an *algorithm* (or *repeated game strategy*).[8] This means that, for all $t = 1, 2, \dots$, X_i^{t+1} is $\mathcal{F}_i^t$-predictable – or, collectively, that X^{t+1} is $\mathcal{F}^t$-predictable.

In the rest of our paper, and unless explicitly mentioned otherwise, our blanket assumptions for the signal process Y_i^t will be as follows:

(1) Y_i^t is a systematically unbiased estimate of $v_i^t(X^t)$, i.e.,

$$\mathbb{E}[Y_i^t \,|\, \mathcal{F}^{t-1}] = v_i^t(X^t)$$

for all $t = 1, 2, \dots$ and all $i \in \mathcal{N}$.

(2) Y_i^t has uniformly bounded second-order moments, i.e.,

$$\mathbb{E}[\|Y_i^t\|_*^2 \,|\, \mathcal{F}^{t-1}] \leq M_i^2 \tag{3.4a}$$

for some $M_i < \infty$ and all $t = 1, 2, \dots$, $i \in \mathcal{N}$.

Alternatively, the above is equivalent to asking that the noise process U_i^t is zero-mean with finite mean square, i.e.,

$$\mathbb{E}[U_i^t \,|\, \mathcal{F}^{t-1}] = 0$$

and

$$\mathbb{E}[\|U_i^t\|_*^2 \,|\, \mathcal{F}^{t-1}] \leq \sigma_i^2 \tag{3.5a}$$

for some finite $\sigma_i < \infty$ and all $t = 1, 2, \dots$, $i \in \mathcal{N}$. Both of these assumptions can be relaxed in various ways (e.g., by asking that Y_i^t is accurate on average only up to some bias term, or by considering higher-order moments of U_i^t), but it will be more convenient to state our results with both these assumptions in play.

As a special case, the "noiseless" regime $U_i^t = 0$ will be sometimes referred to as *perfect information*. However, to avoid clashes with existing terminology (especially within the literature on dynamic and repeated games), we stress here that players are never assumed to observe the actions of other players: even with perfect information, only individual gradient observations are available at each stage.

4. Regret minimization

4.1. **Types of regret.**

As we discussed in the introduction, a minimal worst-case requirement for online decision-making is that of regret minimization. In the non-stationary framework of the previous section, the regret of player $i \in \mathcal{N}$ relative to a test action $x_i \in \mathcal{X}_i$ over a window of play $\mathcal{T} \subseteq \mathbb{N}$ is defined as

$$\mathrm{Reg}_i(\mathcal{T}; x_i) \equiv \sum_{t \in \mathcal{T}} [u_i^t(x_i; X_{-i}^t) - u_i^t(X^t)], \tag{4.1}$$

and the agent's *static* (or *static*) regret is defined as

$$\mathrm{Reg}_i(\mathcal{T}) \equiv \max_{x_i \in \mathcal{X}_i} \mathrm{Reg}_i(\mathcal{T}; x_i) = \max_{x_i \in \mathcal{X}_i} \sum_{t \in \mathcal{T}} [u_i^t(x_i; X_{-i}^t) - u_i^t(X^t)], \tag{4.2}$$

i.e., as the difference between the cumulative payoff of the focal player $i \in \mathcal{N}$ under the sequence of play $X^t \in \mathcal{X}$, $t = 1, 2, \dots$, and that of the player's best fixed action

[8] To avoid superfluous notation, we are omitting in (3.3) the dependence of X and Y on ω, and we are treating Alg_i as a function of variable arity (so as to drop its dependence on t).

over the time window $\mathcal{T}$.[9] Then, specializing to the case where $\mathcal{T} = \{1, \dots, T\}$, we will say that the sequence X^t leads to *no (static) regret* if

$$\limsup_{T\to\infty} \frac{1}{T} \operatorname{Reg}_i(\mathcal{T}) \leq 0 \quad \text{for all } i \in \mathcal{N}, \tag{4.3}$$

i.e., if every player's regret grows at most sublinearly with the horizon of play:

$$\operatorname{Reg}_i(\mathcal{T}) = o(T) \quad \text{for all } i \in \mathcal{N}.$$

By the individual concavity of the players' payoff functions, the payoff difference in the definition of the regret can be bounded from above as

$$u_i^t(x_i; X_{-i}^t) - u_i^t(X^t) \leq \langle v_i^t(X^t), x_i - X_i^t \rangle,$$

for any reference action $x_i \in \mathcal{X}_i$ and all $t \in \mathcal{T}$. Consequently, a player's regret can be itself bounded from above as

$$\operatorname{Reg}_i(\mathcal{T}) \leq \operatorname{Gap}_i(\mathcal{T}),$$

where

$$\operatorname{Gap}_i(\mathcal{T}) \equiv \max_{x_i \in \mathcal{X}_i} \sum_{t\in\mathcal{T}} \langle v_i^t(X^t), x_i - X_i^t \rangle \tag{4.4}$$

represents a linearized, player-specific regret measure that we call the *gap function* of player i. Hence, to achieve no regret, it suffices to design an algorithm guaranteeing that $\operatorname{Gap}_i(\mathcal{T}) = o(T)$ for every player $i \in \mathcal{N}$. This linearization device has been the starting point of most no-regret strategies in the literature (see e.g., Cesa-Bianchi and Lugosi [8], Shalev-Shwartz [45], Mertikopoulos and Zhou [30], Nesterov [35], Bubeck and Cesa-Bianchi [6], and references therein), and we will use it freely in the rest of our paper.

Of course, an important limitation in the definition (4.2) of a player's regret is that it compares the sequence of accrued rewards to that of the best *fixed* action in hindsight. Since the players' payoff functions evolve over time, a player following a policy satisfying (4.3) may still incur a substantial loss relative to a *non-constant* sequence of actions. Thus, to get a finer performance benchmark, we consider instead the *dynamic regret* of player i relative to a *test sequence* $x_i^t \in \mathcal{X}_i$, $t = 1, 2, \dots$, defined as

$$\operatorname{DynReg}_i(\mathcal{T}; x_i^{\mathcal{T}}) \equiv \sum_{t\in\mathcal{T}} [u_i^t(x_i^t; X_{-i}^t) - u_i^t(X^t)].$$

Then, as in the static case, we say that a sequence of play X^t leads to no dynamic regret relative to a test sequence $x^t \in \mathcal{X}$, $t = 1, 2, \dots$,[10] if

$$\limsup_{T\to\infty} \frac{1}{T} \operatorname{DynReg}_i(\mathcal{T}; x_i^{\mathcal{T}}) \leq 0 \quad \text{for all } i \in \mathcal{N},$$

i.e., if every player's dynamic regret relative to x^t grows at most sublinearly with the horizon of play T.

As a special case, if $x_i^t = x_i \in \mathcal{X}_i$ for all $t = 1, 2, \dots$, we recover the agent's static regret relative to x_i, as given by (4.1). At the other end of the spectrum,

[9] By "window" we refer here to an interval of successive positive integers, i.e., $\mathcal{T}$ is of the form $\mathcal{T} = \{a, a+1, \dots, b\}$ for some $a, b \in \mathbb{N}$. Unless explicitly mentioned otherwise, we will only work with intervals of this type.

[10] Hall and Willett [18] and Jadbabaie et al. [21] instead use the term "comparator sequence".

if $x_i^t \in \operatorname{argmax}_{x_i \in \mathcal{X}_i} u_i^t(x_i; X_{-i}^t)$ is a sequence of individual best responses to the realized sequence of play X_{-i}^t of all other players, we get

$$\operatorname{DynReg}_i(\mathcal{T}; x_i^{\mathcal{T}}) = \sum_{t \in \mathcal{T}} \max_{x_i \in \mathcal{X}_i} [u_i^t(x_i; X_{-i}^t) - u_i^t(X^t)].$$

Comparing this expression to the definition (4.2) of the static regret of player i, we see that the order of summation and maximization have been exchanged. In this way, we recover the original definition of Besbes et al. [4], suitably extended to our multi-agent setting: in the long run, under a policy leading to no dynamic regret, each player's accrued rewards are no worse than what the player would have obtained by best-responding to the sequence of play of all other players.

4.2. **Dynamic regret minimization.** Our main goal in the rest of this section will be to provide a a universal bound for the players' dynamic regret relative to arbitrary test sequences. To do so, we will again rely on the individual concavity of the players' payoff functions to write

$$\operatorname{DynReg}_i(\mathcal{T}; x_i^{\mathcal{T}}) \leq \sum_{t \in \mathcal{T}} \langle v_i^t(X^t), x_i^t - X_i^t \rangle,$$

just as in the static case. Then, motivated by the recent analysis of Besbes et al. [4], we will decompose a player's dynamic regret into two components: one driven by the gap function (4.4) over smaller windows of play, and the other measuring the *variation* of the test sequence x_i^t over time, as defined below:

Definition 4.1. The variation of a test sequence $x_i^t \in \mathcal{X}_i$, $t = 1, 2, \dots$, over the window $\mathcal{T} \subseteq \mathbb{N}$ is defined as

$$\operatorname{Var}_i(\mathcal{T}; x_i^{\mathcal{T}}) = \sum_{t \in \mathcal{T}} \|x_i^{t+1} - x_i^t\|,$$

with the convention that $x_i^{t+1} = x_i^t$ if $t = T$.

To proceed with the decomposition outlined above, let $\mathcal{T}_{i,1}, \dots, \mathcal{T}_{i,m_i}$ be a partition of the time window $\mathcal{T} = \{1, \dots, T\}$ into m_i *batches*, each of size

$$\Delta_i = \lfloor T/m_i \rfloor,$$

with the possible exception of the last one (which might be smaller). We then have:

Lemma 4.2. *The dynamic regret of the i-th player relative to a test sequence $x_i^t \in \mathcal{X}_i$, $t = 1, 2, \dots$, satisfies*

$$\operatorname{DynReg}_i(\mathcal{T}; x_i^{\mathcal{T}}) \leq \sum_{\ell=1}^{m_i} \operatorname{Gap}_i(\mathcal{T}_{i,\ell}) + G_i \Delta_i \operatorname{Var}(\mathcal{T}, x_i^{\mathcal{T}}). \tag{4.5}$$

Proof. The proof is an elementary computation building on an idea of Besbes et al. [4]. Indeed, dropping the player index i for notational clarity, individual concavity yields

$$\operatorname{DynReg}(\mathcal{T}; x^{\mathcal{T}}) = \sum_{\ell=1}^{m} \operatorname{DynReg}(\mathcal{T}_\ell; x^{\mathcal{T}_\ell}) \leq \sum_{\ell=1}^{m} \sum_{t \in \mathcal{T}_\ell} \langle v^t(X^t), x^t - X^t \rangle.$$

Now, fixing a reference action $p_\ell \in \mathcal{X}$ for each batch $\ell = 1, \dots, m$, let

$$I_\ell = \sum_{t \in \mathcal{T}_\ell} \langle v^t(X^t), p_\ell - X^t \rangle,$$

and

$$J_\ell = \sum_{t \in \mathcal{T}_\ell} \langle v^t(X^t), x^t - p_\ell \rangle,$$

so that the linearized dynamic regret over each batch can be written as

$$\sum_{t \in \mathcal{T}_\ell} \langle v^t(X^t), x^t - X^t \rangle = I_\ell + J_\ell.$$

To bound I_ℓ, note that

$$I_\ell = \sum_{t \in \mathcal{T}_\ell} \langle v^t(X^t), p_\ell - X^t \rangle \leq \max_{x \in \mathcal{X}} \sum_{t \in \mathcal{T}_\ell} \langle v^t(X^t), x - X^t \rangle = \operatorname{Gap}(\mathcal{T}_\ell).$$

Subsequently, to bound J_ℓ, let p_ℓ be the first element of the test sequence x^t over the ℓ-th batch $\mathcal{T}_\ell$. Then,

$$\begin{aligned}\sum_{t \in \mathcal{T}_\ell} \langle v^t(X^t), x^t - p_\ell \rangle &\leq \sum_{t \in \mathcal{T}_\ell} \|v^t(X^t)\|_* \cdot \|x^t - p_\ell\| \\ &\leq G \sum_{t \in \mathcal{T}_\ell} \|x^t - p_\ell\| \\ &\leq G\Delta \max_{t \in \mathcal{T}_\ell} \|x^t - p_\ell\| \\ &\leq G\Delta \sum_{t \in \mathcal{T}_\ell} \|x^{t+1} - x^t\| = G\Delta \operatorname{Var}(\mathcal{T}_\ell; x^{\mathcal{T}_\ell}),\end{aligned}$$

where we used Young's inequality in the first line and the triangle inequality in the last one. Our claim then follows by summing over each batch $\ell = 1, \dots, m$. □

Lemma 4.2 suggests that minimizing a player's regret relative to a rapidly-varying test sequence x_i^t may be difficult (if not downright impossible). On the other hand, if the test sequence under study is *slowly-varying* in the sense that

$$\operatorname{Var}(\mathcal{T}; x^{\mathcal{T}}) = o(|\mathcal{T}|),$$

then it might be feasible to attain no (dynamic) regret by properly tweaking the batch size Δ_i in the regret decomposition (4.5).

This observation was the starting point of the analysis of Besbes et al. [4] who proposed breaking the horizon of play into batches of a carefully chosen size, and then running on each batch an algorithm that incurs low static regret (i.e., sublinear relative to the size of the batch).[11] In our game-theoretic setting, this boils down to each player choosing a batch size Δ_i and breaking play up into $m_i = \lceil T/\Delta_i \rceil$ successive time windows $\mathcal{T}_{i,1}, \dots, \mathcal{T}_{i,m_i}$, each of size Δ_i (except possibly the last one). Then, at every window $\mathcal{T}_{i,\ell}$, each player $i \in \mathcal{N}$ updates their actions following an (as yet unspecified) algorithm Alg_i which is restarted every Δ_i stages. Formally, this restart procedure can be encoded in pseudocode form as follows:

[11] By contrast, Hall and Willett [18] and Shahrampour and Jadbabaie [44] do not employ a restart procedure and instead specialize to gradient/mirror descent to obtain sublinear regret.

Algorithm 1 Batch restart (player indices suppressed)

```
Require: Horizon $T$, batch size $\Delta$, choice algorithm Alg as in (3.3)
set $t \leftarrow 1$                                                # step counter
choose $X^1 \in \mathcal{X}$                                        # initialization
repeat
    set $\tau \leftarrow \lfloor (t-1)/\Delta \rfloor \Delta + 1$   # augment every $\Delta$ stages
    play $X^t \in \mathcal{X}$                                      # play chosen action
    get signal $Y^t$                                                # receive feedback
    set $X^{t+1} \leftarrow \mathsf{Alg}(X^\tau, Y^\tau, \dots, X^t, Y^t)$   # update action
    $t \leftarrow t+1$                                              # next stage
until $t > T$                                                       # end play
```

In view of all this, if the restart frequency is chosen as a function of the variation of the test sequence under study, we have:

Theorem 4.3. *Consider a time-varying game $\mathcal{G}^t \equiv \mathcal{G}^t(\mathcal{N}, \mathcal{X}, u^t)$, $t = 1, 2, \dots,$ and let Alg_i be an algorithm of the general form (3.3) such that*

$$\mathbb{E}[\mathrm{Gap}_i(\mathcal{T})] \leq C_i \sqrt{T}$$

for some $C_i > 0$ and all time intervals $\mathcal{T} \subseteq \mathbb{N}$ of length T. Suppose further that $x_i^t \in \mathcal{X}_i$, $t = 1, 2, \dots,$ is a test sequence enjoying the variation bound

$$\mathrm{Var}_i(\mathcal{T}; x_i^{\mathcal{T}}) \leq V_T,$$

for some $V_T \geq 1$. Then, if Alg_i is rebooted every $\Delta_i = \lceil (T/V_T)^{2/3} \rceil$ stages following Algorithm 1, we have

$$\mathbb{E}[\mathrm{DynReg}_i(\mathcal{T}; x_i^{\mathcal{T}})] \leq (2C_i + 3G_i) T^{2/3} (V_T)^{1/3}.$$

In particular, if x_i^t is slowly-varying (i.e., $V_T/T \to 0$ as $T \to \infty$), we have

$$\limsup_{T\to\infty} \mathbb{E}[\mathrm{DynReg}_i(\mathcal{T}; x_i^{\mathcal{T}})] \leq 0.$$

Remark 3. In the above, expectations are taken with respect to the randomness of the players' signals (and induced action sequences).

Proof. Taking expectations on both sides of the bound (4.5) yields

$$\mathbb{E}[\mathrm{DynReg}_i(\mathcal{T}; x_i^{\mathcal{T}})] \leq \sum_{\ell=1}^{m_i} \mathbb{E}[\mathrm{Gap}_i(\mathcal{T}_{i,\ell})] + G_i \Delta_i \,\mathrm{Var}(\mathcal{T}, x_i^{\mathcal{T}}),$$

where $m_i = \lceil T/\Delta_i \rceil$ is the number of restarts up to stage T (inclusive). Then, with $\mathrm{Var}_i(\mathcal{T}; x_i^{\mathcal{T}}) \leq V_T$ and $|\mathcal{T}_{i,\ell}| \leq \Delta_i = \lceil (T/V_T)^{2/3} \rceil$, this bound becomes:

$$\begin{aligned}
\mathbb{E}[\mathrm{DynReg}_i(\mathcal{T}; x_i^{\mathcal{T}}] &\leq m_i C_i \sqrt{\Delta_i} + G_i \Delta_i V_T \\
&\leq (T/\Delta_i + 1) C_i \sqrt{\Delta_i} + G_i \Delta_i V_T \\
&\leq C_i [1 + T^{2/3} (V_T)^{1/3} + (T/V_T)^{1/3}] + G_i [V_T + T^{2/3} (V_T)^{1/3}] \\
&\leq (3C_i + 2G_i) T^{2/3} (V_T)^{1/3}
\end{aligned}$$

where, in the last line, we used the fact that $V_T \geq 1$. □

As in the work of Besbes et al. [4], Theorem 4.3 can be seen as a "meta-principle" that allows players to leverage a policy with no *static* regret to obtain a policy with no *dynamic* regret. In this way, it should be contrasted to the work of Hall and Willett [18] and Shahrampour and Jadbabaie [44] who focus on a specific learning policy (based on mirror descent) and circumvent the need for restarting by exploiting its specific properties. We discuss this issue in more detail in Section 6.

5. Distributed learning

In this section we present a class of distributed learning algorithms based on *online mirror descent* (OMD), a meta-algorithm which, together with the closely related "follow the regularized leader" (FTRL) protocol, comprises one of the most widely used algorithmic schemes for no-regret learning in online optimization – for a partial survey, see Nemirovski and Yudin [34], Beck and Teboulle [3], Nemirovski et al. [33], Teboulle [48], Chen and Teboulle [10], Nesterov [35], Shalev-Shwartz [45], Mertikopoulos and Zhou [30], and references therein.

Viewed abstractly, the basic idea of mirror descent (or, in our case, "ascent") is as follows: if player $i \in \mathcal{N}$ plays $x_i \in \mathcal{X}_i$ and receives the gradient signal $y_i \in \mathcal{Y}_i$, the algorithm generates a new action x_i^+ by taking an "approximate gradient" step from x_i along y_i. Formally, this can be written as

$$x_i^+ = P_i(x_i, \gamma y_i) \tag{5.1}$$

where

(1) γ is a step-size parameter controlling the weight attributed to the signal y_i.
(2) $P_i \colon \mathcal{X}_i \times \mathcal{Y}_i \to \mathcal{X}_i$ is a "proximal mapping" (discussed in detail below) which determines the exact way in which the step along y_i is taken.

Remark 4. Because the prox-mapping P_i plays a defining role in the players' action selection process, and to avoid clashes between the term "descent" and the fact that players are treated as maximizers in our setting, we will refer to (5.1) as a *prox-method* (PM).

Now, given a convex subset $\mathcal{C}$ of some ambient vector space $\mathcal{V} \cong \mathbb{R}^d$, the prototypical example of a prox-mapping is the Euclidean projector

$$\begin{aligned} P(x,y) = \Pi_{\mathcal{C}}(x+y) &\equiv \operatorname*{argmin}_{x' \in \mathcal{C}} \{\|x+y-x'\|_2^2\} \\ &= \operatorname*{argmin}_{x' \in \mathcal{C}} \{\langle y, x-x'\rangle + \tfrac{1}{2}\|x'-x\|_2^2\} \end{aligned} \tag{5.2}$$

i.e., the closest-point projection of $x+y$ onto $\mathcal{C}$.[12] Going beyond the Euclidean case, the key novelty of prox-methods is to replace the distance term $\frac{1}{2}\|x'-x\|_2^2$ in (5.2) with a (possibly non-symmetric) "divergence" defined by means of a *distance-generating function* (DGF) $h \colon \mathcal{C} \to \mathbb{R}$, itself assumed to be continuous and K-strongly convex, i.e.,

$$h(tx + (1-t)x') \leq th(x) + (1-t)h(x') - \frac{K}{2} t(1-t)\|x'-x\|^2$$

[12] Note here that, in writing $x+y$, we are blurring the lines between primal vectors $x \in \mathcal{V}$ and dual vectors $y \in \mathcal{V}^*$. This distinction is reinstated in the second line of (5.2) where $y \in \mathcal{V}^*$ is paired properly to $x - x' \in \mathcal{V}$.

for all $x, x' \in \mathcal{C}$ and all $t \in [0,1]$. For technical reasons (and in a slight abuse of notation), we further assume that the subdifferential $\partial h(x) = \{y \in \mathcal{V}^* : h(x') \geq h(x) + \langle y, x' - x \rangle\}$ of h admits a *continuous selection*: specifically, letting

$$\mathcal{C}^\circ \equiv \operatorname{dom} \partial h = \{x \in \mathcal{C} : \partial h(x) \neq \varnothing\}$$

denote the domain of subdifferentiability of h, we posit that there exists a continuous function $\nabla h \colon \mathcal{C}^\circ \to \mathcal{V}^*$ such that $\nabla h(x) \in \partial h(x)$ for all $x \in \mathcal{C}^\circ$.[13] The *Bregman divergence* induced by h is then defined as

$$D_h(x', x) = h(x') - h(x) - \langle \nabla h(x), x' - x \rangle \quad \text{for all } x' \in \mathcal{C},\, x \in \mathcal{C}^\circ,$$

and the associated *prox-mapping* $P \colon \mathcal{C} \times \mathcal{V}^* \to \mathcal{C}$ is given by

$$P(x, y) = \operatorname*{argmin}_{x' \in \mathcal{C}} \{\langle y, x - x' \rangle + D_h(x', x)\} \quad \text{for all } x \in \mathcal{C}^\circ,\, y \in \mathcal{V}^*.$$

Before continuing, it will be instructive to provide some standard examples of prox-mappings:

Example 5.1 (Euclidean projections). We begin by recovering the archetypal example of Euclidean projections. To do so, let $h(x) = \frac{1}{2}\|x\|^2$. Since h is subdifferentiable throughout $\mathcal{C}$, we have $\mathcal{C}^\circ = \mathcal{C}$; moreover, $\nabla h(x) = x$ is a continuous selection of $\partial h(x)$ for all $x \in \mathcal{C}$. Hence, the associated Bregman divergence is

$$D_h(x', x) = \tfrac{1}{2}\|x'\|_2^2 - \tfrac{1}{2}\|x\|_2^2 - \langle x, x' - x \rangle = \tfrac{1}{2}\|x' - x\|_2^2,$$

and the induced prox-mapping is given by (5.2) for all $x \in \mathcal{C}$, $y \in \mathcal{V}^*$.

Example 5.2 (Entropic regularization). Let $\mathcal{C} = \{x \in \mathbb{R}_+^d : \sum_{j=1}^d x_j = 1\}$ denote the unit simplex of $\mathcal{V} = \mathbb{R}^d$. A very widely used distance-generating function for this geometry is the (negative) *Gibbs-Shannon entropy* $h(x) = \sum_{j=1}^d x_j \log x_j$. By inspection, the domain of (sub)differentiability of h is $\mathcal{C}^\circ = \operatorname{ri} \mathcal{C}$, and the resulting Bregman divergence is given by the *Kullback–Leibler* (KL) expression

$$D_h(x', x) = \sum_{j=1}^d x'_j \log\left(\frac{x'_j}{x_j}\right),$$

valid for all $x \in \mathcal{C}^\circ$, $x' \in \mathcal{C}$. In turn, this gives rise to the prox-mapping

$$P(x, y) = \frac{(x_j \exp(-y_j))_{j=1}^d}{\sum_{j=1}^d x_j \exp(-y_j)}$$

for all $x \in \mathcal{C}^\circ$, $y \in \mathcal{V}^*$. The update rule $x^+ = P(x, y)$ is widely known in the literature as the *multiplicative weights* (MW) algorithm and plays a central role for learning in multi-armed bandit problems and finite games. For a survey, see Freund and Schapire [17], Arora et al. [1], Cohen et al. [11], Palaiopanos et al. [38], and references therein.

[13] By standard results in convex analysis [39], we have $\operatorname{ri} \mathcal{C} \subseteq \mathcal{C}^\circ \subseteq \mathcal{C}$. Note also that we are making use of the standard convention that $h(x) = +\infty$ if $x \in \mathcal{V} \setminus \{\mathcal{C}\}$.

Example 5.3 (Fermi-Dirac regularization)*.* Let $\mathcal{C} = [0,1]$ and let $h(x) = x\log(x) + (1-x)\log(1-x)$ be the (negative) Fermi-Dirac entropy. Then, $\mathcal{C}^\circ = (0,1)$ and the induced prox-mapping is given by the expression

$$P(x,y) = \frac{x\exp(-y)}{1-x+x\exp(-y)},$$

valid for all $x \in (0,1)$, $y \in \mathbb{R}$.

With all this at hand, the general class of game-theoretic learning algorithms that we will consider will be given by the recursion

$$X_i^{t+1} = P_i(X_i^t, \gamma_i^t Y_i^t)$$

where, in more detail:

1. $t = 1, 2, \dots$ denotes the stage of the process.
2. $X_i^t \in \mathcal{X}_i$ is the action played by player i at stage t.
3. $Y_i^t \in \mathcal{Y}_i$ is the signal received by player i at stage t, assumed throughout to satisfy the unbiasedness assumption $\mathbb{E}[Y_i^t \,|\, \mathcal{F}^{t-1}] = v_i^t(X^t)$ (cf. Section 3).
4. γ_i^t is a player-specific step-size sequence (assumed nonincreasing).
5. $P_i \colon \mathcal{X}_i \times \mathcal{Y}_i \to \mathcal{X}_i$ denotes the prox-mapping of player i, itself derived from some distance-generating function $h_i \colon \mathcal{X}_i \to \mathbb{R}$ as above.

In particular, unless explicitly mentioned otherwise, all repeated game strategies described in Section 3 will be henceforth assumed to be of the (Markovian) form

$$\mathsf{Alg}_i(X_i^1, Y_i^1, \dots, X_i^t, Y_i^t) = P_i(X_i^t, \gamma_i^t Y_i^t).$$

For concreteness, we also provide a pseudocode implementation of this prox-based learning protocol as Algorithm 2 below:

PM

Removed the initialization and offloaded it to the results that needed it.

Algorithm 2 Prox-method for distributed learning (player indices suppressed)

```
Require: prox-mapping $P\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{X}$, step-size sequence $\gamma^t \geq 0$,
         sequence of stage games $\mathcal{G}^t \equiv \mathcal{G}^t(\mathcal{N},\mathcal{X},u^t)$
initialize $X^1 = \operatorname{argmin}_{x\in\mathcal{X}} h(x)$        # initialization
for $t = 1, 2, \dots$ do
    set $\mathcal{G} \leftarrow \mathcal{G}^t(\mathcal{N},\mathcal{X},u^t)$        # stage game definition
    play $X^t \in \mathcal{X}$        # play chosen action
    get signal $Y^t \in \mathcal{Y}$        # receive feedback
    set $X^{t+1} \leftarrow P(X^t, \gamma^t Y^t)$        # update action
    $t \leftarrow t+1$        # next stage
end for
```

6. Explicit regret bounds

We show in this section how Theorem 4.3 can be applied in the specific case where each player adheres to the prox-method described in the previous section. To that end, suppose that the players face a sequence of games $\mathcal{G}^t$, $t = 1, 2, \dots$, and seek to minimize their regret following Algorithm 2. Following the meta-principle outlined in Section 4, our dynamic regret analysis begins with a basic inequality bounding the static regret of any fixed action over a finite horizon of play:

Proposition 6.1. *Let $\mathcal{G}^t \equiv \mathcal{G}^t(\mathcal{N}, \mathcal{X}, u^t)$ be a sequence of concave games. Assume further that Algorithm 2 is initialized at $p_i \equiv \operatorname{argmin}_{x_i \in \mathcal{X}_i} h_i(x_i)$ and run with step-size γ_i. Then, the regret incurred by player i relative to a fixed test action $x_i \in \mathcal{X}_i$ over the window of play $\mathcal{T} = \{1, \dots, T\}$ enjoys the bound*

$$\operatorname{Reg}_i(\mathcal{T}; x_i) \leq \frac{\mathcal{D}[\mathcal{X}_i, h_i]}{\gamma_i} + \sum_{t=1}^{T} \langle U_i^t, X_i^t - x_i \rangle + \frac{\gamma_i}{2K_i} \sum_{t=1}^{T} \|Y_i^t\|_*^2,$$

with $\mathcal{D}[\mathcal{X}_i, h_i] = \max h_i - \min h_i$.

Results of this flavor are fairly well known in the online learning literature; still, for the sake of completeness, we present a quick proof below.

Proof of Proposition 6.1. The basic starting point of our analysis is the following inequality, taken from Eq. (A.5) in Appendix A:

$$D_{h_i}(x_i, X_i^{t+1}) - D_{h_i}(x_i, X_i^t) \leq -\langle \gamma_i^t Y_i^t, x_i - X_i^t \rangle + \frac{(\gamma_i^t)^2}{2K_i} \|Y_i^t\|_*^2.$$

Using the decomposition of the signal at stage t and rearranging the above expression yields

$$\begin{aligned} \langle \gamma_i^t v_i^t(X^t), x_i - X_i^t \rangle \leq D_{h_i}(x_i, X_i^t) - D_{h_i}(x_i, X_i^{t+1}) \\ - \langle \gamma_i^t U_i^t, x_i - X_i^t \rangle + \frac{(\gamma_i^t)^2}{2K_i} \|Y_i^t\|_*^2. \end{aligned} \tag{6.1}$$

Taking a constant step-size $\gamma_i^t = \gamma_i$ and telescoping gives

$$\begin{aligned} \sum_{t=1}^{T} \langle v_i^t(X^t), x_i - X_i^t \rangle \leq \frac{1}{\gamma_i} \big(D_{h_i}(x_i, X_i^1) - D_{h_i}(x_i, X_i^{T+1}) \big) \\ - \sum_{t=1}^{T} \langle U_i^t, x_i - X_i^t \rangle + \sum_{t=1}^{T} \frac{\gamma_i}{2K_i} \|Y_i^t\|_*^2. \end{aligned}$$

Since $X_i^1 = \operatorname{argmin}_{x_i \in \mathcal{X}_i} h_i(x_i)$ and the Bregman divergence is non-negative, we then get

$$\begin{aligned} \sum_{t=1}^{T} \langle v_i^t(X^t), x_i - X_i^t \rangle &\leq \frac{1}{\gamma_i} D_{h_i}(x_i, X_i^1) + \sum_{t=1}^{T} \langle U_i^t, X_i^t - x_i \rangle + \sum_{t=1}^{T} \frac{\gamma_i}{2K_i} \|Y_i^t\|_*^2 \\ &\leq \frac{1}{\gamma_i} \mathcal{D}[\mathcal{X}_i, h_i] + \sum_{t=1}^{T} \langle U_i^t, X_i^t - x_i \rangle + \sum_{t=1}^{T} \frac{\gamma_i}{2K_i} \|Y_i^t\|_*^2 \end{aligned}$$

where, in the last line, we used the Bregman divergence bound (A.2) derived in Appendix A. □

Proposition 6.1 is the main stepping stone towards obtaining a static regret bound that holds *in expectation*; we provide the details for this derivation in the next section. Beyond this basic bound, it is possible to use Markov's inequality to obtain a bound that holds with high probability; however, as we show in Section 6.2, it is possible to prove a much tighter large deviation principle for the algorithm's static regret by using a series of exponential concentration arguments for martingales.

6.1. **Bounding the expected static regret.** In order to bound the static regret of Algorithm 2, we recall the basic assumptions we imposed on the observational noise of the players' feedback signal. First, by combining Eqs. (3.1) and (3.5a) and letting $M_i^2 = 2G_i^2 + 2\sigma_i^2$, we readily get

$$\mathbb{E}[\|Y_i^t\|_* \,|\, \mathcal{F}^{t-1}] \leq M_i^2 \quad \text{for all } t = 1, 2, \dots$$

Under these assumptions, we have:

Proposition 6.2. *Suppose that Algorithm 2 is run with assumptions as in Proposition 6.1 and step-size*

$$\gamma_i = 2\sqrt{\frac{\mathcal{D}[\mathcal{X}_i; h_i]K_i}{T(M_i^2 + \sigma_i^2)}}. \tag{6.2}$$

We then have the following regret bound over the time window $\mathcal{T} = \{1, \dots, T\}$*:*

$$\mathbb{E}[\mathrm{Reg}_i(\mathcal{T})] \leq 2\sqrt{T(M_i^2 + \sigma_i^2)\mathcal{D}[\mathcal{X}_i; h_i]/K_i}. \tag{6.3}$$

In particular, $\limsup_{T\to\infty} \mathbb{E}[\mathrm{Reg}_i(\mathcal{T})/T] = 0$.

Remark 5. Before proving Proposition 6.2, it is worth noting that (6.3) is an upper bound on the *expected regret* of Algorithm 2, not the algorithm's *pseudo-regret*

$$\mathrm{PReg}_i(\mathcal{T}) = \max_{x_i \in \mathcal{X}_i} \mathbb{E}[\mathrm{Reg}_i(\mathcal{T}; x_i)]. \tag{6.4}$$

Bounding this latter, weaker notion of regret is more common in the online learning literature (see e.g., Bubeck and Cesa-Bianchi [6] and references therein) because the expectation in (6.4) is taken over a fixed action (as opposed to the best possible action in hindsight); as such, the pseudo-regret is considerably easier to treat than the actual, expected regret. Moreover, by Jensen's inequality, a bound on the expected regret can be automatically translated to a bound on the pseudo-regret, so Proposition 6.2 serves a dual purpose in this regard.

Proof of Proposition 6.2. By the bound (6.1), we have

$$\begin{aligned}\langle \gamma_i^t v_i^t(X^t), x_i - X_i^t\rangle \leq D_{h_i}(x_i, X_i^t) &- D_{h_i}(x_i, X_i^{t+1}) \\ &- \langle \gamma_i^t U_i^t, x_i - X_i^t\rangle + \frac{(\gamma_i^t)^2}{2K_i}\|Y_i^t\|_*^2.\end{aligned}$$

For each player i define the auxiliary process $\{Z_i^t\}_{t\in\mathbb{N}}$ by

$$Z_i^1 = X_i^1 \text{ and } Z_i^{t+1} = P_i(Z_i^t, \gamma_i^t U_i^t).$$

A simple induction argument shows that the process $\{Z_i^t\}_{t\geq 1}$ is $\{\mathcal{F}_i^t\}_{t\geq 1}$ measurable, for all $i \in \mathcal{N}$. Using this process, the previous inequality can be rewritten as

$$\begin{aligned}\langle \gamma_i^t v_i^t(X^t), x_i - X_i^t\rangle &\leq D_{h_i}(x_i, X_i^t) - D_{h_i}(x_i, X_i^{t+1}) \\ &\quad - \langle \gamma_i^t U_i^t, Z_i^t - X_i^t\rangle + \frac{(\gamma_i^t)^2}{2K_i}\|Y_i^t\|_*^2 + \langle \gamma_i^t U_i^t, Z_i^t - x_i\rangle.\end{aligned}$$

Hence, after summing and telescoping, we arrive at the bound

$$\begin{aligned}\sum_{t=1}^T \langle \gamma_i^t v_i^t(X^t), x_i - X_i^t\rangle \leq D_{h_i}(x_i, X^1) - D_{h_i}(x_i, X_i^{T+1}) + \sum_{t=1}^T \langle \gamma_i^t U_i^t, X_i^t - Z_i^t\rangle \\ + \sum_{t=1}^T \frac{(\gamma_i^t)^2}{2K_i}\|Y_i^t\|_*^2 + \sum_{t=1}^T \langle \gamma_i^t U_i^t, Z_i^t - x_i\rangle.\end{aligned}$$

Lemma A.2, proven in Appendix A, shows that

$$\sum_{t=1}^{T} \langle \gamma_i^t U_i^t, Z_i^t - x_i \rangle \leq D_{h_i}(x_i, X_i^1) + \frac{1}{2K_i} \sum_{t=1}^{T} \| \gamma_i^t U_i^t \|_*^2.$$

Combining these two inequalities gives

$$\sum_{t=1}^{T} \langle \gamma_i^t v_i^t(X^t), x_i - X_i^t \rangle \leq 2 D_{h_i}(x_i, X_i^1) + \sum_{t=1}^{T} \langle \gamma_i^t U_i^t, X_i^t - Z_i^t \rangle + \sum_{t=1}^{T} \frac{(\gamma_i^t)^2}{2K_i} \big(\|Y_i^t\|_*^2 + \|U_i^t\|_*^2 \big). \tag{6.5}$$

In the case of a constant step-size $\gamma_i^t = \gamma_i$, this implies

$$\sum_{t=1}^{T} \langle v_i^t(X^t), x_i - X_i^t \rangle \leq \frac{2 D_{h_i}(x_i, X_i^1)}{\gamma_i} + \sum_{t=1}^{T} \langle U_i^t, X_i^t - Z_i^t \rangle + \frac{\gamma_i}{2K_i} \sum_{t=1}^{T} \big(\|Y_i^t\|_*^2 + \|U_i^t\|_*^2 \big).$$

Since $X_i^1 \in \operatorname{argmin}_{x_i \in \mathcal{X}_i} h_i(x_i)$, taking the supremum over actions $x_i \in \mathcal{X}_i$ on both sides of this inequality and using (4.4), we conclude

$$\mathrm{Reg}_i(\mathcal{T}) \leq \mathrm{Gap}_i(\mathcal{T}) \leq \frac{2}{\gamma_i} \mathcal{D}[\mathcal{X}_i, h_i] + \sum_{t=1}^{T} \langle U_i^t, X_i^t - Z_i^t \rangle + \sum_{t=1}^{T} \frac{\gamma_i}{2K_i} \big(\|Y_i^t\|_*^2 + \|U_i^t\|_*^2 \big).$$

The process $\sum_{t=1}^{T} \langle U_i^t, X_i^t - Z_i^t \rangle$ is a martingale with respect to the filtration $\mathbb{F} := \{\mathcal{F}_t\}_{t \geq 1}$, which is also bounded in $L^2(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, thanks to (6.9). The process $\sum_{t=1}^{T} \big(\|Y_i^t\|_*^2 + \|U_i^t\|_*^2 \big)$ is a non-negative submartingale, with expected value bounded by $T(M_i^2 + \sigma_i^2)$. Hence, taking expectations on both sides, we obtain

$$\mathbb{E}[\mathrm{Reg}_i(\mathcal{T})] \leq \mathbb{E}[\mathrm{Gap}_i(\mathcal{T})] \leq \frac{2}{\gamma_i} \mathcal{D}[\mathcal{X}_i, h_i] + \frac{T \gamma_i}{2K_i} (M_i^2 + \sigma_i^2).$$

Optimizing with respect to γ_i yields the step-size expression (6.2). We thus get

$$\mathbb{E}[\mathrm{Reg}_i(\mathcal{T})] \leq \mathbb{E}[\mathrm{Gap}_i(\mathcal{T})] \leq 2 \sqrt{\mathcal{D}[\mathcal{X}_i, h_i] T (M_i^2 + \sigma_i^2) / K_i}, \tag{6.6}$$

and our proof is complete. □

We close this section by noting that, when the feedback signal is deterministic, Proposition 6.2 immediately delivers a deterministic $\mathcal{O}(T^{1/2})$ regret bound:

Corollary 6.3. *Suppose that Algorithm 2 is run with assumptions as in Proposition 6.2 and perfect gradient observations* $(\sigma = 0)$*. We then have*

$$\mathrm{Reg}_i(\mathcal{T}) \leq 2 \sqrt{\mathcal{D}[\mathcal{X}_i, h_i] T M_i^2 / K_i}.$$

In particular, $\limsup_{T \to \infty} \mathrm{Reg}_i(\mathcal{T})/T = 0$ *for all* $i \in \mathcal{N}$*.*

6.2. **Bounding the static regret with high probability.** The analysis of the previous section provides bounds on the expected regret of Algorithm 2. However, in many real-world applications, a player typically only gets a single realization of their strategy, so it is important to have bounds that hold, not only on average, but also *with high probability.* Based on Proposition 6.2, a simple Markov bound on the magnitude of the realized regret readily yields

$$\mathbb{P}(\mathrm{Reg}_i(\mathcal{T}) \geq \varepsilon T) \leq \frac{1}{T\varepsilon}\mathbb{E}[\mathrm{Reg}_i(\mathcal{T})] \leq \frac{2}{\varepsilon}\sqrt{\frac{\mathcal{D}[\mathcal{X}_i, h_i](M_i^2 + \sigma_i^2)}{K_i T}}. \tag{6.7}$$

However, the $\Theta(1/\varepsilon)$ tails of this distribution are both heavy and long, implying in turn that there is significant probability of incurring regret that is order of magnitudes higher than the bound (6.3) would indicate. To counter this, we provide below a large deviations principle which shows that the agent's realized regret is exponentially unlikely to significantly exceed the mean bound (6.3).

To do so, we will need to slightly strengthen the finite mean square hypothesis (3.5a) and posit that there exists a constant $M_* > 0$ such that

$$\mathbb{E}\left[\exp\left(\frac{\|Y_i^t\|_*^2}{M_*^2}\right)\right] \leq \exp(1) \quad \text{for all } t = 1, 2, \dots, \; i \in \mathcal{N}. \tag{6.8}$$

This "ψ_2-type" bound has a long history in the optimization literature (see e.g., Nemirovski et al. [33], Juditsky et al. [22] and references therein) and, essentially, it means that the distribution of the feedback signal sequence Y^t is (uniformly) sub-Gaussian. As such, it holds under standard Gaussian observation noise sequences, Rademacher-distributed errors, and all noise distributions with bounded support. We also note that (6.8) implies that the observational noise has finite moments of all orders, so it is stronger than (3.4a).

Clearly, this assumption on the feedback imposes a similar structure on the observational noise process U^t. Indeed, by definition, we have

$$\|U_i^t\|_*^2 = \|Y_i^t - v_i^t(X^t)\|_*^2 \leq 2\|Y_i^t\|_*^2 + 2\|v_i^t(X^t)\|_*^2.$$

Since

$$\|v_i^t(X^t)\|_* = \|\mathbb{E}[Y_i^t \,|\, \hat{\mathcal{F}}_t]\|_* \leq \sqrt{\mathbb{E}[\|Y_i^t\|_*^2 \,|\, \hat{\mathcal{F}}_t]} \leq M_*,$$

we conclude that

$$\|U_i^t\|_*^2 \leq 2\|Y_i^t\|_*^2 + 2M_*^2,$$

and hence

$$\mathbb{E}\left[\exp\left(\frac{\|U_i^t\|_*^2}{4M_*^2}\right)\right] \leq \exp(1). \tag{6.9}$$

These bounds can be used to derive an exponential concentration inequality for the *realized* regret of Algorithm 2 as follows:

Proposition 6.4. *Fix a tolerance level $\varepsilon \in (0,1)$ and a horizon of play $T \geq 1$. Suppose further that (6.8) holds and Algorithm 2 is initialized at $p_i \equiv \operatorname{argmin}_{x_i \in \mathcal{X}_i} h_i(x_i)$ and run with step-size*

$$\gamma_i = \sqrt{\frac{2\mathcal{D}[\mathcal{X}_i, h_i]}{\Omega_i(\varepsilon) T}},$$

where $\Omega_i(\varepsilon) = 5(1 + \log(2/\varepsilon))M_*^2/(2K_i)$. Then, the incurred regret of player $i \in \mathcal{N}$ enjoys the bound

$$\mathrm{Reg}_i(\mathcal{T}) \leq 2\sqrt{2\mathcal{D}[\mathcal{X}_i; h_i]\Omega_i(\varepsilon) T} + 8M_*\sqrt{2K_i^{-1}\mathcal{D}[\mathcal{X}_i, h_i]\log(2/\varepsilon) T}, \tag{6.10}$$

with probability at least $1 - \varepsilon$. Thus, in the limit $\varepsilon \to 0$, $T \to \infty$, we have

$$\mathrm{Reg}_i(\mathcal{T}) = \mathcal{O}(\sqrt{\log(1/\varepsilon) T}) \quad \textit{with probability at least } 1 - \varepsilon.$$

PM

I put the explicit step-size here. The original formulation suggested that the step-size was variable but the proof is for a constant, ex post optimized step-size. Could you check I didn't miss anything?

Since the proof of Proposition 6.4 is fairly technical, we relegate it to Appendix B. What is more important for our purposes is to contrast the bound (6.10) with the Markov bound (6.7). Indeed, inverting (6.10), it follows that there exists some constant $\alpha > 0$ such that

$$\mathbb{P}(\operatorname{Reg}_i(\mathcal{T}) \geq \varepsilon T) \leq \exp(-\varepsilon T/\alpha).$$

Compared to (6.7), this shows that the probability of incurring high regret under (6.8) is exponentially small in both ε and T. This represents a considerable reduction from the $\varepsilon^{-1}T^{-1/2}$ dependence of the bound (6.7) and shows that, under the sub-Gaussian noise assumption (6.8), "black swan" realizations with significantly high regret are exponentially rare.

6.3. **Bounding the expected dynamic regret.** With all this groundwork at hand, we are in a position to derive a bound for the players' expected *dynamic* regret via the meta-principle provided by Theorem 4.3. To do so, the required ingredients are (i) the restart procedure of Besbes et al. [4] (cf. Algorithm 1); and (ii) an algorithm guaranteeing sublinear asymptotic behavior of the gap function. In this section we carry out this program, using Algorithm 2 as the driving subroutine for updating the players' decisions between restarts.

To make all this precise, suppose that each player breaks up the window of play $\mathcal{T} = \{1, \dots, T\}$ into blocks $\mathcal{T}_{i,1}, \mathcal{T}_{i,2}, \dots$, each of size Δ_i (except possibly the last one). We then have:

Proposition 6.5. *Let $x_i^t \in \mathcal{X}_i$, $t = 1, 2, \dots$, be a test sequence enjoying the variation bound*

$$\operatorname{Var}_i(\mathcal{T}; x_i^{\mathcal{T}}) \leq V_T$$

for some $V_T \geq 1$. Suppose further that Algorithm 1 is run with batch size $\Delta_i = \lceil (T/V_T)^{2/3} \rceil$ and, within each block, Algorithm 2 is run with step-size

$$\gamma_i = 2\sqrt{\frac{K_i \mathcal{D}[\mathcal{X}_i, h_i]}{\Delta_i (M_i^2 + \sigma_i^2)}}. \tag{6.11}$$

Then, the dynamic regret incurred by player $i \in \mathcal{N}$ enjoys the bound

$$\mathbb{E}[\operatorname{DynReg}_i(\mathcal{T}, x_i^{\mathcal{T}})] \leq (3C_i + 2G_i) T^{2/3} V_T^{1/3}, \tag{6.12}$$

where $C_i = 2\sqrt{\mathcal{D}[\mathcal{X}_i, h_i](M_i^2 + \sigma_i^2) K_i}$.

Proof. By Lemma 4.2, we can bound the dynamic regret incurred by player i against x_i^t as

$$\operatorname{DynReg}_i(\mathcal{T}; x_i^{\mathcal{T}}) \leq \sum_{\ell=1}^{m_i} \operatorname{Gap}_i(\mathcal{T}_{i,\ell}) + G_i \Delta_i \operatorname{Var}_i(\mathcal{T}; x_i^{\mathcal{T}}),$$

where $m_i = \lceil T/\Delta_i \rceil$ is the number of blocks in the partition of $\mathcal{T}$. By the bound (6.6) in the proof of Proposition 6.2, we can further bound the gap function of player i over the ℓ-th batch as

$$\mathbb{E}[\operatorname{Gap}_i(\mathcal{T}_{i,\ell})] \leq 2\sqrt{\mathcal{D}[\mathcal{X}_i, h_i] \Delta_i (M_i^2 + \sigma_i^2)/K_i}.$$

Hence, summing the bounds of the gap function over each block, we can bound the player's expected dynamic regret as

$$\mathbb{E}[\operatorname{DynReg}_i(\mathcal{T}; x_i^{\mathcal{T}})] \leq 2m_i \sqrt{\mathcal{D}[\mathcal{X}_i, h_i] \Delta_i (M_i^2 + \sigma_i^2)/K_i} + G_i \Delta_i V_T.$$

Our claim then follows in the same way as in the proof of Theorem 4.3. □

In closing this section, there are two points worth noting. First, by following the large deviations analysis of Section 6.2, it is possible to show that the players' dynamic regret is exponentially unlikely to exceed the bound (6.12). However, because of the required restart procedure, the corresponding exact expressions end up being considerably cumbersome to derive (and of comparably little interest), so we omit them.

Second, we should contrast the $\mathcal{O}(T^{2/3}V_T^{1/3})$ bound (6.12) to the corresponding $\mathcal{O}(T^{1/2}V_T^{1/2})$ bound of Hall and Willett [18] and Shahrampour and Jadbabaie [44] for mirror descent without restarting. Ignoring the fact that the latter bounds require perfect gradient observations, their main advantage lies in their better dependence on the horizon of play T. In particular, these bounds capture the (min-max optimal) $\mathcal{O}(T^{1/2})$ rate of regret minimization for static comparator sequences (by contrast, restarting would only provide a $\mathcal{O}(T^{2/3})$ regret minimization rate in the static case). At the same time however, the $\mathcal{O}(T^{2/3}V_T^{1/3})$ bound provided the restart heuristic carries a better dependence on the variation V_T of the chosen test sequence; as such, it is more adapted to highly dynamic environments where the chosen comparator sequence may be rapidly varying. This observation will play an important role in the game-theoretic analysis of the next section.

7. Regret minimization and Nash equilibrium

In this section, we examine the equilibrium convergence properties of the players' long-run behavior in two distinct regimes: *a*) when the sequence of stage games $\mathcal{G}^t$ encountered by the players evolves over time without converging; and *b*) when $\mathcal{G}^t$ converges to some limit game $\mathcal{G}$. In what follows, we will treat the process defining the time-varying game as a "black box" and we will not scruitinize its origins in detail; we do so in order to focus directly on the interplay between the fluctuations of the stage game and the induced sequence of play.

7.1. **Tracking Nash equilibria.** We begin by considering the case where $\mathcal{G}^t$ evolves without converging. Building on the discussion in Section 2, we will assume in what follows that each $\mathcal{G}^t$ is β-strongly monotone in the sense of (2.3), i.e.,

$$\langle v^t(x') - v^t(x), x' - x\rangle \leq -\beta\|x' - x\|^2$$

for all $t = 1, 2, \dots,$ and all $x, x' \in \mathcal{X}$. In particular, this implies that each stage game $\mathcal{G}^t$ admits a unique Nash equilibrium, which we will denote by $\hat{x}^t$. Then, to quantify the degree to which the players' chosen actions $X^t \in \mathcal{X}$ "track" the Nash equilibrium sequence $\hat{x}^t$ over the window of play $\mathcal{T} \subseteq \mathbb{N}$, we will use the *error function*

$$\operatorname{err}_i(\mathcal{T}) = \sum_{t\in\mathcal{T}} \|X_i^t - \hat{x}_i^t\|^2,$$

or, aggregating over all players $i \in \mathcal{N}$,

$$\operatorname{err}(\mathcal{T}) = \sum_{i\in\mathcal{N}} \operatorname{err}_i(\mathcal{T}) = \sum_{t\in\mathcal{T}} \|X^t - \hat{x}^t\|^2.$$

By construction, if this error function grows sublinearly with the size $T = |\mathcal{T}|$ of the window of play, the sequence X^t will be close to Nash equilibrium for most of the time (as determined by the asymptotic growth of $\operatorname{err}(\mathcal{T})$ over time). That being said, it should be intuitively clear that if the sequence of Nash equilibria varies arbitrarily from one stage to the next, then there is no way of achieving $\operatorname{err}(\mathcal{T}) = o(T)$.[14] For this reason, we will focus in what follows on time-varying games with *slowly-varying equilibria*, i.e., such that

$$\sum_{t=1}^{T-1} \|\hat{x}^{t+1} - \hat{x}^t\| = o(T) \quad \text{as } T \to \infty. \tag{7.1}$$

Under this assumption we have:

Theorem 7.1. *Let $\mathcal{G}^t \equiv \mathcal{G}^t(\mathcal{N}, \mathcal{X}, u^t)$ be a sequence of strongly monotone games. Assume further that the variation of each player's equilibrium component over the window of play $\mathcal{T} = \{1, \dots, T\}$ satisfies $\sum_{t=1}^{T-1} \|\hat{x}_i^{t+1} - \hat{x}_i^t\| \leq V_i^T$ for some $V_i^T > 0$. Then, if players follow Algorithm 2 for batches of size $\Delta_i = \lceil (T/V_i^T)^{2/3} \rceil$ (as per Algorithm 1) with step-size given by (6.11), we have*

$$\mathbb{E}[\operatorname{err}_i(\mathcal{T})] = \mathcal{O}(T^{2/3} (V_i^T)^{1/3})$$

for all $i \in \mathcal{N}$. In particular, if the sequence of stage equilibria is slowly-varying in the sense of (7.1), we have $\mathbb{E}[\operatorname{err}(T)] = o(T)$ as $T \to \infty$.

Proof. Our proof strategy will be to leverage the dynamic regret minimization properties of the restart schedule of Algorithm 1 (cf. Theorem 4.3). To that end, note first that, for every reference action $p_i \in \mathcal{X}_i$, strong monotonicity yields

$$\begin{aligned} \beta \|X_i^t - \hat{x}_i^t\|^2 &\leq \langle v_i^t(X^t), \hat{x}_i^t - X_i^t \rangle \\ &\leq \langle v_i^t(X^t), p_i - X_i^t \rangle + \langle v_i^t(X^t), \hat{x}_i^t - p_i \rangle. \end{aligned}$$

Now, letting $\mathcal{T}_{i,\ell}$ be a batch of size at most $\Delta_i = \lceil (T/V_i^T)^{2/3} \rceil$ (as per the restart procedure of Algorithm 1), we obtain the local error bound

$$\begin{aligned} \sum_{t \in \mathcal{T}_{i,\ell}} \beta \|X_i^t - \hat{x}_i^t\|^2 &\leq \sum_{t \in \mathcal{T}_{i,\ell}} \langle v_i^t(X^t), p_i - X_i^t \rangle + \sum_{t \in \mathcal{T}_{i,\ell}} \langle v_i^t(X^t), \hat{x}_i^t - p_i \rangle \\ &\leq \operatorname{Gap}_i(\mathcal{T}_{i,\ell}) + \sum_{t \in \mathcal{T}_{i,\ell}} \langle v_i^t(X^t), \hat{x}_i^t - p_i \rangle. \end{aligned}$$

Hence, writing $\tau_{i,\ell} = \min \mathcal{T}_{i,\ell}$ for the first index of batch $\mathcal{T}_{i,\ell}$ and taking $p_i = \hat{x}_i^{\tau_{i,\ell}}$ as a reference action for the ℓ-th batch, we can bound the second term above as

$$\begin{aligned} \sum_{t \in \mathcal{T}_{i,\ell}} \langle v_i^t(X^t), \hat{x}_i^t - p_i \rangle &\leq G_i |\mathcal{T}_{i,\ell}| \max_{t \in \mathcal{T}_{i,\ell}} \|\hat{x}_i^t - \hat{x}_i^{\tau_{i,\ell}}\| \\ &\leq G_i \Delta_i \operatorname{Var}(\mathcal{T}_{i,\ell}, \hat{x}_i^{\tau_{i,\ell}}). \end{aligned}$$

Then, taking expectations and summing over all batches as in the proof of Lemma 4.2, we get

$$\mathbb{E}\left[\sum_{t=1}^{T} \|X_i^t - \hat{x}_i^t\|^2\right] \leq \frac{1}{\beta} \sum_{\ell=1}^{m_i} \mathbb{E}[\operatorname{Gap}_i(\mathcal{T}_{i,\ell})] + \frac{G_i}{\beta} \Delta_i V_i^T,$$

By Proposition 6.2, we have $\operatorname{Gap}_i(\mathcal{T}_{i,\ell}) = \mathcal{O}(\Delta_i^{1/2})$. Thus, with $\Delta_i = \mathcal{O}((T/V_i^T)^{2/3})$ and $m_i = \mathcal{O}(T/\Delta_i) = \mathcal{O}(T^{1/3} (V_i^T)^{2/3})$, we finally get

$$\mathbb{E}[\operatorname{err}_i(\mathcal{T})] = \mathcal{O}(m_i \Delta_i) + \mathcal{O}(\Delta_i V_i^T)$$

[14] For a rigorous statement and proof in the single-player setting, see the recent paper [4].

$$= \mathcal{O}(T^{1/3}(V_i^T)^{2/3} \cdot T^{1/3}(V_i^T)^{-1/3}) + \mathcal{O}((T/V_i^T)^{2/3} \cdot V_i^T)$$
$$= \mathcal{O}(T^{2/3}(V_i^T)^{1/3}),$$

as claimed. Our second assertion then follows by noting that $T^{2/3}(V_i^T)^{1/3} = o(T)$ if $V_i^T = o(T)$. □

Note that the strategy used to bound the tracking error depends on the variation of the sequence of Nash equilibria of each stage game $\mathcal{G}^t$. We emphasize that this does not mean that the players actually *know* this precise variation: it suffices to have a bound thereof (even pessimistic). For instance, such information could be available to a player who knows that the sequence of stage games encountered comes from a family of games that follow some sufficiently slow variation, but not the exact realization of the game (and, much less, their equilibria). It thus follows to reason that a sharper bound of this form leads to a better tracking error (as evidenced by the $(V_i^T)^{1/3}$ dependence of $\operatorname{err}_i(T)$ on the variation bound V_i^T).

7.2. **Convergence to Nash equilibrium.** We now turn to the case where the sequence of stage games $\mathcal{G}^t \equiv \mathcal{G}^t(\mathcal{N}, \mathcal{X}, u^t)$ converges to some (monotone) limit game $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{X}, u)$.[15] Formally, it will be convenient to characterize this convergence in terms of the quantity

$$B_i^t = \max_{x \in \mathcal{X}} \|v_i^t(x) - v_i(x)\|_*,$$

i.e., via the maximum difference in (individual) payoff gradients between stage t and the limit $t \to \infty$. We will then say that the sequence of games $\mathcal{G}^t$ *converges effectively* to $\mathcal{G}$ if

$$B^t \equiv \sum_{i \in \mathcal{N}} B_i^t \to 0 \quad \text{as } t \to 0. \tag{7.2}$$

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 variational characterization (2.2) shows that the Nash equilibria of a (concave) game are actually determined by the players' individual payoff gradients, not their payoff functions; as such, characterizing the convergence of a sequence of stage games in terms of payoff gradients is closer to the true primitives that define equilibrium behavior in our setting.

Now, as in the previous section, we will focus on the prox-based learning protocol outlined in Algorithm 2. However, since we are now interested in the convergence of the generated sequence of play X^t to a specific target point in $\mathcal{X}$, we will require in what follows that the Bregman divergence of $h = \sum_i h_i$ satsify the additional "reciprocity" condition

$$x^t \to p \quad \text{whenever} \quad D_h(p, x^t) \to 0, \tag{RC}$$

for every sequence of actions $x^t \in \mathcal{X}^\circ \equiv \prod_i \mathcal{X}_i^\circ$. This requirement is known in the literature as "Bregman reciprocity" [10, 27] and, essentially, it ensures that the sublevel sets of $D_h(p, \cdot)$ constitute a neighborhood basis for p in $\mathcal{X}$, i.e., every

[15] To be clear, we are not assuming that each stage game $\mathcal{G}^t$ is a priori monotone.

Bregman zone of the form $\mathbb{D}_\varepsilon(p) \equiv \{x \in \mathcal{X} : D_h(p,x) \leq \varepsilon\}$ contains some δ-ball $\mathbb{B}_\delta(p) = \{x \in \mathcal{X} : \|p - x\| \leq \delta\}$.[16]

With all this at hand, our main Nash equilibrium convergence result in this setting is as follows:

Theorem 7.2. *Let $\mathcal{G}^t \equiv \mathcal{G}^t(\mathcal{N}, \mathcal{X}, u^t)$ be a sequence of concave games converging to a strictly monotone limit game $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{X}, u)$ in the sense of (7.2). Assume further that each player follows Algorithm 2 with a prox-mapping satisfying (RC) and a step-size sequence γ^t such that*

$$\sum_{t=1}^{\infty} \gamma^t = \infty, \quad \sum_{t=1}^{\infty} (\gamma^t)^2 < \infty, \quad \text{and} \quad \sum_{t=1}^{\infty} \gamma^t B^t < \infty. \tag{7.3}$$

Then, with probability 1, the sequence of realized actions X^t converges to the (necessarily unique) Nash equilibrium $\hat{x}$ of the limit game $\mathcal{G}$.

PM

I changed the step-size to be common across all players, otherwise the proof of Proposition 7.4 fails. In particular, I did not see a way of proving that there exists a player $i \in \mathcal{N}$ such that $\langle v_i(X^t), X^t - \hat{x}_i \rangle < 0$ if X^t does not have $\hat{x}$ as a limit point (it's true for the sum, but not necessarily for any given player).

Before discussing the proof of Theorem 7.2, some remarks are in order: First, the requirement $\sum_{t=1}^{\infty} \gamma_i^t B_i^t < \infty$ should be interpreted as a bound on how slow the convergence of $\mathcal{G}^t$ can be in order for convergence to equilibrium to be guaranteed. For instance, as long as $B^t = \mathcal{O}(1/(\log t)^\varepsilon)$ for some $\varepsilon > 0$, the step-size conditions (7.9) can all be satisfied by taking $\gamma^t \propto 1/(t \log t)$. Second, as in Theorem 7.1, the players of the game are not required to know the exact value of B_i^t (which would require a very detailed knowledge of the game at hand): as in all our variation results so far, it suffices to work with an upper bound thereof (even a loose, pessimistic one).

Our proof strategy will be based on two intermediate results, both of independent interest. First, we will show that the sequence of generated actions converges (a.s.) to a level set of the Bregman divergence $D_h(\hat{x}, \cdot)$ relative to $\hat{x}$. Subsequently, we show that X^t cannot remain a bounded distance away from $\hat{x}$ for all sufficiently large t. Combining these results will then suffice to show that X^t can only converge to the zero-level set of the Bregman divergence, i.e., $\lim_{t\to\infty} X^t = \hat{x}$.

We begin by establishing the convergence of X^t to a level set of the Bregman divergence:

Proposition 7.3. *With assumptions as in Theorem 7.2, the Bregman divergence $D_h(\hat{x}, X^t)$ converges (a.s.) to a random variable D^∞ with $\mathbb{P}(D^\infty < \infty) = 1$.*

Proof. To begin, it will be convenient to decompose the signal process Y_i^t as

$$Y_i^t = v_i^t(X^t) + U_i^t = v_i(X^t) + U_i^t + b_i^t, \tag{7.4}$$

where we have set $b_i^t = v_i^t(X^t) - v_i(X^t)$. Then, letting $D_i^t = D_{h_i}(\hat{x}_i, X_i^t)$, the descent inequality (6.1) for the Bregman divergence readily yields

$$\begin{aligned} D_i^{t+1} &\leq D_i^t + \gamma^t \langle Y_i^t, X_i^t - \hat{x}_i \rangle + \frac{(\gamma^t)^2}{2K_i} \|Y_i^t\|_*^2 \\ &\leq D_i^t + \gamma^t \xi_i^t + \gamma^t \beta_i^t + \frac{(\gamma^t)^2}{2K_i} \|Y_i^t\|_*^2, \end{aligned} \tag{7.5}$$

where, in the second line, we have set $\xi_i^t = \langle U_i^t, X_i^t - \hat{x}_i \rangle$ and $\beta_i^t = \langle b_i^t, X_i^t - \hat{x}_i \rangle$, and we used the fact that $\hat{x}$ is a Nash equilibrium of the limit game $\mathcal{G}$ (implying in turn

[16] The converse to this condition (i.e., that $X^t \to p$ whenever $D_h(p, X^t) \to 0$) holds automatically as a simple consequence of the fact that $D_h(p,x) \geq (K/2)\|p - x\|^2$ (cf. Appendix A).

that $\langle v(x_i), x_i - \hat{x}_i \rangle \leq 0$ for all $x_i \in \mathcal{X}_i$ and all $i \in \mathcal{N}$). Thus, taking expectations, we obtain:

$$\begin{aligned}
\mathbb{E}[D_i^{t+1} \,|\, \mathcal{F}^{t-1}] &\leq \mathbb{E}\left[D_i^t + \xi_i^t + \beta_i^t + \frac{(\gamma^t)^2}{2K_i} \|Y_i^t\|_*^2 \,\middle|\, \mathcal{F}^{t-1} \right] \\
&\leq D_i^t + \gamma^t \,\mathbb{E}[\|b_i^t\|_* \|X_i^t - \hat{x}_i\| \,|\, \mathcal{F}^{t-1}] + \frac{(\gamma^t)^2}{2K} \mathbb{E}[\|Y_i^t\|_*^2 \,|\, \mathcal{F}^{t-1}] \\
&\leq D_i^t + \gamma^t B_i^t \operatorname{diam}(\mathcal{X}_i) + \frac{1}{2K} (\gamma^t)^2 M_i^2 \qquad (7.6)
\end{aligned}$$

where *a*) in the second line, we used the fact that X^t is predictable relative to $\mathcal{F}^t$, the definition of β_i^t, and the fact that $\mathbb{E}[U^t \,|\, \mathcal{F}^{t-1}] = 0$; and *b*) in the last line, we used (3.4a) and the definition of B_i^t.

To proceed, let $\varepsilon_i^t = \gamma^t B_i^t$, so the last line of (7.6) can be written as

$$\mathbb{E}[D^{t+1} \,|\, \mathcal{F}^{t-1}] \leq D_i^t + \varepsilon_i^t.$$

Consider now the auxiliary process $\zeta_i^t = D_i^{t+1} + \sum_{s=t+1}^{\infty} \varepsilon_i^s$. Then, taking expectations yields

$$\mathbb{E}[\zeta_i^t \,|\, \mathcal{F}^{t-1}] \leq D_i^t + \varepsilon_i^t + \sum_{s=t+1}^{\infty} \varepsilon_i^s = D_i^t + \sum_{s=t}^{\infty} \varepsilon_i^s = \zeta_i^{t-1},$$

i.e., ζ_i^t is a supermartingale relative to $\mathcal{F}^t$. Furthermore, since $\sum_{t=1}^{\infty} \varepsilon_i^t < \infty$ by the step-size assumption (7.9), we also get $\mathbb{E}[\zeta_i^t] \leq \mathbb{E}[\zeta_i^1] < \infty$, i.e., ζ_i^t is bounded in L^1. Thus, by Doob's (sub)martingale convergence theorem, it follows that ζ_i^t converges almost surely to some random variable ζ_i that is itself finite (almost surely and in L^1). Since $D_i^t = \zeta_i^{t-1} - \sum_{s=t}^{\infty} \varepsilon_i^s$ and $\lim_{t\to\infty} \sum_{s=t}^{\infty} \varepsilon_i^s = 0$ (again, by the step-size summability assumption), we conclude that D_i^t converges itself to ζ_i. Our claim then follows by noting that $D_h(\hat{x}, X^t) = \sum_i D_{h_i}(\hat{x}_i, X_i^t)$. □

Moving on, our next result shows that the sequence of play X^t gets arbitrarily close to the Nash equilibrium $\hat{x}$ of the limit game:

Proposition 7.4. *With probability* 1*, there exists a* (*random*) *subsequence* X^{t_k} *of* X^t *which converges to* $\hat{x}$*.*

Proof. Our proof is by contradiction. To that end, suppose that, with positive probability, the sequence of play X^t does not admit $\hat{x}$ as a limit point. Conditioning on this event, there exists a ball $\mathbb{B}_\delta(\hat{x})$ such that $X^t \notin \mathbb{B}_\delta(\hat{x})$ for all sufficiently large t, implying in turn that X^t is contained in some compact set $\mathcal{K} \subseteq \mathcal{X}$ such that $\hat{x} \notin \mathcal{K}$. By (2.2), we have $\langle v(x), x - \hat{x} \rangle < 0$ whenever $x \in \mathcal{K}$. Therefore, by the continuity of v and the compactness of $\mathcal{K}$, there exists some $c > 0$ such that

$$\langle v(x), x - \hat{x} \rangle \leq -c \quad \text{for all } x \in \mathcal{K}. \qquad (7.7)$$

To proceed, let $D^t = D_h(\hat{x}, X^t)$ as in the proof of Proposition 7.3. Then, telescoping (7.5) yields the estimate

$$D^{t+1} \leq D^1 + \sum_{s=1}^{t} \gamma^s \langle v(X^t), X^t - \hat{x} \rangle + \sum_{s=1}^{t} \gamma^s \xi^s + \sum_{s=1}^{t} \gamma^s \beta^s + \sum_{s=1}^{t} \frac{(\gamma^s)^2}{2K} \|Y^t\|_*^2,$$

where the strong convexity modulus K is defined as $K = \min_i K_i$, and, as in the proof of Proposition 7.3, we set $\xi^t = \langle U^t, X^t - \hat{x} \rangle$ and $\beta^t = \langle b^t, X^t - \hat{x} \rangle$. Hence, setting $S^t = \sum_{s=1}^t \gamma^s$ and using (7.7), we obtain

$$D^{t+1} \leq D^1 - S^t \left[c - \frac{\sum_{s=1}^t \gamma^s \xi^s}{S^t} - \frac{\sum_{s=1}^t \gamma^s \beta^s}{S^t} - \frac{(2K)^{-1} \sum_{s=1}^t (\gamma^s)^2 \|Y^s\|_*^2}{S^t} \right]. \tag{7.8}$$

We proceed to analyze this bound term-by-term:

- First, by definition, we have $\mathbb{E}[\xi^t \,|\, \mathcal{F}^{t-1}] = 0$, so the second term in the brackets of (7.8) is itself a martingale. Furthermore, by (3.4a), we have

$$\begin{aligned} \sum_{t=1}^\infty (\gamma^t)^2 \, \mathbb{E}[(\xi^t)^2 \,|\, \mathcal{F}^{t-1}] &\leq \sum_{t=1}^\infty (\gamma^t)^2 \|X^t - \hat{x}\|^2 \, \mathbb{E}[\|U^t\|_*^2 \,|\, \mathcal{F}^{t-1}] \\ &\leq \operatorname{diam}(\mathcal{X})^2 \sigma^2 \sum_{t=1}^\infty (\gamma^t)^2 < \infty. \end{aligned}$$

  Therefore, by the law of large numbers for martingale difference sequences [19, Theorem 2.18], we conclude that $(1/S^t) \sum_{s=1}^t \gamma^s \xi^s$ converges to 0 with probability 1.

- For the third term in the brackets of (7.8), we have $\beta^t \leq \sum_{i \in \mathcal{N}} \operatorname{diam}(\mathcal{X}_i) B_i^t$, so $\beta^t \to 0$ as $t \to \infty$. Since $\sum_{t=1}^\infty \gamma^t = \infty$, it follows that $\sum_{s=1}^t \gamma^s \beta^s \big/ S^t \to 0$.

- Finally, for the last term, let $R^t = (1/2K) \sum_{s=1}^t (\gamma^s)^2 \|Y^s\|_*^2$. We then have

$$\begin{aligned} \mathbb{E}[R^t \,|\, \mathcal{F}^{t-1}] &= \frac{1}{2K} \, \mathbb{E}\left[\sum_{s=1}^{t-1} (\gamma^s)^2 \|Y^s\|_*^2 + \frac{(\gamma^t)^2}{2K} \|Y^t\|_*^2 \,\middle|\, \mathcal{F}^{t-1} \right] \\ &= R^t + (\gamma^t)^2 \, \mathbb{E}[\|Y^t\|_*^2 \,|\, \mathcal{F}^{t-1}] \geq R^t, \end{aligned}$$

  i.e., R^t is a submartingale relative to $\mathcal{F}^t$. Furthermore, by the law of total expectation, we also have

$$\mathbb{E}[R^t] = \mathbb{E}[\mathbb{E}[R^t \,|\, \mathcal{F}^{t-1}]] \leq \frac{M^2}{2K} \sum_{s=1}^\infty (\gamma^s)^2 < \infty,$$

  where we set $M^2 = \sum_{i \in \mathcal{N}} M_i^2$. In turn, this implies that R^t is uniformly bounded in L^1 so, by Doob's (sub)martingale convergence theorem [19, Theorem 2.5], we conclude that R^t converges to some (almost surely finite) random variable R^∞ with $\mathbb{E}[R^\infty] < \infty$. Consequently, we get $\lim_{t \to \infty} R^t / S^t = 0$ with probability 1.

Combining all of the above, we infer that there exists some (possibly random, but almost surely finite) t_0 such that $D^t \leq D^1 - c/2 \cdot S^t$ for all $t \geq t_0$. In turn, this implies that $D_h(\hat{x}, X^t) \to -\infty$ with probability 1, a contradiction. Going back to our original assumption, this shows that, with probability 1, $\hat{x}$ is a limit point of X^t, so our proof is complete. □

With these two results at hand, we are finally in a position to prove our Nash equilibrium convergence theorem:

Proof of Theorem 7.2. Proposition 7.4 shows that, with probability 1, there exists a (possibly random) subsequence t_k such that $X^{t_k} \to \hat{x}$. By the reciprocity condition (RC), this implies that $\liminf_{t\to\infty} D_h(\hat{x}, X^t) = 0$ (a.s.). However, since $\lim_{t\to\infty} D_h(\hat{x}, X^t)$ exists with probability 1 by Proposition 7.3, it follows that

$$\lim_{t\to\infty} D_h(\hat{x}, X^t) = \liminf_{t\to\infty} D_h(\hat{x}, X^t) = 0$$

i.e., X^t converges to $\hat{x}$. □

7.3. **Convergence in two-player zero-sum games.** We close this section with a convergence result for two-player zero-sum games in the spirit of time-average guarantees that are common in the online learning literature. To state it, assume as in Example 2.2 that the sequence of stage games encountered by the players is determined by a sequence of smooth, convex-concave saddle functions $f^t\colon \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$ so that $u_1^t = -f^t = -u_2^t$. We then have:

Theorem 7.5. *Let f^t be a sequence of convex-concave saddle functions converging to $f\colon \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$ in the sense of (7.2). Assume further that each player follows Algorithm 2 with a step-size sequence γ^t such that*

$$\sum_{t=1}^{\infty} \gamma^t = \infty, \quad \sum_{t=1}^{\infty} (\gamma^t)^2 < \infty, \quad \textit{and} \quad \sum_{t=1}^{\infty} \gamma^t B^t < \infty. \tag{7.9}$$

Then, with probability 1, the sequence of ergodic averages $\bar{X}^t = \sum_{s=1}^t \gamma^s X^s \big/ \sum_{s=1}^t \gamma^s$ converges to the set of saddle-points of f.

Before discussing the proof of Theorem 7.5, some remarks are in order:

Remark 6. It should be noted that, if the limit game is strictly monotone (for instance, if f is *strictly* convex-concave), Theorem 7.5 is essentially subsumed by Theorem 7.2: if the sequence of play X^t converges to the game's unique equilibrium, then so does the corresponding ergodic average $\bar{X}^t = \sum_{s=1}^t \gamma^s X^s \big/ \sum_{s=1}^t \gamma^s$. On the other hand, this leaves open the non-strict case: for instance, if the game at hand is the mixed extension of a two-player zero-sum finite game (i.e., $f(x_1, x_2) = x_1^\top A x_2$ for some matrix A of appropriate dimensions), the limit game is *not* strictly monotone, so Theorem 7.2 does not apply (buTheorem 7.5 does).

Remark 7. We should also note that Theorem 7.5 does not invoke the Bregman reciprocity condition (RC). The reason for this is that the analysis of the ergodic average is not as delicate as that of the actual sequence of play, but this (technical) simplification comes at a price: specifically, Theorem 7.5 says little for the convergence of X^t. In fact, even in the static case ($f^t = f$ for all $t = 1, 2, \dots$), the sequence of chosen actions might be recurrent or cycle around the game's limit Nash equilibrium without converging [26, 29]: this is a qualitative difference in behavior which cannot be detected by the convergence of $\bar{X}^t$.

We now proceed with the proof of Theorem 7.5:

Proof of Theorem 7.5. Let $\hat{x} \in \mathcal{X}$ be a Nash equilibrium of the limit game induced by f, and, as in the proof of Proposition 7.4, let $D^t = D_h(\hat{x}, X^t)$. Then, working as in Eq. (7.5), we obtain the basic estimate

$$D^{t+1} \leq D^t + \gamma^t \langle Y^t, X^t - \hat{x} \rangle + \frac{(\gamma^t)^2}{2K} \|Y^t\|_*^2,$$

where we have set $K = \min\{K_1, K_2\}$. Then, decomposing the input signal Y^t as in (7.4) and rearranging, we get:

$$\gamma^t \langle v(X^t), \hat{x} - X^t \rangle \leq D^t - D^{t+1} + \gamma^t \langle U^t + b^t, X^t - \hat{x} \rangle + \frac{(\gamma^t)^2}{2K} \|Y^t\|_*^2.$$

Then, summing over t gives

$$\begin{aligned}\sum_{s=1}^{t} \gamma^s \langle v(X^s), \hat{x} - X^s \rangle &\leq D^1 - D^{t+1} + \sum_{s=1}^{t} \gamma^s \langle U^s + b^s, X^s - \hat{x} \rangle + \sum_{s=1}^{t} \frac{(\gamma^s)^2}{2K} \|Y^s\|_*^2 \\ &\leq D^1 + \sum_{s=1}^{t} \gamma^s \xi^s + \sum_{s=1}^{t} \gamma^s \beta^s + \frac{1}{2K} R^t\end{aligned}$$

where, as in the proof of Proposition 7.4, we set $\xi^t = \langle U^t, X^t - \hat{x} \rangle$, $\beta^t = \langle b^t, X^t - \hat{x} \rangle$, and $R^t = \sum_{s=1}^{t} (\gamma^s)^2 \|Y^s\|_*^2$. Hence, letting $S^t = \sum_{s=1}^{t} \gamma^s$ and arguing in the same way as in the proof of Proposition 7.4 (which has the same step-size requirements), we deduce that

$$\lim_{t\to\infty} \frac{\sum_{s=1}^{t} \gamma^s \xi^s}{S^t} = 0, \quad \lim_{t\to\infty} \frac{\sum_{s=1}^{t} \gamma^s \beta^s}{S^t} = 0, \quad \text{and} \quad \lim_{t\to\infty} \frac{R^t}{S^t} = 0,$$

with probability 1. On the other hand, given that f is convex-concave, we also have

$$\begin{aligned}\frac{\sum_{s=1}^{t} \gamma^s \langle v(X^s), \hat{x} - X^s \rangle}{S^t} &\geq u_1(\hat{x}_1, \hat{x}_2) - u_1(\bar{X}_1^t, \hat{x}_2) + u_2(\hat{x}_1, \hat{x}_2) - u_2(\hat{x}_1, \bar{X}_2^t) \\ &= f(\bar{X}_1^t, \hat{x}_2) - f(\hat{x}_1, \bar{X}_2^t).\end{aligned}$$

Therefore, combining all of the above, we conclude that

$$f(\bar{X}_1^t, \hat{x}_2) - f(\hat{x}_1, \bar{X}_2^t) \leq \frac{D^1}{S^t} + o(1),$$

i.e., $f(\bar{X}_1^t, \hat{x}_2) - f(\hat{x}_1, \bar{X}_2^t) \to 0$ as $t \to \infty$. Since $\hat{x}$ is a Nash equilibrium, this shows that $\bar{X}^t$ attains the value of f, i.e., $\bar{X}^t$ converges itself to the set of saddle-points of f, as claimed. □

8. Concluding remarks

There are many interesting points for future research. First, we have been very agnostic towards the data generating process of the game problem. With an eye towards simulation based solution techniques, it is important to study time-varying games generated by ergodic processes in the spirit of [13]. For monotone variational inequality problems subjected to Brownian noise, a first step in this direction has been done by [28]. More effort is needed to understand the asymptotic properties of the repeated game process in this approach, possibly also using different assumptions on the driving random process. We are currently investigating this issue.

PM

I don't understand what this means.

In view of applications to problems in engineering and control, the study of a bona fide continuous-time version of the present approach is also a priority. At a higher level, we have made many strong regularity assumptions on the time-varying games in this paper (concavity in own action, and differentiability). Relaxing the smoothness of the individual player function is an important extension of the present approach. Finally, introducing coupled constraints into the player's action set is an important and challenging extension of the present framework, which is also currently under investigation.

Appendix A. Prox-Mappings

In this appendix we collect some basic technical facts on the prox-method. In the following we let $\mathcal{C}$ be a convex compact domain in a finite-dimensional normed vector space $(\mathcal{V}, \|\cdot\|)$, and $h\colon \mathcal{C} \to \mathbb{R}$ be a distance generating function with convexity parameter K. The corresponding Bregman divergence is

$$D_h(x,x') = h(x) - h(x') - \langle \nabla h(x'), x - x' \rangle$$

for $x \in \operatorname{dom}(h)$, $x' \in \operatorname{dom}(h)^\circ$. Define $\Theta_h(a_i) = \max_{x\in\mathcal{C}} D_h(a_i, x_i)$, and

$$x^h = \operatorname{argmin}\{h(x) : x \in \mathcal{C}\}. \tag{A.1}$$

Note that x^h is uniquely defined thanks to the strong-convexity of h, and we have

$$\langle \nabla h(x^h), a - x^h \rangle \geq 0$$

for all $a \in \mathcal{C}$. From this it follows immediately that

$$\Theta(x^h) \leq \max_{x\in\mathcal{C}} h(x) - \min_{x\in\mathcal{C}} h(x) =: \mathcal{D}[\mathcal{C}; h]. \tag{A.2}$$

Furthermore, by K-strong convexity,

$$\frac{K}{2}\|x - x^h\|^2 \leq D_h(x, x^h) \leq \Theta_h(x^h). \qquad \forall x \in \mathcal{C}.$$

Hence,

$$\|x - x^h\| \leq \sqrt{\frac{2}{K}\mathcal{D}[\mathcal{C}, h]} \qquad \forall x \in \mathcal{C},$$

and

$$\mathcal{C} \subseteq \{a \in \mathcal{V} | \; \|a - x^h\| \leq \sqrt{2\mathcal{D}[\mathcal{C}, h]/K}\}. \tag{A.3}$$

Lemma A.1. *Let*

$$P(x,y) = \operatorname*{argmin}_{a\in\mathcal{A}}\{\langle y, a - x\rangle + D_h(a,x)\}.$$

Then

(1) $\mathcal{V}^* \ni s \mapsto P(x,y)$ is single-valued.
(2) $\|P(x,y') - P(x,y)\| \leq \frac{1}{K}\|y' - y\|_*$.
(3) For all $a, x \in \mathcal{A}$ and all $y, y' \in \mathcal{V}^*$, we have

$$D_h(a, P(x,y)) \leq D_h(a,x) + \langle y, a - P(x,y)\rangle - D_h(P(x,y), x). \tag{A.4}$$

Proof. (1) This is clear by strong convexity.

(2) Let $a = P(x,y)$ and $b = P(x,v)$. The optimality conditions at these points are

$$\langle \nabla h(a) - \nabla h(x) + s, x' - a \rangle \geq 0,$$

and

$$\langle \nabla h(b) - \nabla h(x) + v, x' - b \rangle \geq 0$$

for all $x, x' \in \mathcal{X}$. Evaluating the first inequality at the point $x' = b$ and the second inequality at the point $x' = a$ gives

$$\langle \nabla h(a) - \nabla h(x) + s, b - a \rangle \geq 0,$$
$$\langle \nabla h(b) - \nabla h(x) + v, b - a \rangle \leq 0.$$

Hence,

$$\langle \nabla h(a) - \nabla h(x) + s, b - a \rangle \geq \langle \nabla h(b) - \nabla h(x) + v, b - a \rangle$$

$$\Leftrightarrow \langle s - v, b - a\rangle \geq \langle \nabla h(b) - \nabla h(a), b - a\rangle \geq K\|a - b\|^2$$

where the last inequality uses the K-strong convexity of the distance generating function h. Hence,

$$\|s - v\|_* \geq K\|a - b\|.$$

(3) The three-point identity [10] gives

$$D_h(a, x) - D_h(a, c) - D_h(c, x) = \langle \nabla h(c) - \nabla h(x), a - c\rangle$$

Combined with the optimality condition satisfied by the point $c = P(x, y)$, we get

$$\begin{aligned} D_h(a, x) - D_h(a, P(x, y)) - D_h(P(x, y), x) &= \langle \nabla h(P(x, y)) - \nabla h(x), a - P(x, y)\rangle \\ &\geq \langle -y, a - P(x, y)\rangle. \end{aligned}$$

Rearranging gives the desired inequality.

□

Eq. (A.4) provides the first step to derive regret bounds for the prox-method as done in the main text. Using the Fenchel Young inequality

$$\langle y, a - b\rangle \leq \frac{1}{2K}\|y\|_*^2 + \frac{K}{2}\|a - b\|^2$$

this inequality can be refined to

$$\begin{aligned} D_h(a, P(x, y)) - D_h(a, x) &\leq \langle y, a - x\rangle + \langle y, x - P(x, y)\rangle - D_h(P(x, y), x) \\ &\leq \langle y, a - x\rangle + \frac{1}{2K}\|y\|_*^2. \end{aligned} \tag{A.5}$$

The next Lemma provides a slight refinement of the previous one.

Lemma A.2. *Let $\{v^t\}_{t\in\mathbb{N}}$ be a sequence in $\mathcal{V}^*$. Define the process $\{Y^t\}_{t\in\mathbb{N}}$ by*

$$Y^{t+1} = P(Y^t, v^t), \quad Y^1 \in \mathcal{C}^\circ \text{ given.}$$

Then, for all $T \geq 1$ and all $a \in \mathcal{C}^\circ$, we have

$$\sum_{t=1}^{T}\langle v^t, Y^t - a\rangle \leq D_h(a, Y^1) + \frac{1}{2K}\sum_{t=1}^{T}\|v^t\|_*^2.$$

Proof. From Lemma A.1 and Eq. (A.5), we get

$$D_h(a, Y^{t+1}) \leq D_h(a, Y^t) + \langle v^t, a - Y^t\rangle + \frac{\|v^t\|_*^2}{2K}.$$

Rearranging and telescoping gives

$$\begin{aligned} \sum_{t=1}^{T}\langle v^t, Y^t - a\rangle &\leq D_h(a, Y^1) - D_h(a, Y^{T+1}) + \frac{1}{2K}\sum_{t=1}^{T}\|v^t\|_*^2 \\ &\leq D_h(a, Y^1) + \frac{1}{2K}\sum_{t=1}^{T}\|v^t\|_*^2. \end{aligned}$$

The last inequality uses the non-negativity of the Bregman divergence. □

Appendix B. Proof of Proposition 6.4

The purpose of this section is to prove Proposition 6.4, under the assumption that the signal process satisfies Eq. (6.8). We need two intermediate technical results.

Lemma B.1. *Define*

$$\Phi_{i,T} = \sum_{t=1}^{T} \frac{(\gamma_i^t)^2}{2K_i} \left(\|Y_i^t\|_*^2 + \|U_i^t\|_*^2 \right),$$

and let $\xi_i^t = \langle \gamma_i^t U_i^t, X_i^t - Z_i^t \rangle$. *Set*

$$\Xi_{i,T} = \sum_{t=1}^{T} \xi_i^t. \tag{B.1}$$

Then, for all constants $C_1, C_2 > 0$, *we have*

$$\mathbb{P}(\Phi_{i,T} + \Xi_{i,T} \geq (1 + C_1)\Gamma_{i,T} + C_2 \Psi_{i,T}) \leq \exp(-C_1) + \exp(-C_2^2/4). \tag{B.2}$$

where $\Gamma_{i,T} := \sum_{t=1}^{T} \gamma_i^t$, *and* $\Psi_{i,T} = 4M_* \sqrt{2\mathcal{D}_i[\mathcal{X}_i; h_i]/K_i} \sqrt{\sum_{t=1}^{T} (\gamma_i^t)^2}$.

Proof. By definition we have

$$\begin{aligned} \Phi_{i,T} &= \sum_{t=1}^{T} \frac{(\gamma_i^t)^2}{2K_i} \left(\|Y_i^t\|_*^2 + \|U_i^t\|_*^2 \right) \\ &\leq \sum_{t=1}^{T} \frac{(\gamma_i^t)^2}{2K_i} \left(3\|Y_i^t\|_*^2 + 2M_*^2 \right) \end{aligned}$$

Calling $\gamma_i^t = 5M_*^2 (\gamma_i^t)^2/(2K_i)$, gives

$$\mathbb{E}\left[\exp\left(\frac{3(\gamma_i^t)^2 \|Y_i^t\|_*^2}{2K_i \gamma_i^t} \right) \right] \leq \exp(3/5)$$

and

$$\mathbb{E}\left[\exp\left(\frac{(\gamma_i^t)^2 M_*^2}{K_i \gamma_i^t} \right) \right] \leq \exp(2/5).$$

Hence,

$$\mathbb{E}\left[\exp\left(\frac{3(\gamma_i^t)^2 \|Y_i^t\|_*^2 + 2(\gamma_i^t)^2 M_*^2}{2K_i \gamma_i^t} \right) \right] \leq \exp(1).$$

Call $\Gamma_{i,T} := \sum_{t=1}^{T} \gamma_i^t$. Then, Jensen's inequality shows that[17]

$$\mathbb{E}[\exp(\Phi_{i,T}/\Gamma_{i,T})] \leq \exp(1).$$

Therefore, for all $C_1 > 0$, Markov's inequality readily implies

$$\begin{aligned} \mathbb{P}(\Phi_{i,T} \geq (1 + C_1)\Gamma_{i,T}) &= \mathbb{P}(\exp(\Phi_{i,T}/\Gamma_{i,T}) \geq \exp(1 + C_1)) \\ &\leq \exp(-1 - C_1)\mathbb{E}\left[\exp(\Phi_{i,T}/\Gamma_{i,T})\right] \end{aligned}$$

[17] The convexity of the mapping $x \mapsto \exp(x)$ shows the following: Let $\{a_t\}_{t \geq 1}, \{b_t\}_{t \geq 1}$ be sequences in $(0, \infty)$. Then, by Jensen's inequality, we have

$$\sum_{t=1}^{T} \frac{a_t}{\sum_{\ell=1}^{T} a_\ell} \exp\left(\frac{b_t}{a_t} \right) \geq \exp\left(\sum_{t=1}^{T} \frac{b_t}{\sum_{\ell=1}^{T} a_\ell} \right).$$

We apply this inequality with the identification $b_t = \frac{3(\gamma_i^t)^2}{2K_i} \|Y_i^t\|_*^2 + \frac{(\gamma_i^t)^2}{K_i} M_*^2$ and $a_t = \gamma_i^t$.

$$\leq \exp(-C_1). \tag{B.4}$$

Now, let $\xi_i^t = \langle \gamma_i^t U_i^t, X_i^t - Z_i^t \rangle$ and set $\Xi_{i,T} = \sum_{t=1}^T \xi_i^t$. Observe that $\mathbb{E}[\xi_i^t \mid \mathcal{F}_{t-1}] = 0$ for all $t \geq 1$. Therefore $\Xi_{i,T}$ is a martingale with respect to the filtration $\mathbb{F} := \{\mathcal{F}_t\}_{t\geq 1}$, which is also bounded in $L^2(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, thanks to (6.9).

Via the Cauchy-Schwarz inequality, the K_i-strong convexity of the distance generating function, as well as eqs. (A.1) and (A.3), we see that

$$\begin{aligned} |\xi_i^t| &\leq \gamma_i^t \|U_i^t\|_* \cdot \|X_i^t - Y_i^t\|_i \\ &\leq \gamma_i^t \|U_i^t\|_* \left[\|X_i^t - x^{h_i}\|_i + \|x^{h_i} - Y_i^t\|_i\right] \\ &\leq 2\gamma_i^t \|U_i^t\|_* \sqrt{\frac{2}{K_i}\mathcal{D}[\mathcal{X}_i, h_i]} \end{aligned}$$

Hence,

$$\|U_i^t\|_*^2 \geq \frac{K_i |\xi_i^t|^2}{8(\gamma_i^t)^2 \mathcal{D}[\mathcal{X}_i, h_i]}.$$

Consequently,

$$\mathbb{E}\left[\exp\left(\frac{K_i|\xi_i^t|^2}{32(\gamma_i^t)^2\mathcal{D}[\mathcal{X}_i, h_i] M_*^2}\right) \middle| \mathcal{F}_t\right] \leq \mathbb{E}\left[\exp\left(\frac{\|U_i^t\|_*^2}{4M_*^2}\right)\right].$$

For all $t \geq 1$, denote by $\tau_i^t := 4\gamma_i^t M_* \sqrt{2\mathcal{D}[\mathcal{X}_i, h_i]/K_i}$. Using (6.9), this shows that

$$\mathbb{E}\left[\exp\left((\xi_i^t/\tau_i^t)^2\right) \middle| \hat{\mathcal{F}}_t\right] \leq \exp(1).$$

Since $\Xi_{i,t} = \Xi_{i,t-1} + \xi_i^t$, we get for all $\delta > 0$

$$\mathbb{E}[\exp(\delta \Xi_{i,t})] = \mathbb{E}[\exp(\delta \Xi_{i,t-1}) \exp(\delta \xi_i^t)] = \mathbb{E}[\exp(\delta \Xi_{i,t-1}) \mathbb{E}[\exp(\delta \xi_i^t) \mid \hat{\mathcal{F}}_t]]$$

Following [33], we see that for all $\delta > 0$ and $t \geq 1$

$$\mathbb{E}[\exp(\delta \xi_i^t) \mid \hat{\mathcal{F}}_t] \leq \exp\left(\delta^2 (\tau_i^t)^2\right).$$

Proceeding by induction, we observe that for all $\delta > 0$,

$$\mathbb{E}[\exp(\delta \Xi_{i,T})] \leq \exp\left(\delta^2 \sum_{t=1}^T (\tau_i^t)^2\right).$$

Therefore, if we set $\Psi_{i,T} = \sqrt{\sum_{t=1}^T (\tau_i^t)^2}$, Markov's inequality yields the immediate bound

$$\begin{aligned} \mathbb{P}(\Xi_{i,T} \geq C_2 \Psi_{i,T}) &\leq \exp\left(-\delta C_2 \Psi_{i,T}\right) \mathbb{E}[\exp(\delta \Xi_{i,T})] \\ &\leq \exp\left(-\delta C_2 \Psi_{i,T} + \delta^2 \Psi_{i,T}^2\right). \end{aligned}$$

Then, setting $\delta = \frac{C_2}{2\Psi_{i,T}}$ yields

$$\mathbb{P}(\Xi_{i,T} \geq C_2 \Psi_{i,T}) \leq \exp(-C_2^2/2 + C_2^2/4) = \exp(-C_2^2/4)$$

for all $C_2 > 0$. Combining this with the bound (B.4), we conclude that for all $C_1, C_2 > 0$,

$$\mathbb{P}(\Phi_{i,T} + \Xi_{i,T} \geq (1 + C_1)\Gamma_{i,T} + C_2 \Psi_{i,T}) \leq \exp(-C_1) + \exp(-C_2^2/4).$$

To see this, introduce the events $E_3 = \{\Xi_{i,T} + \Phi_{i,T} \geq (1 + C_1)\Gamma_{i,T} + C_2 \Psi_{i,T}\}$, $E_1 = \{\Phi_{i,T} \geq (1 + C_1)\Gamma_{i,T}\}$ and $E_2 = \{\Xi_{i,T} \geq C_2 \Psi_{i,T}\}$. Then $E_3 \subseteq E_1 \cup E_2$, so that $\mathbb{P}(E_3) \leq \mathbb{P}(E_1 \cup E_2) \leq \mathbb{P}(E_1) + \mathbb{P}(E_2)$. □

Lemma B.2. *For $C > 0$, define*

$$\begin{aligned}\mathcal{Q}_{i,T}(C) &:= 2\mathcal{D}[\mathcal{X}_i, h_i] + (1+C)\Gamma_{i,T} + 2\sqrt{C}\Psi_{i,T} \\ &= 2\mathcal{D}[\mathcal{X}_i, h_i] + (1+C)\frac{5}{2K_i}M_*^2\sum_{t=1}^{T}(\gamma_i^t)^2 \\ &\quad + 8\sqrt{2C\mathcal{D}[\mathcal{X}_i, h_i]/K_i}M_*\sqrt{\sum_{t=1}^{T}(\gamma_i^t)^2}.\end{aligned}$$

For all $T \geq 1$ and for all $\varepsilon \in (0,1)$, with probability at least $1-\varepsilon$, we have

$$\operatorname{Gap}_i(\mathcal{T}) \leq \mathcal{Q}_{i,T}(\log(2/\varepsilon)).$$

Proof. Observe that $\mathbb{E}[\xi_i^t \,|\, \mathcal{F}_{t-1}] = 0$ for all $t \geq 1$. Therefore $\Xi_{i,T}$, defined in (B.1), is a martingale with respect to the filtration $\mathbb{F} := \{\mathcal{F}_t\}_{t\geq 1}$, which is also bounded in $L^2(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, thanks to (6.9).

Now, Eq. (6.5) implies that

$$\operatorname{Gap}_i(\mathcal{T}) \leq 2\mathcal{D}[\mathcal{X}_i, h_i] + \Phi_{i,T} + \Xi_{i,T},$$

so $\{\operatorname{Gap}_i(\mathcal{T}) \geq \mathcal{Q}_{i,T}(C)\} \subseteq \{\Phi_{i,T} + \Xi_{i,T} \geq (1+C)\Gamma_{i,T} + 2\sqrt{C}\Psi_{i,T}\}$. Consequently, from Lemma B.1, we deduce that for all $C > 0$,

$$\mathbb{P}(\operatorname{Gap}_i(\mathcal{T}) \geq \mathcal{Q}_{i,T}(C)) \leq 2\exp(-C).$$

Choosing $C = \log(2/\varepsilon)$ proves our claim. □

Proof of Proposition 6.4. From the variational characterization (4.4) for the external regret, the Prox-strategy with a constant step-size $\gamma_i^t \equiv \gamma_i$ gives

$$\operatorname{Reg}_i(\mathcal{T}) \leq \max_{x_i \in \mathcal{X}_i}\sum_{t=1}^{T}\langle v_i^t(X^t), x_i - X_i^t\rangle \leq \frac{2\mathcal{D}[\mathcal{X}_i, h_i]}{\gamma_i} + \frac{1}{\gamma_i}(\Phi_{i,T} + \Xi_{i,T}).$$

Hence, for all $\rho > 0$,

$$\{\operatorname{Reg}_i(\mathcal{T}) \geq \rho\} \subseteq \{\Phi_{i,T} + \Xi_{i,T} \geq \gamma_i\rho - 2\mathcal{D}[\mathcal{X}_i, h_i]\}.$$

Therefore, choosing $\rho = \mathcal{Q}_{i,T}(C)/\gamma_i$ we deduce from Eq. (B.2) that

$$\mathbb{P}(\operatorname{Reg}_i(\mathcal{T}) \geq \mathcal{Q}_{i,T}(C)/\gamma_i) \leq \mathbb{P}(\Phi_{i,T} + \Xi_{i,T} \geq (1+C)\Gamma_{i,T} + 2\sqrt{C}\Psi_{i,T}) \leq 2\exp(-C).$$

Picking $C = \log(2/\varepsilon)$, for any $\varepsilon \in (0,1)$ fixed, we get the desired $(1-\varepsilon)$-probability bound. Now, observe that for a constant step-size, we have

$$\begin{aligned}\frac{\mathcal{Q}_{i,T}(\log(2/\varepsilon))}{\gamma_i} &= \frac{2\mathcal{D}[\mathcal{X}_i, h_i]}{\gamma_i} + \frac{5(1+\log(2/\varepsilon))M_*^2}{2K_i}\gamma_i T \\ &\quad + 8M_*\sqrt{2\log(2/\varepsilon)\mathcal{D}[\mathcal{X}_i, h_i]/K_i}\sqrt{T}.\end{aligned}$$

Call $\Omega_i(\varepsilon) := \frac{5(1+\log(2/\varepsilon))M_*^2}{2K_i}$, and optimizing the above expression with respect to γ_i, gives the optimal constant step-size

$$\gamma_i = \sqrt{\frac{2\mathcal{D}[\mathcal{X}_i, h_i]}{\Omega_i(\varepsilon)T}}.$$

Using this step size in the previous display gives

$$\frac{\mathcal{Q}_{i,T}(\log(2/\varepsilon))}{\gamma_i} = 2\sqrt{2T\mathcal{D}[\mathcal{X}_i, h_i]\Omega_i(\varepsilon)} + 8M_*\sqrt{2T\log(2/\varepsilon)\mathcal{D}[\mathcal{X}_i, h_i]/K_i}.$$

This shows that, with probability at least $1 - \varepsilon$, we have

$$\mathrm{Reg}_i(\mathcal{T}) \leq 2\sqrt{2T\mathcal{D}[\mathcal{X}_i, h_i]\Omega_i(\varepsilon)} + 8M_* \sqrt{2T \log(2/\varepsilon)\mathcal{D}[\mathcal{X}_i, h_i]/K_i}$$

and our proof is complete. □

References

[1] Arora, Sanjeev, Elad Hazan, Satyen Kale. 2012. The multiplicative weights update method: A meta-algorithm and applications. *Theory of Computing* **8**(1) 121–164.

[2] Auer, Peter, Nicolò Cesa-Bianchi, Yoav Freund, Robert E. Schapire. 1995. Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*.

[3] Beck, Amir, Marc Teboulle. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters* **31**(3) 167–175.

[4] Besbes, Omar, Yonatan Gur, Assaf Zeevi. 2015. Non-stationary stochastic optimization. *Operations Research* **63**(5) 1227–1244. doi:10.1287/opre.2015.1408. URL http://dx.doi.org/10.1287/opre.2015.1408.

[5] Bravo, Mario, David S. Leslie, Panayotis Mertikopoulos. 2018. Bandit learning in concave N-person games. *NIPS '18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

[6] Bubeck, Sébastien, Nicolò Cesa-Bianchi. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* **5**(1) 1–122.

[7] Cesa-Bianchi, Nicolò, Pierre Gaillard, Gábor Lugosi, Gilles Stoltz. 2012. Mirror descent meets fixed share (and feels no regret). 989-997, ed., *Advances in Neural Information Processing Systems*, vol. 25.

[8] Cesa-Bianchi, Nicolò, Gábor Lugosi. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

[9] Cesa-Bianchi, Nicolò, Gábor Lugosi, Gilles Stoltz. 2006. Regret minimization under partial monitoring. *Mathematics of Operations Research* **31**(3) 562–580. doi:10.1287/moor.1060.0206. URL https://doi.org/10.1287/moor.1060.0206.

[10] Chen, G., M. Teboulle. 1993. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization* **3**(3) 538–543. doi:10.1137/0803026. URL https://doi.org/10.1137/0803026.

[11] Cohen, Johanne, Amélie Héliou, Panayotis Mertikopoulos. 2017. Learning with bandit feedback in potential games. *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*.

[12] Debreu, Gerard. 1952. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences* **38**(10) 886–893.

[13] Duchi, J., A. Agarwal, M. Johansson, M. Jordan. 2012. Ergodic mirror descent. *SIAM Journal on Optimization* **22**(4) 1549–1578. doi:10.1137/110836043. URL https://doi.org/10.1137/110836043.

[14] Facchinei, Francisco, Andreas Fischer, Veronica Piccialli. 2007. On generalized nash games and variational inequalities. *Operations Research Letters* **35**(2) 159–164. doi:https://doi.org/10.1016/j.orl.2006.03.004. URL http://www.sciencedirect.com/science/article/pii/S0167637706000484.

[15] Facchinei, Francisco, Jong-shi Pang. 2003. *Finite-Dimensional Variational Inequalities and Complementarity Problems - Volume I and Volume II*. Springer Series in Operations Research.

[16] Flaxman, Abraham D., Adam Tauman Kalai, H. Brendan McMahan. 2005. Online convex optimization in the bandit setting: gradient descent without a gradient. *SODA '05: Proceedings of the 16th annual ACM-SIAM Symposium on Discrete Algorithms*. 385–394.

[17] Freund, Yoav, Robert E. Schapire. 1999. Adaptive game playing using multiplicative weights. *Games and Economic Behavior* **29** 79–103.

[18] Hall, Eric C, Rebecca M Willett. 2015. Online convex optimization in dynamic environments. *IEEE Journal of Selected Topics in Signal Processing* **9**(4) 647–662

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

[20] Hofbauer, Josef, William H. Sandholm. 2009. Stable games and their dynamics. *Journal of Economic Theory* **144**(4) 1665–1693.

[21] Jadbabaie, Ali, Alexander Rakhlin, Shahin Shahrampour, Karthik Sridharan. 2015. *Online optimization: Competing with dynamic comparators*.

[22] Juditsky, Anatoli, Arkadi Nemirovski, Claire Tauvel. 2011. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems* **1**(1) 17–58.

[23] Kannan, A., U. Shanbhag. 2012. Distributed computation of equilibria in monotone nash games via iterative regularization techniques. *SIAM Journal on Optimization* **22**(4) 1177–1205. doi:10.1137/110825352. URL https://doi.org/10.1137/110825352.

[24] Lugosi, Gábor, Shie Mannor, Gilles Stoltz. 2008. Strategies for prediction under imperfect monitoring. *Mathematics of Operations Research* **33**(3) 513–528.

[25] Mertikopoulos, Panayotis, E Veronica Belmega, Romain Negrel, Luca Sanguinetti. 2017. Distributed stochastic optimization via matrix exponential learning. *IEEE Transactions on Signal Processing* **65**(9) 2277–2290.

[26] Mertikopoulos, Panayotis, Christos H. Papadimitriou, Georgios Piliouras. 2018. Cycles in adversarial regularized learning. *SODA '18: Proceedings of the 29th annual ACM-SIAM Symposium on Discrete Algorithms*.

[27] Mertikopoulos, Panayotis, Mathias Staudigl. 2018. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization* **28**(1) 163–197.

[28] Mertikopoulos, Panayotis, Mathias Staudigl. 2018. Stochastic mirror descent dynamics and their convergence in monotone variational inequalities. *Journal of Optimization Theory and Applications* (to appear).

[29] Mertikopoulos, Panayotis, Houssam Zenati, Bruno Lecouat, Chuan-Sheng Foo, Vijay Chandrasekhar, Georgios Piliouras. 2018. Mirror descent in saddle-point problems: Going the extra (gradient) mile. https://arxiv.org/abs/1807.02629.

[30] Mertikopoulos, Panayotis, Zhengyuan Zhou. 2018. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming* doi:10.1007/s10107-018-1254-8. URL https://doi.org/10.1007/s10107-018-1254-8.

[31] Monderer, Dov, Lloyd S. Shapley. 1996. Potential games. *Games and Economic Behavior* **14**(1) 124 – 143.

[32] Nemirovski, Arkadi, Shmuel Onn, Uriel G. Rothblum. 2009. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research* **35**(1) 52–78. doi:10.1287/moor.1090.0427. URL https://doi.org/10.1287/moor.1090.0427.

[33] Nemirovski, Arkadi Semen, Anatoli Juditsky, Guangui (George) Lan, Alexander Shapiro. 2009. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* **19**(4) 1574–1609.

[34] Nemirovski, Arkadi Semen, David Berkovich Yudin. 1983. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, NY.

[35] Nesterov, Yurii. 2009. Primal-dual subgradient methods for convex problems. *Mathematical Programming* **120**(1) 221–259.

[36] Nikaido, Hukukane, Kazuo Isoda. 1955. Note on non-cooperative convex games. *Pacific Journal of Mathematics* **5** 807–815. URL https://projecteuclid.org:443/euclid.pjm/1171984836.

[37] Orda, Ariel, Raphael Rom, Nahum Shimkin. 1993. Competitive routing in multi-user communication networks. *IEEE/ACM Trans. Netw.* **1**(5) 614–627.

[38] Palaiopanos, Gerasimos, Ioannis Panageas, Georgios Piliouras. 2017. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*.

[39] Rockafellar, Ralph Tyrrell. 1970. *Convex Analysis*. Princeton University Press, Princeton, NJ.

[40] Rosen, J Ben. 1965. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society* 520–534.

[41] Rustichini, Aldo. 1999. Minimizing regret: The general case. *Games and Economic Behavior* **29**(1) 224 – 243. doi:https://doi.org/10.1006/game.1998.0690. URL http://www.sciencedirect.com/science/article/pii/S089982569890690X.

[42] Sandholm, William H. 2015. Population games and deterministic evolutionary dynamics. H. Peyton Young, Shmuel Zamir, eds., *Handbook of Game Theory IV*. Elsevier, 703–778.

[43] Scutari, Gesualdo, Daniel P Palomar, Francisco Facchinei, Jong-shi Pang. 2010. Convex optimization, game theory, and variational inequality theory. *IEEE Signal Processing Magazine* **27**(3) 35–49 1053–5888.

[44] Shahrampour, Shahin, Ali Jadbabaie. 2018. Distributed online optimization in dynamic environments using mirror descent. *IEEE Transactions on Automatic Control* **63**(3) 714–725.

[45] Shalev-Shwartz, Shai. 2011. Online learning and online convex optimization. *Foundations and Trends in Machine Learning* **4**(2) 107–194.

[46] Sorin, Sylvain, Cheng Wan. 2016. Finite composite games: Equilibria and dynamics. *Journal of Dynamics and Games* **3**(1) 101–120.

[47] Spall, James C. 1997. A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica* **33**(1) 109–112.

[48] Teboulle, Marc. 1992. Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research* **17** 670–690.

[49] Tsuda, Koji, Gunnar Rätsch, Manfred K. Warmuth. 2005. Matrix exponentiated gradient updates for on-line Bregman projection. *Journal of Machine Learning Research* **6** 995–1018.

[50] Viossat, Yannick, Andriy Zapechelnyuk. 2013. No-regret dynamics and fictitious play. *Journal of Economic Theory* **148**(2) 825–842.

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