2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Bandit Online Learning in Merely Coherent Games with Multi-Point Pseudo-Gradient Estimate

Yuanhanqing Huang[1] and Jianghai Hu[1]

Abstract— **Non-cooperative games serve as a powerful framework for capturing the interactions among self-interested players and have broad applicability in modeling a wide range of practical scenarios, ranging from power management to drug delivery. Although most existing solution algorithms assume the availability of first-order information or full knowledge of the objectives and others' action profiles, there are situations where the only accessible information at players' disposal is the realized objective function values. In this paper, we devise a bandit online learning algorithm for merely coherent games that integrates the optimistic mirror descent scheme and multi-point pseudo-gradient estimates. We further demonstrate that the generated actual sequence of play can converge a.s. to a critical point if the sequences of query radius and sample size are chosen properly, without resorting to extra Tikhonov regularization terms or additional norm conditions. Finally, we illustrate the validity of the proposed algorithm via a Rock-Paper-Scissors game and a least square estimation game.**

I. INTRODUCTION

Recent years have witnessed considerably increasing interest in the analysis of multi-agent systems and large-scale networks, which find a wide range of applications such as thermal load management of autonomous buildings [1], power management in sensor network [2], optimal drug delivery in the treatment of disease [3], control of environmental pollution [4], etc. One primary objective in multi-agent systems is to devise local protocols for each agent, by following which, the resulting group behavior is optimal as measured by a certain system-level metric [5]. With its origins in [6], game theory offers the theoretical tools to model and examine the strategic choices and associated outcomes of rational players who make decisions in a non-cooperative manner. In particular, in the Nash equilibrium problem (NEP), this group of players seeks to reach a stationary point known as Nash equilibrium (NE), where no rational player has any incentive to unilaterally deviate from it.

In order to devise an algorithm for the NEP or its variants, it is crucial to have access to the first-order information, i.e., the partial gradient of the local objective function of each player, the evaluation of which nevertheless usually requires the action profile from all players. In view of this, in some studies [7], [8], [9], the availability of first-order oracles is taken as a given, whereas some other studies [10], [11], [12] investigate network games where a communication network exists and players are willing to communicate with their trusted neighbors and keep local estimates of others' action profiles. Despite the notable progress discussed above, there are many real-world scenarios where players only have access to the observed objective values of selected actions, which makes the bandit/zeroth-order learning strategy a compelling choice. Our primary objective in this work is to develop an online learning algorithm for multi-player continuous games that possess mere coherence with bandit information.

Related Work: There have been several recent notable contributions to the field of bandit learning in games. In their work [13], Bravo et al. proposed a bandit version of mirror descent (MD), which guarantees a.s. convergence to an NE when the game is strictly monotone and achieves a convergence rate of $O(1/t^{1/3})$ for strongly monotone cases. Concerning the study of convergence rates in the realm of strongly monotone games or strongly variationally stable Nash equilibrium seeking, [14], [15], [16], [17] have succeeded in elevating the convergence rates from $O(1/t^{1/3})$ to $O(1/t^{1/2})$. Huang et al. [18] developed two bandit learning algorithms by integrating residual pseudo-gradient estimates into single-call extra-gradient schemes that ensure a.s. convergence to critical points of pseudo-monotone plus games. Moreover, in strongly pseudo-monotone plus games, by employing the proposed algorithms, the convergence rate is further elevated to $O(1/t^{1-\epsilon})$.

To extend the analysis beyond the realm of strictly monotone and pseudo-monotone plus games, Tatarenko et al. [19] utilized the single time-scale Tikhonov regularization and a doubly regularized approximate gradient descent strategy to develop an algorithm that converges to NEs in probability when the game is monotone and four decaying sequences are tuned properly. In a recent study [20], Gao et al. introduced an algorithm that integrates second-order learning dynamics and Tikhonov regularization and established the a.s. convergence of the sequence of play under the assumption that there exists at least one interior variationally stable state (VSS). Yet, the convergence is contingent on the norm condition that the ℓ_2-norm of the state sequence should be greater than that of the VSS, which can be challenging to verify during the iterative process.

In the literature of variational inequalities (VIs) and their stochastic versions (SVIs), Mertikopoulos et al. [21] showed that the vanilla MD converges when the problem is strictly coherent, a relaxed variant of strict monotonicity, but fails to converge in merely coherent VIs. In contrast, the extra-gradient (EG) method is capable of achieving convergence to a solution in all coherent VIs, but it requires the exact operator values. In the presence of random noise in

This work was supported by the National Science Foundation under Grant No. 2014816 and No.2038410.

[1]The authors are with the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 47907, USA {huan1282, jianghai}@purdue.edu

2023 62nd IEEE Conference on Decision and Control (CDC) | 979-8-3503-0124-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CDC49753.2023.10384240

operator values, strict coherence is necessary to establish the convergence of the EG iteration. Similar convergence analysis is also reported in [22] for pseudo-monotone plus SVIs. To address the challenges posed by random noise, Iusem et al. [23] developed an extra-gradient method for pseudo-monotone SVIs that incorporates an iterative variance reduction procedure and established both asymptotic convergence and non-asymptotic convergence rates for the proposed algorithm.

Contributions: In this work, we develop a bandit online learning algorithm and establish the a.s. convergence of the generated sequence of play under the regularity condition that the game is merely coherent, which is broader and more general than the games investigated in [9], [13], [14], [18]. The proposed algorithm leverages the optimistic mirror descent (OMD) [24], [25], a single-call extra-gradient scheme, as the backbone, which enables us to contend with the absence of strict coherence and reduces the query cost induced by the extra step. Alongside the OMD updates, the multi-point pseudo-gradient estimation is employed and the decaying rate of the variance of zeroth-order estimations can be controlled by properly tuning the query count per iteration. Furthermore, the validity of the proposed algorithm is verified through a Rock-Paper-Scissors game and a least square estimation game. All the proofs are included in [26] due to the page limit.

Basic Notations: For a set of vectors $\{v_i\}_{i\in\mathcal{S}}$, $[v_i]_{i\in\mathcal{S}}$ or $[v_1;\cdots;v_{|\mathcal{S}|}]$ denotes their vertical stack. For a vector v and a positive integer i, $[v]_i$ denotes the i-th entry of v. We let $\|\cdot\|$ denote the ℓ_2-norm and $\langle,\rangle$ represent the canonical dot product. Let $\mathrm{cl}(\mathcal{S})$ denote the closure of set $\mathcal{S}$, $\mathrm{int}(\mathcal{S})$ the interior, and $\partial\mathcal{S}$ the boundary.

II. SETUP AND PRELIMINARIES

A. Game Formulation

In a multi-player non-cooperative game $\mathcal{G}$ with the presence of N players, indexed by $\mathcal{N} \coloneqq \{1,\ldots,N\}$, each player $i \in \mathcal{N}$ aims to optimize its own local objective J^i by adjusting its action $x^i \in \mathcal{X}^i \subseteq \mathbb{R}^{n^i}$, which can be described as follows:

$$\text{minimize}_{x^i\in\mathcal{X}^i}\ J^i(x^i;x^{-i}), \tag{1}$$

where $x^{-i} \coloneqq [x^j]_{j\in\mathcal{N}_{-i}}$ denotes the stack action of other players that parameterizes the objective J^i with $\mathcal{N}_{-i} \coloneqq \mathcal{N}\backslash\{i\}$ and $x \coloneqq [x^j]_{j\in\mathcal{N}}$; $\mathcal{X}^i$ denotes the feasible set of player i, and for brevity, we let $\mathcal{X} \coloneqq \prod_{j\in\mathcal{N}}\mathcal{X}^j \subseteq \mathbb{R}^n$ represent the global strategy space and $\mathcal{X}^{-i} \coloneqq \prod_{j\in\mathcal{N}}\mathcal{X}^j \subseteq \mathbb{R}^{n^{-i}}$ with $n \coloneqq \sum_{j\in\mathcal{N}} n^j$ and $n^{-i} \coloneqq \sum_{j\in\mathcal{N}^{-i}} n^j$. Our blanket assumptions for the objective functions J^i's and the local feasible sets $\mathcal{X}^i$'s will be as follows:

Assumption 1: For each player i, the local objective function J^i is continuously differentiable in x over the global strategy space $\mathcal{X}$. Moreover, its individual strategy space $\mathcal{X}^i$ is compact and convex, and has a non-empty interior.

Given the smoothness posited in Assumption 1, a single-valued operator that we will leverage extensively throughout is the pseudo-gradient operator $F : \mathbb{R}^n \to \mathbb{R}^n$. It is defined as the concatenation of all the partial gradient operators, i.e.,

$$F : x \mapsto [\nabla_{x^i} J^i(x^i;x^{-i})]_{i\in\mathcal{N}}. \tag{2}$$

Before proceeding, we remark that Assumption 1 implicitly implies that F is Lipschitz continuous on $\mathcal{X}$ with some constant L, i.e., for any x and $x' \in \mathcal{X}$, we have

$$\|F(x) - F(x')\| \le L\|x - x'\|. \tag{3}$$

As for the solution concept, we focus on critical points (CPs) [27, Sec. 2.2], a more relaxed solution concept than Nash equilibria (NEs), whose definition is given as follows.

Definition 1: (Critical Points) A decision profile $x_* \in \mathcal{X}$ is a critical point of the game $\mathcal{G}$ if it is a solution to the associated (Stampacchia) variational inequality (VI), i.e.,

$$\langle F(x_*), x - x_*\rangle \ge 0,\ \forall x \in \mathcal{X}. \tag{4}$$

We postulate that the games discussed in this work admit at least one critical point inside $\mathcal{X}$. A well-known result is that CPs coincide with NEs when J^i is convex and continuously differentiable in x^i for all i [28, Sec. 1.4.2].

In this work, our aim is to propose a new algorithm that is applicable to a broader class of games as compared to strictly monotone games and pseudo-monotone plus games. Moreover, we intend to further relax pseudo-monotonicity assumptions that are usually imposed upon the structure of the game to the ones merely upon equilibria.

Assumption 2: (Mere Coherence) The game $\mathcal{G}$ is merely coherent if every critical point (CP) x_* of $\mathcal{G}$ is merely variationally stable, i.e., $\langle F(x), x - x_*\rangle \ge 0$ for all $x \in \mathcal{X}$.

Before we proceed, it is pertinent to make a few comments. Our analysis primarily lies within Euclidean space; however, we recognize the potential for extending its applicability to finite-dimensional Hilbert spaces. In addition, we employ mere coherence rather than pseudo-monotonicity as the standing assumption, as the former one is less restrictive. Recall that an operator F is pseudo-monotone if for all $x, y \in \mathcal{X}$, $\langle F(y), x - y\rangle \ge 0 \implies \langle F(x), x - y\rangle \ge 0$. Nonetheless, the latter is generally the more readily verifiable assumption in practical applications, since it does not needs the CPs x_*'s to be known a priori.

B. Optimistic Mirror Descent

In this subsection, we shall provide a brief overview of the optimistic mirror descent algorithm, as well as related concepts and results. As an extension of the Euclidean projection, the mirror map $\nabla\psi^* : \mathbb{R} \to \mathbb{R}$ is defined as:

$$\nabla\psi^*(z) = \mathrm{argmax}_{x\in\mathcal{X}}\{\langle z, x\rangle - \psi(x)\}, \tag{5}$$

where $\psi : \mathrm{dom}\,\psi \to \mathbb{R}$ is a so-called distance-generating function (DGF) with $\mathrm{dom}\,\psi$ denoting a convex and open set where ψ is well-defined. The DGF fulfills the following conditions [29, Sect. 4.1]: *(i)* ψ is differentiable and $\tilde{\mu}$-strongly convex for some $\tilde{\mu} > 0$; *(ii)* $\nabla\psi(\mathrm{dom}\,\psi) = \mathbb{R}^n$; *(iii)* $\mathrm{cl}(\mathrm{dom}\,\psi) \supseteq \mathcal{X}$ and $\lim_{x\to\partial(\mathrm{dom}\,\psi)} \|\nabla\psi(x)\|_* = +\infty$. The

definition of DGF ψ allows us to introduce a pseudo-distance called the Bregman divergence, which is defined as:

$$D(p,x) = \psi(p) - \psi(x) - \langle \nabla\psi(x), p - x\rangle, \forall p, x \in \operatorname{dom}\psi. \quad (6)$$

To let $D(p,\cdot)$ represent a certain distance measure to p and use this measure to define a neighborhood of p, we make the following assumption.

Assumption 3: (Bregman Reciprocity) The chosen DGF ψ satisfies that if the sequence $(x_k)_{k\in\mathbb{N}_+}$ converges to some point p, i.e., $\|x_k - p\| \to 0$, then $D(p,x_k) \to 0$.

Then, the Bregman divergence generates the prox-mapping $P_{x,\mathcal{X}} : \mathcal{H} \to \operatorname{dom}\psi \cap \mathcal{X}$ for some fixed $x \in \mathcal{X} \cap \operatorname{dom}\psi$ that plays a critical role in mirror descent and its variants:

$$P_{x,\mathcal{X}}(y) = \operatorname{argmin}_{x'\in\mathcal{X}}\{\langle y, x - x'\rangle + D(x',x)\}. \quad (7)$$

With all these in hand, the optimistic mirror descent (OMD) [24], [25] can be expressed as below:

$$\begin{aligned} X_{k+1/2} &= P_{X_k,\mathcal{X}}(-\tau_k F(X_{k-1/2})) \\ X_{k+1} &= P_{X_k,\mathcal{X}}(-\tau_k F(X_{k+1/2})), \end{aligned} \quad (8)$$

where $(\tau_k)_{k\in\mathbb{N}_+}$ denotes a proper sequence of step sizes. The update consists of the following two steps. Given the base state X_k at step k, in the look-forward step, the leading state $X_{k+1/2}$ is procured by updating X_k with the proxy $F(X_{k-1/2})$ queried at $X_{k-1/2}$ rather than the exact pseudo-gradient $F(X_k)$ queried at X_k to reduce the oracle call per iteration. This step is essential in anticipating the landscape of F and facilitating the convergence when F is merely monotone, i.e., $\langle F(x) - F(y), x - y\rangle \geq 0$, for all x and y feasible. In the state-updating step, the base state X_k is revised to X_{k+1} following the pseudo-gradient information $F(X_{k+1/2})$. The OMD falls into the single-call category, distinguishing itself from the conventional extra gradient algorithm [23] by exclusively utilizing the first-order information at $X_{k+1/2}$, without requiring information from both X_k and $X_{k+1/2}$.

III. MULTI-POINT PSEUDO-GRADIENT ESTIMATION

In this paper, we examine the scenario where the first-order information at the leading state, i.e., $F(X_{k+1/2})$ is not readily available, and players need to estimate them based on the realized objective function values. A prevalent technique in the literature of first-order information estimation methods is the simultaneous perturbation stochastic approximation (SPSA) approach [13]. For each $i \in \mathcal{N}$, let $\mathbb{B}_i, \mathbb{S}_i \subseteq \mathbb{R}^{n^i}$ denote the unit ball and the unit sphere centered at the origin. At each iteration k, before implementing the SPSA estimate, we initially undertake the following perturbation step:

$$\hat{X}^i_{k+1/2} = (1 - \frac{\delta_k}{r^i})X^i_{k+1/2} + \frac{\delta_k}{r^i}(p^i + r^i u^i_k) = \bar{X}^i_{k+1/2} + \delta_k u^i_k, \quad (9)$$

where u^i_k is randomly sampled from $\mathbb{S}_i \subseteq \mathbb{R}^{n^i}$ and we define $u_k := [u^i_k]_{i\in\mathcal{N}}$; δ_k represents the random query radius at iteration k; $\mathbb{B}(p^i, r^i) \subseteq \mathcal{X}^i$ is an arbitrary fixed Euclidean ball within the feasible set $\mathcal{X}^i$ that centers at p^i with radius r^i; $\bar{X}^i_{k+1/2} := (1 - \delta_k/r^i)X^i_{k+1/2} + (\delta_k/r^i)p^i$. Denote $\bar{X}_{k+1/2} := [\bar{X}^i_{k+1/2}]_{i\in\mathcal{N}}$. In the merit of the feasibility adjustment in (9), the action to be taken will sit within the feasible set, i.e., $\hat{X}^i_{k+1/2} \in \mathcal{X}^i$ and $\hat{X}_{k+1/2} := [\hat{X}^i_{k+1/2}]_{i\in\mathcal{N}} \in \mathcal{X}$. With this in hand, the SPSA estimation can be expressed as $\frac{n^i}{\delta_k} J^i(\hat{X}_{k+1/2})u^i_k$. Nevertheless, as previously noted in [13], the SPSA approach incurs a larger estimation variance with a decrease in query radius aimed at improving estimation accuracy, which results in conservative choices of updating step sizes τ_k and significant degradation of the convergence rate. To resolve this conundrum, there has been increased consideration given to schemes such as two-point estimation and residual estimation to keep the variance bounded. On account of this, we consider the multi-point pseudo-gradient estimation (MPG) scheme, the counterparts of which in the field of optimization can be found in [30]. At every iteration k, each player i executes the perturbation step in (9) $(T_k + 1)$ times in an independent manner, takes the action $\hat{X}^i_{k+1/2,t}$, and observes the associated realized objective function values $J^i(\hat{X}_{k+1/2,t})$, where the variable $t \in \mathbb{N}$ is an index of the multiple samples taken per iteration. The multi-point pseudo-gradient estimate can be formulated as below:

$$G^i_k := \frac{n^i}{\delta_k T_k}\sum_{t=1}^{T_k}\big(J^i(\hat{X}_{k+1/2,t}) - J^i(\hat{X}_{k+1/2,0})\big)u^i_{k,t}, \quad \text{(MPG)}$$

where $(u^i_{k,t})_{t=0,\ldots,T_k}$ are i.i.d. random variables uniformly distributed over $\mathbb{S}_i$; the action taken by player i is given by $\hat{X}^i_{k+1/2,t} := (1 - \frac{\delta_k}{r^i})X^i_{k+1/2} + \frac{\delta_k}{r^i}(p^i + r^i u^i_{k,t}) = \bar{X}^i_{k+1/2} + \delta_k u^i_{k,t}$; $\hat{X}_{k+1/2,t} := [\hat{X}^i_{k+1/2,t}]_{i\in\mathcal{N}}$. To simplify the presentation, we will henceforth use $\hat{J}^i_{k,t}$ to represent the realized objective value $J^i(\hat{X}_{k+1/2,t})$ for the t-th sample at iteration k. Prior to delving into the properties of MPG, we first outline the probability setup to streamline our later discussion. Let $(\Omega, \mathcal{F}, \mathcal{P})$ denote the underlying probability space. The filtration $(\mathcal{F}_k)_{k\in\mathbb{N}_+}$ is constructed as $\mathcal{F}_k := \sigma\{X_0, \{u_{1,t}\}_{t=0}^{T_1}, \ldots, \{u_{k-1,t}\}_{t=0}^{T_{k-1}}\}$, which captures the update that results in X_k, i.e., the entire information up to and including iteration $k-1$. . Then to characterize MPG, we start by considering the following decomposition of it:

$$\begin{aligned} G^i_k =& \nabla_{x^i} J^i(X_{k+1/2}) + \big(G^i_k - \mathbb{E}[G^i_k \mid \mathcal{F}_k]\big) \\ &+ \big(\mathbb{E}[G^i_k \mid \mathcal{F}_k] - \nabla_{x^i} J^i(X_{k+1/2})\big). \end{aligned}$$

For brevity, we let $B^i_k := \mathbb{E}[G^i_k \mid \mathcal{F}_k] - \nabla_{x^i} J^i(X_{k+1/2})$ represent the systematic error and $V^i_k := G^i_k - \mathbb{E}[G^i_k \mid \mathcal{F}_k]$ the stochastic error. To facilitate later analysis, for each J^i, we introduce the δ-smoothed objective function $\tilde{J}^i_\delta$:

$$\tilde{J}^i_\delta(x^i; x^{-i}) := \frac{1}{\mathbb{V}^i_\delta}\int_{\delta\mathbb{S}_{-i}}\int_{\delta\mathbb{B}_i} J^i(x^i + \tilde{\tau}^i; x^{-i} + \tau^{-i})d\tilde{\tau}^i d\tau^{-i}, \quad (10)$$

where $\mathbb{S}_{-i} := \prod_{j\in\mathcal{N}^{-i}}\mathbb{S}_j \subseteq \mathbb{R}^{n^{-i}}$; $\mathbb{V}^i_\delta := \operatorname{vol}(\delta\mathbb{B}_i)\cdot\operatorname{vol}(\delta\mathbb{S}_{-i})$. The lemmas presented below provide an examination of the properties of B^i_k and V^i_k, which will be later employed in the proof of the main theorem.

Lemma 1: Suppose that Assumption 1 holds. Then at each iteration k, the conditional expectation satisfies $\mathbb{E}[G^i_k \mid \mathcal{F}_k] = \nabla_{x^i}\tilde{J}^i_{\delta_k}(\bar{X}_{k+1/2})$ a.s. for every $i \in \mathcal{N}$. Moreover the systematic error $B_k := [B^i_k]_{i\in\mathcal{N}}$ possesses a decaying upper bound $\|B_k\| \leq \alpha_B\delta_k$ for some positive constant α_B.

Proof: See [26, Appendix A]. ∎

In contrast to the single-point or two-point estimates, the advantage of utilizing MPG is primarily demonstrated in the following lemma, which measures the decaying rate of the stochastic error w.r.t. the number of samples.

Lemma 2: Suppose that Assumption 1 holds. Then at each iteration k, the squared norm of $V_k \coloneqq [V_k^i]_{i\in\mathcal{N}}$ satisfies $\mathbb{E}[\|V_k\|^2 \mid \mathcal{F}_k] \le \alpha_V/T_k$ for some positive constant α_V.

Proof: See [26, Appendix B]. ∎

IV. A VARIANCE-REDUCTION LEARNING ALGORITHM AND CONVERGENCE ANALYSIS

In view of the convergence properties of OMD introduced in Sec. II-B, we design a zeroth-order algorithm for merely monotone games by incorporating MPG into OMD, the precision of which can be controlled by adjusting the sample size per iteration. Each player of the group possesses their own local $\tilde{\mu}^i$-strongly convex DGF, denoted by ψ^i. Additionally, the function $\psi(x) \coloneqq \sum_{i\in\mathcal{N}} \psi^i(x^i)$ with $x \coloneqq [x^i]_{i\in\mathcal{N}}$ represents the group DGF, which is $\tilde{\mu}$-strongly convex. The proposed approach is outlined in Algorithm 1.

Algorithm 1 Zeroth-Order Variance-Reduced Learning of CPs Based on Optimistic Mirror Descent (Player i)

Initialize: $X_0^i = X_{1/2}^i = X_1^i \in \mathcal{X}^i \cap \operatorname{dom}\psi^i$ arbitrarily; $G_0^i = \mathbf{0}_{n^i}$; p^i, r^i to be the center and radius of an arbitrary ball within the set $\mathcal{X}^i$
procedure AT THE k-TH ITERATION ($k \in \mathbb{N}_+$)
  $X_{k+1/2}^i \leftarrow P_{X_k^i,\mathcal{X}^i}(-\tau G_{k-1}^i)$
  for $t = 0, \dots, T_k$ **do**
    Randomly sample the direction $u_{k,t}^i$ from $\mathbb{S}_i$
    $\hat{X}_{k+1/2,t}^i \leftarrow (1 - \frac{\delta_k}{r^i})X_{k+1/2,t}^i + \frac{\delta_k}{r^i}(p^i + r^i u_{k,t}^i)$
    Take action $\hat{X}_{k+1/2,t}^i$
    Observe the realized objective function value $\hat{J}_{k,t}^i \coloneqq J^i(\hat{X}_{k+1/2,t}^i; \hat{X}_{k+1/2,t}^{-i})$
  end for
  $G_k^i \leftarrow \frac{n^i}{\delta_k T_k}\sum_{t=1}^{T_k}(\hat{J}_{k,t}^i - \hat{J}_{k,0}^i)u_{k,t}^i = \frac{1}{T_k}\sum_{t=1}^{T_k} G_{k,t}^i$
  $X_{k+1}^i \leftarrow P_{X_k^i,\mathcal{X}^i}(-\tau G_k^i)$
end procedure
Return: $\{\hat{X}_{k+1/2}^i\}_{i\in\mathcal{N}}$

The Robbins-Siegmund (R-S) theorem serves as a heavy-lifting tool in the field of stochastic optimization to examine the convergence of sequences. Its formal statement is presented as follows.

Lemma 3: ([31, Thm. 1]) Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and $(\mathcal{F}_k)_k$ a filtration of $\mathcal{F}$. For each $k = 1, 2, \dots$, Z_k, β_k, ξ_k, and ζ_k are non-negative $\mathcal{F}_k$-measurable random variables that satisfy $\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] \le (1+\beta_k)Z_k + \xi_k - \zeta_k$. If $\sum_{k\in\mathbb{N}_+}\beta_k < \infty$ a.s. and $\sum_{k\in\mathbb{N}_+}\xi_k < \infty$ a.s., then $\lim_{k\to\infty} Z_k$ exists and is finite a.s. and $\sum_{k\in\mathbb{N}_+}\zeta_k < \infty$ a.s.

To employ the theorem, it is necessary to guarantee that $\sum_{k\in\mathbb{N}_+}\xi_k$ is finite a.s. Recall from Lemma 2, in the variance reduction scenario, the decaying upper bound is constructed for $\mathbb{E}[\|V_k\|^2 \mid \mathcal{F}_k]$ rather than the random variable $\|V_k\|^2$. In the meantime, unlike the typical extra-gradient method, OMD leverages the pseudo-gradient $F(X_{k-1/2})$ from the last iteration when updating to the leading state $X_{k+1/2}$. This approximation brings the stochastic error $\|V_{k-1}\|^2$ into the recurrent inequality which, due to the absence of the averaging effect, does not possess a decaying upper bound and prevents us from applying the R-S theorem. Motivated by the consideration above, our next step will be establishing a variant of the R-S theorem by relaxing the condition imposed upon the sequence $(\xi_k)_{k\in\mathbb{N}_+}$.

Theorem 1: Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and $(\mathcal{F}_k)_k$ a filtration of $\mathcal{F}$. For each $k = 1, 2, \dots$, Z_k, ξ_k, and ζ_k are non-negative $\mathcal{F}_k$-measurable random variables that satisfy $\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] \le Z_k + \xi_k - \zeta_k$ with $\mathbb{E}[Z_1] < \infty$. If $\sum_{k\in\mathbb{N}_+}\mathbb{E}[\xi_k] < \infty$, then Z_k converges a.s. to some random variable Z_∞ with $\mathbb{E}[Z_\infty] < \infty$ and $\sum_{k\in\mathbb{N}_+}\zeta_k < \infty$ a.s.

Proof: See [26, Appendix C]. ∎

With this conclusion available, we can establish the following results about the convergence of Algorithm 1 and the sufficient conditions to guarantee it.

Theorem 2: Consider a multi-player game $\mathcal{G}$. Suppose that Assumptions 1 to 3 hold. In addition, the sequence of query radius $(\delta_k)_{k\in\mathbb{N}_+}$ and the sequence of the reciprocal of sample size $(1/T_k)_{k\in\mathbb{N}_+}$ are monotonically decreasing and satisfy

$$\sum_{k\in\mathbb{N}_+}\delta_k < \infty, \quad \sum_{k\in\mathbb{N}_+} 1/T_k < \infty. \tag{11}$$

The step size τ satisfies $(\tau L/\tilde{\mu})^2 \le 1/12$. Then the base state $(X_k)_{k\in\mathbb{N}_+}$ as well as the leading state $(X_{k+1/2})_{k\in\mathbb{N}_+}$ converge a.s. to a CP x_* of $\mathcal{G}$. Moreover, the actual sequence of play also satisfy $\lim_{k\to\infty}\hat{X}_{k+1/2,t} = x_*$ a.s., for arbitrary sample t.

Proof: See [26, Appendix D]. ∎

V. NUMERICAL EXPERIMENTS

A. The Rock-Paper-Scissors (RPS) Game

Consider the zero-sum rock-paper-scissors game between two players. The payoff matrices A^a and A^b of player a and b are set respectively as

$$A^a \coloneqq \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix} \text{ and } A^b \coloneqq \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix} = -A^a,$$

which further give rise to the objective functions: $J^a(x^a; x^b) = -(x^a)^T A^a x^b$ and $J^b(x^b; x^a) = -(x^a)^T A^b x^b$. The associated strategy spaces are the probability simplices, i.e., $\mathcal{X}^a = \mathcal{X}^b \coloneqq \{x \in \mathbb{R}^3 \mid 0 \le x \le 1, \mathbf{1}^T x = 1\}$. The RPS game is merely monotone and admits a unique CP/NE at $[1/3; 1/3; 1/3]$ for both players. To fulfill the requirement about the non-empty interior in Assumption 1, taking player a as an example, we can employ a simple coordinate transformation $\varphi : \mathbb{R}^2 \to \mathbb{R}^3$ with $\varphi : [y_1; y_2] \mapsto [y_1; y_2; 1 - y_1 - y_2] = [x_1; x_2; x_3]$ and $\varphi^{-1}(\mathcal{X}^a) = \tilde{\mathcal{X}}^a = \{y \in \mathbb{R}^2 \mid 0 \le y \le 1, \mathbf{1}^T y \le 1\}$. Then MPG is applied to obtain a pseudo-gradient estimate $\tilde{G}_k \in \mathbb{R}^2$, and we use another map $\phi : \mathbb{R}^2 \to \mathbb{R}^3$ to pull the pseudo-gradient from y-coordinate system back to x-coordinate system. The map ϕ is defined as $\phi : [\tilde{g}_1; \tilde{g}_2] \mapsto [2/3\tilde{g}_1 - 1/3\tilde{g}_2; -1/3\tilde{g}_1 + 2/3\tilde{g}_2; -1/3\tilde{g}_1 - 1/3\tilde{g}_2]$, which is derived from the observation

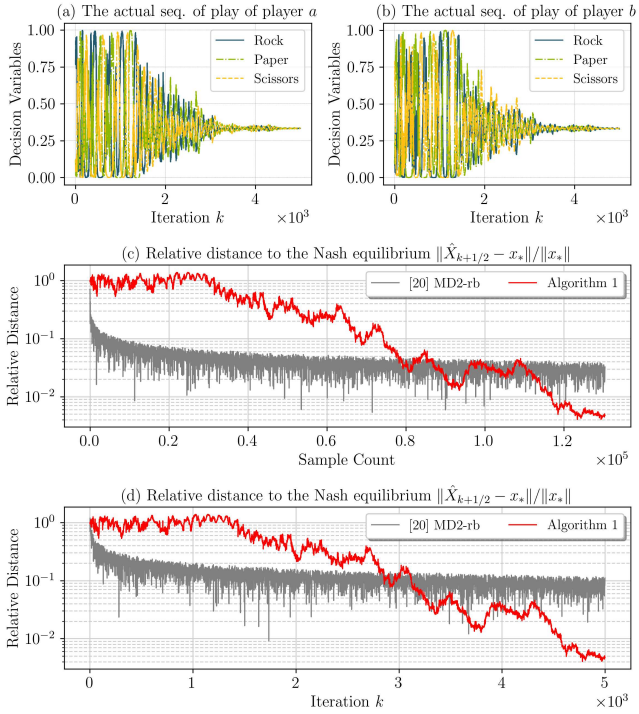

Fig. 1: Performance of Algorithm 1 in the RPS Game

that

$$\tilde{g}_i = \sum_{j=1,2,3} \frac{\partial J^a}{\partial x_j} \cdot \frac{\partial x_j}{\partial y_i} = \sum_{j=1,2,3} g_j \cdot \frac{\partial x_j}{\partial y_i} \text{ and } \sum_{j=1,2,3} g_j = 0.$$

A similar procedure can be applied to player b to guarantee the fulfillment of the assumption.

In the numerical simulation, we choose $\tau = 0.1$, the decaying query radius $\delta_k = 0.1\times(k+20)^{-1.1}$, and the increasing number of queries per iteration $T_k = \lceil 10^{-3} \times k^{1.1} + 20 \rceil$. Since the negative entropy $h(x) = \sum_{i=1,2,3}[x]_i \log[x]_i$ is 1-strongly convex in $\|\cdot\|$ and satisfies all the requirements discussed in Sec. II-B, it can be chosen as a DGF for player a and b. The simulation results are illustrated in Fig. 1, with Fig. 1 (a) and (b) visualizing the actual sequences of play of player a and b. To compare with [20] (MD2-rb), Fig. 1 (c) and (d) illustrate the relative distance $\|\hat{X}_{k+1/2} - x_*\|/\|x_*\|$ to the CP/NE x_*, where the x-axis denotes the sample count and iteration index, respectively. The selection of parameters for [20] (MD2-rb) adheres to the specifications provided in its corresponding section. As depicted in the figure, [20] (MD2-rb) displays a faster decline in the early iterations, whereas Algorithm 1 achieves a superior convergence rate as the progress advances.

B. Least Square Estimation in Linear Models

In this numerical experiment, we convert the linear regression to a zero-sum bilinear game between two players [32, Sec. VI]. Given a set of data samples $\{(z_j, y_j)\}_{j=1}^M$ where $z_j \in \mathbb{R}^N$ and $y_j \in \mathbb{R}$ represent the input feature vector and the output scalar, respectively. In addition, $y_j = w_0 + w^T z_j + \xi_j$, with $\tilde{w} := [w_0; w] \in [-\bar{w}, \bar{w}]^{N+1} \subseteq \mathbb{R}^{N+1}$ denoting the parameters to be determined and ξ_j some random noise. Here, the region $[-\bar{w}, \bar{w}]^{N+1}$ with $\bar{w} \in \mathbb{R}^+$ is enforced to ensure the strategy space is bounded. For brevity, denote $\tilde{z}_j := [1; z_j]$, $\tilde{Z} := [\tilde{z}_1, \ldots, \tilde{z}_M]$ and $y = [y_1; \cdots; y_M]$. We can then formulate this least square estimation problem as:

$$\underset{\tilde{w}\in[-\bar{w},\bar{w}]^{N+1}}{\text{minimize}} \ \frac{1}{2}\|\tilde{Z}^T\tilde{w} - y\|_2^2. \tag{12}$$

To convert it into a two-player game, we leverage an auxiliary variable $\lambda \in \mathbb{R}^M$ and the fact that $\frac{1}{2}\|\tilde{Z}^T\tilde{w} - y\|_2^2 = \max_{\lambda\in\mathbb{R}^M} \lambda^T(\tilde{Z}^T\tilde{w} - y) - \frac{1}{2}\|\lambda\|_2^2 = \max_{\lambda\in\mathbb{R}^M} J(\tilde{w}, \lambda)$. Taking the boundedness of $\tilde{w}$ into account, it can be asserted that there exists a bounded set $[-\bar{\lambda}, \bar{\lambda}]^M$ such that the solution λ to the maximization problem above satisfies $\lambda \in [-\bar{\lambda}, \bar{\lambda}]^M$. As such, let $J^1(x^1; x^2) = J(x^1, x^2)$ and $J^2(x^2; x^1) = -J(x^1, x^2)$, and this game can be formulated as follows:

$$\text{Player 1: } \underset{-\bar{w}\le x^1\le\bar{w}}{\text{minimize}} \ J^1(x^1; x^2), \ \text{Player 2: } \underset{-\bar{\lambda}\le x^2\le\bar{\lambda}}{\text{minimize}} \ J^2(x^2; x^1).$$

For the verification of the remaining assumptions, showing the uniqueness of the CP, and other detailed discussions, we refer the interested reader to [18, Sec. V-B][32, Sec. VI].

When implementing the experiments, we choose $N = 5$, $M = 20$, and $\bar{w} = \bar{\lambda} = 5$. Then random noise ξ_i is uniformly distributed over the interval $[-0.6, 0.6]$. We compare different sets of the sequences of query radius δ_k and query samples per iteration T_k. In Fig. 2 (a), the original curve to fit, the noisy data samples used, the optimal solution that can be procured from the existing data, and one OMD solution generated by Algorithm 1 are illustrated. Comparing the results with different choices of δ_k, we note that for this problem when δ_k decays comparable to or faster than $O(k^{-1.1})$, further increasing the decaying rate contributes little to speed up the convergence rate of the sequence. As for the influence of different T_k, when T_k is a small constant, the generated sequences will diverge; when $T_k = 10$ increases to some sufficiently large constant $T_k = 15$, the associated sequences demonstrate the trend of convergence to some ϵ-neighborhood of the CP; when T_k decays no slower than $O(k^{-1.1})$, as reflected in Fig. 2 (c) and (e), the fluctuations of the relative step sizes are mitigated; yet little difference can be observed regarding the relative distance to the CP, as shown in Fig. 2 (b) and (d).

VI. CONCLUSION

In this work, we investigate bandit learning in multi-player continuous games with an emphasis on handling merely coherent cases. A new learning algorithm is proposed, by integrating the idea of optimistic mirror descent and multi-point pseudo-gradient estimation. Under the assumptions posited and the conditions that the sequences of query radius δ_k and the reciprocal of sample size T_k are absolutely summable, the actual sequence of play generated by the proposed algorithm is shown to converge a.s. to a CP of the game. There are several potential directions for future exploration. The first one is relaxing the requirements for the number of samples per iteration T_k, since the superlinear growth of T_k may prevent the application of the proposed algorithm when the

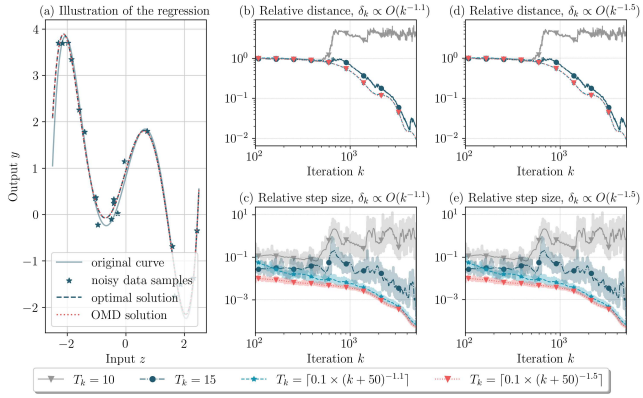

Fig. 2: Performance of Algorithm 1 in the Least Square Estimation: in Panel (a), the optimal solution is obtained by solving (12) analytically; the OMD solution corresponds to the case when $\delta_k = O(k^{-1.1})$ and $T_k = \lceil 0.1 \times (k+50)^{-1.1} \rceil$; Panel (b) and (d) visualize the relative distance to the unique CP, i.e., the metric is given by $\|\hat{X}_{k+1/2} - x_*\|_2 / \|x_*\|_2$; Panel (c) and (e) report the relative updating step sizes per iteration, i.e., $\|\hat{X}_{k+1/2} - \hat{X}_{k-1/2}\|_2 / \|\hat{X}_{k-1/2}\|_2$. The rolling averages with a window size of 100 and the original fluctuations are illustrated in solid curves and semi-transparent curves, respectively.

bandit feedback is inadequate. Furthermore, when it comes to a large-scale player network, the asynchronicity of the updates is a prevalent issue and the cost of synchronization is prohibitive, which is further exacerbated by the multi-point scheme considered. We intend to address these questions in future work.

References

[1] Z. Jiang and J. Cai, "Game theoretic control of thermal loads in demand response aggregators," in *2021 American Control Conference (ACC)*. IEEE, 2021, pp. 4141–4147.

[2] E. Campos-Nanez, A. Garcia, and C. Li, "A game-theoretic approach to efficient power management in sensor networks," *Operations Research*, vol. 56, no. 3, pp. 552–561, 2008.

[3] Y. Wu, M. Zhang, J. Wu, X. Zhao, and L. Xia, "Evolutionary game theoretic strategy for optimal drug delivery to influence selection pressure in treatment of hiv-1," *Journal of mathematical biology*, vol. 64, pp. 495–512, 2012.

[4] S. Du, F. Ma, Z. Fu, L. Zhu, and J. Zhang, "Game-theoretic analysis for an emission-dependent supply chain in a 'cap-and-trade'system," *Annals of Operations Research*, vol. 228, pp. 135–149, 2015.

[5] N. Li and J. R. Marden, "Designing games for distributed optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 2, pp. 230–242, 2013.

[6] J. F. Nash Jr, "Equilibrium points in n-person games," *Proceedings of the national academy of sciences*, vol. 36, no. 1, pp. 48–49, 1950.

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

[8] P. Yi and L. Pavel, "An operator splitting approach for distributed generalized Nash equilibria computation," *Automatica*, vol. 102, pp. 111–121, 2019.

[9] T. Tatarenko, W. Shi, and A. Nedić, "Geometric convergence of gradient play algorithms for distributed nash equilibrium seeking," *IEEE Transactions on Automatic Control*, vol. 66, no. 11, pp. 5342–5353, 2020.

[10] L. Pavel, "Distributed GNE seeking under partial-decision information over networks via a doubly-augmented operator splitting approach," *IEEE Transactions on Automatic Control*, vol. 65, no. 4, pp. 1584–1597, 2019.

[11] M. Bianchi, G. Belgioioso, and S. Grammatico, "Fast generalized Nash equilibrium seeking under partial-decision information," *Automatica*, vol. 136, p. 110080, 2022.

[12] Y. Huang and J. Hu, "Distributed computation of stochastic GNE with partial information: An augmented best-response approach," *IEEE Transactions on Control of Network Systems*, vol. 10, no. 2, pp. 947–959, 2023.

[13] M. Bravo, D. Leslie, and P. Mertikopoulos, "Bandit learning in concave N-person games," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[14] T. Lin, Z. Zhou, W. Ba, and J. Zhang, "Optimal no-regret learning in strongly monotone games with bandit feedback," *arXiv preprint arXiv:2112.02856*, 2021.

[15] T. Tatarenko and M. Kamgarpour, "On the rate of convergence of payoff-based algorithms to Nash equilibrium in strongly monotone games," *arXiv preprint arXiv:2202.11147*, 2022.

[16] ——, "Convergence rate of learning a strongly variationally stable equilibrium," *arXiv preprint arXiv:2304.02355*, 2023.

[17] D. Drusvyatskiy, M. Fazel, and L. J. Ratliff, "Improved rates for derivative-free gradient play in strongly monotone games," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 3403–3408.

[18] Y. Huang and J. Hu, "Zeroth-order learning in continuous games via residual pseudogradient estimates," *arXiv preprint arXiv:2301.02279*, 2023.

[19] T. Tatarenko and M. Kamgarpour, "Bandit online learning of Nash equilibria in monotone games," *arXiv preprint arXiv:2009.04258*, 2020.

[20] B. Gao and L. Pavel, "Bandit learning with regularized second-order mirror descent," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 5731–5738.

[21] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras, "Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile," in *International Conference on Learning Representations*, 2019.

[22] A. Kannan and U. V. Shanbhag, "Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants," *Computational Optimization and Applications*, vol. 74, no. 3, pp. 779–820, 2019.

[23] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson, "Extragradient method with variance reduction for stochastic variational inequalities," *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 686–724, 2017.

[24] W. Azizian, F. Iutzeler, J. Malick, and P. Mertikopoulos, "The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities," in *Conference on Learning Theory*. PMLR, 2021, pp. 326–358.

[25] Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos, "On the convergence of single-call stochastic extra-gradient methods," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[26] Y. Huang and J. Hu, "Bandit online learning in merely coherent games with multi-point pseudo-gradient estimate," *arXiv preprint arXiv:2303.16430*, 2023.

[27] P. Mertikopoulos, Y.-P. Hsieh, and V. Cevher, "Learning in games from a stochastic approximation viewpoint," *arXiv preprint arXiv:2206.03922*, 2022.

[28] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

[29] S. Bubeck, "Theory of convex optimization for machine learning," *arXiv preprint arXiv:1405.4980*, vol. 15, 2014.

[30] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, "Optimal rates for zero-order convex optimization: The power of two function evaluations," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.

[31] H. Robbins and D. Siegmund, "A convergence theorem for non-negative almost supermartingales and some applications," in *Optimizing methods in statistics*. Elsevier, 1971, pp. 233–257.

[32] B. Gao and L. Pavel, "Continuous-time discounted mirror descent dynamics in monotone concave games," *IEEE Transactions on Automatic Control*, vol. 66, no. 11, pp. 5451–5458, 2020.