2022 IEEE 61st Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

Multi-Agent Learning via Markov Potential Games in Marketplaces for Distributed Energy Resources

Dheeraj Narasimha, Kiyeob Lee, Dileep Kalathil and Srinivas Shakkottai

2022 IEEE 61st Conference on Decision and Control (CDC) | 978-1-6654-6761-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CDC51059.2022.9992762

***Abstract*— Much change is happening in electricity markets due to the entrance of small-scale prosumers that both generate and consume electricity. Both large and small consumers can also be incentivized to reduce their demand during peak load periods, referred to as demand-response. The net effect of such distributed energy resources (DERs) on the grid can be quite substantial, and designing secondary markets wherein such DERs can participate repeatedly over time has become important. Many such marketplaces have a so-called potential game structure, in that a unilateral change in the strategy of an agent causes equivalent changes in both its own reward and a global potential function. We consider a dynamic setting in which each stage is a potential game, but is accompanied by Markovian state transitions, which we call Markov Potential Games (MPG). It is well known that it is formidably challenging to compute or learn Nash Equilibria (NE) in Markov Games. We develop a key concept that we term as the potential value function that ties together the potential function in the stage game with the value function in a Markov Decision Process. We first show that an NE can be computed in a centralized manner by maximizing the potential value function. We also show NE can also be obtained in a multi-agent manner via asynchronous better (not necessarily best) response updates that are consistent with a simple multi-agent reinforcement learning algorithm. Finally, we show several examples wherein the MPG framework applies to DER dynamics in an electricity marketplace, and numerically study the efficiency of the equilibria attained.**

I. Introduction

With increasing adoption of photo-voltaic cells in homes the number of individuals who are both producers and consumers within electric grids have greatly increased recently. We call these types of individuals *prosumers* who wish to engage with the electricity marketplace [1]. Other entrants into the marketplace are small and large consumers that can modify their consumption based on the peaks in power consumption by the rest of the grid or during emergency events. *Demand Response* (DR) programs seek to engage such consumers by incentivizing such demand shaping [2], [3]. The aggregate effect of including these distributed energy resources (DERs) can be substantial in lowering the variability of demand, thereby stabilizing prices and making the grid more reliable, and potentially also reducing carbon emissions through inclusion of cleaner energy sources. Indeed, the Electric Reliability Council of Texas (ERCOT) organizes a variety of programs to engage with DERs, including demand response approaches for flexible loads arising from emerging applications such as Bitcoin mining. Given their increasing importance, modeling and analyzing marketplaces for DER participation presents an interesting problem space. For example, we can consider a demand-response marketplace as a Cournot competition, where the price per unit-reduction depends on the offers for reduction made by all the participating individuals. The equilibrium price in such cases dictates the level of reduction for the agents taking part in these markets. It is well known that in these kinds of Cournot games, there is a special function called the *potential function* such that a unilateral change in reward by any player is aligned across all players through the potential function. The game is then referred to as potential game (PG), and the game's Nash Equilibrium (NE) is associated with the maximizing strategy of the potential function. Finite PG can attain such a NE point with a simple class of dynamics, namely, asynchronous best response dynamics, under which the agents asynchronously take best responses until no further better response is possible. However, while these games can model the demand as a function of prosumer participation, they are poor at modeling external states and their evolution. For instance, it is insufficient to model the demand-response marketplace as a stateless Cournot game, since the reduction in demand might not necessarily lower the price due to a high baseline demand in a particular time period. We are therefore interested in the setting where the states themselves follow a Markovian structure. Specifically, if we consider the state of the system as the current demand, a demand reduction made in a particular time period could result in a net demand increase in the next time period. Thus, the demand at any time period could depend both on exogenous factors such as the weather, as well the previous demand reduction actions.

In this Markov marketplace setting, each agent essentially solves a Markov Decision process (MDP) to identify its best response to the other agents' strategies, with the fixed point being an NE. Thus, we are interested in Markov potential games (MPG), which appear in a variety of resource sharing problems. These are Markov games in which the difference in utility (over the time horizon) induced by a change in strategy for any player is exactly equal to the difference in the value of an *auxiliary potential function*. Note, in our case each state induces its own potential game with a corresponding potential function, thus, we refer to the value (over the time horizon) engendered by the stage potential functions in this case as the *potential value function (PVF)*.

This work was supported in part by NSF grant ECCS-2038963 and NSF- CAREER-EPCN-2045783. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Authors are with the Dept.of Electrical and Computer Engineering, 3128 Texas A&M University, College Station, Texas email: dheeraj.narasimha, kiyeoblee, dileep.kalathil,sshakkot @tamu.edu

Concretely, we focus on an attractive structure where the auxiliary potential function is identical to the potential value function—a condition that we call *Strong Markov Potential Games (S-MPG)*. The goal of this work is first to establish conditions for existence of S-MPGs, and then to propose approaches for computing and learning an associated NE.

Main Contributions

Our main contributions are as follows:
(i) Our first result is on identifying criteria for the existence of S-MPGs. There has been recent work in the area of MPGs under variants of the notion of what makes a strong MPG. However, work such as [4] present incorrect criteria for their existence, [5] has an inherently unverifiable condition on the structure that allows an S-MPG to exist, while work such as [6] simply assume their existence. Thus, following the work of [7] we collect criteria for the existence of an auxiliary potential function and we show verifiable sufficiency conditions for S-MPGs.
(ii) We present three computation algorithms for identifying a ϵ-Nash equilibrium. First, following the seminal work of [8], it is natural to think of a finite improvement path as a sequence of best response policies. We extend this notion to the case of S-MPG and show how the PVF can be obtained by standard techniques such as policy or value iteration. Second, we propose a centralized approach to maximizing the PVF using value iteration. Third, we design a sequence of asynchronous one step better responses by each agent, forming a convergent decentralized value iteration process.
(iii) We present a model-free learning algorithm for learning an ϵ-Nash equilibrium under an unknown model. The algorithm follows a structure of *asynchronous play* which reduces learning in a possibly non-stationary environment of MPG to learning in a sequence of stationary environments of Markov Decision Processes.
(iv) Finally, we show under several settings on demand response, pollution management, and a generic stage-by-stage S-MPG the nature of what an S-MPG looks like in practice, and illustrate the performance of our computation and learning algorithms in discovering NE.

II. Related Work

Potential games were introduced in seminal work [8]. Learning in single stage potential games has a rich history with the works of [9], [10] and [11] being notable examples.

An extension of the one-shot strategic game is the multi-stage dynamic game, where agents have an underlying state and the problem faced by each agent can be modeled as a Markov decision process, first studied in [12]. Examples of such games can be found in [13], [14], and [15]. The natural question of multi-agent reinforcement learning (MARL) is considered in well-known work such as [16], [17] and [18], which all assume that the other agents' strategies are known well enough to determine a best response in the manner of a min-max strategy. A survey of MARL is available in [19]. However, the results are quite limited without introducing additional structure on the nature of the game.

Under the structural assumption of a potential game, one train of work considers the state-based game approach, wherein agent actions do not impact transitions, enabling them to take *myopic* actions to maximize their immediate reward [20], [21]. Similarly, [22] considers the specific case of deterministic transitions for greater tractability. Other work assumes that a condition similar to S-MPG holds, without attempting to determine if it does so for any game, and shows that gradient play converges under this setting [6]. [23] describes a solution to the Markov potential game without any requirement of the "strong" property, however they are unable to characterize the rate of convergence. Further, they provide no sufficiency condition for the structure either.

Closest to our work are approaches that determine structural properties on the rewards and transitions that enable tractability of a Markov potential game. Unfortunately, some work such as [7] and [4] have analytical errors (that we detailed in the supplementary material [24]), while [5] utilizes a strong condition on the nature of the value function that cannot realistically be verified for a given game. Thus, the question of sufficient structure that enables provable convergence of multi-agent computation or learning in dynamic potential games is still unaddressed.

III. Preliminaries

We consider a Markov dynamic game where each agent interacts with others by taking actions in a dynamic environment over an infinite horizon with discrete time steps. The agents collect a stage payoff depending on their actions and the current state of the environment, and which also together determine the next state. A Markov dynamic game is a tuple $\Gamma =< \mathcal{S}, \mathcal{A}, \{r_i\}_{i\in[I]}, P, \gamma >$ defined as follows: (1) agents are denoted by $i \in [I] := \{1, 2, \ldots, I\}$, (2) the globally observed state space denoted as $\mathcal{S}$ is a set of finitely many states, (3) the action space for agent $i \in [I]$ denoted as $\mathcal{A}_i$ is a set of finitely many actions and the set of joint actions is denoted as $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_I$, (4) the transition probability is denoted as $\mathbb{P}(s'|s, a)$ from state s to state s' given a joint action $a = (a_1, \ldots, a_I) \in \mathcal{A}$, (5) the stage payoff function of agent i is denoted as $r_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for all $i \in [I]$. We focus on stationary policies, i.e., for each agent i, denote by $\pi_i(s, a_i) \in [0, 1]$ the probability that agent i takes action $a_i \in \mathcal{A}_i$ at state $s \in \mathcal{S}$. We denote a joint policy by $\pi = (\pi_1, \ldots, \pi_I)$, and the set of joint policies by $\Pi := \Pi_1 \times \cdots \times \Pi_I$ where $\Pi_i := \prod_{s\in\mathcal{S}} \Delta(\mathcal{A}_i)$ for all $i \in [I]$ where $\Delta(X)$ is a distribution over X. We also denote a joint policy of all agents other than agent i by π_{-i}, (by abuse of notation) denote $\pi = (\pi_i, \pi_{-i})$ and also denote $\Pi_{-i} = \prod_{j\neq i} \Pi_j$. Moreover, we denote a joint policy in step $k \in \mathbb{N}$ by $\pi^k = (\pi_1^k, \ldots, \pi_I^k)$ when necessary. Define a value function V_i^π as

$$V_i^\pi(s) := \mathbb{E}_s^\pi \big[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) \big] \tag{1}$$

for all $i \in [I]$, $s \in \mathcal{S}$, $\pi \in \Pi$.

Definition 1. *A joint policy* $\pi^* := (\pi_1^*, \dots, \pi_I^*) \in \Pi$ *is called a Nash equilibrium (NE) if* $V_i^{\pi^*}(s) \geq V_i^{\pi_i, \pi_{-i}^*}(s)\ \forall s \in S,\ i \in [I],\ \pi_i \in \Pi_i$. *We also say that* $\pi^\epsilon := (\pi_1^\epsilon, \dots, \pi_I^\epsilon) \in \Pi$ *is an* ϵ*-NE if* $V_i^{\pi^\epsilon}(s) + \epsilon \geq V_i^{\pi_i, \pi_{-i}^\epsilon}(s) \quad \forall s \in S,\ i \in [I],\ \pi_i \in \Pi_i$.

It is well-known that there exists a Nash equilibrium in discounted stochastic games [25]. A static game can be considered as a simplified type of stochastic game that has a single state and no state transitions. Hence, a reward function r_i reduces to depend only on actions, i.e., $r_i : \mathcal{A} \to \mathbb{R}$ for all $i \in [I]$. We introduce a (static) potential game, which we extend to a dynamic setting in the next section.

Definition 2. *A (static) game is called a potential game (PG) if there exists a function* $\phi : \mathcal{A} \to \mathbb{R}$*, called potential, that satisfies the following condition:* $r_i(a_i, a_{-i}) - r_i(a_i', a_{-i}) = \phi(a_i, a_{-i}) - \phi(a_i', a_{-i})$ *for all* $a_i, a_i' \in \mathcal{A}_i, a_{-i} \in \mathcal{A}_{-i}$ *and* $i \in [I]$.

In words, in a potential game, a unilateral change of action by agent i gives him the same reward with r_i as a unilateral change in ϕ. Since, ϕ is agent independent, one can view this as all the agents cooperating to maximize a joint potential function ϕ.

IV. Markov Potential Games

Consider the Markov dynamic game from the previous section, $\Gamma =< \mathcal{S}, \mathcal{A}, \{r_i\}_{i\in[I]}, P, \gamma >$, with finite action sets and state spaces played over an infinite horizon with discount factor γ. We call this game a **Markov Potential Game** (MPG) if there exists a function $\Lambda = \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we will call it the **Auxiliary Potential Function** such that

$$\Lambda^{\pi_i, \pi_{-i}}(s) - \Lambda^{\pi_i', \pi_{-i}}(s) = V_i^{\pi_i, \pi_{-i}}(s) - V_i^{\pi_i', \pi_{-i}}(s) \quad (2)$$

for every i in each state s. We denote the MPG by $\Gamma =< \mathcal{S}, \mathcal{A}, \{R_i\}_{i\in[I]}, P, \Lambda, \gamma >$.
The following definition of a path will be useful to prove optimality results in MPGs.

Definition 3. *A path is a sequence of joint policies* $\rho = (\pi^1, \pi^2, \pi^3, \dots)$ *such that for every* $k \in \mathbb{N}$, $\pi^{k+1} = \{\pi_{i(k)}', \pi_{-i(k)}^k\}$ *obtained from* π^k *by allowing a single agent* $i(k)$ *(single deviator in step* k*) to change its policy of agent* $i(k)$.

We can then make an assertion on the existence of a deterministic NE in the potential game setting.

Theorem 1. *Every MPG,* Γ *has a deterministic Nash policy.*

While this result seems intuitive, the fact that we have state as well as action means that we cannot directly use the argument of finite improvements from [8], but need a slightly more nuanced extension to the Markov game setting. The proof is presented in supplementary material [24].

The next result tells us that if we can find a maximizer for the auxiliary potential function, then we have found a Nash equilibrium. Since Λ maps from $\mathcal{S} \times \Pi \to \mathbb{R}^{|S|}$, we define a partial ordering on $\mathbb{R}^{|S|}$; $x \leq y$, $x = \{x_1, x_2 \dots x_{|S|}\}, y = \{y_1, y_2 \dots y_{|S|}\} \in \mathbb{R}^{|S|}$ if and only if $x_i \leq y_i$ for $i \in \{1, 2, \dots |S|\}$. If z is the *maximizer* of Λ, then for any $x \in \Pi$, $\Lambda^z(s) \geq \Lambda^x(s)$ for every $s \in \mathcal{S}$. Suppose Λ admits a maximizer and suppose further that this maximizer happens to be deterministic, then we will call this policy the *optimal deterministic joint policy*.

Corollary 1. *An optimal deterministic joint policy* $\pi^* = (\pi_1^*, \dots, \pi_I^*)$ *implies that* π^* *is a Nash equilibrium of Markov potential game* Γ.

Corollary 2. *An* ϵ*-optimal deterministic joint policy* $\pi^\epsilon = (\pi_1^\epsilon, \dots, \pi_I^\epsilon)$ *implies that* π^ϵ *is an* ϵ*-Nash equilibrium of Markov potential game* Γ.

The next lemma from [5] provides us with a characterization of value functions and serves as an equivalent definition of MPG.

Lemma 1. *If* $V_i^{\pi_i, \pi_{-i}}(s)$ *is the pay-off to go for agent* i *at state* s *under policy* $\pi = \{\pi_i, \pi_{-i}\}$ *and* $\Lambda^{\pi_i, \pi_{-i}}(s)$ *is the Auxiliary Potential function for the MPG, then*

$$V_i^{\pi_i, \pi_{-i}}(s) = \Lambda^{\pi_i, \pi_{-i}}(s) + U_i^{\pi_{-i}}(s) \quad (3)$$

In words, the lemma states that any value function in an MPG can be decomposed into a "Potential component" where the players essentially collaborate to improve the joint value and a second component which is both implicitly and explicitly independent of the player's actions. We present the proof in the supplementary material [24].

Corollary 3. *Suppose each stage is a potential game, (we will call them* ***stage potential games****), then*

$$r_i(s, a_i, a_{-i}) = \phi^{(a_i, a_{-i})}(s) + u_i(s, a_{-i})$$

We will define the *potential value function*, $\Phi : \mathcal{S} \times \Pi \to \mathbb{R}$ by

$$\Phi^\pi(s) = \mathbb{E}_{a\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t \phi_{s_t,}(a) | s_0 = s\right] \quad (4)$$

We wish to find sufficient conditions under which, Φ given by (4) is a potential function for the Markov game $\Gamma =< \mathcal{S}, \mathcal{A}, \{r_i\}_{i\in[I]}, P, \gamma >$. In other words, we are looking for sufficient conditions under which $\Phi = \Lambda$.

When Φ is a potential function we know a deterministic Nash equilibrium exists. Moreover, maximizing Φ, is equivalent to maximizing a Markov Decision Process, here it is well known that MDPs have a deterministic solution that yields a global maxima for Φ. Such a solution can be obtained from value or policy iterations, thus, greatly simplifying the computation of Nash equilibria. We will call the class of Markov Potential games for whom the Potential value function is an auxiliary potential function, Strong Markov Potential games (S-MPG).

A. Sufficiency conditions for Strong Markov Potential Games

As discussed in the previous section, we are interested in the conditions under which our game is a S-MPG. We will begin by examining the conditions outlined in [7].

SER-SIT condition:

Following [7], [26] and [27], we consider *Separable Reward - State independent transition* games with the following additional conditions.

Condition 1. *The rewards are separable in the following sense,*

$$r_i(s, a_i, a_{-i}) = r_i^0(s) + r_i^1(a_i, a_{-i})$$

and the probability transition matrix is state independent,

$$\mathbb{P}(s, a_i, a_{-i}) = \mathbb{P}(a_i, a_{-i})$$

1) The partial reward, $r_i^1(a_i, a_{-i})$ is a potential game. 2) The pay-off, $\langle \mathbb{P}(.|a_i, a_{-i}), r_i^0(.) \rangle$ for each i follows a potential game.

The following theorem from [7] gives sufficient conditions under which a game with the SER-SIT structure is a potential game.

Proposition 1. *Under Condition 1, the SER-SIT game is a S-MPG.*

The proof can be found in the supplementary material [24].

As a consequence of the SERSIT conditions we now have at least one verifiable condition under which an S-MPG exist.

Action Independent Transitions:

We call games where the probability transition kernel does not depend on the action set of the players state-based games [20] and the probability kernels, action independent transition kernels.

Condition 2. *The rewards at each stage follow a stage potential game and the transition probabilities are action independent.*

Proposition 2. *Under Condition 2, the reward at each stage can be written as:*

$$r_i(s, a_i, a_{-i}) = \phi(s, a_i, a_{-i}) + u_i(a_{-i}, s)$$

and the state-based game is an S-MPG.

We now have two independent verifiable conditions under which a Strong Markov Potential Game may exist.

B. Examples:

While there are many practical examples of *state based games*, see for example [20] there are relatively few examples of SER-SIT games. We provide two important examples, as mentioned in the introduction.

Demand Response Marketplace: We consider a marketplace with N demand response providers who can choose to contribute $a_i \in \{0, 1, \dots |A|\}$ levels of energy reduction to the grid. The state of the system reflects the demand for that time period. We denote the states by $\mathcal{S} = \{0, 1, \dots |\mathcal{S}|\}$, with 0 indicating that the grid is in dire need of demand reductions, to $|\mathcal{S}|$ indicating that the current supply is sufficient to meet the current demand. In return for producing a level of output (demand reduction) a_i, the DR provider receives reward $r_i(s, a_i, a_{-i}) := a_i(f(a_i, a_{-i}) - \delta^s)$ for an appropriate function f. The states of this system evolve according to $s' = g(\sum_{i=1}^N a_i) + w$ where w is a common discrete noise term that models uncertainty due to a variety of factors including weather events. Such a game can be modeled as a SER-SIT game when f and g are chosen appropriately. The system simulation is presented in Section VII.

Pollution Tax Model: The environment has two states, *pollution free* and *polluted*, and we use s_0 and s_1 to denote these states. There are two generating firms that have two actions available to them, *clean, C* and *dirty, D*. If either one of the firms takes the *dirty* action, the environment in the next time slot goes to state, s_1. If both firms choose the clean action, the firm goes to, s_0. The firms are taxed equally in the polluted state by T. The firms earn reward g_1 and g_2 per unit produced, action C produces one unit while D produces two units. The payoff function is therefore given by $r_i(s, a_1, a_2) = g_i + \mathbf{1}_{\{a_i=D\}} g_i - \mathbf{1}_{\{s=s_1\}} T$ where $\mathbf{1}_{\{E\}}$ is the indicator function of E. The transition kernel, $\mathbb{P}(s_0|s, a_1, a_2) = \mathbb{P}(s_0|a_1, a_2) = 1$ only when $a_1 = C$ and $a_2 = C$. Clearly, the reward obeys a SER structure while the transitions are state independent and it is easy to check that the reward at each stage obeys a potential structure. The system simulation is presented in Section VII.

V. Computation of NE in Strong Markov Potential Games

Throughout this section, we assume that the reward at each stage follows a stage potential game. Further we assume that either Assumption (1) or (2) hold. Hence, our potential value function, (4) is an auxiliary potential function. We now have a Strong Markov Potential Game denoted by, $\Gamma =< \mathcal{S}, \mathcal{A}, \{R_i\}_{i\in[I]}, P, \Phi, \gamma >$. We will present computational methods for convergence of Markov potential games and prove that these methods identify an ϵ-Nash equilibrium. Note that we use $\|\cdot\|$ to refer to the $\|\cdot\|_\infty$ norm.

A. Centralized Computation

Centralized Value Iteration:

Our first setting can be thought of as the case when a central administrator wants to ensure the optimal outcome for all the agents and maximizes the potential value function in order to obtain this outcome.

Since our potential value function, Φ is the value of an MDP, we may use value or policy iteration to find the joint $\epsilon-$optimal policy. We know by Corollary 2 that it suffices to find an ϵ-optimal joint policy, π^ϵ with respect to Φ when finding an ϵ-Nash equilibrium. To this end we use value iteration over the joint actions of all the agents.

Let $\Phi(s) = \Phi^\pi(s)$ be the potential value function under some fixed policy π. Define a mapping H by,

$$(H\Phi)(s) = \max_{a\in\mathcal{A}} (\phi(s, a) + \gamma \sum_{s'\in S} \mathbb{P}(s'|s, a)\Phi(s'))$$

for all $s \in \mathcal{S}$. Readers will note that this is the familiar Bellman optimality operator on Φ.

Let us begin with some initialization for $\Phi_0(s)$ for all $s \in \mathcal{S}$ for some initial policy π^0. At any time $t > 0$, let $\Phi_t = H^t\Phi_0$ i.e, the optimality operator applied to our initial

point t times. It is well known that the optimality operator is monotonic and a contraction map in the $\|\cdot\|$ norm.

Algorithm 1 Centralized Computation

Input: A Markov potential game (Γ and ϕ) and ϵ
Let $\Phi_0(s) = 0$ for all $s \in \mathcal{S}$, $t = 0$, $c := \|H\Phi_0 - \Phi_0\|$
repeat
  $\Phi_{t+1}(s) \leftarrow H\Phi_t(s)$ for all $s \in \mathcal{S}$
  Increment t by 1
until $\max(\|\Phi_{t+1}(s) - \Phi_t(s)\|, \frac{\gamma^t}{1-\gamma}c) < \frac{1}{2}\epsilon$
return $\pi^\epsilon \leftarrow \arg\max_{a\in\mathcal{A}} \phi(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a)\Phi_t(s')$

Theorem 2. *For any $\epsilon > 0$, Algorithm 1 generates an ϵ-Nash equilibrium π^ϵ.*

The proof for this theorem may be found in the supplementary material [24].

Component-wise Value Iteration:

Suppose an administrator wishes to maximize his cost but may only improve a single component at a time. Under such a setting we are interested in the existence of critical points i.e, points where there can be no improvement by changing a single component of the value function. In this setting, we consider an MDP with value function

$$\Phi^\pi(s) := \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) | s_0 = s, a_t \sim \pi(s_t)\Big].$$

The state space of our MDP is $\mathcal{S}$; the action set $\mathcal{A}$ can be broken into I independent components such that $\mathcal{A} = \prod_{i=1}^{I} \mathcal{A}_i$. Formally, our objective is to find policies $\pi^* := \{\pi_1^*, \pi_2^*, \dots \pi_I^*\}$ such that $\Phi^{\pi^*}(s) \geq \Phi^{\pi_i, \pi_{-i}^*}(s)$ for any $s \in \mathcal{S}$, of our function Φ. Here, we show a possible decentralized component-wise iteration procedure to find critical points in our problem.

Define $C_i^\pi \Phi$ as,

$$C_i^\pi \Phi(s) = \max_{b_i \in \mathcal{A}_i} \phi(s, b_i, \pi_{-1}(s)) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s, b_i, \pi_{-1}(s))\Phi(s'). \tag{5}$$

$$\pi_i^{\text{br}}(\pi, \Phi) = \arg\max_{b_i \in \mathcal{A}_i} \phi(s, b_i, \pi_{-1}(s)) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s, b_1, \pi_{-1}(s))\Phi(s'). \tag{6}$$

In words, we begin with some estimate of our value function Φ and a policy π. We select a component in a round robin fashion and improve its value using the operator C_i^π for the value and π_i^{br}. Note, during regular value iteration it is unnecessary to maintain the policy at time t, however, in the component-wise case this becomes necessary since our next value not only depends on the one-step improvement of the i^{th} component but also the policy π_{-i}.

Lemma 2. *Suppose we begin with a deterministic policy π and a corresponding potential value function for this policy given by Φ^π. Then Algorithm 2 converges to a Nash equilibrium asymptotically.*

Algorithm 2 Component-wise Value Iteration

Input: An MDP $M := \langle \mathcal{S}, \prod_{i=1}^{I} \mathcal{A}_i, \phi, P, \gamma \rangle$,
Initialize : $\Phi_0(s)$ for all $s \in \mathcal{S}$, $t = 1$ and $\pi^0 = \pi$
repeat
  $i \leftarrow t$ modulo I
  $\Phi_{t+1}(s) \leftarrow C_i^{\pi^t} \Phi_t$ for all $s \in \mathcal{S}$
  $\pi^{t+1} \leftarrow \pi_i^{br}(\pi^t, \Phi_t)$
  Increment t by 1
until $\max(\|\Phi_{t+1}(s) - \Phi_t(s)\|, \frac{\gamma^t}{1-\gamma}c) < \epsilon$
return Φ^t, π^t

The previous lemma showed us that the component-wise value iteration converges asymptotically to a Nash equilibrium but did not give us a rate of convergence. The next lemma shows that the rate is in fact geometric.

Lemma 3. *C_i^π is a contraction with Lipschitz constant γ in the infinity norm when π is fixed; as a result, different trajectories of value functions will converge to a common trajectory at a geometric rate.*

The proofs for both Lemma 2 and 3 may be found in the supplementary material [24].

B. Decentralized Computation

The following section consists of the more realistic setting where the agents have different utility functions but are playing a Markov potential game.

Asynchronous Computation:

We begin by using a natural extension of the concept of finite improvement paths from [8]. Define the Bellman optimality operator in one component by: $T^i V_i(s) := \max_{a_i \in \mathcal{A}_i} R_i(s, a_i, \pi_{-i}) + \gamma \sum_{s'\in S} P(s'|s, a_i, \pi_{-i}) V_i(s'))$.

Denote the Q-function $Q_i^{\pi_i^*, \pi_{-i}}(s, a_i) = R_i(s, a_i, \pi_{-i}) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s, a_i, \pi_{-i}) V_i^{\pi_i^*, \pi_{-i}}(s')$ for all $s \in \mathcal{S}$, $a_i \in \mathcal{A}_i$, $i \in [I]$ and $\pi_{-i} \in \Pi_{-i}$ where π_i^* is an optimal policy with respect to π_{-i}. Let $\bar{\epsilon}$ denote the minimum separation between agents' optimal Q-functions with respect to π_{-i}, defined as

$$\bar{\epsilon} = \min_{Q_i^{\pi_i^*, \pi_{-i}} \neq Q_i^{\pi_i^*, \pi_{-i}}} |Q_i^{\pi_i^*, \pi_{-i}}(s, a_i) - Q_i^{\pi_i^*, \pi_{-i}}(s, b_i)| \tag{7}$$

minimized over all s, a_i, b_i.

Suppose we choose ϵ such that $0 < \epsilon < \frac{1}{2}\bar{\epsilon}$ and if computation and learning (such as VI, PI or Q-learning) of asynchronous play is within the tolerance level of ϵ, then it characterizes an ϵ-improvement path (defined analogously to an improvement path) because ϵ is smaller than the minimum separation $\bar{\epsilon}$ which distinguishes optimal and sub-optimal actions.

Note that the following theorem holds true for any Markov potential game and not just S-MPGs.

Theorem 3. *Algorithm 3 generates an ϵ-Nash equilibrium π^ϵ for all $0 < \epsilon < \frac{1}{2}\bar{\epsilon}$.*

Algorithm 3 Asynchronous Computation

Input : A Markov potential game (Γ and ϕ) and ϵ
Initialize : A joint policy π_0 and $k = 0$
repeat
  Choose an agent $i = i(k) \in [I]$ and fix $\pi_{-i(k)}$
  Initialize : $V_{i,0}(s) \leftarrow 0$ for all $s \in \mathcal{S}$, $c \leftarrow \|T^i V_{i,0} - V_{i,0}\|$, $t \leftarrow 0$
  repeat
    $V_{i,t+1}(s) \leftarrow T^i V_{i,t}(s)$ for all $s \in \mathcal{S}$
    Increment t by 1
  until $\max(\|V_{i,t+1}(s) - V_{i,t}(s)\|, \frac{\gamma^t}{1-\gamma}c) < \frac{1}{8}\epsilon$
  $\pi_{i,k+1} \leftarrow \arg\max_{a_i \in \mathcal{A}_i} R_i(s, a_i, \pi_{-i(k)}) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a_i, \pi_{-i(k)}) V_{i,t}(s')$
  $\pi_{k+1} \leftarrow (\pi_{i,k+1}, \pi_{-i,k+1}) := (\pi_{i,k+1}, \pi_{-i(k)})$
  $\hat{V}_i^{\pi_{k+1}} \leftarrow V_{i,t}$
  Increment k by 1
until $\hat{V}_{i(k)}^{\pi_{k+1}}(s) \leq \hat{V}_{i(k)}^{\pi_k}(s) + \epsilon \quad \forall s \in \mathcal{S}$ and $i(k) \in [I]$
return $\pi^\epsilon = \pi_K = (\pi_{1,K}, \ldots, \pi_{I,K})$

Distributed Better Response:

Algorithm 2 was a central algorithm to compute critical points on our MDP. Here, we present a way to perform an equivalent operation in the decentralized case where each agent will try to perform a one-step maximization of their value function analogously to the one-step component-wise update. As before, we will need to define a few operators for our algorithm $D_i^\pi \Phi$ as,

$$D_i^\pi V_i(s) = \max_{b_i \in \mathcal{A}_i} r_i(s, b_i, \pi_{-1}(s)) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s, b_i, \pi_{-1}(s)) V_i(s')$$

$$\pi_i^{\text{D-br}}(\pi, V_i) = \arg\max_{b_i \in \mathcal{A}_i} r_i(s, b_i, \pi_{-1}(s)) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s, b_1, \pi_{-1}(s)) V_i(s').$$

And let,

$$\bar{D}_i^\pi V_i(s) = r_i(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s, \pi(s)) V_i(s').$$

Algorithm 4 describes one way to asynchronously perform value iteration in order to obtain a Nash equilibrium. In words, at each time step t, an agent is chosen in a round robin fashion. At time 1, agent 1 is chosen and performs a one-step update of its own utility function V_1^π using D_1^π. Agent 1 can simultaneously compute, $\pi_1^{\text{br}}(\pi, V_1^\pi)$ and broadcast this update to every other agent. Next, every agent, j updates its value using a one step evaluation $\bar{D}_j^{\pi^1} V_j^\pi$, where $\pi^1 := \{\pi_1^{\text{br}}, \pi_{-1}\}$. At any time step $t > 1$, this process repeats until convergence.

Theorem 4. *Algorithm 4 converges to an ϵ Nash equilibrium.*

The proof (please refer [24]), proceeds by establishing the equivalence between centralized component-wise operators and the distributed agent-wise operators, C_i^π and D_i^π respectively.

Algorithm 4 Distributed Better Response

Input: An MPG $M := \langle \mathcal{S}, \prod_{i=1}^I \mathcal{A}_i, \Phi, \{r_i\}_{i \in [I]}, P, \gamma \rangle$,
Initialize : $V_i(s)$ for all $s \in \mathcal{S}$, $t = 1$ and $\pi^0 = \pi$
repeat
  $i \leftarrow t$ modulo I
  $V_{i,t+1}(s) \leftarrow D_i^{\pi^t} V_{i,t}(s)$ for all $s \in \mathcal{S}$
  $\pi^{t+1} \leftarrow (\pi_i^{D-br}(\pi^t, V_{i,t}), \pi_{-i})$
  for $j = \{1, 2, \ldots i-1, i+1, \ldots I\}$ **do**
    $V_{j,t+1}(s) \leftarrow \bar{D}_j^{\pi^{t+1}} V_{j,t}$
  Increment t by 1
until $\max(\max_i \|V_{i,t+1}(s) - V_{i,t}(s)\|, \frac{\gamma^t}{1-\gamma}c) < \epsilon$
return $\{V_{i,t}\}_{i \in \{0,1,\ldots I-1\}}, \pi^t$

VI. Learning in MPGs

Our learning algorithm follows directly from our asynchronous computation algorithm. We will use an off-policy Q-learning method,i.e., at round k, an agent, $i(k)$ is chosen at random. All other agents are assumed to keep their policy constant while the agent learns her (near) best response through Q-learning. Then we proceed to the next agent.

Agent i at time t observes the tuple $\langle s_t, a_t, r_i, s_{t+1} \rangle$ and updates her Q-table as follows,

$$Q_i(s_t, a_t, t+1) = Q_j(s_t, a_t, t) + \alpha_t \big[r_j(s_t, a_t) + \gamma \max_{a'} Q_j(s_{t+1}, a', t) \big] \tag{8}$$

The behavior policy $\pi \in \Delta(\mathcal{A})^{|S|}$ (note this is a joint policy on all the agents) is ergodic in the Markov chain (s, a) with stationary distribution μ. Define, $\mu_{min} := \min_{(s,a)} \mu(s, a) > 0$ and mixing time t_{mix}. We let τ be the time at which total variation distance from the stationary distribution is $\frac{1}{4}$, $\tau = t_{mix} \log \frac{2}{\mu_{min}}$. The step size, $\alpha_t := \frac{h}{t+t_0}$ and $h \geq \frac{4}{\mu_{min}(1-\gamma)}$, $t_0 \geq \max(4h, \tau)$.

Lemma 4 (Theorem 7 in [28])**.** *Let M(π_{-j}) be the MDP for agent j and let ϵ and δ be two positive constants. If the Q-table for j is updated using* (8)*, then: $\|Q(T) - Q^*\| < \epsilon$ with probability at least $1 - \delta$ whenever, $T \geq T_0$. Where $T_0 = \tilde{O}\left(\frac{t_{mix}}{\epsilon^2 (1-\gamma)^5 \mu_{min}^2}\right)$, $\tilde{O}$ suppresses logarithmic factors of $1/\epsilon, 1/(1-\gamma), 1/\mu_{min}$ and t_{mix}.*

Here, note, t_{mix} will scale as $\log |\mathcal{S}||\mathcal{A}|$ and $\frac{1}{\mu_{min}}$ scales as $O(|\mathcal{S}||\mathcal{A}|)$. The learning algorithm extended from asynchronous computation Algorithm 3 that under the framework of *asynchronous play* is:

Theorem 5. *For $0 < \epsilon < \frac{1}{2}\bar{\epsilon}$, $0 < \delta < 1$, suppose each agent, chosen in a round robin schedule, updates their $Q-$table according to* (8) *for an order $\tilde{O}\left(\frac{t_{mix}}{\epsilon^2 (1-\gamma)^5 \mu_{min}^2}\right)$ at each step k. Algorithm 5 generates an ϵ-Nash equilibrium π^ϵ with probability at least $1 - \delta$.*

Algorithm 5 Asynchronous Learning

Input : ϵ, δ
Initialize : A joint policy π_0, $k \leftarrow 0$, $R_{\max} \leftarrow \max_{i,s,a} R_i(s,a)$, $K \leftarrow \epsilon \frac{R_{\max}}{1-\gamma} I$, $\delta_0 \leftarrow \frac{1}{K}\delta$
repeat
  Choose an agent $i = i(k) \in [I]$ and fix $\pi_{-i(k)}$
  Execute update (8) in MDP $M(\pi_{-i(k)})$ with input $\epsilon, \delta_0, \pi_{-i(k)}$
  Obtain ϵ-optimal policy $\pi_{i(k),k+1} \leftarrow \pi_i$ and ϵ-optimal value function $V_{i(k)} \leftarrow V_i$
  $\pi_{k+1} \leftarrow (\pi_{i,k+1}, \pi_{-i,k+1}) := (\pi_{i(k),k+1}, \pi_{-i(k)})$
  $\hat{V}_{i(k)}^{\pi_{k+1}} \leftarrow V_{i(k)}$
  Increment k by 1.
until $\hat{V}_{i(k)}^{\pi_{k+1}}(s) \leq \hat{V}_{i(k)}^{\pi_k}(s) + \epsilon \quad \forall s \in \mathcal{S}$ and $i(k) \in [I]$
return $\pi^\epsilon = \pi_K = (\pi_{1,K}, \ldots, \pi_{I,K})$

Note that since each agent needs to update their policy using Q-learning, the total number of iterations required to reach Nash equilibrium will be multiplied by the number of agents.

VII. Numerical Studies

SER-SIT game (Demand Response Marketplace): Our first case is DR Market described in Section IV-B. Suppose there are N grid assets such as demand response aggregators. Each grid asset can produce $a_i \in \{0, 1, \ldots, 4\}$ amount of products, for example, load reduction from demand response aggregators. State space $\mathcal{S} = \{0, 1, \ldots, 4\}$ represents levels of emergency where state 0 represents that the system is most strained and that state 4 represents that the system is least stressed. Reward for each asset (agent) i is given by $r_i(s, a_i, a_{-i}) = a_i(f(a_i, a_{-i}) - c^s)$ where $f(a_i, a_{-i})$ is a function of every asset's action and c is a constant. It immediately follows that $\phi(a_i, a_{-i}) = (\prod_{i=1}^N a_i)(f(a_i, a_{-i}) - c^s)$ is an ordinal potential function for each $s \in \mathcal{S}$. If $f(a_i, a_{-i}) = a_i(\alpha - \beta \sum_{i=1}^N a_i)$ with constants $\alpha = 2, \beta = 0.25$ and $c = 1.25$, then ϕ is an exact potential function where ϕ is defined as $\phi(a_i, a_{-i}) = \alpha \sum_{i=1}^N a_i - \beta \sum_{i=1}^N a_i^2 - \beta \sum_{1 \leq i < j < N}^N a_i a_j - \sum_{i=1}^N c^s$. Suppose that next state s' is given by $s' = a_1 + a_2 - w$ where $w \sim uniform\{0, 1, 2, 3, 4\}$ with probability 0.9 and $s' \sim uniform\{0, 1, 2, 3, 4\}$ with probability 0.1. Given the structures, it can be shown that the grid asset management is a SER-SIT game and each state is a potential game. For ease of presentation of the heat map of actions we restrict the assets to 2. We show convergence of Algorithms 4 and 5 in Figure 1 and 2 respectively and heat map of action-value function for grid assets in Figure 3 (heat map for grid assets is identical due to identical reward structure). It is observed that both grid assets take the largest load reduction in state 0 (most stressed system state) while they take the smallest load reduction in state 4 (least stressed system state).

We further study this market over a synthetic Texas transmission grid model [29]. Here, the state of the system is mapped to the offered load (demand) at each of the nodes in the grid, and the market maker solves the DC Optimal Power Flow (DC-OPF) problem to determine the location marginal prices (LMP) as the (wholesale) electricity price at each node. It is well known that large demand at certain nodes can trigger very high LMPs [29]. We consider whether the demand-response market will mitigate these high LMPs.

Thus, for each $s \in \mathcal{S}$ in our setup, there is a corresponding nodal demand vector d_s. If grid asset i takes an action $a_i = j$, it is equivalent to its commitment to reduce $j\%$ of its demand in selected locations. We consider the impact on average LMP (over the grid) according to (i) a do-nothing policy denoted as π^0, i.e., the two DR agents take action $a_i = 0$ for all states, and (ii) the Nash equilibrium policy π^* of the grid asset market as shown in the heat map of Figure 3. The trajectories of LMPs in the two cases are illustrated in Fig. 4, where is clear that the learned strategy for DR significantly reduces the average LMPs.

SER-SIT game (Pollution Tax Model): Our second case is the pollution tax model from Section IV-B. We consider $g_1 = g_2 = 2$ and the tax incurred as $T = 4$. As seen in Figures 6 and 7, the system converges to a Nash equilibrium where players choose to use the *clean* actions in both states.

State based game: We finally present an illustration of

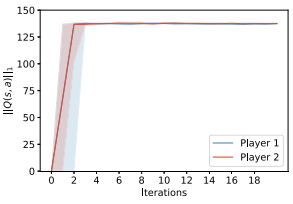

Fig. 1: Distributed Better Response

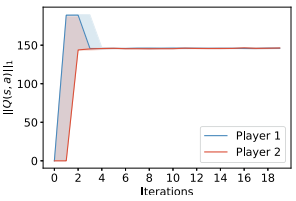

Fig. 2: Asynchronous Learning

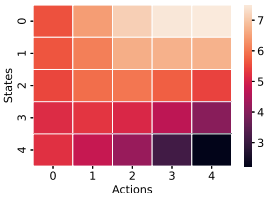

Fig. 3: Heat map of action-value

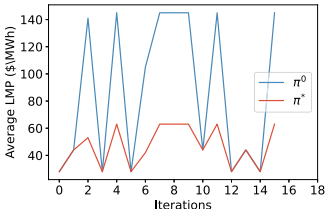

Fig. 4: LMP trajectories

Fig. 5: Note, the convergence of the $L1$ norm in the figure above can happen if and only if the corresponding potential function converges.

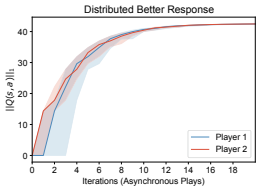

Fig. 6: Distributed Better Response (Algorithm 4)

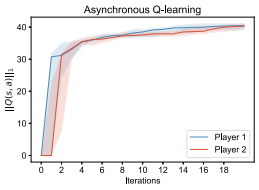

Fig. 7: Asynchronous Learning (Algorithm 5)

a state-based Markov Potential Game in which the state transitions are due to an exogenous Markov chain that does not depend on the player actions. A variety of energy markets can be modelled in this manner, where the state (such as the current weather) provides a context for player interaction.

For illustration, we consider a state-based MPG that is composed of two canonical strategic games, with the state being which game is currently being played. The games chosen are *Prisoner's Dilemma* and *Bach or Stravinsky*, denoted as s_0 and s_1, respectively. The players actions and rewards are shown in Tables I and II.

Actions	C	D
C	-1, -1	-6, 0
D	0, -6	-4, -4

TABLE I: State s_0

Actions	C	D
C	2,1	0,0
D	0,0	1,2

TABLE II: State s_1

The transition probability matrix is given by $\mathbb{P}[s' = s_0|s, a_1, a_2] = \mathbb{P}[s' = s_0] = 0.6$ and $\mathbb{P}[s' = s_1|s, a_1, a_2] = \mathbb{P}[s' = s_1] = 0.4$. In figures 8 and 9 we plot Q-values with respect to L1 norm over iterations to show the convergence of Algorithms 4 and 5.

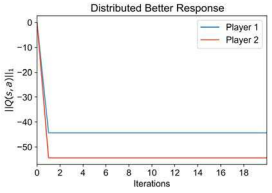

Fig. 8: Distributed Better Response (Algorithm 4)

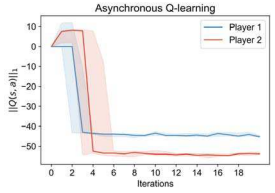

Fig. 9: Asynchronous Learning (Algorithm 5)

VIII. Conclusion

We considered multi-agent marketplaces in the context of DERs. We modeled the system as a Markov Potential Game and characterized sufficiency conditions under which an MPG can be treated as an MDP with multidimensional actions controlled by different agents. We constructed centralized and distributed algorithms to compute Nash equilibria, and developed an MARL variant based on these algorithms. This enables the expansion of the problem space over which we can determine Nash equilibria via MARL beyond simple min-max approaches applicable to zero sum situations. We used several games in the context of DER marketplaces as examples to demonstrate the efficacy of our methods.

References

[1] B. Xia, S. Shakkottai, and V. Subramanian, "Small-scale markets for a bilateral energy sharing economy," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 3, pp. 1026–1037, 2019.

[2] J. Li, B. Xia, X. Geng, H. Ming, S. Shakkottai, V. Subramanian, and L. Xie, "Energy coupon: A mean field game perspective on demand response in smart grids," in *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, 2015, pp. 455–456.

[3] B. Xia, H. Ming, K.-Y. Lee, Y. Li, Y. Zhou, S. Bansal, S. Shakkottai, and L. Xie, "Energycoupon: A case study on incentive-based demand response in smart grid," in *Proceedings of the Eighth International Conference on Future Energy Systems*, 2017, pp. 80–90.

[4] D. H. Mguni, Y. Wu, Y. Du, Y. Yang, Z. Wang, M. Li, Y. Wen, J. Jennings, and J. Wang, "Learning in nonzero-sum stochastic games with potentials," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021.

[5] S. Leonardos, W. Overman, I. Panageas, and G. Piliouras, "Global convergence of multi-agent policy gradient in Markov potential games," in *International Conference on Learning Representations*, 2022.

[6] R. Zhang, Z. Ren, and N. Li, "Gradient play in multi-agent Markov stochastic games: Stationary points and convergence," 2021.

[7] J. A. Potters, T. Raghavan, and S. H. Tijs, "Pure equilibrium strategies for stochastic games via potential functions," in *Advances in Dynamic Games and Their Applications*. Springer, 2009, pp. 1–12.

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

[9] D. Fudenberg, F. Drew, D. K. Levine, and D. K. Levine, *The theory of learning in games*. MIT press, 1998, vol. 2.

[10] Y. Ermoliev and S. Flam, "Learning in potential games," IIASA, Laxenburg, Austria, IIASA Interim Report, June 1997.

[11] A. Heliou, J. Cohen, and P. Mertikopoulos, "Learning with bandit feedback in potential games," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[12] L. S. Shapley, "Stochastic games," *Proceedings of the National Academy of Sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.

[13] J. Filar and K. Vrieze, *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

[14] N. Hemachandra and K. S. M. Rao, "On pure Nash equilibria in stochastic games," https://www.ieor.iitb.ac.in/files/VVS_TR_Jan2011.pdf, [Online].

[15] A. Das, S. N. Krishna, L. Manasa, A. Trivedi, and D. Wojtczak, "On pure Nash equilibria in stochastic games," in *TAMC*, 2015.

[16] J. Hu and M. P. Wellman, "Multiagent reinforcement learning: Theoretical framework and an algorithm," in *ICML*, 1998.

[17] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *ICML*, 1994.

[18] L. Buşoniu, R. Babuška, and B. D. Schutter, "Multi-agent reinforcement learning: An overview," *Innovations in multi-agent systems and applications-1*, pp. 183–221, 2010.

[19] Y. Yang and J. Wang, "An overview of multi-agent reinforcement learning from game theoretical perspective," *arXiv preprint arXiv:2011.00583*, 2020.

[20] J. R. Marden, "State based potential games," *Automatica*, vol. 48, no. 12, pp. 3075–3088, 2012.

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

[22] S. Zazo, S. V. Macua, M. Sánchez-Fernández, and J. Zazo, "Dynamic potential games with constraints: Fundamentals and applications in communications," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3806–3821, 2016.

[23] R. Fox, S. McAleer, W. Overman, and I. Panageas, "Independent natural policy gradient always converges in Markov potential games," *arXiv preprint arXiv:2110.10614*, 2021.

[24] D. Narasimha, K. Lee, D. Kalathil, and S. Shakkottai, "Multi-agent learning via Markov potential games in electricity marketplaces," https://www.dropbox.com/s/2pi4z61qszs8tur/Arxiv_Techreport_potential_games.pdf?dl=0.

[25] A. M. Fink *et al.*, "Equilibrium in a stochastic n-person game," *Journal of science of the Hiroshima University, series ai (mathematics)*, vol. 28, no. 1, pp. 89–93, 1964.

[26] T. Parthasarathy, S. H. Tijs, and O. J. Vrieze, "Stochastic games with state independent transitions and separable rewards," in *Selected Topics in Operations Research and Mathematical Economics*, 1984.

[27] M. J. Sobel, "Myopic solutions of Markov decision processes and Stochastic games," *Oper. Res.*, vol. 29, pp. 995–1009, 1981.

[28] G. Qu and A. Wierman, "Finite-time analysis of asynchronous stochastic approximation and Q-Learning," in *Conference on Learning Theory*. PMLR, 2020, pp. 3185–3205.

[29] K. Lee, X. Geng, S. Sivaranjani, B. Xia, H. Ming, S. Shakkottai, and L. Xie, "Targeted demand response for mitigating price volatility and enhancing grid reliability in synthetic Texas electricity markets," *iScience*, vol. 25, no. 2, p. 103723, 2022.