IEEE CSS

Received 1 May 2023; revised 27 July 2023; accepted 29 August 2023. Date of publication 15 September 2023; date of current version 20 October 2023. Recommended by Senior Editor Prof. Lacra Pavel.

Digital Object Identifier 10.1109/OJCSYS.2023.3316071

Global and Local Convergence Analysis of a Bandit Learning Algorithm in Merely Coherent Games

YUANHANQING HUANG (Graduate Student Member, IEEE), AND JIANGHAI HU

(Intersection of Machine Learning with Control)

School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA

CORRESPONDING AUTHOR: YUANHANQING HUANG (e-mail: huan1282@purdue.edu).

This work was supported by the National Science Foundation under Grants 2014816 and 2038410.

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 path planning of self-driving vehicles. 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 article, we devise a bandit online learning algorithm that integrates the optimistic mirror descent scheme and multi-point pseudo-gradient estimates. We further prove that the generated actual sequence of play converges a.s. to a critical point if the game under study is globally merely coherent, without resorting to extra Tikhonov regularization terms or additional norm conditions. We also discuss the convergence properties of the proposed bandit learning algorithm in locally merely coherent games. Finally, we illustrate the validity of the proposed algorithm via two two-player minimax problems and a cognitive radio bandwidth allocation game.

INDEX TERMS Game theory, learning theory, optimization under uncertainties, stochastic systems.

I. INTRODUCTION

Recent years have witnessed 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], and path planning and control of self-driving cars [3], with prospects for further applicability in optimal drug delivery in the treatment of diseases [4] and control of environmental pollution [5], 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 [6]. With its origins in [7], 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 profiles from all players. In view of this, in some studies [8], [9], [10], the availability of first-order oracles is taken as a given, whereas some other studies [11], [12], [13] investigate scenarios 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 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 are globally or locally merely coherent with bandit information.

Related Work: There have been several recent notable contributions to the field of bandit learning in games. In their work [14], 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. By employing a barrier-based method, Lin et al. [15] improved the convergence rate for strongly monotone games from $O(1/t^{1/3})$ to $O(1/t^{1/2})$. Similar convergence rates have also been reported in [16], [17], [18]. Huang et al. [19] 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. [20] 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 [21], 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 on variational inequalities (VIs) and their stochastic versions (SVIs), Mertikopoulos et al. [22] 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 operator values, strict coherence is necessary to establish the convergence of the EG iteration. Similar convergence analysis is also reported in [23] for pseudo-monotone plus SVIs. To address the challenges posed by random noise, Iusem et al. [24] developed an extra-gradient method for pseudo-monotone SVIs that incorporates an iterative variance reduction procedure and established both asymptotic convergence and convergence rates in terms of the residual function for the proposed algorithm.

In the realm of multi-player games without global monotonicity or coherence, there exists a body of research that delves into games satisfying the weak Minty variational inequality or negative comonotonicity: Pethick et al. [25] and Cai et al. [26], [27] devised algorithm that ensure convergence under the deterministic setting; Diakonikolas et al. [28] proposed a generalization of the extra-gradient method that ensures convergence to a stationary point for unconstrained problems; Pethick et al. [29] extended their previous work to stochastic cases and designed algorithms that converge to solutions for constrained problems for a random iterate. Another significant body of research has focused on local solutions when global regularity conditions are absent: Mertikopoulos and Zhou [8] investigated the local convergence properties of mirror descent in deterministic and stochastic cases; Hsieh et al. [30] focused on a class of single-call extra-gradient methods in Euclidean space and established local geometric convergence results for deterministic cases and a local convergence rate of $O(1/t)$ for stochastic cases. These local convergence rate results are later generalized by Azizian et al. [31] to Banach spaces over a range of Legendre exponents.

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 [14], [15], [16], [17], [18]. The proposed algorithm leverages the optimistic mirror descent (OMD) [30], [31] and a single-call extra-gradient scheme as the backbone, which allows us to deal 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. In contrast to [21], despite the requirement in our approach that every solution is globally merely variationally stable, we avoid enforcing the additional norm condition in [21, Thm. 1]. Additionally, we investigate games with only local mere coherence and establish that, by utilizing appropriate initializations, the generated actual sequences of play can converge to critical points (CPs) with sufficiently high probability. Furthermore, the validity of the proposed algorithm is verified through two two-player minimax problems and a cognitive radio bandwidth allocation game.

Organization: In Section II, we provide a formal formulation of the multi-player games and briefly introduce optimistic mirror descent. Section III presents the multi-point pseudo-gradient estimate and offers insights into the associated systematic and stochastic errors. Subsequently, in Section IV, we present the proposed algorithm and provide the main convergence results in globally merely coherent games. Section V is dedicated to the examination of local convergence for the proposed learning algorithm. In Section VI, to demonstrate the theoretical findings and the effectiveness of the proposed algorithm, we conduct simulations for two-player zero-sum games and the cognitive radio bandwidth allocation game. Section VII concludes the article and highlights potential extensions and applications.

Basic Notations: Let $\mathbb{R}_{++} := (0, +\infty)$ and $\mathbb{N}_+ := \mathbb{N}\backslash\{0\}$. For a set of vectors $\{v_i\}_{i\in S}$, $[v_i]_{i\in S}$ or $[v_1; \cdots; v_{|S|}]$ 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 $\langle\cdot,\cdot\rangle$ 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. The symbols $a\wedge b$ and $a\vee b$ stand for the lesser and the greater of the two real numbers a and b, respectively.

A conference version of this article can be found in [32], which mainly focuses on the convergence analysis under the global mere coherence assumption.

II. SETUP AND PRELIMINARIES

A. GAME FORMULATION

In a multi-player non-cooperative game $\mathcal{G}$ with N players, indexed by $\mathcal{N}:=\{1,\dots,N\}$, 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:

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

where $x^{-i}:=[x^j]_{j\in\mathcal{N}_{-i}}$ denotes the stack action of other players that parameterizes the objective J^i with $\mathcal{N}_{-i}:=\mathcal{N}\backslash\{i\}$ and $x:=[x^j]_{j\in\mathcal{N}}$; $\mathcal{X}^i$ denotes the feasible set of player i, and for brevity, we let $\mathcal{X}:=\prod_{j\in\mathcal{N}}\mathcal{X}^j\subseteq\mathbb{R}^n$ represent the global strategy space and $\mathcal{X}^{-i}:=\prod_{j\in\mathcal{N}}\mathcal{X}^j\subseteq\mathbb{R}^{n^{-i}}$ with $n:=\sum_{j\in\mathcal{N}}n^j$ and $n^{-i}:=\sum_{j\in\mathcal{N}^{-i}}n^j$. Our analysis primarily lies within Euclidean space; however, it has the potential to be extended to finite-dimensional Hilbert spaces. 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) [33, 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.,

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

where $\langle\cdot,\cdot\rangle$ represents the canonical inner product.

CPs are the fixed points of the "linearized" best-response iterate $x\mapsto\mathrm{argmin}_{x'\in\mathcal{X}}\langle F(x),x'\rangle$ and can be perceived as local NEs [33]. CPs form a superset of NEs and coincide with them when J^i is convex and continuously differentiable in x^i for all i [34, Sec. 1.4.2]. We postulate that the games discussed in this work admit at least one CP inside $\mathcal{X}$.

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. Two assumptions are employed in Sections IV and V to facilitate the analysis of global and local convergence, respectively.

Assumption 2 (Global Mere Coherence [22], [33]): The game $\mathcal{G}$ is globally merely coherent if every CP x_* of $\mathcal{G}$ is globally merely variationally stable, i.e., $\langle F(x),x-x_*\rangle\ge 0$ for all $x\in\mathcal{X}$.

Assumption 3 (Local Mere Coherence): The game $\mathcal{G}$ is locally merely coherent (around a CP set $\mathcal{X}_*\subseteq\mathcal{X}$) if there exists a neighborhood U with a non-empty interior, such that $\mathcal{X}_*\subseteq\mathrm{int}(U)$ and for every CP $x_*\in\mathcal{X}_*$, $\langle F(x),x-x_*\rangle\ge 0$ for all $x\in U\cap\mathcal{X}$.

We can infer that the set $\mathcal{X}_*$ is compact due to the inherent properties of the problem setup, i.e., the feasible set is compact, and a CP should fulfill (4).

Remark 1: The reason why we assume that every CP is merely variationally stable in the above assumptions is that we leverage the residual function $\varepsilon(\cdot)$ defined in Lemma 4 to prove convergence. Since the convergence of $\varepsilon(\cdot)$ only implies the existence of a convergent subsequence to a CP, this condition is needed to pass the subsequence convergence to the whole-sequence convergence. In contrast, [21] only requires that there exists a variationally stable x_*, and constructs an energy function specified for x_* to prove the convergence. Yet, another norm condition $\|X_k\|^2\ge\|x_*\|$ for all k is posited regarding the generated sequence $(X_k)_{k\in\mathbb{N}_+}$ and the verification of it can be challenging. If there are multiple solutions satisfying variational stability, at the very beginning of the iteration, it might be unclear which solution one should focus on and the sequence will converge to, and the choice of energy function requires some extra care.

B. OPTIMISTIC MIRROR DESCENT

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

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

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 ψ fulfills the following conditions [35, Sec. 4.1]: i) ψ is continuously 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:

$$D(p, x) = \psi(p) - \psi(x) - \langle \nabla\psi(x), p - x\rangle, \tag{6}$$

$\forall p, x \in \mathrm{dom}\,\psi$. 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 4 (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 $P_{x,\mathcal{X}} : \mathbb{R}^n \to \mathrm{dom}\,\psi \cap \mathcal{X}$ for some fixed $x \in \mathrm{dom}\,\psi \cap \mathcal{X}$ that plays a critical role in mirror descent and its variants:

$$P_{x,\mathcal{X}}(y) = \underset{x'\in\mathcal{X}}{\mathrm{argmin}}\{\langle y, x - x'\rangle + D(x', x)\}. \tag{7}$$

With all these in hand, the OMD [30], [31] can be expressed as below.

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

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 [36]. 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 [24] by exclusively utilizing the first-order information at the leading state $X_{k+1/2}$, rather than at both X_k and $X_{k+1/2}$.

III. MULTI-POINT PSEUDO-GRADIENT ESTIMATION

In this article, 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 [14]. 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:

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

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 $\hat{X}_{k+1/2} := [\hat{X}^i_{k+1/2}]_{i\in\mathcal{N}} \in \mathcal{X}$. Then 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 [14], the SPSA approach incurs an increasing estimation variance when the query radius is reduced to improve the 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, whose counterparts in the field of optimization can be found in [37]. 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:

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

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 $(\Omega, \mathcal{F}, \mathbb{P})$ 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:

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

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 $\tilde{J}^i_\delta$:

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

where $\mathbb{S}_{-i} := \prod_{j\in\mathcal{N}^{-i}} \mathbb{S}_j \subseteq \mathbb{R}^{n^{-i}}$; $\mathbb{V}^i_\delta := \mathrm{vol}(\delta\mathbb{B}_i) \cdot \mathrm{vol}(\delta\mathbb{S}_{-i})$. The lemmas presented below provide an examination of the

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 $\mathcal{X}^i$; T_k satisfies $\sum_{k\in\mathbb{N}_+} 1/T_k < \infty$.
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 $t = 0, \ldots, T_k$ **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 $\hat{f}_{k,t}^i := J^i(\hat{X}_{k+1/2,t}^i; \hat{X}_{k+1/2,t}^{-i})$
  end for
  $G_k^i \leftarrow \frac{n^i}{\delta_k T_k}\sum_{t=1}^{T_k}(\hat{f}_{k,t}^i - \hat{f}_{k,0}^i)u_{k,t}^i = \frac{1}{T_k}\sum_{t=1}^{T_k} G_{k,t}^i$
  $X_{k+1}^i \leftarrow P_{X_k^i,\mathcal{X}^i}(-\tau G_k^i)$
end procedure
Return: $\{\hat{X}_{k+1/2}^i\}_{i\in\mathcal{N}}$

properties of B_k^i and V_k^i, which will be later employed in the proof of the main theorem. Their proofs are reported in Appendix A.

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

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 $V_k := [V_k^i]_{i\in\mathcal{N}}$ satisfies $\mathbb{E}[\|V_k\|^2 \mid \mathcal{F}_k] \le \alpha_V/T_k$ for some positive constant α_V.

IV. A VARIANCE-REDUCTION LEARNING ALGORITHM AND CONVERGENCE ANALYSIS

In view of the properties of OMD introduced in Section II-B, we design a zeroth-order algorithm for merely coherent 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 $\psi(x) := \sum_{i\in\mathcal{N}}\psi^i(x^i)$ with $x := [x^i]_{i\in\mathcal{N}}$ represents the group DGF, which is $\tilde{\mu}$-strongly convex. The proposed approach is outlined in Algorithm 1.

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 ([38, Th. 1]): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_k)_k$ a filtration of $\mathcal{F}$. For each $k = 1, 2, \ldots$, 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 $(\xi_k)_{k\in\mathbb{N}_+}$. The proofs for the results in this section are reported in Appendix B

Theorem 1: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_k)_k$ a filtration of $\mathcal{F}$. For each $k = 1, 2, \ldots$, 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.

Lemma 4 (Standing Inequality): Suppose Assumption 1 holds and the step size τ satisfies $(\tau L/\tilde{\mu})^2 \le 1/12$. For the iteration $k \ge 3$, the following recurrent relation holds:

$$\begin{aligned} D(x_*, X_{k+1}) + \frac{\tilde{\mu}}{15}\|X_{k+1/2} - X_{k-1/2}\|^2 &\le D(x_*, X_k) \\ &+ \frac{\tilde{\mu}}{15}\|X_{k-1/2} - X_{k-3/2}\|^2 - \frac{\tilde{\mu}}{30}\|X_{k+1/2} - X_{k-1/2}\|^2 \\ &- \frac{\tilde{\mu}}{30}\|X_k - X_{k+1/2}\|^2 - \frac{\tilde{\mu}}{40}\varepsilon(X_k) \\ &- \tau\langle F(X_{k+1/2}), X_{k+1/2} - x_*\rangle + \hat{\Delta}_k, \end{aligned} \tag{11}$$

where the residual function is defined as $\varepsilon(x) := \|x - P_{x,\mathcal{X}}(-\tau F(x))\|^2$ and the errors are captured by $\hat{\Delta}_k := |\tau\langle B_k, X_{k+1/2} - x_*\rangle| - \tau\langle V_k, X_{k+1/2} - x_*\rangle + \tilde{\mu}/(12L^2)\|B_k - B_{k-1} + V_k - V_{k-1}\|^2 + \tilde{\mu}/(120L^2)\cdot\|B_{k-1} - V_{k-1}\|^2 + \tilde{\mu}/(15L^2)\cdot\|B_{k-1} - B_{k-2} + V_{k-1} - V_{k-2}\|^2$.

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

Theorem 2: Consider a game $\mathcal{G}$. Suppose that Assumptions 1, 2, and 4 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

$$\sum_{k\in\mathbb{N}_+}\delta_k<\infty,\quad \sum_{k\in\mathbb{N}_+}1/T_k<\infty. \tag{12}$$

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 t.

V. LOCAL CONVERGENCE OF THE BANDIT LEARNING ALGORITHM

This section is dedicated to exploring the scenario in which the mere coherence property does not hold on the whole feasible set $\mathcal{X}$ but instead on a limited vicinity of certain CPs. In preparation for further analysis, we postulate the following Lipschitz assumption on the group DGF ψ. As a reminder, the group DGF ψ is defined as the sum of individual DGFs, i.e., $\psi(x):=\sum_{i\in\mathcal{N}}\psi^i(x^i)$.

Assumption 5: The group DGF ψ is $\tilde{L}$-smooth on $\mathcal{X}$, i.e., for arbitrary x_a and x_b in $\mathcal{X}$,

$$\psi(x_a)\le\psi(x_b)+\langle\nabla\psi(x_b),x_a-x_b\rangle+\frac{\tilde{L}}{2}\|x_a-x_b\|^2.$$

An equivalent condition is that $\nabla\psi:\mathcal{X}\to\mathbb{R}^n$ is $\tilde{L}$-Lipschitz:

$$\langle\nabla\psi(x_a)-\nabla\psi(x_b),x_a-x_b\rangle\le\tilde{L}\|x_a-x_b\|^2.$$

As a result, for each player $i\in\mathcal{N}$, its DGF ψ^i is $\tilde{L}^i$-smooth with the constant $\tilde{L}^i\le\tilde{L}$.

Assuming Assumption 3 holds, we can identify a smaller region around the CP set $\mathcal{X}_*$ as $\tilde{U}_\epsilon:=\{x:D(\mathcal{X}_*,x)\le\epsilon\}\subseteq U$, where $D(\mathcal{X}_*,x):=\inf_{x'\in\mathcal{X}_*}D(x',x)$. For each $x_*\in\mathcal{X}_*$, we also let $\tilde{U}_\epsilon(x_*):=\{x:D(x_*,x)\le\epsilon\}\subseteq\tilde{U}_\epsilon$. It is straightforward to verify that $\tilde{U}_\epsilon=\cup_{x_*\in\mathcal{X}_*}\tilde{U}_\epsilon(x_*)$. In light of the relation $D(x_*,x)\ge\frac{\tilde{\mu}}{2}\|x-x_*\|^2$, we can deduce that $\|x_*-x\|^2\le\frac{2\epsilon}{\tilde{\mu}}$ holds for all $x\in\tilde{U}_\epsilon(x_*)$ as well. In the forthcoming analysis, we will center around the following two sets that take feasibility into account:

$$U_\epsilon:=\tilde{U}_\epsilon\cap\mathcal{X}\text{ and }U_\epsilon(x_*):=\tilde{U}_\epsilon(x_*)\cap\mathcal{X}. \tag{13}$$

To facilitate our analysis, for an arbitrary $x_*\in\mathcal{X}_*$, we define $\hat{\Delta}^B_{k,x_*}$ and $\hat{\Delta}^V_{k,x_*}=\tilde{\Delta}^V_{k,x_*}+\bar{\Delta}^V_{k,x_*}$ such that $\hat{\Delta}_k\le\hat{\Delta}^B_{k,x_*}+\hat{\Delta}^V_{k,x_*}$ with $\hat{\Delta}_k$ given in Lemma 4 as

$$\begin{aligned}\hat{\Delta}^B_{k,x_*}&:=|\tau\langle B_k,X_{k+1/2}-x_*\rangle|+(\tilde{\mu}/L^2)\cdot\\&\quad\left((1/3)\|B_k\|^2+(37/60)\|B_{k-1}\|^2+(4/15)\|B_{k-2}\|^2\right),\\\tilde{\Delta}^V_{k,x_*}&:=-\tau\langle V_k,X_{k+1/2}-x_*\rangle,\ \bar{\Delta}^V_{k,x_*}:=(\tilde{\mu}/L^2)\cdot\\&\quad\left((1/3)\|V_k\|^2+(37/60)\|V_{k-1}\|^2+(4/15)\|V_{k-2}\|^2\right).\end{aligned}$$

For conciseness, we shall henceforth drop the subscript x_* in $\tilde{\Delta}^V_{k,x_*}$, $\bar{\Delta}^V_{k,x_*}$, etc., for notational simplicity and use the following notations:

$$S_k:=\sum_{t=3}^{k}\tilde{\Delta}^V_{k,x_*}\in\mathcal{F}_{k+1},\text{ and }R_k:=\sum_{t=3}^{k}\bar{\Delta}^V_{k,x_*}\in\mathcal{F}_{k+1}.$$

The variables S_k and R_k represent upper bounds for the cumulated errors introduced by the stochastic error $(V_t)_{t=1,\dots,k}$. Define the event $E^{x_*}_k$ for $k\ge3$ as follows:

$$E^{x_*}_k:=\left\{\omega:\max_{3\le\ell\le k}\left[|S_k|+R_k\right](\omega)\le\frac{\epsilon}{16}\right\}. \tag{14}$$

In particular, we set $E^{x_*}_2=\Omega$. It is worth mentioning that $E^{x_*}_k\in\mathcal{F}_{k+1}$ since $X_{k+1/2}\in\mathcal{F}_k$ yet $V_k\notin\mathcal{F}_k$ and it forms a contracting sequence of events, i.e., $E^{x_*}_2\supseteq E^{x_*}_3\supseteq E^{x_*}_4\supseteq\cdots\supseteq E^{x_*}_k\supseteq\cdots$. Furthermore, we draw attention to that the values of S_k and R_k are dependent on x_* and $E^{x_*}_k$ is tied to ϵ, although this dependence is not explicitly captured in the notations. The proofs of this section are reported in Appendix C.

Lemma 5: Suppose Assumption 1 holds and $X_{k+1/2}\in U_\epsilon$. Then the MPG is upper bounded by a constant $\bar{G}$, i.e.,

$$\|G_k\|\le\bar{G}:=2N\sum_{i\in\mathcal{N}}n^i\bar{\nabla}^i_\epsilon, \tag{15}$$

where for each $i\in\mathcal{N}$, $\bar{\nabla}^i_\epsilon:=\max_{z\in U_\epsilon}\|\nabla_x J^i(z)\|$.

Lemma 6: Suppose Assumptions 1, 3, and 5 hold. Moreover, there exists an $x_*\in\mathcal{X}_*$ such that the leading state $X_{k-1/2}$ and the action profile X_k satisfy $X_{k-1/2}\in U_\epsilon(x_*)$ and $D(x_*,X_k)\le(7/8)\epsilon$, respectively. Additionally, τ is chosen sufficiently small such that $\tau\tilde{L}\bar{G}/\tilde{\mu}^{3/2}\le\sqrt{2}/16\sqrt{\epsilon}$. Then $D(x_*,X_{k+1/2})\le\epsilon$, i.e., $X_{k+1/2}\in U_\epsilon(x_*)$.

The event $E^{x_*}_k$ represents that up to iteration k, the cumulated error induced by the stochastic error never goes beyond the chosen threshold $\epsilon/16$. In Lemma 7, we will prove that if $E^{x_*}_k$ happens, then the leading state $X_{t+1/2}$ will stay within the region of attraction $U_\varepsilon(x_*)$ for $t=1,\dots,k+1$.

Lemma 7: Suppose Assumptions 1, 3, and 5 hold, and there exists an $x_*\in\mathcal{X}_*$ such that $X_1\in U_{\epsilon/2}(x_*)\subseteq U_\epsilon$. Moreover, τ and the monotonically decreasing sequence $(\delta_k)_{k\in\mathbb{N}_+}$ are properly selected such that

$$\begin{aligned}&\tau\tilde{L}\bar{G}/\tilde{\mu}^{3/2}\le\sqrt{2}/16\sqrt{\epsilon},\\&\tau\bar{G}\cdot\left(\frac{\epsilon}{\tilde{\mu}}\right)^{1/2}+\frac{\tau^2}{\tilde{\mu}}\bar{G}^2\le\epsilon/16,\text{ and}\\&\sum_{k\ge3}\left(\tau\alpha_B\left(\frac{2\epsilon}{\tilde{\mu}}\right)^{1/2}\delta_k+\frac{5\tilde{\mu}\alpha_B^2}{4L^2}(\delta_{k-2})^2\right)<(1/16)\epsilon.\end{aligned} \tag{16}$$

Then on the event $E^{x_*}_K$, the sequence $(X_{t+3/2})_{t\le K}$ will not escape $U_\epsilon(x_*)$.

To leverage the conditional invariance of $U_\epsilon(x_*)$ regarding the whole sequence $(X_{k+1/2})_{k\in\mathbb{N}_+}$, we construct the limiting event that imposes an upper bound on stochastic errors:

$$E^{x_*}_\infty:=\left\{\omega:\sup_{\ell\ge3}[|S_\ell|+R_\ell](\omega)\le\frac{\epsilon}{16}\right\}. \tag{17}$$

In Theorem 3, we will prove that the probability measure of event $E_\infty^{x_*}$ can be made arbitrarily close to 1 by letting the sample size sequence $(T_k)_{k\in\mathbb{N}_+}$ increase rapidly enough.

Theorem 3: Suppose Assumptions 1, 3, and 5 hold, $X_1 \in U_{\epsilon/2}$, and τ and $(\delta_k)_{k\in\mathbb{N}_+}$ satisfy the conditions listed in (16). Let $p \in (0, 1)$ be an arbitrary but fixed constant. The sequence $(T_k)_{k\in\mathbb{N}_+}$ is monotonically increasing and fulfills $\frac{\alpha_V}{\tilde{\epsilon}}(\frac{2\tau^2\epsilon}{\tilde{\mu}} + \frac{5\tilde{\mu}}{4L^2}) \cdot \sum_{k=1}^{\infty} \frac{1}{T_k} \le p$ with $\tilde{\epsilon} := ((\frac{\epsilon}{16} - \frac{1}{4}) \vee \frac{1}{4}) \wedge (\frac{\epsilon}{16})^2$. Then for any $x_* \in \mathcal{X}_*$ with $X_1 \in U_{\epsilon/2}(x_*)$, the probability of event $E_\infty^{x_*}$ satisfies $\mathbb{P}(E_\infty^{x_*}) \ge 1 - p$.

While the occurrence of the event $E_\infty^{x_*}$ depends on the particular $x_* \in \mathcal{X}_*$ selected, the conditions outlined in Theorem 3 that ensure its probability can be close to 1 are uniform across $\mathcal{X}_*$ and do not rely on x_*. Likewise, the conditions stated in Lemma 7 to guarantee the invariance of $X_{t+3/2}$ regarding $U_\epsilon(x_*)$ do not depend on x_*.

Finally, we will show in Theorem 4 that if the random sample ω belongs to event $E_\infty^{x_*}$, the actual sequence of play will locally converge to a CP $x_\star \in U_\epsilon(x_*)$.

Theorem 4: Suppose Assumptions 1, 3, 4, and 5 hold, $X_1 \in U_{\epsilon/2}$. Moreover, the selected τ, $(\delta_k)_{k\in\mathbb{N}_+}$, and $(T_k)_{k\in\mathbb{N}_+}$ satisfy $(\tau L/\tilde{\mu})^2 \le 12$, the conditions listed in (16), and $\sum_{k\in\mathbb{N}_+} 1/T_k < \infty$. Then for any $x_* \in \mathcal{X}_*$ that satisfies $X_1 \in U_{\epsilon/2}(x_*)$, on event $E_\infty^{x_*}$, the actual sequence of play will converge a.s. to a CP $x_\star \in U_\epsilon$.

Remark 2: Combining the results of Theorems 3 and 4 yields that if all the conditions given in these two theorems are fulfilled and $(T_k)_{k\in\mathbb{N}_+}$ is chosen to satisfy $\frac{\alpha_V}{\tilde{\epsilon}}(\frac{2\tau^2\epsilon}{\tilde{\mu}} + \frac{5\tilde{\mu}}{4L^2}) \cdot \sum_{k=1}^{\infty} \frac{1}{T_k} \le p < 1$, then for arbitrary initialization $X_1 \in U_{\epsilon/2}$, the generated sequence of play $\hat{X}_{k+1/2,t}$ will converge to a CP with probability no less than $1 - p$.

VI. NUMERICAL EXPERIMENTS

In the conference version [32, Sec. V], a rock-paper-scissors (RPS) game and a least square estimation game are examined, both of which satisfy global mere coherence. The RPS game leverages the negative entropy as DGF and its mirror map can be reduced to a softmax function, where the numerical comparison with [21] is also included. In this section, we conduct two sets of numerical experiments that only satisfy local mere coherence but not global mere coherence. We note that the scope of these two games is not covered by the results in [10], [15], [21].

A. TWO-PLAYER MINIMAX PROBLEMS

In this subsection, we use two two-player minimax saddle-point problems to illustrate the effectiveness of the proposed method. Similar numerical examples have been previously discussed in [25], [39], which takes the following form:

$$\underset{x^1\in\mathcal{X}^1}{\text{minimize}}\ \underset{x^2\in\mathcal{X}^2}{\text{maximize}}\ f(x^1, x^2). \tag{18}$$

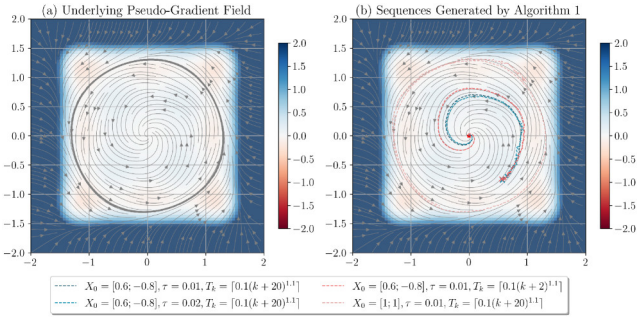

FIGURE 1. The pseudo-gradient field F of (19) and the actual sequences of play $\hat{X}_{k+1/2}$ generated by Algorithm 1.

Specifically, we consider a minimax problem that is formulated as follows:

$$f_a(x^1, x^2) = x^1 \cdot x^2 + \psi_a(x^1) - \psi_a(x^2), \tag{19}$$

where $\psi_a(z) = \frac{2}{21}z^6 - \frac{1}{3}z^4 + \frac{1}{3}z^2$, $\mathcal{X}^1 = \mathcal{X}^2 = [-2, 2]$. As demonstrated in [25, Example 4], $x_* := [0; 0]$ is a global CP for the feasible region. The pseudo-gradient field underneath this saddle point problem is given by $F : \begin{bmatrix} x^1 \\ x^2 \end{bmatrix} \mapsto \begin{bmatrix} \nabla_{x^1} f_a(x^1, x^2) \\ -\nabla_{x^2} f_a(x^1, x^2), \end{bmatrix}$. For the corresponding ODE $\dot{x} = -F(x)$, the region $\mathcal{X}$ under consideration contains both an attracting and a repellent limit cycle, as proved in [25, Prop. 2]. The experimental results are depicted in Fig. 1. The background color displays the value of $\langle F(x), x - x_*\rangle$ with $x \in \mathcal{X}^1 \times \mathcal{X}^2$. The underlying pseudo-gradient field and the attracting limit cycle are graphically presented in Fig. 1(a). In the simulation, we choose the query radius as $\delta_k = 0.1(k + 10)^{-1.1}$. Fig. 1(b) displays the actual sequences of play, with the legends providing a comprehensive account of the parameters selected. Fig. 1 indicates that the appropriate selection of the initial point within the basin of attraction results in a converging sequence towards the CP x_*. When the sample count T_k per iteration is insufficient, the estimation error may temporarily or even permanently drive the sequence away from the solution, as evidenced by the red curve.

In a similar vein, another example featuring a smaller basin of attraction is formulated in the following manner:

$$f_b(x^1, x^2) = (x^1 - 0.05)(x^2 - 0.3) + \psi_b(x^1) - \psi_b(x^2), \tag{20}$$

where $\psi_b(x) = \frac{1}{6}x^6 - \frac{1}{2}x^4 + \frac{1}{4}x^2$, $\mathcal{X}^1 = \mathcal{X}^2 = [-3/2, 3/2]$. We can procure the CP $x_\star = [0.1422; 0.2346]$ via direct numerical computation. In Fig. 2(a), we visualize the underlying pseudo-gradient field and the values of $\langle F(x), x - x_*\rangle$ and highlight the attractive limit cycle with the solid grey curve. In our simulations, we manipulate T_k and X_0, and the results in Fig. 2(b) indicate that while increasing T_k decreases the estimation error, proper selection of X_0 remains a crucial factor for achieving convergence to the CP.

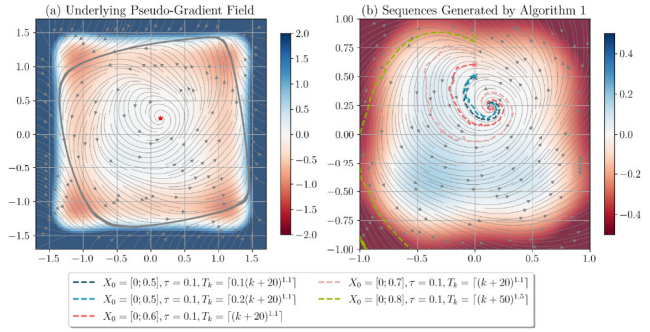

FIGURE 2. The pseudo-gradient field F of (20) and the actual sequences of play $\hat{X}_{k+1/2}$ generated by Algorithm 1.

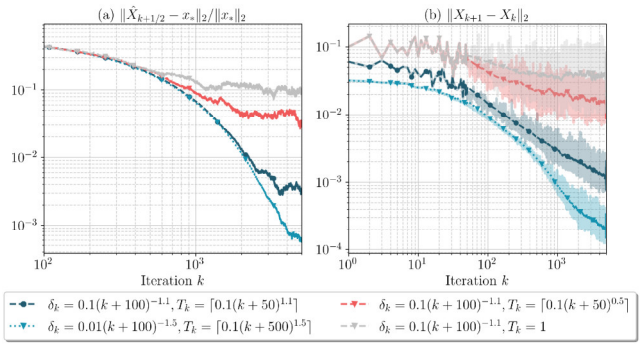

FIGURE 3. Performance of Algorithm 1 in the cognitive radio allocation game: (a) relative distance to the cp under study, i.e., $\|\hat{X}_{k+1/2} - x_*\|_2 / \|x_*\|_2$; (b) updating step lengths per iteration, i.e., $\|X_{k+1} - X_k\|_2$.

B. COGNITIVE RADIO BANDWIDTH ALLOCATION PROBLEM

We consider a cognitive radio bandwidth allocation game whose transmissions are over single-input single-output (SISO) frequency-selective channels [36], [40]. It is composed of P primary users (PUs) and N secondary users (SUs), with the SUs indexed by $\mathcal{N} := \{1, \dots, N\}$. Each SU i competes against each other to maximize its own information rate, while simultaneously accounting for the cost incurred by determining its power allocation vector $x^i \in \mathcal{X}^i \subseteq \mathbb{R}^S$ over the $S \in \mathbb{N}_+$ subcarriers. The objective for each SU $i \in \mathcal{N}$ can be characterized by the following expression:

$$J^i(x^i; x^{-i}) = (p^i)^T x^i - \sum_{s=1}^{S} r_s^i(x^i; x^{-i}) \text{ with}$$

$$r_s^i(x^i; x^{-i}) = \log\left(1 + \frac{|H_s^{ii}|^2 [x^i]_s}{(\sigma_s^i)^2 + \sum_{j \in \mathcal{N}_{-i}} |H_s^{ij}|^2 [x^j]_s}\right), \tag{21}$$

where $(\sigma_s^i)^2$ represents the thermal noise power over the subcarrier s; H_s^{ij} denotes the channel transfer function between the secondary transmitter j and the receiver i; $[x^i]_s$ represents the s-th entry of the vector x^i, which accounts for the power allocation decision of subcarrier s. Additionally, each SU i must adhere to a set of local constraints, which include prescribed transmit power and acceptable levels of degradation on the performance of the PUs. The local feasible set of SU i is described as $\mathcal{X}^i := \{x^i \in \mathbb{R}^S : 0 \le x^i \le b^i, \mathbf{1}^T x^i \le \bar{b}^i, \sum_{s=1}^{S} |Q_s^{pi}|^2 [x^i]_s \le I_{\text{tot}}^{pi}, \forall p \in \{1, \dots, P\}\}$, where we let $b^i \in \mathbb{R}_{++}^S$ and $\bar{b}^i \in \mathbb{R}_{++}$; Q_s^{pi} denote the channel transfer function between the secondary transmitter i and the primary receiver p over the subcarrier s; I_{tot}^{pi} is the maximum interference allowed to be generated by the SU i at the PU p over the whole spectrum.

When conducting the experiments, we consider a game with $P = 3$ PUs, $N = 10$ SUs, and $S = 5$ subcarriers. Let $\tau = 0.01$. An interior CP x_* is found and we verify numerically that the symmetric part of the Jacobian of the pseudo-gradient operator at x_* is positive definite, which entails that it fulfills Assumption 3. The starting point X_0 is initialized in a proper neighborhood of x_*. Four different sets of query radius δ_k and query count T_k have been chosen for implementation. In Fig. 3(a), a comparison of the relative distances to x_* reveals that the convergence rate of the actual sequence of play is positively correlated with the rates of increase in T_k and decrease in δ_k. When T_k remains a constant or merely grows sublinearly, the actual sequence of play will be bounded away from x_* and fail to converge to it. Fig. 3(b) displays a comparison of updating step lengths for various choices of parameters, indicating that the curves associated with summable $(\delta_k)_{k \in \mathbb{N}_+}$ and $(1/T_k)_{k \in \mathbb{N}_+}$ exhibit fewer fluctuations and maintain a decreasing trend. The rolling averages with a window size of 100 are depicted through the opaque curves, while the original fluctuations are illustrated by semi-transparent curves.

VII. 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 globally merely coherent game. For games featuring only local mere coherence, we establish the convergence of actual sequences of play in some neighborhoods of CPs with high probability. 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 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.

APPENDIX

A. PROOF OF SECTION III

1) PROOF OF LEMMA 1

By the tower property $\tilde{\mathcal{F}}_k := \sigma\{\mathcal{F}_k \cup \sigma\{u_{k,0}\}\} \supseteq \mathcal{F}_k$ and the linearity of conditional expectation, we have $\mathbb{E}[G_k^i \mid \mathcal{F}_k] = \mathbb{E}[\mathbb{E}[G_k^i \mid \tilde{\mathcal{F}}_k] \mid \mathcal{F}_k] = \frac{1}{T_k}\sum_{t=1}^{T_k} \mathbb{E}[\frac{n^i}{\delta_k}\mathbb{E}[(\hat{J}_{k,t}^i - \hat{J}_{k,0}^i)u_{k,t}^i \mid \tilde{\mathcal{F}}_k] \mid \mathcal{F}_k]$. For every $t \in \{1, \ldots, T_k\}$, it follows from Lemma 1 of [19] that $\nabla_{x^i}\tilde{J}_{\delta_k}^i(\bar{X}_{k+1/2})$ is a version of the conditional expectation $\frac{n^i}{\delta_k}\mathbb{E}[(\hat{J}_{k,t}^i - \hat{J}_{k,0}^i)u_{k,t}^i \mid \tilde{\mathcal{F}}_k]$. Based on the fact that $\bar{X}_{k+1/2} \in \mathcal{F}_k$, we have the following relation holds a.s.: $\mathbb{E}[G_k^i \mid \mathcal{F}_k] = \frac{1}{T_k}\sum_{t=1}^{T_k}\mathbb{E}[\nabla_{x^i}\tilde{J}_{\delta_k}^i(\bar{X}_{k+1/2}) \mid \mathcal{F}_k] = \nabla_{x^i}\tilde{J}_{\delta_k}^i(\bar{X}_{k+1/2})$. With the above results in hand, the norm of systematic error $\|B_k^i\|$ can be reformulated as $\|B_k^i\| = \|\nabla_{x^i}\tilde{J}_{\delta_k}^i(\bar{X}_{k+1/2}) - \nabla_{x^i}J^i(X_{k+1/2})\|$, and the proof for Lemma 2 of [19] directly carries over.

2) PROOF OF LEMMA 2

Using the definition of MPG and the linearity of conditional expectation, we have:

$$\mathbb{E}[\|G_k^i\|^2 \mid \tilde{\mathcal{F}}_k] = \left(\frac{n^i}{\delta_k T_k}\right)^2 \mathbb{E}\left[\left\|\sum_{t=1}^{T_k}(\hat{J}_{k,t}^i - \hat{J}_{k,0}^i)u_{k,t}^i\right\|^2 \mid \tilde{\mathcal{F}}_k\right]$$

$$= \left(\frac{n^i}{\delta_k T_k}\right)^2 \left(\sum_{t=1}^{T_k}\mathbb{E}[\|(\hat{J}_{k,t}^i - \hat{J}_{k,0}^i)u_{k,t}^i\|^2 \mid \tilde{\mathcal{F}}_k] + \sum_{1\le s,t\le T_k, s\neq t}\mathbb{E}[(\hat{J}_{k,s}^i - \hat{J}_{k,0}^i)(\hat{J}_{k,t}^i - \hat{J}_{k,0}^i)\cdot\langle u_{k,s}^i, u_{k,t}^i\rangle \mid \tilde{\mathcal{F}}_k]\right).$$

For each pair (s,t) with $s \neq t$, denote $\tilde{\mathcal{F}}_{k,s} := \sigma\{\tilde{\mathcal{F}}_k \cup \sigma\{u_{k,s}\}\}$ and the conditional expectation of the inner product can be reformulated as follows:

$$\begin{aligned}
&\left(\frac{n^i}{\delta_k}\right)^2 \mathbb{E}\left[\langle(\hat{J}_{k,s}^i - \hat{J}_{k,0}^i)u_{k,s}^i, (\hat{J}_{k,t}^i - \hat{J}_{k,0}^i)u_{k,t}^i\rangle \mid \tilde{\mathcal{F}}_k\right] \\
&\overset{(a)}{=} \mathbb{E}\Bigg[\Bigg\langle\frac{n^i}{\delta_k}(\hat{J}_{k,s}^i - \hat{J}_{k,0}^i)u_{k,s}^i, \\
&\quad\mathbb{E}\left[\frac{n^i}{\delta_k}(\hat{J}_{k,t}^i - \hat{J}_{k,0}^i)u_{k,t}^i \mid \tilde{\mathcal{F}}_{k,s}\right]\Bigg\rangle\Bigg|\tilde{\mathcal{F}}_k\Bigg] \\
&\overset{(b)}{=} \mathbb{E}\left[\left\langle\frac{n^i}{\delta_k}(\hat{J}_{k,s}^i - \hat{J}_{k,0}^i)u_{k,s}^i, \nabla_{x^i}\tilde{J}_{\delta_k}^i(\bar{X}_{k+1/2})\right\rangle\middle|\tilde{\mathcal{F}}_k\right] \\
&\overset{(c)}{=} \left\langle\mathbb{E}\left[\frac{n^i}{\delta_k}(\hat{J}_{k,s}^i - \hat{J}_{k,0}^i)u_{k,s}^i \mid \tilde{\mathcal{F}}_k\right], \nabla_{x^i}\tilde{J}_{\delta_k}^i(\bar{X}_{k+1/2})\right\rangle \\
&= \|\nabla_{x^i}\tilde{J}_{\delta_k}^i(\bar{X}_{k+1/2})\|^2 \text{ a.s.},
\end{aligned}$$

where (a) follows from the fact that $\tilde{\mathcal{F}}_{k,s} \supseteq \tilde{\mathcal{F}}_k$ and $(\hat{J}_{k,s}^i - \hat{J}_{k,0}^i)u_{k,s}^i$ is $\tilde{\mathcal{F}}_{k,s}$-measurable; (b) and (c) can be deduced by applying the same arguments in Lemma 1. Combining the observations above yields:

$$\mathbb{E}[\|G_k^i\|^2 \mid \tilde{\mathcal{F}}_k] = \left(\frac{n^i}{\delta_k T_k}\right)^2 \sum_{t=1}^{T_k}\mathbb{E}\left[\|(\hat{J}_{k,t}^i - \hat{J}_{k,0}^i)u_{k,t}^i\|^2 \mid \tilde{\mathcal{F}}_k\right] + \left(1 - \frac{1}{T_k}\right)\|\nabla_{x^i}\tilde{J}_{\delta_k}^i(\bar{X}_{k+1/2})\|^2, \text{ a.s.}$$

For the stochastic error $V_k^i := G_k^i - \mathbb{E}[G_k^i \mid \mathcal{F}_k] = G_k^i - \nabla_{x^i}\tilde{J}_{\delta_k}^i(\bar{X}_{k+1/2})$, applying the results above gives:

$$\begin{aligned}
\mathbb{E}[\|V_k^i\|^2 \mid \tilde{\mathcal{F}}_k] &= \mathbb{E}[\|G_k^i - \nabla_{x^i}\tilde{J}_{\delta_k}^i(\bar{X}_{k+1/2})\|^2 \mid \tilde{\mathcal{F}}_k] \\
&= \mathbb{E}[\|G_k^i\|^2 \mid \tilde{\mathcal{F}}_k] - 2\mathbb{E}[\langle G_k^i, \nabla_{x^i}\tilde{J}_{\delta_k}^i(\bar{X}_{k+1/2})\rangle \mid \tilde{\mathcal{F}}_k] \\
&\quad + \|\nabla_{x^i}\tilde{J}_{\delta_k}^i(\bar{X}_{k+1/2})\|^2 \\
&= \mathbb{E}[\|G_k^i\|^2 \mid \tilde{\mathcal{F}}_k] - \|\nabla_{x^i}\tilde{J}_{\delta_k}^i(\bar{X}_{k+1/2})\|^2 \\
&= \left(\frac{n^i}{\delta_k T_k}\right)^2 \sum_{t=1}^{T_k}\mathbb{E}[\|(\hat{J}_{k,t}^i - \hat{J}_{k,0}^i)u_{k,t}^i\|^2 \mid \tilde{\mathcal{F}}_k] \\
&\quad - \frac{1}{T_k}\|\nabla_{x^i}\tilde{J}_{\delta_k}^i(\bar{X}_{k+1/2})\|^2 \\
&\le \left(\frac{n^i}{\delta_k}\right)^2 \frac{1}{T_k}\mathbb{E}[(\hat{J}_{k,t}^i - \hat{J}_{k,0}^i)^2\|u_{k,t}^i\|^2 \mid \tilde{\mathcal{F}}_k] \text{ a.s.}
\end{aligned}$$

The difference $(\hat{J}_{k,t}^i - \hat{J}_{k,0}^i)^2$ can be further bounded as:

$$\begin{aligned}
(\hat{J}_{k,t}^i - \hat{J}_{k,0}^i)^2 &\overset{(a)}{=} (\langle\nabla_x J^i(Z), \hat{X}_{k+1/2,t} - \hat{X}_{k+1/2,0}\rangle)^2 \\
&\le \|\nabla_x J^i(Z)\|^2 \cdot \|\hat{X}_{k+1/2,t} - \hat{X}_{k+1/2,0}\|^2 \\
&\overset{(b)}{\le} \bar{\nabla}_i^2 \cdot \delta_k^2\|u_{k,t} - u_{k,0}\|^2 = 4N\bar{\nabla}_i^2\delta_k^2, \qquad \text{(A.1)}
\end{aligned}$$

where, in (a), we apply the mean value theorem for differentiable function and let Z denote some convex combination of $\hat{X}_{k+1/2,t}$ and $\hat{X}_{k+1/2,0}$; for the relation (b) we let $\bar{\nabla}_i := \max_{x\in\mathcal{X}}\|\nabla_x J^i(x)\|$ and apply the definition in (9). Consequently, it can be directly inferred that $\mathbb{E}[\|V_k^i\|^2 \mid \tilde{\mathcal{F}}_k] \le 4N(\bar{\nabla}_i n^i)^2/T_k$ and $\mathbb{E}[\|V_k\|^2 \mid \tilde{\mathcal{F}}_k] \le 4N\sum_{i\in\mathcal{N}}(\bar{\nabla}_i n^i)^2/T_k$.

B. PROOF OF SECTION IV

1) PROOF OF THEOREM 1

Before proceeding, we attribute the proving technique leveraged below to that of [41, Th. 2.3.5], while we provide complete proof for a simplified version and fill out some omitted steps of the reference for the completeness of this work. By letting $\hat{\zeta}_k := \sum_{t=2}^{k}\zeta_{t-1}$ for $k \ge 2$ and $\hat{\zeta}_1 = 0$, the recurrent inequality can be expressed as

$$\mathbb{E}[Z_{k+1} + \hat{\zeta}_{k+1} \mid \mathcal{F}_k] \le Z_k + \hat{\zeta}_k + \xi_k, \forall k \in \mathbb{N}_+. \qquad \text{(B.1)}$$

Likewise, let $\hat{\xi}_k := \sum_{t=2}^{k}\xi_{t-1}$ for $k \ge 2$ and $\hat{\xi}_1 = 0$, and we have $0 \le \hat{\xi}_k \nearrow \hat{\xi}_\infty$. It follows from the monotone convergence theorem that $\mathbb{E}[\hat{\xi}_k] \nearrow \mathbb{E}[\hat{\xi}_\infty]$ and $\sum_{k\in\mathbb{N}_+}\mathbb{E}[\xi_k] < \infty$ implies $\mathbb{E}[\hat{\xi}_\infty] < \infty$. Through the integration of this definition into (B.1), we can construct a new recurrent inequality as

follows:

$$\mathbb{E}[Z_{k+1}+\hat{\zeta}_{k+1}+\mathbb{E}[\hat{\xi}_\infty \mid \mathcal{F}_{k+1}]-\hat{\xi}_{k+1} \mid \mathcal{F}_k] \leq Z_k+\hat{\zeta}_k+\mathbb{E}[\hat{\xi}_\infty \mid \mathcal{F}_k]-\hat{\xi}_k. \tag{B.2}$$

Based on the observation that $\hat{\xi}_\infty-\hat{\xi}_k\geq 0$, we can let $\tilde{Z}_k:=Z_k+\hat{\zeta}_k+\mathbb{E}[\hat{\xi}_\infty \mid \mathcal{F}_k]-\hat{\xi}_k$, which forms a sequence of non-negative random variables, and deduce that:

$$\mathbb{E}[\tilde{Z}_{k+1} \mid \mathcal{F}_k]\leq \tilde{Z}_k. \tag{B.3}$$

Furthermore, for each $k\in\mathbb{N}_+$, $\mathbb{E}[\tilde{Z}_k]\leq \mathbb{E}[\tilde{Z}_1]=\mathbb{E}[Z_1]+\mathbb{E}[\hat{\xi}_\infty]<\infty$, which together with the preceding observations indicates that $(\tilde{Z}_k)_{k\in\mathbb{N}_+}$ is a non-negative super-martingale. Straightforward application of the martingale convergence theorem yields: $\lim_{k\to\infty}\tilde{Z}_k=\tilde{Z}_\infty$ a.s., where $\tilde{Z}_\infty$ is a L^1 random variable, i.e., $\mathbb{E}[|\tilde{Z}_\infty|]<\infty$. Denote $\hat{\xi}_k^c:=\mathbb{E}[\hat{\xi}_\infty \mid \mathcal{F}_k]-\hat{\xi}_k\in\mathcal{F}_k$. Note that $(\hat{\xi}_k^c)_{k\in\mathbb{N}_+}$ is a non-negative super-martingale and $\lim_{k\to\infty}\mathbb{E}[\hat{\xi}_k^c]=\lim_{k\to\infty}\mathbb{E}[\mathbb{E}[\hat{\xi}_\infty \mid \mathcal{F}_k]-\hat{\xi}_k]=\lim_{k\to\infty}(\mathbb{E}[\hat{\xi}_\infty]-\mathbb{E}[\hat{\xi}_k])=\mathbb{E}[\hat{\xi}_\infty]-\lim_{k\to\infty}\mathbb{E}[\hat{\xi}_k]=0$ as demonstrated earlier, and thus $\hat{\xi}_k^c\overset{k\to\infty}{\to}0$ a.s. As a result, $\lim_{k\to\infty}(Z_k+\hat{\zeta}_k)=\tilde{Z}_\infty$ a.s. Since the sequence $(\hat{\zeta}_k)_{k\in\mathbb{N}_+}$ is non-negative, monotonically increasing and bounded from above, its limit exists a.s., i.e., $\lim_{k\to\infty}\hat{\zeta}_k=\hat{\zeta}_\infty$ a.s. Moreover, due to the surrogate relation that $\hat{\zeta}_\infty\leq\tilde{Z}_\infty$ and $\mathbb{E}[\tilde{Z}_\infty]<\infty$, we then obtain $\mathbb{E}[\hat{\zeta}_\infty]<\infty$. Therefore, we arrive at the conclusion that $\sum_{k\in\mathbb{N}_+}\zeta_k=\lim_{k\to\infty}\hat{\zeta}_k<\infty$ a.s. and $\lim_{k\in\mathbb{N}_+}Z_k=\tilde{Z}_\infty-\hat{\zeta}_\infty$ a.s. and the limit is L^1, i.e., $\mathbb{E}[\tilde{Z}_\infty-\hat{\zeta}_\infty]<\infty$.

2) PROOF OF LEMMA 4

By applying the standing recurrent inequality of OMD [19, Lemma A.2][22, Prop. B.3] and letting x_* denote one CP of $\mathcal{G}$, we can obtain the following relation for the k-th iteration:

$$\begin{aligned}
&D(x_*,X_{k+1})\leq D(x_*,X_k)-\tau\langle G_k,X_{k+1/2}-x_*\rangle\\
&+\frac{\tau^2}{2\tilde{\mu}}\|G_k-G_{k-1}\|^2-\frac{\tilde{\mu}}{2}\|X_{k+1/2}-X_k\|^2\\
&\leq D(x_*,X_k)-\tau\langle F(X_{k+1/2}),X_{k+1/2}-x_*\rangle-\frac{\tilde{\mu}}{2}\\
&\cdot\|X_{k+1/2}-X_k\|^2-\tau\langle B_k,X_{k+1/2}-x_*\rangle-\tau\langle V_k,X_{k+1/2}-x_*\rangle\\
&+\frac{\tau^2}{2\tilde{\mu}}\|F(X_{k+1/2})-F(X_{k-1/2})+B_k-B_{k-1}+V_k-V_{k-1}\|^2,
\end{aligned}$$

where we apply the decomposition of MPG. By appealing to the Cauchy-Schwarz inequality and the L-Lipschitz continuity of F, we can derive that

$$\begin{aligned}
&D(x_*,X_{k+1})\leq D(x_*,X_k)-\tau\langle F(X_{k+1/2}),X_{k+1/2}-x_*\rangle\\
&-\frac{\tilde{\mu}}{2}\|X_{k+1/2}-X_k\|^2+\frac{(\tau L)^2}{\tilde{\mu}}\|X_{k+1/2}-X_{k-1/2}\|^2+\hat{\Delta}_{k,1},
\end{aligned}\tag{B.4}$$

where $\hat{\Delta}_{k,1}:=-\tau\langle B_k,X_{k+1/2}-x_*\rangle-\tau\langle V_k,X_{k+1/2}-x_*\rangle+\tau^2/\tilde{\mu}\|B_k-B_{k-1}+V_k-V_{k-1}\|^2$ represents the error term, which we aim to demonstrate as being suitably diminutive.

To facilitate the convergence analysis in the merely coherent scenario, we upper bound $-\|X_{k+1/2}-X_k\|^2$ as follows:

$$\begin{aligned}
&-\|X_{k+1/2}-X_k\|^2\leq-\frac{1}{2}\|X_k-P_{X_k,\mathcal{X}}(-\tau F(X_k))\|^2\\
&+\|P_{X_k,\mathcal{X}}(-\tau G_{k-1})-P_{X_k,\mathcal{X}}(-\tau F(X_k))\|^2\\
&\overset{(a)}{\leq}-\frac{1}{2}\varepsilon(X_k)+\frac{\tau^2}{\tilde{\mu}^2}\|G_{k-1}-F(X_k)\|^2\leq-\frac{1}{2}\varepsilon(X_k)\\
&+2\left(\frac{\tau L}{\tilde{\mu}}\right)^2\|X_{k-1/2}-X_k\|^2+2\left(\frac{\tau}{\tilde{\mu}}\right)^2\|B_{k-1}+V_{k-1}\|^2\\
&\leq-\frac{1}{2}\varepsilon(X_k)+4\left(\frac{\tau L}{\tilde{\mu}}\right)^2\|X_{k-1/2}-X_{k+1/2}\|^2\\
&+4\left(\frac{\tau L}{\tilde{\mu}}\right)^2\|X_{k+1/2}-X_k\|^2+2\left(\frac{\tau}{\tilde{\mu}}\right)^2\|B_{k-1}+V_{k-1}\|^2,
\end{aligned}$$

where in (a), $\varepsilon(x):=\|x-P_{x,\mathcal{X}}(-\tau F(x))\|^2$ serves as a residual function and we leverage the $1/\tilde{\mu}$-Lipscthiz continuity of $P_{X_k,\mathcal{X}}$ [19, Lemma A.1 iv)]. By the observation that $\varepsilon(x_\star)=0$ is equivalent to the zero inclusion that $0\in N_{\mathcal{X}}(x_\star)+\tau F(x_\star)$, we can assert that $x_\star$ is a CP of $\mathcal{G}\iff\varepsilon(x_\star)=0$. In light of the upper bound derived above and the choice of step size $(\tau L/\tilde{\mu})^2\leq 1/12$, (B.4) can be reformulated as:

$$\begin{aligned}
&D(x_*,X_{k+1})\leq D(x_*,X_k)-\frac{\tilde{\mu}}{2}\left(1-\frac{1}{10}\right)\|X_{k+1/2}-X_k\|^2\\
&+\frac{1}{10}\Bigg(-\frac{\tilde{\mu}}{4}\varepsilon(X_k)+\frac{\tilde{\mu}}{6}\|X_{k-1/2}-X_{k+1/2}\|^2+\frac{\tilde{\mu}}{6}\\
&\cdot\|X_{k+1/2}-X_k\|^2\Bigg)+\frac{\tilde{\mu}}{12}\|X_{k+1/2}-X_{k-1/2}\|^2\\
&-\tau\langle F(X_{k+1/2}),X_{k+1/2}-x_*\rangle+\hat{\Delta}_{k,1}+\hat{\Delta}_{k,2},
\end{aligned}\tag{B.5}$$

where $\hat{\Delta}_{k,2}:=\tilde{\mu}/(120L^2)\cdot\|B_{k-1}-V_{k-1}\|^2$. We reapply the Cauchy-Schwarz inequality to $\|X_{k+1/2}-X_{k-1/2}\|^2$, yielding

$$\|X_{k+1/2}-X_{k-1/2}\|^2\leq 2\|X_{k+1/2}-X_k\|^2+2\|X_k-X_{k-1/2}\|^2,$$

while it can be recursively obtained that for all $k\geq 3$,

$$\begin{aligned}
&\|X_k-X_{k-1/2}\|^2=\|\nabla\psi^*(\nabla\psi(X_{k-1})-\tau G_{k-1})\\
&-\nabla\psi^*(\nabla\psi(X_{k-1})-\tau G_{k-2})\|^2\leq(\tau/\tilde{\mu})^2\|G_{k-1}-G_{k-2}\|^2\\
&\leq 2(\tau L/\tilde{\mu})^2\|X_{k-1/2}-X_{k-3/2}\|^2+2(\tau/\tilde{\mu})^2\Delta_{k,3},
\end{aligned}$$

with $\Delta_{k,3}:=\|B_{k-1}-B_{k-2}+V_{k-1}-V_{k-2}\|^2$. Adding $(\tilde{\mu}/10)\cdot\|X_{k+1/2}-X_{k-1/2}\|^2$ to both sides of (B.5) and substituting $\|X_{k+1/2}-X_{k-1/2}\|^2$ of the R.H.S. with the proceeding inequality produce:

$$\begin{aligned}
&D(x_*,X_{k+1})+\frac{\tilde{\mu}}{10}\|X_{k+1/2}-X_{k-1/2}\|^2\leq D(x_*,X_k)\\
&-\tau\langle F(X_{k+1/2}),X_{k+1/2}-x_*\rangle-\frac{13\tilde{\mu}}{30}\|X_{k+1/2}-X_k\|^2
\end{aligned}$$

$$
\begin{aligned}
&- \frac{\tilde{\mu}}{40}\varepsilon(X_k) + \frac{\tilde{\mu}}{5}\|X_{k+1/2} - X_{k-1/2}\|^2 + \hat{\Delta}_{k,1} + \hat{\Delta}_{k,2} \\
&\leq D(x_*, X_k) + \frac{\tilde{\mu}}{15}\|X_{k-1/2} - X_{k-3/2}\|^2 \\
&\quad - \tau\langle F(X_{k+1/2}), X_{k+1/2} - x_*\rangle - \frac{\tilde{\mu}}{30}\|X_k - X_{k+1/2}\|^2 \\
&\quad - \frac{\tilde{\mu}}{40}\varepsilon(X_k) + \hat{\Delta}_{k,1} + \hat{\Delta}_{k,2} + \hat{\Delta}_{k,3},
\end{aligned}
$$

where $\hat{\Delta}_{k,3} := \tilde{\mu}/(15L^2)\Delta_{k,3}$. Further manipulating the coefficients of $\|X_{k+1/2} - X_{k-1/2}\|^2$ gives $\forall k \geq 3$:

$$
\begin{aligned}
&D(x_*, X_{k+1}) + \frac{\tilde{\mu}}{15}\|X_{k+1/2} - X_{k-1/2}\|^2 \leq D(x_*, X_k) \\
&\quad + \frac{\tilde{\mu}}{15}\|X_{k-1/2} - X_{k-3/2}\|^2 - \frac{\tilde{\mu}}{30}\|X_{k+1/2} - X_{k-1/2}\|^2 \\
&\quad - \frac{\tilde{\mu}}{30}\|X_k - X_{k+1/2}\|^2 - \frac{\tilde{\mu}}{40}\varepsilon(X_k) \\
&\quad - \tau\langle F(X_{k+1/2}), X_{k+1/2} - x_*\rangle + \hat{\Delta}_k, \qquad \text{(B.6)}
\end{aligned}
$$

where we set $\hat{\Delta}_k := |\tau\langle B_k, X_{k+1/2} - x_*\rangle| - \tau\langle V_k, X_{k+1/2} - x_*\rangle + \tilde{\mu}/(12L^2)\|B_k - B_{k-1} + V_k - V_{k-1}\|^2 + \tilde{\mu}/(120L^2) \cdot \|B_{k-1} - V_{k-1}\|^2 + \tilde{\mu}/(15L^2) \cdot \|B_{k-1} - B_{k-2} + V_{k-1} - V_{k-2}\|^2 \geq \hat{\Delta}_{k,1} + \hat{\Delta}_{k,2} + \hat{\Delta}_{k,3}$.

3) PROOF OF THEOREM 2

Utilizing the fact that x_* is a critical point of the game and Assumption 2 is satisfied, we can infer that, when $X_{k+1/2} \in \mathcal{X}$, $\langle F(X_{k+1/2}), X_{k+1/2} - x_*\rangle \geq 0$. By invoking Lemma 4 and taking the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}_k]$ of both sides, we can deduce that

$$
\begin{aligned}
&\mathbb{E}[D(x_*, X_{k+1}) + \frac{\tilde{\mu}}{15}\|X_{k+1/2} - X_{k-1/2}\|^2 \mid \mathcal{F}_k] \leq D(x_*, X_k) \\
&\quad + \frac{\tilde{\mu}}{15}\|X_{k-1/2} - X_{k-3/2}\|^2 - \frac{\tilde{\mu}}{30}\|X_{k+1/2} - X_{k-1/2}\|^2 \\
&\quad - \frac{\tilde{\mu}}{30}\|X_k - X_{k+1/2}\|^2 - \frac{\tilde{\mu}}{40}\varepsilon(X_k) + \mathbb{E}[\hat{\Delta}_k \mid \mathcal{F}_k]. \qquad \text{(B.7)}
\end{aligned}
$$

The parameters satisfy $\sum_{k\in\mathbb{N}_+}\delta_k < \infty$ and $\sum_{k\in\mathbb{N}_+} 1/T_k < \infty$, which together with Lemma 1 implies $\sum_{k\geq 3}|\tau\langle B_k, X_{k+1/2} - x_*\rangle| \leq \sum_{k\geq 3}\tau\|B_k\| \cdot D_{\mathcal{X}} \leq \sum_{k\geq 3}\tau D_{\mathcal{X}}\alpha_B\delta_k < \infty$. Likewise, $\sum_{k\geq 3}\|B_k\|^2 \leq \sum_{k\geq 3}(\alpha_B\delta_k)^2 < \infty$. The application of Lemma 2 allows us to characterize the squared norm of the stochastic error V_k, resulting in $\sum_{k\geq 3}\mathbb{E}[\|V_k\|^2] \leq \sum_{k\geq 3}\alpha_V/T_k < \infty$. Moreover, the inner product involving the stochastic error V_k satisfies

$$
\mathbb{E}[\langle V_k, X_{k+1/2} - x_*\rangle] = \mathbb{E}[\langle \mathbb{E}[V_k \mid \mathcal{F}_k], X_{k+1/2} - x_*\rangle] = 0.
$$

Through the synthesis of the aforementioned findings, we can ascertain that $\sum_{k\geq 3}\mathbb{E}[\hat{\Delta}_k] < \infty$. Then the application of Theorem 1 allows us to assert the following:

i) $\sum_{k\geq 3}\tilde{\mu}/40 \cdot \varepsilon(X_k) < \infty$ a.s.;
ii) $\sum_{k\geq 3}\tilde{\mu}/30 \cdot \|X_{k+1/2} - X_{k-1/2}\|^2 < \infty$ a.s.;
iii) $\sum_{k\geq 3}\tilde{\mu}/30 \cdot \|X_{k+1/2} - X_k\|^2 < \infty$ a.s.;
iv) $D(x_*, X_{k+1}) + \tilde{\mu}/15 \cdot \|X_{k+1/2} - X_{k-1/2}\|^2$ converges a.s. to some L^1 random variable.

These results entail that there exists a sample set $\hat{\Omega} \subseteq \Omega$ and $\mathbb{P}(\hat{\Omega}) = 1$ such that for any $\omega \in \hat{\Omega}$, the above statements i) − iv) hold true for the deterministic sequences $(X_k(\omega))_{k\in\mathbb{N}_+}$ and $(X_{k+1/2}(\omega))_{k\in\mathbb{N}_+}$. Moreover, since $(X_k(\omega))_{k\in\mathbb{N}} \in \mathcal{X}$ and the map $P_{x,\mathcal{X}}(-\tau F(x))$ is continuous in x, there exists a subsequence $(k_m)_{m\in\mathbb{N}_+}$ such that $X_{k_m}(\omega) \overset{m\to\infty}{\to} x_\star$ and $\lim_{m\to\infty}\varepsilon(X_{k_m}(\omega)) = \varepsilon(x_\star) = 0$, i.e., $x_\star$ is a CP of $\mathcal{G}$. We can then substitute $x_\star$ for x_* in iv). Since ii) suggests that $\|X_{k+1/2} - X_{k-1/2}\|^2(\omega) \overset{k\to\infty}{\to} 0$, we can assert from iv) that $D(x_\star, X_k(\omega))$ admits a finite limit. In conjunction with Assumption 4, it follows that $D(x_\star, X_{k_m}(\omega)) \overset{m\to\infty}{\to} 0$ and hence $D(x_\star, X_k(\omega)) \overset{k\to\infty}{\to} 0$, i.e., the base states $(X_k(\omega))_{k\in\mathbb{N}_+}$ converge to $x_\star$. Combining this result with iii) yields that the leading states $(X_{k+1/2}(\omega))_{k\in\mathbb{N}_+}$ converge to $x_\star$, and the a.s. convergence of the actual sequence of play $(\hat{X}_{k+1/2,t}(\omega))_{k\in\mathbb{N}_+}$ to $x_\star$ is directly derived from (9) and $\delta_k \overset{k\to\infty}{\to} 0$.

C. PROOF OF SECTION V

1) PROOF OF LEMMA 5

By employing the definition of MPG, it can be attained that

$$
\begin{aligned}
\|G_k^i\| &\leq \frac{n^i}{\delta_k T_k}\sum_{t=1}^{T_k}|J^i(\hat{X}_{k+1/2,t}) - J^i(\hat{X}_{k+1/2,0})|\|u_{k,t}^i\| \\
&\overset{(a)}{\leq} \frac{n^i}{\delta_k T_k} \cdot \sum_{t=1}^{T_k}|\langle\nabla_x J^i(Z), \delta_k(u_{k,t} - u_{k,0})\rangle| \\
&\overset{(b)}{\leq} \frac{n^i}{T_k}\sum_{t=1}^{T_k}\bar{\nabla}_\epsilon^i \cdot \|u_{k,t} - u_{k,0}\| \overset{(c)}{\leq} 2Nn^i\bar{\nabla}_\epsilon^i,
\end{aligned}
$$

where regarding (a), it stems from the mean value theorem for differentiable function and letting Z denote some convex combination of $\hat{X}_{k+1/2,t}$ and $\hat{X}_{k+1/2,0}$; in (b), we apply the Cauchy-Schwarz inequality and let $\bar{\nabla}_\epsilon^i := \max_{z\in U_\epsilon}\|\nabla_x J^i(z)\|$; (c) ensues from that $\|u_{k,t} - u_{k,0}\| \leq \sum_{i\in\mathcal{N}}\|u_{k,t}^i - u_{k,0}^i\| \leq 2N$. Consequently, $\|G_k\| \leq \sum_{i\in\mathcal{N}}\|G_k^i\| \leq 2N\sum_{i\in\mathcal{N}}n^i\bar{\nabla}_\epsilon^i$.

2) PROOF OF LEMMA 6

By applying the "three-point identity" of the Bregman divergence [35, Sec. 4.1], we can relate $X_{k+1/2}$ to X_k as follows:

$$
\begin{aligned}
D(x_*, X_{k+1/2}) &= D(x_*, X_k) - D(X_{k+1/2}, X_k) \\
&\quad + \langle\nabla\psi(X_{k+1/2}) - \nabla\psi(X_k), X_{k+1/2} - x_*\rangle \\
&\leq D(x_*, X_k) + \|\nabla\psi(X_{k+1/2}) - \nabla\psi(X_k)\| \cdot \|X_{k+1/2} - x_*\| \\
&\overset{(a)}{\leq} D(x_*, X_k) + \tilde{L}\|X_{k+1/2} - X_k\| \cdot \|X_{k+1/2} - x_*\| \\
&\overset{(b)}{\leq} D(x_*, X_k) + \frac{\tau\tilde{L}}{\tilde{\mu}} \cdot \|G_{k-1}\| \cdot \|X_{k+1/2} - x_*\|
\end{aligned}
$$

$$\overset{(c)}{\leq} \frac{7}{8}\epsilon + \frac{\tau\tilde{L}\bar{G}}{\tilde{\mu}} \cdot \left(\frac{2}{\tilde{\mu}} D(x_*, X_{k+1/2})\right)^{1/2}$$
$$\leq \frac{7}{8}\epsilon + \frac{1}{8}\sqrt{\epsilon} \cdot \sqrt{D(x_*, X_{k+1/2})},$$

where (a) is the outcome of Assumption 5; (b) can be deduced from that $X_{k+1/2} = P_{X_k,\mathcal{X}}(-\tau G_{k-1})$ and $P_{X_k,\mathcal{X}}$ is $1/\tilde{\mu}$-Lipschitz continuous; in (c), we employ Lemma 5 and $D(p,x) \geq \tilde{\mu}/2\|p-x\|^2$. It immediately entails that $D(x_*, X_{k+1/2}) \leq \epsilon$.

3) PROOF OF LEMMA 7

We prove this property by induction. For the first iteration, $X_{3/2} = X_1$ and $X_2 = P_{X_1,\mathcal{X}}(-\tau G_1)$, and it follows that

$$D(x_*, X_2) \leq D(x_*, X_1) - \tau\langle G_1, X_{3/2} - x_*\rangle + \frac{\tau^2}{2\tilde{\mu}}\|G_1\|^2$$
$$\leq \epsilon/2 + \tau\bar{G} \cdot \left(\frac{\epsilon}{\tilde{\mu}}\right)^{1/2} + \frac{\tau^2}{2\tilde{\mu}}\bar{G}^2 \leq \epsilon/2 + \epsilon/16 = 9\epsilon/16,$$

where we note that $\|X_{3/2} - x_*\|^2 \leq 2/\tilde{\mu} \cdot D(x_*, X_{3/2}) \leq \epsilon/\tilde{\mu}$. Lemma 6 implies that $X_{5/2} \in U_\epsilon(x_*)$. For the second iteration, $X_{5/2} = P_{X_2,\mathcal{X}}(-\tau G_1)$ and $X_3 = P_{X_2,\mathcal{X}}(-\tau G_2)$, and by similar arguments, it follows that

$$D(x_*, X_3) \leq D(x_*, X_2) - \tau\langle G_2, X_{5/2} - x_*\rangle + \frac{\tau^2}{2\tilde{\mu}}\|G_2 - G_1\|^2$$
$$\leq \epsilon/2 + \epsilon/16 + \tau\bar{G} \cdot \left(\frac{2\epsilon}{\tilde{\mu}}\right)^{1/2} + \frac{2\tau^2}{\tilde{\mu}}\bar{G}^2 \leq 11\epsilon/16.$$

Again using Lemma 6, we have $X_{7/2} \in U_\epsilon(x_*)$.

To prove the statement, we will utilize an inductive argument. For an arbitrary $k \in \{3, 4, \ldots, K\}$, suppose that $X_{t+1/2} \in U_\epsilon(x_*)$ holds for all $3 \leq t \leq k$, and we aim to show $X_{k+3/2} \in U_\epsilon(x_*)$. By applying Lemma 4, neglecting the negative terms on the R.H.S., and telescoping them across $t = 3, \ldots, k$, we have

$$D(x_*, X_{k+1}) + \frac{\tilde{\mu}}{15}\|X_{k+1/2} - X_{k-1/2}\|^2 \leq D(x_*, X_3) + \frac{\tilde{\mu}}{15}$$
$$\cdot \|X_{5/2} - X_{3/2}\|^2 - \sum_{t=3}^{k} \tau\langle F(X_{t+1/2}), X_{t+1/2} - x_*\rangle + \sum_{t=3}^{k} \hat{\Delta}_k.$$

Since by the inductive hypothesis, $X_{t+1/2} \in U_\epsilon(x_*)$ for $3 \leq t \leq k$, $\langle F(X_{t+1/2}), X_{t+1/2} - x_*\rangle \geq 0$, for all $3 \leq t \leq k$. In addition, $\|X_{5/2} - X_{3/2}\|^2 \leq 2\|X_{5/2} - X_2\|^2 + 2\|X_2 - X_1\|^2 \leq 4(\tau\bar{G}/\tilde{\mu})^2$. Combining the properties above yields:

$$D(x_*, X_{k+1}) \leq D(x_*, X_3) + \frac{4}{15} \cdot \frac{\tau^2\bar{G}^2}{\tilde{\mu}} + \sum_{t=3}^{k} \hat{\Delta}_k.$$

We then proceed to upper bound $\hat{\Delta}_k$ by separating it into the parts associated with systematic errors and stochastic errors, i.e., $\hat{\Delta}_k \leq \hat{\Delta}^B_{k,x_*} + \hat{\Delta}^V_{k,x_*}$. After applying Cauchy-Schwarz inequality and triangle inequality, $\hat{\Delta}^B_{k,x_*}$ can be upper bounded as:

$$\hat{\Delta}^B_{k,x_*} \leq \tau\alpha_B\delta_k\left(\frac{2\epsilon}{\tilde{\mu}}\right)^{1/2} + \frac{\tilde{\mu}\alpha_B^2}{L^2}\left(\frac{1}{3}\delta_k^2 + \frac{37}{60}\delta_{k-1}^2 + \frac{4}{15}\delta_{k-2}^2\right)$$
$$\leq \tau\alpha_B\delta_k\left(\frac{2\epsilon}{\tilde{\mu}}\right)^{1/2} + \frac{5\tilde{\mu}\alpha_B^2}{4L^2}\delta_{k-2}^2 = \bar{\Delta}^B_{k,x_*}.$$

On account of the postulated summability $\sum_{k\in\mathbb{N}_+}\delta_k < \infty$, we can choose a proper sequence of query radius such that $\sum_{k\geq 3}\bar{\Delta}^B_{k,x_*} \leq (1/16)\epsilon$. We then move on to examine $\sum_{t=3}^{k}\hat{\Delta}^V_{k,x_*} \leq |S_k| + R_k$. On the event $E_k^{x_*}$, $|S_k| + R_k \leq (1/16)\epsilon$ and hence $D(x_*, X_{k+1}) \leq 11\epsilon/16 + 4/15 \cdot \epsilon/16 + \epsilon/8 < 7\epsilon/8$. By Lemma 6, $D(x_*, X_{k+3/2}) \leq \epsilon$ and $X_{k+3/2} \in U_\epsilon(x_*)$.

4) PROOF OF THEOREM 3

Under the condition that $X_1 \in U_{\epsilon/2}$, it ensues that $\{x \in \mathcal{X}_* : D(x, X_1) \leq \epsilon/2\} \neq \varnothing$, and we can select an arbitrary $x_* \in \mathcal{X}_*$ that satisfies $X_1 \in U_{\epsilon/2}(x_*)$. In the subsequent proof, unless otherwise stated, we will adopt the shorthand notation $\tilde{\Delta}^V_k$ and $\bar{\Delta}^V_k$ to refer to $\tilde{\Delta}^V_{k,x_*}$ and $\bar{\Delta}^V_{k,x_*}$ for brevity. With this in hand, we construct the following recurrent relation for $k \geq 3$

$$\left[(S_{k+1})^2 + R_{k+1}\right] \cdot \mathbb{1}_{E_k^{x_*}} = \left[(S_k)^2 + R_k\right] \cdot \mathbb{1}_{E_k^{x_*}}$$
$$+ \left[2S_k\tilde{\Delta}^V_{k+1} + (\tilde{\Delta}^V_{k+1})^2 + \bar{\Delta}^V_{k+1}\right] \cdot \mathbb{1}_{E_k^{x_*}}, \tag{C.1}$$

where we further expand $\mathbb{1}_{E_k^{x_*}} := \mathbb{1}_{E_{k-1}^{x_*}} - \mathbb{1}_{E_{k-1}^{x_*}\backslash E_k^{x_*}}$ to procure telescoping terms. On event $E_{k-1}^{x_*}\backslash E_k^{x_*}$, we can construct a lower bound as $[(S_k)^2 + R_k] \geq ((\frac{\epsilon}{16} - \frac{1}{4}) \vee \frac{1}{4}) \wedge (\frac{\epsilon}{16})^2 = \tilde{\epsilon}$ since $|S_k| + R_k > \epsilon/16$. Computing expectation by conditioning yields $\mathbb{E}[S_k\tilde{\Delta}^V_{k+1}\mathbb{1}_{E_k^{x_*}}] = \mathbb{E}[S_k\mathbb{1}_{E_k^{x_*}} \cdot \mathbb{E}[\tilde{\Delta}^V_{k+1} \mid \mathcal{F}_{k+1}]] = 0$. In addition, we have $\mathbb{E}[(\tilde{\Delta}^V_{k+1})^2\mathbb{1}_{E_k^{x_*}}] \leq \tau^2\mathbb{E}[\|V_{k+1}\|^2\|X_{k+3/2} - x_*\|^2 \cdot \mathbb{1}_{E_k^{x_*}}] \leq \frac{2\tau^2\epsilon\alpha_V}{\tilde{\mu}T_{k+1}}$, since $X_{k+3/2} \in U_\epsilon$ on event $E_k^{x_*}$ as proved in Lemma 7; $\mathbb{E}[\bar{\Delta}^V_{k+1}] \leq \frac{5\tilde{\mu}}{4L^2} \cdot \frac{\alpha_V}{T_{k-1}}$. Then taking the expectation of both sides of (C.1) gives:

$$\mathbb{E}\left[[(S_{k+1})^2 + R_{k+1}] \cdot \mathbb{1}_{E_k^{x_*}}\right] \leq \mathbb{E}\left[[(S_k)^2 + R_k] \cdot \mathbb{1}_{E_{k-1}^{x_*}}\right]$$
$$- \tilde{\epsilon} \cdot \mathbb{E}[\mathbb{1}_{E_{k-1}^{x_*}\backslash E_k^{x_*}}] + \frac{2\tau^2\epsilon\alpha_V}{\tilde{\mu}T_{k+1}} + \frac{5\tilde{\mu}}{4L^2} \cdot \frac{\alpha_V}{T_{k-1}}. \tag{C.2}$$

Using the results above, it can be shown that $\mathbb{E}[(S_3)^2 + R_3] \leq \frac{2\tau^2\epsilon\alpha_V}{\tilde{\mu}T_3} + \frac{5\tilde{\mu}}{4L^2} \cdot \frac{\alpha_V}{T_1}$. By telescoping (C.2), we obtain

$$\tilde{\epsilon} \cdot \sum_{k=3}^{K}\mathbb{P}\left(E_{k-1}^{x_*}\backslash E_k^{x_*}\right) \leq \mathbb{E}\left[[(S_3)^2 + R_3] \cdot \mathbb{1}_{E_2^{x_*}}\right]$$
$$+ \sum_{k=4}^{K}\left(\frac{2\tau^2\epsilon\alpha_V}{\tilde{\mu}T_k} + \frac{5\tilde{\mu}}{4L^2} \cdot \frac{\alpha_V}{T_{k-2}}\right)$$
$$\leq \sum_{k=3}^{K}\left(\frac{2\tau^2\epsilon\alpha_V}{\tilde{\mu}T_k} + \frac{5\tilde{\mu}}{4L^2} \cdot \frac{\alpha_V}{T_{k-2}}\right) \leq \sum_{k=1}^{K-2}\left(\frac{2\tau^2\epsilon}{\tilde{\mu}} + \frac{5\tilde{\mu}}{4L^2}\right)\frac{\alpha_V}{T_k}.$$

Since $(E_k^{x_*})_{k\geq 2}$ is a contracting sequence of events, we have $\sum_{k=3}^{K}\mathbb{P}(E_{k-1}^{x_*}\backslash E_k^{x_*})=\mathbb{P}(\Omega\backslash E_K^{x_*})=\mathbb{P}((E_K^{x_*})^c)$ and hence

$$\mathbb{P}\left((E_K^{x_*})^c\right)\leq\frac{\alpha_V}{\tilde{\epsilon}}\left(\frac{2\tau^2\epsilon}{\tilde{\mu}}+\frac{5\tilde{\mu}}{4L^2}\right)\cdot\sum_{k=1}^{K-2}\frac{1}{T_k}.$$

Then $((E_k^{x_*})^c)_{k\geq 2}$ is an expanding sequence of events and $(E_k^{x_*})^c\nearrow(E_\infty^{x_*})^c$. By the continuity of probability measure, $\mathbb{P}((E_k^{x_*})^c)\nearrow\mathbb{P}((E_\infty^{x_*})^c)$ and choosing proper $(T_k)_{k\in\mathbb{N}_+}$ yields

$$\mathbb{P}\left((E_\infty^{x_*})^c\right)\leq\frac{\alpha_V}{\tilde{\epsilon}}\left(\frac{2\tau^2\epsilon}{\tilde{\mu}}+\frac{5\tilde{\mu}}{4L^2}\right)\cdot\sum_{k=1}^{\infty}\frac{1}{T_k}\leq p,$$

and hence $\mathbb{P}(E_\infty^{x_*})\geq 1-p$.

5) PROOF OF THEOREM 4

We start by fixing an arbitrary $x_*\in\mathcal{X}_*$ such that $X_1\in U_{\epsilon/2}(x_*)$. Applying the standing inequality from Lemma 4 regarding $x_\star\in\mathcal{X}_*$ that can be different from x_* and taking the indicator function $\mathbb{1}_{E_{k-1}^{x_*}}\in\mathcal{F}_k$ and the inequality $\mathbb{1}_{E_k^{x_*}}\leq\mathbb{1}_{E_{k-1}^{x_*}}$ into account, we have

$$\begin{aligned}
&\mathbb{E}\left[D(x_\star,X_{k+1})\mathbb{1}_{E_k^{x_*}}+\frac{\tilde{\mu}}{15}\|X_{k+1/2}-X_{k-1/2}\|^2\mathbb{1}_{E_k^{x_*}}\mid\mathcal{F}_k\right]\\
&\leq D(x_\star,X_k)\mathbb{1}_{E_{k-1}^{x_*}}+\frac{\tilde{\mu}}{15}\|X_{k-1/2}-X_{k-3/2}\|^2\mathbb{1}_{E_{k-1}^{x_*}}\\
&-\left(\frac{\tilde{\mu}}{30}\|X_{k+1/2}-X_{k-1/2}\|^2+\frac{\tilde{\mu}}{30}\|X_k-X_{k+1/2}\|^2\right.\\
&\left.+\frac{\tilde{\mu}}{40}\varepsilon(X_k)\right)\mathbb{1}_{E_{k-1}^{x_*}}-\tau\langle F(X_{k+1/2}),X_{k+1/2}-x_\star\rangle\mathbb{1}_{E_{k-1}^{x_*}}\\
&+\mathbb{E}\left[\hat{\Delta}_k\mathbb{1}_{E_{k-1}^{x_*}}\mid\mathcal{F}_k\right],
\end{aligned}\tag{C.3}$$

where on the event $E_{k-1}^{x_*}$, we immediately have $X_{k+1/2}\in U_\epsilon(x_*)\subseteq U_\epsilon$ and $\langle F(X_{k+1/2}),X_{k+1/2}-x_\star\rangle\geq 0$. Since $\mathbb{E}[\langle V_k,X_{k+1/2}-x_\star\rangle\mathbb{1}_{E_{k-1}^{x_*}}\mid\mathcal{F}_k]=\langle\mathbb{E}[V_k\mid\mathcal{F}_k],X_{k+1/2}-x_\star\rangle\mathbb{1}_{E_{k-1}^{x_*}}=0$, $\sum_{k\in\mathbb{N}_+}\delta_k<\infty$, and $\sum_{k\in\mathbb{N}_+}1/T_k<\infty$, we can conclude that $\sum_{k\in\mathbb{N}_+}\mathbb{E}[\mathbb{E}[\hat{\Delta}_k\mathbb{1}_{E_{k-1}^{x_*}}\mid\mathcal{F}_k]]<\infty$, regardless of the specific choice of x_*. By applying the extended version of the R-S theorem (Theorem 1), we arrive at the following claims:

i) $\sum_{k\geq 3}\frac{\tilde{\mu}}{40}\varepsilon(X_k)\mathbb{1}_{E_{k-1}^{x_*}}<\infty$ a.s.;
ii) $\sum_{k\geq 3}\frac{\tilde{\mu}}{30}\|X_{k+1/2}-X_{k-1/2}\|^2\mathbb{1}_{E_{k-1}^{x_*}}<\infty$ a.s.;
iii) $\sum_{k\geq 3}\frac{\tilde{\mu}}{30}\|X_k-X_{k+1/2}\|^2\mathbb{1}_{E_{k-1}^{x_*}}<\infty$ a.s.;
iv) $D(x_\star,X_k)\mathbb{1}_{E_{k-1}^{x_*}}+\frac{\tilde{\mu}}{15}\|X_{k-1/2}-X_{k-3/2}\|^2\mathbb{1}_{E_{k-1}^{x_*}}$ converges a.s. to some L^1 random variable.

These results entail that there exists a sample set $\hat{\Omega}\subseteq\Omega$ and $\mathbb{P}(\hat{\Omega})=1$ such that for any $\omega\in\hat{\Omega}\cap E_\infty^{x_*}$, the above statements i) – iv) hold true for the deterministic sequences $(X_k(\omega))_{k\in\mathbb{N}_+}\subseteq U_{7\epsilon/8}(x_*)$ and $(X_{k+1/2}(\omega))_{k\in\mathbb{N}_+}\subseteq U_\epsilon(x_*)$ and all the indicator functions admit the constant value 1.

On account of the continuity of $P_{x,\mathcal{X}}(-\tau F(x))$ in x, there exists a subsequence $(k_m)_{m\in\mathbb{N}_+}$ such that $X_{k_m}(\omega)\overset{m\to\infty}{\to}x_\#$ and $\lim_{m\to\infty}\varepsilon(X_{k_m}(\omega))=\varepsilon(x_\#)=0$, i.e., $x_\#$ is a CP of $\mathcal{G}$. We can then substitute $x_\#$ for $x_\star$ in iv). Since ii) suggests that $\|X_{k+1/2}-X_{k-1/2}\|^2(\omega)\overset{k\to\infty}{\to}0$, we can assert from iv) that $D(x_\#,X_k(\omega))$ admits a finite limit. In conjunction with Assumption 4, it follows that $D(x_\#,X_{k_m}(\omega))\overset{m\to\infty}{\to}0$ and hence $D(x_\#,X_k(\omega))\overset{k\to\infty}{\to}0$, i.e., the base states $(X_k(\omega))_{k\in\mathbb{N}_+}$ converge to $x_\#$. Combining this result with (iii) yields that the leading states $(X_{k+1/2}(\omega))_{k\in\mathbb{N}_+}$ converge to $x_\#$, and the convergence of the actual sequence of play $(\hat{X}_{k+1/2,t}(\omega))_{k\in\mathbb{N}_+}$ to $x_\#$ is directly derived from (9) and $\delta_k\overset{k\to\infty}{\to}0$.

REFERENCES

[1] Z. Jiang and J. Cai, "Game theoretic control of thermal loads in demand response aggregators," in *Proc. Amer. Control Conf.*, 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 Res.*, vol. 56, no. 3, pp. 552–561, 2008.
[3] A. Liniger and J. Lygeros, "A noncooperative game approach to autonomous racing," *IEEE Trans. Control Syst. Technol.*, vol. 28, no. 3, pp. 884–897, May 2020.
[4] 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," *J. Math. Biol.*, vol. 64, pp. 495–512, 2012.
[5] 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," *Ann. Operations Res.*, vol. 228, pp. 135–149, 2015.
[6] N. Li and J. R. Marden, "Designing games for distributed optimization," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 2, pp. 230–242, Apr. 2013.
[7] J. F. Nash Jr., "Equilibrium points in n-person games," *Proc. Nat. Acad. Sci.*, vol. 36, no. 1, pp. 48–49, 1950.
[8] P. Mertikopoulos and Z. Zhou, "Learning in games with continuous action sets and unknown payoff functions," *Math. Program.*, vol. 173, no. 1, pp. 465–507, 2019.
[9] P. Yi and L. Pavel, "An operator splitting approach for distributed generalized Nash equilibria computation," *Automatica*, vol. 102, pp. 111–121, 2019.
[10] T. Tatarenko, W. Shi, and A. Nedić, "Geometric convergence of gradient play algorithms for distributed Nash equilibrium seeking," *IEEE Trans. Autom. Control*, vol. 66, no. 11, pp. 5342–5353, Nov. 2021.
[11] L. Pavel, "Distributed GNE seeking under partial-decision information over networks via a doubly-augmented operator splitting approach," *IEEE Trans. Autom. Control*, vol. 65, no. 4, pp. 1584–1597, Apr. 2020.
[12] M. Bianchi, G. Belgioioso, and S. Grammatico, "Fast generalized Nash equilibrium seeking under partial-decision information," *Automatica*, vol. 136, 2022, Art. no. 110080.
[13] Y. Huang and J. Hu, "Distributed computation of stochastic GNE with partial information: An augmented best-response approach," *IEEE Trans. Control Netw. Syst.*, vol. 10, no. 2, pp. 947–959, Jun. 2023.
[14] M. Bravo, D. Leslie, and P. Mertikopoulos, "Bandit learning in concave N-person games," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 5666–5676.
[15] T. Lin, Z. Zhou, W. Ba, and J. Zhang, "Doubly optimal no-regret online learning in strongly monotone games with bandit feedback," 2021, *arXiv:2112.02856*.
[16] T. Tatarenko and M. Kamgarpour, "On the rate of convergence of payoff-based algorithms to Nash equilibrium in strongly monotone games," 2022, *arXiv:2202.11147*.
[17] D. Drusvyatskiy, M. Fazel, and L. J. Ratliff, "Improved rates for derivative free gradient play in strongly monotone games," in *Proc. IEEE 61st Conf. Decis. Control*, 2022, pp. 3403–3408.
[18] T. Tatarenko and M. Kamgarpour, "Convergence rate of learning a strongly variationally stable equilibrium," 2023, *arXiv:2304.02355*.
[19] Y. Huang and J. Hu, "Zeroth-order learning in continuous games via residual pseudogradient estimates," 2023, *arXiv:2301.02279*.
[20] T. Tatarenko and M. Kamgarpour, "Bandit learning in convex non-strictly monotone games," 2020, *arXiv:2009.04258*.

[21] B. Gao and L. Pavel, "Bandit learning with regularized second-order mirror descent," in *Proc. IEEE 61st Conf. Decis. Control*, 2022, pp. 5731–5738.
[22] 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 *Proc. Int. Conf. Learn. Representations*, 2019. [Online]. Available: https://openreview.net/pdf?id=Bkg8jjC9KQ
[23] A. Kannan and U. V. Shanbhag, "Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants," *Comput. Optim. Appl.*, vol. 74, no. 3, pp. 779–820, 2019.
[24] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson, "Extragradient method with variance reduction for stochastic variational inequalities," *SIAM J. Optim.*, vol. 27, no. 2, pp. 686–724, 2017.
[25] T. Pethick, P. Latafat, P. Patrinos, O. Fercoq, and V. Cevher, "Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://openreview.net/pdf?id=2_vhkAMARk
[26] Y. Cai, A. Oikonomou, and W. Zheng, "Accelerated algorithms for monotone inclusions and constrained nonconvex-nonconcave min-max optimization," 2022, *arXiv:2206.05248*.
[27] Y. Cai and W. Zheng, "Accelerated single-call methods for constrained min-max optimization," in *Proc. Int. Conf. Learn. Representations*, 2023.
[28] J. Diakonikolas, C. Daskalakis, and M. I. Jordan, "Efficient methods for structured nonconvex-nonconcave min-max optimization," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 2746–2754.
[29] T. Pethick, O. Fercoq, P. Latafat, P. Patrinos, and V. Cevher, "Solving stochastic weak minty variational inequalities without increasing batch size," in *Proc. Int. Conf. Learn. Representations*, 2023.
[30] Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos, "On the convergence of single-call stochastic extra-gradient methods," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 6938–6948.
[31] W. Azizian, F. Iutzeler, J. Malick, and P. Mertikopoulos, "The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities," in *Proc. Conf. Learn. Theory*, 2021, pp. 326–358.
[32] Y. Huang and J. Hu, "Bandit online learning in merely coherent games with multi-point pseudo-gradient estimate," 2023, *arXiv:2303.16430*.
[33] P. Mertikopoulos, Y.-P. Hsieh, and V. Cevher, "A unified stochastic approximation framework for learning in games," *Math. Program.*, pp. 1–51, 2023.
[34] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Berlin, Germany: Springer, 2003.
[35] S. Bubeck, "Theory of convex optimization for machine learning," 2014, *arXiv:1405.4980*.
[36] G. Scutari, D. P. Palomar, F. Facchinei, and J.-S. Pang, "Monotone games for cognitive radio systems," in *Distributed Decision Making and Control*. Berlin, Germany: Springer, 2012, pp. 83–112.
[37] 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 Trans. Inf. Theory*, vol. 61, no. 5, pp. 2788–2806, May 2015.
[38] H. Robbins and D. Siegmund, "A convergence theorem for non-negative almost supermartingales and some applications," in *Optimizing Methods in Statistics*. Amsterdam, The Netherlands: Elsevier, 1971, pp. 233–257.
[39] Y.-P. Hsieh, P. Mertikopoulos, and V. Cevher, "The limits of min-max optimization algorithms: Convergence to spurious non-critical sets," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, vol. 139, pp. 4337–4348.
[40] J.-S. Pang, G. Scutari, D. P. Palomar, and F. Facchinei, "Design of cognitive radio systems under temperature-interference constraints: A variational inequality approach," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3251–3271, Jun. 2010.
[41] S. Gadat, "Stochastic optimization algorithms, non asymptotic and asymptotic behaviour," vol. 14, Univ. Toulouse, 2017, pp. 16. [Online]. Available: https://perso.math.univ-toulouse.fr/gadat/files/2012/12/cours_Algo_Stos_M2R5.pdf

YUANHANQING HUANG (Graduate Student Member, IEEE) received the B.E. degree in automatic control from Tongji University, Shanghai, China, in 2017. She is currently a graduate student at the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. Her research interests include game theory and network optimization.

JIANGHAI HU received the B.E. degree in automatic control from Xi'an Jiaotong University, Xi'an, China, in 1994, the M.A. degree in mathematics, and the Ph.D. degree in electrical engineering from the University of California, Berkeley, in 2002 and 2003, respectively. He is currently a Professor with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. His research interests include multi-agent systems, hybrid systems, and control applications.