Decentralized Nash Equilibrium Learning for Online Game With Bandit Feedback

Min Meng, Xiuxian Li, Senior Member, IEEE, and Jie Chen, Fellow, IEEE

Abstract—This article studies distributed online bandit learning of generalized Nash equilibria for online games, where the cost functions of all players and coupled constraints are time-varying. The function values, rather than full information about cost and local constraint functions, are revealed to local players with time delays. The goal of each player is to selfishly minimize its own cost function with no future information, subject to a strategy set constraint and time-varying coupled inequality constraints. To this end, a distributed online algorithm based on mirror descent and one-point delayed bandit feedback is designed for seeking generalized Nash equilibria in the online game. It is shown that the devised online algorithm achieves sublinear expected regrets and accumulated constraint violation if the path variation of the generalized Nash equilibrium sequence is sublinear. Simulations are presented to illustrate the efficiency of the theoretical result.

Index Terms—Distributed online learning, generalized Nash equilibrium, mirror descent, one-point delayed bandit feedback, online game.

I. INTRODUCTION

Multiplayer noncooperative games with self-interested decision-makers have found a remarkable breadth of applications, such as social networks [1], smart grid [2], sensor networks [3], and so on. A vital concept for this kind of game is Nash equilibrium (NE) [4], from which no player has an incentive to deviate. If the strategy set of each player depends on other players’ strategies, which often emerge in a wide range of real-world applications, e.g., limited resources among all players, then the NE is called a generalized NE (GNE).

Recently, distributed NEs and GNEs seeking in noncooperative games have received increasing attention. In contrast to centralized methods [5], [6], distributed NEs or GNEs seeking algorithms [7], [8], [9], without the need of a global coordinator bidirectionally communicating with all the players, only depend on partial players’ decisions instead of full actions’ information. As a result, distributed algorithms can reduce communication burden, increase robustness to link failures or malicious attacks, and preserve individual players’ private information to some extent.

By now, the above discussions have been on offline games, where cost and constraint functions are time-invariant. Nevertheless, the surrounding environments in various practical situations, such as real-time traffic networks, online auction, and allocation of radio resources, often change over time, incurring time-varying cost functions and/or constraints, which is usually called an online game [10]. In online games, the cost and constraint functions are revealed to local players only after making their decisions. In this setting, a distributed GNE learning algorithm for online games with time-invariant constraints was designed in [11] based on primal-dual strategies and distributed consensus. Subsequently, Meng et al. [12] considered time-varying constraints and proposed a distributed online algorithm on the basis of minimizer ascent and primal-dual strategies.

In [11] and [12], each player can access the gradient information for its local cost and local constraint functions sequentially. However, the gradient information in many realistic applications cannot be grabbed by local players. Instead, only function values exist, even with time-delays due to latency in communication and computation, are available to local players, i.e., bandit feedback, making online games more considerably challenging [13], [14].

In this article, distributed online learning for GNE of online games with time-varying coupled constraints and delayed bandit feedback is investigated. The main contributions are summarized as follows.

1) To the best of our knowledge, this article is the first to study online games with time-varying coupled inequality constraints and delayed bandit feedback, while existing works on online games only consider the case with full information on cost and constraint functions [11], [12] without delays. The delays here are assumed to be the same for all the players, which is reasonable in practice, such as in an auction, where bidders acquire their payoffs with an equal time delay.

2) A distributed GNE-seeking algorithm for online games is devised by mirror descent and one-point delayed bandit feedback. It is demonstrated that this algorithm can achieve sublinear expected regrets and accumulated constraint violation if the path variation of the GNE sequence is sublinear.

The rest of this article is organized as follows. In Section II, the problem formulation and some preliminaries are introduced. Section III proposes distributed bandit online algorithms for seeking GNEs with delayed rewards. A numerical example is presented to support the obtained results in Section IV. Finally, Section V concludes this article.

Notations: $\mathbb{R}$, $\mathbb{R}^n$, and $\mathbb{R}^{m \times n}$ represent the sets of real numbers, real column vectors of dimension $n$, and real matrices of dimension $m \times n$, respectively. $B_n$ and $S_n$ represent the unit ball and the sphere centered at the origin in $\mathbb{R}^n$, respectively. For a positive integer $m$, $[m] := \{1, 2, \ldots, m\}$. Denote by $1_n$ (resp. $0_n$) the $n$-dimensional vector with all elements being 1 (resp. 0). The transpose of a vector or matrix $P$ is denoted as $P^\top$. $\text{col}(z_1, \ldots, z_n) := (z_1, \ldots, z_n)^\top$. $A \otimes B$ is the Kronecker product of matrices $A$ and $B$. $(x, y)$ represents the inner product of vectors $x$ and $y$. $\col_\geq (z_1, \ldots, z_n)$ denotes the projection of $x$ onto $\mathbb{R}_+$, where $\mathbb{R}_+ := \{y \in \mathbb{R}^n \mid y \geq 0\}$. The symbol $h_1(\cdot)$ and $h_2(\cdot)$ means that...
there exists a constant $C$ such that $|h_1(x)| \leq C h_2(x)$ and $h_1 = o(T)$ connotes that $\lim_{T \to \infty} h_1(T) = 0$.

## II. PROBLEM FORMULATION AND PRELIMINARIES

### A. Problem Formulation

The online game with time-varying coupled inequality constraints studied in this article is denoted as $\Gamma(V, X, f, E)$. Where $f_{i,t}: X_i \to \mathbb{R}$ is the private time-varying cost function of player $i$ at time $t \geq 1$, $X_i = X_{0,i} \cap \{X_1, \ldots, X_N\}$ is the time-varying constraint set, where $0 \leq t \leq \mathbb{R} [g(x)] \subseteq \mathbb{R}^n$ is time-varying coupled inequality constraints and $X_i \subseteq \mathbb{R}^n$ implies the time-invariant private strategy set of player $i$. Here, $n := \sum_{i=1}^{N} n_i$ and $g_{i,t}: \mathbb{R}^{n_i} \to \mathbb{R}^m$ is the private constraint function of player $i$.

Full information on cost and constraint functions may be impossible or computationally heavy to be disclosed to local players. Hence, this article focuses on the case where only delayed function values, $f_{i,t}\{x_{i,t}, x_{i,t-1}, \ldots, x_{i,0}\} \in X_i$, are revealed privately to player $i$ after a strategy profile $x_{i} := (x_{i,1}, \ldots, x_{i,K}) \in X_i$ is chosen, where $x_{i,t} \in X_i$ is the strategy selected by player $i$ at time $t$, $i, t \in E$ is called a neighbor of $i$. Denote by $\mathcal{N}_i := \{j : \{j, i\} \in E\}$ the set of neighbors of node $i$. A path from node $a$ to node $b$ is a sequence of edges $(a_i, b_i), h = 1, 2, \ldots, h - 1$, is a sequence of distinct vertices. An undirected graph $\mathcal{G}$ is said to be connected if there is a path between any two nodes. For this online game $\Gamma(V, X, f, E)$, it is impossible for agents to precompute a GNE $(x_{i,1}^*, x_{i,2}^*, \ldots, x_{i,K}^*)$ at every time $t$, i.e., a strategy profile $(x_{i,1}^*, x_{i,2}^*, \ldots, x_{i,K}^*)$ satisfying

$$f_{i,t}(x_{i,t}, x_{i,t}, \ldots, x_{i,t}) \leq f_{i,t}(x_{i,t}, x_{i,t}, \ldots, x_{i,t}) \quad \forall x_{i} \in X_i(x_{i,t})$$

(1)

where $X_{i,t}(x_{i,t}) := \{x_{i,t} \in X_i | (x_{i,t}, x_{i,t}) \in X_i\}$, but agents can improve their performance by learning through play. Note that (1) is equivalent to the following optimization problem:

$$\min_{x_{i,t}} f_{i,t}(x_{i,t}, x_{i,t}, \ldots, x_{i,t}) \quad \text{subject to} \quad x_{i,t} \in X_i(x_{i,t}).$$

(2)

Then in this article, based on (2), an important performance measure called dynamic regret, is adopted, defined as follows [11]:

$$\text{Reg}_T(T) := \sum_{t=1}^{T} (f_{i,t}(x_{i,t}, x_{i,t}, \ldots, x_{i,t}) - f_{i,t}(x_{i,t}^*, x_{i,t}, \ldots, x_{i,t}^*))$$

(3)

where $T$ is the total learning time horizon.

In the meantime, to measure the violation of constraints, the commonly used constraint violation measure is presented as

$$R_{\sigma}(T) := \left\| \sum_{t=1}^{T} g_{i,t}(x_i) \right\|.$$

(4)

The metrics in (3) and (4) provide a meaningful method for quantifying the ability of an online algorithm to adapt to unknown and unpredictable environments. From this viewpoint, an online algorithm is “no-regret” if $\text{Reg}_T(T)$ and $R_{\sigma}(T)$ sublinear with respect to $T$, i.e., $\text{Reg}_T(T) = o(T)$, $i \in [N]$, and $R_{\sigma}(T) = o(T)$.

However, if the GNE sequence $\{x_{i,t}\}_{t=1}^{T}$ fluctuates drastically, then learning the exact GNE may be impossible, as studied in online optimization [15]. To depict the fluctuation of the GNE sequence $\{x_{i,t}\}$, the path variation (or path length), $\Phi_T$, is used, usually defined, as [11, 16]:

$$\Phi_T := \sum_{t=1}^{T} \| x_{i,t+1}^* - x_{i,t} \|.$$
Bregman divergence is widely applied in machine learning and game theory, generalizing the standard Euclidean distance. Specifically, when $\phi_\theta (\cdot) = \| \cdot \|^2$, Bregman divergence $D_{\phi_\theta} (\theta, \vartheta)$ amounts to $|\theta - \vartheta|^2$. Two mild assumptions on Bregman divergence are made.

**Assumption 5:** $D_{\phi_\theta} (\cdot, \cdot)$ is Lipschitz with respect to the first variable, i.e., there is a constant $K > 0$ such that for any $\theta_1, \theta_2 \in X_1$, $|D_{\phi_\theta} (\theta_1, \vartheta) - D_{\phi_\theta} (\theta_2, \vartheta)| \leq K |\theta_1 - \theta_2|$. Assumption 6 is that $D_{\phi_\theta} (\cdot, \cdot)$ is convex with the second variable, i.e., for any $\alpha \in [0, 1]$ and $\theta_1, \theta_2 \in X_1$, $\alpha D_{\phi_\theta} (\theta_1, \vartheta) + (1-\alpha) D_{\phi_\theta} (\theta_2, \vartheta)$ is positive semidefinite for $\theta, \vartheta \in X$, where $H_{\phi_\theta}$ denotes the Hessian matrix of $\phi_\theta$ [21]. Assumptions 5 and 6 play an essential role in deriving the bounds of the regrets and accumulated constraint violation.

**C. Gradient Approximation**

Let $f : X \to \mathbb{R}$ be a function, where the domain $X \subseteq \mathbb{R}^n$ is a convex and bounded set, and has a nonempty interior. Assume that $X$ is contained in the ball $R(X)^B$ and contains the ball $r(X)^B$, that is, $r(X)^B \subseteq X \subseteq R(X)^B$, where $r(X)$ and $R(X)$ are positive constants. A one-point gradient estimator is proposed as [22]

$$\nabla f(x) = \frac{1}{\delta} f(x + \delta u) \quad \forall \vartheta \in (1 - \eta)X \quad (11)$$

where $u \in \mathbb{S}^n$ is a uniformly distributed random vector, $\eta \in (0, 1)$ is a shrinkage coefficient and $\delta \in (0, r(X)^B)$ is an exploration parameter. It should be noted that the perturbations in (11) can be ensured to still remain in the set $X$ by defining the estimator (11) over the shrinking set $(1 - \eta)X$ instead of $X$. Let us define a smooth version of $f$ as $\hat{f} (x) := E_{u \in \mathbb{S}^n} [f (x + \delta v)]$, where $v$ is selected uniformly at random from the unit ball $\mathbb{B}^n$.

**Lemma 1 (See [23]):**

1. It holds that $x + \delta u \in X$ for any $x \in (1 - \eta)X$, $u \in \mathbb{S}^n$ and $\delta \in (0, r(X)^B]$.  
2. $\hat{f} (x)$ is differentiable on $(1 - \eta)X$ even when $f(x)$ is not, and it holds that $\nabla \hat{f} (x) = E_{u \in \mathbb{S}^n} [\nabla f (x)]$, $\forall \vartheta \in (1 - \eta)X$.  
3. $\hat{f} (x)$ is convex on $(1 - \eta)X$ if $f(x)$ is convex on $X$, and $f(x) \leq \hat{f} (x)$ for any $x \in (1 - \eta)X$.  
4. If $f(x)$ is $l_0$-Lipschitz on the set $X$, then $\hat{f}$ and $\nabla \hat{f}$ are $l_0$-Lipschitz and $\eta l_0 / \delta$-Lipschitz, respectively. Additionally, there holds that $|f(x) - \hat{f}(x)| \leq l_0 \vartheta$, $\forall \vartheta \in (1 - \eta)X$.  
5. If $|f(x)| \leq F_0$ for any $x \in X$, then $|f(x)| \leq F_0$, and $\|\nabla \hat{f} (x)\| \leq n F_0 / \delta$, $\forall \vartheta \in (1 - \eta)X$.  

In fact, replacing the original function $f$ by the function $\hat{f}$ is the critical idea of this gradient-free method since it is seen from 4) of Lemma 1 that the function $f(x)$ is extremely close to $\hat{f}(x)$ when $\delta$ is small. In addition, $\nabla \hat{f}$ can be regarded as an unbiased estimator of $\nabla \hat{f}$ from 2) of Lemma 1.

**III. DISTRIBUTED BANDIT FEEDBACK WITH DELAYS**

In this section, a distributed online algorithm for tracking variational GNE sequence of the studied online game is proposed based on delayed bandit feedback and mirror descent. We also analyze the expected regrets and constraint violation bounds for the proposed algorithm.

A distributed bandit online algorithm is given in Algorithm 1, where each player $i \in [N]$ maintains four local variables: the local strategy variable $x_{i,t} \in X_i$ for each $t \geq 0$, and the local dual strategy variables $z_{i,t} \in (1 - \eta_t)X_i$, and the local dual variable $\lambda_{i,t} \in \mathbb{R}^n_{\geq 0}$ for each $t \in [1, T]$. These variables are updated based on (12a)–(12d).

**Algorithm 1:** Distributed Bandit Online Primal-Dual Mirror Descent With One-point Delayed Bandit Feedback.

Each player $i$ maintains vector variables $x_{i,t} \in X_i$, $z_{i,t} \in (1 - \eta_t)X_i$, and $\lambda_{i,t} \in \mathbb{R}^n_{\geq 0}$ at iteration $t$.

**Initialization:** For any $i \in [N]$, initialize $z_{i,t} = \lambda_{i,t} = 0$, $t = 1, \ldots, \tau$.

**At each $t > \tau,$** every player $i$ receives the delayed function values $f_{i,t-r}(x_{i,t-r})$ and $g_{i,t-r}(x_{i,t-r})$, and process the following update:

$$\begin{align*}
\tilde{z}_{i,t+1} &= \arg \min_{z \in (1 - \eta_t)X_i} \left\{ \alpha_i(z, \nabla f_{i,t-r}(x_{i,t-r})) + \alpha_i(z, \nabla g_{i,t-r}(x_{i,t-r})) \right\} \\
&+ \alpha_i(z, \nabla g_{i,t-r}(x_{i,t-r})) \hat{\lambda}_{i,t} + \delta_{i,t} u_{i,t+1}
\end{align*}$$

where $z_{i,t} := \text{col}(z_{i,t-1}, \ldots, z_{i,T})$, $\nabla f_{i,t-r}(z_{i,t-r}) := \frac{1}{\delta} f_{i,t-r}(x_{i,t-r})$, $u_{i,t+1} \nabla g_{i,t-r}(z_{i,t-r}) := \frac{1}{\delta} g_{i,t-r}(x_{i,t-r})$, $\alpha_i \in \left[0, 1\right]$, $\delta_i \in [0, 1]$ are nonincreasing sequences to be determined, and $u_{i,t+1} \in \mathbb{S}^n_i$, $i \in [N]$, are uniformly distributed random vectors.

The intuition of the update rules in (12a)–(12d) is explained as follows. A regularized/penalized Lagrangian function at time $t$ associated to each player $i \in [N]$ is defined as

$$A_{i,t}(x_{i,t}, \lambda_{i,t}; x_{-i,t}) := f_{i,t}(x_{i,t}, x_{-i,t}) + \lambda_{i,t} g_{i,t}(x_{i,t}) - \beta_2 \| \lambda_{i,t} \|^2$$

where $x_{i,t} \in X_i$, and $\lambda_{i,t} \in \mathbb{R}^n_{\geq 0}$ is the Lagrange multiplier or dual variable, $\beta_2 > 0$ is the regularization parameter. Based on (13), a mirror-descent-based algorithm is designed as

$$\begin{align*}
x_{i,t+1} &= \arg \min_{z \in X_i} \left\{ \alpha_i(x, \nabla f_{i,t-r}(x_{i,t-r})) + \alpha_i(x, \nabla g_{i,t-r}(x_{i,t-r})) \right\} \\
&+ \alpha_i(x, \nabla g_{i,t-r}(x_{i,t-r})) \hat{\lambda}_{i,t} + D_{\phi_\theta}(x_i, x_{i,t})
\end{align*}$$

where $\alpha_i > 0$ and $\eta_t \in (0, 1)$ are step sizes used in the primal and dual updates, respectively. In (14) and (15), to update the strategy $x_{i,t}$ of player $i$ needs the information of all players’ strategies, the gradients of $f_{i,t}(x_{i,t}, x_{-i,t})$ and $g_{i,t}(x_{i,t})$, and the global nonlinear constraint function $g_i(x)$, which is always not available to any single player if there does not exist a central coordinator. In contrast, Algorithm 1 is fully distributed and gradient-free, where player $i$ only can receive delayed values of the cost and constraint functions, $f_{i,t-r}(x_{i,t-r}, x_{-i,t-r})$ and $g_{i,t-r}(x_{i,t-r})$, at time $t$, where the integer $\tau \geq 0$ denotes the same delay for all the players.

In Algorithm 1, $u_{i,t}$ is randomly chosen following a uniform distribution. Let $\tilde{\delta}_t$ represent the $\sigma$-algorithm generated by $\{u_{1,t}, \ldots, u_{N,t}\}$, and $F_t := \bigcup_{t=1}^{\infty} \tilde{\delta}_t$. It can be seen that the random sequences $z_{i,t} \in (1 - \eta_t)X_i$, $\lambda_{i,t} \in \mathbb{R}^n_{\geq 0}$ generated by Algorithm 1 are independent of $\tilde{\delta}_t$ for $s > t - \tau$ and depend on $F_s$. All inequalities involving random variables in this article are satisfied almost surely (a.s.) if not stated explicitly.

In what follows, some necessary lemmas are first presented.

**Lemma 2:** Under Assumptions 1 and 2, for any $i \in [N]$ and $t > \tau$, $\lambda_{i,t}$ and $\hat{\lambda}_{i,t}$ generated by Algorithm 1 satisfy

$$\begin{align*}
\|x_{i,t}\| &\leq B_i \beta_{t-1}^{-1} \\
\|\lambda_{i,t}\| &\leq B_i \beta_{t-1}^{-1}
\end{align*}$$
\[ \| \hat{\lambda}_{i,t} - \lambda_t \| \leq 2\sqrt{N} B_\eta \sum_{s=0}^{t-r-1} \sigma_m \gamma_{t-1-s} \]  
(18) 
\[ \frac{A_{t+1}}{2\tau} \leq 2N^2 B_\eta^2 (\hat{\lambda}_t - \lambda)^g g_t, (x_t - r) + 0.5N \beta_t \| \lambda \|_2^2 + 2N \sqrt{N} B_\eta^2 \sum_{s=0}^{t-r-1} \sigma_m \gamma_{t-1-s} \]  
(19) 
where \( \lambda \in \mathbb{R}_+^n \), \( \tau_t := \sum_{i=1}^{N} \lambda_{i,t}/N \), and \( A_{t+1} := \sum_{i=1}^{N} \| \lambda_{i,t+1} - \lambda^* \|^2 - (1 - \beta_0 \gamma_t) \lambda_{i,t} - \lambda^* \|^2. \)  

**Proof:** See Appendix A. 

**Lemma 3:** Under Assumptions 1–6, for all \( i \in [N] \), \( z_{i,t} \) generated by Algorithm 1 satisfies 
\[ \mu \sum_{t=1}^{T} \mathbb{E} [ \| \lambda_t^* - z_t \|^2 ] \leq \sum_{t=1}^{T} \frac{1}{\alpha_{t+1}} \mathbb{E} \left[ D_{\beta}(\hat{\lambda}_t, z_{i,t}) - D_{\beta}(\hat{\lambda}_t, x_{i,t+\tau+1}, z_{i,t+\tau+1}) \right] \]  
+ \[ \sum_{t=1}^{T} \frac{K_T}{\alpha_{t+1}} \mathbb{E} \left[ \| \hat{\lambda}_{i,t+\tau+1} - \lambda_t \|^2 + C_1 \tau + C_2 \tau \right. \]  
(20) 
where \( \lambda_t^* := (1 - \eta) x_t, \lambda_t^* := \text{col}(\lambda_{i,t}^*, \ldots, \lambda_{n,t}^*), \lambda^* := \sup_{t \in [1, T]} \mathbb{E} [ x_t ] \), \( \| \lambda_t \|, \lambda_t \| := \text{arg max}_{x \in \mathbb{R}^n_t} \{ D_{\beta}(x, x_t^*) + \gamma_t \lambda^* g_t(x_t) \} \), and 
\[ C_{1, \tau} := 2N \sqrt{N} B \lambda \sum_{t=1}^{T} \delta_{t+\tau} + 2NB_L \lambda \sum_{t=1}^{T} \frac{\delta_{t+\tau}}{\alpha_{t+1}} + 2N(B_L \lambda + B_\eta^2) \sum_{t=1}^{T} \eta_t + \frac{\eta_t^2 B_\lambda^2}{\mu_0} \sum_{t=1}^{T} \alpha_{t+\tau} \delta_{t+\tau}^2 \]  
+ \[ N \left( 2\sqrt{N} B_\lambda^2 L + B_L \lambda \right) \sum_{t=1}^{T} \eta_t + \frac{\eta_t^2 B_\lambda^2}{\mu_0} \sum_{t=1}^{T} \alpha_{t+\tau} \delta_{t+\tau}^2 \]  
+ \[ N B_L \lambda L \sum_{t=1}^{T} \frac{\eta_t}{\alpha_{t+\tau}} + 6N \sqrt{N} B_\lambda^2 \sum_{t=1}^{T} \sum_{s=0}^{t-r-1} \sigma_m \gamma_{t-1-s} \]  
(21) 
\[ C_{2, \tau} := \sum_{t=1}^{T} \| g_t(x_t) \| + 2 \left( \frac{1 + (1 - \sum_{t=1}^{T} \beta_{t+\tau})}{2} \right) \| \lambda_t \|^2 \]  
+ \[ \sum_{t=1}^{T} \left( \frac{1}{2\tau} - \frac{1}{2\tau} - \frac{\beta_{t+\tau}}{2} \right) \sum_{t=1}^{T} \| \lambda_{i,t+\tau} - \lambda_t \|^2 \]  

**Proof:** See Appendix B. 

**Lemma 4:** Under Assumptions 2–6, for any \( i \in [N] \), the expected dynamic regret (3) and constraint violation (4) generated by Algorithm 1 is bounded by 
\[ \mathbb{E} [ \text{Reg}_r(T) ] \leq \frac{L_f}{\sqrt{T}} \sqrt{TC_{2, T} + T \tau_{r+1}} + \frac{\sqrt{N} \tau}{\alpha_{T+r}} (\tau + 1) \Phi_{T+r}^* \]  
(22) 
where \( \lambda_{r+1} := 2C_{2, T} \mathbb{E} \left[ \sum_{t=1}^{T} g_t(x_t) \right] \| \lambda_t \|^2 \) and 
\[ C_{1, T} := \frac{\sqrt{N} \tau}{\alpha_{T+r}} (\tau + 1) \Phi_{T+r}^* \]  
(23) 
\[ C_{2, T} := \frac{\sqrt{N} \tau}{\alpha_{T+r}} (\tau + 1) \Phi_{T+r}^* \]  
**Corollary 1:** Under Assumptions 2–6, there hold for Algorithm 1 that the expected dynamic regrets and accumulated constraint violation are sublinearly bounded if both \( (\tau + 1) \log T \) and \( (\tau + 1) \Phi_{T+r}^* \) are sublinear with respect to \( T \).  
**Proof:** If \( \Phi_T = O(T^\alpha) \), let \( a_1 \in [0, \frac{1}{2} - \frac{\alpha}{2}] \) and choose \( a_1, a_2, a_3 \) satisfying the conditions in Theorem 1, then \( \mathbb{E} [ \text{Reg}_r(T) ] / T \to 0 \), and \( \mathbb{E} [ \text{Reg}_r(T) ] / T \to 0 \) as \( T \) goes to infinity. 

**Corollary 2:** Under Assumptions 2–6, the bounds of the expected dynamic regrets and accumulated constraint violation generated by Algorithm 1 can be explicitly given as 
\[ \mathbb{E} [ \text{Reg}_r(T) ] \leq \frac{\sqrt{\tau + 1} T^{\frac{3}{2}}}{\sqrt{T}} + O \left( T^{\frac{3}{2}} \Phi_{T+r}^* \right) \]  
(24) 
\[ \mathbb{E} [ \text{Reg}_s(T) ] \leq \frac{\sqrt{\tau + 1} T^{\frac{3}{2}}}{\sqrt{T}} + O \left( T^{\frac{3}{2}} \Phi_{T+r}^* \right). \]  

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
of firm $i$ is written as $f_{i,t}(x_{i,t}, x_{-i,t}) = p_{i,t}(x_{i,t}) - x_{i,t}d_{i,t}(x_{i,t})$ with the production cost being $p_{i,t}(x_{i,t}) = x_{i,t}(\sin(t/12) + 1)$ and the demand price of firm $i$ being $d_{i,t}(x_{i,t}) = 22 + i/9 - 0.5i\sin(t/12) - \sum_{t=1}^{N} x_{i,t}$. The market capacity constraint is considered as the shared inequality constraint $\sum_{t=1}^{N} x_{i,t} \leq \sum_{t=1}^{N} l_{i,t}$, where $l_{i,t} = 10 + \sin(t/12)$ is the local bound only available to firm $i$. If considering the offline and centralized setting, the time-varying GNE can be computed as $x_{i,t}^* = P_{X_i}(\xi_t)$, where $\xi_t := \frac{1}{2}(i - 1) + (5 - 1/21 - i/2)\sin \frac{t}{2}$. In the distributed and delayed bandit feedback setting, choose $a_1 = 0.45$, $a_2 = 0.1$, $a_3 = 0.11$, and set initial states $z_{i,1} \in (1 - \eta_1)X_i$ randomly, and $\lambda_{i,1} = 0_{{m}}$, then by Algorithm 1, the averages of dynamic regrets and constraint violation with different delays are shown in Figs. 1 and 2, respectively, from which one can see that Algorithm 1 achieves sublinear expected regrets and constraint violation. Moreover, feedback delay makes the convergence of the averages of dynamic regrets and constraint violation a little bit slower than the delayless case.

V. CONCLUSION

In this article, distributed online learning of GNEs in online games was studied based on bandit feedback and mirror descent. Each player in an online game aims to minimize its own self-interested time-varying cost function subject to time-varying coupled inequality constraints with no information of gradients of cost and constraint functions. To this end, distributed online algorithms based on mirror descent and one-point bandit feedback were designed for online games with feedback delays, which can achieve sublinear dynamic regrets and accumulated constraint violations if some quantities are sublinear. Future research directions can be placed on developing a method to improve the regret bounds, to consider the scenario of bandit feedback with different time-delays for different agents and to study the case with communication delays when transmitting information among agents, besides delayed bandit feedback.

APPENDIX

A. Proof of Lemma 2

Equation (16) and (17) can be proved by mathematical induction. Specifically, note that $\|\lambda_{i,t+1}\| = 0 \leq B_\beta/\beta_{t+1}$ and $\|\hat{\lambda}_{i,t+1}\| = 0 \leq B_\beta/\beta_{t+1}$. Assume that $\|\lambda_{j,t}\| \leq B_\beta/\beta_t$ and $\|\hat{\lambda}_{j,t}\| \leq B_\beta/\beta_t$ for $t > \tau$. Then, one has from (12d) that

$$
\|\lambda_{i,t+1}\| \leq \|\lambda_{i,t} + \gamma_t(g_{i,t}(x_{i,t} - \beta_t\hat{\lambda}_{i,t}))\|

\leq \|\lambda_{i,t}\| + \gamma_t\|g_{i,t}(x_{i,t} - \beta_t\hat{\lambda}_{i,t})\|

\leq (1 - \beta_t\gamma_t)\|\lambda_{i,t}\| + \gamma_t\|g_{i,t}(x_{i,t} - \beta_t\hat{\lambda}_{i,t})\|

\leq \frac{B_\beta}{\beta_{t+1}}
$$

where the first inequality is derived based on $\|a\| \leq \|a\|$ for any vector $a$, and the last inequality is obtained by inductive hypothesis, the boundedness of $g_{i,t}(x_{i,t})$ in (5) and $\beta_t \geq \beta_{t+1}$. Subsequently, by the convexity of norms and $\sum_{t=1}^{\infty} a_t = 1$, it is obtained that $\|\hat{\lambda}_{i,t+1}\| \leq \sum_{t=1}^{\infty} \lambda_{i,t} \|\hat{\lambda}_{i,t}\| \leq B_\beta/\beta_{t+1}$. Therefore, (16) and (17) hold.

Let us now prove (18). From (12d), one has

$$
\lambda_{i,t+1} = \hat{\lambda}_{i,t} + \epsilon_{i,t}
$$

where $e_{i,t} := \lambda_{i,t} + \gamma_t(g_{i,t}(x_{i,t} - \beta_t\hat{\lambda}_{i,t})) + \lambda_{i,t} - \hat{\lambda}_{i,t}, i \in [N]$. Let $\lambda_{i,t} := (\lambda_{1,t}, \ldots, \lambda_{N,t})$, $\lambda_{i,t} := col(\lambda_{1,t}, \ldots, \lambda_{N,t})$, and $\epsilon_{i,t} = \epsilon_1, \ldots, \epsilon_{N,t}$. Then, the iteration of $\lambda_{i,t}$ is given as

$$
\hat{\lambda}_{i,t+1} = \left(\left(A - \frac{1}{N}I_N\right) \otimes I_m\right) \hat{\lambda}_t + \left(\left(I - \frac{1}{N}I_N\right) \otimes I_m\right) \epsilon_{i,t}
$$

(30)
Under Assumption 1, one has $0 \leq \sigma_m < 1$, where $\sigma_m = \|A - \frac{1}{\eta}B\|$. Taking norm on both sides of (30) leads to that $\|\hat{\lambda}_{t+1}\| \leq \sigma_m \|\hat{\lambda}_t\| + \|\xi_t\|$. Note that $\|\hat{\xi}_t\| = \|\hat{\lambda}_t + \gamma_1 (g_t(x_{t-1}) - \beta \hat{\lambda}_t)\| +\|\hat{\lambda}_t\|$. Therefore $\|\hat{\xi}_t\| \leq \gamma_1 (g_t(x_{t-1}) - \beta \hat{\lambda}_t) + \|\hat{\lambda}_t\|$. Thus, (4) is proved.

As for (19), it can be derived from (12d) that for $\lambda \in \mathbb{R}^n$,

$$
\|\hat{\lambda}_{i,t+1} - \lambda\|^2 \geq \|\hat{\lambda}_{i,t} + \gamma_1 (g_t(x_{t-1}) - \beta \hat{\lambda}_t)\|^2 \geq \|\hat{\lambda}_{i,t} - \lambda\|^2 + 2\gamma \|\hat{\lambda}_t\| - \|\hat{\lambda}_t\|^2 - \|\hat{\lambda}_t\|^2 = \|\hat{\lambda}_{i,t} - \lambda\|^2 + 4\gamma^2 \|\hat{\lambda}_t\|^2
$$

(31)

where the last inequality is based on (5) and (16).

For the last term in (31)

$$
2\gamma (\hat{\lambda}_{i,t} - \lambda)^\top (g_t(x_{t-1}) - \beta \hat{\lambda}_t)
$$

$$= 2\gamma (\hat{\lambda}_{i,t} - \lambda)^\top g_t(x_{t-1}) - 2\beta \gamma (\hat{\lambda}_{i,t} - \lambda)^\top (\hat{\lambda}_{i,t} - \lambda)
$$

$$\leq 2\gamma \|\hat{\lambda}_{i,t} - \lambda\|^2 + 2\gamma^2 \|\hat{\lambda}_t\|^2 + \beta^2 \gamma \|\hat{\lambda}_t\|^2 - \|\hat{\lambda}_t\|^2
$$

(32)

Substituting (32) into (31), summing over $i \in [N]$ on both sides of (31), and then a simple computation leads to (19).

**Proof of Lemma 3**

For any $t > 0$ and $i \in [N]$, it can be derived based on the optimality of $\hat{z}_{i,t+1} + 1$ in (12a) that for any $z \in (1 - \eta_t)X_i$

$$
\langle \hat{z}_{i,t+1} - z, \alpha_{i,t+1} \nabla f_{i,t}(z) + (\nabla g_{i,t}(z_{i,t+1}))\hat{\lambda}_{i,t+1} \rangle
$$

(33)

Taking $z = \hat{z}_{i,t}$ in (33) yields that

$$
\alpha_{i,t+1} \langle \hat{z}_{i,t} - z_{i,t}, \nabla f_{i,t}(z_{i,t}) + (\nabla g_{i,t}(z_{i,t}))\hat{\lambda}_{i,t}\rangle
$$

(34)

where (8) is applied to obtain the last inequality. Then, one has

$$
D_v(\hat{z}_{i,t}, z_{i,t+1})
$$

(35)

Taking expectation on both sides of (35) yields

$$
\mathbb{E}_{u_{i,t}}[D_v(\hat{z}_{i,t}, z_{i,t+1})] = \mathbb{E}_{u_{i,t}}[D_v(\hat{z}_{i,t}, z_{i,t})] + \mathbb{E}_{u_{i,t}}[\nabla f_{i,t}(z_{i,t})\hat{\lambda}_{i,t}]
$$

(36)

where the equality holds since $z_{i,t}$ is independent of $u_{i,t}$ and $\mathbb{E}_v(\hat{f}_{i,t}(z_{i,t})) = \nabla f_{i,t}(z_{i,t})$ by Lemma 1 with $\hat{f}_{i,t}(z_{i,t}) := \mathbb{E}_{u_{i,t}}[f_{i,t}(z_{i,t}) + \delta u_{i,t}, x_{i,t}]]$. Next, we discuss the bound of each term on the right-hand side of (36).

For the first term on the right-hand side of (36), one has that

$$
\alpha_{i,t+1} \langle \hat{z}_{i,t} - z_{i,t}, \nabla f_{i,t}(z_{i,t}) - \nabla f_{i,t}(x_{i,t})\rangle = \alpha_{i,t+1} \langle \hat{z}_{i,t} - z_{i,t}, \nabla f_{i,t}(z_{i,t}) - \nabla f_{i,t}(x_{i,t})\rangle
$$

(37)

For the second term on the right-hand side of (36), it is obtained that

$$
\mathbb{E}_{u_{i,t}}[\nabla f_{i,t}(z_{i,t}) - \nabla f_{i,t}(x_{i,t})] = \mathbb{E}_{u_{i,t}}[\nabla f_{i,t}(z_{i,t}) - \nabla f_{i,t}(x_{i,t})]
$$

(38)

where $z_{i,t}$ is col($z_{1,t}, \ldots, z_{N,t}$). The equality is based on the Leibniz integral rule, the first inequality is based on the Cauchy-Schwarz inequality and (5), and the second inequality holds by Assumption 3.

For the third term on the right-hand side of (37), one has that

$$
\alpha_{i,t+1} \langle \hat{z}_{i,t} - z_{i,t}, \nabla f_{i,t}(x_{i,t}) - \nabla f_{i,t}(z_{i,t})\rangle
$$

(39)

For the fourth term on the right-hand side of (37), it is derived that

$$
\alpha_{i,t+1} \langle \hat{z}_{i,t} - z_{i,t}, \nabla f_{i,t}(x_{i,t}) - \nabla f_{i,t}(z_{i,t})\rangle
$$

(40)

where the inequality is derived based on the optimality of $(\hat{z}_{i,t}, \hat{\lambda}_{i,t})$ and (5). Substituting (38), (39), and (40) into (37) implies that

$$
\alpha_{i,t+1} \langle \hat{z}_{i,t} - z_{i,t}, \nabla f_{i,t}(z_{i,t})\rangle
$$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
\[
\leq \alpha_{t+r}(\bar{z}_{i,t} - z_{i,t}, \nabla f_i, (z_{i,t}) - \nabla f_i, (\bar{z}_{i}) + 2\sqrt{N}B_z \Lambda \delta_t + \\
\left(2\sqrt{N}B_z^2 L + B_z L_f\right) \alpha_{t+r} \eta_t + 2B_z L_g \Lambda \alpha_{t+r+1} \gamma_t.
\]

(41)

For the second term on the right-hand side of (36), there holds
\[
\alpha_{t+r} E_{u_{i,t}}[(z_{i,t} - \tilde{z}_{i,t+1}, \nabla f_i, (z_{i,t}))]
\leq E_{u_{i,t}} \left[ \frac{\alpha_{t+r}^2}{\mu_0} \|\nabla f_i, (z_{i,t})\|^2 + \frac{\mu_0}{4} \|z_{i,t} - \tilde{z}_{i,t+1}\|^2 \right]
\leq \frac{n_i^2 B_z^2 \alpha_{t+r}^2}{\mu_0} \frac{\delta_t}{\delta_t^2} + \frac{\mu_0}{4} E_{u_{i,t}}[\|z_{i,t} - \tilde{z}_{i,t+1}\|^2]
\]

(42)

where the first and second inequalities are derived by the Cauchy-Schwarz inequality and S in Lemma 1, respectively.

As to the third term on the right-hand side of (36), one gets that
\[
\alpha_{t+r} E_{u_{i,t}}[(\tilde{z}_{i,t} - z_{i,t+1}, (\tilde{g}_t(z_{i,t}))^T \hat{\lambda}_{i,t+r})]
\]
\[
\leq \alpha_{t+r} \hat{\lambda}_{i,t+r} (\tilde{g}_t(z_{i,t})) - \tilde{g}_t(z_{i,t}))
\leq \frac{n_i^2 B_z^2 \alpha_{t+r}^2}{\mu_0} \frac{\delta_t}{\delta_t^2} + \frac{\mu_0}{4} E_{u_{i,t}}[\|z_{i,t} - \tilde{z}_{i,t+1}\|^2]
\]

(43)

Then, combining with (36) and (41)–(45) gives that
\[
E_{u_{i,t}}[D_{\psi}, (\tilde{x}_{i,t+1}, z_{i,t,r+1})]
\leq D_{\psi}(\tilde{x}_{i,t+1}, z_{i,t}) + K \|\tilde{x}_{i,t+1} - \tilde{x}_{i}\|^2
\leq \alpha_{t+r} (\tilde{x}_{i,t} - z_{i,t}, \nabla f_i, (z_{i,t}) - \nabla f_i, (\tilde{x}_{i})
\leq 2\sqrt{N}B_z \Lambda \delta_t + \\
\left(2\sqrt{N}B_z^2 L + B_z L_f\right) \alpha_{t+r} \eta_t + 2B_z L_g \Lambda \alpha_{t+r+1} \gamma_t
\]

(44)

(45)

via the properties of conditional expectation, one has that $E_{\tilde{f}_{t+r}} [E_{u_{i,t}}[D_{\psi}, (\tilde{x}_{i,t+1}, z_{i,t+1})]] = E_{\tilde{f}_{t+r}}[E_{u_{i,t}}[D_{\psi}, (\tilde{x}_{i,t+1}, z_{i,t+1})]]$.

(46)

Summing over $i = 1, \ldots, N, t = 1, \ldots, T$, taking expectation on its both sides, relying on Assumption 4, and rearranging the inequality, one has that
\[
\frac{\mu}{2} \sum_{i=1}^{T} E [\|x_{i,t} - z_{i,t}\|^2]
\leq \sum_{i=1}^{T} \frac{1}{\alpha_{t+r}^2} \sum_{i=1}^{N} E[D_{\psi}, (x_{i,t+1}, z_{i,t})] - D_{\psi}, (x_{i,t+r+1}, z_{i,t+1})]
\leq \frac{n_i^2 B_z^2 \alpha_{t+r}^2}{\mu_0} \frac{\delta_t}{\delta_t^2} + \frac{\mu_0}{4} \sum_{i=1}^{N} E_{u_{i,t}}[\|z_{i,t} - \tilde{z}_{i,t+1}\|^2]
\]

\[
\leq \frac{n_i^2 B_z^2 \alpha_{t+r}^2}{\mu_0} \frac{\delta_t}{\delta_t^2} + \frac{\mu_0}{4} \sum_{i=1}^{N} E_{u_{i,t}}[\|z_{i,t} - \tilde{z}_{i,t+1}\|^2]
\]

(47)

where the inequality holds based on $\sum_{i=1}^{N} \|x_{i,t} - z_{i,t}\|^2 \leq 0$.

Based on the definition of $\Lambda$, in Lemma 2, there holds
\[
-\frac{\Lambda_{t+1}}{2\gamma_t} = - \sum_{i=1}^{N} \left[ \frac{1}{2\gamma_t} \|x_{i,t+1} - \lambda\|^{2} - \frac{1}{2\gamma_t} \|x_{i,t} - \lambda\|^{2} \right]
\]

(48)

Summing over $t \in [T]$ yields that
\[
- \sum_{t=1}^{T} \frac{\Lambda_{t+1}}{2\gamma_t}
\leq - \sum_{t=1}^{N} \left[ \frac{1}{2\gamma_t} \|x_{i,T+1} - \lambda\|^{2} - \frac{1}{2\gamma_t} \|x_{i,t} - \lambda\|^{2} \right]
\]

(49)

(50)

where $\lambda_{i,t} = 0_{m}$ is applied to derive the inequality. Combining the above inequality with (19) in Lemma 2 and (47) results in (20).

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
C. Proof of Lemma 4

Note that the first term on the right hand-side of (20) satisfies

\[
\sum_{t=1}^{T} \frac{1}{\alpha_{t+r}} \sum_{i=1}^{N} \mathbb{E}[D_{\phi_i}(\tilde{x}_{i,t}, z_{i,t}) - D_{\phi_i}(\tilde{x}_{i,t+r+1}, z_{i,t+r+1})] = \sum_{t=1}^{T} \sum_{i=1}^{N} \left( \frac{1}{\alpha_{t+r}} - \frac{1}{\alpha_{t-1}} \right) \mathbb{E}[D_{\phi_i}(\tilde{x}_{i,t}, z_{i1})]
\]

and

\[
\sum_{t=1}^{T} \sum_{i=1}^{N} \frac{1}{\alpha_{t+r}} \sum_{\tau=1}^{T} \sum_{\xi=1}^{N} \left[ \tau \sum_{i=1}^{N} \mathbb{E}[D_{\phi_i}(\tilde{x}_{i,t}, z_{i,t})] \right]
\]

\[\leq \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{1}{\alpha_{t+r}} \sum_{\tau=1}^{T} \sum_{\xi=1}^{N} \left[ \tau \sum_{i=1}^{N} \mathbb{E}[D_{\phi_i}(\tilde{x}_{i,t}, z_{i,t})] \right]
\]

\[= \frac{\sqrt{NK}}{\alpha_{t+r}} \sum_{t=1}^{T} \sum_{\xi=0}^{N} \left[ B_{\xi} K \sum_{\tau=1}^{T} \frac{\eta_{-t+1-\xi}}{\alpha_{t+r}} \right] \leq \frac{\sqrt{NK}}{\alpha_{t+r}} \sum_{t=1}^{T} \sum_{\xi=0}^{N} \left[ B_{\xi} K \sum_{\tau=1}^{T} \frac{\eta_{-t+1-\xi}}{\alpha_{t+r}} \right]
\]

(49)

where Assumption 5 is adopted to get the first inequality. In addition, it holds that

\[
\sum_{t=1}^{T} \frac{1}{\alpha_{t+r}} \sum_{i=1}^{N} \left[ \frac{1}{N} \sum_{\tau=1}^{T} \sum_{\xi=1}^{N} \left( \frac{1}{\alpha_{t+r}} - \frac{1}{\alpha_{t-1}} \right) \mathbb{E}[D_{\phi_i}(\tilde{x}_{i,t}, z_{i,t})] \right]
\]

\[\leq \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{1}{\alpha_{t+r}} \sum_{\tau=1}^{T} \sum_{\xi=1}^{N} \left[ \tau \sum_{i=1}^{N} \mathbb{E}[D_{\phi_i}(\tilde{x}_{i,t}, z_{i,t})] \right]
\]

\[= \frac{\sqrt{NK}}{\alpha_{t+r}} \sum_{t=1}^{T} \sum_{\xi=0}^{N} \left[ B_{\xi} K \sum_{\tau=1}^{T} \frac{\eta_{-t+1-\xi}}{\alpha_{t+r}} \right] \leq \frac{\sqrt{NK}}{\alpha_{t+r}} \sum_{t=1}^{T} \sum_{\xi=0}^{N} \left[ B_{\xi} K \sum_{\tau=1}^{T} \frac{\eta_{-t+1-\xi}}{\alpha_{t+r}} \right]
\]

(50)

The last inequality is based on the decrease of \(\eta_t - \eta_{t+1}\) with \(t\). Then, taking \(\lambda = 0.0\) and combining (20), (49), and (50), (21) is derived by

\[
\mathbb{E}[\operatorname{Reg}_i(T)] = \frac{1}{C_4T} \left[ \sum_{t=1}^{T} \left( f_{i,t}(x_{i,t}, x_{i,t-1}) - f_{i,t}(x_{i,t}, x_{i,t-1}) \right) \right]
\]

\[\leq L_f \left[ \sum_{t=1}^{T} \left| x_{i,t} - x_{i,t-1} \right| \right]
\]

\[= L_f \left[ \sum_{t=1}^{T} \left| x_{i,t} - \tilde{x}_{i,t} + \tilde{x}_{i,t} - \tilde{x}_{i,t} + \delta_i u_{i,t} - \delta_i \eta_{t+1} \right| \right]
\]

Taking \(\lambda = \lambda_r = 2C_4^{-1} \frac{\sqrt{T}}{\alpha_{t+r}} \sum_{i=1}^{T} g_i(x_i)\), and \(20\), it holds that

\[
\frac{1}{C_4T} \left[ \sum_{t=1}^{T} g_i(x_i) \right]^2 \leq C_{2,T} + \frac{\sqrt{NK}}{\alpha_{t+r}} \left( \tau + 1 \right) \Phi_{\tau+1} + \Theta_{\tau,2}
\]

The proof is thus completed.