Maastricht University

Multiagent Online Learning in Time-Varying Games

Citation for published version (APA):

Duvocelle, B., Mertikopoulos, P., Staudigl, M., & Vermeulen, D. (2023). Multiagent Online Learning in Time-Varying Games. *Mathematics of Operations Research*, *48*(2), 914-941. https://doi.org/10.1287/moor.2022.1283

Document status and date:
Published: 01/05/2023

DOI:
10.1287/moor.2022.1283

Document Version:
Publisher's PDF, also known as Version of record

Document license:
Taverne

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

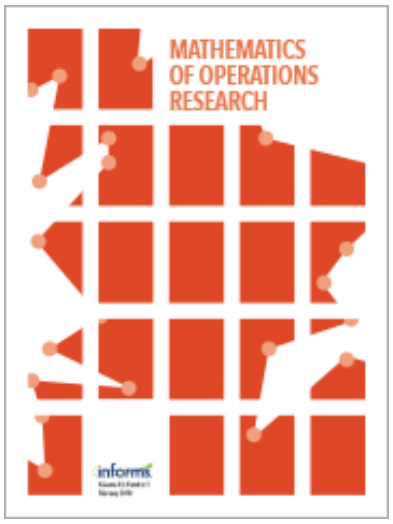
Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 01 May. 2024

This article was downloaded by: [137.120.148.153] On: 09 September 2022, At: 01:56
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA



Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Multiagent Online Learning in Time-Varying Games

Benoit Duvocelle, Panayotis Mertikopoulos, Mathias Staudigl, Dries Vermeulen

To cite this article:
Benoit Duvocelle, Panayotis Mertikopoulos, Mathias Staudigl, Dries Vermeulen (2022) Multiagent Online Learning in Time-Varying Games. Mathematics of Operations Research

Published online in Articles in Advance 01 Jul 2022

. https://doi.org/10.1287/moor.2022.1283

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2022, INFORMS

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

http://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Articles in Advance, pp. 1–28
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Multiagent Online Learning in Time-Varying Games

Benoit Duvocelle,[a] Panayotis Mertikopoulos,[b,c] Mathias Staudigl,[d,*] Dries Vermeulen[a]

[a] Department of Quantitative Economics, Maastricht University, NL–6200 MD Maastricht, Netherlands; [b] Université Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France; [c] Criteo AI Lab, 38130 Echirolles, France; [d] Department of Advanced Computing Sciences, Maastricht University, NL–6200 MD Maastricht, Netherlands
*Corresponding author

Contact: b.duvocelle@maastrichtuniversity.nl, https://orcid.org/0000-0002-7191-389X (BD); panayotis.mertikopoulos@imag.fr, https://orcid.org/0000-0003-2026-9616 (PM); m.staudigl@maastrichtuniversity.nl, https://orcid.org/0000-0003-2481-0019 (MS); d.vermeulen@maastrichtuniversity.nl (DV)

Received: September 8, 2018
Revised: May 19, 2020; September 7, 2021; March 7, 2022
Accepted: May 7, 2022
Published Online in Articles in Advance: July 1, 2022

MSC2020 Subject Classification: Primary: 91A10, 91A26; secondary: 68Q32, 90C25

https://doi.org/10.1287/moor.2022.1283

Copyright: © 2022 INFORMS

Abstract. We examine the long-run behavior of multiagent online learning in games that evolve over time. Specifically, we focus on a wide class of policies based on mirror descent, and we show that the induced sequence of play (a) converges to a Nash equilibrium in time-varying games that stabilize in the long run to a strictly monotone limit, and (b) it stays asymptotically close to the evolving equilibrium of the sequence of stage games (assuming they are strongly monotone). Our results apply to both gradient- and payoff-based feedback—that is, when players only get to observe the payoffs of their chosen actions.

Funding: This research was partially supported by the European Cooperation in Science and Technology COST Action [Grant CA16228] "European Network for Game Theory" (GAMENET). P. Mertikopoulos is grateful for financial support by the French National Research Agency (ANR) in the framework of the "Investissements d'avenir" program [Grant ANR-15-IDEX-02], the LabEx PERSYVAL [Grant ANR-11-LABX-0025-01], MIAI@Grenoble Alpes [Grant ANR-19-P3IA-0003], and the ALIAS [Grant ANR-19-CE48-0018-01].

Keywords: dynamic regret • Nash equilibrium • mirror descent • time-varying games

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

1. Introduction

Consider a repeated multiagent decision process that unfolds as follows:

1. At each stage $t = 1, 2, \dots$, every agent selects an action from some continuous set.

2. Each agent receives a reward based on the chosen action and the actions of all other players. These rewards are determined by a normal form game $\mathcal{G}_t$ that evolves over time and is a priori unknown to the players.

3. Based on the reward received (and/or any other payoff-relevant information), the players update their actions and the process repeats.

The main questions that we seek to address in this paper are the following: First, are there online learning policies that allow players to track a Nash equilibrium (NE) over time (or to converge to one if the stage games stabilize)? And, if so, what is the impact of the information available to the players and the variability of the sequence of stage games?

1.1. Background

One of the most widely used policies for learning in games is the *mirror descent* (MD) class of algorithms and its variants (cf. Bubeck and Cesa-Bianchi [14], Shalev-Shwartz [57], and references therein). This family of first order methods dates back to Nemirovski and Yudin [45], and contains as special cases standard (sub)gradient descent methods; entropic gradient descent (Beck and Teboulle [4]); the "hedge" (or exponential/multiplicative weights) algorithm in finite games (Auer et al. [3], Littlestone and Warmuth [38], Vovk [67]); and in games with a linear payoff structure, the follow-the-regularized-leader (FTRL) class of policies (Shalev-Shwartz [57], Shalev-Shwartz and Singer [58]). These methods are applied to a wide range of games—from min-max to potential games—leading to a vast literature that is impossible to survey here; for an appetizer, see Juditsky et al. [32], Nemirovski et al. [46], Nesterov [47], and references therein.

In the single-player case, the standard figure of merit is the minimization of the learner's *regret*, that is, the cumulative payoff difference between the player's chosen policy and the best policy in hindsight (static or dynamic, depending on the precise notion of regret under consideration). In this context, when the payoff functions encountered by the learner are concave, MD methods guarantee an $\mathcal{O}(\sqrt{T})$ static regret bound that is

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

well-known to be order-optimal (Abernethy et al. [1]); moreover, if the problem has a favorable geometry (e.g., when the learner's action set is a simplex or a spectrahedron), these bounds are "almost" dimension-free, a fact that is of crucial importance in practical applications.

In view of these appealing guarantees, one might expect this picture to carry over effortlessly to multiagent decision problems as well. However, game-theoretic learning can be considerably more involved because, in addition to the *exogenous* variability of the stage game $\mathcal{G}_t$ as a function of t, the players' individual reward functions also vary *endogenously* as a function of the actions chosen by the other players at any given time t. Moreover, the standard solution concept in game theory is that of a *Nash equilibrium*—not the players' regret (external, internal, or dynamic). As a result, even though the algorithms under study are essentially the same in both single agent and multiagent environments, the analysis and the results obtained in these two settings are often markedly different.

Our paper focuses on multiagent problems and aims to analyze the equilibrium tracking and convergence properties of MD-based policies in time-varying games. In so doing, we seek to partially fill a gap in the existing literature on game-theoretic learning, which focuse almost exclusively on the case in which there are no exogenous variations in the players' payoff functions—that is, when the stage game $\mathcal{G}_t$ remains *fixed* for all t. To provide the necessary context, we begin by discussing some relevant works, and we outline our main contributions right after.

1.2. Related Work

The well-known impossibility result of Hart and Mas-Colell [28] shows that there are no uncoupled dynamics leading to a Nash equilibrium in all games. Thus, given that no-regret dynamics are unilateral by construction—and, hence, uncoupled a fortiori—it is not possible to establish a blanket causal link between no-regret play and convergence to a Nash equilibrium; in fact, even in the relatively simple context of bilinear zero-sum games, no-regret learning may cycle indefinitely without converging, always remaining a uniform distance away from the game's Nash equilibria (Mertikopoulos et al. [41, 43]).

For this reason, deriving the equilibrium convergence properties of multiagent learning processes requires a more specialized look, typically zooming in on specific classes of games. In the case of mixed extensions of finite games, Cominetti et al. [17], Coucheney et al. [18], Cohen et al. [16], and Leslie and Collins [37] show that certain variants of the exponential weights algorithm converge to perturbed Nash equilibria with probability one in potential and $2\times 2\times\cdots\times 2$ games. More recently, in the case of *continuous* potential games, Perkins et al. [50] show that a lifted variant of MD-based methods converges weakly to an ε-neighborhood of the game's set of Nash equilibria. Importantly, in all these works, convergence is established by first showing that a naturally associated continuous-time dynamical system converges and then using the so-called ordinary differential equation (ODE) method of stochastic approximation (Benaïm [6], Benaïm et al. [7]) to translate this result to discrete time.

More relevant for our purposes is the recent work of Mertikopoulos and Zhou [40], who focus on the class of *monotone games*, that is, continuous games that satisfy the so-called *diagonal strict concavity* (DSC) condition of Rosen [53]. Specifically, using the same ODE stochastic approximation tools discussed, Mertikopoulos and Zhou [40] show that the sequence of play generated by a specific version of the dual averaging algorithm of Nesterov [47] converges to a Nash equilibrium with probability one (w.p.1) even in the presence of noise and uncertainty. The analysis of Mertikopoulos and Zhou [40] is subsequently extended by Bravo et al. [13] to learning with payoff-based, "bandit feedback"—that is, when players observe only the payoff of the action that they played. At around the same time, Tatarenko and Kamgarpour [63, 64] use a Tikhonov regularization approach to obtain a series of comparable results for "merely monotone" games (i.e., monotone games that are not necessarily *strictly* monotone), whereas more recently, Drusvyatskiy and Ratliff [21] improve the rate of convergence in strongly monotone games to $\mathcal{O}(1/T^{1/2})$. Finally, in a very recent paper, Bervoets et al. [9] use stochastic approximation methodologies to prove the convergence of a payoff-based, *dampened gradient approximation* scheme in two other classes of one-dimensional concave games: games with strategic complements and ordinal potential games with isolated equilibria.

1.3. Our Contributions

In all the works described, the game faced by the players remains fixed throughout the learning process, and the variation in the players' individual payoff functions is strictly endogenous—that is, it is only a result of the other players' evolving action choice. By contrast, our paper seeks to tackle problems in which the sequence of games encountered by the players also evolves exogenously—that is, players encounter a *time-varying game*.

In this general context, we examine the equilibrium tracking and convergence properties of a wide class of MD-based policies—encoded as Algorithm 1 in Section 3—in two distinct regimes:

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

a. When the sequence of stage games converges to some well-defined limit (in our case, a strictly monotone game).

b. When $\mathcal{G}_t$ evolves over time without converging.

In terms of feedback, we consider a flexible *stochastic first order oracle* (SFO) model that provides noisy payoff gradient estimates to the players based on the actions they choose at each stage of the process, and we establish the following results (stated in an informal, simplified version).

Theorem 1 (Informal Version). *Suppose that $\mathcal{G}_t$ converges to a strictly monotone game $\mathcal{G}$. If each player runs Algorithm 1 with a suitable step size, the, with probability one, the sequence of realized actions X_t converges to the (unique) Nash equilibrium x^* of $\mathcal{G}$.*

Theorem 2 (Informal Version). *Suppose that $\mathcal{G}_t$ is strongly monotone and varies smoothly over time, that is, $\sum_{t=1}^{T} \|x^*_{t+1} - x^*_t\| = \mathcal{O}(T^r)$ for some $r < 1$ (where x^*_t is the Nash equilibrium of $\mathcal{G}_t$). If each player runs Algorithm 1 with a suitable step size, the sequence of realized actions X_t enjoys the equilibrium tracking error $\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T} \|X_t - x^*_t\|^2] = \mathcal{O}(T^{-\frac{1-r}{3}})$.*

In words, Theorem 1 shows that, if the sequence of stage games $\mathcal{G}_t$ stabilizes to some well-defined limit $\mathcal{G}$, the induced sequence of play converges to a Nash equilibrium of $\mathcal{G}$ with probability one. On the other hand, if $\mathcal{G}_t$ does not stabilize, there is no equilibrium state to which to converge (either static or in the mean); in this case, Theorem 2 shows that the average distance from equilibrium vanishes over time, and it provides an explicit estimate of the resulting "equilibrium tracking error" in terms of the equilibrium variation of the sequence of stage games encountered by the players.

Finally, to account for environments in which gradient information is not available, we also consider the case of learning with *payoff-based* feedback. By considering a one-shot gradient estimation process based on single-point stochastic approximation techniques (Bravo et al. [13], Flaxman et al. [26], Héliou et al. [29], Spall [61]), we map the problem of payoff-based learning to our generic oracle model, and we show that our convergence and equilibrium tracking results still apply in this case (though the corresponding rates are worsened because of the players' having less information at their disposal).

In terms of proof techniques, the exogenous dependence of $\mathcal{G}_t$ on t means that the continuous-time limit of the players' learning process is likewise nonautonomous (i.e., it also depends on t). As a result, there is no longer a well-defined "mean field equation" to approximate, so it is not possible to employ the ODE method of Benaïm [6] that underlies the series of papers discussed earlier. Instead, to establish convergence to an equilibrium in the "stable limit" regime, we work directly in discrete time, and we employ a mix of submartingale limit theory and quasi-Fejér arguments. Finally, our equilibrium tracking result relies on decomposing the horizon of play into batches of appropriately chosen lengths and, subsequently, utilizes a batch comparison technique that is introduced by Besbes et al. [10] to analyze the dynamic regret of *single-agent* online learning algorithms.

2. Preliminaries

2.1. Notation

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

2.2. Continuous Games

Throughout our paper, we focus on games with a finite number of players and continuous action sets. Specifically, every player $i \in \mathcal{N} = \{1, \dots, N\}$ is assumed to select an *action* x_i from a compact convex subset $\mathcal{K}_i$ of a finite-dimensional normed space $\mathcal{X}_i$; subsequently, every player receives a *reward* based on each player's individual objective and the *action profile* $x = (x_i; x_{-i}) \equiv (x_1, \dots, x_i, \dots, x_N)$ of all players' actions. In more detail, writing $\mathcal{K} := \prod_{i\in\mathcal{N}} \mathcal{K}_i$ for the game's *action space*, we assume that each player's reward is determined by an associated *payoff* (or *utility*) *function* $u_i : \mathcal{K} \to \mathbb{R}$. The tuple $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{K}, u)$ is then referred to as a *continuous game*.

In terms of regularity, we assume throughout that the players' payoff functions are continuously differentiable, and we write $v_i(x)$ for the individual payoff gradient of the ith player, that is,

$$v_i(x) = \nabla_{x_i} u_i(x_i; x_{-i}) \tag{2.1}$$

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

or, putting all players together,

$$v(x) = (v_1(x), \dots, v_N(x)). \tag{2.2}$$

In this, we are tacitly assuming that u_i is defined on an open neighborhood of $\mathcal{K}$ in the ambient space $\mathcal{X} := \prod_{i\in\mathcal{N}} \mathcal{X}_i$ of the game; none of our results depend on this device, so we do not make this assumption explicit. Wealso adopt the established convention of treating $v_i(x)$ as an element of the dual space $\mathcal{Y}_i := \mathcal{X}_i^*$ of $\mathcal{X}_i$. Finally, we assume that $\mathcal{X}$ is endowed with the norm $\|x\|^2 = \sum_i \|x_i\|^2$, where, for ease of notation, we write $\|\cdot\|$ for the norm of each factor space $\mathcal{X}_i$ and rely on the context to resolve any ambiguities.

2.3. Nash Equilibria and Monotonicity

The most prevalent solution concept in game theory is that of an NE. This is an action profile $x^* \in \mathcal{K}$ that is resilient to unilateral deviations, that is,

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

The set of Nash equilibria of $\mathcal{G}$ is denoted in the sequel as $x^* := NE(\mathcal{G})$.

By virtue of this definition, it is straightforward to check that Nash equilibria satisfy the Stampacchia variational inequality

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

As a result, finding a Nash equilibrium of a continuous game typically involves solving the Stampacchia problem (SVI). This observation forms the basis of an important link between game theory and optimization (cf. Facchinei and Pang [24], Laraki et al. [36], and references therein).

Now, starting with the seminal work of Rosen [53], much of the literature focuses on games that satisfy the *diagonal concavity* (DC) condition

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

Owing to the link between (DC) and the theory of monotone operators in optimization, games that satisfy (DC) are commonly referred to as monotone games.[1] In particular, mirroring the corresponding terminology from convex analysis, we say that $\mathcal{G}$ is

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

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

Obviously, we have the inclusions "stronglymonotone" $\subsetneq$ "strictlymonotone" $\subsetneq$ "monotone", mirroring the corresponding chain of inclusions "stronglyconcave" $\subsetneq$ "strictlyconcave" $\subsetneq$ "concave" for concave functions.

Examples of monotone games include Kelly auctions and Tullock markets (Kelly et al. [34], Tullock [66]), signal covariance and power control problems in signal processing (D'Oro et al. [20], Mertikopoulos and Moustakas [39]), Cournot oligopolies (Monderer and Shapley [44]), and many other problems in which online decision making is the norm. For a diverse list of applications in different contexts, see Facchinei and Kanzow [23] and Scutari et al. [55].

3. The Learning Model

To account for the possibility of exogenous variations in the game-theoretic setup of the previous section, we assume that the players face a different stage game $\mathcal{G}_t$ at each decision opportunity. More explicitly, the envisioned sequence of play unfolds as follows:

1. At each stage $t = 1, 2, \dots$, every agent $i \in \mathcal{N}$ selects an action $X_{i,t} \in \mathcal{K}_i$.
2. Each player receives the associated reward based on $\mathcal{G}_t$ and observes—or otherwise constructs—an estimate $\hat{v}_{i,t} \in \mathcal{Y}_i$ of the individual payoff gradients.
3. Subsequently, players update their actions and the process repeats.

The core ingredients of this framework are

a. The sequence of stage games $\mathcal{G}_t$ encountered by the players.
b. The sequence of gradient signals $\hat{v}_{i,t} \in \mathcal{Y}_i$ observed (or inferred) at each stage.
c. The way that players update their actions as a function of the observed information.

We discuss each of these elements in detail.

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

3.1. The Stage Game Sequence

The only blanket assumption that we make for the sequence of stage games $\mathcal{G}_t$ is that the players' payoff functions are Lipschitz continuous and smooth. More precisely, we posit the following requirement for the players' tth stage payoff field $v_t(x) = (v_{i,t}(x))_{i\in\mathcal{N}}$.

Assumption 1. *The game's payoff functions are C^2-smooth; in particular, there exist constants $G_i, L_i > 0$ such that*

$$\|v_{i,t}(x)\|_* \leq G_i \tag{3.1a}$$

$$\|v_{i,t}(x') - v_{i,t}(x)\|_* \leq L_i\|x' - x\| \tag{3.1b}$$

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

For posterity, we also write $G := \max_i G_i$ and $L_i := \max_i L_i$. Beyond this mild regularity assumption, the sequence of stage games is assumed arbitrary. For instance, the evolution of $\mathcal{G}_t$ could be random (i.e., $\mathcal{G}_t$ could be determined by some randomly drawn parameter θ_t at each stage), it could be governed by an underlying (hidden) Markov chain model, etc. In particular, we do not assume that the stage game $\mathcal{G}_t$ is revealed to the players before choosing an action: from their individual viewpoint, the players are involved in a repeated decision process in which the choice of an action returns a reward, but they have no knowledge of the game generating this reward. This "agnostic" approach is motivated by the fact that the standard rationality postulates of game theory (full rationality, common knowledge of rationality, etc.) are not satisfied in many cases of practical interest. We briefly discuss two concrete examples of this framework.

Example 1 (Repeated Kelly Auctions). Consider a Kelly auction in which a splittable resource (advertising time on a website, a catch of fish in a fish market, etc.) is auctioned off, day after day, to a set of N buyers (Kelly et al. [34], Tullock [66]). In more detail, each player can place a monetary bid $x_i \in [0, b_i]$ to acquire a unit of said resource, up to the player's total budget b_i. Then, once all bids are in, the resource is allocated proportionally to each player's bid; that is, the ith player gets a fraction $\rho_i = x_i/[c + \sum_{j\in N} x_j]$ of the auctioned resource (with $c > 0$ denoting an "entry barrier" for participating in the auction). Thus, if $g_{i,t}$ denotes the marginal gain that the ith player acquires per resource unit, the player's prorated utility at the tth epoch is

$$u_{i,t}(x_i; x_{-i}) = \frac{g_{i,t}x_i}{c + \sum_{j\in\mathcal{N}} x_j} - x_i. \tag{3.2}$$

Clearly, the players' utility functions evolve as a function of the intrinsic value $g_{i,t}$ associated to a unit of the auctioned resource. Because this value may be subject to arbitrary exogenous fluctuations (for instance, depending on the traffic coming to the website at any given time in the advertising example), we obtain a time-varying game as before.

Example 2 (Power Control). As another example, consider N wireless users transmitting a stream of packets to a common receiver over a shared wireless channel (Mertikopoulos et al. [42], Scutari et al. [55], Tse and Viswanath [65]). If the channel gain for the ith user at the tth frame is $g_{i,t}$ and the user transmits with power $p_i \in [0, P_{\max}]$, the user's information transmission rate is given by the celebrated Shannon formula

$$R_{i,t}(p_i; p_{-i}) = \log\left(1 + \frac{g_{i,t}p_i}{\sigma + \sum_{j\neq i} g_{j,t}p_j}\right), \tag{3.3}$$

where $\sigma > 0$ denotes the ambient noise in the channel (Tse and Viswanath [65]). Because the users' channel gains evolve over time (e.g., because of fading, user mobility, or other fluctuations in the wireless medium), we obtain a time-varying game in which each user seeks to maximize the individual communication rate.

3.2. The Feedback Signal

The second basic ingredient of our model is the feedback available to the players after choosing an action. In tune with the limited information setting outlined earlier, we only posit that, at each stage $t = 1,2,\dots$, every player $i \in \mathcal{N}$ receives—or otherwise constructs—a "gradient signal" $\hat{v}_{i,t} \in \mathcal{Y}_i$. Analytically, this signal is treated as if generated from an SFO, that is, an abstract mechanism that provides an estimate of each player's individual payoff gradient at the chosen action profile. Specifically, if called at $X_t = (X_{1,t}, \dots, X_{N,t}) \in \mathcal{K}$, we assume that $\hat{v}_{i,t}$ is of the form

$$\hat{v}_{i,t} = v_{i,t}(X_t) + Z_{i,t}, \tag{SFO}$$

where the "observational error" $Z_{i,t}$ captures all sources of uncertainty in the received input.

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

To differentiate further between "random" (zero-mean) and "systematic" (nonzero-mean) errors in $\hat{v}_{i,t}$, it is convenient to decompose the error process $Z_{i,t}$ as

$$Z_{i,t} = U_{i,t} + b_{i,t}, \tag{3.4}$$

where $U_{i,t}$ is zero-mean and $b_{i,t}$ denotes the mean of $Z_{i,t}$. Formally, writing $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ for the natural filtration of X_t, we set

$$b_{i,t} = \mathbb{E}[Z_{i,t} \mid \mathcal{F}_t] \quad \text{and} \quad U_{i,t} = Z_{i,t} - b_{i,t}, \tag{3.5}$$

so by definition, $\mathbb{E}[U_{i,t} \mid \mathcal{F}_t] = 0$. In this way, the oracle feedback received by each player $i \in \mathcal{N}$ can be classified according to the following statistics:

1. Bias:

$$\|b_{i,t}\|_* \leq B_{i,t}. \tag{3.6a}$$

2. Variance:

$$\mathbb{E}[\|U_{i,t}\|_*^2 \mid \mathcal{F}_t] \leq \sigma_{i,t}^2. \tag{3.6b}$$

3. Second moment:

$$\mathbb{E}[\|\hat{v}_{i,t}\|_*^2 \mid \mathcal{F}_t] \leq M_{i,t}^2. \tag{3.6c}$$

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

$$S_{i,t}^2 = M_{i,t}^2 + \sigma_{i,t}^2. \tag{3.6d}$$

In this, $B_{i,t}$, $\sigma_{i,t}$, and $M_{i,t}$ are to be construed as deterministic upper bounds on the bias, variance, and magnitude of the oracle signal $\hat{v}_{i,t}$ that player $i \in \mathcal{N}$ receives at time t. We also assume throughout that $B_{i,t}$ is nonincreasing, whereas $\sigma_{i,t}$ and $M_{i,t}$ are nondecreasing. Finally, in obvious notation, we write $\hat{v}_t$, b_t, U_t, and so forth for the corresponding profiles $\hat{v}_t = (\hat{v}_{i,t})_{i \in \mathcal{N}}$ and the like.

Remark 1. To streamline our presentation, we first present our results in a model-agnostic manner, that is, without specifying the origins of the oracle model (SFO); subsequently, in Section 5, we provide an explicit construction of such an oracle from payoff-based observations, and we discuss in detail what this entails for our analysis and results.

3.3. Learning via Mirror Descent

The last element of the players' learning process concerns the way that players update their actions based on the received feedback. For concreteness, we focus throughout on the widely used family of algorithms known as MD, which posits that players update their actions by taking a "proximal" gradient step from their current action.[2] Formally, this can be modeled via the basic recursion

$$X_{i,t+1} = \mathcal{P}_i(X_{i,t}; \gamma_{i,t}\hat{v}_{i,t}), \tag{MD}$$

where

1. $t = 1, 2, \dots$ denotes the stage of the process.
2. $X_{i,t}$ denotes the action chosen by player i at stage t.
3. $\hat{v}_{i,t}$ is the oracle signal of player i at stage t.
4. $\gamma_{i,t} > 0$ is a player-specific step-size sequence (assumed nonincreasing).
5. $\mathcal{P}_i$ denotes the "prox-mapping" of player $i \in N$ (see a detailed definition as follows).

For a pseudocode implementation from the viewpoint of a generic player, see Algorithm 1.

Algorithm 1 (Learning via Mirror Descent (Player Indices Suppressed))

Require: `prox-mapping` $\mathcal{P}$, `step-size` $\gamma_t > 0$

`initialize` $X_1 \leftarrow \arg\min h$ `# initialization`

for $t = 1, 2, \dots$ **do**

  `play` $X_t \in \mathcal{K}$ `# play action`

  `get gradient signal` $\hat{v}_t$ `# get feedback`

  `set` $X_{t+1} \leftarrow \mathcal{P}(X_t; \gamma_t \hat{v}_t)$ `# update action`

end for

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

Methods based on mirror descent have received intense scrutiny ever since the pioneering work of Nemirovski and Yudin [45]; for an appetizer, see Beck and Teboulle [4], Bravo and Mertikopoulos [12], Nemirovski et al. [46], Nesterov [47], Shalev-Shwartz [57], and references therein. For intuition, the archetypal example of the method is based on the Euclidean prox-mapping

$$\mathcal{P}(x;y) = \Pi_{\mathcal{C}}(x+y) = \arg\min_{x'\in\mathcal{C}}\left\{\|x+y-x'\|_2^2\right\} = \arg\min_{x'\in\mathcal{C}}\left\{\langle y, x-x'\rangle + \frac{1}{2}\|x'-x\|_2^2\right\}, \tag{3.7}$$

where $\Pi_{\mathcal{C}}$ denotes the closest point projection onto a given convex set $\mathcal{C}$. Going beyond this familiar example, the key novelty of mirror descent is to replace the quadratic term in (3.7) by the so-called *Bregman divergence*

$$D(x',x) = h(x') - h(x) - \langle \nabla h(x), x'-x\rangle, \tag{3.8}$$

induced by a distance-generating function (DGF) h on $\mathcal{C}$. This function plays the role of the squared Euclidean norm in (3.11), and following Juditsky et al. [32], we define it as follows.

Definition 1. Let $\mathcal{C}$ be a compact convex subset of $\mathcal{X} \cong \mathbb{R}^d$. A convex function $h : \mathcal{X} \to \mathbb{R}\cup\{\infty\}$ is said to be a DGF on $\mathcal{C}$ if

1. h is continuous and supported on $\mathcal{C}$, that is, $\operatorname{dom} h := \{x\in\mathcal{X} : h(x) < \infty\} = \mathcal{C}$.
2. h is K-strongly convex relative to $\|\cdot\|$ on $\mathcal{C}$, that is,

$$h(\lambda x + (1-\lambda)x') \le \lambda h(x) + (1-\lambda)h(x') - \frac{1}{2}K\lambda(1-\lambda)\|x'-x\|^2 \tag{3.9}$$

for all $x, x' \in \mathcal{C}$ and all $\lambda\in[0,1]$.

3. The subdifferential ∂h of h admits a *continuous selection*, that is, there exists a continuous mapping $\nabla h : \operatorname{dom}\partial h \to \mathcal{Y}$ such that $\nabla h(x) \in \partial h(x)$ for all $x \in \operatorname{dom}\partial h$.[3]

For concision, given a DGF h on $\mathcal{C}$, we refer to $\mathcal{C}_h := \operatorname{dom}\partial h$ as the *prox-domain* of h. The Bregman divergence $D : \mathcal{C}_h \times \mathcal{C} \to \mathbb{R}$ induced by h is then given by (3.8), and the associated prox-mapping $\mathcal{P} : \mathcal{C}_h \times \mathcal{Y} \to \mathcal{C}$ is defined as

$$\mathcal{P}(x;y) = \arg\min_{x'\in\mathcal{C}}\{\langle y, x-x'\rangle + D(x',x)\} \quad \text{for all } x\in\mathcal{C}_h, y\in\mathcal{Y}. \tag{3.10}$$

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

Throughout the sequel, we assume that each player $i\in\mathcal{N}$ is endowed with an individual distance-generating function $h_i : \mathcal{K}_i \to \mathbb{R}$. In obvious notation, we also write K_i for the strong convexity modulus of h_i, $\mathcal{K}_{h_i}$ for its prox-domain, $D_i : \mathcal{K}_i \times \mathcal{K}_{h_i} \to \mathbb{R}$ for the associated Bregman divergence, and $\mathcal{P}_i : \mathcal{K}_{h_i} \times \mathcal{Y}_i \to \mathcal{K}_i$ for the induced prox-mapping. For concreteness, we provide two standard examples.

Example 3 (Euclidean Projections). We begin by revisiting Euclidean projections on a compact convex subset $\mathcal{C}$ of $\mathbb{R}^d$. The corresponding DGF is $h(x) = \frac{1}{2}\|x\|^2$ for $x\in\mathcal{X}$, so $\mathcal{C}_h = \mathcal{C}$ and $\nabla h(x) = x$ for all $x\in\mathcal{C}$. Hence, the associated Bregman divergence is

$$D(x',x) = \frac{1}{2}\|x'\|_2^2 - \frac{1}{2}\|x\|_2^2 - \langle x, x'-x\rangle = \frac{1}{2}\|x'-x\|_2^2, \tag{3.11}$$

and the resulting recursion $x^+ = \Pi(x+\gamma v)$ is just a standard projected forward step.

Example 4 (Entropic Regularization). Let $\mathcal{C} = \Delta_d := \{x\in\mathbb{R}^d_+ : \sum_{j=1}^d x_j = 1\}$ denote the unit simplex of $\mathcal{X} = \mathbb{R}^d$. A very widely used distance-generating function for this geometry is the (negative) *Gibbs-Shannon entropy* $h(x) = \sum_{j=1}^d x_j\log x_j$ (with the standard notational convention $0\cdot\log 0 = 0$). By inspection, the prox-domain of h is $\mathcal{C}_h := ri\,\mathcal{C}$, and the resulting Bregman divergence is just the *Kullback–Leibler* (KL) divergence

$$D(x',x) = D_{\mathrm{KL}}(x',x) := \sum_{j=1}^d x_j'\log\left(\frac{x_j'}{x_j}\right) \quad \text{for all } x\in\mathcal{C}_h, x'\in\mathcal{C}. \tag{3.12}$$

In turn, a standard calculation leads to the prox-mapping

$$\mathcal{P}(x;y) = \frac{(x_1e^{y_1},\dots,x_ne^{y_n})}{x_1e^{y_1}+\cdots+x_ne^{y_n}} \tag{3.13}$$

for all $x\in\mathcal{C}_h$, $y\in\mathcal{Y}$. The corresponding update rule $x^+ = \mathcal{P}(x;\gamma v)$ is widely known in optimization as *entropic gradient descent* (Beck and Teboulle [4], Kivinen and Warmuth [35]), and as hedge (or exponential/multiplicative

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

weights update) in game theory and online learning (Arora et al. [2], Auer et al. [3], Littlestone and Warmuth [38], Vovk [67]).

4. Equilibrium Tracking and Convergence Analysis

We are now in a position to state our main results for the equilibrium tracking and convergence properties of (MD) in time-varying games. For concreteness, we focus on two distinct—and, to a large extent, complementary—regimes:

a. When the sequence of stage games $\mathcal{G}_t$ converges to some limit game $\mathcal{G} \equiv \mathcal{G}_\infty$.

b. When $\mathcal{G}_t$ evolves over time without converging.

In both cases, we treat the process defining the time-varying game as a "black box," and we do not scrutinize its origins in detail; we do so in order to focus on the interplay between the variability of the sequence $\mathcal{G}_t$ and the induced sequence of play.

4.1. Stabilization and Convergence to Equilibrium

We begin with the case in which the sequence of stage games stabilizes to some monotone limit game $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{K}, u)$. Formally, it is convenient to characterize this stabilization in terms of the quantity

$$R_{i,t} = \max_{x \in \mathcal{K}} \|v_{i,t}(x) - v_i(x)\|_*, \tag{4.1}$$

and we say that the sequence of games $\mathcal{G}_t$, $t = 1, 2, \dots$ converges to $\mathcal{G}$ if

$$\lim_{t \to \infty} R_{i,t} = 0 \quad \text{for all } i \in \mathcal{N}. \tag{4.2}$$

To state our equilibrium convergence result, we require two further assumptions. The first is a technical "reciprocity condition" for the players' DGF, namely,

$$D(p, x_t) \to 0 \quad \text{whenever} \quad x_t \to p \tag{RC}$$

for every sequence of actions $x_t \in \mathcal{K}_h$. This requirement is fairly standard in the trajectory analysis of mirror descent algorithms (Beck and Teboulle [4], Chen and Teboulle [15]) and, taken together with the strong convexity of h, it implies that $x_t \to p$ if and only if $D(p, x_t) \to 0$ (hence, the name).[4] In particular, if h is Lipschitz, we have

$$D(p, x_t) \leq h(p) - h(x_t) + \|\nabla h(x_t)\|_* \|x_t - p\| = \mathcal{O}(\|x_t - p\|), \tag{4.3}$$

so (RC) always holds in that case. A further easy check shows that Example 4 also satisfies this condition, so (RC) is not restrictive in this regard.

The second set of conditions concerns the players' step-size sequence. First, we assume throughout that

$$\sum_{t=1}^{\infty} \gamma_{i,t} = \infty \quad \text{for all } i \in \mathcal{N}, \tag{S1}$$

that is, each player's learning process cannot stop prematurely. Second, we assume that the step-size policies of any two players $i, j \in \mathcal{N}$ are mutually compatible in the sense that

$$\sum_{t=1}^{\infty} |\gamma_{i,t} - \lambda_{ij} \gamma_{j,t}| < \infty \quad \text{for some } \lambda_{ij} > 0. \tag{S2}$$

Informally, the compatibility assumption (S2) means that the players' step-size policies exhibit a comparable asymptotic behavior as $t \to \infty$, that is, $\gamma_{i,t}/\gamma_{j,t} = \Theta(1)$ for all $i, j \in \mathcal{N}$. The rationale for this is fairly straightforward: if a player employs a step-size policy that vanishes much faster than that of all other players, this player effectively becomes a "constant externality" in the timescale of the other players. On that account, it makes more sense to consider convergence in a "reduced" game in which this player is effectively removed from the game—and so on, until only the "slower" timescale players remain. Assumption (S2) rules out such cases and ensures that all players remain active throughout the horizon of play.

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

Theorem 1. *Let $\mathcal{G}_t$ be a time-varying game converging to a strictly monotone game $\mathcal{G}$. Suppose further that each player $i \in \mathcal{N}$ runs Algorithm 1 with a DGF satisfying (RC) and a step-size policy satisfying (S1), (S2), and*

$$\sum_{t=1}^{\infty} \gamma_{i,t}(R_{i,t} + B_{i,t}) < \infty \quad \textit{and} \quad \sum_{t=1}^{\infty} \gamma_{i,t}^2 S_{i,t}^2 < \infty. \tag{S3}$$

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

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

In particular, if the feedback and stabilization metrics $B_{i,t}$, $S_{i,t}$, and $R_{i,t}$ behave asymptotically as $B_{i,t} = \mathcal{O}(1/t^{b_i})$, $S_{i,t} = \mathcal{O}(t^{s_i})$ and $R_{i,t} = \mathcal{O}(1/t^{r_i})$ for some $b_i, s_i, r_i \geq 0$, we have the following immediate corollaries.

Corollary 1. *With assumptions as before, if each player follows Algorithm 1 with $\gamma_{i,t} \propto 1/t^p$ for some $p > \max\{1 - r_i, 1 - b_i, 1/2 + s_i\}$, $p \leq 1$, the induced sequence of play X_t converges to a Nash equilibrium with probability one.*

Corollary 2. *If Algorithm 1 is run with perfect oracle feedback and assumptions as before, taking $p > \max_i(1 - r_i)$ guarantees that X_t converges to a Nash equilibrium with probability one.*

To streamline our discussion, we postpone the proof of Theorem 1 until later in this section, and we proceed with some remarks.

4.1.1. Learning in Static Games and Stochastic Approximation. The special case $\mathcal{G}_t \equiv \mathcal{G}$ for all $t = 1, 2, \dots$ can be seen as learning in a repeated, *static* game. As we discuss in the introduction, this case is extensively studied in the literature, usually via the so-called ODE method of stochastic approximation (Benaïm [6], Benaïm et al. [7], Benveniste et al. [8]). In this literature, convergence of a learning process is typically established by showing that an underlying mean field dynamical system converges and then using a series of asymptotic pseudotrajectory approximation results to infer that the same applies to the discrete-time algorithm under study as well.

In this direction, the closest result to our own is the recent paper of Mertikopoulos and Zhou [40] in which the authors show that a specific, multiagent version of Nesterov's [47] dual averaging algorithm converges to a Nash equilibrium in static, strictly monotone games. However, there are several key obstacles that arise when trying to adapt the proof techniques of Mertikopoulos and Zhou [40] to our setting. First and foremost, the prox-mappings $\mathcal{P}_i$ are, in general, *discontinuous* across different faces of $\mathcal{K}_i$, so (MD) cannot be seen as the discretization of an ODE (consider, for example, the Euclidean case in which $\mathcal{P}_i$ is the closest point projection to $\mathcal{K}_i$). An approach based on the theory of differential inclusions (DIs) (Benaïm et al. [7]) could help overcome this obstacle, but even then, the exogenous dependence of $\mathcal{G}_t$ on t means that the DI approximation of the players' learning process is likewise nonautonomous. Thus, given that there is no longer a well-defined continuous-time system to approximate, it is not possible to employ a dynamical systems approach as in Mertikopoulos and Zhou [40].

Finally, we also note that the use of player-specific step-size sequences complicates the discretization landscape even further. In the stochastic approximation literature, player-specific step sizes are usually treated within a multiple-timescale framework, for example, as in Borkar [11], Leslie and Collins [37], and Perkins and Leslie [49]. However, in this case, the underlying ODE must also separate the faster from the slower timescales, which means that the players with the smaller step sizes end up being effectively removed from the game. This is an important part of the reason that the literature on learning in static games traditionally focuses on learning algorithms with the same step size across players and also an important reason that the stochastic approximation approach of Mertikopoulos and Zhou [40] does not apply in our setting.

4.1.2. Step-Size Requirements and Tuning. In the literature on learning in games, a common choice for the step size of iterative methods is the policy $\gamma_{i,t} \propto 1/t$ (cf. Beggs [5], Bervoets et al. [9], Cominetti et al. [17], Coucheney et al. [18], Erev and Roth [22], Hofbauer and Sandholm [30], and references therein). In view of Corollary 1, if the players' oracle feedback is unbiased and bounded in mean square (i.e., $b_i = \infty$, $s_i = 0$ for all $i \in \mathcal{N}$), this step-size policy guarantees convergence to a Nash equilibrium as long as the game stabilizes at a power law rate—that is, provided that $R_t := \max_i R_{i,t} = \mathcal{O}(1/t^r)$ for some $r > 0$.[5] In fact, if (MD) is run with $\gamma_{i,t} \propto 1/(t \log t)$, convergence is guaranteed even if the game stabilizes at a slower, sublogarithmic rate $R_t = \mathcal{O}(1/(\log t)^\varepsilon)$ for some $\varepsilon > 0$.

The policies $\gamma_{i,t} \propto 1/t$ and $\gamma_{i,t} \propto 1/(t \log t)$ should be seen as conservative "fail-safes": it stands to reason that, if more information about the asymptotic behavior of $R_{i,t}$ is available, a more aggressive step-size policy (as per Corollary 1) might be more efficient. Specifically, if we focus as before on the case in which the players' oracle feedback is unbiased and bounded in mean square ($b = \infty$, $s = 0$), the second moment term $\sum_t \gamma_{i,t}^2 S_{i,t}^2$ is subleading in (S3) relative to the stabilization error term $\sum_t \gamma_{i,t} R_{i,t}$ whenever $p \geq r_i$ for *some* $i \in \mathcal{N}$. Because the summability condition (S3) further requires $p > 1 - r_i$ for *all* $i \in \mathcal{N}$, this would suggest taking $p = \min_i r_i$ if $\min_i r_i > 1/2$ and p larger than $1/2$ by an arbitrarily small amount otherwise.

By contrast, if no prior information on $R_{i,t}$ is available, it is not clear how to choose the exponent p in an optimal manner relative to the variability of $\mathcal{G}_t$. In particular, because r_i depends on the entire (infinite) tail of $R_{i,t}$, adaptive policies that rely on the (finite) history of play up to time t—for example, in the spirit of Rakhlin and

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

Sridharan [51] and Syrgkanis et al. [62]—do not seem well-suited for this purpose. We are not aware of any way to circumvent this difficulty in terms of almost sure convergence of the sequence of play.

4.2. Tracking Nash Equilibria

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

To that end, let $\mathcal{G}_t$ be a sequence of strongly monotone games, and consider the equilibrium tracking error

$$err(T) := \sum_{t=1}^{T} \|X_t - x_t^*\|^2 = \sum_{t=1}^{T} \sum_{i\in\mathcal{N}} \|X_{i,t} - x_{i,t}^*\|^2, \tag{4.4}$$

where x_t^* is the (unique) Nash equilibrium of $\mathcal{G}_t$.[6] By construction, if $err(T)$ is small relative to T, the sequence of chosen actions X_t is close to equilibrium for most of the window of interest. However, if the variability of $\mathcal{G}_t$ is too high, it is not reasonable to expect a tracking error that grows sublinearly in T, even in the single-player case.[7] To quantify this, we define the game's *equilibrium variation* as

$$V(T) := \sum_{t=1}^{T} \|x_{t+1}^* - x_t^*\|, \tag{4.5}$$

and we say that $\mathcal{G}_t$ varies *smoothly* if

$$V(T) = o(T) \quad \text{as } T \to \infty. \tag{4.6}$$

In what follows, we seek to establish conditions under which Algorithm 1 guarantees $err(T) = o(T)$ when (4.6) holds. Our main result in this direction is as follows.

Theorem 2. *Let $\mathcal{G}_t$ be a sequence of strongly monotone games satisfying Assumption 1. Suppose further that each player $i \in \mathcal{N}$ runs Algorithm 1 with step size $\gamma_{i,t} \propto t^{-p_i}$, $p_i \in (0,1)$, a Lipschitz distance-generating function, and feedback of the form (SFO) with $B_{i,t} = \mathcal{O}(1/t^{b_i})$ and $S_{i,t}^2 = \mathcal{O}(t^{2s_i})$ for some $b_i, s_i \geq 0$, $i \in \mathcal{N}$. Then, the players' tracking error is bounded as*

$$\mathbb{E}[err(T)] = \mathcal{O}(T^{1-\min_i(p_i-2s_i)} + T^{1-\min_i b_i} + T^{\max_i p_i + \min_i(p_i-2s_i)} V(T)). \tag{4.7}$$

Corollary 3. *Suppose that the players' oracle feedback is unbiased and bounded in mean square ($b_i = \infty$, $s_i = 0$ for all $i \in \mathcal{N}$). If the equilibrium variation of the game is $V(T) = \mathcal{O}(T^r)$ for some $r > 0$, Algorithm 1 enjoys the bound*

$$\mathbb{E}[err(T)] = \mathcal{O}(T^{1-p_{\min}} + T^{2p_{\max}+r}). \tag{4.8}$$

Here, $p_{\min} = \min_i p_i$ and $p_{\max} = \max_i p_i$. In particular, if each player runs Algorithm 1 with $\gamma_{i,t} \propto 1/t^{(1-r)/3}$, then

$$\mathbb{E}[err(T)] = \mathcal{O}(T^{\frac{2+r}{3}}). \tag{4.9}$$

Theorem 2 is our basic equilibrium tracking result, so we proceed with some remarks.

4.2.1. Step-Size Requirements and Tuning. If the players' gradient oracle is unbiased and bounded in mean square ($b_i = \infty$ and $s_i = 0$ for all $i \in \mathcal{N}$), Corollary 3 shows that equilibrium tracking is possible as long as

$$p_i < \frac{1-r}{2} \quad \text{for all } i \in \mathcal{N}. \tag{4.10}$$

Comparing this condition with the step-size requirements for equilibrium convergence (cf. Theorem 1 and Corollary 1), we may infer that equilibrium tracking is more lightweight in terms of prerequisites: specifically, because Theorem 2 does not require the step-size compatibility condition (S2), each player can pick p_i independently of one another. The reason for this difference has to do with the fact that equilibrium tracking focuses on the players' *average* behavior over the horizon of play; by contrast, the convergence of the sequence of play depends on the entire tail of $\gamma_{i,t}$, so the asymptotic behavior of the players' step-size policies cannot be too different.

4.2.2. Equilibrium Tracking and Dynamic Regret Minimization: Similarities. In our setup, the dynamic regret incurred by the ith player up to time T under the sequence of play $X_t \in \mathcal{K}$, $t = 1, 2, \dots$ can be defined as

$$DynReg_i(T) = \sum_{t=1}^{T} [u_{i,t}(\hat{x}_{i,t}; X_{-i,t}) - u_{i,t}(X_t)] = \sum_{t=1}^{T} [\tilde{u}_{i,t}(\hat{x}_{i,t}) - \tilde{u}_{i,t}(X_{i,t})], \tag{4.11}$$

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

where $\tilde{u}_{i,t} := u_i(\cdot; X_{-i,t})$ denotes the *effective* payoff function encountered by player $i \in \mathcal{N}$ at stage t given the chosen action profile $X_{-i,t}$ of all other players, and

$$\hat{x}_{i,t} \in \arg\max_{x_i \in \mathcal{K}_i} u_{i,t}(x_i; X_{-i,t}) = \arg\max_{x_i \in \mathcal{K}_i} \tilde{u}_{i,t}(x_i) \tag{4.12}$$

denotes the ith player's "counterfactual" *best response* to $X_{-i,t}$ in the game $\mathcal{G}_t$ (with $\mathcal{G}_t$, $t = 1, 2, \dots$, assumed fixed as a sequence but otherwise arbitrary and unknown to the players). Obviously, if there are no other players in the game, $\hat{x}_t$ coincides with the Nash equilibrium of the tth stage game against nature, so a natural question that arises is whether the equilibrium tracking guarantees of Theorem 2 can be related to a dynamic regret bound.

In this regard, a slight modification of the proof of Theorem 2 yields the following: if an agent with a convex compact action set $\mathcal{K}$ runs Algorithm 1 with step size $\gamma_t \propto 1/t^p$ against a stream of concave—though not necessarily strongly concave—payoff functions $u_t : \mathcal{K} \to \mathbb{R}$ with drift $V(T)$, then

$$\mathbb{E}[DynReg(T)] = \mathcal{O}(T^{1+2s-p} + T^{1-b} + T^{2p-2s}V(T)). \tag{4.13}$$

In particular, if $V(T) = \mathcal{O}(T^r)$ and the player's oracle feedback is unbiased and bounded in mean square ($b = \infty$, $s = 0$), the choice $p = (1 - r)/3$ guarantees

$$\mathbb{E}[DynReg(T)] = \mathcal{O}(T^{\frac{2+r}{3}}). \tag{4.14}$$

For a precise statement and proof, we refer the reader to Section 6.2.[8]

4.2.3. Equilibrium Tracking and Dynamic Regret Minimization: Differences. Going back to the multiagent case, the sequence $\hat{x}_t = (\hat{x}_{i,t})_{i \in \mathcal{N}}$ with $\hat{x}_{i,t}$ given by (4.12) may be very different from the Nash equilibrium sequence x_t^*: the former best responds to the actual sequence of play X_t, whereas the latter best responds to itself (so it depends only on $\mathcal{G}_t$ and is otherwise *independent* of X_t). As we saw earlier, this distinction is redundant in the single-player case, but it is crucial in the multiagent one: the sequence $\hat{x}_t$ may vary rapidly even if x_t^* is constant. For example, even if the sequence of base payoff functions $u_{i,t}$ does not depend on t exogenously (i.e., $u_{i,t} \equiv u_i$ for all t), the effective payoff functions $\tilde{u}_{i,t} := u_i(\cdot; X_{-i,t})$ encountered individually by each agent still depend on t endogenously via $X_{-i,t}$. As a result, the single-agent bound (4.13) does not a priori apply: because $\|X_{-i,t+1} - X_{-i,t}\| = \mathcal{O}(\gamma_t) = \mathcal{O}(t^{-p})$, the variation of $\tilde{u}_{i,t}$ over a window of length T under Algorithm 1 could be as high as $\Theta(T^{1-p})$ even though the underlying game is *constant* (so the equilibrium variation of the game is *zero*). This shows that, unless the play of other agents perfectly follows their individual component of a Nash equilibrium, there may be a significant conceptual gap between the single- and multiagent settings.

This subtlety is also reflected on the strong monotonicity assumption in Theorem 2, which invites the question whether the bound (4.7) is tight. To wit, when faced with a sequence of strongly concave payoff functions, Besbes et al. [10] show that an adversary can always impose $DynReg(T) = \Omega(V(T)^{1/2}T^{1/2})$. This bound is strictly better than the $\mathcal{O}(T^{1-p} + V(T)T^{2p})$ guarantee of Corollary 3, suggesting that there may be room for improvement. Nevertheless, there are two important roadblocks to achieve this:

1. First, in the single-agent case, the key to attaining faster regret minimization is the basic inequality

$$u_t(x_t^*) - u_t(x) \leq \langle v_t(x), x_t^* - x \rangle - \frac{\mu}{2}\|x - x_t^*\|^2, \tag{4.15}$$

where x_t^* denotes the (necessarily unique) maximizer of u_t. As a result, the growth of $Gap(T)$—which is driven by gradient terms of the form $\langle v_t(X_t), x_t^* - X_t \rangle$—is mitigated by the quadratic correction terms: by balancing these two terms, it is possible to obtain sharper bounds for $DynReg(T)$ when each u_t is strongly concave.

On the other hand, in a multiagent, game-theoretic setting, (4.15) becomes

$$u_{i,t}(x_{i,t}^*; x_{-i}) - u_{i,t}(x) \leq \langle v_{i,t}(x), x_{i,t}^* - x_i \rangle - \frac{\mu}{2}\|x_i - x_{i,t}^*\|^2, \tag{4.16}$$

where x_t^* now denotes the (necessarily unique) Nash equilibrium of the strongly monotone stage game $\mathcal{G}_t \equiv \mathcal{G}_t(\mathcal{N}, \mathcal{K}, u_t)$. Arguing as in the single-agent setting indeed yields a sharper bound on the quantity

$$\sum_{t=1}^{T} \sum_{i \in \mathcal{N}} [u_{i,t}(x_{i,t}^*; X_{-i,t}) - u_{i,t}(X_t)], \tag{4.17}$$

but in general, the minimization of this quantity does not provide a certificate that X_t is in any way close to equilibrium. In particular, in contrast to the single-agent case, (4.16) could be either positive or negative, so it cannot act as a merit function for tracking an evolving equilibrium.

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

2. Second, the optimal static regret minimization rate in strongly convex problems is attained when $\gamma_t \propto 1/t$. However, Besbes et al. [10] provide a counterexample in which this step-size policy produces *linear* dynamic regret. In view of this, achieving an $\mathcal{O}(V(T)^{1/2}T^{1/2})$ dynamic regret minimization rate seems to require a different approach and/or assumptions—for example, an adaptive policy in the spirit of Jadbabaie et al. [31] in the case of perfect gradient feedback.

We mention these to emphasize that bounding the equilibrium tracking error $err(T)$ is significantly different than bounding the dynamic regret of an individual agent in the unilateral setting (even though the obtained guarantees look similar). It is an open question whether it is possible to close the gap between the $\mathcal{O}(T^{\frac{2+r}{3}})$ equilibrium tracking error of Theorem 2 for multiagent online learning in strongly monotone games and the corresponding $\mathcal{O}(T^{\frac{1+r}{2}})$ dynamic regret bound of Besbes et al. [10] for single-agent online strongly convex problems. This gap suggests that the lower bound for equilibrium tracking in general games may require a distinction to be made between games that admit a potential function (which is always the case in the single-agent setting) and those that do not. In particular, it is reasonable to conjecture that the bound (4.7) may be improved in strongly concave potential games (perhaps through the use of a finely tuned restart mechanism); however, the general case seems considerably more difficult, so we defer it to future work.

4.2.4. Legendre DGFs. We should also note that the reciprocity condition (RC) is replaced in the statement of Theorem 2 by the stronger requirement $\sup_{x_i}\|\nabla h_i(x_i)\|_* < \infty$, which rules out Legendre-like DGFs (such as the entropic setup of Example 4). This condition is needed in Proposition 3, which requires a finite Bregman diameter $\mathcal{D}_i := \sup_{x_i,x_i'} D_i(x_i, x_i')$ to bound the "regret-like" quantity $\sum_{t=1}^{T}\langle v_{i,t}(X_t), x_i - X_{i,t}\rangle$. Orabona and Pál [48] recently show that (MD) may incur linear regret when run with a variable step size in problems with infinite Bregman diameter, so this requirement is not an artifact of the analysis.

That being said, there are several ways to overcome this hurdle: First, the players could run (MD) with a constant step size over windows of a specified length and use a restart mechanism to achieve a sublinear equilibrium tracking error; this approach is proposed by Besbes et al. [10] for the minimization of dynamic regret, and we discuss it in more detail in Section 6.2. Another way is to add an "anchoring term" in the definition of the prox-mapping $\mathcal{P}_i$ and play the so-called *dual-stabilized* mirror descent policy

$$X_{i,t+1} = \arg\min_{x_i\in\mathcal{K}_i}\{\gamma_{i,t}\langle\hat{v}_{i,t}, X_{i,t} - x_i\rangle + D_i(x_i, X_{i,t}) + (\gamma_{i,t+1}^{-1} - \gamma_{i,t}^{-1})D_i(x_i, X_{i,1})\}. \tag{DS-MD}$$

This policy is introduced by Fang et al. [25], who show that (DS-MD) achieves sublinear regret even in domains with an infinite Bregman diameter. Finally, another—and arguably simpler—approach is to switch to the dual averaging policy of Nesterov [47], which instead prescribes

$$X_{i,t+1} = \arg\max_{x_i\in\mathcal{K}_i}\left\{\sum_{s=1}^{t}\langle\hat{v}_{i,s}, x_i\rangle - \gamma_{i,t}h_i(x_i)\right\}. \tag{DA}$$

This algorithm has the advantage of attaining order-optimal regret guarantees with the Bregman diameter $\mathcal{D}_i$ replaced by the range $\mathcal{R}_i := \max h_i - \min h_i$ of h_i (which is always finite because $\mathcal{K}_i$ is compact and the domain of h_i contains $\mathcal{K}_i$). Either of these algorithmic tweaks ultimately yields a sublinear tracking error in domains with an infinite Bregman diameter, but the details lie beyond the scope of our work, so we do not discuss them here.

4.3. Proof of Theorem 1

The rest of this section is devoted to proving the results stated earlier, starting with the proof of Theorem 1. The first key step in this direction is the definition of a suitable "energy-like" function that is—on average and up to small, second order errors—decreasing along the trajectory of play X_t. In the analysis of mirror descent algorithms, this role is usually played by the Bregman divergence relative to the target point under study (in our case, the Nash equilibrium of $\mathcal{G}$). However, because each player $i \in \mathcal{N}$ now learns at a different pace (as determined by their individual step-size policy $\gamma_{i,t}$), the definition of a suitable energy function for Algorithm 1 is not as straightforward.

To that end (and with a fair amount of hindsight), we begin by introducing the player-specific weights

$$\lambda_i = \left(\prod_{j\in\mathcal{N}}\lambda_{ij}\right)^{1/N} \quad \text{for all } i \in \mathcal{N}, \tag{4.18}$$

with $\lambda_{ij} > 0$, $i, j \in \mathcal{N}$ given by the mutual compatibility condition (S2). As we show, these weights enjoy a decomposition property that is key for the sequel.

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

Lemma 1. *Suppose that $\gamma_{i,t}$ satisfies (S1) and (S2). Then, $\lambda_{ij} = \lambda_i/\lambda_j$ for all $i, j \in \mathcal{N}$.*

Proof. Our proof relies on the following two intermediate claims.

Claim 1. *The weights λ_{ij} are uniquely defined. Indeed, suppose that (S2) holds also with $\lambda'_{ij} \neq \lambda_{ij}$ for some $i, j \in \mathcal{N}$. Then, for all $t = 1$, we have*

$$|\lambda_{ij} - \lambda'_{ij}|\gamma_{j,t} = |\lambda_{ij}\gamma_{j,t} - \lambda'_{ij}\gamma_{j,t}| \leq |\gamma_{i,t} - \lambda_{ij}\gamma_{j,t}| + |\gamma_{i,t} - \lambda'_{ij}\gamma_{j,t}|, \tag{4.19}$$

so $|\lambda_{ij} - \lambda'_{ij}|\gamma_{j,t}$ is summable given that both $|\gamma_{i,t} - \lambda_{ij}\gamma_{j,t}|$ and $|\gamma_{i,t} - \lambda'_{ij}\gamma_{j,t}|$ are summable (by assumption). This contradicts (4.4), so our claim follows.

Claim 2. *The weights λ_{ij} satisfy the chain rule $\lambda_{ik} = \lambda_{ij}\lambda_{jk}$ for all $i, j, k \in \mathcal{N}$. Indeed,*

$$\begin{aligned}\sum_{t=1}^{\infty}|\gamma_{i,t} - \lambda_{ij}\lambda_{jk}\gamma_{k,t}| &= \sum_{t=1}^{\infty}|\gamma_{i,t} - \lambda_{ij}\gamma_{j,t} + \lambda_{ij}\gamma_{j,t} - \lambda_{ij}\lambda_{jk}\gamma_{k,t}| \\ &\leq \sum_{t=1}^{\infty}|\gamma_{i,t} - \lambda_{ij}\gamma_{j,t}| + \lambda_{ij}\sum_{t=1}^{\infty}|\gamma_{j,t} - \lambda_{jk}\gamma_{k,t}| \\ &< \infty\end{aligned} \tag{4.20}$$

with the last inequality following from (S2). Our claim then follows from the definition of λ_{ik} and our preceding uniqueness claim.

Thus, with these two claims in hand, we readily obtain

$$\frac{\lambda_i}{\lambda_j} = \frac{(\prod_{k\in\mathcal{N}}\lambda_{ik})^{1/N}}{(\prod_{k\in\mathcal{N}}\lambda_{jk})^{1/N}} = \prod_{k\in\mathcal{N}}(\lambda_{ik}/\lambda_{jk})^{1/N} = \prod_{k\in\mathcal{N}}\lambda_{ij}^{1/N} = \lambda_{ij}, \tag{4.21}$$

where, in the third step, we use the preceding chain rule to write $\lambda_{ij} = \lambda_{ik}/\lambda_{jk}$. This establishes our assertion and completes our proof.

This lemma shows that (S2) can be rewritten as $\sum_{t=1}^{\infty}|\gamma_{i,t}/\lambda_i - \gamma_{j,t}/\lambda_j| < \infty$, which, in turn, implies that λ_i can be interpreted as the relative "learning speed" of player $i \in \mathcal{N}$. In view of this, we consider the effective step size

$$\gamma_t = \frac{1}{N}\sum_{i\in\mathcal{N}}\frac{\gamma_{i,t}}{\lambda_i} \tag{4.22}$$

and the energy function

$$E(x) = \sum_{i\in\mathcal{N}}\frac{D_i(x_i^*, x_i)}{\lambda_i}, \tag{4.23}$$

where $x_i^* \in \mathcal{K}_i$ denotes the ith component of the Nash equilibrium x^* of $\mathcal{G}$. We then have the following quasi-descent inequality for E under (MD).

Lemma 2. *Suppose that each player $i \in \mathcal{N}$ runs Algorithm 1 with a step-size policy $\gamma_{i,t}$ satisfying (S1) and (S2). Then, the iterates $E_t := E(X_t)$ of E under X_t enjoy the bound*

$$\begin{aligned}E_{t+1} \leq E_t + \gamma_t\langle v(X_t), X_t - x^*\rangle + \sum_{i\in\mathcal{N}}\frac{\gamma_{i,t}}{\lambda_i}\langle r_{i,t} + Z_{i,t}, X_{i,t} - x_i^*\rangle + \sum_{i\in\mathcal{N}}\frac{\gamma_{i,t}^2}{2\lambda_i K_i}\|\hat{v}_{i,t}\|_*^2 \\ + \frac{\max_i G_i \operatorname{diam}(\mathcal{K}_i)}{N}\sum_{i,j\in\mathcal{N}}\left|\frac{\gamma_{i,t}}{\lambda_i} - \frac{\gamma_{j,t}}{\lambda_j}\right|\end{aligned} \tag{4.24}$$

with $r_{i,t} = v_{i,t}(X_t) - v_i(X_t)$.

Proof. By Lemma A.4 in Appendix A, the Bregman divergence $D_{i,t} := D_i(x_i^*, X_{i,t})$ satisfies the inequality

$$D_{i,t+1} \leq D_{i,t} + \gamma_{i,t}\langle\hat{v}_{i,t}, X_{i,t} - x_i^*\rangle + \frac{\gamma_{i,t}^2}{2K_i}\|\hat{v}_{i,t}\|_*^2. \tag{4.25}$$

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

Therefore, with $\hat{v}_{i,t} = v_{i,t}(X_t) + Z_{i,t} = v_i(X_t) + r_{i,t} + Z_{i,t}$ and $E_t = \sum_{i\in\mathcal{N}} \lambda_i^{-1} D_{i,t}$, we get

$$E_{t+1} \leq E_t + \sum_{i\in\mathcal{N}} \frac{\gamma_{i,t}}{\lambda_i} \langle Z_{i,t} + r_{i,t}, X_{i,t} - x_i^* \rangle + \sum_{i\in\mathcal{N}} \frac{\gamma_{i,t}^2}{2\lambda_i K_i} \|\hat{v}_{i,t}\|_*^2. \tag{4.26a}$$

$$+ \sum_{i\in\mathcal{N}} \frac{\gamma_{i,t}}{\lambda_i} \langle v_i(X_t), X_{i,t} - x_i^* \rangle, \tag{4.26b}$$

so it suffices to upper bound the term (4.26b) of the preceding inequality. To that end, we have

$$\begin{aligned}
(4.26\text{b}) &= \gamma_t \langle v(X_t), X_t - x^* \rangle + \sum_{i\in\mathcal{N}} \left(\frac{\gamma_{i,t}}{\lambda_i} - \gamma_t \right) \langle v_i(X_t), X_{i,t} - x_i^* \rangle \\
&\leq \gamma_t \langle v(X_t), X_t - x^* \rangle + \sum_{i\in\mathcal{N}} \left| \frac{\gamma_{i,t}}{\lambda_i} \right| - \gamma_t \cdot G_i \operatorname{diam}(\mathcal{K}_i) \\
&= \gamma_t \langle v(X_t), X_t - x^* \rangle + \sum_{i\in\mathcal{N}} \frac{G_i \operatorname{diam}(\mathcal{K}_i)}{N} \left| \sum_{j\in\mathcal{N}} \left(\frac{\gamma_{i,t}}{\lambda_i} - \frac{\gamma_{j,t}}{\lambda_j} \right) \right| \\
&\leq \gamma_t \langle v(X_t), X_t - x^* \rangle + \frac{\max_i G_i \operatorname{diam}(\mathcal{K}_i)}{N} \sum_{i,j\in\mathcal{N}} \left| \frac{\gamma_{i,t}}{\lambda_i} - \frac{\gamma_{j,t}}{\lambda_j} \right|.
\end{aligned} \tag{4.27}$$

Our claim then follows by substituting this bound back in (4.26).

The importance of the energy-like bound (4.24) lies in that the "drift term" $\gamma_t \langle v(X_t), X_t - x^* \rangle$ provides a leading negative contribution to E_t (because x^* is a Nash equilibrium of $\mathcal{G}$), whereas all other terms become vanishingly small over time. The next proposition formalizes this idea and shows that E_t converges to some (random) finite value.

Proposition 1. *Suppose that each player $i \in \mathcal{N}$ runs Algorithm 1 with a step-size $\gamma_{i,t}$ satisfying (S1)–(S3). Then, E_t converges (almost surely (a.s.)) to a random variable E_∞ with $\mathbb{E}[E_\infty] < \infty$.*

Proof. We begin by decomposing each player's oracle signal as

$$\hat{v}_{i,t} = v_{i,t}(X_t) + b_{i,t} + U_{i,t} = v_i(X_t) + r_{i,t} + b_{i,t} + U_{i,t}, \tag{4.28}$$

and we set, respectively,

$$\rho_{i,t} = \langle r_{i,t}, X_{i,t} - x_i^* \rangle \qquad \rho_t = \sum_{i\in\mathcal{N}} \frac{\gamma_{i,t}}{\lambda_i \gamma_t} \rho_{i,t}, \tag{4.29a}$$

$$\beta_{i,t} = \langle b_{i,t}, X_{i,t} - x_i^* \rangle \qquad \beta_t = \sum_{i\in\mathcal{N}} \frac{\gamma_{i,t}}{\lambda_i \gamma_t} \beta_{i,t}, \tag{4.29b}$$

and

$$\psi_{i,t} = \langle U_{i,t}, X_{i,t} - x_i^* \rangle \qquad \psi_t = \sum_{i\in\mathcal{N}} \frac{\gamma_{i,t}}{\lambda_i \gamma_t} \psi_{i,t} \tag{4.29c}$$

with γ_t given by (4.22) and $r_{i,t} = v_{i,t}(X_{i,t}) - v_i(X_t)$ defined as in Lemma 2. The energy inequality (4.24) then gives

$$E_{t+1} \leq E_t + \gamma_t \langle v(X_t), X_t - x^* \rangle + \gamma_t (\rho_t + \beta_t + \psi_t) + \chi_t + \sum_{i\in\mathcal{N}} \frac{\gamma_{i,t}^2}{2\lambda_i K_i} \|\hat{v}_{i,t}\|_*^2, \tag{4.30}$$

where we set

$$\chi_t = \frac{\max_i G_i \operatorname{diam}(\mathcal{K}_i)}{N} \sum_{i,j\in\mathcal{N}} \left| \frac{\gamma_{i,t}}{\lambda_i} - \frac{\gamma_{j,t}}{\lambda_j} \right|. \tag{4.31}$$

Therefore, conditioning on the history $\mathcal{F}_t$ of X_t up to stage t (inclusive) and taking expectations, we get

$$\begin{aligned}
\mathbb{E}[E_{t+1} | \mathcal{F}_t] &\leq \mathbb{E}\left[E_t + \gamma_t \langle v(X_t), X_t - x^* \rangle + \gamma_t (\rho_t + \beta_t + \psi_t) + \chi_t + \sum_{i\in\mathcal{N}} \frac{\gamma_{i,t}^2 \|\hat{v}_{i,t}\|_*^2}{2\lambda_i K_i} \,\middle|\, \mathcal{F}_t \right] \\
&\leq E_t + \gamma_t (\rho_t + \beta_t) + \chi_t + \sum_{i\in\mathcal{N}} \frac{\gamma_{i,t}^2}{2\lambda_i K_i} M_{i,t}^2,
\end{aligned} \tag{4.32}$$

where we use the definition (3.6c) of $M_{i,t}$ and the facts that

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

a. x^* is a Nash equilibrium of $\mathcal{G}$ (so $\langle v(X_t), X_t - x^* \rangle \leq 0$).
b. ρ_t and β_t are both $\mathcal{F}_t$-measurable (by definition).
c. $\mathbb{E}[\psi_t|\mathcal{F}_t] = \langle \mathbb{E}[U_t|\mathcal{F}_t], X_t - x^* \rangle = 0$.
To proceed, note that

$$\rho_{i,t} = \langle r_{i,t}, X_{i,t} - x_i^* \rangle \leq \|r_{i,t}\|_* \|X_{i,t} - x_i^*\| \leq \operatorname{diam}(\mathcal{K}_i) R_{i,t}, \tag{4.33}$$

and similarly, $\beta_{i,t} \leq \operatorname{diam}(\mathcal{K}_i) B_{i,t}$. The bound (4.32) may then be written as $\mathbb{E}[E_{t+1}|\mathcal{F}_t] \leq E_t + \varepsilon_t$, where

$$\varepsilon_t = \sum_{i \in \mathcal{N}} \left[\frac{\gamma_{i,t}}{\lambda_i} \operatorname{diam}(\mathcal{K}_i) \cdot (R_{i,t} + B_{i,t}) + \frac{\gamma_{i,t}^2}{2\lambda_i K_i} M_{i,t}^2 \right]. \tag{4.34}$$

Consider now the auxiliary process $\zeta_t = E_{t+1} + \sum_{s=t+1}^{\infty} \varepsilon_s$. Taking expectations yields

$$\mathbb{E}[\zeta_t|\mathcal{F}_t] \leq E_t + \varepsilon_t + \sum_{s=t+1}^{\infty} \varepsilon_s = E_t + \sum_{s=t}^{\infty} \varepsilon_s = \zeta_{t-1}, \tag{4.35}$$

that is, ζ_t is a supermartingale relative to $\mathcal{F}_t$. Moreover, because $\sum_{t=1}^{\infty} \varepsilon_t < \infty$ by (S3) and Lemma 1, we also get $\mathbb{E}[\zeta_t] \leq \mathbb{E}[\zeta_1] < \infty$, that is, ζ_t is bounded in L^1. Therefore, by Doob's (sub)martingale convergence theorem (Hall and Heyde [27, theorem 2.5]), it follows that ζ_t converges almost surely to some random variable ζ that is itself finite (almost surely and in L^1). Because $E_t = \zeta_{t-1} - \sum_{s=t}^{\infty} \varepsilon_s$ and $\lim_{t\to\infty} \sum_{s=t}^{\infty} \varepsilon_s = 0$, we conclude that E_t converges (a.s.) to ζ, and our proof is complete.

Moving forward, our next result shows that we can extract a subsequence of X_t that converges to a Nash equilibrium of the limit game $\mathcal{G}$.

Proposition 2. *With assumptions as in Proposition 1, we have* $\liminf_t \|X_t - x^*\| = 0$ *(a.s.).*

Proof. We begin by showing that, for all $\varepsilon > 0$, the hitting time

$$\tau_\varepsilon = \inf\{t \in \mathbb{N} : \|X_t - x^*\| \leq \varepsilon\} \tag{4.36}$$

is finite with probability one; formally, we show that the event $\mathcal{N}_\varepsilon = \{\tau_\varepsilon = \infty\}$ has $\mathbb{P}(\mathcal{N}_\varepsilon) = 0$ for all $\varepsilon > 0$.

To do so, fix some $\varepsilon > 0$ and let $c_\varepsilon = -\inf\{\langle v(x), x - x^* \rangle : \|x - x^*\| \geq \varepsilon\}$, so $c_\varepsilon > 0$ by the strict monotonicity of $\mathcal{G}$ and the fact that v is continuous and $\mathcal{K}$ is compact. Then, with notation as in the proof of Proposition 1, telescoping the bound (4.30) yields

$$E_{t+1} \leq E_1 - c_\varepsilon \sum_{s=1}^{t} \gamma_s + \underbrace{\sum_{s=1}^{t} \gamma_s(\rho_s + \beta_s) + \sum_{s=1}^{t} \chi_s}_{\mathrm{I}_t} + \underbrace{\sum_{s=1}^{t} \gamma_s \psi_s}_{\mathrm{II}_t} + \underbrace{\sum_{s=1}^{t} \sum_{i \in \mathcal{N}} \frac{\gamma_{i,s}^2}{2\lambda_i K_i} \|\hat{v}_{i,s}\|_*^2}_{\mathrm{III}_t} \tag{4.37}$$

for all $t \leq \tau_\varepsilon$. We now proceed to bound each of the underscored terms:

1. First, for the term I_t, we show in the proof of Proposition 1 that

$$\gamma_t(\rho_t + \beta_t) \leq \sum_{i \in \mathcal{N}} \frac{\gamma_{i,t}}{\lambda_i} \operatorname{diam}(\mathcal{K}_i) \cdot (R_{i,t} + B_{i,t}), \tag{4.38}$$

so $\sum_{t=1}^{\infty} \gamma_t(\rho_t + \beta_t) < \infty$ by (S3). Condition (S2) further gives $\sum_{t=1}^{\infty} \chi_t < \infty$, so I_t is uniformly bounded from above by (Tex translation failed).

2. For the noise term $\mathrm{II}_t = \sum_{s=1}^{t} \gamma_s \psi_s$, we have $\mathbb{E}[\psi_t|\mathcal{F}_t] = 0$, so II_t is a martingale. Furthermore, by (3.6b) and the step-size assumption (S3) of Theorem 1, we have

$$\begin{aligned} \sum_{t=1}^{\infty} \gamma_{i,t}^2 \mathbb{E}[\psi_{i,t}^2|\mathcal{F}_t] &\leq \sum_{t=1}^{\infty} \gamma_{i,t}^2 \|X_{i,t} - x_i^*\|^2 \mathbb{E}[\|U_{i,t}\|_*^2|\mathcal{F}_t] \\ &\leq \operatorname{diam}(\mathcal{K}_i)^2 \sum_{t=1}^{\infty} \gamma_{i,t}^2 \sigma_{i,t}^2 < \infty. \end{aligned} \tag{4.39}$$

In turn, this implies that $\sum_{t=1}^{\infty} \gamma_t^2 \mathbb{E}[\psi_t^2|\mathcal{F}_t] < \infty$, so by the law of large numbers for martingale difference sequences (Hall and Heyde [27, theorem 2.18]), we conclude that $\sum_{s=1}^{t} \gamma_s \psi_s / \sum_{s=1}^{t} \gamma_s \to 0$ (a.s.).

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

3. Finally, for the last term, let $\Psi_{i,t} = \sum_{s=1}^{t} \gamma_{i,s}^2 \|\hat{v}_{i,s}\|_*^2$, so $\mathrm{III}_t = \sum_{i\in\mathcal{N}} (2\lambda_i K_i)^{-1} \Psi_{i,t}$. We then have

$$\begin{aligned}\mathbb{E}[\Psi_{i,t}|\mathcal{F}_t] &= \mathbb{E}\left[\sum_{s=1}^{t-1} \gamma_{i,s}^2 \|\hat{v}_{i,s}\|_*^2 + \gamma_{i,t}^2 \|\hat{v}_{i,t}\|_*^2 \,\middle|\, \mathcal{F}_t\right] \\ &= \Psi_{i,t-1} + \gamma_{i,t}^2 \mathbb{E}[\|\hat{v}_{i,t}\|_*^2 | \mathcal{F}_t] \geq \Psi_{i,t-1},\end{aligned} \tag{4.40}$$

that is, $\Psi_{i,t}$ is a submartingale relative to $\mathcal{F}_t$ (recall that $\hat{v}_t$ is generated after X_t, so it is not $\mathcal{F}_t$-measurable). Furthermore, by the law of total expectation, we also have

$$\mathbb{E}[\Psi_{i,t}] = \mathbb{E}[\mathbb{E}[\Psi_{i,t}|\mathcal{F}_t]] \leq \sum_{t=1}^{\infty} \gamma_{i,t}^2 M_{i,t}^2 < \infty. \tag{4.41}$$

This shows that $\Psi_{i,t}$ is uniformly bounded in L^1 so, by Doob's (sub)martingale convergence theorem (Hall and Heyde [27, theorem 2.5]), it follows that $\Psi_{i,t}$ converges to some (almost surely finite) random variable $\Psi_{i,\infty}$ with $\mathbb{E}[\Psi_{i,\infty}] < \infty$. We, thus, conclude that III_t is likewise bounded from above by $\mathrm{III}_\infty = \sum_{i\in\mathcal{N}} (2\lambda_i K_i)^{-1} \Psi_{i,\infty} < \infty$ (a.s.).

Suppose now that $\mathbb{P}(\mathcal{N}_\varepsilon) = \mathbb{P}(\tau_\varepsilon = \infty) > 0$. Then, there exists a realization of X_t such that

$$E_{t+1} \leq E_1 - \left[c_\varepsilon - \frac{\mathrm{I}_t + \mathrm{II}_t + \mathrm{III}_t}{\sum_{s=1}^{t} \gamma_s}\right] \cdot \sum_{s=1}^{t} \gamma_s \quad \text{for all } t = 1, 2, \ldots, \tag{4.42}$$

and in addition, $(\mathrm{I}_t + \mathrm{II}_t + \mathrm{III}_t)/\sum_{s=1}^{t} \gamma_s \to 0$ (because we show that this last event occurs w.p.1). However, by (S1), this gives $\lim_{t\to\infty} E_t = -\infty$, a contradiction that shows $\tau_\varepsilon < \infty$ w.p.1 for all $\varepsilon > 0$. Hence, given that each $\mathcal{N}_{1/k}$ is a zero-probability event and there is a countable number thereof, we conclude that

$$\begin{aligned}\mathbb{P}(\liminf_t \|X_t - x^*\| = 0) &= \mathbb{P}(\tau_{1/k} < \infty \text{ for all } k = 1, \ldots, \infty) \\ &= \mathbb{P}\left(\bigcap\nolimits_{k=1}^{\infty} \{\tau_{1/k} < \infty\}\right) = 1 - \mathbb{P}\left(\bigcup\nolimits_{k=1}^{\infty} \mathcal{N}_{1/k}\right) = 1,\end{aligned} \tag{4.43}$$

and our proof is complete.

With these two intermediate results at hand, we are finally in a position to prove Theorem 1.

Proof of Theorem 1. By Proposition 2, X_t admits a (possibly random) subsequence X_{t_k} such that $X_{t_k} \to x^*$ (a.s.). By the reciprocity condition (RC), this further implies that $\liminf_{t\to\infty} E_t = 0$ (a.s.). However, because $\lim_{t\to\infty} E_t$ exists (by Proposition 1), we conclude that

$$\mathbb{P}\left(\lim_{t\to\infty} X_t = x^*\right) = \mathbb{P}\left(\lim_{t\to\infty} E_t = 0\right) = \mathbb{P}\left(\liminf_{t\to\infty} E_t = 0\right) = 1, \tag{4.44}$$

and our proof is complete.

4.4. Proof of Theorem 2

We now proceed to prove the equilibrium tracking guarantees of Algorithm 1. To that end, given a sequence of action profiles $X_t \in \mathcal{K}$, $t = 1, 2, \ldots$ and a window of interest $\mathcal{T} = [\tau_{\text{start}} \,..\, \tau_{\text{end}}]$ with $1 \leq \tau_{\text{start}} \leq \tau_{\text{end}} \leq T$, it is useful to consider the gap functions

$$\mathit{Gap}_{x_i}(\mathcal{T}) = \sum_{t\in\mathcal{T}} \langle v_{i,t}(X_t), x_i - X_{i,t}\rangle \qquad \mathit{Gap}_x(\mathcal{T}) = \sum_{i\in\mathcal{N}} \mathit{Gap}_{x_i}(\mathcal{T}), \tag{4.45a}$$

and

$$\mathit{Gap}_i(\mathcal{T}) = \max_{x_i\in\mathcal{K}_i} \mathit{Gap}_{x_i}(\mathcal{T}) \qquad \mathit{Gap}(\mathcal{T}) = \sum_{i\in\mathcal{N}} \mathit{Gap}_i(\mathcal{T}). \tag{4.45b}$$

By convention, we also write $\mathit{Gap}_{x_i}(T)$, $\mathit{Gap}_x(T)$, etc., when the window of interest is of the form $\mathcal{T} = [1 \,..\, T]$.

Now, by the strong monotonicity of $\mathcal{G}_t$, we have $\mu\|X_t - x_t^*\|^2 \leq \langle v_t(X_t), x_t^* - X_t\rangle$, so $\mathit{Gap}(T)$ may act as a surrogate for bounding the equilibrium tracking error $\mathit{err}(T)$ of Algorithm 1. In view of this, we begin with a technical bound for the gap under (MD).

Proposition 3. *Suppose that player $i \in \mathcal{N}$ runs Algorithm 1 with step size $\gamma_{i,t}$ and oracle feedback of the form (SFO). Then, for any window of the form $\mathcal{T} = [\tau_{\text{start}} \,..\, \tau_{\text{end}}]$, we have*

$$\mathit{Gap}_{x_i}(\mathcal{T}) \leq \sum_{t\in\mathcal{T}} \left(\frac{1}{\gamma_{i,t}} - \frac{1}{\gamma_{i,t-1}}\right) D_i(x_i, X_{i,t}) + \sum_{t\in\mathcal{T}} \langle Z_{i,t}, X_{i,t} - x_i\rangle + \frac{1}{2K_i} \sum_{t\in\mathcal{T}} \gamma_{i,t} \|\hat{v}_{i,t}\|_*^2 \tag{4.46}$$

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

with the convention $\gamma_{i,\tau_{\text{start}}-1} = \infty$ *in the sum. In addition, if* $\gamma_{i,t}$ *is nonincreasing, then*

$$\mathbb{E}[Gap_i(\mathcal{T})] \leq \frac{2H_i(x_i)}{\gamma_{i,\tau_{\text{end}}}} + 2\operatorname{diam}(\mathcal{K}_i)\sum_{t\in\mathcal{T}} B_{i,t} + \frac{1}{2K_i}\sum_{t\in\mathcal{T}} \gamma_{i,t} S_{i,t}^2, \tag{4.47}$$

where $H_i(x_i) = \sup_{x_i'\in\mathcal{K}_{h_i}} D(x_i, x_i')$.

Proof. We first focus on the pointwise bound (4.46). To that end, because $X_{i,t+1} = \mathcal{P}(X_{i,t}; \gamma_{i,t}\hat{v}_{i,t})$ for all $t = 1, 2, \dots,$ invoking Lemma A.4 with $Y_t \leftarrow \gamma_{i,t}\hat{v}_{i,t}$ and $\alpha_t \leftarrow 1/\gamma_{i,t}$ yields

$$\sum_{t\in\mathcal{T}} \langle \hat{v}_{i,t}, x_i - X_{i,t}\rangle \leq \sum_{t\in\mathcal{T}} \left(\frac{1}{\gamma_{i,t}} - \frac{1}{\gamma_{i,t-1}}\right) D_i(x_i, X_{i,t}) + \frac{1}{2K_i}\sum_{t\in\mathcal{T}} \gamma_{i,t}\|\hat{v}_{i,t}\|_*^2. \tag{4.48}$$

By the feedback model (SFO), we have $\hat{v}_{i,t} = v_{i,t}(X_t) + Z_{i,t}$, so

$$Gap_{x_i}(\mathcal{T}) = \sum_{t\in\mathcal{T}} \langle v_{i,t}(X_t), x_i - X_{i,t}\rangle = \sum_{t\in\mathcal{T}} \langle \hat{v}_{i,t}, x_i - X_{i,t}\rangle + \sum_{t\in\mathcal{T}} \langle Z_{i,t}, X_{i,t} - x_i\rangle. \tag{4.49}$$

Our claim then follows by adding (4.48) and (4.49).

For the bound (4.47), maximizing over $x_i \in \mathcal{K}_i$ in (4.46) and taking expectations, we get

$$\mathbb{E}[Gap_i(\mathcal{T})] = \mathbb{E}\left[\max_{x_i\in\mathcal{K}_i} Gap_{x_i}(\mathcal{T})\right] \leq \mathbb{E}\left[\sum_{t\in\mathcal{T}} \left(\frac{1}{\gamma_{i,t}} - \frac{1}{\gamma_{i,t-1}}\right) D_i(x_i, X_{i,t})\right] \tag{4.50a}$$

$$+ \frac{1}{2K_i}\sum_{t\in\mathcal{T}} \gamma_{i,t}\mathbb{E}[\|\hat{v}_{i,t}\|_*^2] \tag{4.50b}$$

$$+ \mathbb{E}\left[\max_{x_i\in\mathcal{K}_i} \sum_{t\in\mathcal{T}} \langle Z_{i,t}, X_{i,t} - x_i\rangle\right]. \tag{4.50c}$$

With $\gamma_{i,t}$ nonincreasing, the first two terms are readily bounded as

$$(4.50a) \leq \sum_{t\in\mathcal{T}} \left(\frac{1}{\gamma_{i,t}} - \frac{1}{\gamma_{i,t-1}}\right) H_i(x_i) \leq \frac{H_i(x_i)}{\gamma_{i,\tau_{\text{end}}}}, \tag{4.51a}$$

$$(4.50b) \leq \frac{K_i}{2}\sum_{t\in\mathcal{T}} \gamma_{i,t} M_{i,t}^2, \tag{4.51b}$$

so we are left to bound (4.50c). To that end, introduce the auxiliary process

$$\tilde{X}_{i,t+1} = \mathcal{P}(\tilde{X}_{i,t}; -\gamma_{i,t}U_{i,t}) \tag{4.52}$$

with $\tilde{X}_1 = X_1$. We then have

$$\begin{aligned}\sum_{t\in\mathcal{T}} \langle Z_{i,t}, X_{i,t} - x_i\rangle &= \sum_{t\in\mathcal{T}} \langle Z_{i,t}, (X_{i,t} - \tilde{X}_{i,t}) + (\tilde{X}_{i,t} - x_i)\rangle \\ &= \sum_{t\in\mathcal{T}} \langle Z_{i,t}, X_{i,t} - \tilde{X}_{i,t}\rangle + \sum_{t\in\mathcal{T}} \langle b_{i,t}, \tilde{X}_{i,t} - x_i\rangle + \sum_{t\in\mathcal{T}} \langle U_{i,t}, \tilde{X}_{i,t} - x_i\rangle,\end{aligned} \tag{4.53}$$

so it suffices to derive a bound for each of these terms. This can be done as follows:

1. The first term of (4.53) does not depend on x_i, so we have

$$\begin{aligned}\mathbb{E}\left[\max_{x_i\in\mathcal{K}_i} \sum_{t\in\mathcal{T}} \langle Z_{i,t}, X_{i,t} - \tilde{X}_{i,t}\rangle\right] &= \sum_{t\in\mathcal{T}} \mathbb{E}[\mathbb{E}[\langle Z_{i,t}, X_{i,t} - \tilde{X}_{i,t}\rangle | \mathcal{F}_t]] \\ &= \sum_{t\in\mathcal{T}} \mathbb{E}[\langle b_{i,t}, X_{i,t} - \tilde{X}_{i,t}\rangle] \leq \operatorname{diam}(\mathcal{K}_i) B_{i,t},\end{aligned} \tag{4.54}$$

where, in the last step, we use the definition (3.6a) of $B_{i,t}$ and the bound

$$\langle b_{i,t}, X_{i,t} - \tilde{X}_{i,t}\rangle \leq \|X_{i,t} - \tilde{X}_{i,t}\|\|b_{i,t}\|_* \leq \operatorname{diam}(\mathcal{K}_i)\|b_{i,t}\|_*. \tag{4.55}$$

2. The second term of (4.53) can be bounded in a similar way as

$$\mathbb{E}\left[\max_{x_i\in\mathcal{K}_i} \sum_{t\in\mathcal{T}} \langle b_{i,t}, \tilde{X}_{i,t} - x_i\rangle\right] \leq \mathbb{E}[\operatorname{diam}(\mathcal{K}_i)\|b_{i,t}\|_*] \leq \operatorname{diam}(\mathcal{K}_i) B_{i,t}. \tag{4.56}$$

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

3. Finally, for the last term, Lemma A.4 with $Y_t \leftarrow -\gamma_{i,t}U_{i,t}$ and $\alpha_t = 1/\gamma_{i,t}$ gives

$$\begin{aligned}\sum_{t\in\mathcal{T}}\langle U_{i,t}, \tilde{X}_{i,t} - x_i\rangle &= \sum_{t\in\mathcal{T}}\alpha_{i,t}\langle -\gamma_{i,t}U_{i,t}, x_i - \tilde{X}_{i,t}\rangle \\ &\leq \sum_{t\in\mathcal{T}}\left(\frac{1}{\gamma_{i,t}} - \frac{1}{\gamma_{i,t-1}}\right)D(x_i, \tilde{X}_{i,t}) + \frac{1}{2K_i}\sum_{t\in\mathcal{T}}\gamma_{i,t}\|U_{i,t}\|_*^2. \end{aligned} \tag{4.57}$$

Thus, after taking expectations and telescoping, we obtain

$$\mathbb{E}\left[\max_{x_i\in\mathcal{K}_i}\langle U_{i,t}, \tilde{X}_{i,t} - x_i\rangle\right] \leq \frac{H_i(x_i)}{\gamma_{i,\tau_{\text{end}}}} + \frac{1}{2K_i}\sum_{t\in\mathcal{T}}\gamma_{i,t}\sigma_{i,t}^2. \tag{4.58}$$

The bound (4.47) then follows by plugging back all of the above in (4.50c).

We are now in a position to prove our equilibrium tracking result. Our proof strategy is to leverage the gap minimization guarantees of Algorithm 1 (as encoded in Proposition 3) together with a batch comparison idea from Besbes et al. [10].

Proof of Theorem 2. For the sake of the analysis (and only the analysis), partition the horizon of play $\mathcal{T} = [1\,..\,T]$ in m contiguous batches $\mathcal{T}_k$, $k = 1,\dots,m$, each of length Δ (except possibly the mth one, which might be smaller). We prove the error bound (4.7) by linking $err(\mathcal{T}_k)$ to $Gap(\mathcal{T}_k) = \sum_{i\in\mathcal{N}} Gap_i(\mathcal{T}_k)$ for all $k = 1,\dots,m = \lceil T/\Delta\rceil$.

More explicitly, take the batch length to be of the form $\Delta = \lceil T^q\rceil$ for some constant $q \in [0,1]$ to be determined later. In this way, the number of batches is $m = \lceil T/\Delta\rceil = \Theta(T^{1-q})$, and the kth batch is of the form $\mathcal{T}_k = [(k-1)\Delta + 1\,..\,k\Delta]$ for all $k = 1,\dots,m-1$ (the value $k = m$ is excluded as the mth batch might be smaller). Then, to bound the players' equilibrium tracking error within $\mathcal{T}_k$, the strong monotonicity property (2.2) for $\mathcal{G}_t$ gives

$$\mu\|X_t - x_t^*\|^2 \leq \langle v_t(X_t), x_t^* - X_t\rangle = \langle v_t(X_t), \hat{x} - X_t\rangle + \langle v_t(X_t), x_t^* - \hat{x}\rangle \tag{4.59}$$

for every reference action profile $\hat{x} \in \mathcal{K}$ and all $t \in \mathcal{T}$. Thus, letting $err(\mathcal{T}_k) = \sum_{t\in\mathcal{T}_k}\|X_t - x_t^*\|^2$, we obtain the batch bound

$$\begin{aligned}\mu\, err(\mathcal{T}_k) = \mu\sum_{t\in\mathcal{T}_k}\|X_t - x_t^*\|^2 &\leq \sum_{t\in\mathcal{T}_k}\langle v_t(X_t), x_t^* - X_t\rangle \\ &= \sum_{t\in\mathcal{T}_k}\langle v_t(X_t), \hat{x} - X_t\rangle + \sum_{t\in\mathcal{T}_k}\langle v_t(X_t), x_t^* - \hat{x}\rangle \\ &\leq Gap(\mathcal{T}_k) + \sum_{t\in\mathcal{T}_k}\langle v_t(X_t), x_t^* - \hat{x}\rangle. \end{aligned} \tag{4.60}$$

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

$$C_k = \sum_{t\in\mathcal{T}_k}\langle v_t(X_t), x_t^* - \hat{x}_k\rangle, \tag{4.61}$$

for the last term of (4.60). A meaningful bound for C_k can then be obtained by taking $\hat{x}_k$ to be the (unique) Nash equilibrium of the first game encountered in the batch $\mathcal{T}_k$, that is, setting $\hat{x}_k = x^*_{\min\mathcal{T}_k}$. Doing this, we obtain the series of estimates

$$\begin{aligned} C_k &\leq \sum_{t\in\mathcal{T}_k}\|v_t(X_t)\|_*\cdot\|x_t^* - \hat{x}_k\| && \text{(by Cauchy–Schwarz)} \\ &\leq \sum_{t\in\mathcal{T}_k} G\|x_t^* - \hat{x}_k\| && \text{(by Assumption 1)} \\ &\leq G\Delta\max_{t\in\mathcal{T}_k}\|x_t^* - \hat{x}_k\| && \text{(term-by-term bound)} \\ &\leq G\Delta\sum_{t\in\mathcal{T}_k}\|x_{t+1}^* - x_t^*\| && \{\text{by definition of } \hat{x}_k\} \\ &= G\Delta V(\mathcal{T}_k), && \end{aligned} \tag{4.62}$$

where, in obvious notation, we set $V(\mathcal{T}_k) = \sum_{t\in\mathcal{T}_k}\|x_{t+1}^* - x_t^*\|$. Then, plugging everything back in (4.60) and summing over all batches $k = 1,\dots,m$, we get the total bound

$$\mathbb{E}[err(T)] \leq \frac{1}{\mu}\mathbb{E}[Gap(T)] + \frac{G\Delta}{\mu}V(T). \tag{4.63}$$

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

With this estimate in hand, let $\mathcal{D}_i := \sup_{x_i,x_i'} D_i(x_i,x_i') = \max_{x_i\in\mathcal{K}_i} H_i(x_i)$, so $\mathcal{D}_i < \infty$ by Lemma A.5. Then, with $\gamma_{i,t}$ decreasing, summing the second part of Proposition 3 over all $i \in \mathcal{N}$ yields

$$\begin{aligned}\sum_{k=1}^{m}\mathbb{E}[Gap(\mathcal{T}_k)] &\leq \sum_{i\in\mathcal{N}}\left[\sum_{k=1}^{m}\frac{2\mathcal{D}_i}{\gamma_{i,k\Delta}} + 2\operatorname{diam}(\mathcal{K}_i)\sum_{t=1}^{T}B_{i,t} + \frac{1}{2K_i}\sum_{t=1}^{T}\gamma_{i,t}S_{i,t}^2\right]\\ &= \mathcal{O}\left(\Delta^{\max_i p_i}\sum_{k=1}^{m}k^{\max_i p_i} + \sum_{t=1}^{T}t^{-\min_i b_i} + \sum_{t=1}^{T}t^{-\min_i(p_i-2s_i)}\right)\\ &= \mathcal{O}\left(\Delta^{\max_i p_i}m^{1+\max_i p_i} + T^{1-\min_i b_i} + T^{1-\min_i(p_i-2s_i)}\right),\end{aligned} \tag{4.64}$$

where, in the second line, we use the fact that $\gamma_{i,t} = \Theta(1/t^{p_i})$. Because $\Delta = \mathcal{O}(T^q)$ and $m = \mathcal{O}(T/\Delta) = \mathcal{O}(T^{1-q})$, we get

$$\Delta^{\max_i p_i}m^{1+\max_i p_i} = \mathcal{O}(T^{q\max_i p_i}T^{(1-q)(1+\max_i p_i)} = \mathcal{O}(T^{1+\max_i p_i - q}). \tag{4.65}$$

In turn, this yields the error bound

$$\mathbb{E}[err(T)] = \mathcal{O}(T^{1+\max_i p_i - q} + T^{1-\min_i b_i} + T^{1-\min_i(p_i-2s_i)} + T^q V(T)), \tag{4.66}$$

so the guarantee (4.7) follows by setting $q = \max_i p_i + \min_i(p_i - 2s_i)$.

5. Learning with Payoff-Based Information

In this section, we proceed to examine a payoff-based learning scheme, that is, a method that relies only on observations of the players' realized, in-game payoffs (the so-called "bandit setting"). The first step is to introduce a payoff-based stochastic first order oracle in the spirit of Spall [60, 61]; subsequently, by mapping this oracle to the general feedback model of Section 3, we leverage the analysis of Section 4 to derive the algorithm's properties in time-varying games.

5.1. Payoff-Based Feedback and Estimation of Payoff Gradients

Heuristically, the main idea of the player's gradient estimation process is easiest to describe in one-dimensional environments. In particular, suppose that an agent wishes to estimate the derivative of an unknown function $f : \mathbb{R} \to \mathbb{R}$ at some point $x \in \mathbb{R}$. Then, by definition, given an accuracy target δ, the derivative of f at x can be approximated by two queries of f as

$$f'(x) \approx \frac{f(x+\delta) - f(x-\delta)}{2\delta}. \tag{5.1}$$

Building on this idea, $f'(x)$ can be estimated from a *single* function evaluation as follows: let w be a random variable taking the value +1 or –1 with probability 1/2, and consider the estimator

$$\hat{v} = \frac{f(x+\delta w)}{\delta}w. \tag{5.2}$$

In expectation, this gives

$$\mathbb{E}[\hat{v}] = \frac{1}{2\delta}f(x+\delta) - \frac{1}{2\delta}f(x-\delta). \tag{5.3}$$

Thus, if f' is Lipschitz continuous, we readily get $\mathbb{E}[\hat{v} - f'(x)] = \mathcal{O}(\delta)$, that is, the estimator (5.2) is accurate up to $\mathcal{O}(\delta)$.

This idea is the starting point of the so-called single-point stochastic approximation (SPSA) method that was pioneered by Spall [60, 61]. Its extension to a multidimensional setting is straightforward: if an agent seeks to estimate the gradient of a function $f : \mathbb{R}^d \to \mathbb{R}$, it suffices to sample a perturbation direction w uniformly at random from $\mathcal{E} \equiv \{\pm e_1, \dots, \pm e_d\}$ and consider the estimator

$$\hat{v} = \frac{d}{\delta}f(x+\delta w)\,w. \tag{5.4}$$

The only difference between (5.2) and (5.4) is the dimensional scaling factor d, which compensates for the fact that each principal direction of $\mathbb{R}^d$ is sampled with probability $1/d$. Then, the same reasoning as earlier shows that $\mathbb{E}[\|\hat{v} - \nabla f(x)\|] = \mathcal{O}(\delta)$.

In the presence of constraints, a caveat that arises is that the query point $\hat{x} = x + \delta w$ must remain feasible. To guarantee this, let $\mathcal{C}$ be a convex body in $\mathbb{R}^d$, and let $f : \mathcal{C} \to \mathbb{R}$ be a function whose gradient we want to estimate

at some point $x \in \mathcal{C}$. To avoid the occurrence $x + \delta w \notin \mathcal{C}$, we first transfer x toward the interior of $\mathcal{C}$ by a homothetic transformation of the form

$$x \longmapsto x^\delta \equiv x - \frac{\delta}{r}(x - p), \tag{5.5}$$

where $p \in int(\mathcal{C})$ is an interior point of $\mathcal{C}$ and $r > 0$ is such that

a. The ball $\mathcal{B}_r(p)$ is entirely contained in $\mathcal{C}$.

b. $\delta/r < 1$.

Taken together, these conditions ensure that the query point

$$\hat{x} = x^\delta + \delta w = (1 - \delta/r)x + (\delta/r)(p + r\,w) \tag{5.6}$$

belongs itself to $\mathcal{C}$ (simply note that $p + rw \in \mathcal{B}_r(p) \subseteq \mathcal{C}$).

With all this in mind, we obtain the following process for estimating individual payoff gradients in the context of a continuous game $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{K}, u)$:

1. Every player $i \in \mathcal{N}$ selects a *pivot point* $x_i \in \mathcal{K}_i$ and draws a *perturbation vector* w_i uniformly at random from $\mathcal{E}_i := \{\pm e_1, \dots, \pm e_{d_i}\}$. Subsequently, each player plays

$$\hat{x}_i = x_i + \delta_i w_i + (\delta_i/r_i)(p_i - x_i) \tag{5.7}$$

and receives the associated payoffs $\hat{u}_i := u_i(\hat{x}_1, \dots, \hat{x}_N)$, $i \in \mathcal{N}$.

2. Each player constructs the single-point stochastic approximation estimate

$$\hat{v}_i = \frac{d_i}{\delta_i} \hat{u}_i \cdot w_i, \tag{5.8}$$

and the process repeats.

In this, the sampling radius δ_i and the homothety parameters $p_i \in \mathcal{K}_i$, $r_i > 0$, are chosen arbitrarily by each player $i \in \mathcal{N}$, only subject to the requirements $\mathcal{B}_{r_i}(p_i) \subseteq \mathcal{K}_i$ and $\delta_i/r_i < 1$ (to guarantee that $\hat{x}_i$ is a feasible action). Also, when unfolding over the course of a learning process, we assume that players employ a variable sampling radius $\delta_{i,t}$ (similar to the players' individual step-size policy $\gamma_{i,t}$). In this way, the estimator (5.8) can be seen as a payoff-based oracle that can be coupled with Algorithm 1 to generate a new candidate action and continue playing. For a pseudocode implementation of the resulting policy, see Algorithm 2.

Remark 2. Throughout this section, we tacitly assume that the players' action spaces are convex bodies, that is, they have nonempty topological interior. This assumption is only made for convenience: if this is not the case, it suffices to replace the basis vectors $\{\pm e_k\}$ with a basis of the affine hull of each player's action space and proceed in the same way.

Algorithm 2 (Payoff-Based Learning via Mirror Descent)

Require: `step-size` $\gamma_{i,t} > 0$`; sampling radius` $\delta_{i,t} > 0$`; homothety parameters` $p_i \in \mathcal{K}_i$, $r_i > 0$

`initialize` $X_{i,1} \in \mathcal{K}_{h_i}$ `# initialize pivot`

for $t = 1, 2, \dots$ **do** `simultaneously for all` $i = 1, \dots, N$

  `draw` $W_{i,t}$ `uniformly from` $\{\pm e_1, \dots, \pm e_{d_i}\}$ `# random perturbation`

  `play` $\hat{X}_{i,t} = X_{i,t} + \delta_{i,t} W_{i,t} + (\delta_{i,t}/r_{i,t})(p_i - X_{i,t})$ `# select action`

  `receive` $\hat{u}_{i,t} \equiv u_{i,t}(\hat{X}_{i,t}; \hat{X}_{-i,t})$ `# get payoff`

  `set` $\hat{v}_{i,t} = (d_i/\delta_{i,t})\,\hat{u}_{i,t} W_{i,t}$ `# estimate gradient`

  `set` $X_{i,t+1} \leftarrow \mathcal{P}_i(X_{i,t}; \gamma_{i,t} \hat{v}_{i,t})$ `# update pivot`

end for

5.2. Analysis and Results

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

Lemma 3. *The SPSA estimator* (5.8) *satisfies*

$$\|\mathbb{E}[\hat{v}_i - v_i(x)]\|_* = \mathcal{O}(\delta_{\max}^2/\delta_i) \quad \text{and} \quad \mathbb{E}[\|\hat{v}_i\|_*^2] = \mathcal{O}(1/\delta_i^2). \tag{5.9}$$

Here, $\delta_{\max} = \max_i \delta_i$.

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

Proof. The second moment bound $\mathbb{E}[\|\hat{v}_i\|_*^2] = \mathcal{O}(1/\delta_i^2)$ follows trivially from the definition (5.8) of $\hat{v}$ and the boundedness of u_i. As for our first claim, let

$$\xi_i = \hat{x}_i - x_i = \delta_i w_i + (\delta_i/r_i)(p_i - x_i). \tag{5.10}$$

Set $\xi = (\xi_i)_{i\in\mathcal{N}}$. Then, by the smoothness of u_i, a first order Taylor expansion with integral remainder gives

$$\hat{v}_i = \frac{d_i}{\delta_i} u_i(\hat{x}) \cdot w_i = \frac{d_i}{\delta_i} u_i(x) \cdot w_i + \frac{d_i}{\delta_i} \sum_{j\in\mathcal{N}} \langle \nabla_{x_j} u_i(x), \xi_j \rangle w_i \tag{5.11a}$$

$$+ \sum_{j,k\in\mathcal{N}} \int_0^1 (1-t)\, \xi_j^\top \nabla^2_{x_j x_k} u_i(x + t\xi)\, \xi_k dt \cdot w_i. \tag{5.11b}$$

Hence, taking expectations, the first term becomes

$$\begin{aligned}
\mathbb{E}[(5.11a)] &= \frac{d_i}{\delta_i} \mathbb{E}[\langle v_i(x), \xi_i \rangle w_i] + \frac{d_i}{\delta_i} \sum_{j\neq i} \langle \nabla_{x_j} u_i(x), \mathbb{E}[\xi_j] \rangle \mathbb{E}[w_i] \\
&= d_i \mathbb{E}[\langle v_i(x), w_i \rangle w_i] = d_i \cdot \frac{1}{2d_i} \sum_{\ell=1}^{d_i} [v_{i\ell}(x) e_{i\ell} - v_{i\ell}(x)(-e_{i\ell})] \\
&= v_i(x),
\end{aligned} \tag{5.12}$$

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

$$\mathbb{E}[(5.11b)] = \frac{d_i}{\delta_i} \sum_{j,k\in\mathcal{N}} \delta_j \delta_k \mathbb{E}\left[\int_0^1 (1-t)\, \xi_j^\top \nabla^2_{x_j x_k} u_i(x + t\xi)\, \xi_k dt \cdot w_i \right] = \mathcal{O}(\delta_{\max}^2/\delta_i), \tag{5.13}$$

where we use the fact that $\mathcal{K}$ is compact and u_i is C^2-smooth over $\mathcal{K}$. Our claim then follows by combining the bounds (5.12) and (5.13).

We are now in a position to state and prove our main result for the payoff-based learning policy outlined in Algorithm 2.

Theorem 3. *Let $\mathcal{G}_t$ be a time-varying game satisfying Assumption 1. Suppose further that each player $i \in \mathcal{N}$ runs Algorithm 1 with step size $\gamma_{i,t} \propto t^{-p_i}$ and sampling radius $\delta_{i,t} \propto t^{-q_i}$ for some $p_i, q_i \in (0,1]$. Then,*

1. *If $\mathcal{G}_t$ stabilizes to a strictly monotone game $\mathcal{G}$ at a rate $R_{i,t} = \mathcal{O}(1/t^{r_i})$, $r_i > 0$, and $p_i = p > \max\{1 - r_i, 1 + q_i - 2q_{\min}, 1/2 + q_i\}$ for all $i \in \mathcal{N}$, the sequence of chosen actions $\hat{X}_t$, $t = 1, 2, \dots$, converges to the Nash equilibrium of $\mathcal{G}$ with probability one. In particular, convergence to a Nash equilibrium is guaranteed under the choice $p_i = 1$, $q_i = 1/3$.*
2. *If $\mathcal{G}_t$ is strongly monotone and its drift is bounded as $V(T) = \mathcal{O}(T^r)$ for some $r < 1$, the sequence of chosen actions $\hat{X}_t$, $t = 1, 2, \dots$ enjoys the equilibrium tracking guarantee:*

$$\mathbb{E}\left[\sum_{t=1}^T \|\hat{X}_t - x_t^*\|^2 \right] = \mathcal{O}\Big(T^{1 - \min_i (p_i - 2q_i)} + T^{1 + q_{\max} - 2q_{\min}} + T^{r + p_{\max} + \min_i (p_i - 2q_i)} \Big), \tag{5.14}$$

where x_t^* denotes the (necessarily unique) Nash equilibrium of $\mathcal{G}_t$, and we set $p_{\min/\max} = \min/\max_i p_i$ and $q_{\min/\max} = \min/\max_i q_i$. In particular, for $p_i = 3(1-r)/5$ and $q_i = (1-r)/5$, we get the optimized tracking guarantee:

$$\mathbb{E}\left[\sum_{t=1}^T \|\hat{X}_t - x_t^*\|^2 \right] = \mathcal{O}(T^{\frac{4+r}{5}}). \tag{5.15}$$

Theorem 3 combines two regimes: part 1 treats time-varying games that stabilize to a well-defined limit, whereas part 2 concerns the case in which the game evolves without converging. This is in direct analogy to Theorems 1 and 2 for the case of generic stochastic first order oracle (SFO) feedback and, indeed, Theorem 3 draws heavily on these results. However, there is now a discrepancy between the actions $\hat{X}_t$ chosen by the players and the candidate actions X_t on which the SPSA estimator (5.8) returns feedback. We explain this difference in the proof of Theorem 3.

Proof of Theorem 3. Let X_t, $t = 1, 2, \dots$, be the sequence of pivot points generated by Algorithm 2: specifically, X_t is given by (MD), but the players' realized action profile $\hat{X}_t$ is given by (5.6). Then, by Lemma 3, it follows that

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

the SPSA estimator $\hat{v}_t$ of (5.8) returns feedback of the form (SFO) on X_t with bias and variance bounded as $B_t = \mathcal{O}(\delta_{i,t}) = \mathcal{O}(1/t^{q_i})$ and $M_t^2 = \mathcal{O}(1/\delta_{i,t}^2) = \mathcal{O}(t^{2q_i})$, respectively. Because the sequence X_t is generated via the prox-rule $X_{t+1} = \mathcal{P}(X_t; \gamma_t \hat{v}_t)$ of Algorithm 1, we have

1. If $\mathcal{G}_t$ stabilizes to a strictly monotone game $\mathcal{G}$, invoking Corollary 1 with $b_i = s_i = q_i$ shows that the sequence of pivot points X_t converges (a.s.) to the (necessarily unique) equilibrium of $\mathcal{G}$ as long as $p_i = p > \max\{1 - r_i, 1 - q_i, 1/2 + q_i\}$ for all $i \in \mathcal{N}$. Because $\|\hat{X}_t - X_t\| = \mathcal{O}(\delta_{i,t})$ and $\delta_{i,t} \to 0$, our claim follows.
2. If $\mathcal{G}_t$ is strongly monotone with drift $V(T) = \mathcal{O}(T^r)$, Theorem 2 gives

$$\mathbb{E}[err(T)] = \mathcal{O}(T^{1-\min_i(p_i-2q_i)} + T^{1+q_{\max}-2q_{\min}} + T^{p_{\max}+\min_i(p_i-2q_i)}V(T)), \tag{5.16}$$

where, by virtue of Lemma 3, we set $s_i = q_i$ and $b_i = 2q_{\min} - q_i$ in (4.8). However, by (5.6) and the compactness of $\mathcal{K}$, we also have $\|\hat{X}_t - X_t\| = \mathcal{O}(\delta_{i,t}) = \mathcal{O}(1/t^{q_{\min}})$, implying, in turn, that

$$\begin{aligned}\frac{1}{2}\sum_{t=1}^{T}\|\hat{X}_t - x_t^*\|^2 &\leq \sum_{t=1}^{T}\|\hat{X}_t - X_t\|^2 + \sum_{t=1}^{T}\|X_t - x_t^*\|^2\\ &= \mathcal{O}(T^{1-2q_{\min}}) + \sum_{t=1}^{T}\|X_t - x_t^*\|^2.\end{aligned} \tag{5.17}$$

Putting all this together, we conclude that $\mathbb{E}[\sum_{t=1}^{T}\|\hat{X}_t - x_t^*\|^2]$ is bounded as per (5.14), and our proof is complete.

As a special case, part 1 of Theorem 3 implies that the sequence of play induced by Algorithm 2 in a fixed strictly monotone game $\mathcal{G}_t \equiv \mathcal{G}$ converges to a Nash equilibrium with probability one as long as $p > \max\{1 - q, 1/2 + q\}$. In this way, we recover a recent result by Bravo et al. [13], who use a different form of the SPSA estimator (5.8) to establish the convergence of payoff-based no-regret learning in constant, monotone games. It is also possible to undertake a finer analysis for the method's rate of convergence in the case in which the limit game $\mathcal{G}$ is strongly monotone, but this lies beyond the scope of this work.

6. Further Results and Discussion

In this section, we proceed to discuss some extensions and applications of our results that otherwise disrupt the flow of our paper.

6.1. Games with Randomly Evolving Payoffs

We begin by discussing some applications of our results to games that evolve randomly over time—that is, when $\mathcal{G}_t$ is determined by some randomly drawn parameter ω_t describing the "state of the world." Randomly evolving games of this type are commonly referred to as *stochastic Nash games* in the mathematical optimization, control, and engineering literatures (Cui et al. [19], Ravat and Shanbhag [52]), where they are sometimes analyzed within a more general framework featuring joint coupling constraints. For example, in the wireless communications problem we describe earlier (Example 2), this corresponds to the case in which the users' channel gains $g_{i,t}$ fluctuate randomly between transmission frames—the so-called fast-fading channel model (Mertikopoulos et al. [42], Tse and Viswanath [65]).

To define this game-theoretic setting in detail, suppose that the players' utilities are determined by an ensemble of random functions of the form $\tilde{u}_i : \mathcal{K} \times \Omega \to \mathbb{R}$, where Ω has the structure of a complete probability space and each $\tilde{u}_i(x;\omega)$ is assumed to be

a. Measurable in ω.
b C^2-smooth in x with uniformly bounded derivatives.
c. Individually concave in the ith component of x.

Then, at each stage $t = 1, 2, \dots$, an independent and identically distributed state variable ω_t is drawn from Ω according to $\mathbb{P}$, and the players face the game $\mathcal{G}_t$ with payoff functions

$$u_{i,t}(x) = \tilde{u}_i(x;\omega_t) \quad \text{for all } i \in \mathcal{N}. \tag{6.1}$$

Given the randomness involved, it is meaningful to consider the associated mean game $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{K}, u)$ with payoff functions

$$u_i(x) = \mathbb{E}[\tilde{u}_i(x;\omega)] \quad \text{for all } i \in \mathcal{N}, \tag{6.2}$$

where the expectation $\mathbb{E}[\cdot]$ is taken relative to the (common) law of the state variables ω_t. It is then natural to ask whether the players' behavior under Algorithm 1 approaches a Nash equilibrium of the mean $\mathcal{G}$ as the game

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

unfolds. Our next result provides a positive result in this direction under the assumption that the players' individual payoff gradients have finite variance, that is,

$$\mathbb{E}[\|\nabla_{x_i}\tilde{u}_i(x;\omega) - \nabla_{x_i}u_i(x)\|_*^2] \leq \Sigma^2 \quad \text{for all } x \in \mathcal{K}. \tag{6.3}$$

Under this assumption, we have the following equilibrium convergence guarantee.

Theorem 4. *Let $\mathcal{G}_t$, $t = 1,2,\dots$ be a sequence of random games as before, and assume that the mean game $\mathcal{G}$ is strictly monotone. Suppose further that each player $i \in \mathcal{N}$ runs Algorithm 1 with a DGF satisfying (RC) and a step-size policy satisfying (S1), (S2), and*

$$\sum_{t=1}^{\infty}\gamma_{i,t}B_{i,t} < \infty \quad \text{and} \quad \sum_{t=1}^{\infty}\gamma_{i,t}^2 S_{i,t}^2 < \infty. \tag{S3′}$$

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

Remark 3. There are two distinct and conditionally independent sources of stochasticity in Theorem 4:

1. The randomness coming from ω_t (which determines the tth stage game $\mathcal{G}_t$).
2. The randomness in the players' oracle feedback.

In particular, we tacitly assume here that the filtration $\mathcal{F}_t$ underlying the definition (3.6) of the players' feedback process refers to the joint history of X_t and ω_t, and the statement "with probability one" likewise refers to both sources of randomness taken together.

Proof. Let $\tilde{v}_i(x;\omega) = \nabla_{x_i}\tilde{u}_i(x;\omega)$ and $v_i(x) = \nabla_{x_i}u_i(x)$. Then, by differentiating under the integral sign, we have $\mathbb{E}[\tilde{v}_i(x;\omega)] = \mathbb{E}[\nabla_{x_i}\tilde{u}_i(x;\omega)] = \nabla_{x_i}\mathbb{E}[\tilde{v}_i(x;\omega)] = v_i(x)$, so the players' oracle signal may be decomposed as

$$\hat{v}_{i,t} = v_i(X_t;\omega_t) + U_{i,t} + b_{i,t} = v_i(X_t) + \bar{U}_{i,t} + b_{i,t}, \tag{6.4}$$

where $\bar{U}_{i,t} = U_{i,t} + v_i(X_t;\omega_t) - v_i(X_t)$. Then, in a slight abuse of notation, we obtain

$$\mathbb{E}[\bar{U}_{i,t} \mid X_t,\dots,X_1] = \mathbb{E}[\mathbb{E}[\bar{U}_{i,t} \mid \mathcal{F}_t]] = 0 + \mathbb{E}[v_i(X_t;\omega_t) - v_i(X_t)] = 0, \tag{6.5}$$

and furthermore,

$$\begin{aligned}\mathbb{E}[\|\bar{U}_{i,t}\|_*^2 \mid X_t,\dots,X_1] &\leq 2\mathbb{E}[\|U_{i,t}\|_*^2 + \|v_i(X_t;\omega_t) - v_i(X_t)\|_*^2 \mid X_t,\dots,X_1] \\ &\leq 2\sigma_{i,t}^2 + 2\Sigma^2 = \mathcal{O}(S_{i,t}^2).\end{aligned} \tag{16}$$

Finally, letting $\bar{b}_{i,t} = \mathbb{E}[b_{i,t}]$, we also get $\|\bar{b}_{i,t}\|_* \leq B_{i,t}$ by definition. Accordingly, given that $\mathbb{E}[\hat{v}_{i,t}|X_t,\dots,X_1] = v_i(X_t) + \bar{b}_{i,t}$, our claim follows by applying Theorem 1 to the sequence of (strictly monotone) games $\bar{\mathcal{G}}_t \equiv \mathcal{G}$ for all $t \geq 1$.

Even though Theorem 1 plays a major role in the proof of Theorem 4, the latter is conceptually distinct from the former because it provides an equilibrium convergence result in a setting in which the sequence of stage games does not stabilize over time. Analogous results for equilibrium tracking or payoff-based learning (in the direction of Theorem 2 or 3, respectively) can also be derived, but this takes us too far afield, so we do not carry out the detailed analysis here.

6.2. Regret Bounds

We close this section with a precise statement and derivation of the dynamic regret bound (4.13) that is alluded to in Section 4.4.

Proposition 4. *Suppose that a single player runs Algorithm 1 against a sequence of concave payoff functions $u_t : \mathcal{K} \to \mathbb{R}$ with a Lipschitz DGF and step-size and oracle feedback parameters as in Theorem 2. Then, the player's dynamic regret is bounded as*

$$\mathbb{E}[DynReg(T)] = \mathcal{O}(T^{1+2s-p} + T^{1-b} + T^{2p-2s}V(T)). \tag{4.13, redux}$$

In particular, if $V(T) = \mathcal{O}(T^r)$ and the algorithm's feedback is unbiased and bounded in mean square ($b = \infty$, $s = 0$), the player enjoys the bound $\mathbb{E}[DynReg(T)] = \mathcal{O}(T^{1-p} + T^{2p+r})$. Hence, for $p = (1-r)/3$, the player achieves

$$\mathbb{E}[DynReg(T)] = \mathcal{O}(T^{\frac{2+r}{3}}). \tag{4.14, redux}$$

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

Proof of Proposition 4. As in the proof of Theorem 2, partition the horizon of play $\mathcal{T} = [1\,..\,T]$ in m contiguous batches $\mathcal{T}_k$, $k = 1, \dots, m$, each of length Δ (except possibly the mth one, which might be smaller). Then, letting $DynReg(\mathcal{T}_k) = \sum_{t\in\mathcal{T}_k}[u_{i,t}(\hat{x}_{i,t}; X_{-i,t}) - u_{i,t}(X_t)]$, we have

$$\begin{aligned} DynReg(\mathcal{T}_k) = \sum_{t\in\mathcal{T}_k}[u_t(x_t^*) - u_t(X_t)] &\leq \sum_{t\in\mathcal{T}_k}\langle v_t(X_t), x_t^* - X_t\rangle \\ &= \sum_{t\in\mathcal{T}_k}\langle v_t(X_t), \hat{x}_k - X_t\rangle + \sum_{t\in\mathcal{T}_k}\langle v_t(X_t), x_t^* - \hat{x}_k\rangle \\ &\leq Gap(\mathcal{T}_k) + \sum_{t\in\mathcal{T}_k}\langle v_t(X_t), x_t^* - \hat{x}_k\rangle, \end{aligned}$$

where $\hat{x}_k \in \mathcal{K}$ is a test action specific to each batch $k = 1, \dots, m$. Hence, repeating the series of arguments leading up to (4.65), we get the dynamic regret bound

$$\mathbb{E}[DynReg(T)] \leq \mathbb{E}[Gap(T)] + G\Delta V(T) \tag{6.8}$$

and our claim by invoking the bounds (4.66) and (4.67).

Dynamic regret guarantees of the form (4.15) already exist in the literature. Specifically, Besbes et al. [10] obtain a similar bound by exploiting the following meta-principle:

i. First, break the horizon of play into batches of size Δ.

ii. Over each batch, run an algorithm that guarantees low static regret relative to Δ.

iii. Then, fine-tune these steps in terms of the horizon T and the variation $V(T)$ of the agent's payoff functions in order to get low dynamic regret.

In our setting, if Algorithm 1 is rebooted every $\Delta \sim [T/V(T)]^{2/3}$ iterations and is run with constant step size $\gamma \sim 1/\sqrt{\Delta}$ between reboots, the meta-principle of Besbes et al. [10] guarantees the dynamic regret bound

$$\mathbb{E}[DynReg(T)] = \mathcal{O}(T^{2/3}V(T)^{1/3}). \tag{6.9}$$

Besbes et al. [10] further show that this bound is unimprovable under the blanket feedback model (SFO), so (4.15) is tight in this regard.[9]

A disadvantage of this restart approach is that

i. The batch length Δ must be chosen carefully relative to the total variation of the sequence of payoff functions encountered.

ii. At every reboot, the algorithm begins tabula rasa, essentially forgetting all knowledge it had accumulated up to the point in question.

Besbes et al. [10] already discuss some possible ways to avoid restarts, and we are aware of at least two related approaches in the literature: Jun et al. [33] propose a meta-aggregator based on coin betting, whereas Jadbabaie et al. [31] and Shahrampour and Jadbabaie [56] take an approach based on optimistic mirror descent. Importantly, both policies achieve $DynReg(T) = \mathcal{O}(V(T)^{1/2}T^{1/2})$ without prior knowledge of $V(T)$: because $V(T)^{1/2}T^{1/2} = V(T)^{1/3}V(T)^{1/6}T^{1/2} = o(V(T)^{1/3}T^{2/3})$ whenever $V(T) = o(T)$, these guarantees seem to contradict the optimality of the bound $\mathcal{O}(T^{2/3}V(T)^{1/3})$.[10] The resolution of this apparent incongruity is that Jun et al. [33] and Jadbabaie et al. [31] assume access to a *perfect* gradient oracle, whereas the discussion herein only assumes access to a *stochastic* one.

To the best of our knowledge, the perfect oracle requirement cannot be relaxed: if the players' gradient feedback is noisy, successive oracle calls cannot provide reliable information about the variation of the agent's payoff functions from one stage to the next, so the learning process cannot adapt to $V(T)$. Designing a policy that provably interpolates between the stochastic and deterministic regimes is a very fruitful question for further research, but one that lies beyond the scope of this paper.

7. Concluding Remarks

There are many interesting points for future research. A particularly promising one is to bridge the gap between the step-size policies that guarantee an optimal equilibrium tracking error and the policies that guarantee convergence to a Nash equilibrium in the case in which $\mathcal{G}_t$ stabilizes to a well-defined limit. As we see, these considerations are not always in tune: when the rules of the game fluctuate constantly, players can use very different step sizes and still track the game's equilibrium on average; by contrast, when the game stabilizes, convergence to a Nash equilibrium requires a certain compatibility between the players' step-size policies (and requires finer

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

tuning). Balancing these two objectives in an adaptive, context-agnostic manner is a rich and promising direction for future research.

When on the topic of adaptivity, it should be recalled that players with access to perfect gradient information can achieve better rates of dynamic regret minimization without any prior knowledge of the game's drift over time (Jadbabaie et al. [31], Shahrampour and Jadbabaie [56]). Whether this is still possible in the stochastic (or, worse, bandit) case is another fruitful open question for further research.

Acknowledgments

The authors are deeply grateful to the associate editor and two anonymous referees for providing many insightful comments and remarks that greatly improved the manuscript.

Appendix A. Basic Properties of Bregman Proximal Mappings

In this appendix, we collect some basic technical facts on distance-generating functions and prox-mappings. These results are not new, but given the range of conventions and definitions in the literature, we find it useful to provide here precise statements and proofs. For a detailed discussion, we refer the reader to Nemirovski et al. [46], Juditsky et al. [32], and references therein.

In what follows, h denotes a distance-generating function on a compact convex subset $\mathcal{C}$ of an d-dimensional normed space $\mathcal{X} \cong \mathbb{R}^d$ with dual $\mathcal{Y} = \mathcal{X}^*$ as per Definition 1. We begin with a basic subgradient comparison lemma:

Lemma A.1. *For all $p \in \mathcal{C}$ and all $y \in \partial h(x)$, $x \in \mathcal{C}_h$, we have*

$$\langle \nabla h(x), x - p \rangle \leq \langle y, x - p \rangle. \tag{A.1}$$

Proof. By continuity, it suffices to show that (A.1) holds for all $p \in ri\, \mathcal{C}$. To show this, fix $p \in ri\, \mathcal{C}$, and let

$$\phi(t) = h(x + t(p - x)) - [h(x) + \langle y, x + t(p - x) \rangle] \quad \text{for all } t \in [0,1]. \tag{A.2}$$

Given that h is strongly convex and $y \in \partial h(x)$, it follows that $\phi(t) \geq 0$ with equality if and only if $t = 0$. Because $\psi(t) = \langle \nabla h(x + t(p - x)) - y, p - x \rangle$ is a continuous selection of subgradients of ϕ and both ϕ and ψ are continuous over $[0,1]$, it follows that ϕ is continuously differentiable with $\phi' = \psi$ on $[0,1]$. Hence, with ϕ convex and $\phi(t) \geq 0 = \phi(0)$ for all $t \in [0,1]$, we conclude that $\phi'(0) = \langle \nabla h(x) - y, p - x \rangle \geq 0$, and our proof is complete.

We continue with a basic property of Bregman divergences known as the "three-point identity" (Chen and Teboulle [15]):

Lemma A.2 (Three-Point Identity). *For all $p \in \mathcal{C}$ and all $x, x' \in \mathcal{C}_h$, we have*

$$D(p, x) = D(p, x') + D(x', x) + \langle \nabla h(x) - \nabla h(x'), x' - p \rangle. \tag{A.3}$$

The proof of this lemma is a straightforward expansion, so we omit it. We employ this identity to estimate the Bregman divergence relative to a base point $p \in \mathcal{C}$ before and after a prox-step.

Lemma A.3. *Fix some $p \in \mathcal{C}$ and consider the recursive update rule*

$$x^+ = \mathcal{P}(x; y) \tag{A.4}$$

for $x \in \mathcal{C}_h$, $y \in \mathcal{Y}$. Then,

$$D(p, x^+) \leq D(p, x) - D(x^+, x) + \langle y, x^+ - p \rangle \tag{A.5a}$$

$$\leq D(p, x) + \langle y, x - p \rangle + \frac{1}{2K} \|y\|_*^2. \tag{A.5b}$$

Proof. By the definition (3.10) of $\mathcal{P}$, we have $y + \nabla h(x) \in \partial h(x^+)$. This means that $x^+ \in \operatorname{dom} \partial h \equiv \mathcal{C}_h$, so the three-point identity (Lemma A.2) applies. We, thus, get

$$D(p, x) = D(p, x^+) + D(x^+, x) + \langle \nabla h(x) - \nabla h(x^+), x^+ - p \rangle \tag{A.6}$$

or, after rearranging,

$$D(p, x^+) = D(p, x) - D(x^+, x) + \langle \nabla h(x^+) - \nabla h(x), x^+ - p \rangle. \tag{A.7}$$

Because $\nabla h(x) + y \in \partial h(x^+)$, Lemma A.1 yields $\langle \nabla h(x^+), x^+ - p \rangle \leq \langle y + \nabla h(x), x^+ - p \rangle$, so (A.5a) follows by plugging this bound back to (A.7).

For the second part of the lemma, first rewrite (A.5a) as

$$D(p, x^+) \leq D(p, x) + \langle y, x - p \rangle + \langle y, x^+ - x \rangle - D(x^+, x). \tag{A.8}$$

By Young's inequality, we also have

$$\langle y, x^+ - x \rangle \leq \frac{1}{2K} \|y\|_*^2 + \frac{K}{2} \|x^+ - x\|^2, \tag{A.9}$$

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

so (A.8) becomes

$$D(p,x^+) \le D(p,x) + \langle y, x-p\rangle + \frac{1}{2K}\|y\|_*^2 + \frac{K}{2}\|x^+ - x\|^2 - D(x^+,x). \tag{A.10}$$

Then, by the strong convexity of h, we obtain $D(x^+,x) = h(x^+) - h(x) - \langle \nabla h(x), x^+ - x\rangle \ge (K/2)\|x^+ - x\|^2$, and our claim follows.

This basic lemma allows us to derive the following "template inequality" for processes of the general form (A.4).

Lemma A.4. *Consider a sequence of dual vectors $Y_t \in \mathcal{Y}$, $t = 1,2,\dots$, and let*

$$X_{t+1} = \mathcal{P}(X_t; Y_t) \tag{A.11}$$

with $X_1 \in \mathcal{C}_h$ initialized arbitrarily. Then, for all $x \in \mathcal{C}$ and every nonnegative sequence $\alpha_t \ge 0$ defined over the window $\mathcal{T} = [\tau_{\text{start}} \,..\, \tau_{\text{end}}]$, we have

$$\sum_{t\in\mathcal{T}} \alpha_t \langle Y_t, x - X_t\rangle \le \sum_{t\in\mathcal{T}} (\alpha_t - \alpha_{t-1}) D(x, X_t) + \frac{1}{2K}\sum_{t\in\mathcal{T}} \alpha_t \|Y_t\|_*^2, \tag{A.12}$$

with the convention that $\alpha_{\tau_{\text{start}}-1} = 0$ in the preceding sum.

Proof. Let $D_t = D(x, X_t)$. Then, (A.5b) readily yields

$$D_{t+1} \le D_t + \langle Y_t, X_t - x\rangle + \frac{1}{2K}\|Y_t\|_*^2, \tag{A.13}$$

so after multiplying by $\alpha_t \ge 0$ and rearranging, we get

$$\alpha_t \langle Y_t, x - X_t\rangle \le \alpha_t (D_t - D_{t+1}) + \frac{\alpha_t}{2K}\|Y\|_*^2. \tag{A.14}$$

Therefore, by bringing $\langle Y_t, X_t - x\rangle$ to the left-hand side and summing over $t \in \mathcal{T}$, we get

$$\begin{aligned}\sum_{t\in\mathcal{T}} \alpha_t \langle Y_t, x - X_t\rangle &\le \sum_{t\in\mathcal{T}} \alpha_t (D_t - D_{t+1}) + \frac{1}{2K}\sum_{t\in\mathcal{T}} \alpha_t \|Y\|_*^2 \\ &= \sum_{t\in\mathcal{T}} (\alpha_t - \alpha_{t-1}) D_t - \alpha_{\tau_{\text{end}}} D_{\tau_{\text{end}}+1} + \frac{1}{2K}\sum_{t\in\mathcal{T}} \alpha_t \|Y\|_*^2.\end{aligned} \tag{A.15}$$

Because $D_{\tau_{\text{end}}+1} \ge 0$, our claim follows.

Finally, we make frequent use of the following straightforward result.

Lemma A.5. *Suppose that h is Lipschitz. Then,* $\sup_{x\in\mathcal{C}, x'\in\mathcal{C}_h} D(x,x') < \infty$.

Proof. For all $x \in \mathcal{C}$ and all $x' \in \mathcal{C}_h$, we have

$$D(x,x') = h(x) - h(x') - \langle \nabla h(x'), x - x'\rangle \le h(x) - h(x') + \|\nabla h(x')\|_* \|x - x'\|. \tag{A.16}$$

By assumption, $L \equiv \sup_{x'} \|\nabla h(x')\|_* < \infty$. Hence, with $\mathcal{C}$ compact, we readily get

$$D(x,x') \le h(x) - h(x') + L\,\text{diam}(\mathcal{C}). \tag{A.17}$$

Because $h(x) - h(x') \le \max h - \min h < \infty$, our assertion follows.

Endnotes

[1] More precisely, Rosen [53] uses the name DSC for a weighted variant of (DC) that holds as a strict inequality when $x' \neq x$. Hofbauer and Sandholm [30] use the term "stable" to refer to a class of population games that satisfy a condition similar to (DC), whereas Sandholm [54] and Sorin and Wan [59], respectively, call such games "contractive" and "dissipative." We use the term "monotone" throughout to underline the connection of (DC) with operator theory and variational inequalities.

[2] The terminology "descent" alludes to the fact that (MD) is originally studied in the context of convex minimization (as opposed to reward *maximization*). We should also mention here that "mirror descent" is sometimes used synonymously with the popular FTRL protocol of Shalev-Shwartz and Singer [58]. The two methods coincide in linear problems but not otherwise; in general, FTRL requires access to a best response oracle, so it is beyond the scope of this paper.

[3] We recall here that the subdifferential ∂h of h at x is defined as $\partial h(x) = \{y \in \mathcal{Y} : h(x') \ge h(x) + \langle y, x' - x\rangle \text{ for all } x' \in \mathcal{X}\}$. The notation $\text{dom}\,\partial h := \{x \in \text{dom}\, h : \partial h(x) \neq \varnothing\}$ stands for the domain of subdifferentiability of h, and by standard results in convex analysis, we have $ri\ \text{dom}\, h \subseteq \text{dom}\,\partial h \subseteq \text{dom}\, h$.

[4] Indeed, $D(p,x_t) = h(p) - h(x_t) - \langle \nabla h(x_t), p - x_t\rangle \ge (K/2)\|x_t - p\|^2$, so $x_t \to p$ whenever $D(p,x_t) \to 0$.

[5] More generally, the policy $\gamma_{i,t} \propto 1/t$ guarantees convergence as long as the bias decays as $B_{i,t} = \mathcal{O}(1/t^{b_i})$ for some $b_i > 0$ and the variance grows at most sublinearly ($\sigma_{i,t}^2 = \mathcal{O}(t^{2s_i})$ for some $s_i < 1/2$).

[6] In games with multiple equilibria, the norm should be replaced by the Hausdorff distance of the corresponding equilibrium sets; we focus on strongly monotone games to avoid such complications.

[7] For a concrete statement along these lines, see Besbes et al. [10, theorem 2].

[8] The guarantee (4.14) is not a consequence of Theorem 2 because it concerns function values and it makes no strong concavity assumptions for the payoff functions faced by the agent; the proof, however, is similar.

[9] Strictly speaking, Besbes et al. [10] define $V(T)$ as $V(T) = \sum_{t=1}^{T} \|u_{t+1} - u_t\|_\infty$, but this distinction is not important for our purposes.

[10] We thank one of the anonymous reviewers for bringing this point to our attention.

References

[1] Abernethy J, Bartlett PL, Rakhlin A, Tewari A (2008) Optimal strategies and minimax lower bounds for online convex games. *Proc. 21st Annual Conf. Learn. Theory.*
[2] Arora S, Hazan E, Kale S (2012) The multiplicative weights update method: A meta-algorithm and applications. *Theory Comput.* 8(1):121–164.
[3] Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (1995) Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Proc. 36th Annual Sympos. Foundations Comput. Sci.*
[4] Beck A, Teboulle M (2003) Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.* 31(3): 167–175.
[5] Beggs AW (2005) On the convergence of reinforcement learning. *J. Econom. Theory* 122(1):1–36.
[6] Benaïm M (1999) Dynamics of stochastic approximation algorithms. Azéma J, Émery M, Ledoux M, Yor M, eds. *Séminaire de Probabilités XXXIII, Lecture Notes in Mathematics*, vol. 1709 (Springer, Berlin Heidelberg), 1–68.
[7] Benaïm M, Hofbauer J, Sorin S (2005) Stochastic approximations and differential inclusions. *SIAM J. Control Optim.* 44(1):328–348.
[8] Benveniste A, Métivier M, Priouret P (1990) *Adaptive Algorithms and Stochastic Approximations* (Springer).
[9] Bervoets S, Bravo M, Faure M (2020) Learning with minimal information in continuous games. *Theoretical Econom.* 15:1471–1508.
[10] Besbes O, Gur Y, Zeevi A (2015) Non-stationary stochastic optimization. *Oper. Res.* 63(5):1227–1244.
[11] Borkar VS (1997) Stochastic approximation with two time scales. *Systems Control Lett.* 29(5):291–294.
[12] Bravo M, Mertikopoulos P (2017) On the robustness of learning in games with stochastically perturbed payoff observations. *Games Econom. Behav.* 103:41–66.
[13] Bravo M, Leslie DS, Mertikopoulos P (2018) Bandit learning in concave *N*-person games. *Proc. 32nd Internat. Conf. Neural Inform. Processing Systems.*
[14] Bubeck S, Cesa-Bianchi N (2012) *Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems*, Foundations and Trends in Machine Learning, vol. 5 (Now Publishers Inc., Boston).
[15] Chen G, Teboulle M (1993) Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM J. Optim.* 3(3):538–543.
[16] Cohen J, Héliou A, Mertikopoulos P (2017) Learning with bandit feedback in potential games. *Proc. 31st Internat. Conf. Neural Inform. Processing Systems.*
[17] Cominetti R, Melo E, Sorin S (2010) A payoff-based learning procedure and its application to traffic games. *Games Econom. Behav.* 70(1): 71–83.
[18] Coucheney P, Gaujal B, Mertikopoulos P (2015) Penalty-regulated dynamics and robust learning procedures in games. *Math. Oper. Res.* 40(3):611–633.
[19] Cui S, Franci B, Grammatico S, Shanbhag UV, Staudigl M (2021) A relaxed-inertial forward-backward-forward algorithm for stochastic generalized Nash equilibrium seeking. Preprint, submitted March 24, https://arxiv.org/abs/2103.13115.
[20] D'Oro S, Mertikopoulos P, Moustakas AL, Palazzo S (2015) Interference-based pricing for opportunistic multi-carrier cognitive radio systems. *IEEE Trans. Wireless Comm.* 14(12):6536–6549.
[21] Drusvyatskiy D, Ratliff LJ (2021) Improved rates for derivative free play in convex games. Preprint, submitted November 18, https://arxiv.org/abs/2111.09456.
[22] Erev I, Roth AE (1998) Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *Amer. Econom. Rev.* 88(4):848–881.
[23] Facchinei F, Kanzow C (2007) Generalized Nash equilibrium problems. *4OR* 5(3):173–210.
[24] Facchinei F, Pang J-S (2003) *Finite-Dimensional Variational Inequalities and Complementarity Problems.* Springer Series in Operations Research (Springer).
[25] Fang H, Harvey N, Portella V, Friedlander M (2020) Online mirror descent and dual averaging: Keeping pace in the dynamic case. *Proc. 37th Internat. Conf. Machine Learn.*
[26] Flaxman AD, Kalai AT, McMahan HB (2005) Online convex optimization in the bandit setting: Gradient descent without a gradient. *Proc. 16th Annual ACM-SIAM Sympos. Discrete Algorithms*, 385–394.
[27] Hall P, Heyde CC (1980) Martingale limit theory and its application. *Probability and Mathematical Statistics* (Academic Press, New York).
[28] Hart S, Mas-Colell A (2003) Uncoupled dynamics do not lead to Nash equilibrium. *Amer. Econom. Rev.* 93(5):1830–1836.
[29] Héliou A, Mertikopoulos P, Zhou Z (2020) Gradient-free online learning in continuous games with delayed rewards. *Proc. 37th Internat. Conf. Machine Learn.*
[30] Hofbauer J, Sandholm WH (2009) Stable games and their dynamics. *J. Econom. Theory* 144(4):1665–1693.
[31] Jadbabaie A, Rakhlin A, Shahrampour S, Sridharan K (2015) Online optimization: Competing with dynamic comparators. *Proc. 18th Internat. Conf. Artificial Intelligence Statist.*
[32] Juditsky A, Nemirovski AS, Tauvel C (2011) Solving variational inequalities with stochastic mirror-prox algorithm. *Stoch. Syst.* 1(1):17–58.
[33] Jun K-S, Orabona F, Wright SJ, Willett RM (2017) Improved strongly adaptive online learning using coin betting. *Proc. 20th Internat. Conf. Artificial Intelligence Statist.*
[34] Kelly FP, Maulloo AK, Tan DKH (1998) Rate control for communication networks: Shadow prices, proportional fairness and stability. *J. Oper. Res. Soc.* 49(3):237–252.
[35] Kivinen J, Warmuth MK (1997) Exponentiated gradient vs. gradient descent for linear predictors. *Inform. Comput.* 132(1):1–63.
[36] Laraki R, Renault J, Sorin S (2019) *Mathematical Foundations of Game Theory* (Springer).

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.

[37] Leslie DS, Collins EJ (2003) Convergent multiple-timescales reinforcement learning algorithms in normal form games. *Ann. Appl. Probab.* 13(4):1231–1251.
[38] Littlestone N, Warmuth MK (1994) The weighted majority algorithm. *Inform. Comput.* 108(2):212–261.
[39] Mertikopoulos P, Moustakas AL (2016) Learning in an uncertain world: MIMO covariance matrix optimization with imperfect feedback. *IEEE Trans. Signal Processing* 64(1):5–18.
[40] Mertikopoulos P, Zhou Z (2019) Learning in games with continuous action sets and unknown payoff functions. *Math. Programming* 173(1–2):465–507.
[41] Mertikopoulos P, Papadimitriou CH, Piliouras G (2018) Cycles in adversarial regularized learning. *Proc. 29th Annual ACM-SIAM Sympos. Discrete Algorithms.*
[42] Mertikopoulos P, Belmega EV, Negrel R, Sanguinetti L (2017) Distributed stochastic optimization via matrix exponential learning. *IEEE Trans. Signal Processing* 65(9):2277–2290.
[43] Mertikopoulos P, Lecouat B, Zenati H, Foo C-S, Chandrasekhar V, Piliouras G (2019) Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *Proc. 2019 Internat. Conf. Learn. Representations.*
[44] Monderer D, Shapley LS (1996) Potential games. *Games Econom. Behav.* 14(1):124–143.
[45] Nemirovski AS, Yudin DB (1983) *Problem Complexity and Method Efficiency in Optimization* (Wiley, New York).
[46] Nemirovski AS, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* 19(4):1574–1609.
[47] Nesterov Y (2009) Primal-dual subgradient methods for convex problems. *Math. Programming* 120(1):221–259.
[48] Orabona F, Pál D (2018) Scale-free online learning. *Theoretical Comput. Sci.* 716:50–69.
[49] Perkins S, Leslie DS (2014) Stochastic fictitious play with continuous action sets. *J. Econom. Theory* 152:179–213.
[50] Perkins S, Mertikopoulos P, Leslie DS (2017) Mixed-strategy learning with continuous action sets. *IEEE Trans. Automatic Control* 62(1): 379–384.
[51] Rakhlin A, Sridharan K (2013) Optimization, learning, and games with predictable sequences. *Proc. 27th Internat. Conf. Neural Inform. Processing Systems.*
[52] Ravat U, Shanbhag U (2011) On the characterization of solution sets of smooth and nonsmooth convex stochastic Nash games. *SIAM J. Optim.* 21(3):1168–1199.
[53] Rosen JB (1965) Existence and uniqueness of equilibrium points for concave *N*-person games. *Econometrica* 33(3):520–534.
[54] Sandholm WH (2015) Population games and deterministic evolutionary dynamics. Young HP, Zamir S, eds. *Handbook of Game Theory IV* (Elsevier), 703–778.
[55] Scutari G, Facchinei F, Palomar DP, Pang J-S (2010) Convex optimization, game theory, and variational inequality theory in multiuser communication systems. *IEEE Signal Processing Magazine* 27(3):35–49.
[56] Shahrampour S, Jadbabaie A (2018) Distributed online optimization in dynamic environments using mirror descent. *IEEE Trans. Automatic Control* 63(3):714–725.
[57] Shalev-Shwartz S (2011) Online learning and online convex optimization. *Foundations Trends Machine Learn.* 4(2):107–194.
[58] Shalev-Shwartz S, Singer Y (2006) Convex repeated games and Fenchel duality. *Proc. 19th Annual Conf. Neural Inform. Processing Systems* (MIT Press, Cambridge, MA), 1265–1272.
[59] Sorin S, Wan C (2016) Finite composite games: Equilibria and dynamics. *J. Dynamic Games* 3(1):101–120.
[60] Spall JC (1992) Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automatic Control* 37(3):332–341.
[61] Spall JC (1997) A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica J. IFAC* 33(1):109–112.
[62] Syrgkanis V, Agarwal A, Luo H, Schapire RE (2015) Fast convergence of regularized learning in games. *Proc. 29th Internat. Conf. Neural Inform. Processing Systems*, 2989–2997.
[63] Tatarenko T, Kamgarpour M (2019a) Learning generalized Nash equilibria in a class of convex games. *IEEE Trans. Automatic Control* 64(4):1426–1439.
[64] Tatarenko T, Kamgarpour M (2019b) Learning Nash equilibria in monotone games. *Proc. 58th IEEE Annual Conf. Decision Control.*
[65] Tse D, Viswanath P (2005) *Fundamentals of Wireless Communication* (Cambridge University Press, Cambridge, UK).
[66] Tullock G (1980) Efficient rent seeking. Buchanan JM, Tollison RD, Tullock G, eds. *Toward a Theory of the Rent-Seeking Society* (Texas A&M University Press, College Station, TX).
[67] Vovk VG (1990) Aggregating strategies. *Proc. Third Workshop Comput. Learn. Theory*, 371–383.

Downloaded from informs.org by [137.120.148.153] on 09 September 2022, at 01:56 . For personal use only, all rights reserved.