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

A Bandit Learning Method for Continuous Games under Feedback Delays with Residual Pseudo-Gradient Estimate

Yuanhanqing Huang[1] and Jianghai Hu[1]

Abstract— **Learning in multi-player games can model a large variety of practical scenarios, where each player seeks to optimize its own local objective function, which at the same time relies on the actions taken by others. Motivated by the frequent absence of first-order information such as partial gradients in solving local optimization problems and the prevalence of asynchronicity and feedback delays in multi-agent systems, we introduce a bandit learning algorithm, which integrates mirror descent, residual pseudo-gradient estimates, and the priority-based feedback utilization strategy, to contend with these challenges. We establish that for pseudo-monotone plus games, the actual sequences of play generated by the proposed algorithm converge a.s. to critical points. Compared with the existing method, the proposed algorithm yields more consistent estimates with less variation and allows for more aggressive choices of parameters. Finally, we illustrate the validity of the proposed algorithm through a thermal load management problem of building complexes.**

I. INTRODUCTION

With the proliferation of cyber-physical engineering systems and modern network applications, the non-cooperative multi-player game has emerged as a valuable tool for modeling and investigating the decision-making process of agents with interest conflicts [1]. Each participant in the game seeks to unilaterally optimize its own objective, whose value also depends on the action taken by others. Notable practical applications include thermal load management of autonomous buildings [2], supply-side risk management in power markets [3], power control in wireless communication [4], path planning and control of self-driving cars [5], etc.

Over the past few decades, the control and optimization communities have devoted significant effort to developing solution algorithms for non-cooperative games by reformulating them as variational inequalities [6]. Recently, there has been growing interest in distributed solutions under partial information settings, as they offer advantages in scalability and privacy preservation [7], [8], [9]. Despite their promise in some cases, the applicability of these methods is often limited by the requirement for the existence of first-order/pseudo-gradient oracles or the full knowledge of the objectives, which may not be available in practical settings. Prompted by the need to relax the information requirement, researchers approximate the missing pseudo-gradient information with the actions taken and the resulting objective values. This problem can then be fit into the framework of bandit online learning [10], where at every updating step, each player selects an action, observes the realized objective value, and updates its strategy according to the observed result and the process repeats.

Another practical challenge that hinders the implementation in real-world scenarios is the latency between taking action and receiving bandit feedback, which is further exacerbated in multi-agent systems, where agents could experience heterogeneous delays. Latency can arise as a result of significant communication delays or the fundamental limitation that certain actions take time to manifest their effects. In the context of routing problems [11], assessing the effectiveness of a navigation strategy entails waiting for a driver to execute the instructions, operate the vehicle, and record the time elapsed. In light of the preceding consideration, the primary objective of this work is to propose a bandit online learning algorithm for multi-player continuous games that can ensure convergence despite the presence of feedback delays.

Related Work: In the context of bandit learning in games with instantaneous feedback, Bravo et al. [12] introduced a bandit mirror descent (MD) method that ensures a.s. convergence when the game is strictly monotone. The single-point pseudo-gradient estimate is obtained via the simultaneous perturbation stochastic approximation (SPSA) approach [13]. In the context of strongly monotone games and their variants, the algorithms proposed in [14], [15], [16], [17] similarly employ single-point estimates of the pseudo-gradient and attain a $O(1/t^{1/2})$ convergence rate. The single-point estimates are also applied in [18] and [19] for merely monotone games and their variants. Given the susceptibility of single-point estimates to large variances, a critical factor impacting the efficiency of algorithms, Tatarenko et al. [15] introduced the two-point estimate. This strategy mitigates variance-related issues and enhances the convergence rate to $O(1/t)$ for strongly monotone games. In the field of zeroth-order optimization, Zhang et al. [20] considered a residual feedback scheme to control the estimation variance. By integrating residual pseudo-gradient estimate into the single-call extra-gradient scheme, Huang et al. [21] developed two bandit algorithms. The proposed algorithms only require a single query per iteration and ensure a.s. convergence for pseudo-monotone plus games and achieve $O(1/t^{1-\epsilon})$ convergence rate for strongly monotone games.

To contend with the feedback delays in games, Huang et al. [22] proposed an algorithm based on the improved accelerated gradient descent for potential games, which can tackle cases ranging from sublinear delays to superlinear delays. Zhang et al. [23] focused on the general-sum Markov

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/CDC49753.2023.10383332

games where the agents are impacted by heterogeneous reward delays and proposed the delay-adaptive multi-agent V-learning to procure coarse-correlated equilibria. Of particular relevance is [24], in which Helious et al. delved into the development of a no-regret bandit learning algorithm for strictly monotone games corrupted by homogeneous sublinear reward delays. Nevertheless, the delicate balance between bias and variance of the proposed method is elusive and requires careful calibration. Moreover, its stringent requirements on step sizes and query radius hinder its applicability.

Contributions: First, we propose a bandit learning algorithm under feedback delays, where the delays can be heterogeneous but upper-bounded by a constant or homogeneous with a sublinearly growing upper bound. Our algorithm integrates mirror descent, residual pseudo-gradient estimates, and the priority-based feedback utilization strategy. It is the first algorithm that employs the variance control strategy via single-point residual estimates in the scenario of bandit learning with delays. Second, we establish the a.s. convergence of the proposed algorithm for pseudo-monotone plus games. While some of the proving techniques have been previously established in [21], this paper places additional emphasis on addressing the error caused by delays, which can complicate the problem, particularly when two subsequent realized objective values are required for each single estimate. Compared to the existing method in [24], the proposed algorithm in this work maintains a constant upper bound for the estimation variance and relaxes the conditions on step size and query radius by incorporating the residual pseudo-gradient estimates. In addition, we evaluate the performance of the solution algorithms using the thermal load management problem of buildings. Compared to the existing work, the proposed algorithm achieves faster and more consistent convergence. Due to the page limit, complete proofs are presented in [25].

Basic Notations: 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. Denote $\mathbb{N}_+ \coloneqq \mathbb{N}\backslash\{0\}$ and $\mathbb{R}_{++} \coloneqq (0,+\infty)$. We let $\|\cdot\|_2$ represent the Euclidean norm, $\|\cdot\|$ a general norm, and $\|\cdot\|_*$ its dual. For a set $\mathcal{S}$, let $\mathbb{1}_{\mathcal{S}}$ denote the indicator function for this set, i.e., $\mathbb{1}_{\mathcal{S}}(x) = 1$ if $x \in \mathcal{S}$ and 0 otherwise. 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.

II. SETUP AND PRELIMINARIES

A. Problem Setup

In this subsection, we formalize the multi-player continuous game with feedback delays that we will investigate and introduce the assumptions to impose. In this N-player game $\mathcal{G}$, with the player set given by $\mathcal{N} \coloneqq \{1,\dots,N\}$, each player i needs to optimize its own local objective by determining its local action $x^i \in \mathcal{X}^i$, where $\mathcal{X}^i \subseteq \mathbb{R}^{n^i}$ represents the local strategy space of player i. For brevity, we let the stack vector $x \coloneqq [x^j]_{j\in\mathcal{N}}$ denote the global action, the stack vector $x \coloneqq [x^j]_{j\in\mathcal{N}_{-i}}$ denote the action taken by all players except player i with $\mathcal{N}_{-i} \coloneqq \mathcal{N}\backslash\{i\}$. Similarly, denote the global strategy space $\mathcal{X} \coloneqq \prod_{j\in\mathcal{N}}\mathcal{X}^j \subseteq \mathbb{R}^n$ with $n \coloneqq \sum_{j\in\mathcal{N}} n^j$. Formally, given the action x^{-i} taken by other players, each player i aims to solve the following local problem:

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

The following conditions are imposed regarding the smoothness of objective J^i's and the properties of $\mathcal{X}^i$'s.

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

The underlying probability space is given by $(\Omega,\mathcal{F},\mathbb{P})$. One operator we will leverage throughout is the pseudo-gradient operator $F : \mathcal{X} \to \mathbb{R}^n$, which is defined as the stack of the partial gradient given the smoothness imposed in Assumption 1, i.e.,

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

The Lipschitz continuity of F then entails the fact that each J^i is C^1 and $\mathcal{X}^i$ compact, i.e., there exists some constant L, such that for arbitrary x and $y \in \mathcal{X}$, $\|F(x) - F(y)\|_* \le L\|x - y\|$. In the same vein, the gradient $\nabla_x J^i : \mathcal{X} \to \mathbb{R}^{n^i}$ is also Lipschitz continuous and admits a tighter Lipschitz constant denoted by L^i. Throughout this work, we will concentrate on the solution concept known as critical points (CPs) [26, Section 2], whose definition is given as follows.

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

$$\langle F(x_*), x - x_*\rangle \ge 0,\ \forall x \in \mathcal{X}, \tag{3}$$

which is typically denoted by the abbreviation $\mathrm{VI}(\mathcal{X},F)$. Besides, the following assumption is postulated regarding the monotonicity of F to facilitate the convergence analysis.

Assumption 2: The pseudo-gradient F is pseudo-monotone plus on $\mathcal{X}$, i.e., F is pseudo-monotone, i.e., for all $x, y \in \mathcal{X}$, $\langle F(y), x - y\rangle \ge 0 \implies \langle F(x), x - y\rangle \ge 0$, and satisfies for any action profiles $x, y \in \mathcal{X}$, $\langle F(y), x - y\rangle \ge 0$ and $\langle F(x), x - y\rangle = 0 \implies F(x) = F(y)$.

Pseudo-monotone plus games are a broader class of games than strictly monotone games, but they are not a subset or a superset of merely monotone games. Examples of pseudo-monotone plus games that are not merely monotone can be found in [21, Section V.A][27].

B. Setup for Feedback Delays

In this work, we consider the scenario where there exists some time lag between the time when an action is taken and the time when the associated realized objective value is received by the player. To simplify notation, we let the realized objective value of player i at the k-th iteration be denoted by $\hat{J}^i_k$. Then, for player i, the delay time of $\hat{J}^i_k$ is denoted by d^i_k, and this piece of bandit information is available at iteration $\lceil k + d^i_k\rceil$. We impose that the delay time should grow at most sublinearly in the iteration k when the delays are homogeneous or be upper bounded by

some constant when the delays are heterogeneous, which is formally stated in the assumptions below.

Assumption 3: For each player i, the feedback delay d_k^i associated with the realized objective value $\hat{J}_k^i$ is a random variable and $d_k^i \in [0, \bar{d}(k)]$, where $\bar{d}(k) := k^{\alpha_d} + \bar{d}$, for some constants $\bar{d} \geq 0$ and $0 \leq \alpha_d < 1$.

Assumption 4: Either one of the following statements holds:

(i) the delay d_k^i is upper-bounded by a constant $\bar{d}$;
(ii) all the players experience the same delay, i.e., $d_k^1 = \cdots = d_k^N = d_k$.

The issue of handling delays that grow sublinearly or even superlinearly relative to a global clock is receiving increasing attention in the realm of distributed systems [28]. For example, in volunteer computing grids, the participation of new and faster workers in the network can undermine the performance of slower workers, causing their computation requests to accumulate quickly over time and resulting in growing delays.

C. Mirror Map and Mirror Descent

To streamline our subsequent discussion, we briefly introduce mirror descent and related concepts in this subsection. The interested readers are referred to [29, Ch. 4] for more detailed information. Let $\mathcal{B}$ denote a Banach space and $\mathcal{B}^*$ its dual. We first let $\psi : \operatorname{dom}\psi \to \mathbb{R}$ with $\operatorname{dom}\psi \subseteq \mathcal{B}$ denote a distance generating function (DGF). Here, $\operatorname{dom}\psi$ refers to the set where ψ is well-defined and is assumed to be convex and open. The DGF ψ satisfies: (i) ψ is differentiable and $\tilde{\mu}$-strongly convex for some $\tilde{\mu} > 0$; (ii) $\nabla\psi(\operatorname{dom}\psi) = \mathbb{R}^n$; (iii) $\operatorname{cl}(\operatorname{dom}\psi) \supseteq \mathcal{X}$ and $\lim_{x\to\partial(\operatorname{dom}\psi)} \|\nabla\psi(x)\|_* = +\infty$. With the DGF ψ in hand, the mirror map $\nabla\psi^*$ can be defined as:

$$\nabla\psi^*(z) = \operatorname{argmax}_{x\in\mathcal{X}}\{\langle z, x\rangle - \psi(x)\}, \tag{4}$$

which can be regarded as an extension of projection in general spaces. We let $D(\cdot,\cdot) : \mathcal{B}\times\mathcal{B} \to \mathbb{R}$ represent the Bregman divergence, whose formal expression is given by:

$$D(p,x) = \psi(p) - \psi(x) - \langle\nabla\psi(x), p - x\rangle, \forall p, x \in \operatorname{dom}\psi. \tag{5}$$

Assumption 5: (Bregman Reciprocity) The chosen DGF ψ satisfies that when the sequence $(x_k)_{k\in\mathbb{N}^+}$ converges to some point p, i.e., $\|x_k - p\| \to 0$, then $D(p, x_k) \to 0$.

The above assumption is introduced to enable the Bregman divergence $D(p,\cdot)$ to function as a specific distance metric with respect to p, thereby delineating a particular vicinity around p. The prox-mapping $P_{x,\mathcal{X}} : \mathcal{B}^* \to \operatorname{dom}\psi \cap \mathcal{X}$, induced by the Bregman divergence, is defined as:

$$P_{x,\mathcal{X}}(y) = \operatorname{argmin}_{x'\in\mathcal{X}}\{\langle y, x - x'\rangle + D(x', x)\}, \tag{6}$$

which plays an essential role in mirror descent and its variants. A lemma characterizing mirror maps and prox-mappings that will be frequently used in the subsequent analysis is given below.

Lemma 1: Consider the ambient Banach space $\mathcal{B}$ equipped with norm $\|\cdot\|$ and a closed and convex feasible set $\mathcal{X} \subseteq \operatorname{cl}(\operatorname{dom}\psi) \subseteq \mathcal{B}$. Suppose $\psi : \operatorname{dom}\psi \to \mathbb{R}$ is a DGF, then the mirror map $\nabla\psi^*$ is $1/\tilde{\mu}$-Lipschitz continuous, i.e., $\forall y_1, y_2 \in \mathcal{B}^*$, $\|\nabla\psi^*(y_1) - \nabla\psi^*(y_2)\| \leq (1/\tilde{\mu})\|y_1 - y_2\|_*$.

Proof: See [21, Lemma A.1]. ∎

To solve $\operatorname{VI}(\mathcal{X}, F)$, the mirror descent can be expressed as:

$$X_{k+1} = P_{X_k,\mathcal{X}}(-\gamma_k g_k) = \nabla\psi^*(\nabla\psi(x_k) - \gamma_k g_k), \tag{7}$$

where in the literature of stochastic VI, g_k usually denotes some noise-corrupted first-order information queried at X_k and γ_k an appropriate updating step size. One prevalent assumption is that there exists a first-order oracle to generate g_k after observing X_k, and given some proper filtration $(\mathcal{F}_k)_{k\in\mathbb{N}_+}$, it holds that $\mathbb{E}[g_k \mid \mathcal{F}_k] = F(X_k)$ and $\mathbb{E}[\|g_k\|_*^2 \mid \mathcal{F}_k]$ is a.s. bounded. The convergence properties of the actual sequences and the ergodic sequences have been extensively studied in [30], [31], [32].

III. BANDIT MIRROR DESCENT WITH FEEDBACK DELAYS

A. Residual Pseudo-Gradient Estimate

Our blanket assumption throughout is that the first-order oracle that returns g_k is unavailable, and each player can only observe its realized objective value associated with the action taken. To address the absence of first-order information, we leverage a pseudo-gradient estimate called the residual pseudo-gradient estimate (RPG) [21] to approximate the missing information from the observed objective values. At each iteration k, initially, it is necessary to undertake the following perturbation step:

$$\hat{X}_k^i = (1 - \frac{\delta_k}{r^i})X_k^i + \frac{\delta_k}{r^i}(p^i + r^i u_k^i) = \bar{X}_k^i + \delta_k u_k^i, \tag{8}$$

where u_k^i is randomly sampled from the unit sphere in the n^i−dimensional Euclidean space and we define $u_k := [u_k^i]_{i\in\mathcal{N}}$; δ_k represents the random query radius at iteration k; $\mathbb{B}(p^i, r^i) \subseteq \mathcal{X}^i$ is an arbitrary fixed ball within the feasible set $\mathcal{X}^i$ that centers at p^i with radius r^i; $\bar{X}_k^i := (1 - \delta_k/r^i)X_k^i + (\delta_k/r^i)p^i$. The RPG associated with the states X_k at k-th iteration leverages the realized objective values from the current iteration $\hat{J}_k^i := J^i(\hat{X}_k^i; \hat{X}_k^{-i})$ and the previous iteration $\hat{J}_{k-1}^i := J^i(\hat{X}_{k-1}^i; \hat{X}_{k-1}^{-i})$, which is formally given by

$$G_k^i := \frac{n^i}{\delta_k}(\hat{J}_k^i - \hat{J}_{k-1}^i)u_k^i. \tag{9}$$

To analyze the properties of RPG, a smoothed version for each local objective function J^i is leveraged:

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

where $\mathbb{S}_{-i} := \prod_{j\in\mathcal{N}^{-i}}\mathbb{S}_j \subseteq \mathbb{R}^{n^{-i}}$ with each $\mathbb{S}_j$ representing a unit sphere centered at the origin within $\mathbb{R}^{n^j}$; $\mathbb{B}_i$ denotes the unit ball centered at the origin inside $\mathbb{R}^{n^i}$; $\mathbb{V}_\delta^i := \operatorname{vol}(\delta\mathbb{B}_i)\cdot\operatorname{vol}(\delta\mathbb{S}_{-i})$ is the normalizing volume constant of the area that we are integrating over. One widely employed decomposition in the existing literature is that

$$G_k^i = \nabla_{x^i}J^i(X_k) + \big(G_k^i - \mathbb{E}[G_k^i \mid \mathcal{F}_k]\big) + \big(\mathbb{E}[G_k^i \mid \mathcal{F}_k] - \nabla_{x^i}J^i(X_k)\big),$$

where we let $B_k^i := \mathbb{E}[G_k^i \mid \mathcal{F}_k] - \nabla_{x^i}J^i(X_k)$ represent the systematic error and $V_k^i := G_k^i - \mathbb{E}[G_k^i \mid \mathcal{F}_k]$ the stochastic

error. Denote $B_k := [B_k^i]_{i\in\mathcal{N}}$ and $V_k := [V_k^i]_{i\in\mathcal{N}}$. Let $(\mathcal{F}_k)_{k\in\mathbb{N}_+}$ be the filtration concerning the random exploration factor, i.e., $\mathcal{F}_k := \sigma\{X_0, u_1, \ldots, u_{k-1}\}$. Then we have the following lemma to characterize the properties of B_k.

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

Proof: See the proof of [21, Lemma 1 & Lemma 2]. ∎

B. Feedback Utilization Strategy

The systematic error B_k^i and stochastic error V_k^i rooted in the estimate (9) make it inappropriate to merely leverage the most recent first-order estimate multiple times until a more recent one arrives as what is done in [22]; otherwise, the error will accumulate and endanger the convergence of the iterations. In view of this, we adopt the priority-based feedback utilization strategy: at each update, the first-order estimate with the earliest timestamp will be used and then discarded, similar to the approach employed in [24]. However, the single-point estimate strategy used in [24] mandates solely one realized function value, in which case it suffices to maintain a priority queue exclusively for these values. In contrast, the RPG adopted in this work requires two consecutive realized function values to obtain one estimate, which necessitates maintaining a cache to store observed function values and another priority queue for the resulting RPG estimates.

In our feedback utilization strategy, two information caches $\mathcal{P}_J^i$ and $\mathcal{P}_G^i$ are endowed for each player i. As reflected in (9), two subsequent objective values ($\hat{J}_k^i$ and $\hat{J}_{k-1}^i$) are a prerequisite to compute G_k^i, and it is possible that one arrives much earlier than the other. As such, cache $\mathcal{P}_J^i$ will store all the objective values received and pop out the ones that have been used twice in computing (9). For another thing, caused by the uncertainty in the feedback delay d_k^i, it is possible that at some iteration, player i has no available first-order estimates, while for some other iterations, multiple estimates are at player i's disposal. This motivates us to design $\mathcal{P}_G^i$ as a priority queue with the timestamp of each pseudo-gradient estimate as the key value. For notational convenience, we introduce a map $s^i : \mathbb{N}_+ \to \mathbb{N}_+$ that maps from the current iteration to the iteration where the first-order estimate originates from. When $\mathcal{P}_G^i$ is empty at iteration k, $s^i(k) = 1$ and the action remains unchanged. We also note that the map s^i is implicitly parameterized by the random sample $\omega \in \Omega$ and could vary across this group of players under Assumption 4 (i). To account for the heterogeneity in feedback delay $(d_k^i)_{i\in\mathcal{N}}$, we introduce a group iteration index map $s : \mathbb{N}_+ \to \mathbb{N}_+^N$, that projects from a certain iteration index k to the stack of originated indices $[s^i(k)]_{i\in\mathcal{N}}$.

Below, we present two lemmas that characterize the priority-based feedback utilization strategy, which our subsequent convergence analysis hinges upon. The proof is reported in [25, Appendix A].

Lemma 3: For each player i and arbitrary iteration $k \in \mathbb{N}_+$, we have the following:
(i) $K_\varnothing^i(k) := |\{s : \mathcal{P}_G^i = \varnothing, 1 \le s \le k\}| \le \min\{k, \bar{d}(k)+1\}$;
(ii) if $s^i(k) \ne 1$, then $s^i(k) + \bar{d}(s^i(k)) \ge k$.

C. The MD Algorithm with Feedback Delays

Algorithm 1 Bandit Learning with Reward Delays of CPs Based on Mirror Descent (Player i)

Initialize: $X_1^i \in \mathcal{X}^i \cap \operatorname{dom}\psi^i$ chosen arbitrary; take action $\hat{X}_1^i$ and $\hat{J}_1^i = J^i(X_1^i; X_1^{-i})$ will arrive $\lceil d_1^i \rceil$ iterations later; $G_1^i = \mathbf{0}_{n^i}$; 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 ($k \in \mathbb{N}_+$)
  Receive $\mathcal{R}_k^i := \{(t, \hat{J}_t^i, u_t^i) : k-1 < t + d_t^i \le k\}$
  $\mathcal{P}_J^i \leftarrow \mathcal{P}_J^i \cup \mathcal{R}_k^i$
  for $(t, \hat{J}_t^i, u_t^i) \in \mathcal{R}_k^i$ **do**
    if $(t+1, \hat{J}_{t+1}^i, u_{t+1}^i) \in \mathcal{P}_J^i$ **then**
      $G_{t+1}^i \leftarrow \frac{n^i}{\delta_{t+1}}(\hat{J}_{t+1}^i - \hat{J}_t^i)u_{t+1}^i$, $\mathcal{P}_G^i := \mathcal{P}_G^i \cup \{G_{t+1}^i\}$
    end if
    if $(t-1, \hat{J}_{t-1}^i, u_{t-1}^i) \in \mathcal{P}_J^i$ **then**
      $G_t^i \leftarrow \frac{n^i}{\delta_t}(\hat{J}_t^i - \hat{J}_{t-1}^i)u_t^i$, $\mathcal{P}_G^i := \mathcal{P}_G^i \cup \{G_t^i\}$
    end if
    $\mathcal{P}_J^i$ clears up the received feedback that has been utilized twice
  end for
  if $\mathcal{P}_G^i \ne \varnothing$ **then**
    $s^i(k) \leftarrow$ earliest index in $\mathcal{P}_G^i$, $\mathcal{P}_G^i \leftarrow \mathcal{P}_G^i \backslash \{G_{s^i(k)}^i\}$
  else
    $s^i(k) \leftarrow 1$ ▷ No update at this iteration
  end if
  $X_{k+1}^i \leftarrow P_{X_k^i, \mathcal{X}^i}(-\gamma_k G_{s^i(k)}^i)$
  Randomly sample the direction u_{k+1}^i from $\mathbb{S}_i$
  $\hat{X}_{k+1}^i \leftarrow (1 - \frac{\delta_{k+1}}{r^i})X_{k+1}^i + \frac{\delta_{k+1}}{r^i}(p^i + r^i u_{k+1}^i)$
  Take action $\hat{X}_{k+1}^i$ and the realized objective value $\hat{J}_{k+1}^i := J^i(\hat{X}_{k+1}^i; \hat{X}_{k+1}^{-i})$ will arrive $\lceil d_{k+1}^i \rceil$ iterations later
end procedure
Return: $\{\hat{X}_k^i\}_{i\in\mathcal{N}}$

The fusion of MD, RPG, and the priority-based feedback utilization strategy results in the proposed algorithm for bandit learning in continuous games with feedback delays, which is detailed in Algorithm 1. As has been discussed in [21], one prominent benefit we can reap from RPG is that the associated stochastic error V_k enjoys bounded variance if the decaying rate of step size is faster than that of query radius. It is worth mentioning that, Algorithm 1 leverages $\hat{G}_k = G_{s(k)} := [G_{s^i(k)}^i]_{i\in\mathcal{N}}$ rather than G_k to implement the action update at the k-th iteration, which is susceptible to the approximation errors stemming from bandit estimation and feedback delays. The existence of feedback delays then disrupts the recurrent relation characterizing $(\hat{G}_k)_{k\in\mathbb{N}_+}$, as a result of which, the analysis of the boundedness of the stochastic error and the estimates G_k^i in [21] cannot be directly carried over. To facilitate later analysis, we set $\hat{G}_1 = G_1 = G_0 = \mathbf{0}_n$ and $\hat{J}_0^i = \hat{J}_1^i$. In the lemma below, we will present

the sufficient condition to guarantee that the estimates $\hat{G}_k$ enjoy a uniform upper bound across $k \in \mathbb{N}_+$ and $\omega \in \Omega$. The proof is reported in [25, Appendix B].

Lemma 4: Suppose that Assumptions 1 and 3 hold. Moreover, step size $(\gamma_k)_{k\in\mathbb{N}_+}$ and query radius $(\delta_k)_{k\in\mathbb{N}_+}$ are monotonically decreasing and satisfy: $\lim_{k\to\infty}\gamma_k = 0$, $\sum_{k\in\mathbb{N}_+}\gamma_k = \infty$, $\lim_{k\to\infty}\delta_k = 0$, δ_k/δ_{k+1} is uniformly bounded for all $k \in \mathbb{N}_+$, $\lim_{k\to\infty}\gamma_k/\delta_k = 0$. Considering $(\hat{G}_k)_{k\in\mathbb{N}_+}$ generated by Algorithm 1, we have $\sup_{k\in\mathbb{N}_+}\|\hat{G}_k\|_* < \infty$.

For the feedback-delay scenario, the randomness originates from two sources: the random exploration factor at each iteration u_k^i and the feedback delay d_k^i associated with the realized objective value $\hat{J}_k^i$. Let the σ-field reflecting the delay information up to iteration k be denoted as:

$$\mathcal{F}_k^d \coloneqq \sigma\{d_t^i : \forall i \in \mathcal{N}, 1 \le t \le k\} \tag{11}$$

Note that $s^i(t) \in \mathcal{F}_k^d$ for all $1 \le t \le k$ and the available information respecting random exploration factors u_t^i depends on $\mathcal{F}_k^d$. Based on the observation, we are prompted to consider a more suitable σ-field $\tilde{\mathcal{F}}_{s(k)}$ for this specific problem, rather than the σ-field $\mathcal{F}_k$ previously discussed in Sec. III-A, which is defined as:

$$\tilde{\mathcal{F}}_k \coloneqq \sigma\big(\mathcal{F}_k^d \cup \{u^i_{s^i(t)} : \forall i \in \mathcal{N}, 1 \le t \le k-1\}\big). \tag{12}$$

With this definition in hand, we can then proceed to discuss the asymptotic convergence results for the actual sequence of play generated by Algorithm 1. The proof can be found in [25, Appendix C].

Theorem 1: Suppose the game $\mathcal{G}$ under consideration satisfies Assumptions 1 to 5 and all the players of $\mathcal{G}$ follow Algorithm 1 throughout the process. Moreover, the step size $(\gamma_k)_{k\in\mathbb{N}_+}$ and the query radius $(\delta_k)_{k\in\mathbb{N}_+}$ are chosen as $\gamma_k = \gamma_0/(k+K_\gamma)^{\alpha_\gamma}$ and $\delta_k = \delta_0/(k+K_\delta)^{\alpha_\delta}$, respectively. The selected parameters satisfy $0.5 < \alpha_\gamma \le 1, \alpha_\gamma > \alpha_\delta, \alpha_\gamma + \alpha_\delta > 1, 2\alpha_\gamma - \alpha_d > 1$. Then the actual sequence of play $(\hat{X}_k)_{k\in\mathbb{N}_+}$ converges to one of the CP x_* almost surely.

IV. Numerical Experiments

To illustrate the effectiveness of the proposed algorithm, we provide a numerical example of the thermal load management problem in a building complex. Suppose the load aggregator under study consisting of N buildings, indexed by $\mathcal{N} \coloneqq \{1,\dots,N\}$. Over a given time horizon $\mathcal{T} \coloneqq \{1,\dots,T\}$, we use x_t^i to represent the power consumption of building i at a certain time slot $t \in \mathcal{T}$. Moreover, the concatenations $x^i \coloneqq [x_t^i]$ and $x \coloneqq [x^i]$ denote the power profile of building i for all time slots and the energy profile of all buildings in this load aggregator, respectively. The internal pricing mechanism under consideration [2] discourages peak-demand usage by incorporating an approximate version of Shapley value, where each building i's share of peak demand is defined as $R^i(x) = \sum_{C_j: i\in C_j} \frac{(N-|C_j|)!(|C_j|-1)!}{N!}\Big(V(C_j,x) - V(C_j\backslash\{i\},x)\Big)$, where $C \coloneqq \{C_1,\dots C_{n_c}\}$ with each $C_j \subseteq \mathcal{N}(j=1,\dots,n_c)$ denotes the clique set; the function V is defined as $V(C_j,x) = \frac{1}{C}\log\Big(\sum_{t\in\mathcal{T}}\exp\big(\sum_{l\in C_j} Cx_t^l\big)\Big)$, where $C \in \mathbb{R}_{++}$ is a constant sufficiently large to make the log-sum-exp function a proper smooth approximation to the maximum function.

With knowledge of the power profile x^{-i} of other buildings, each building i seeks to find an optimal power control strategy, which can be expressed as follows:

$$\begin{aligned} &\text{minimize}_{x^i\in\mathcal{X}^i}\ (p_e)^T x^i + Q^i(x^i) + p_d \cdot R^i(x) \\ &\text{subject to } r_t^i = a^i r_{t-1}^i + b^i x_t^i,\ y_t^i = c^i r_t^i, \\ &\qquad \underline{y}_t^i \le y_t^i \le \bar{y}_t^i,\ 0 \le x_t^i \le \bar{x}^i, \forall t \in \mathcal{T}, \end{aligned} \tag{13}$$

where $p_e \in \mathbb{R}_{++}^T$ denotes the energy price and $p_d \in \mathbb{R}_{++}$ penalized the peak electricity usage of the aggregator; a strongly convex quadratic function Q^i is introduced for the convergence purpose; y_t^i denotes the temperature of building i at the t-th time slot and its dynamics are characterized by the first and second equality constraints; the third constraint enforces that the temperature y_t^i should be within a comfort zone $[\underline{y}_t^i, \bar{y}_t^i]$; the last constraint reflects the system power capacity for each building. It can be proved that this multi-player game admits a potential function $\Phi(x) = \sum_{i\in\mathcal{N}}\Big((p_e)^T x^i + Q^i(x^i)\Big) + p_d \cdot \sum_{C_j\in C}\frac{(N-|C_j|)!(|C_j|-1)!}{N!}V(C_j,x)$.

In the experiments, twenty buildings ($N = 20$) are involved in this game, and each building needs to determine its energy profile for four different time slots ($T = 4$). Suppose there are six cliques and the number of buildings within each clique ranges from three to eight. For $Q^i(x^i) = (x^i)^T \mathrm{diag}(\lambda_{i1},\dots,\lambda_{in^i})x^i$, each diagonal entry λ_{ij} is randomly sampled from $[0.04, 0.06]$. The query radius δ_k and the step size γ_k are set to be $\delta_k = 1/(k+10)^{0.6}$ and $\gamma_k = 1/(k+10^3)^{0.9}$, respectively. Regarding the feedback delay d_k^i, we consider the case when d_k^i is upper bounded by $\bar{d}_k = 10^3$ while the realized values of d_k^i vary across different buildings. In addition, several experiments are conducted under the setup that d_k^i is homogeneous in this group of buildings and grows sublinearly. To compare with the existing work, we implement the method in [24] with $\delta_k = 1/(k+10)^{0.35}$ and $\gamma_k = 1/(k+10^3)^{0.9}$ as required by the associated convergence theorem. Two metrics are employed to measure the performance of Algorithm 1, which include the relative distance between the NE and the perturbed actions, $\|\hat{X}_k - x_*\|_2/\|x_*\|_2$, and the difference between the potential function's optimal value and the values at the perturbed actions, $\Phi(\hat{X}_k) - \Phi_*$.

The numerical results are illustrated in Fig 1. It can be observed that when the feedback delay d_k^i grows no faster than $O(\sqrt{k})$, the convergence rates of the generated sequences are dominated by the first-order estimation error and no significant difference is noted among $\bar{d}_k = 10^3$, $d_k = k^{0.1}$, $d_k = 5k^{0.5}$, and $d_k = 10k^{0.5}$. When the delay time d_k^i grows faster and even approaches the rate of $O(k)$, the errors induced by the feedback delay outweigh those induced by the estimation error, as reflected in the curves associated with $d_k = 5k^{0.75}$ and $d_k = 5k^{0.99}$. Furthermore, the results in Fig. 1 indicate that Algorithm 1 exhibits reduced variance, more consistent sequences of play, and faster convergence compared to the existing method in [24].

V. Conclusion

This paper studies the problem of bandit learning in multi-player continuous games, which is further complicated

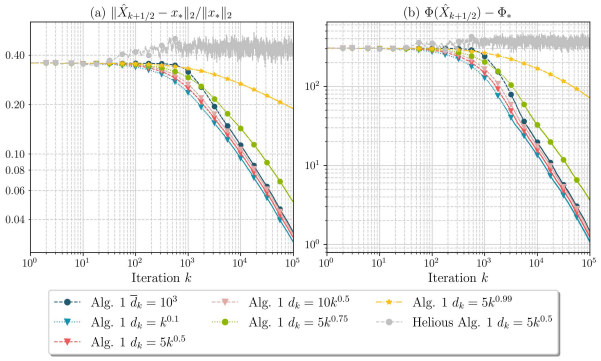

Fig. 1: Performance of the Proposed Algorithm Confronted with Homogeneous and Heterogeneous Feedback Delays

by information delays. Compared with the existing method introduced in [24], the algorithm proposed in this paper incorporates the residual pseudo-gradient estimation strategy and the mirror descent iteration, which loosens the conditions imposed upon the query radius and the step sizes. The a.s. convergence of the actual sequences of play generated by the proposed algorithm is established for pseudo-monotone plus games. One important direction for future research concerns the case where the feedback delays grow as the iteration proceeds and at the same time, they are heterogeneous across the participants. Another potential future direction resides in designing an algorithm that could tackle a more general class of multi-player games, such as merely monotone games, which are prevalent in the modeling of practical problems. Nevertheless, when applied to merely monotone games, mirror descent and most of its variants fail to converge and are prone to be trapped in spurious solutions.

References

[1] T. Li, G. Peng, Q. Zhu, and T. Başar, "The confluence of networks, games, and learning a game-theoretic framework for multiagent decision making over networks," *IEEE Control Systems Magazine*, vol. 42, no. 4, pp. 35–67, 2022.

[2] 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.

[3] A. Kannan, U. V. Shanbhag, and H. M. Kim, "Addressing supply-side risk in uncertain power markets: stochastic nash models, scalable algorithms and error analysis," *Optimization Methods and Software*, vol. 28, no. 5, pp. 1095–1138, 2013.

[4] Z. Zhou, P. Mertikopoulos, A. L. Moustakas, N. Bambos, and P. Glynn, "Robust power management via learning and game design," *Operations Research*, vol. 69, no. 1, pp. 331–345, 2021.

[5] A. Liniger and J. Lygeros, "A noncooperative game approach to autonomous racing," *IEEE Transactions on Control Systems Technology*, vol. 28, no. 3, pp. 884–897, 2019.

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

[7] 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.

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

[9] 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.

[10] S. Shalev-Shwartz *et al.*, "Online learning and online convex optimization," *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[11] D. Q. Vu, K. Antonakopoulos, and P. Mertikopoulos, "Fast routing under uncertainty: Adaptive learning in congestion games via exponential weights," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 708–14 720, 2021.

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

[13] A. Agarwal, O. Dekel, and L. Xiao, "Optimal algorithms for online convex optimization with multi-point bandit feedback." in *Colt*. Citeseer, 2010, pp. 28–40.

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

[19] 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.

[20] Y. Zhang, Y. Zhou, K. Ji, and M. M. Zavlanos, "A new one-point residual-feedback oracle for black-box learning and control," *Automatica*, vol. 136, p. 110006, 2022.

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

[22] ——, "On the convergence rates of a nash equilibrium seeking algorithm in potential games with information delays," *arXiv preprint arXiv:2209.12078*, 2022.

[23] Y. Zhang, R. Zhang, G. Li, Y. Gu, and N. Li, "Multi-agent reinforcement learning with reward delays," *arXiv preprint arXiv:2212.01441*, 2022.

[24] A. Héliou, P. Mertikopoulos, and Z. Zhou, "Gradient-free online learning in continuous games with delayed rewards," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 4172–4181.

[25] Y. Huang and J. Hu, "A bandit learning method for continuous games under feedback delays with residual pseudo-gradient estimate," *arXiv preprint arXiv:2303.16433*, 2023.

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

[27] 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.

[28] Z. Zhou, P. Mertikopoulos, N. Bambos, P. Glynn, and Y. Ye, "Distributed stochastic optimization with large delays," *Mathematics of Operations Research*, vol. 47, no. 3, pp. 2082–2111, 2022.

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

[30] 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.

[31] 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.

[32] A. Juditsky, J. Kwon, and É. Moulines, "Unifying mirror descent and dual averaging," *Mathematical Programming*, pp. 1–38, 2022.