2020 59th IEEE Conference on Decision and Control (CDC)
Jeju Island, Republic of Korea, December 14-18, 2020

Multi-Agent Reinforcement Learning in Cournot Games

Yuanyuan Shi and Baosen Zhang

Abstract— **In this work, we study the interaction of strategic agents in continuous action Cournot games with limited information feedback. Cournot game is the essential market model for many socio-economic systems where agents learn and compete without the full knowledge of the system or each other. We consider the dynamics of the policy gradient algorithm, which is a widely adopted continuous control reinforcement learning algorithm, in concave Cournot games. We prove the convergence of policy gradient dynamics to the Nash equilibrium when the price function is linear or the number of agents is two. This is the first result (to the best of our knowledge) on the convergence property of learning algorithms with continuous action spaces that do not fall in the no-regret class.**

I. Introduction

Reinforcement Learning (RL) has yielded impressive results in various sequential decision-making problems in recent years. These successes include playing games with super-human performance [1], solving complex robotic tasks [2], [3] and autonomous driving [4]. Some of these applications focus on a single agent, but many applications of interest consider groups of the agents (or players). In the latter case, the agents operate in a common environment, each of them interacting with the environment and other agents. This multi-agent setting contains a rich set of models and has received significant attention in the past several years (see [5] and references within).

In this paper, we focus on the dynamics of learning agents, where each agent aims to optimize its long-term expected return by repeatedly participating in the game. This question has been mostly studied at two extremes, where the agents are either fully cooperative [6] or they are fully competitive (i.e. zero-sum games) [7]. Instead of these extremes, we focus on the case of *general-sum games*, where the agents are self-interested but no adversarially so.

General-sum games have been widely used to model the interactions and competition in cyber-physical systems because of the wide range of individual goals and possible relationships between agents [8]. Each agent in the games is self-interested, and reward may be conflicting with others, but often not in a zero-sum manner. Compared to the two extreme cases, there have been relatively few convergence results for general-sum games, partly because of the technical challenge caused by heterogeneous goals and limited information.

This work is partially supported by the National Science Foundation awards CNS-1931718 and ECCS-1807142.

The authors are with the Department of Electrical and Computer Engineering, University of Washington, Seattle, WA 98115 {yyshi,zhangbao}@uw.edu

In this work, we focus on a specific class of general-sum games where the agents undergo *Cournot competitions* [9]. Cournot game has been used to model the energy systems [10], transportation networks [11] and healthcare systems. It is also one of the most prevalent models of firm competition in economics. In the Cournot game model, firms control their production level, which influences the market price. For example, some electricity markets can be thought of as a Cournot game, where the production level is the amount of power produced by the generators, and the price is decided by the total generation bid and the demand. Each generator's payoff is then calculated as the market price multiplying its share of the supply, subtracting its production cost [10].[1] In this example the generators do not cooperate with each other, but the total profit is also not zero.

When learning is not needed, there is a wealth of results for the Cournot competition. For example, when each agent has full information about the game, including the price function and the cost function of all other agents, there are many works characterizing the properties of the Nash equilibrium of the game [12], [13]. However, when learning is involved and agents do not have full information, the properties of the game are not well understood. This is even the case in the simplest setting, where the agents only receive the price from the system as the feedback but do not know the price function form nor the actions of other agents.

To answer what happens when agents learn, we must model how they learn - or more precisely, what type of learning algorithms is used. A key technical challenge is that when learning is used, the Cournot game becomes stochastic. Currently, most works focus on no-regret algorithms [14], [15] because they only require a minimal set of assumptions on the game. In addition, the no-regret definition could be directly translated to the coarse correlated equilibrium condition [16] for a wide range of algorithms (e.g., multiplicative-weight [17], online mirror descent [18], Follow-the-Regularized-Leader [19]). However, while the theoretical properties of no-regret algorithms are attractive, they also limit the applicability of these algorithms. In practice, systems and agents are often *not adversarial* to each other, and the competition is often designed to have specific structures. In many games, it is more natural for players to use myopic policies such as reinforcement learning algorithms that directly aim for profit maximization. In addition, the notation of coarse correlated equilibrium can be quite weak, and sharper results are often desired.

[1] In this a first-order approximation of the locational marginal pricing used by markets in the United States.

978-1-7281-7447-1/20/$31.00 ©2020 IEEE

These algorithms can lead to much better performances than no-regret algorithms, but proving their convergence has proven to be challenging [5] since the coupling between the (continuous) actions of the players must be carefully analyzed. Attempts have been made to discretize the space(e.g., Q-learning [20]) then studying the resulting discrete game, but the dimensionally quickly grows and important features (e.g., convexity) are hard to retain [21], [22].

In this work, we directly work with the continuous action and state space by considering agents use policy gradient learning algorithms. In particular, we assume the class of policies where the actions are parameterized by the mean of distributions (e.g., Gaussian policies). The *major contribution* of this work is in the following: we prove that when the price function is linear or when there are two agents, there is a unique Nash equilibrium (NE) in the *stochastic* Cournot game, and the policy gradient converges exponentially quickly to the NE. This is the first result (to the best of our knowledge) on the convergence property of algorithms with continuous action spaces that do not fall in the no-regret class.

The rest of the paper is organized as follows. Section II covers the background and prior works on standard Cournot games and policy gradient algorithm. Section III describes the stochastic Cournot game and sketches the convergence proof for policy gradient agents in the games. Section IV provides detailed proof of the convergence result. We provide several case studies and investigative experiments in Section V that demonstrate the convergence behavior of agents in various settings. Finally, Section VI concludes the paper and outlines directions for future work.

II. Problem Setup and Preliminaries

A. Cournot Game

Definition 1 (Cournot Game): Consider N players produce homogeneous products in a limited market, where the action space of player i is its production level $x_i \geq 0$. The utility function of player i is denoted as $\pi_i(\boldsymbol{x}) = p(\sum_{j=1}^{N} x_j)x_i - C_i(x_i)$, where p is the market price (inverse demand) function that maps the total production quantity to a price in $\mathbb{R}$ and $C_i(\cdot)$ is the cost function of player i.

The goal of each player i in the Cournot game is to choose the best production quantity x_i such that maximizes his utility π_i. An important concept in game theory is the *Nash equilibrium*, at which state no player can increase their payoffs by unilaterally changing their strategies. A Nash equilibrium of the Cournot game defined by $(\pi_1, ..., \pi_N)$ is a vector $\boldsymbol{x}^* \geq 0$ such that for all i:

$$\pi_i(x_i^*, \boldsymbol{x}_{-i}^*) \geq \pi_i(\tilde{x}_i, \boldsymbol{x}_{-i}^*), \quad \text{for all} \quad \tilde{x}_i, \tag{1}$$

where $\boldsymbol{x}_{-i}$ denotes the actions of all players except i. In this paper, we restrict our attention to Cournot games satisfying the following assumptions:

Assumption: We assume the price function and cost functions:

(A1) The price function p is concave, strictly decreasing and twice differentiable on $[0, y_{\max}]$, where $y_{\max}$ is the first point where p becomes 0. For $y > y_{\max}$, $p(y) = 0$. In addition, $p(0) > 0$.

(A2) The cost function $C_i(x_i)$ is convex, strictly increasing, twice differentiable and $p(0) > C_i'(0)$, for all i.

These assumptions are standard in the literature (e.g., see [23] and references within). The assumption $p(0) > C_i'(0)$ is to avoid the triviality of a player never participating in the game. The following proposition shows that Cournot game satisfying the above assumptions has an unique Nash equilibrium.

Proposition 1: A Cournot game satisfying (A1) and (A2) has exactly one Nash equilibrium.

Proof of Proposition 1 refers to Theorem 1 in [24].

B. Policy-based Reinforcement Learning

In this work, we adopt a *policy-based* methods of how agents would learn and act. For each agent, we assume that it has (possibly noisy) information of the system states at time t, which we denote by $\mathbf{s_t}$. This agent maintains a policy $\pi_\theta(\cdot|\mathbf{s_t})$, which is a probability distribution on the action it would take, conditioned on the agent's information $\mathbf{s_t}$. At each time step, after the agent picks actions $\boldsymbol{a}_t \sim \pi_\theta(\cdot|\mathbf{s_t})$, the system releases reward. Subsequently, players update their policy parameters along the gradient direction of their long-term expected reward. Such learning procedure is called *policy gradient* method [25] in the literature. As a key permise for the idea, the policy long-term reward is,

$$J(\theta) = E_{\tau \sim p_\theta(\tau)}[\sum_t r(\mathbf{s}_t, \mathbf{a}_t)] \tag{2}$$

and the gradient is given by,

$$\nabla_\theta J(\theta) = E_{\tau \sim p_\theta(\tau)}[(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t))(\sum_{t=1}^{T} r(\boldsymbol{s}_t, \boldsymbol{a}_t))], \tag{3}$$

where τ and $J(\theta)$ are trajectories and the expected trajectory return under policy π_θ, respectively, and $\nabla_\theta \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t))$ is the score function of the policy. Various of policy gradient methods have been proposed by estimating the gradient (3) in different ways, including REINFORCE [25], natural policy gradient [26] and actor-crtic algorithms [27].

III. Stochastic Cournot Game

We discuss the main convergence results in this section. As briefly mentioned, we consider policy-based models of how agents choose and evolve their actions. In particular, as the agents only get the reward as feedback and nothing else, the dynamics in Section II-B reduces to the *stateless* version. This is consistent with the practice of many social-economic systems (e.g., energy markets [28]), where providing full feedback is either impractical or explicitly disallowed due to privacy and market power concerns.

In particular, we consider a policy that is parameterized by the mean of a distribution. This model includes many popular algorithms, for example, the ubiquitous Gaussian policies and their extensions [29]. Let θ_i denote the *mean* of player i's action, and X_i to be a zero-mean random variable. For

convenience, we assume it is continuous and has a bounded density function denoted by $f_i(X_i)$. We say X_i is unimodal at mean if f_i has a global maximum at the mean and no other isolated local maxima [2].

At each time step, player i choose the action to play as $a_i \sim \pi_{\theta_i}(\cdot) = \theta_i + X_i$. Note that in most Cournot games, the action is interpreted as quantity, that cannot be negative. Therefore, player i has to play by drawing a quantify from the rectified distribution $(\theta_i + X_i)^+$, where $a^+ = \max(a, 0)$. Under the Cournot game setup, the expected profit in Eq. (2) can be written out as the follows,

$$J_i(\theta_i; \boldsymbol{\theta}_{-i}) = E_{\boldsymbol{X}}\left[p\left(\sum_{j=1}^{N}(\theta_j + X_j)^+\right)(\theta_i + X_i)^+ - C_i((\theta_i + X_i)^+)\right]. \quad (4)$$

and the gradient value in Eq. (3) equals,

$$\nabla_{\theta_i} J_i = E\left[1(\theta_i + X_i \geq 0)\left\{p'\left(\sum_{j=1}^{N}(\theta_j + X_j)^+\right)(\theta_i + X_i) + p\left(\sum_{j=1}^{N}(\theta_j + X_j)^+\right) - C_i'(\theta_i + X_i)\right\}\right], \quad (5)$$

where $1(\cdot)$ is the indicator function.

We call the game associated with these J_i's the *stochastic Cournot game*, where player i chooses θ_i, observe the profit J_i, and update θ_i according to the payoff gradient. The Nash equilibrium of the stochastic Cournot game is defined as, $(\theta_1^*, ..., \theta_N^*)$ such that $\nabla_{\theta_i^*} J_i = 0, \forall i$. The form of (5) has made analyzing the system dynamics difficult compared to standard Cournot games. Firstly, because the actions are rectified, the profit of player i not long just depends on the sum of the other players (as in a Cournot game), but it actually depends on each of the other players' parameters. This rules out many elegant and simple results on the existence and uniqueness of Nash equilibria [30], [31], [32], [33]. Secondly, although the realization fo the actions are nonnegative, it is not obvious that θ_i's need to be nonnegative, or even bounded. The main result in the paper is to overcome these challenges and show that under some assumptions, the game is well-behaved and policy gradient updates converge exponentially quickly to the Nash equilibrium in Stochastic Cournot games.

Theorem 1: Consider a stochastic Cournot game satisfying the assumptions (A1) and (A2). Suppose each player's policy is parameterized as the mean θ_i and a zero-mean random variable X_i that is unimodal at the mean with infinite support, and suppose that all players follow policy gradient in (5) to update their mean. Then the policies converge to the Nash equilibrium exponentially quickly for all initializations either of the following condition holds:

1) The price function is linear.
2) The number of players equals two.

[2]For example, Gaussian and uniform distributions are unimodal under this definition.

The condition of the theorem includes Gaussian policies, which is a natural choice for continuous action spaces, and such a form also includes popular neural network policies [2] where the mean can be parameterized via a neural network. We also do not restrict the players to be symmetric, and each of the players would adopt different variances or even have completely different classes of distributions. The infinite support requirement of the distribution is a technicality and can be weakened, although it would make the proofs much more cumbersome.

The proof of Theorem 1 proceeds in three lemmas. We defer the full proofs of these lemmas to Section IV and sketch the steps in the proof here. The first step in the proof is to show that we can restrict the actions of the players to a compact region using the following lemma:

Lemma 1: Under the assumptions of Theorem 1, θ_i can be restricted to $[\underline{\theta}_i, y_{\max}]$, where $\underline{\theta}_i$ is a constant.

This lemma essentially confines the choices of the players to a compact interval, which sets up the rest of the proof. As a reminder, $y_{\max}$ is the point where the price function becomes 0. The proof of this lemma is based on showing that player i's profit will be suboptimal if it chooses an θ_i outside of the interval, regardless of other players' choices. The detailed proof of Lemma 1 could be found in the extended version [34].

Interestingly, to show the parameters of the policy gradients converges to the Nash equilibrium of the stochastic Cournot game for the two cases stated in Theorem 1, we need two different proof techniques. Therefore, we separate them into two lemmas as stated below.

Lemma 2: Under the assumptions of (A1)-(A2) and suppose the market price is linear, the policy gradient updates converge to the unique Nash equilibrium exponentially fast under all initial conditions.

Lemma 3: Under the assumptions of (A1)-(A2) and suppose there are only two players, the policy gradient updates converge to the unique Nash equilibrium exponentially fast under all initial conditions.

We close this section with two remarks. Firstly, our proof provides the sufficient conditions for the convergence of policy gradient in Cournot games, that is either the price function is linear, or the player number is no more than two for general price function. However, these may not be necessary. We provide a three-player example with quadratic price function in Section V-C, where we also observe convergence behavior. Secondly, it should be noted that in practice, some players may decide to not follow the policy gradient updates and use other learning algorithms (or act in adversarial manners). We provide some empirical evaluations of the system robustness in Section V-C, by assuming a small portion of players is acting randomly. Both directions, 1) generalizing the convergence proof to a broader class of games and 2) dynamics under heterogeneous/adversarial learning agents are important as future works.

IV. Proof of Theorem 1

In this section, we prove the two major lemmas regarding the convergence result stated in the previous section.

A. Proof of Lemma 2

Let $\boldsymbol{G}$ denote the Hessian of the game, so

$$G_{ij} = \frac{\partial^2 J_i}{\partial\theta_i\partial\theta_j} = \frac{\partial g_i}{\partial\theta_j}.$$

Focusing on the diagonal terms, we have

$$G_{ii} = \frac{\partial^2 J_i}{\partial\theta_i^2} = E\left[1(\theta_i + X_i \geq 0)\cdot\left\{p''\left(\sum_{j=1}^{N}(\theta_j + X_j)^+\right)\right.\right.$$
$$\left.\left.\cdot(\theta_i + X_i) + 2p'\left(\sum_{j=1}^{N}(\theta_j + X_j)^+\right) - C_i''(\theta_i + X_i)\right\}\right] < 0, \tag{6}$$

which is negative by assumptions (A1) and (A2). A game is said to be a *concave N-player game* [35] if $G_{ii} < 0, \forall i$ and the action space is convex and compact. Therefore, it is a concave N-player game. The following proposition is given in [35] as a sufficient condition to show when gradient-based algorithms converge to Nash equilibriums:

Proposition 2: Let $\boldsymbol{G}$ denote the Hessian of a concave N-player game. If $\boldsymbol{G}^T + \boldsymbol{G}$ is negative definite over the space of actions, there is a unique Nash equilibrium and the policy gradient dynamics approach it exponentially quickly for all initializations.

Using Proposition 2, it suffices for us to show that the negative definiteness of $\boldsymbol{G}^T + \boldsymbol{G}$ under the stochastic Cournot game. The main challenge of proving the negative definiteness lies in the *expectation* term, and we need to relate the properties of the Hessian of a function to its expectation. The following proposition tackles the aforementioned challenge and relates the Hessian property to its expectation.

Proposition 3: Let $f : \mathbb{R}^N \to \mathbb{R}$ be a continuous function and suppose that the first and second order partial derivatives exist for all points except possibly for a set of measure 0. Let $\boldsymbol{G}$ be the Hessian of f, whenever it exists. Now consider the function $\hat{f} : \mathbb{R}^N \to \mathbb{R}$, where $\hat{f}(y) = E_{\boldsymbol{X}} f(y + \boldsymbol{X})$ and $\boldsymbol{X}$ is random vector in $\mathbb{R}^N$, with continuous and bounded density function and infinite support. Let $\hat{\boldsymbol{G}}$ be the Hessian of $\hat{f}$. Then: i) If $\boldsymbol{G}^T + \boldsymbol{G}$ is negative semidefinite at all points where $\boldsymbol{G}$ exists, then $\hat{\boldsymbol{G}}^T + \hat{\boldsymbol{G}}$ is negative semidefinite. ii) If $\boldsymbol{G}^T + \boldsymbol{G}$ is negative definite for a set of measure larger than 0, then $\hat{\boldsymbol{G}}^T + \hat{\boldsymbol{G}}$ is negative definite.

Proof: The proof of this proposition is straightforward. By the assumption on the random vector $\boldsymbol{X}$, we can switch the order of differentiation and the expectation. In addition, the density being continuous allows us to ignore the points where $\boldsymbol{G}$ does not exist. Then

$$\begin{aligned}&\boldsymbol{v}^T(\boldsymbol{G}^T(\boldsymbol{y}) + \boldsymbol{G}(\boldsymbol{y}))\boldsymbol{v}\\ &= E_{\boldsymbol{X}}\left[\boldsymbol{v}^T(\boldsymbol{G}^T(\boldsymbol{y} + \boldsymbol{X}) + \boldsymbol{H}(\boldsymbol{y} + \boldsymbol{X}))\boldsymbol{v}\right] \leq 0,\end{aligned}$$

for any $\boldsymbol{v}$. Now suppose $\boldsymbol{G}^T(\boldsymbol{y}+\boldsymbol{x}) + \boldsymbol{G}(\boldsymbol{y}+\boldsymbol{x}))$ is negative definite for set of positive measure, then by the continuity of the density function, $\boldsymbol{v}^T(\boldsymbol{G}^T(\boldsymbol{y}) + \boldsymbol{G}(\boldsymbol{y}))\boldsymbol{v} < 0$ for all nonzero $\boldsymbol{v}$ and $\hat{\boldsymbol{G}}^T + \hat{\boldsymbol{G}}$ is negative definite. ∎

Let $f_i(\boldsymbol{x}) = p(\sum_i x_i^+)x_i^+$, which is continuous and twice differentiable except for a measure zero set on $\mathbb{R}^N$. Since $J_i = E[f_i(\boldsymbol{\theta} + \boldsymbol{X})]$, we need to show $f = \left[f_1 \dots f_N\right]$ satisfies the condition of Proposition 3. Given a vector $\boldsymbol{x}$, without loss of generality, assume that $x_1, \dots, x_k \geq 0$ and $x_{k+1}, \dots, x_N < 0$. The second order derivatives of f are,

$$\frac{\partial^2 f_i}{\partial x_i\partial x_j} = 1(x_i \geq 0)\cdot \begin{cases} p''(\sum_l x_l^+)x_i + 2p'(\sum_l x_l^+) - C_i''(x_i), i = j \\ 1(x_j > 0)\left(p''(\sum_l x_l^+)x_i + p'(\sum_l x_l^+)\right), i \neq j \end{cases}.$$

Because of the indicator on both $x_i \geq 0$ and $x_j \geq 0$, the Hessian is only nonzero for the *upper left block*. In this block, we have $\forall i, j \leq k$,

$$\frac{\partial^2 f_i}{\partial x_i\partial x_j} = \begin{cases} p''(\sum_{l=1}^k x_l)x_i + 2p'(\sum_{l=1}^k x_l) - C_i''(x_i), i = j \\ p''(\sum_{l=1}^k x_l)x_i + p'(\sum_{l=1}^k x_l), i \neq j, \end{cases}.$$

When the price function is linear, the second order derivative term vanishes, i.e. $p''(\sum_{l=1}^k x_l) = 0$. Therefore, we have,

$$\frac{\partial^2 f_i}{\partial x_i\partial x_j} = \begin{cases} 2p'(\sum_{l=1}^k x_l) - C_i''(x_i) \text{ if } i = j, i, j \leq k \\ p'(\sum_{l=1}^k x_l) \text{ if } i \neq j, i, j \leq k \end{cases}$$

Now we can write $\boldsymbol{G}$ as $\boldsymbol{G}_1 + \boldsymbol{G}_2 + \boldsymbol{G}_3$ where

$$\boldsymbol{G}_1 = \left[\begin{array}{ccc|c} \begin{bmatrix} p'(\sum_{l=1}^k x_l) & \cdots & p'(\sum_{l=1}^k x_l) \\ \vdots & \ddots & \vdots \\ p'(\sum_{l=1}^k x_l) & \cdots & p'(\sum_{l=1}^k x_l) \end{bmatrix} & & & \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \\ \hline & \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} & & \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \end{array}\right] \tag{7}$$

Both $\boldsymbol{G}_2, \boldsymbol{G}_3$ are diagonal matrices. For $\boldsymbol{G}_2$, it has the i'th component being $p'(\sum_{l=1}^k x_l)$ for $i \leq k$ and 0 for $i > k$. Since X_i have infinite support, there exists cases where $k = N$ (all player sample non-negative actions) for a *measure larger than 0 set*, in which $\boldsymbol{G}_2$ is *negative definite*. For $\boldsymbol{G}_3$, it has the i'th component being $-C_i''(x_i) \leq 0$ for $i \leq k$ and 0 for $i > k$, thus it is negative semi-definite. Therefore, it suffices for us to show the negative semi-definess of $\boldsymbol{G}_1$.

The eigenvalues of $\boldsymbol{G}_1$ in Eq. (7) are the combination of eigenvalues of the upper left and lower-right matrices. Eigenvalues of the upper left matrix are $kp'(\sum_{l=1}^k x_l) < 0$ and 0 ($k-1$ repeats), and eigenvalues of the lower right are all zeros. Thus $\boldsymbol{G}_1$ is negative semi-definite.

Therefore, there exists an measure nonzero set of actions, such that $\boldsymbol{G} = \boldsymbol{G}_1 + \boldsymbol{G}_2 + \boldsymbol{G}_3$ is negative definite. By Proposition 3, we have the Hessian of $(J_1, ..., J_N)$, that is $\hat{\boldsymbol{G}}$ is negative definite.

B. Proof of Lemma 3

Now, consider the two-player Cournot games with general price function $p(\cdot)$ under assumption (A1) and (A2). There are four cases considering the positiveness of x_1 and x_2.

a) $x_1, x_2 \geq 0$:

$$\boldsymbol{G}_a = \begin{cases} p''(x_1+x_2)x_i + 2p'(x_1+x_2) - C_i''(x_i), i = j, \\ p''(x_1+x_2)x_i + p'(x_1+x_2), i \neq j, \end{cases}$$

b) $x_1 < 0, x_2 \geq 0$:

$$\boldsymbol{G}_b = \begin{bmatrix} 0 & 0 \\ 0 & p''(x_1+x_2)x_2 + 2p'(x_1+x_2) - C_2''(x_2) \end{bmatrix}$$

c) $x_1 \geq 0, x_2 < 0$:

$$\boldsymbol{G}_c = \begin{bmatrix} p''(x_1+x_2)x_1 + 2p'(x_1+x_2) - C_1''(x_1) & 0 \\ 0 & 0 \end{bmatrix}$$

d) $x_1 < 0, x_2 < 0$:

$$\boldsymbol{G}_d = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

The game Hessian matrix thus follows,

$$\hat{\boldsymbol{G}} = E\left[1(x_1, x_2 \geq 0)\boldsymbol{G}_a + 1(x_1 < 0, x_2 \geq 0)\boldsymbol{G}_b + 1(x_1 \geq 0, x_2 < 0)\boldsymbol{G}_c + 1(x_1 < 0, x_2 < 0)\boldsymbol{G}_d\right], \quad (8)$$

$\boldsymbol{G}_a$ is a strictly diagonally dominant matrix since the magnitude of the diagonal entry is strictly larger than the *sum* of the magnitudes of all the other (non-diagonal) entries in each row, i.e., $|p''(x_1+x_2)x_i + 2p'(x_1+x_2) - C_i''(x_i)| > |p''(x_1+x_2)x_i + p'(x_1+x_2)|, \forall i$. Given that $\boldsymbol{G}_b, \boldsymbol{G}_c, \boldsymbol{G}_d$ are all diagonally dominant matrices, $\hat{\boldsymbol{G}}$ in Eq. (8) is a strictly diagonally dominant matrix . Therefore, the eigenvalues of matrix $\hat{\boldsymbol{G}}$ are all in the left–half plane (i.e., the real parts of eigenvalues are negative) by the Gershgorin circle theorem [36]. Proposition 4 in [37] showed that when all eigenvalues of the Hessian are in the open left–half plane, then the Nash equilibrium is an exponentially stable fixed point of the dynamical system generated by the gradient descend algorithm.

V. Numerical Experiments

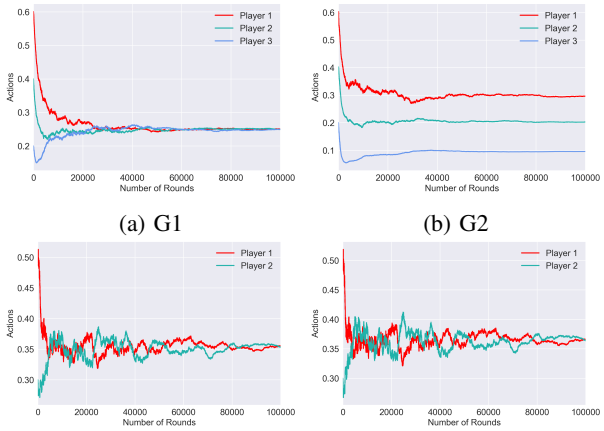
In this section, we exam the performance of policy gradient algorithms in various of Cournot games. We first verify the convergence behavior under linear price and two-player cases. Next, we provide investigative studies on the system behavior under multi-player and players with random actions scenarios is well as intuitions for the robustness behavior.

A. Experiment Setup

We perform all the experiments using the natural policy gradient algorithm [26] with a Gaussian policy. Following the derivations in Section II-B, the gradient with respect to the policy parameter θ_i follows,

$$\nabla_{\theta_i} J_i(\theta_i) = E_x[\pi_i(x_i, x_{-i}) \nabla_{\theta_i} \log f_{\theta_i}(x_i)] = \frac{1}{N} \sum_{i=1}^{N} \hat{\pi}_i \nabla_\theta \log f_\theta(x_i), \quad (9)$$

where $\pi_i(x_i, x_{-i})$ is the payoff function of player i and $\hat{\pi}_i$ is the observed payoff. In the above formula, $f_{\theta_i}(x_i) = \theta_i + X_i$ is the decision making policy for player i and $X_i \sim N(0, \sigma_i)$. For the action, we have $x_i = (\mu_i + X_i)^+$, where the action is truncated to be non-negative. We choose the standard deviation for each player the same as $\sigma = 0.05$. All experiments are run using a 2.2 GHz Intel Core i7 Macbook Pro with 16 GB memory.



(a) G1 (b) G2

(c) G3 (d) G4

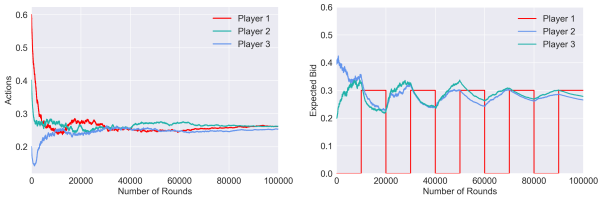
Fig. 1: Convergence behavior of policy gradient in stochastic Cournot games: (a)-(b) are games with linear price and (c)-(d) are two-player games with general price functions.

B. Cournot Game Examples

In this section, we verify the convergence behavior of the proposed algorithm in four example Cournot games, with different price and individual cost settings. **G1:** three-player with linear price function $p(\boldsymbol{x}) = 1 - (x_1 + x_2 + x_3)$ and no individual cost $C_i(x_i) = 0, \forall i$. The Nash equilibrium is $x_1^* = x_2^* = x_3^* = \frac{1}{4}$. **G2:** three-player with linear price function $p(\boldsymbol{x}) = 1 - (x_1 + x_2 + x_3)$ and differnet individual cost $C_i(x_i) = 0.1 \cdot i \cdot x_i$ for player i. The Nash equilibrium is $x_1^* = 0.3, x_2^* = 0.2, x_3^* = 0.1$. **G3:** two-player quadratic price function $p(\boldsymbol{x}) = 1 - (x_1 + x_2)^2$ without cost. The Nash equilibrium is $x_1^* = x_2^* = \sqrt{1/8} \approx 0.3536$. **G4:** two-player cubic price function $p(\boldsymbol{x}) = 1 - \frac{1}{2}(x_1 + x_2)^3$ without cost. The Nash equilibrium is $x_1^* = x_2^* = \sqrt[3]{1/20} \approx 0.3684$. In all of the games, each player simultaneously picks a production level. The price is determined by the sum of productions and broadcasted back to all players. This game is repeated multiple times with all players use policy gradient to learn and act. The dynamics of the policy parameter (i.e., the mean) are plotted in Figure 1. In all simulated games with different initializations and settings, the policy parameters converge to the Nash equilibrium, which verifies the theoretical results in Section III.

C. Investigative Studies

In this section, we provide two investigative studies relating to system performance under more general setups: 1) multi-agent Cournot game with non-linear price function; 2) hetergenous players that do not follow policy gradient updates. Note that our theoretical result in Section III does not apply to the following two cases. **G5:** three-player with quadratic price function $p(\boldsymbol{x}) = 1 - (x_1 + x_2 + x_3)^2$ and no cost. **G6:** three-player with linear price function $p(\boldsymbol{x}) =$



(a) G5 (b) G6

Fig. 2: Dynamics of policy gradient beyond the convergence condition provided in Theorem 1.

$1-(x_1+x_2+x_3)$ and no individual cost. One player does not follow policy gradient updates.

Fig 2 shows that both the three-player general price and heterogeneous players cases also converge to some equilibria, though they do not satisfy the convergence conditions in Theorem 1, These results are promising in the sense that our results might be able to generalize to a broader class of games, and theoretically proving these would be valuable future work. There may also be settings where players are malicious, but designing optimal adversarial tactics and the detection algorithms, by themselves are topics that contain a vast body of literature and is beyond the scope of this work.

VI. Conclusion

In this paper, we study the interaction of strategic players in Cournot games with limited feedback. We proved the convergence of policy gradient reinforcement learning to the Nash equilibrium, where player's policy is parameterized by the mean, under two conditions: either the price function is linear or there are two players. Extending the results to more general conditions such as multi-player general price functions would be an important future direction.

References

[1] N. Brown and T. Sandholm, "Superhuman ai for multiplayer poker," *Science*, vol. 365, no. 6456, pp. 885–890, 2019.

[2] S. Levine and V. Koltun, "Guided policy search," in *Proceedings of 30th International Conference on International Conference on Machine Learning (ICML)*, 2013.

[3] D. J. Mankowitz, N. Levine, R. Jeong, Y. Shi, J. Kay, A. Abdolmaleki, J. T. Springenberg, T. Mann, T. Hester, and M. Riedmiller, "Robust reinforcement learning for continuous control with model misspecification," *International Conference on Learning Representations (ICLR)*, 2020.

[4] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "Safe, multi-agent, reinforcement learning for autonomous driving," *arXiv preprint arXiv:1610.03295*, 2016.

[5] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *arXiv preprint arXiv:1911.10635*, 2019.

[6] A. OroojlooyJadid and D. Hajinezhad, "A review of cooperative multi-agent deep reinforcement learning," *arXiv preprint arXiv:1908.03963*, 2019.

[7] F. Schäfer and A. Anandkumar, "Competitive gradient descent," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 7623–7633.

[8] J. W. Crandall and M. A. Goodrich, "Learning to compete, compromise, and cooperate in repeated general-sum games," in *Proceedings of the 22nd International Conference on Machine learning (ICML)*, 2005, pp. 161–168.

[9] A. A. Cournot, *Recherches sur les principes mathématiques de la théorie des richesses*, 1838.

[10] D. S. Kirschen and G. Strbac, *Fundamentals of power system economics*. John Wiley & Sons, 2018.

[11] K. Bimpikis, S. Ehsani, and R. İlkılıç, "Cournot competition in networked markets," *Management Science*, vol. 65, no. 6, pp. 2467–2481, 2019.

[12] A. Kannan and U. V. Shanbhag, "Distributed computation of equilibria in monotone nash games via iterative regularization techniques," *SIAM Journal on Optimization*, vol. 22, no. 4, pp. 1177–1205, 2012.

[13] A. F. Daughety, *Cournot oligopoly: characterization and applications*. Cambridge university press, 2005.

[14] U. Nadav and G. Piliouras, "No regret learning in oligopolies: cournot vs. bertrand," in *International Symposium on Algorithmic Game Theory (EC)*. Springer, 2010, pp. 300–311.

[15] Y. Shi and B. Zhang, "No-regret learning in cournot games," *arXiv preprint arXiv:1906.06612*, 2019.

[16] T. Roughgarden, "Algorithmic game theory," *Communications of the ACM*, vol. 53, no. 7, pp. 78–86, 2010.

[17] S. Arora, E. Hazan, and S. Kale, "The multiplicative weights update method: a meta-algorithm and applications," *Theory of Computing*, vol. 8, no. 1, pp. 121–164, 2012.

[18] E. Hazan, "Introduction to online convex optimization," *Found. Trends Optim.*, vol. 2, no. 3-4, pp. 157–325, Aug. 2016. [Online]. Available: https://doi.org/10.1561/2400000013

[19] A. Kalai and S. Vempala, "Efficient algorithms for online decision problems," *Journal of Computer and System Sciences*, vol. 71, no. 3, pp. 291–307, 2005.

[20] D. S. Leslie and E. J. Collins, "Individual q-learning in normal form games," *SIAM Journal on Control and Optimization*, vol. 44, no. 2, pp. 495–514, 2005.

[21] E. Rodrigues Gomes and R. Kowalczyk, "Dynamic analysis of multiagent q-learning with ε-greedy exploration," in *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009, pp. 369–376.

[22] G. Arslan and S. Yüksel, "Decentralized q-learning for stochastic teams and games," *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1545–1558, 2016.

[23] R. Johari and J. N. Tsitsiklis, "Efficiency loss in cournot games," *Harvard University*, 2005.

[24] F. Szidarovszky and S. Yakowitz, "A new proof of the existence and uniqueness of the cournot equilibrium," *International Economic Review*, pp. 787–789, 1977.

[25] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proceedings of the 12th International Conference on Neural Information Processing Systems (NeurIPS)*, 1999.

[26] S. M. Kakade, "A natural policy gradient," in *Proceedings of the 15th International Conference on Neural Information Processing Systems (NeurIPS)*, 2002, pp. 1531–1538.

[27] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International Conference on Machine Learning (ICML)*, 2016, pp. 1928–1937.

[28] E. L. Quinn, "Privacy and the new energy infrastructure," *Available at SSRN 1370731*, 2009.

[29] P.-W. Chou, D. Maturana, and S. Scherer, "Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 834–843.

[30] F. Szidarovszky and S. Yakowitz, "Contributions to cournot oligopoly theory," *Journal of Economic Theory*, vol. 28, no. 1, pp. 51–70, 1982.

[31] W. Novshek, "On the existence of cournot equilibrium," *The Review of Economic Studies*, vol. 52, no. 1, pp. 85–98, 1985.

[32] R. Amir, "Cournot oligopoly and the theory of supermodular games," *Games and Economic Behavior*, vol. 15, no. 2, pp. 132–148, 1996.

[33] C. Ewerhart, "Cournot games with biconcave demand," *Games and Economic Behavior*, vol. 85, pp. 37–47, 2014.

[34] Y. Shi and B. Zhang, "Multi-agent reinforcement learning in cournot games," *arXiv preprint*, 2020.

[35] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave n-person games," *Econometrica: Journal of the Econometric Society*, pp. 520–534, 1965.

[36] E. W. Weisstein, "Gershgorin circle theorem," 2003.

[37] L. J. Ratliff, S. A. Burden, and S. S. Sastry, "Characterization and computation of local nash equilibria in continuous games," in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2013, pp. 917–924.