2021 American Control Conference (ACC)
New Orleans, USA, May 25-28, 2021

KL Divergence Regularized Learning Model for Multi-Agent Decision Making

Shinkyu Park and Naomi Ehrich Leonard

Abstract— **This paper investigates a multi-agent decision making model in large population games. We consider a population of agents that select strategies of interaction with one another. Agents repeatedly revise their strategy choices using revisions defined by the decision-making model. We examine the scenario in which the agents' strategy revision is subject to time delay. This is specified in the problem formulation by requiring the decision-making model to depend on delayed information of the agents' strategy choices. The main goal of this work is to find a multi-agent decision-making model under which the agents' strategy revision converges to equilibrium states, which in our population game formalism coincide with the Nash equilibrium set of underlying games. As key contributions, we propose a new decision-making model called the Kullback-Leibler (KL) divergence regularized learning model, and we establish stability of the Nash equilibrium set under the new model. Using a numerical example and simulations, we illustrate strong convergence properties of our new model.**

2021 American Control Conference (ACC) | 978-1-6654-4197-1/21/$31.00 ©2021 IEEE | DOI: 10.23919/ACC50511.2021.9483414

I. Introduction

Consider a multi-agent decision problem in large population games as follows. Given a finite set of strategies, each agent in the population makes a decision on which strategy to select for strategic interactions with other agents. We adopt the population game formalism [1, Chapter 2] in which a payoff function assigns payoffs to agents based on their strategy profile – the distribution of the agents' strategy choices – and a decision-making model prescribes how the agents revise and improve their strategy choices. In engineering research communities, such multi-agent decision problems are prevalent and some of existing works focus on seeking decision-making models that enable a large population of agents to learn and self-organize to an effective strategy profile [2]–[9].

The main purpose of this paper is to develop and analyze a new decision-making model that ensures convergence of the strategy profile to an equilibrium state when the payoff function is subject to time delay. Population games with time delay can be used to formulate real-world multi-agent decision problems in which there is delay, in, e.g., propagation of traffic congestion in congestion games [9], communication between the electric power utility and demand response agents in demand response games [6], and information transmission between agents in network games [10].

This work is supported by Princeton University's School of Engineering and Applied Science through the generosity of Lydia and William Addy '82.

S. Park and N. E. Leonard are with the Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544, USA. shinkyu@princeton.edu, naomi@princeton.edu

Prior works in the game theory literature suggest that when population games are subject to time delay, the strategy profile will exhibit oscillations when existing decision-making models are used [11]–[21]. Such oscillations in multi-agent decision making would prevent agents from selecting an effective strategy profile.

The key contribution of this paper is the proposal of a new class of decision-making models called the *Kullback-Leibler (KL) divergence regularized learning model*. Under this model, the agents' strategy revision is insensitive to small fluctuations in the payoffs, which prevents the strategy profile from exhibiting oscillations, and is effective in successively improving agent strategy choices. As a consequence, despite time delay in the payoff function, the proposed model enables the agents' strategy profile to converge to an equilibrium state, which, in our context, coincides with the Nash equilibrium of the underlying population games.

The proposed KL divergence regularized model can be viewed as a generalization of the logit model [22], [23]. As explained in [24], [25], the logit model attains stability in a larger class of population games, including games with time delay, than do other classes of decision-making models. However, the equilibrium state of the logit model is a perturbed version of the Nash equilibrium. This forces the agents to select a sub-optimal strategy profile in certain population games such as potential games [26], [27]. This downside motivated the present investigation of a new model.

Many works reported in the game theory literature focus on investigating the effect of time delay on stability of equilibrium states of existing models and establishing conditions on the time delay that guarantee convergence of the strategy profile. Instead, in our work using the KL divergence regularized model, we establish stability of the Nash equilibrium set in an important class of population games subject to any fixed time delay. We summarize the main contributions of this paper as follows:

- We propose a new class of decision-making models in large population games – KL divergence regularized model – and provide an algorithm that iteratively updates the model's parameters. Leveraging stability results from a recent work on higher-order learning in large population games [24], we establish, under the KL divergence regularized model, the convergence of the strategy profile to the Nash equilibrium set in an important class of population games, widely known as *contractive population games* [28].
- We simulate a congestion game in which the payoff function is subject to time delay. Using the results, we

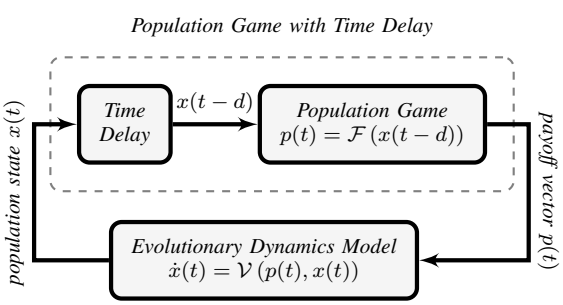

Fig. 1. Diagram of the framework defined as a feedback interconnection of an evolutionary dynamics model and a population game with time delay.

compare the performance of existing decision-making models with that of the KL divergence regularized model to highlight its stronger convergence property.

The paper is organized as follows. In §II, we explain preliminaries and the multi-agent decision problem. In §III, we introduce the KL divergence regularized model that describes how the agents can improve their strategy choices in population games with time delay. We present the main results of the paper, which establish the convergence of the strategy profile to the set of Nash equilibria in contractive population games. To illustrate the key contribution of this work, in §IV, we present simulations using an example to demonstrate the effectiveness of the KL divergence regularized model compared to other existing decision-making models. We conclude and discuss future directions in §V.

II. Preliminaries and Problem Description

Consider a large population of agents where each agent selects a strategy to engage in a strategic interaction with other agents. We denote by $\{1, \cdots, n\}$ the set of n strategies available to the agents. Given that each agent can select one strategy at a time, let $x = (x_1, \cdots, x_n)$ be an n-dimensional nonnegative real-valued vector with i-th entry x_i denoting the portion of the population adopting strategy i. The vector x specifies the strategy profile of the population. Following well-established convention [1], we refer to x as the *population state*. The set of feasible population states is defined as

$$\mathbb{X} = \left\{ z \in \mathbb{R}^n_+ \;\middle|\; \sum_{i=1}^{n} z_i = 1 \right\}. \tag{1}$$

In this section we present the framework we adopt to formulate the multi-agent decision problem investigated. The framework is illustrated in Fig. 1.

A. Population Games with Time Delay

In our framework, agents can revise their strategy choices in response to an n-dimensional real-valued vector $p = (p_1, \cdots, p_n)$, where each p_i represents the payoff assigned to the agents selecting strategy i. In the population game formalism [1, Chapter 2], given the population state $x \in \mathbb{X}$, the payoff vector p is defined by a continuously differentiable function $\mathcal{F} : \mathbb{X} \to \mathbb{R}^n$ as $p = \mathcal{F}(x)$. We refer to $\mathcal{F}$ as the payoff function.

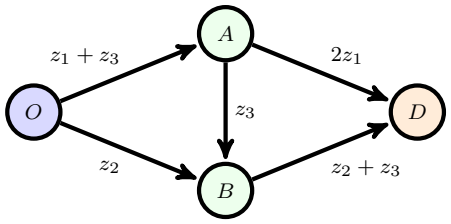

Fig. 2. Congestion Game: Agents traverse from the origin O to the destination D using one of the following three routes: $O \to A \to D$ (Route 1), $O \to B \to D$ (Route 2), and $O \to A \to B \to D$ (Route 3). Each z_i denotes the portion of the population using Route i and the weight on each edge denotes the level of congestion on the edge.

Throughout the paper, we identify population games with their associated payoff functions $\mathcal{F}$ and adopt the following definition of the Nash equilibrium.

Definition 1 (Nash Equilibrium of Population Game $\mathcal{F}$): An element z^{NE} in $\mathbb{X}$ is called the *Nash equilibrium* of the population game $\mathcal{F}$ if it satisfies the following condition:

$$\left(z^{\text{NE}} - z\right)^T \mathcal{F}\left(z^{\text{NE}}\right) \geq 0, \ \forall z \in \mathbb{X}. \tag{2}$$

Population games can have multiple Nash equilibria. Denote $\mathbb{NE}(\mathcal{F})$ as the set of Nash equilibria of $\mathcal{F}$.

We adopt the following definition of a norm for the differential map $D\mathcal{F}$ of $\mathcal{F}$: For $\|\cdot\|_2$ the Euclidean norm in $\mathbb{R}^n$ and $\mathbb{TX}$ the tangent space of $\mathbb{X}$,

$$\|D\mathcal{F}(z)\| = \max_{\tilde{z} \in \mathbb{TX}} \frac{\|D\mathcal{F}(z)\tilde{z}\|_2}{\|\tilde{z}\|_2}. \tag{3}$$

We make the following assumption on $\mathcal{F}$.

Assumption 1: We assume the differential map $D\mathcal{F}$ of $\mathcal{F}$ exists and is bounded: $\|D\mathcal{F}(z)\| \leq B_{\mathcal{F}}, \ \forall z \in \mathbb{X}$. Positive constant $B_{\mathcal{F}}$ is the bound of $D\mathcal{F}$.

A population game $\mathcal{F}$ is defined as contractive as follows.

Definition 2 (Contractive Population Game [28]): A population game $\mathcal{F}$ is *contractive* if the following holds:

$$(w - z)^T \left(\mathcal{F}(w) - \mathcal{F}(z)\right) \leq 0, \ \forall w, z \in \mathbb{X}. \tag{4}$$

We present an example of the contractive population game, which we adopt from [24]. We will use it to validate our main results through simulations in §IV.

Example 1 (Congestion Game): Each agent in a population traverses between pre-assigned origin and destination using one of 3 available routes (see Fig. 2). Each strategy in the game is defined as an agent taking one of the routes and its associated payoff quantifies the level of congestion (such as the time it takes to traverse the route), which depends on the number of agents using the same or other overlapping routes. The factor 2 on $A \to D$ reflects a narrow path. To formalize, we adopt the payoff function $\mathcal{F}^{\text{Congestion}}$ defined as

$$\mathcal{F}^{\text{Congestion}}(z) = -\begin{pmatrix} 3z_1 + z_3 \\ 2z_2 + z_3 \\ z_1 + z_2 + 3z_3 \end{pmatrix}, \tag{5}$$

which represents (the negative of) the level of congestion on all 3 routes. We note that (5) has the unique Nash equilibrium $(4/11,\ 6/11,\ 1/11)$ at which the average level of congestion across the three routes is minimized. We can identify that the congestion game (5) qualifies as contractive.

We consider the scenario, distinct from the standard population formalism, where the population game is subject to time delay. Given a population state trajectory $x(t),\ t \geq -d$, the payoff vector $p(t)$ at each time instant $t \geq 0$ is

$$p(t) = \mathcal{F}\left(x(t-d)\right), \tag{6}$$

where d is a positive constant denoting the time delay. We assume that the time delay d is an unknown parameter but is upper bounded by a constant B_d. Note that as in the standard population game formalism, when the population state $x(t)$ converges to x^*, the payoff vector $p(t)$ converges to $\mathcal{F}(x^*)$.

B. Strategy Revision Protocol and Evolutionary Dynamics Model

In our framework, each agent repeatedly revises its strategy choice based on the payoffs associated with the available strategies. To formalize this, we adopt the evolutionary dynamics model [1, Part II] in which the so-called *strategy revision protocol* $\mathcal{T}_{ji}(r,z)$ describes the probability that each agent switches its strategy from j to i provided that the payoff vector and population state are assigned as $r \in \mathbb{R}^n$ and $z \in \mathbb{X}$, respectively. Following discussions in [1, Chapter 10], when the population size is large and each agent is randomly selected for the strategy revision based on an exponential distribution, the population state $x(t) = (x_1(t), \cdots, x_n(t))$ at each time instant $t \geq 0$ can be approximated by a solution of the following ordinary differential equation:

$$\dot{x}_i(t) = \sum_{j=1}^{n} x_j(t)\mathcal{T}_{ji}\left(p(t), x(t)\right) - x_i(t)\sum_{j=1}^{n}\mathcal{T}_{ij}\left(p(t), x(t)\right) \tag{7}$$

where the payoff vector $p(t)$ is determined by the payoff function of the underlying population game, e.g., (6). Following the same naming convention as in [24], we refer to (7) as the *evolutionary dynamics model (EDM)*.

Reference [1, Chapter 5] summarizes well-known protocols proposed in the game theory literature. Among existing strategy revision protocols, of relevance to our new model, which we introduce in §III, is the logit protocol:

$$\mathcal{T}_i^{\text{Logit}}(r) = \frac{\exp\left(\eta^{-1} r_i\right)}{\sum_{l=1}^{n}\exp\left(\eta^{-1} r_l\right)},\ r \in \mathbb{R}^n \tag{8}$$

where η is a positive constant and r is the value assigned to the payoff vector. Agents that adopt the logit protocol, i.e., $\mathcal{T}_{ji}(r,z) = \mathcal{T}_i^{\text{Logit}}(r)$, revise their strategy choices depending only on the payoffs. As discussed in [23], the logit protocol is regarded as a perturbed version of the best response protocol where the level of perturbation is quantified by constant η.

To explain the motivation behind proposing a new model, we briefly discuss the limitation of some of the existing strategy revision protocols including (8). Analysis presented in [11]–[13], [17], [18], [20], [21] suggests that stability of the Nash equilibrium set under existing models, such as best response dynamics, replicator dynamics, and their variants, is attained when the time delay term d in (6) is sufficiently small. Also, Hopf bifurcation analysis from [20], [21] shows that if d is large, the population state trajectory forms a limit cycle and does not converge to the Nash equilibrium set.

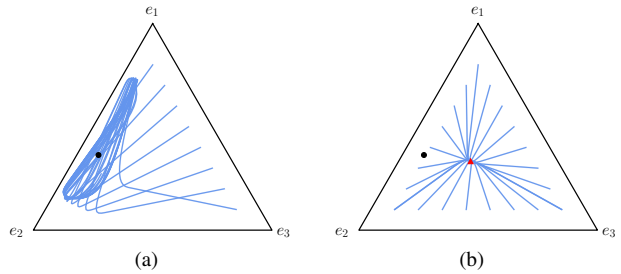

Fig. 3. Population state trajectories under the logit protocol (8) with (a) $\eta = 0.1$ and (b) $\eta = 3.0$ in the congestion game (5) with unit time delay ($d = 1$). The black circle represents the Nash equilibrium $(4/11, 6/11, 1/11)$ of the congestion game and the red triangle represents the limit point of the population state trajectories.

And yet, according to stability results from [24, §VIII], [25, Table II], the EDM (7) under the logit protocol (8) attains stability in a larger class of population games compared to other protocols. As we will show in §III-A, if the constant η in (8) is sufficiently large, the resulting population state trajectory converges to the stationary points of (7). However, such stationary points do not coincide with the Nash equilibria of the underlying population games. In fact, as we explain in §III-A, the population state converges to a perturbed version of the set of Nash equilibria.

To illustrate this, in Fig. 3, using the logit model (8) with $\eta = 0.1, 3.0$ in the congestion game (5) with unit time delay ($d = 1$), we show that when η is small ($\eta = 0.1$), the resulting population state trajectories oscillate, whereas when η is large ($\eta = 3.0$), the trajectories converge to a stationary point, located away from the Nash equilibrium of (5).

C. Problem Formulation

Our main goal in this paper is to design a new class of protocols $\mathcal{T}_{ji}$ under which the population state $x(t)$ of (7) converges to the Nash equilibrium set $\mathbb{NE}(\mathcal{F})$ of the underlying population game $\mathcal{F}$ subject to time delay. We formally state the main problem as follows.

Problem 1: Design a strategy revision protocol $\mathcal{T}_{ji}$ and find conditions on the EDM (7) and the population game $\mathcal{F}$ with time delay (6) under which the population state $x(t)$ asymptotically attains the Nash equilibrium set $\mathbb{NE}(\mathcal{F})$:

$$\lim_{t\to\infty}\inf_{z\in\mathbb{NE}(\mathcal{F})}\|x(t) - z\|_2 = 0 \tag{9}$$

We emphasize that unlike the logit protocol (8), which only ensures the convergence to a perturbed version of the Nash equilibrium set, we seek a new strategy revision protocol that guarantees the convergence of the population state exactly to the Nash equilibrium set.

III. KL Divergence Regularized Protocol

Recall that each agent in the population revises its strategy choice based on the protocol $\mathcal{T}_{ji}$. In this section, we begin by proposing a new protocol called the *KL divergence regularized protocol*, described as follows. Given an element

$y = (y_1, \cdots, y_n)$ in $\mathrm{int}(\mathbb{X})$, the interior of the population state set $\mathbb{X}$, each agent makes a revision to strategy i with probability given by

$$\mathcal{T}_i^{\mathrm{KLReg}}(r) = \frac{y_i \exp\left(\eta^{-1} r_i\right)}{\sum_{l=1}^n y_l \exp\left(\eta^{-1} r_l\right)} \tag{10}$$

where $r \in \mathbb{R}^n$ represents the value assigned to the payoff vector and η is a positive constant. The parameter y can be viewed as a bias imposed on the protocol (10) in the sense that when the payoffs assigned to the strategies are identical, i.e., $r_1 = \cdots = r_n$, the agents favor strategy i with probability proportional to y_i. Also, we note that when the parameter y satisfies $y_1 = \cdots = y_n$, the protocol (10) coincides with the logit protocol (8); hence, (10) qualifies as a generalization of (8).

One key aspect of the KL divergence regularized protocol is in having flexibility in selecting the parameter y. In §III-A, we describe an algorithm that iteratively updates y and show that, under the algorithm, the KL divergence regularized protocol ensures the convergence of the population state $x(t)$ to the Nash equilibrium set of underlying population games.

From (7) and (10), by applying $\mathcal{T}_{ji}(r, z) = \mathcal{T}_i^{\mathrm{KLReg}}(r)$, we define the KL divergence regularized EDM as follows:

$$\dot{x}_i(t) = \frac{y_i \exp\left(\eta^{-1} p_i(t)\right)}{\sum_{l=1}^n y_l \exp\left(\eta^{-1} p_l(t)\right)} - x_i(t). \tag{11}$$

Note that for a bounded payoff vector $p(t)$, we can infer that if the population state $x(t)$ starts from the interior set $\mathrm{int}(\mathbb{X})$, then it remains in $\mathrm{int}(\mathbb{X})$.

To explain our decision to name (10) the KL divergence regularized protocol, note that (10) can be expressed as

$$\begin{aligned}\mathcal{T}^{\mathrm{KLReg}}(r) &= \begin{pmatrix} \mathcal{T}_1^{\mathrm{KLReg}}(r) \\ \vdots \\ \mathcal{T}_n^{\mathrm{KLReg}}(r) \end{pmatrix} \\ &= \arg\max_{z \in \mathbb{X}} \left[r^T z - \eta\, \mathcal{D}(z \,\|\, y) \right], \end{aligned} \tag{12}$$

where $\mathcal{D}(z \,\|\, y)$ is the Kullback-Leibler (KL) divergence defined as $\mathcal{D}(z \,\|\, y) = \sum_{i=1}^n z_i \ln \frac{z_i}{y_i}$. The KL divergence penalizes the difference between the variable z and the parameter y. By viewing the KL divergence as a regularization in the maximization (12), we can observe that the protocol (10) computes the maximizer of the *regularized* cost function that combines the average payoff $r^T z$ and (the negative of) the regularization term $\mathcal{D}(z \,\|\, y)$, weighted by η.

A. Convergence Properties

Consider the KL divergence regularized EDM (11) in which the payoff vector $p(t)$ is defined by a population game with time delay (6), i.e., $p(t) = \mathcal{F}(x(t-d))$. We recall that such configuration can be defined as a feedback interconnection of (6) and (11), as illustrated in Fig. 1. In this section, we establish the convergence of the population state $x(t)$ from (11) to the Nash equilibrium set $\mathbb{NE}(\mathcal{F})$.

The central idea in establishing the convergence result is a suitable selection of the parameter y of (11). To see this, as in [23], we can infer that the stationary point of (11) is the Nash equilibrium of the perturbed payoff function $\tilde{\mathcal{F}}$ defined by $\tilde{\mathcal{F}}(z) = \mathcal{F}(z) - \eta \nabla_z \mathcal{D}(z \,\|\, y)$. Therefore, given that the population state $x(t)$ converges, if the parameter y coincides with the Nash equilibrium of the original payoff function $\mathcal{F}$, then $x(t)$ converges to the Nash equilibrium. In what follows, we present an algorithmic scheme that iteratively updates the parameter y and show that, under the algorithm, the KL divergence regularized EDM ensures the convergence of both y and $x(t)$ to the Nash equilibrium set $\mathbb{NE}(\mathcal{F})$.

Our analysis hinges on a stability technique recently proposed in evolutionary game contexts [24], [25], [29], where the notion of passivity [30] from feedback control theory plays a pivotal role. The key ideas behind the passivity-based stability technique are defining proper notions of passivity for (6) and (11), and then establishing stability of the Nash equilibrium set by leveraging the well-established principle in control theory that a feedback interconnection of two passive dynamical systems results in stability.

We begin by briefly reviewing two notions of passivity – *weak δ-antipassivity* and *δ-passivity* – adopted for (6) and (11). We then proceed with establishing stability of the feedback interconnection of (6) and (11). We refer the interested reader to [24], [25], [29] for more in-depth discussions on passivity in population games and also [31] that investigates different notions of passivity in game theory.

Definition 3 (Weak δ-Antipassivity with Deficit [24]): The population game with time delay (6) is *weak δ-antipassive with deficit ν^** if there is $\alpha > 0$ for which

$$\alpha > \int_{t_0}^t \left[\dot{x}^T(\tau) \dot{p}(\tau) - \nu \dot{x}^T(\tau) \dot{x}(\tau) \right] \mathrm{d}\tau, \ \forall t \geq t_0 \geq 0$$

holds for every population state trajectory $x(t),\ t \geq 0$ and for every nonnegative constant $\nu > \nu^*$.

The constant ν^* is used as a measure of passivity deficit in (6). In the following lemma, we establish weak δ-antipassivity of the population game with time delay (6).

Lemma 1: The population game with time delay (6) is weak δ-antipassive with positive deficit $\nu^* = B_{\mathcal{F}}$, where $B_{\mathcal{F}}$ is the upper bound of $D\mathcal{F}$ as defined in Assumption 1.

Definition 4 (δ-Passivity with Surplus [24]): The EDM (7) is *δ-passive with surplus η^** if there is a continuously differentiable function $\mathcal{S} : \mathbb{X} \times \mathbb{R}^n \to \mathbb{R}_+$ for which

$$\begin{aligned}&\mathcal{S}(x(t), p(t)) - \mathcal{S}(x(t_0), p(t_0)) \leq \\ &\quad \int_{t_0}^t \left[\dot{x}^T(\tau) \dot{p}(\tau) - \eta\, \dot{x}^T(\tau) x(\tau) \right] \mathrm{d}\tau, \ \forall t \geq t_0 \geq 0 \end{aligned} \tag{13}$$

holds for every payoff vector trajectory $p(t)$, $t \geq 0$, and for every nonnegative constant $\eta < \eta^*$. We refer to $\mathcal{S}$ as the *δ-storage function*.

The constant η^* is used as a measure of passivity surplus in (11). Definition 4 states a stronger notion of passivity, as compared to Definition 3, since it requires the existence of the δ-storage function $\mathcal{S}$. The following definition of informative δ-storage functions will play an important role in establishing the main convergence result.

Definition 5 (Informative δ-Storage Function [24]): Let $\mathcal{S}$ be a δ-storage function of a given δ-passive EDM (7). $\mathcal{S}$ is *informative* if the following three conditions are equivalent:

$$\mathcal{S}(z,r) = 0 \tag{14a}$$

$$\mathcal{V}(z,r) = 0 \tag{14b}$$

$$\nabla_z^T \mathcal{S}(z,r)\, \mathcal{V}(z,r) = 0 \tag{14c}$$

where each element $\mathcal{V}_i : \mathbb{X} \times \mathbb{R}^n \to \mathbb{R}$ of $\mathcal{V} = (\mathcal{V}_1, \cdots, \mathcal{V}_n)$ is the vector field of (7), defined as

$$\mathcal{V}_i(z,r) = \sum_{j=1}^{n} z_j \mathcal{T}_{ji}(r,z) - z_i \sum_{j=1}^{n} \mathcal{T}_{ij}(r,z).$$

The following lemma establishes the δ-passivity of the KL divergence regularized EDM (11).

Lemma 2: Given $\eta > 0$ and $y \in \operatorname{int}(\mathbb{X})$, the KL divergence regularized EDM (11) is δ-passive with surplus η and has an informative δ-storage function $\mathcal{S} : \mathbb{X} \times \mathbb{R}^n \to \mathbb{R}_+$,

$$\mathcal{S}(z,r) = \max_{\bar{z} \in \mathbb{X}} \left(\bar{z}^T r - \eta\, \mathcal{D}(\bar{z} \,\|\, y)\right) - \left(z^T r - \eta\, \mathcal{D}(z \,\|\, y)\right). \tag{15}$$

Using Lemmas 1 and 2, we proceed to establish our main convergence result. As a preliminary step, the following proposition, which is based on [24, Lemma 1], establishes the convergence of the population state $x(t)$ of (11) with the parameter y fixed and the payoff vector $p(t)$ determined by the population game with time delay (6). To state the proposition, we need the following definition of the perturbed Nash equilibrium set of the population game $\mathcal{F}$.

Definition 6: Given $\eta > 0$ and $y \in \operatorname{int}(\mathbb{X})$, let $\mathbb{PE}(\mathcal{F}, \eta\mathcal{D})$ be the set of perturbed Nash equilibria of $\mathcal{F}$:

$$\mathbb{PE}(\mathcal{F}, \eta\mathcal{D}) = \left\{ z^{\mathrm{PE}} \in \mathbb{X} \,\middle|\, \left(z^{\mathrm{PE}} - z\right)^T \tilde{\mathcal{F}}\left(z^{\mathrm{PE}}\right) \geq 0,\ \forall z \in \mathbb{X} \right\} \tag{16}$$

where $\tilde{\mathcal{F}}(z) = \mathcal{F}(z) - \eta\, \nabla_z \mathcal{D}(z \,\|\, y)$ is the perturbed payoff function and $\mathcal{D}(z \,\|\, y)$ is the KL divergence.

Proposition 1: Consider the KL divergence regularized EDM (11) for which the payoff vector $p(t)$ is determined by a population game $\mathcal{F}$ with time delay (6). Provided that the parameter η of (11) satisfies $\eta > B_{\mathcal{F}}$, the population state $x(t)$ of (11) converges to $\mathbb{PE}(\mathcal{F}, \eta\mathcal{D})$:

$$\lim_{t \to \infty} \inf_{z \in \mathbb{PE}(\mathcal{F}, \eta\mathcal{D})} \|x(t) - z\|_2 = 0. \tag{17}$$

The condition $\eta > B_{\mathcal{F}}$ in Proposition 1 implies that the convergence (17) holds if the lack of passivity in (6), quantified by $B_{\mathcal{F}}$, is compensated for by the surplus of passivity in (11), measured by η. In the remainder of this section, using Proposition 1, we propose and study Algorithm 1 that allows the population to repeatedly update the parameter y of (11). We show that the algorithm enables the convergence of the population state to the Nash equilibrium set $\mathbb{NE}(\mathcal{F})$.

Algorithm 1: Given initial values of the parameter $y \in \operatorname{int}(\mathbb{X})$ and a time instant variable $t_0 = 0$, update y and t_0 as follows. At every time instant t_1 at which the following conditions (18) and (19) hold, assign $y = x(t_1)$ and $t_0 = t_1$:

$$t_1 \geq t_0 + B_d \tag{18}$$

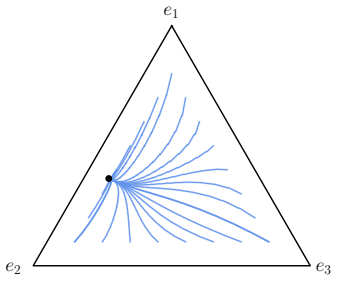

Fig. 4. Population state trajectories derived by the KL divergence regularized model with $\eta = 3.0$. The black circle represents the Nash equilibrium $(4/11, 6/11, 1/11)$ of the congestion game.

$$\max_{z \in \mathbb{X}} (z - x(t_1))^T \left(p(t_1) - \eta\, \nabla \mathcal{D}(x(t_1) \,\|\, x(t_0))\right) + B_{\mathcal{F}} \max_{\tau \in [t_1 - 2B_d, t_1]} \|x(t_1) - x(\tau)\|_2 \leq \frac{\eta}{2} \mathcal{D}(x(t_1) \,\|\, x(t_0)) \tag{19}$$

where $p(t)$ and $x(t)$ are the payoff vector and population state, respectively, of (6) and (11).

Time instant t_1 satisfying (19) always exists since, according to Proposition 1, the population state $x(t)$ converges to the set $\mathbb{PE}(\mathcal{F}, \eta\mathcal{D})$ and hence the left hand side of (19) vanishes as t_1 tends to infinity. Also, to realize Algorithm 1, the agents only need to know the upper bounds B_d and $B_{\mathcal{F}}$ of the time delay d and the payoff function $\mathcal{F}$, respectively, but not the exact forms of d and $\mathcal{F}$.

We remark that by the definition of the EDM (11), when the population state $x(t)$ starts from the interior set $\operatorname{int}(\mathbb{X})$, it remains in $\operatorname{int}(\mathbb{X})$. Hence, we conclude that Algorithm 1 ensures the updated parameter y to satisfy $y \in \operatorname{int}(\mathbb{X})$.

When the parameter y of (11) is updated according to Algorithm 1, we can establish the convergence of the population state to the Nash equilibrium set $\mathbb{NE}(\mathcal{F})$ in contractive population games (see Definition 2) The following theorem states the convergence result.

Theorem 1: Let the KL divergence regularized EDM (11), for which the payoff vector is $p(t)$, be determined by a contractive population game $\mathcal{F}$ with time delay (6). Suppose that $\eta > B_{\mathcal{F}}$ holds and the parameter y of (11) is repeatedly updated according to Algorithm 1. The population state $x(t)$ of (11) converges to the Nash equilibrium set $\mathbb{NE}(\mathcal{F})$:

$$\lim_{t \to \infty} \inf_{z \in \mathbb{NE}(\mathcal{F})} \|x(t) - z\|_2 = 0. \tag{20}$$

Note that since the population state $x(t)$ remains in the interior set $\operatorname{int}(\mathbb{X})$, if the population game $\mathcal{F}$ has a Nash equilibrium in the boundary $\operatorname{bd}(\mathbb{X})$ of $\mathbb{X}$, then $x(t)$ would not reach the equilibrium in finite time, but it will asymptotically converge to it.

IV. Simulations with Numerical Example

We illustrate our main results using a numerical example of the congestion population game $\mathcal{F}^{\mathrm{Congestion}}$ defined in (5). We adopt the KL divergence regularized model (11) with $\eta = 3.0$ where the parameter y in (11) is updated according to Algorithm 1. We carry out simulations in the congestion

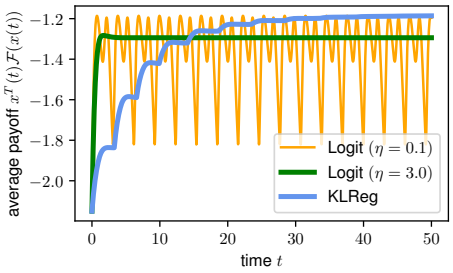

Fig. 5. Comparison of average payoff $x^T(t)\mathcal{F}(x(t))$ for the logit with $\eta = 0.1, 3.0$ and the KL divergence regularized with $\eta = 3.0$.

population game $\mathcal{F}^{\text{Congestion}}$ subject to unit time delay ($d = 1$), where for simplicity we assign $B_d = d$.

As can be observed from simulation outcomes, depicted in Fig. 4, the resulting population state trajectories converge to the Nash equilibrium of $\mathcal{F}^{\text{Congestion}}$. Recall that for the logit model case, depicted in Fig. 3, the population state trajectories either exhibit oscillations around the Nash equilibrium (when η is small) or converge to a stationary point located away from the Nash equilibrium (when η is large).

To assess the efficacy of the KL divergence regularized model in attaining an effective strategy profile of the agents, we compare the average payoff $x^T(t)\mathcal{F}(x(t))$, derived from the logit (8) with $\eta = 0.1, 3.0$ and from the KL divergence regularized protocol (10) with $\eta = 3.0$. As shown in Fig. 5 the population state determined by (10) asymptotically attains the largest average payoff.

V. Conclusions

We proposed and analyzed the KL divergence regularized model, conceived for a population of agents to attain an effective strategy profile in large population games. Key contributions of the paper include proposing the algorithmic scheme that successively updates the parameter of the model, and, by leveraging recent stability results in evolutionary games, establishing that population state trajectories resulting from the model converge to the Nash equilibrium set in contractive population games that are subject to time delay.

In future work, we plan to extend the presented results to other important scenarios where the payoffs are subject to multiple time delays and are derived from dynamically modified payoff models. We also plan to analytically examine the effects of disturbance in the payoffs and time delay on the convergence rate of the KL divergence regularized model.

References

[1] W. H. Sandholm, *Population Games and Evolutionary Dynamics*. MIT Press, 2011.

[2] D. Liu, A. Hafid, and L. Khoukhi, "Population game based energy and time aware task offloading for large amounts of competing users," in *2018 IEEE Global Comm. Conf. (GLOBECOM)*, 2018, pp. 1–6.

[3] S. Moon, H. Kim, and Y. Yi, "Brute: Energy-efficient user association in cellular networks from population game perspective," *IEEE Trans. Wireless Communications*, vol. 15, no. 1, pp. 663–675, 2016.

[4] D. Lee and D. Kundur, "An evolutionary game approach to predict demand response from real-time pricing," in *2015 IEEE Electrical Power and Energy Conference (EPEC)*, 2015, pp. 197–202.

[5] H. Tembine, E. Altman, R. El-Azouzi, and Y. Hayel, "Evolutionary games in wireless networks," *IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 40, no. 3, pp. 634–646, 2010.

[6] P. Srikantha and D. Kundur, "Resilient distributed real-time demand response via population games," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2532–2543, 2017.

[7] A. Pashaie, L. Pavel, and C. J. Damaren, "A population game approach for dynamic resource allocation problems," *International Journal of Control*, vol. 90, no. 9, pp. 1957–1972, 2017.

[8] N. Quijano, C. Ocampo-Martinez, J. Barreiro-Gomez, G. Obando, A. Pantoja, and E. Mojica-Nava, "The role of population games and evolutionary dynamics in distributed control systems: The advantages of evolutionary game theory," *IEEE Control Systems Magazine*, vol. 37, no. 1, pp. 70–97, 2017.

[9] M. J. Smith, "The stability of a dynamic model of traffic assignment—an application of a method of Lyapunov," *Transportation Science*, vol. 18, no. 3, pp. 245–252, 1984.

[10] I. Menache and A. Ozdaglar, "Network Games: Theory, Models, and Dynamics," *Synthesis Lectures on Communication Networks*, vol. 4, no. 1, pp. 1–159, Mar. 2011, publisher: Morgan & Claypool Publishers.

[11] H. Tembine, E. Altman, R. El-Azouzi, and Y. Hayel, "Bio-inspired delayed evolutionary game dynamics with networking applications," *Telecommunication Systems*, vol. 47, pp. 137–152, 2011.

[12] T. Yi and W. Zuwang, "Effect of time delay and evolutionarily stable strategy," *J. Theoretical Biology*, vol. 187, no. 1, pp. 111 – 116, 1997.

[13] S.-C. Wang, J.-R. Yu, S. Kurokawa, and Y. Tao, "Imitation dynamics with time delay," *J. Theoretical Biology*, vol. 420, pp. 8 – 11, 2017.

[14] H. Oaku, "Evolution with delay," *The Japanese Economic Review*, vol. 53, pp. 114–133, 2002.

[15] M. Bodnar, J. Mikisz, and R. Vardanyan, "Three-player games with strategy-dependent time delays," *Dynamic Games and Applications*, vol. 10, pp. 664–675, 2020.

[16] J. Alboszta and J. Mie¸kisz, "Stability of evolutionarily stable strategies in discrete replicator dynamics with time delay," *Journal of Theoretical Biology*, vol. 231, no. 2, pp. 175 – 179, 2004.

[17] N. B. Khalifa, R. El-Azouzi, and Y. Hayel, "Delayed evolutionary game dynamics with non-uniform interactions in two communities," in *53rd IEEE Conf. Decision and Control*, 2014, pp. 3809–3814.

[18] G. Obando, J. I. Poveda, and N. Quijano, "Replicator dynamics under perturbations and time delays," *Mathematics of Control, Signals, and Systems*, vol. 28, no. 20, 2016.

[19] N. Sîrghi and M. Neamţu, "Dynamics of deterministic and stochastic evolutionary games with multiple delays," *International Journal of Bifurcation and Chaos*, vol. 23, no. 07, p. 1350122, 2013.

[20] E. Wesson and R. Rand, "Hopf bifurcations in delayed rock–paper–scissors replicator dynamics," *Dynamic Games and Applications*, vol. 6, pp. 139–156, 2016.

[21] S. Mittal, A. Mukhopadhyay, and S. Chakraborty, "Evolutionary dynamics of the delayed replicator-mutator equation: Limit cycle and cooperation," *Phys. Rev. E*, vol. 101, p. 042410, Apr 2020.

[22] J. Hofbauer and W. H. Sandholm, "On the global convergence of stochastic fictitious play," *Econometrica*, vol. 70, no. 6, pp. 2265–2294, 2002.

[23] ——, "Evolution in games with randomly disturbed payoffs," *Journal of Economic Theory*, vol. 132, no. 1, pp. 47 – 69, 2007.

[24] S. Park, N. C. Martins, and J. S. Shamma, "Payoff dynamics model and evolutionary dynamics model: Feedback and convergence to equilibria (arxiv:1903.02018)," arXiv.org, March 2019.

[25] ——, "From population games to payoff dynamics models: A passivity-based approach," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 6584–6601.

[26] D. Monderer and L. S. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, no. 1, pp. 124 – 143, 1996.

[27] W. H. Sandholm, "Potential games with continuous player sets," *Journal of Economic Theory*, vol. 97, no. 1, pp. 81 – 108, 2001.

[28] ——, "Chapter 13: Population games and deterministic evolutionary dynamics," ser. Handbook of Game Theory with Economic Applications, H. P. Young and S. Zamir, Eds. Elsevier, 2015, vol. 4, pp. 703–778.

[29] M. J. Fox and J. S. Shamma, "Population games, stable games, and passivity," *Games*, vol. 4, pp. 561–583, Oct. 2013.

[30] J. C. Willems, "Dissipative dynamical systems part I: General theory," *Arch. Ration. Mech. Anal.*, vol. 45, no. 5, pp. 321–351, Jan. 1972.

[31] B. Gao and L. Pavel, "On passivity, reinforcement learning and higher-order learning in multi-agent finite games," *IEEE Transactions on Automatic Control*, pp. 1–1, 2020.