Bandit learning in concave N-person games

Mario Bravo, David S. Leslie, Panayotis Mertikopoulos

BANDIT LEARNING IN CONCAVE N-PERSON GAMES

MARIO BRAVO[♯], DAVID S. LESLIE[‡], AND PANAYOTIS MERTIKOPOULOS[∗]

Abstract. This paper examines the long-run behavior of learning with bandit feedback in non-cooperative concave games. The bandit framework accounts for extremely low-information environments where the agents may not even know they are playing a game; as such, the agents' most sensible choice in this setting would be to employ a no-regret learning algorithm. In general, this does not mean that the players' behavior stabilizes in the long run: no-regret learning may lead to cycles, even with perfect gradient information. However, if a standard monotonicity condition is satisfied, our analysis shows that no-regret learning based on mirror descent with bandit feedback converges to Nash equilibrium with probability 1. We also derive an upper bound for the convergence rate of the process that nearly matches the best attainable rate for *single-agent* bandit stochastic optimization.

1. Introduction

The bane of decision-making in an unknown environment is *regret:* noone wants to realize in hindsight that the decision policy they employed was strictly inferior to a plain policy prescribing the same action throughout. For obvious reasons, this issue becomes considerably more intricate when the decision-maker is subject to situational uncertainty and the "fog of war": when the only information at the optimizer's disposal is the reward obtained from a given action (the so-called "bandit" framework), is it even possible to design a no-regret policy? Especially in the context of online convex optimization (repeated decision problems with continuous action sets and convex costs), this problem becomes even more challenging because the decision-maker typically needs to infer gradient information from the observation of a single scalar. Nonetheless, despite this extra degree of difficulty, this question has been shown to admit a positive answer: regret minimization *is* possible, even with bandit feedback (Flaxman et al., 2005; Kleinberg, 2004).

In this paper, we consider a multi-agent extension of this framework where, at each stage $n = 1, 2, \dots$, of a repeated decision process, the reward of an agent is determined by the actions of all agents via a fixed mechanism: *a non-cooperative*

♯ Universidad de Santiago de Chile, Departamento de Matemática y Ciencia de la Computación

‡ Lancaster University & PROWLER.io

∗ Univ. Grenoble Alpes, CNRS, Inria, LIG 38000 Grenoble, France

E-mail addresses: mario.bravo.g@usach.cl, d.leslie@lancaster.ac.uk, panayotis.mertikopoulos@imag.fr.

2010 *Mathematics Subject Classification.* Primary 91A10, 91A26; secondary 68Q32, 68T02.

Key words and phrases. Bandit feedback; concave games; Nash equilibrium; mirror descent.

M. Bravo gratefully acknowledges the support provided by FONDECYT grant 11151003. P. Mertikopoulos was partially supported by the Huawei HIRP flagship grant ULTRON, and the French National Research Agency (ANR) grant ORACLESS (ANR–16–CE33–0004–01). Part of this work was carried out with financial support by the ECOS project C15E03.

N-person game. In general, the agents – or players – might be completely oblivious to this mechanism, perhaps even ignoring its existence: for instance, when choosing how much to bid for a good in an online auction, an agent is typically unaware of who the other bidders are, what are their specific valuations, etc. Hence, lacking any knowledge about the game, it is only natural to assume that agents will at least seek to achieve a minimal worst-case guarantee and minimize their regret. As a result, a fundamental question that arises is *a*) whether the agents' sequence of actions stabilizes to a rationally admissible state under no-regret learning; and *b*) if it does, whether convergence is affected by the information available to the agents.

Related work. In finite games, no-regret learning guarantees that the players' time-averaged, empirical frequency of play converges to the game's set of coarse correlated equilibria (CCE), and the rate of this convergence is $\mathcal{O}(1/n)$ for (λ,μ)-smooth games (Foster et al., 2016; Syrgkanis et al., 2015). In general however, this set might contain highly subpar, rationally inadmissible strategies: for instance, Viossat and Zapechelnyuk (2013) provide examples of CCE that assign positive selection probability *only* to strictly dominated strategies. In the class of potential games, Cohen et al. (2017) recently showed that the *actual* sequence of play (i.e., the sequence of actions that determine the agents' rewards at each stage) converges under no-regret learning, even with bandit feedback. Outside this class however, the players' chosen actions may cycle in perpetuity, even in simple, two-player zero-sum games with full information (Mertikopoulos et al., 2018a,b); in fact, depending on the parameters of the players' learning process, agents could even exhibit a fully unpredictable, aperiodic and chaotic behavior (Palaiopanos et al., 2017). As such, without further assumptions in place, no-regret learning in a multi-agent setting does not necessarily imply convergence to a unilaterally stable, equilibrium state.

In the broader context of games with continuous action sets (the focal point of this paper), the long-run behavior of no-regret learning is significantly more challenging to analyze. In the case of mixed-strategy learning, Perkins and Leslie (2014) and Perkins et al. (2017) showed that mixed-stratgy learning based on stochastic fictitious play converges to an ε-perturbed Nash equilibrium in potential games (but may lead to as much as $\mathcal{O}(\varepsilon n)$ regret in the process). More relevant for our purposes is the analysis of Nesterov (2009) who showed that the time-averaged sequence of play induced by a no-regret dual averaging (DA) process with noisy gradient feedback converges to Nash equilibrium in monotone games (a class which, in turn, contains all concave potential games).

The closest antecedent to our approach is the recent work of Mertikopoulos and Zhou (2018) who showed that the *actual* sequence of play generated by dual averaging converges to Nash equilibrium in the class of variationally stable games (which includes all monotone games). To do so, the authors first showed that a naturally associated continuous-time dynamical system converges, and then used the so-called *asymptotic pseudotrajectory* (APT) framework of Benaïm (1999) to translate this result to discrete time. Similar APT techniques were also used in a very recent preprint by Bervoets et al. (2018) to establish the convergence of a *payoff-based* learning algorithm in two classes of one-dimensional concave games: games with strategic complements, and ordinal potential games with isolated equilibria. The algorithm of Bervoets et al. (2018) can be seen as a special case of mirror descent coupled with a two-point gradient estimation process, suggesting several interesting links with our paper.

Our contributions. In this paper, we drop all feedback assumptions and we focus on the *bandit* framework where the only information at the players' disposal is the payoffs they receive at each stage. As we discussed above, this lack of information complicates matters considerably because players must now estimate their payoff gradients from their observed rewards. What makes matters even worse is that an agent may introduce a significant bias in the (concurrent) estimation process of another, so traditional, multiple-point estimation techniques for derivative-free optimization cannot be applied (at least, not without significant communication overhead between players).

To do away with player coordination requirements, we focus on learning processes which could be sensibly deployed in a single-agent setting and we show that, in monotone games, the sequence of play induced by a wide class of no-regret learning policies converges to Nash equilibrium with probability 1. Furthermore, by specializing to the class of strongly monotone games, we show that the rate of convergence is $\mathcal{O}(n^{-1/3})$, i.e., it is nearly optimal with respect to the attainable $\mathcal{O}(n^{-1/2})$ rate for bandit, *single-agent* stochastic optimization with strongly convex and smooth objectives (Agarwal et al., 2010; Shamir, 2013).

We are not aware of a similar Nash equilibrium convergence result for concave games with general convex action spaces and *bandit* feedback: the analysis of Mertikopoulos and Zhou (2018) requires first-order feedback, while the analysis of Bervoets et al. (2018) only applies to one-dimensional games. We find this outcome particularly appealing for practical applications of game theory (e.g., in network routing) because it shows that in a wide class of (possibly very complicated) nonlinear games, the Nash equilibrium prediction does not require full rationality, common knowledge of rationality, flawless execution, or even the knowledge that a game is being played: a commonly-used, individual no-regret algorithm suffices.

2. Problem setup and preliminaries

2.1. **Concave games.** Throughout this paper, we will focus on games with a finite number of players $i \in \mathcal{N} = \{1, \dots, N\}$ and continuous action sets. During play, every player $i \in \mathcal{N}$ selects an *action* x_i from a compact convex subset $\mathcal{X}_i$ of a d_i-dimensional normed space $\mathcal{V}_i$; subsequently, based on each player's individual objective and the *action profile* $x = (x_i; x_{-i}) \equiv (x_1, \dots, x_N)$ of all players' actions, every player receives a *reward*, and the process repeats. In more detail, writing $\mathcal{X} \equiv \prod_i \mathcal{X}_i$ for the game's *action space*, we assume that each player's reward is determined by an associated *payoff* (or *utility*) *function* $u_i \colon \mathcal{X} \to \mathbb{R}$. Since players are not assumed to "know the game" (or even that they are involved in one) these payoff functions might be a priori unknown, especially with respect to the dependence on the actions of other players. Our only structural assumption for u_i will be that $u_i(x_i; x_{-i})$ is concave in x_i for all $x_{-i} \in \mathcal{X}_{-i} \equiv \prod_{j \neq i} \mathcal{X}_j$, $i \in \mathcal{N}$.

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

Example 2.1 (Cournot competition)**.** In the standard Cournot oligopoly model, there is a finite set of *firms* indexed by $i = 1, \dots, N$, each supplying the market with a quantity $x_i \in [0, C_i]$ of some good (or service), up to the firm's production capacity C_i. By the law of supply and demand, the good is priced as a decreasing function $P(x_{\text{tot}})$ of the total amount $x_{\text{tot}} = \sum_{i=1}^{N} x_i$ supplied to the market, typically

following a linear model of the form $P(x_{\text{tot}}) = a - bx_{\text{tot}}$ for positive constants $a, b > 0$. The utility of firm i is then given by

$$u_i(x_i; x_{-i}) = x_i P(x_{\text{tot}}) - c_i x_i, \tag{2.1}$$

i.e., it comprises the total revenue from producing x_i units of the good in question minus the associated production cost (in the above, $c_i > 0$ represents the marginal production cost of firm i).

Example 2.2 (Resource allocation auctions)**.** Consider a service provider with a number of splittable *resources* $s \in \mathcal{S} = \{1, \dots, S\}$ (bandwidth, server time, GPU cores, etc.). These resources can be leased to a set of N bidders (players) who can place monetary bids $x_{is} \geq 0$ for the utilization of each resource $s \in \mathcal{S}$ up to each player's total budget b_i, i.e., $\sum_{s\in\mathcal{S}} x_{is} \leq b_i$. Once all bids are in, resources are allocated proportionally to each player's bid, i.e., the i-th player gets $\rho_{is} = (q_s x_{is}) \big/ (c_s + \sum_{j\in\mathcal{N}} x_{js})$ units of the s-th resource (where q_s denotes the available units of said resource and $c_s \geq 0$ is the "entry barrier" for bidding on it). A simple model for the utility of player i is then given by

$$u_i(x_i; x_{-i}) = \sum_{s\in\mathcal{S}} [g_i \rho_{is} - x_{is}], \tag{2.2}$$

with g_i denoting the marginal gain of player i from acquiring a unit slice of resources.

2.2. **Nash equilibrium and monotone games.** The most widely used solution concept for non-cooperative games is that of a *Nash equilibrium* (NE), defined here as any action profile $x^* \in \mathcal{X}$ that is resilient to unilateral deviations, viz.

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

By the classical existence theorem of Debreu (1952), every concave game admits a Nash equilibrium. Moreover, thanks to the individual concavity of the game's payoff functions, Nash equilibria can also be characterized via the first-order optimality condition

$$\langle v_i(x^*), x_i - x_i^* \rangle \leq 0 \quad \text{for all } x_i \in \mathcal{X}_i, \tag{2.3}$$

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

$$v_i(x) = \nabla_i u_i(x_i; x_{-i}), \tag{2.4}$$

with ∇_i denoting differentiation with respect to x_i.[1] In terms of regularity, it will be convenient to assume that each v_i is Lipschitz continuous; to streamline our presentation, this will be our standing assumption in what follows.

Starting with the seminal work of Rosen (1965), much of the literature on continuous games and their applications has focused on games that satisfy a condition known as *diagonal strict concavity* (DSC). In its simplest form, this condition posits that there exist positive constants $\lambda_i > 0$ such that

$$\sum_{i\in\mathcal{N}} \lambda_i \langle v_i(x') - v_i(x), x_i' - x_i \rangle < 0 \quad \text{for all } x, x' \in \mathcal{X},\ x \neq x'. \tag{DSC}$$

Owing to the formal similarity between (DSC) and the various operator monotonicity conditions in optimization (see e.g., Bauschke and Combettes, 2017), games that satisfy (DSC) are commonly referred to as (strictly) *monotone*. As was shown by Rosen (1965, Theorem 2), monotone games admit a unique Nash equilibrium $x^* \in \mathcal{X}$,

[1] We adopt here the standard convention of treating $v_i(x)$ as an element of the dual space $\mathcal{Y}_i \equiv \mathcal{V}_i^*$ of $\mathcal{V}_i$, with $\langle y_i, x_i \rangle$ denoting the duality pairing between $y_i \in \mathcal{Y}_i$ and $x_i \in \mathcal{X}_i \subseteq \mathcal{V}_i$.

which, in view of (DSC) and (NE), is also the unique solution of the (weighted) variational inequality

$$\sum_{i\in\mathcal{N}} \lambda_i \langle v_i(x), x_i - x_i^* \rangle < 0 \quad \text{for all } x \neq x^*. \tag{VI}$$

This property of Nash equilibria of monotone games will play a crucial role in our analysis and we will use it freely in the rest of our paper.

In terms of applications, monotonicity gives rise to a very rich class of games. As we show in the paper's supplement, Examples 2.1 and 2.2 both satisfy diagonal strict concavity (with a nontrivial choice of weights for the latter), as do atomic splittable congestion games in networks with parallel links (Orda et al., 1993; Sorin and Wan, 2016), multi-user covariance matrix optimization problems in multiple-input and multiple-output (MIMO) systems (Mertikopoulos et al., 2017), and many other problems where online decision-making is the norm. Namely, the class of monotone games contains all strictly convex-concave zero-sum games and all games that admit a (strictly) concave *potential*, i.e., a function $f\colon \mathcal{X} \to \mathbb{R}$ such that $v_i(x) = \nabla_i f(x)$ for all $x \in \mathcal{X}$, $i \in \mathcal{N}$. In view of all this (and unless explicitly stated otherwise), we will focus throughout on monotone games; for completeness, we also include in the supplement a straightforward second-order test for monotonicity.

3. Regularized no-regret learning

We now turn to the learning methods that players could employ to increase their individual rewards in an online manner. Building on Zinkevich's (2003) online gradient descent policy, the most widely used algorithmic schemes for no-regret learning in the context of online convex optimization invariably revolve around the idea of *regularization*. To name but the most well-known paradigms, "following the regularized leader" (FTRL) explicitly relies on best-responding to a regularized aggregate of the reward functions revealed up to a given stage, while online mirror descent (OMD) and its variants use a linear surrogate thereof. All these no-regret policies fall under the general umbrella of "regularized learning" and their origins can be traced back to the seminal *mirror descent* (MD) algorithm of Nemirovski and Yudin (1983).[2]

The basic idea of mirror descent is to generate a new feasible point x^+ by taking a so-called "mirror step" from a starting point x along the direction of an "approximate gradient" vector y (which we treat here as an element of the dual space $\mathcal{Y} \equiv \prod_i \mathcal{Y}_i$ of $\mathcal{V} \equiv \prod_i \mathcal{V}_i$).[3] To do so, let $h_i\colon \mathcal{X}_i \to \mathbb{R}$ be a continuous and K_i-strongly convex *distance-generating* (or *regularizer*) function, i.e.,

$$h_i(tx_i + (1-t)x_i') \leq t h_i(x_i) + (1-t)h_i(x_i') - \tfrac{1}{2}K_i t(1-t)\|x_i' - x_i\|^2, \tag{3.1}$$

for all $x_i, x_i' \in \mathcal{X}_i$ and all $t \in [0,1]$. In terms of smoothness (and in a slight abuse of notation) we also assume that the subdifferential of h_i admits a *continuous selection*, i.e., a continuous function $\nabla h_i\colon \operatorname{dom} \partial h_i \to \mathcal{Y}_i$ such that $\nabla h_i(x_i) \in \partial h_i(x_i)$ for all

[2] In a utility maximization setting, mirror descent should be called mirror *ascent* because players seek to *maximize* their rewards (as opposed to *minimizing* their losses). Nonetheless, we keep the term "descent" throughout because, despite the role reversal, it is the standard name associated with the method.

[3] For concreteness (and in a slight abuse of notation), we assume in what follows that $\mathcal{V}$ is equipped with the product norm $\|x\|^2 = \sum_i \|x_i\|^2$ and $\mathcal{Y}$ with the dual norm $\|y\|_* = \max\{\langle y, x\rangle : \|x\| \leq 1\}$.

$x_i \in \operatorname{dom} \partial h_i$.[4] Then, letting $h(x) = \sum_i h_i(x_i)$ for $x \in \mathcal{X}$ (so h is strongly convex with modulus $K = \min_i K_i$), we get a *pseudo-distance* on $\mathcal{X}$ via the relation

$$D(p, x) = h(p) - h(x) - \langle \nabla h(x), p - x \rangle, \tag{3.2}$$

for all $p \in \mathcal{X}$, $x \in \operatorname{dom} \partial h$.

This pseudo-distance is known as the *Bregman divergence* and we have $D(p, x) \geq 0$ with equality if and only if $x = p$; on the other hand, D may fail to be symmetric and/or satisfy the triangle inequality so, in general, it is not a bona fide distance function on $\mathcal{X}$. Nevertheless, we also have $D(p, x) \geq \frac{1}{2} K \|x - p\|^2$ (see the paper's supplement), so the convergence of a sequence X_n to p can be checked by showing that $D(p, X_n) \to 0$. For technical reasons, it will be convenient to also assume the converse, i.e., that $D(p, X_n) \to 0$ when $X_n \to p$. This condition is known in the literature as "Bregman reciprocity" (Chen and Teboulle, 1993), and it will be our blanket assumption in what follows (note that it is trivially satisfied by Examples 3.1 and 3.2 below).

Now, as with true Euclidean distances, $D(p, x)$ induces a *prox-mapping* given by

$$P_x(y) = \arg\min_{x' \in \mathcal{X}} \{\langle y, x - x' \rangle + D(x', x)\} \tag{3.3}$$

for all $x \in \operatorname{dom} \partial h$ and all $y \in \mathcal{Y}$. Just like its Euclidean counterpart below, the prox-mapping (3.3) starts with a point $x \in \operatorname{dom} \partial h$ and steps along the dual (gradient-like) vector $y \in \mathcal{Y}$ to produce a new feasible point $x^+ = P_x(y)$. Standard examples of this process are:

Example 3.1 (Euclidean projections)**.** Let $h(x) = \frac{1}{2}\|x\|_2^2$ denote the Euclidean squared norm. Then, the induced prox-mapping is

$$P_x(y) = \Pi(x + y), \tag{3.4}$$

with $\Pi(x) = \arg\min_{x' \in \mathcal{X}} \|x' - x\|^2$ denoting the standard Euclidean projection onto $\mathcal{X}$. Hence, the update rule $x^+ = P_x(y)$ boils down to a "vanilla", Euclidean projection step along y.

Example 3.2 (Entropic regularization and multiplicative weights)**.** Suppressing the player index for simplicity, let $\mathcal{X}$ be a d-dimensional simplex and consider the entropic regularizer $h(x) = \sum_{j=1}^d x_j \log x_j$. The induced pseudo-distance is the so-called *Kullback–Leibler* (KL) divergence $D_{\mathrm{KL}}(p, x) = \sum_{j=1}^d p_j \log(p_j / x_j)$, which gives rise to the prox-mapping

$$P_x(y) = \frac{(x_j \exp(y_j))_{j=1}^d}{\sum_{j=1}^d x_j \exp(y_j)} \tag{3.5}$$

for all $x \in \mathcal{X}^\circ$, $y \in \mathcal{Y}$. The update rule $x^+ = P_x(y)$ is widely known as the *multiplicative weights* (MW) algorithm and plays a central role for learning in multi-armed bandit problems and finite games (Arora et al., 2012; Auer et al., 1995; Freund and Schapire, 1999).

[4]Recall here that the subdifferential of h_i at $x_i \in \mathcal{X}_i$ is defined as $\partial h_i(x_i) \equiv \{y_i \in \mathcal{Y}_i : h_i(x_i') \geq h_i(x_i) + \langle y_i, x_i' - x_i \rangle \text{ for all } x_i' \in \mathcal{V}_i\}$, with the standard convention that $h_i(x_i) = +\infty$ if $x_i \in \mathcal{V}_i \setminus \mathcal{X}_i$. By standard results, the domain of subdifferentiability $\partial h_i \equiv \{x_i \in \mathcal{X}_i : \partial h_i \neq \varnothing\}$ of h_i satisfies $\mathcal{X}_i^\circ \subseteq \operatorname{dom} \partial h_i \subseteq \mathcal{X}_i$.

With all this in hand, the multi-agent *mirror descent* (MD) algorithm is given by the recursion

$$X_{n+1} = P_{X_n}(\gamma_n \hat{v}_n), \tag{MD}$$

where γ_n is a variable step-size sequence and $\hat{v}_n = (\hat{v}_{i,n})_{i\in\mathcal{N}}$ is a generic feedback sequence of estimated gradients. In the next section, we detail how this sequence is generated with first- or zeroth-order (bandit) feedback.

4. First-order vs. bandit feedback

4.1. **First-order feedback.** A common assumption in the literature is that players are able to obtain gradient information by querying a *first-order oracle* (Nesterov, 2004). i.e., a "black-box" feedback mechanism that outputs an estimate $\hat{v}_i$ of the individual payoff gradient $v_i(x)$ of the i-th player at the current action profile $x = (x_i; x_{-i}) \in \mathcal{X}$. This estimate could be either *perfect*, giving $\hat{v}_i = v_i(x)$ for all $i \in \mathcal{N}$, or *imperfect*, returning noisy information of the form $\hat{v}_i = v_i(x) + U_i$ where U_i denotes the oracle's error (random, systematic, or otherwise).

Having access to a perfect oracle is usually a tall order, either because payoff gradients are difficult to compute directly (especially without global knowledge), because they involve an expectation over a possibly unknown probability law, or for any other number of reasons. It is therefore more common to assume that each player has access to a *stochastic oracle* which, when called against a sequence of actions $X_n \in \mathcal{X}$, produces a sequence of gradient estimates $\hat{v}_n = (v_{i,n})_{i\in\mathcal{N}}$ that satisfies the following statistical assumptions:

$$\begin{aligned} &a)\ \textit{Unbiasedness:} && \mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n] = v(X_n). \\ &b)\ \textit{Finite mean square:} && \mathbb{E}[\|\hat{v}_n\|_*^2 \,|\, \mathcal{F}_n] \leq V^2 \ \text{ for some finite } V \geq 0. \end{aligned} \tag{4.1}$$

In terms of measurability, the expectation in (4.1) is conditioned on the history $\mathcal{F}_n$ of X_n up to stage n; in particular, since $\hat{v}_n$ is generated randomly from X_n, it is not $\mathcal{F}_n$-measurable (and hence not adapted). To make this more transparent, we will write $\hat{v}_n = v(X_n) + U_{n+1}$ where U_n is an adapted martingale difference sequence with $\mathbb{E}[\|U_{n+1}\|_*^2 \,|\, \mathcal{F}_n] \leq \sigma^2$ for some finite $\sigma \geq 0$.

4.2. **Bandit feedback.** Now, if players don't have access to a first-order oracle – the so-called *bandit* or *payoff-based* framework – they will need to derive an individual gradient estimate from the only information at their disposal: the actual payoffs they receive at each stage. When a function can be queried at multiple points (as few as two in practice), there are efficient ways to estimate its gradient via directional sampling techniques as in Agarwal et al. (2010). In a game-theoretic setting however, multiple-point estimation techniques do not apply because, in general, a player's payoff function depends on the actions of *all* players. Thus, when a player attempts to get a second query of their payoff function, this function may have already changed due to the query of another player – i.e., instead of sampling $u_i(\cdot; x_{-i})$, the i-th player would be sampling $u_i(\cdot; x'_{-i})$ for some $x'_{-i} \neq x_{-i}$.

Following Spall (1997) and Flaxman et al. (2005), we posit instead that players rely on a simultaneous perturbation stochastic approximation (SPSA) approach that allows them to estimate their individual payoff gradients v_i based off a *single* function evaluation. In detail, the key steps of this one-shot estimation process for each player $i \in \mathcal{N}$ are:

(0) Fix a *query radius* $\delta > 0$.[5]
(1) Pick a *pivot point* $x_i \in \mathcal{X}_i$ where player i seeks to estimate their payoff gradient.
(2) Draw a vector z_i from the unit sphere $\mathbb{S}_i \equiv \mathbb{S}^{d_i}$ of $\mathcal{V}_i \equiv \mathbb{R}^{d_i}$ and play $\hat{x}_i = x_i + \delta z_i$.[6]
(3) Receive $\hat{u}_i = u_i(\hat{x}_i; \hat{x}_{-i})$ and set

$$\hat{v}_i = \frac{d_i}{\delta} \hat{u}_i \, z_i. \tag{4.2}$$

By adapting a standard argument based on Stokes' theorem (detailed in the supplement), it can be shown that $\hat{v}_i$ is an unbiased estimator of the individual gradient of the δ-smoothed payoff function

$$u_i^\delta(x) = \frac{1}{\operatorname{vol}(\delta\mathbb{B}_i) \prod_{j \neq i} \operatorname{vol}(\delta\mathbb{S}_j)} \int_{\delta\mathbb{B}_i} \int_{\prod_{j\neq i} \delta\mathbb{S}_j} u_i(x_i + w_i; x_{-i} + z_{-i}) \, dz_1 \cdots dw_i \cdots dz_N \tag{4.3}$$

with $\mathbb{B}_i \equiv \mathbb{B}^{d_i}$ denoting the unit ball of $\mathcal{V}_i$. The Lipschitz continuity of v_i guarantees that $\|\nabla_i u_i - \nabla_i u_i^\delta\|_\infty = \mathcal{O}(\delta)$, so this estimate becomes more and more accurate as $\delta \to 0^+$. On the other hand, the second moment of $\hat{v}_i$ grows as $\mathcal{O}(1/\delta^2)$, implying in turn that the variability of $\hat{v}_i$ grows unbounded as $\delta \to 0^+$. This manifestation of the bias-variance dilemma plays a crucial role in designing no-regret policies with bandit feedback (Flaxman et al., 2005; Kleinberg, 2004), so δ must be chosen with care.

Before dealing with this choice though, it is important to highlight two feasibility issues that arise with the single-shot SPSA estimate (4.2). The first has to do with the fact that the perturbation direction z_i is chosen from the unit sphere $\mathbb{S}_i$ so it may fail to be tangent to $\mathcal{X}_i$, even when x_i is interior. To iron out this wrinkle, it suffices to sample z_i from the intersection of $\mathbb{S}_i$ with the affine hull of $\mathcal{X}_i$ in $\mathcal{V}_i$; on that account (and without loss of generality), we will simply assume in what follows that each $\mathcal{X}_i$ is a *convex body* of $\mathcal{V}_i$, i.e., it has nonempty topological interior.

The second feasibility issue concerns the size of the perturbation step: even if z_i is a feasible direction of motion, the query point $\hat{x}_i = x_i + \delta z_i$ may be unfeasible if x_i is too close to the boundary of $\mathcal{X}_i$. For this reason, we will introduce a "safety net" in the spirit of Agarwal et al. (2010), and we will constrain the set of possible pivot points x_i to lie within a suitably shrunk zone of $\mathcal{X}$.

In detail, let $\mathbb{B}_{r_i}(p_i)$ be an r_i-ball centered at $p_i \in \mathcal{X}_i$ so that $\mathbb{B}_{r_i}(p_i) \subseteq \mathcal{X}_i$. Then, instead of perturbing x_i by z_i, we consider the *feasibility adjustment*

$$w_i = z_i - r_i^{-1}(x_i - p_i), \tag{4.4}$$

and each player plays $\hat{x}_i = x_i + \delta w_i$ instead of $x_i + \delta z_i$. In other words, this adjustment moves each pivot to $x_i^\delta = x_i - r_i^{-1}\delta(x_i - p_i)$, i.e., $\mathcal{O}(\delta)$-closer to the interior base point p_i, and then perturbs x_i^δ by δz_i. Feasibility of the query point is then ensured by noting that

$$\hat{x}_i = x_i^\delta + \delta z_i = (1 - r_i^{-1}\delta) x_i + r_i^{-1}\delta(p_i + r_i z_i), \tag{4.5}$$

so $\hat{x}_i \in \mathcal{X}_i$ if $\delta/r_i < 1$ (since $p_i + r_i z_i \in \mathbb{B}_{r_i}(p_i) \subseteq \mathcal{X}_i$).

[5]For simplicity, we take δ equal for all players; the extension to player-specific δ is straightforward, so we omit it.

[6]We tacitly assume here that the query directions $z_i \in \mathbb{S}^{d_i}$ are drawn independently across players.

Algorithm 1: Multi-agent mirror descent with bandit feedback (player indices suppressed)

Require: step-size $\gamma_n > 0$, query radius $\delta_n > 0$, safety ball $\mathbb{B}_r(p) \subseteq \mathcal{X}$

choose $X \in \operatorname{dom} \partial h$ # initialization
repeat at each stage $n = 1, 2, \dots$
  draw Z uniformly from $\mathbb{S}^d$ # perturbation direction
  set $W \leftarrow Z - r^{-1}(X - p)$ # query direction
  play $\hat{X} \leftarrow X + \delta_n W$ # choose action
  receive $\hat{u} \leftarrow u(\hat{X})$ # get payoff
  set $\hat{v} \leftarrow (d/\delta_n)\hat{u} \cdot Z$ # estimate gradient
  update $X \leftarrow P_X(\gamma_n \hat{v})$ # update pivot
until end

The difference between this estimator and the oracle framework we discussed above is twofold. First, each player's *realized* action is $\hat{x}_i = x_i + \delta w_i$, not x_i, so there is a disparity between the point at which payoffs are queried and the action profile where the oracle is called. Second, the resulting estimator $\hat{v}$ is not unbiased, so the statistical assumptions (4.1) for a stochastic oracle do not hold. In particular, given the feasibility adjustment (4.4), the estimate (4.2) with $\hat{x}$ given by (4.5) satisfies

$$\mathbb{E}[\hat{v}_i] = \nabla_i u_i^\delta(x_i^\delta; x_{-i}^\delta), \tag{4.6}$$

so there are *two* sources of systematic error: an $\mathcal{O}(\delta)$ perturbation in the function, and an $\mathcal{O}(\delta)$ perturbation of each player's pivot point from x_i to x_i^δ. Hence, to capture both sources of bias and separate them from the random noise, we will write

$$\hat{v}_i = v_i(x) + U_i + b_i \tag{4.7}$$

where $U_i = \hat{v}_i - \mathbb{E}[\hat{v}_i]$ and $b_i = \nabla_i u_i^\delta(x^\delta) - \nabla_i u_i(x)$. We are thus led to the following manifestation of the bias-variance dilemma: the bias term b in (4.7) is $\mathcal{O}(\delta)$, but the second moment of the noise term U is $\mathcal{O}(1/\delta^2)$; as such, an increase in accuracy (small bias) would result in a commensurate loss of precision (large noise variance). Balancing these two factors will be a key component of our analysis.

5. Convergence analysis and results

Combining the learning framework of Section 3 with the single-shot gradient estimation machinery of Section 4, we obtain the following variant of (MD) with payoff-based, *bandit feedback:*

$$\begin{aligned} \hat{X}_n &= X_n + \delta_n W_n, \\ X_{n+1} &= P_{X_n}(\gamma_n \hat{v}_n). \end{aligned} \tag{MD-b}$$

In the above, the perturbations W_n and the estimates $\hat{v}_n$ are given respectively by (4.4) and (4.2), i.e.,

$$W_{i,n} = Z_{i,n} - r_i^{-1}(X_{i,n} - p_i) \qquad \hat{v}_{i,n} = (d_i/\delta_n) u_i(\hat{X}_n)\, Z_{i,n} \tag{5.1}$$

and $Z_{i,n}$ is drawn independently and uniformly across players at each stage n (see also Algorithm 1 for a pseudocode implementation and Fig. 1 for a schematic representation).

In the rest of this paper, our goal will be to determine the equilibrium convergence properties of this scheme in concave N-person games. Our first asymptotic result below shows that, under (MD-b), the players' learning process converges to Nash equilibrium in monotone games:

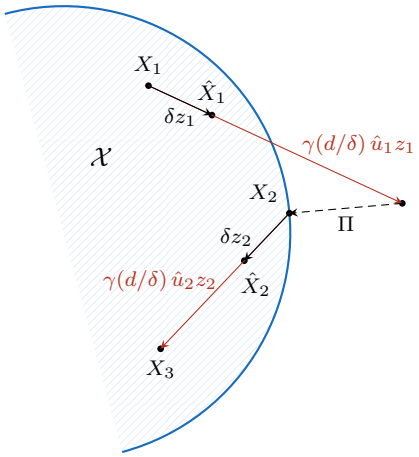

Figure 1: Schematic representation of Algorithm 1 with ordinary, Euclidean projections. To reduce visual clutter, we did not include the feasibility adjustment $r^{-1}(x-p)$ in the action selection step $X_n \mapsto \hat{X}_n$.

Theorem 5.1. *Suppose that the players of a monotone game $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{X}, u)$ follow (MD-b) with step-size γ_n and query radius δ_n such that*

$$\lim_{n\to\infty} \gamma_n = \lim_{n\to\infty} \delta_n = 0, \quad \sum_{n=1}^{\infty} \gamma_n = \infty, \quad \sum_{n=1}^{\infty} \gamma_n \delta_n < \infty, \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{\gamma_n^2}{\delta_n^2} < \infty. \tag{5.2}$$

Then, the sequence of realized actions $\hat{X}_n$ converges to Nash equilibrium with probability 1.

Even though the setting is different, the conditions (5.2) for the tuning of the algorithm's parameters are akin to those encountered in Kiefer–Wolfowitz stochastic approximation schemes and serve a similar purpose. First, the conditions $\lim_{n\to\infty} \gamma_n = 0$ and $\sum_{n=1}^{\infty} \gamma_n = \infty$ respectively mitigate the method's inherent randomness and ensure a horizon of sufficient length. The requirement $\lim_{n\to\infty} \delta_n = 0$ is also straightforward to explain: as players accrue more information, they need to decrease the sampling bias in order to have any hope of converging. However, as we discussed in Section 4, decreasing δ also increases the variance of the players' gradient estimates, which might grow to infinity as $\delta \to 0$. The crucial observation here is that new gradients enter the algorithm with a weight of γ_n so the aggregate bias after n stages is of the order of $\mathcal{O}(\sum_{k=1}^{n} \gamma_k \delta_k)$ and its variance is $\mathcal{O}(\sum_{k=1}^{n} \gamma_k^2/\delta_k^2)$. If these error terms can be controlled, there is an underlying drift that emerges over time and which steers the process to equilibrium. We make this precise in the supplement by using a suitably adjusted variant of the Bregman divergence as a quasi-Féjér energy function for (MD-b) and relying on a series of (sub)martingale convergence arguments to establish the convergence of $\hat{X}_n$ (first as a subsequence, then with probability 1).

Of course, since Theorem 5.1 is asymptotic in nature, it is not clear how to choose γ_n and δ_n so as to optimize the method's convergence rate. Heuristically, if we take schedules of the form $\gamma_n = \gamma/n^p$ and $\delta_n = \delta/n^q$ with $\gamma, \delta > 0$ and

$0 < p, q \leq 1$, the only conditions imposed by (5.2) are $p + q > 1$ and $p - q > 1/2$. However, as we discussed above, the aggregate bias in the algorithm after n stages is $\mathcal{O}(\sum_{k=1}^{n} \gamma_n \delta_n) = \mathcal{O}(1/n^{p+q-1})$ and its variance is $\mathcal{O}(\sum_{k=1}^{n} \gamma_k^2/\delta_k^2) = \mathcal{O}(1/n^{2p-2q-1})$: if the conditions (5.2) are satisfied, both error terms vanish, but they might do so at very different rates. By equating these exponents in order to bridge this gap, we obtain $q = p/3$; moreover, since the single-shot SPSA estimator (4.2) introduces a $\Theta(\delta_n)$ random perturbation, q should be taken as large as possible to ensure that this perturbation vanishes at the fastest possible rate. As a result, the most suitable choice for p and q seems to be $p = 1$, $q = 1/3$, leading to an error bound of $\mathcal{O}(1/n^{1/3})$.

We show below that this bound is indeed attainable for games that are *strongly monotone*, i.e., they satisfy the following stronger variant of diagonal strict concavity:

$$\sum_{i \in \mathcal{N}} \lambda_i \langle v_i(x') - v_i(x), x_i' - x_i \rangle \leq -\frac{\beta}{2} \|x - x'\|^2 \tag{β-DSC}$$

for some $\lambda_i, \beta > 0$ and for all $x, x' \in \mathcal{X}$. Focusing for expository reasons on the most widely used, Euclidean incarnation of the method (Example 3.1), we have:

Theorem 5.2. *Let x^* be the (necessarily unique) Nash equilibrium of a β-strongly monotone game. If the players follow (MD-b) with Euclidean projections and parameters $\gamma_n = \gamma/n$ and $\delta_n = \delta/n^{1/3}$ with $\gamma > 1/(3\beta)$ and $\delta > 0$, we have*

$$\mathbb{E}[\|\hat{X}_n - x^*\|^2] = \mathcal{O}(n^{-1/3}). \tag{5.3}$$

Theorem 5.2 is our main finite-time analysis result, so some remarks are in order. First, the step-size schedule $\gamma_n \propto 1/n$ is not required to obtain an $\mathcal{O}(n^{-1/3})$ convergence rate: as we show in the paper's supplement, more general schedules of the form $\gamma_n \propto 1/n^p$ and $\delta_n \propto 1/n^q$ with $p > 3/4$ and $q = p/3 > 1/4$, still guarantee an $\mathcal{O}(n^{-1/3})$ rate of convergence for (MD-b). To put things in perspective, we also show in the supplement that if (MD) is run with first-order oracle feedback satisfying the statistical assumptions (4.1), the rate of convergence becomes $\mathcal{O}(1/n)$. Viewed in this light, the price for not having access to gradient information is no higher than $\mathcal{O}(n^{-2/3})$ in terms of the players' equilibration rate.

Finally, it is also worth comparing the bound (D.2) to the attainable rates for stochastic convex optimization (the single-player case). For problems with objectives that are both strongly convex and smooth, Agarwal et al. (2010) attained an $\mathcal{O}(n^{-1/2})$ convergence rate with bandit feedback, which Shamir (2013) showed is unimprovable. Thus, in the single-player case, the bound (D.2) is off by $n^{1/6}$ and coincides with the bound of Agarwal et al. (2010) for strongly convex functions that are not necessarily smooth. One reason for this gap is that the $\Theta(n^{-1/2})$ bound of Shamir (2013) concerns the smoothed-out time average $\bar{X}_n = n^{-1} \sum_{k=1}^{n} X_k$, while our analysis concerns the sequence of *realized actions* $\hat{X}_n$. This difference is semantically significant: In optimization, the query sequence is just a means to an end, and only the algorithm's output matters (i.e., $\bar{X}_n$). In a game-theoretic setting however, it is the players' *realized* actions that determine their rewards at each stage, so the figure of merit is the actual sequence of play $\hat{X}_n$. This sequence is more difficult to control, so this disparity is, perhaps, not too surprising; nevertheless, we believe that this gap can be closed by using a more sophisticated single-shot estimate, e.g., as in Ghadimi and Lan (2013). We defer this analysis to the future.

6. Concluding remarks

The most sensible choice for agents who are oblivious to the presence of each other (or who are simply conservative), is to deploy a no-regret learning algorithm. With this in mind, we studied the long-run behavior of individual regularized no-regret learning policies and we showed that, in monotone games, play converges to equilibrium with probability 1, and the rate of convergence almost matches the optimal rates of *single-agent*, stochastic convex optimization. Nevertheless, several questions remain open: whether there is an intrinsic information-theoretic obstacle to bridging this gap; whether our convergence rate estimates hold with high probability (and not just in expectation); and whether our analysis extends to a fully decentralized setting where the players' updates need not be synchronous. We intend to address these questions in future work.

Appendix A. Monotone games

Our aim in this appendix is to show that the game-theoretic examples of Section 2 are both monotone. Before studying them in detail, it will be convenient to introduce a straightforward second-order test for monotonicity based on the game's Hessian matrix.

Specifically, extending the notion of the Hessian of an ordinary (scalar) function, the (λ-*weighted*) *Hessian* of a game $\mathcal{G}$ is defined as the block matrix $H_{\mathcal{G}}(x;\lambda) = (H_{ij}(x;\lambda))_{i,j\in\mathcal{N}}$ with blocks

$$H_{ij}(x;\lambda) = \frac{\lambda_i}{2}\,\nabla_j\,\nabla_i\,u_i(x) + \frac{\lambda_j}{2}(\nabla_i\,\nabla_j\,u_j(x))^\top. \tag{A.1}$$

As was shown by Rosen (1965, Theorem 6), $\mathcal{G}$ satisifes (DSC) with weight vector λ whenever $z^\top H_{\mathcal{G}}(x;\lambda)z < 0$ for all $x \in \mathcal{X}$ and all nonzero $z \in \mathcal{V} \equiv \prod_i \mathcal{V}_i$ that are tangent to $\mathcal{X}$ at x.[7] It is thus common to check for monotonicity by taking $\lambda_i = 1$ for all $i \in \mathcal{N}$ and verifying whether the unweighted Hessian of $\mathcal{G}$ is negative-definite on the affine hull of $\mathcal{X}$.

A.1. **Cournot competition (Example 2.1).** In the standard Cournot oligopoly model described in the main body of the paper, the players' payoff functions are given by

$$u_i(x) = x_i\big(a - b\textstyle\sum_j x_j\big) - c_i x_i. \tag{A.2}$$

Consequently, a simple differentiation yields

$$H_{ij}(x) = \frac{1}{2}\frac{\partial^2 u_i}{\partial x_i \partial x_j} + \frac{1}{2}\frac{\partial^2 u_j}{\partial x_j \partial x_i} = -b(1+\delta_{ij}), \tag{A.3}$$

where $\delta_{ij} = \mathbb{1}\{i=j\}$ is the Kronecker delta. This matrix is clearly negative-definite, so the game is monotone.

A.2. **Resource allocation auctions (Example 2.2).** In our auction-theoretic example, the players' payoff functions are given by

$$u_i(x_i;x_{-i}) = \sum_{s\in\mathcal{S}}\left[\frac{g_i q_s x_{is}}{c_s + \sum_{j\in\mathcal{N}} x_{js}} - x_{is}\right] \tag{A.4}$$

[7] By "tangent" we mean here that z belongs to the tangent cone $\mathrm{TC}(x)$ to $\mathcal{X}$ at x, i.e., the intersection of all supporting (closed) half-spaces of $\mathcal{X}$ at x.

To prove monotonicity in this example, we will consider the following criterion due to Goodman (1980): a game $\mathcal{G}$ satisfies (DSC) with weights λ_i, $i \in \mathcal{N}$, if:

a) Each payoff function u_i is strictly concave in x_i and convex in x_{-i}.
b) The function $\sum_{i\in\mathcal{N}} \lambda_i u_i(x)$ is concave in x.

Since the function $\phi(x) = x/(c+x)$ is strictly concave in x for all $c > 0$, the first condition above is trivial to verify. For the second, letting $\lambda_i = 1/g_i$ gives

$$\begin{aligned}\sum_{i\in\mathcal{N}} \lambda_i u_i(x) &= \sum_{i\in\mathcal{N}} \sum_{s\in\mathcal{S}} \frac{q_s x_{is}}{c_s + \sum_{j\in\mathcal{N}} x_{js}} - \sum_{i\in\mathcal{N}} \sum_{s\in\mathcal{S}} x_{is} \\ &= \sum_{s\in\mathcal{S}} q_s \frac{\sum_{i\in\mathcal{N}} x_{is}}{c_s + \sum_{i\in\mathcal{N}} x_{is}} - \sum_{i\in\mathcal{N}} \sum_{s\in\mathcal{S}} x_{is}. \end{aligned} \tag{A.5}$$

Since the summands above are all concave in their respective arguments, our claim follows.

Appendix B. Properties of Bregman proximal mappings

In this appendix, we provide some auxiliary results and estimates that are used throughout the convergence analysis of Appendix C. Some of the results we present here are not new (see e.g., Nemirovski et al., 2009); however, the set of hypotheses used to obtain them varies widely in the literature, so we provide all proofs for completeness.

In what follows, we will make frequent use of the convex conjugate $h^*\colon \mathcal{Y} \to \mathbb{R}$ of h, defined here as

$$h^*(y) = \max_{x\in\mathcal{X}} \{\langle y, x\rangle - h(x)\}. \tag{B.1}$$

By standard results in convex analysis (Rockafellar, 1970, Chap. 26), h^* is differentiable on $\mathcal{Y}$ and its gradient satisfies the identity

$$\nabla h^*(y) = \arg\max_{x\in\mathcal{X}} \{\langle y, x\rangle - h(x)\}. \tag{B.2}$$

For notational convenience, we will also write

$$Q(y) = \nabla h^*(y) \tag{B.3}$$

and we will refer to $Q\colon \mathcal{Y} \to \mathcal{X}$ as the *mirror map* generated by h.

Together with the prox-mapping induced by h, all these notions are related as follows:

Lemma B.1. *Let h be a regularizer on $\mathcal{X}$. Then, for all $x \in \operatorname{dom}\partial h$, $y \in \mathcal{Y}$, we have:*

$$a)\ \ x = Q(y) \iff y \in \partial h(x). \tag{B.4a}$$

$$b)\ \ x^+ = P_x(y) \iff \nabla h(x) + y \in \partial h(x^+) \iff x^+ = Q(\nabla h(x) + y). \tag{B.4b}$$

Finally, if $x = Q(y)$ and $p \in \mathcal{X}$, we have

$$\langle \nabla h(x), x - p\rangle \leq \langle y, x - p\rangle. \tag{B.5}$$

Remark. Note that (B.4b) directly implies that $\partial h(x^+) \neq \varnothing$, i.e., $x^+ \in \operatorname{dom}\partial h$. An immediate consequence of this is that the update rule $x \leftarrow P_x(y)$ is *well-posed*, i.e., it can be iterated in perpetuity.

Proof of Lemma B.1. To prove (B.4a), note that x solves (B.2) if and only if $y - \partial h(x) \ni 0$, i.e., if and only if $y \in \partial h(x)$. Similarly, for (B.4b), comparing (3.3) and (B.1), we see that x^+ solves (3.3) if and only if $\nabla h(x) + y \in \partial h(x^+)$, i.e., if and only if $x^+ = Q(\nabla h(x) + y)$.

For the inequality (B.5), it suffices to show it holds for interior $p \in \mathcal{X}^\circ$ (by continuity). To do so, let

$$\phi(t) = h(x + t(p - x)) - [h(x) + \langle y, x + t(p - x)\rangle]. \tag{B.6}$$

Since h is strongly convex and $y \in \partial h(x)$ by (B.4a), it follows that $\phi(t) \geq 0$ with equality if and only if $t = 0$. Moreover, note that $\psi(t) = \langle \nabla h(x + t(p-x)) - y, p - x\rangle$ is a continuous selection of subgradients of ϕ. Given that ϕ and ψ are both continuous on $[0, 1]$, it follows that ϕ is continuously differentiable and $\phi' = \psi$ on $[0, 1]$. Thus, with ϕ convex and $\phi(t) \geq 0 = \phi(0)$ for all $t \in [0, 1]$, we conclude that $\phi'(0) = \langle \nabla h(x) - y, p - x\rangle \geq 0$, from which our claim follows. □

We continue with some basic relations connecting the Bregman divergence relative to a target point before and after a prox step. The basic ingredient for this is a generalization of the law of cosines which is known in the literature as the "three-point identity" (Chen and Teboulle, 1993):

Lemma B.2. *Let h be a regularizer on $\mathcal{X}$. Then, for all $p \in \mathcal{X}$ and all $x, x' \in \operatorname{dom} \partial h$, we have*

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

Proof. By definition, we get:

$$\begin{aligned} D(p, x') &= h(p) - h(x') - \langle \nabla h(x'), p - x'\rangle \\ D(p, x) &= h(p) - h(x) - \langle \nabla h(x), p - x\rangle \\ D(x, x') &= h(x) - h(x') - \langle \nabla h(x'), x - x'\rangle. \end{aligned} \tag{B.8}$$

The lemma then follows by adding the two last lines and subtracting the first. □

With all this at hand, we have the following upper and lower bounds:

Proposition B.3. *Let h be a K-strongly convex regularizer on $\mathcal{X}$, fix some $p \in \mathcal{X}$, and let $x^+ = P_x(y)$ for $x \in \operatorname{dom} \partial h$, $y \in \mathcal{Y}$. Then, we have:*

$$D(p, x) \geq \frac{K}{2}\|x - p\|^2. \tag{B.9a}$$

$$D(p, x^+) \leq D(p, x) - D(x^+, x) + \langle y, x^+ - p\rangle \tag{B.9b}$$

$$\leq D(p, x) + \langle y, x - p\rangle + \frac{1}{2K}\|y\|_*^2 \tag{B.9c}$$

Proof of (B.9a). By the strong convexity of h, we get

$$h(p) \geq h(x) + \langle \nabla h(x), p - x\rangle + \frac{K}{2}\|p - x\|^2 \tag{B.10}$$

so (B.9a) follows by gathering all terms involving h and recalling the definition of $D(p, x)$. □

Proof of (B.9b) and (B.9c). By the three-point identity (B.7), we readily obtain

$$D(p, x) = D(p, x^+) + D(x^+, x) + \langle \nabla h(x) - \nabla h(x^+), x^+ - p\rangle, \tag{B.11}$$

and hence:

$$\begin{aligned} D(p,x^+) &= D(p,x) - D(x^+,x) + \langle \nabla h(x^+) - \nabla h(x), x^+ - p\rangle \\ &\leq D(p,x) - D(x^+,x) + \langle y, x^+ - p\rangle, \end{aligned} \tag{B.12}$$

where, in the last step, we used (B.5) and the fact that $x^+ = Q(\nabla h(x) + y)$, by (B.4b), since $x^+ = P_x(y)$. The above is just (B.9b), so the first part of our proof is complete.

To proceed with the proof of (B.9c), note that (B.12) gives

$$D(p,x^+) \leq D(p,x) + \langle y, x - p\rangle + \langle y, x^+ - x\rangle - D(x^+,x). \tag{B.13}$$

By Young's inequality (Rockafellar, 1970), we also have

$$\langle y, x^+ - x\rangle \leq \frac{K}{2}\|x^+ - x\|^2 + \frac{1}{2K}\|y\|_*^2, \tag{B.14}$$

and hence

$$\begin{aligned} D(p,x^+) &\leq D(p,x) + \langle y, x - p\rangle + \frac{1}{2K}\|y\|_*^2 + \frac{K}{2}\|x^+ - x\|^2 - D(x^+,x) \\ &\leq D(p,x) + \langle y, x - p\rangle + \frac{1}{2K}\|y\|_*^2, \end{aligned} \tag{B.15}$$

with the last step following from Lemma B.1 after plugging in x in place of p. □

Appendix C. Asymptotic convergence analysis

Our goal in this appendix is to prove Theorem 5.1. Our proof strategy will be based on a two-pronged approach. First, we will show that the pivot sequence X_n satisfies a "quasi-Fejér" property (Combettes, 2001; Combettes and Pesquet, 2015) with respect to the Bregman divergence. This quasi-Fejér property allows us to show that the Bregman divergence $D(x^*, X_n)$ with respect to a Nash equilibrium x^* of $\mathcal{G}$ converges. To show that this limit is actually zero for *some* Nash equilibrium, we prove that, with probability 1, the sequence X_n admits a (random) subsequence that converges to a Nash equilibrium. The theorem then follows by combining these two results.

To carry all this out, we begin with an auxiliary lemma for the SPSA estimation process of Section 4:

Lemma C.1. *The SPSA estimator $\hat{v} = (\hat{v}_i)_{i\in\mathcal{N}}$ given by (4.2) satisfies*

$$\mathbb{E}[\hat{v}_i] = \nabla_i u_i^\delta, \tag{C.1}$$

with u_i^δ as in (4.3). Moreover, we have $\|\nabla_i u_i^\delta - \nabla_i u_i\|_\infty = \mathcal{O}(\delta)$.

Proof. By the independence of the sampling directions z_i, $i \in \mathcal{N}$, we have

$$\begin{aligned} \mathbb{E}[\hat{v}_i] &= \frac{d_i/\delta}{\prod_j \operatorname{vol}(\mathbb{S}_j)} \int_{\mathbb{S}_1}\cdots\int_{\mathbb{S}_N} u_i(x_1 + \delta z_1, \dots, x_N + \delta z_N) z_i \, dz_1 \cdots dz_N \\ &= \frac{d_i/\delta}{\prod_j \operatorname{vol}(\delta\mathbb{S}_j)} \int_{\delta\mathbb{S}_1}\cdots\int_{\delta\mathbb{S}_N} u_i(x_1 + z_1, \dots, x_N + z_N) \frac{z_i}{\|z_i\|} \, dz_1 \cdots dz_N \\ &= \frac{d_i/\delta}{\prod_j \operatorname{vol}(\delta\mathbb{S}_j)} \int_{\delta\mathbb{S}_i}\int_{\prod_{j\neq i}\delta\mathbb{S}_j} u_i(x_i + z_i; x_{-i} + z_{-i}) \frac{z_i}{\|z_i\|} \, dz_i \, dz_{-i} \\ &= \frac{d_i/\delta}{\prod_j \operatorname{vol}(\delta\mathbb{S}_j)} \int_{\delta\mathbb{B}_i}\int_{\prod_{j\neq i}\delta\mathbb{S}_j} \nabla_i u_i(x_i + w_i; x_{-i} + z_{-i}) \, dw_i \, dz_{-i}, \end{aligned} \tag{C.2}$$

where, in the last line, we used the identity

$$\nabla \int_{\delta\mathbb{B}} f(x+w)\,dw = \int_{\delta\mathbb{S}} f(x+z)\frac{z}{\|z\|}\,dz \tag{C.3}$$

which, in turn, follows from Stokes' theorem (Flaxman et al., 2005; Lee, 2003). Since $\mathrm{vol}(\delta\mathbb{B}_i) = (\delta/d_i)\,\mathrm{vol}(\delta\mathbb{S}_i)$, the above yields $\mathbb{E}[\hat{v}_i] = \nabla_i u_i^\delta$ with u_i^δ given by (4.3).

For the second part of the lemma, let L_i denote the Lipschitz constant of v_i, i.e., $\|v_i(x') - v_i(x)\|_* \leq L_i\|x' - x\|$ for all $x, x' \in \mathcal{X}$. Then, for all $w_i \in \delta\mathbb{B}_i$ and all $z_j \in \delta\mathbb{S}_j$, $j \neq i$, we have

$$\|\nabla_i u_i(x_i + w_i; x_{-i} + z_{-i}) - \nabla_i u_i(x)\| \leq L_i\sqrt{\|w_i\|^2 + \sum\nolimits_{j\neq i}\|z_j\|^2} \leq L_i\sqrt{N}\delta. \tag{C.4}$$

Our assertion then follows by integrating and differentiating under the integral sign. □

With this basic estimate at hand, we proceed to establish the convergence of the Bregman divergence relative to the game's Nash equilibria:

Proposition C.2. *Let x^* be a Nash equilibrium of $\mathcal{G}$. Then, with assumptions as in Theorem 5.1, the Bregman divergence $D(x^*, X_n)$ converges (a.s.) to a finite random variable D_∞.*

Remark. For expository reasons, we tacitly assume above (and in what follows) that $\mathcal{G}$ satisfies (DSC) with weights $\lambda_i = 1$ for all $i \in \mathcal{N}$. If this is not the case, the Bregman divergence $D(p, x)$ should be replaced by the weight-adjusted variant

$$D^\lambda(p, x) = \sum_{i\in\mathcal{N}} \lambda_i D(p_i, x_i). \tag{C.5}$$

Since this adjustment would force us to carry around all player indices, the presentation would become significantly more cumbersome; to avoid this, we stick with the simpler, unweighted case.

Proof. Let $D_n = D(x^*, X_n)$ for some Nash equilibrium x^* of $\mathcal{G}$ and write

$$\hat{v}_n = v(X_n) + U_{n+1} + b_n, \tag{C.6}$$

where, recalling the setup of Section 4 in the main body of the paper, the noise process $U_{n+1} = \hat{v}_n - \mathbb{E}[\hat{v}_n \,|\, \mathcal{F}_n]$ is an $\mathcal{F}_n$-adapted martingale difference sequence and $b_n = v^{\delta_n}(X_n^{\delta_n}) - v(X_n)$ denotes the systematic bias of the estimator $\hat{v}_n$.[8] Then, by Proposition B.3, we have

$$\begin{aligned} D_{n+1} = D(x^*, P_{X_n}(\gamma_n\hat{v}_n)) &\leq D(x^*, X_n) + \gamma_n\langle\hat{v}_n, X_n - x^*\rangle + \frac{\gamma_n^2}{2K}\|\hat{v}_n\|_*^2 \\ &= D_n + \gamma_n\langle v(X_n) + U_{n+1} + b_n, X_n - x^*\rangle + \frac{\gamma_n^2}{2K}\|\hat{v}_n\|_*^2 \\ &\leq D_n + \gamma_n\xi_{n+1} + \gamma_n r_n + \frac{\gamma_n^2}{2K}\|\hat{v}_n\|_*^2, \end{aligned} \tag{C.7}$$

[8] Recall here that X_i^δ, $i \in \mathcal{N}$, denotes the δ-adjusted pivot $X_i^\delta = X_i + r_i^{-1}\delta(X_i - p_i)$, i.e., including the feasibility adjustment $r_i^{-1}(X_i - p_i)$.

where, in the last line, we set $\xi_{n+1} = \langle U_{n+1}, X_n - x^* \rangle$, $r_n = \langle b_n, X_n - x^* \rangle$, and we used the variational characterization (VI) of Nash equilibria of monotone games. Thus, conditioning on $\mathcal{F}_n$ and taking expectations, we get

$$\begin{aligned}\mathbb{E}[D_{n+1} \,|\, \mathcal{F}_n] &\leq D_n + \mathbb{E}[\xi_{n+1} \,|\, \mathcal{F}_n] + \gamma_n \,\mathbb{E}[r_n \,|\, \mathcal{F}_n] + \frac{\gamma_n^2}{2K} \,\mathbb{E}[\|\hat{v}_n\|_*^2 \,|\, \mathcal{F}_n] \\ &\leq D_n + \gamma_n \,\mathbb{E}[r_n \,|\, \mathcal{F}_n] + \frac{V^2}{2K} \frac{\gamma_n^2}{\delta_n^2}. \end{aligned} \tag{C.8}$$

where we set $V^2 = \sum_i d_i^2 \max_{x \in \mathcal{X}} |u_i(x)|^2$ and we used the fact that X_n is $\mathcal{F}_n$-measurable, so

$$\mathbb{E}[\xi_{n+1} \,|\, \mathcal{F}_n] = \langle \mathbb{E}[U_{n+1} \,|\, \mathcal{F}_n], X_n - x^* \rangle = 0. \tag{C.9}$$

Finally, by Lemma C.1, we have

$$\begin{aligned}\|b_n\|_* &= \|v^{\delta_n}(X_n^{\delta_n}) - v(X_n)\|_* \\ &\leq \|v^{\delta_n}(X_n^{\delta_n}) - v(X_n^{\delta_n})\|_* + \|v(X_n^{\delta_n}) - v(X_n)\|_* \\ &= \mathcal{O}(\delta_n), \end{aligned} \tag{C.10}$$

where we used the fact that v is Lipschitz continuous and $\|v^\delta - v\|_\infty = \mathcal{O}(\delta)$. This shows that there exists some $B > 0$ such that $r_n \leq B\delta_n$; as a consequence, we obtain

$$\mathbb{E}[D_{n+1} \,|\, \mathcal{F}_n] \leq D_n + B\gamma_n\delta_n + \frac{V^2}{2K} \frac{\gamma_n^2}{\delta_n^2}. \tag{C.11}$$

Now, letting $R_n = D_n + \sum_{k=n}^\infty [B\gamma_k\delta_k + (2K)^{-1}V^2\gamma_k^2/\delta_k^2]$, the estimate (C.7) gives

$$\begin{aligned}\mathbb{E}[R_{n+1} \,|\, \mathcal{F}_n] &= \mathbb{E}[D_{n+1} \,|\, \mathcal{F}_n] + \sum_{k=n+1}^\infty \left[B\gamma_k\delta_k + \frac{V^2}{2K} \frac{\gamma_k^2}{\delta_k^2} \right] \\ &\leq D_n + B\gamma_n\delta_n + \frac{V^2}{2K} \frac{\gamma_n^2}{\delta_n^2} + \sum_{k=n+1}^\infty \left[B\gamma_k\delta_k + \frac{V^2}{2K} \frac{\gamma_k^2}{\delta_k^2} \right] \\ &\leq D_n + \sum_{k=n}^\infty \left[B\gamma_k\delta_k + \frac{V^2}{2K} \frac{\gamma_k^2}{\delta_k^2} \right] \\ &= R_n, \end{aligned} \tag{C.12}$$

i.e., R_n is an $\mathcal{F}_n$-adapted supermartingale.[9] Since the series $\sum_{n=1}^\infty \gamma_n\delta_n$ and $\sum_{n=1}^\infty \gamma_n^2/\delta_n^2$ are both summable, it follows that

$$\begin{aligned}\mathbb{E}[R_n] &= \mathbb{E}[\mathbb{E}[R_n \,|\, \mathcal{F}_{n-1}]] \\ &\leq \mathbb{E}[R_{n-1}] \leq \cdots \leq \mathbb{E}[R_1] \\ &\leq \mathbb{E}[D_1] + \sum_{n=1}^\infty \left[B\gamma_n\delta_n + \frac{V^2}{2K} \frac{\gamma_n^2}{\delta_n^2} \right] \\ &< \infty \end{aligned} \tag{C.13}$$

i.e., R_n is uniformly bounded in L^1. Thus, by Doob's convergence theorem for supermartingales (Hall and Heyde, 1980, Theorem 2.5), it follows that R_n converges (a.s.) to some finite random variable R_∞. In turn, by inverting the definition of R_n, it follows that D_n converges (a.s.) to some random variable D_∞, as claimed. □

[9] In particular, this shows that $\mathbb{E}[D_n \,|\, \mathcal{F}_{n-1}]$ is quasi-Fejér in the sense of Combettes (2001).

Proposition C.3. *Suppose that the assumptions of Theorem 5.1 hold. Then, with probability* 1*, there exists a (random) subsequence* X_{n_k} *of* (MD-b) *which converges to Nash equilibrium.*

Proof. We begin with the technical observation that the set $\mathcal{X}^*$ of Nash equilibria of $\mathcal{G}$ is closed (and hence, compact). Indeed, let x_n^*, $n = 1, 2, \dots$, be a sequence of Nash equilibria converging to some limit point $x^* \in \mathcal{X}$; to show that $\mathcal{X}^*$ is closed, it suffices to show that $x^* \in \mathcal{X}$. However, since Nash equilibria of $\mathcal{G}$ satisfy the variational characterization (VI), we also have $\langle v(x), x - x_n^* \rangle \leq 0$ for all $x \in \mathcal{X}$. Hence, with $x_n^* \to x^*$ as $n \to \infty$, it follows that

$$\langle v(x), x - x^* \rangle = \lim_{n\to\infty} \langle v(x), x - x_n^* \rangle \leq 0 \quad \text{for all } x \in \mathcal{X}, \tag{C.14}$$

i.e., x^* satisfies (VI). Since $\mathcal{G}$ is monotone, we conclude that x^* is a Nash equilibrium, as claimed.

Suppose now ad absurdum that, with positive probability, the pivot sequence X_n generated by (MD-b) admits no limit points in $\mathcal{X}^*$.[10] Conditioning on this event, and given that $\mathcal{X}^*$ is compact, there exists a (nonempty) compact set $\mathcal{C} \subset \mathcal{X}$ such that $\mathcal{C} \cap \mathcal{X}^* = \varnothing$ and $X_n \in \mathcal{C}$ for all sufficiently large n. Moreover, by (VI), we have $\langle v(x), x - x^* \rangle < 0$ whenever $x \in \mathcal{C}$ and $x^* \in \mathcal{X}^*$. Therefore, by the continuity of v and the compactness of $\mathcal{X}^*$ and $\mathcal{C}$, there exists some $c > 0$ such that

$$\langle v(x), x - x^* \rangle \leq -c \quad \text{for all } x \in \mathcal{C},\ x^* \in \mathcal{X}^*. \tag{C.15}$$

To proceed, fix some $x^* \in \mathcal{X}^*$ and let $D_n = D(x^*, X_n)$ as in the proof of Proposition C.2. Then, telescoping (C.7) yields the estimate

$$D_{n+1} \leq D_1 + \sum_{k=1}^{n} \gamma_k \langle v(X_n), X_n - x^* \rangle + \sum_{k=1}^{n} \gamma_k \xi_{k+1} + \sum_{k=1}^{n} \gamma_k r_k + \sum_{k=1}^{n} \frac{\gamma_k^2}{2K} \|\hat{v}_n\|_*^2, \tag{C.16}$$

where, as in the proof of Proposition C.2, we set

$$\xi_{n+1} = \langle U_{n+1}, X_n - x^* \rangle \tag{C.17}$$

and

$$r_n = \langle b_n, X_n - x^* \rangle. \tag{C.18}$$

Subsequently, letting $\tau_n = \sum_{k=1}^{n} \gamma_k$ and using (C.15), we obtain

$$D_{n+1} \leq D_1 - \tau_n \left[c - \frac{\sum_{k=1}^{n} \gamma_k \xi_{k+1}}{\tau_n} - \frac{\sum_{k=1}^{n} \gamma_k r_k}{\tau_n} - \frac{(2K)^{-1} \sum_{k=1}^{n} \gamma_k^2 \|\hat{v}_k\|_*^2}{\tau_n} \right]. \tag{C.19}$$

Since U_n is a martingale difference sequence with respect to $\mathcal{F}_n$, we have $\mathbb{E}[\xi_{n+1} \,|\, \mathcal{F}_n] = 0$ (recall that X_n is $\mathcal{F}_n$-measurable by construction). Moreover, by construction, there exists some constant $\sigma > 0$ such that

$$\|U_{n+1}\|_*^2 \leq \frac{\sigma^2}{\delta_n^2}, \tag{C.20}$$

[10] We assume here without loss of generality that $\mathcal{X}^* \neq \mathcal{X}$; otherwise, there is nothing to show.

and hence:

$$\sum_{n=1}^{\infty} \gamma_n^2 \, \mathbb{E}[\xi_{n+1}^2 \,|\, \mathcal{F}_n] \leq \sum_{n=1}^{\infty} \gamma_n^2 \|X_n - x^*\|^2 \, \mathbb{E}[\|U_{n+1}\|_*^2 \,|\, \mathcal{F}_n] \\ \leq \operatorname{diam}(\mathcal{X})^2 \sigma^2 \sum_{n=1}^{\infty} \frac{\gamma_n^2}{\delta_n^2} < \infty. \tag{C.21}$$

Therefore, by the law of large numbers for martingale difference sequences (Hall and Heyde, 1980, Theorem 2.18), we conclude that $\tau_n^{-1} \sum_{k=1}^n \gamma_k \xi_{k+1}$ converges to 0 with probability 1.

For the third term in the brackets of (C.19) we have $r_n \to 0$ as $n \to \infty$ (a.s.). Since $\sum_{n=1}^{\infty} \gamma_n = \infty$, it follows $\sum_{k=1}^n \gamma_k r_k \big/ \sum_{k=1}^n \gamma_k \to 0$.

Finally, for the last term in the brackets of (C.19), let $S_{n+1} = \sum_{k=1}^n \gamma_k^2 \|\hat{v}_k\|_*^2$. Since $\hat{v}_k$ is $\mathcal{F}_n$-measurable for all $k = 1, 2, \dots, n-1$, we have

$$\mathbb{E}[S_{n+1} \,|\, \mathcal{F}_n] = \mathbb{E}\left[\sum_{k=1}^{n-1} \gamma_k^2 \|\hat{v}_k\|_*^2 + \gamma_n^2 \|\hat{v}_n\|_*^2 \,\middle|\, \mathcal{F}_n \right] = S_n + \gamma_n^2 \, \mathbb{E}[\|\hat{v}_n\|_*^2 \,|\, \mathcal{F}_n] \geq S_n, \tag{C.22}$$

i.e., S_n is a submartingale with respect to $\mathcal{F}_n$. Furthermore, by the law of total expectation, we also have

$$\mathbb{E}[S_{n+1}] = \mathbb{E}[\mathbb{E}[S_{n+1} \,|\, \mathcal{F}_n]] \leq V^2 \sum_{k=1}^{n} \frac{\gamma_k^2}{\delta_k^2} \leq V^2 \sum_{k=1}^{\infty} \frac{\gamma_k^2}{\delta_k^2} < \infty, \tag{C.23}$$

implying in turn that S_n is uniformly bounded in L^1. Hence, by Doob's submartingale convergence theorem (Hall and Heyde, 1980, Theorem 2.5), we conclude that S_n converges to some (almost surely finite) random variable S_∞ with $\mathbb{E}[S_\infty] < \infty$. Consequently, we have $\lim_{n\to\infty} S_{n+1}/\tau_n = 0$ with probability 1.

Applying all of the above to the estimate (C.19), we get $D_{n+1} \leq D_1 - c\tau_n/2$ for sufficiently large n, and hence, $D(x^*, X_n) \to -\infty$, a contradiction. Going back to our original assumption, this shows that at least one of the limit points of X_n must lie in $\mathcal{X}^*$, so our proof is complete. □

We are finally in a position to prove Theorem 5.1 regarding the convergence of (MD-b):

Proof of Theorem 5.1. By Proposition C.3, there exists a (possibly random) Nash equilibrium x^* of $\mathcal{G}$ such that $\|X_{n_k} - x^*\| \to 0$ for some (random) subsequence X_{n_k}. By the assumed reciprocity of the Bregman divergence, this implies that $\liminf_{n\to\infty} D(x^*, X_n) = 0$ (a.s.). Since $\lim_{n\to\infty} D(x^*, X_n)$ exists with probability 1 (by Proposition C.2), it follows that

$$\lim_{n\to\infty} D(x^*, X_n) = \liminf_{n\to\infty} D(x^*, X_n) = 0, \tag{C.24}$$

i.e., X_n converges to x^* by the first part of Proposition B.3. Since $\delta_n \to 0$ and $\|\hat{X}_n - X_n\| = \delta_n \|W_n\| = \mathcal{O}(\delta_n)$, our claim follows. □

Appendix D. Finite-time analysis and rates of convergence

We now turn to the finite-time analysis of (MD-b). To begin, we briefly recall that a game $\mathcal{G}$ is *β-strongly monotone* if it satisfies the condition

$$\sum_{i\in\mathcal{N}} \lambda_i \langle v_i(x') - v_i(x), x_i' - x_i\rangle \leq -\frac{\beta}{2}\|x - x'\|^2 \tag{β-DSC}$$

for some $\lambda_i, \beta > 0$ and for all $x, x' \in \mathcal{X}$. Our aim in what follows will be to prove the following convergence rate estimate for multi-agent mirror descent in strongly monotone games:

Theorem D.1. *Let x^* be the (unique) Nash equilibrium of a β-strongly monotone game. Then:*

a) If the players have access to a gradient oracle satisfying (4.1) and they follow (MD) with Euclidean projections and step-size sequence $\gamma_n = \gamma/n$ for some $\gamma > 1/\beta$, we have

$$\mathbb{E}[\|X_n - x^*\|^2] = \mathcal{O}(n^{-1}). \tag{D.1}$$

b) If the players only have bandit feedback and they follow (MD-b) with Euclidean projections and parameters $\gamma_n = \gamma/n$ and $\delta_n = \delta/n^{1/3}$ with $\gamma > 1/(3\beta)$ and $\delta > 0$, we have

$$\mathbb{E}[\|\hat{X}_n - x^*\|^2] = \mathcal{O}(n^{-1/3}). \tag{D.2}$$

Remark. Theorem 5.2 is recovered by the second part of Theorem D.1 above; the first part (which was alluded to in the main paper) serves as a benchmark to quantify the gap between bandit and oracle feedback.

For the proof of Theorem D.1 we will need the following lemma on numerical sequences, a version of which is often attributed to Chung (1954):

Lemma D.2. *Let a_n, $n = 1, 2, \dots$, be a non-negative sequence such that*

$$a_{n+1} \leq a_n\left(1 - \frac{P}{n^p}\right) + \frac{Q}{n^{p+q}} \tag{D.3}$$

where $0 < p \leq 1$, $q > 0$, and $P, Q > 0$. Then, assuming $P > q$ if $p = 1$, we have

$$a_n \leq \frac{Q}{R}\frac{1}{n^q} + o\left(\frac{1}{n^q}\right), \tag{D.4}$$

with $R = P$ if $p < 1$ and $R = P - q$ if $p = 1$.

Proof. Clearly, it suffices to show that $\limsup_{n\to\infty} n^q a_n \leq Q/R$. To that end, write $q_n = n[(1 + 1/n)^q - 1]$, so $(1 + 1/n)^q = 1 + q_n/n$ and $q_n \to q$ as $n \to \infty$. Then, multiplying both sides of (D.3) by $(n+1)^q$ and letting $\tilde{a}_n = a_n n^q$, we get

$$\begin{aligned}
\tilde{a}_{n+1} &\leq a_n(n+1)^q\left(1 - \frac{P}{n^p}\right) + \frac{Q(n+1)^q}{n^{p+q}} \\
&= \tilde{a}_n\left(1 + \frac{q_n}{n}\right)\left(1 - \frac{P}{n^p}\right) + \frac{Q(1 + q_n/n)}{n^p} \\
&= \tilde{a}_n\left[1 + \frac{q_n}{n} - \frac{P}{n^p} + \mathcal{O}\left(\frac{1}{n^{p+1}}\right)\right] + \frac{Q_n}{n^p},
\end{aligned} \tag{D.5}$$

where we set $Q_n = Q(1+q_n/n)$, so $Q_n \to Q$ as $n \to \infty$. Then, under the assumption that $P > q$ when $p = 1$, (D.5) can be rewritten as

$$\tilde{a}_{n+1} \leq \tilde{a}_n \left(1 - \frac{R_n}{n^p}\right) + \frac{Q_n}{n^p}, \tag{D.6}$$

for some sequence R_n with $R_n \to R$ as $n \to \infty$.

Now, fix some small enough $\varepsilon > 0$. From (D.6), we readily get

$$\tilde{a}_{n+1} \leq \tilde{a}_n - \frac{R_n \tilde{a}_n - Q_n}{n^p}. \tag{D.7}$$

Since $R_n \to R$ and $Q_n \to Q$ as $n \to \infty$, we will have $R_n > R - \varepsilon$ and $Q_n < Q + \varepsilon$ for all n greater than some n_ε. Thus, if $n \geq n_\varepsilon$ and $(R-\varepsilon)\tilde{a}_n - (Q+\varepsilon) > \varepsilon$, we will also have

$$\tilde{a}_{n+1} \leq \tilde{a}_n - \frac{R_n \tilde{a}_n - Q_n}{n^p} \leq \tilde{a}_n - \frac{(R-\varepsilon)\tilde{a}_n - (Q+\varepsilon)}{n^p} \leq \tilde{a}_n - \frac{\varepsilon}{n^p}. \tag{D.8}$$

The above shows that, as long as $\tilde{a}_n > (Q+2\varepsilon)/(R-\varepsilon)$, $\tilde{a}_n$ will decrease at least by ε/n^p at each step. In turn, since $\sum_{n=1}^{\infty}(1/n^p) = \infty$, it follows by telescoping that $\limsup_{n\to\infty} \tilde{a}_n \leq (Q+2\varepsilon)/(R-\varepsilon)$. Hence, with ε arbitrary, we conclude that $\limsup_{n\to\infty} a_n n^q \leq Q/R$, as claimed. □

Proof of Theorem D.1. We begin with the second part of the theorem; the first part will follow by setting some estimates equal to zero, so the analysis is more streamlined that way. Also, as in the previous section, we tacitly assume that (β-DSC) holds with weights $\lambda_i = 1$ for all $i \in \mathcal{N}$. If this is not the case, the Bregman divergence $D(p,x)$ should be replaced by the weight-adjusted variant (C.5), but this would only make the presentation more difficult to follow, so we omit the details.

The main component of our proof is the estimate (C.7), which, for convenience (and with notation as in the previous section), we also reproduce below:

$$D_{n+1} \leq D_n + \gamma_n \langle v(X_n), X_n - x^* \rangle + \gamma_n \xi_{n+1} + \gamma_n r_n + \frac{\gamma_n^2}{2K} \|\hat{v}_n\|_*^2. \tag{D.9}$$

In the above, since the algorithm is run with Euclidean projections, $D_n = \frac{1}{2}\|X_n - x^*\|^2$; other than that, ξ_n and r_n are defined as in (C.17) and (C.18) respectively. Since the game is β-strongly monotone and x^* is a Nash equilibrium, we further have

$$\langle v(X_n), X_n - x^* \rangle \leq \langle v(X_n) - v(x^*), X_n - x^* \rangle \leq -\frac{\beta}{2}\|X_n - x^*\|^2 = -\beta D_n, \tag{D.10}$$

so (D.9) becomes

$$D_{n+1} \leq (1 - \beta\gamma_n) D_n + \gamma_n \xi_{n+1} + \gamma_n r_n + \frac{\gamma_n^2}{2K} \|\hat{v}_n\|_*^2. \tag{D.11}$$

Thus, letting $\bar{D}_n = \mathbb{E}[D_n]$ and taking expectations, we obtain

$$\bar{D}_{n+1} \leq (1 - \beta\gamma_n)\bar{D}_n + B\gamma_n\delta_n + \frac{V^2}{2K}\frac{\gamma_n^2}{\delta_n^2}, \tag{D.12}$$

with B and V defined as in the proof of Theorem 5.1 in the previous section.

Now, substituting $\gamma_n = \gamma/n^p$ and $\delta_n = \delta/n^q$ in (D.12) readily yields

$$\bar{D}_{n+1} \leq \left(1 - \frac{\beta\gamma}{n^p}\right)\bar{D}_n + \frac{B\gamma\delta}{n^{p+q}} + \frac{V^2\gamma^2\delta^2}{2Kn^{2(p-q)}}. \tag{D.13}$$

Hence, taking $p = 1$ and $q = 1/3$, the last two exponents are equated, leading to the estimate

$$\bar{D}_{n+1} \leq \left(1 - \frac{\beta\gamma}{n}\right)\bar{D}_n + \frac{C}{n^{4/3}}, \tag{D.14}$$

with $C = \gamma\delta B + (2K)^{-1}\gamma^2\delta^2V^2$. Thus, with $\beta\gamma > 1/3$, applying Lemma D.2 with $p = 1$ and $q = 1/3$, we finally obtain $\bar{D}_n = \mathcal{O}(1/n^{1/3})$.

The proof for the oracle case is similar: the key observation is that the bound (D.12) becomes

$$\bar{D}_{n+1} \leq (1 - \beta\gamma_n)\bar{D}_n + \frac{V^2}{2K}\gamma_n^2, \tag{D.15}$$

with V defined as in (4.1). Hence, taking $\gamma_n = \gamma/n$ with $\beta\gamma > 1$ and applying again Lemma D.2 with $p = q = 1$, we obtain $\bar{D}_n = \mathcal{O}(1/n)$ and our proof is complete. □

To conclude, we note that the $\mathcal{O}(1/n^{1/3})$ bound of Theorem D.1 cannot be readily improved by choosing a different step-size schedule of the form $\gamma_n \propto 1/n^p$ for some $p < 1$. Indeed, applying Lemma D.2 to the estimate (D.13) yields a bound which is either $\mathcal{O}(1/n^q)$ or $\mathcal{O}(1/n^{p-2q})$, depending on which exponent is larger. Equating the two exponents (otherwise, one term would be slower than the other), we get $q = p/3$, leading again to a $\mathcal{O}(1/n^{1/3})$ bound. Unless one has finer control on the bias/variance of the SPSA gradient estimator used in (MD-b), we do not see a way of improving this bound in the current context.

References

Agarwal, Alekh, O. Dekel, L. Xiao. 2010. Optimal algorithms for online convex optimization with multi-point bandit feedback. *COLT '10: Proceedings of the 23rd Annual Conference on Learning Theory*.

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

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

Bauschke, Heinz H., Patrick L. Combettes. 2017. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. 2nd ed. Springer, New York, NY, USA.

Benaïm, Michel. 1999. Dynamics of stochastic approximation algorithms. Jacques Azéma, Michel Émery, Michel Ledoux, Marc Yor, eds., *Séminaire de Probabilités XXXIII*, *Lecture Notes in Mathematics*, vol. 1709. Springer Berlin Heidelberg, 1–68.

Bervoets, Sebastian, Mario Bravo, Mathieu Faure. 2018. Learning with minimal information in continuous games. https://arxiv.org/abs/1806.11506.

Chen, Gong, Marc Teboulle. 1993. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization* **3**(3) 538–543.

Chung, Kuo-Liang. 1954. On a stochastic approximation method. *The Annals of Mathematical Statistics* **25**(3) 463–483.

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

Combettes, Patrick L. 2001. Quasi-Fejérian analysis of some optimization algorithms. Dan Butnariu, Yair Censor, Simeon Reich, eds., *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*. Elsevier, New York, NY, USA, 115–152.

Combettes, Patrick L., Jean-Christophe Pesquet. 2015. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization* **25**(2) 1221–1248.

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

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

Foster, Dylan J., Thodoris Lykouris, Kathrik Sridharan, Éva Tardos. 2016. Learning in games: Robustness of fast convergence. *NIPS '16: Proceedings of the 30th International Conference on Neural Information Processing Systems*. 4727–4735.

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

Ghadimi, Saeed, Guanghui Lan. 2013. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization* **23**(4) 2341–2368.

Goodman, John C. 1980. Note on existence and uniqueness of equilibrium points for concave N-person games. *Econometrica* **48**(1) 251.

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

Kleinberg, Robert D. 2004. Nearly tight bounds for the continuum-armed bandit problem. *NIPS' 04: Proceedings of the 18th Annual Conference on Neural Information Processing Systems*.

Lee, John M. 2003. *Introduction to Smooth Manifolds*. No. 218 in Graduate Texts in Mathematics, Springer-Verlag, New York, NY.

Mertikopoulos, Panayotis, E. Veronica Belmega, Romain Negrel, Luca Sanguinetti. 2017. Distributed stochastic optimization via matrix exponential learning. *IEEE Trans. Signal Process.* **65**(9) 2277–2290.

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

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

Mertikopoulos, Panayotis, Zhengyuan Zhou. 2018. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming* .

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

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

Nesterov, Yurii. 2004. *Introductory Lectures on Convex Optimization: A Basic Course*. No. 87 in Applied Optimization, Kluwer Academic Publishers.

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

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

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

Perkins, Steven, David S. Leslie. 2014. Stochastic fictitious play with continuous action sets. *Journal of Economic Theory* **152** 179–213.

Perkins, Steven, Panayotis Mertikopoulos, David S. Leslie. 2017. Mixed-strategy learning with continuous action sets. *IEEE Trans. Autom. Control* **62**(1) 379–384.

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

Rosen, J. B. 1965. Existence and uniqueness of equilibrium points for concave N-person games. *Econometrica* **33**(3) 520–534.

Shamir, Ohad. 2013. On the complexity of bandit and derivative-free stochastic convex optimization. *COLT '13: Proceedings of the 26th Annual Conference on Learning Theory*.

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

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

Syrgkanis, Vasilis, Alekh Agarwal, Haipeng Luo, Robert E. Schapire. 2015. Fast convergence of regularized learning in games. *NIPS '15: Proceedings of the 29th International Conference on Neural Information Processing Systems*. 2989–2997.

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

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