Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 50-1 (2017) 1508–1513

Payoff-Based Approach to Learning Nash Equilibria in Convex Games ⋆

T. Tatarenko * **M. Kamgarpour** **

* *Department of Control Theory and Robotics, TU Darmstadt, Germany (e-mail: tatarenk@rmr.tu-darmstadt.de)*
** *Automatic Control Laboratory, Swiss Federal Institute of Technology, Zurich, Switzerland (e-mail: mkamgar@control.ee.ethz.ch)*

Abstract: We consider multi-agent decision making, where each agent optimizes its cost function subject to constraints. Agents' actions belong to a compact convex Euclidean space and the agents' cost functions are coupled. We propose a distributed payoff-based algorithm to learn Nash equilibria in the game between agents. Each agent uses only information about its current cost value to compute its next action. We prove convergence of the proposed algorithm to a Nash equilibrium in the game leveraging established results on stochastic processes. The performance of the algorithm is analyzed with a numerical case study.

Keywords: Multi-agent decision making, game theory, payoff-based algorithm

1. INTRODUCTION

Decision-making in multi-agent systems arises in applications ranging from electricity market to communication and transportation networks (Arslan et al., 2007; Saad et al., 2012; Scutari et al., 2006). Game theory provides a powerful framework for formulating optimisation problems corresponding to competing or collaborative multi-agent systems. The various notions of equilibria in games characterise desirable and stable solutions to multi-agent optimisation problems. The focus of our paper is on distributed computation of Nash equilibria for a class of multi-agent decision making modeled by non-cooperative games.

There is a large body of work on computation of Nash equilibria in multi-agent games. The approaches differ by the particular structure of agents' cost functions as well as information available to each agent. In a *potential game*, a central optimization problem can be formulated whose minimizers correspond to Nash equilibria of the game. One can then use distributed algorithms for computing the minima of the potential function (Li and Marden, 2013; Salehisadaghiani and Pavel, 2014) to converge to Nash equilibria. Distributed algorithms are also proposed for *aggregative games* (Jensen, 2010; Paccagnan et al., 2016). In general, for implementation of these distributed algorithms communication is needed between individual agents or between each agent and a central coordinator.

Alternative to distributed optimization approaches, learning approaches to computing Nash equilibria proceed by sampling agents' actions from a set of probability distributions. These probability distributions are updated based on the information available in the system. Most of the past work has focused on algorithms that require knowledge of the structure of the cost functions. For example, (Perkins et al., 2015; Marden et al., 2009; Tatarenko, 2014, 2016b) have dealt with learning procedures requiring the so-called oracle information, where each agent can calculate its current cost given any action from its action set.

There are many practical situations in which agents do not know functional form of the objectives. Rather, each agent can only observe their obtained payoffs and be aware of their local actions. In this case, the information structure is referred to as *payoff-based*. A payoff-based learning in potential games is proposed in (Marden and Shamma, 2012) with the guarantee of stochastic stability of potential function minimizers, which coincide with Nash equilibria in potential games. However, to implement this payoff-based algorithm agents need to have some memory. Other algorithms requiring only payoff-based information and memory are proposed in (Goto et al., 2012) and (Zhu and Martínez, 2013). These learning procedures also guarantee stochastic stability of potential function minimizers. Moreover, by tuning a time-dependent parameter the learning procedures converge to a distribution over potential function minimizers in total variation. All aforementioned payoff-based procedures are applicable to games with finite action space. This shortcoming motivated our payoff-based approach to learn local optima continuous action games without memory (Tatarenko, 2016a). There, we addressed potential games in which agents' actions live in $\mathbb{R}$.

Our contributions in this paper are as follows. We develop a payoff-based approach for computing Nash equilibria in a general class of games with pseudo-monotone maps. In contrast to past work, we consider action spaces being compact subsets of a multidimensional Euclidean space. Given the constraints on action sets in a non-potential game setting, the previously proposed learning methods are no longer applicable. Thus, we develop a sampling based approach, with an appropriate update of probability distributions to sample from. We prove that through appropriate choices of step size, the actions converge in probability to Nash equilibria.

⋆ This research is partially supported by M. Kamgarpour's European Union ERC Starting Grant, CONENE.

10.1016/j.ifacol.2017.08.300

$\mathcal{N}(\boldsymbol{\mu}^i(t) = [\mu_1^i(t+1),\dots,\mu_d^i(t+1)]^\top, \sigma(t+1))$ with the density:

$$p_i(x_1^i,\dots,x_d^i;\boldsymbol{\mu}^i(t+1),\sigma(t+1)) = \frac{1}{(\sqrt{2\pi}\sigma(t+1))^d}\exp\left\{-\sum_{k=1}^{d}\frac{(x_k^i-\mu_k^i(t+1))^2}{2\sigma^2(t+1)}\right\}.$$

Our choice of Gaussian distribution is based on the idea of CALA (continuous action-set learning automaton), presented in the literature on learning automata (Thathachar and Sastry, 2003). The mean parameter $\boldsymbol{\mu}^i(t)$ for the state's distribution is updated as follows:

$$\boldsymbol{\mu}^i(t+1) = \mathrm{Proj}_{A_i}\left[\boldsymbol{\mu}^i(t) - \gamma(t+1)\sigma^2(t+1)\hat{J}_i(t)\frac{\mathbf{x}^i(t)-\boldsymbol{\mu}^i(t)}{\sigma^2(t)}\right].$$

In the above, $\mathrm{Proj}_C[\cdot]$ denotes the projection operator on set C. The initial finite value of $\boldsymbol{\mu}(0)$ can be defined arbitrarily. We emphasize the difference between states and actions. In particular, states are intermediary values $\mathbf{x}(t) = [\mathbf{x}^1(t),\dots,\mathbf{x}^N(t)]$ updated during the payoff-based algorithm under consideration. They need not belong to the set of joint actions $\boldsymbol{A}$. We will show that upon convergence of the algorithm, the states will also belong to the joint action set.

To analyze convergence of the proposed algorithm, first, we show that this algorithm is analogous to the Robbins-Monro stochastic approximation procedure (Bharath and Borkar, 1999). Next, we prove convergence of the random vector $\boldsymbol{\mu}(t) = [\boldsymbol{\mu}^1(t),\dots,\boldsymbol{\mu}^N(t)]$ by properly choosing $\{\sigma(t),\gamma(t)\}_{t=0}^{\infty}$.

It is straightforward to show that under Assumption 1

$$\begin{aligned}&\mathrm{E}_{\mathbf{x}(t)}\{\hat{J}_i(t)\frac{x_k^i(t)-\mu_k^i(t)}{\sigma^2(t)}\}\\ =&\mathrm{E}\{\hat{J}_i(t)\frac{x_k^i(t)-\mu_k^i(t)}{\sigma^2(t)}|x_k^i(t)\sim\mathcal{N}(\mu_k^i(t),\sigma(t))\}\\ =&\frac{\partial\tilde{J}_i(\boldsymbol{\mu}^1(t),\dots,\boldsymbol{\mu}^N(t),\sigma(t))}{\partial\mu_k^i}\end{aligned} \tag{2}$$

for any $i\in[N]$, $k\in[d]$, where

$$\tilde{J}_i(\boldsymbol{\mu}^1,\dots,\boldsymbol{\mu}^N,\sigma) = \int_{\mathbb{R}^{Nd}} J_i(\boldsymbol{x})p(\boldsymbol{\mu},\boldsymbol{x})d\boldsymbol{x},$$

$$p(\boldsymbol{\mu},\boldsymbol{x}) = \prod_{j=1}^{N} p_j(x_1^j,\dots,x_d^j;\boldsymbol{\mu}^j,\sigma).$$

Note that $\tilde{J}_i$ can be interpreted as the ith player's cost function in mixed strategies, given that the mixed strategies are multivariate normal distributions $\{\mathcal{N}(\boldsymbol{\mu}^i,\sigma)\}_i$.

We can rewrite the algorithm in the following vector form:

$$\boldsymbol{\mu}(t+1) = \mathrm{Proj}_{\boldsymbol{A}}[\boldsymbol{\mu}(t) - \gamma(t+1)\sigma^2(t+1) \times(\boldsymbol{M}(\boldsymbol{\mu}(t)) + \boldsymbol{Q}(\boldsymbol{\mu}(t),\sigma(t)) + \boldsymbol{R}(\boldsymbol{\mu}(t),\mathbf{x}(t),\sigma(t)))], \tag{3}$$

where

$$\boldsymbol{Q}(\boldsymbol{\mu}(t),\sigma(t)) = \tilde{\boldsymbol{M}}(\boldsymbol{\mu}(t)) - \boldsymbol{M}(\boldsymbol{\mu}(t)),$$

$$\boldsymbol{R}(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma(t)) = \boldsymbol{F}(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma(t)) - \tilde{\boldsymbol{M}}(\boldsymbol{\mu}(t)),$$

$$\boldsymbol{F}(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma(t)) = [\hat{J}_1(t)\frac{\mathbf{x}^1(t)-\boldsymbol{\mu}^1(t)}{\sigma^2(t)},\dots,\hat{J}_N(t)\frac{\mathbf{x}^N(t)-\boldsymbol{\mu}^N(t)}{\sigma^2(t)}],$$

and

$$\tilde{\boldsymbol{M}}(\cdot) = [\tilde{M}_{1,1}(\cdot),\dots,\tilde{M}_{1,d}(\cdot),\dots,\tilde{M}_{N,1}(\cdot),\dots,\tilde{M}_{N,d}(\cdot)]^\top$$

is the Nd-dimensional vector, where $i\in[N]$, $k\in[d]$, and

$$\tilde{M}_{i,k}(\boldsymbol{\mu}(t)) = \frac{\partial\tilde{J}_i(\boldsymbol{\mu}^1(t),\dots,\boldsymbol{\mu}^N(t),\sigma(t))}{\partial\mu_k^i}.$$

The algorithm above in in the framework of Robbins-Monro stochastic approximations procedures (Bharath and Borkar, 1999). In particular, the vector $\boldsymbol{M}(\boldsymbol{\mu}(t))$ corresponds to the gradient term in stochastic approximations procedures, $\boldsymbol{Q}(\boldsymbol{\mu}(t),\sigma(t))$ is a disturbance of the gradient term, whereas $\{\boldsymbol{R}(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma(t))\}_t$, according to (2), is a martingale difference. In our analysis, we will use the following well-known result of Robbins and Siegmund on non-negative random variables, see, for example, Lemma 10 in (Poljak, 1987).

Theorem 4. Let (Ω,F,P) be a probability space and $F_1\subset F_2\subset\dots$ a sequence of sub-σ-algebras of F. Let z_t,β_t,ξ_t, and ζ_t be non-negative F_t-measurable random variables such that $\mathrm{E}(z_{t+1}|F_t)\le z_t(1+\beta_t)+\xi_t-\zeta_t$. Then almost surely $\lim_{t\to\infty} z_t$ exists and is finite. Moreover, $\sum_{t=1}^{\infty}\zeta_t<\infty$ almost surely on $\{\sum_{t=1}^{\infty}\beta_t<\infty,\sum_{t=1}^{\infty}\xi_t<\infty\}$.

Now, we are ready to state our main result.

Theorem 5. Let players in a game $\Gamma(N,\{A_i\},\{J_i\})$ update their states $\{\mathbf{x}^i(t)\}$ at time t according to the normal distribution $\{\mathcal{N}(\boldsymbol{\mu}^i(t),\sigma(t))\}$, where the mean parameters are updated as in (3). Let Assumptions 1-3 hold and the variance parameter $\sigma(t)$ and the step-size parameter $\gamma(t)$ be chosen such that $\sum_{t=0}^{\infty}\gamma(t)\sigma^2(t)=\infty$, $\sum_{t=0}^{\infty}\gamma(t)\sigma^3(t)<\infty$, and $\sum_{t=0}^{\infty}\gamma^2(t)<\infty$. Then, as $t\to\infty$, the mean vector $\boldsymbol{\mu}(t)$ converges almost surely to a Nash equilibrium $\boldsymbol{\mu}^*\in\boldsymbol{A}$ of the game Γ, given any initial mean vector $\boldsymbol{\mu}(0)$, and the joint state $\mathbf{x}(t)$ converges in probability to $\boldsymbol{a}^*=\boldsymbol{\mu}^*$.

The theorem above claims almost sure convergence of the sequence of the mean vectors $\{\boldsymbol{\mu}(t)\}$ and weak convergence of the sequence of the agents' states $\{\mathbf{x}(t)\}$ to a Nash equilibrium in the game under consideration.

Remark 6. Recall that we distinguish between the states $\{\mathbf{x}^i\}_{i\in[N]}$ and actions $\{\boldsymbol{a}^i\}_{i\in[N]}$ of players in games. During the run of the algorithm, players choose their states $\{\mathbf{x}^i\}_{i\in[N]}$ according to the normal distributions $\{\mathcal{N}(\boldsymbol{\mu}^i,\sigma)\}_{i\in[N]}$ and have access to the current value $\{\hat{J}_i(t)\}_{i\in[N]}$ of their cost functions, given the actual joint state: $\hat{J}_i(t) = J_i(\mathbf{x}^1(t),\dots,\mathbf{x}^N(t))$. However, feasibility of the mean vectors $\{\boldsymbol{\mu}^i(t)\}_{i\in[N]}$ in the proposed procedure justifies the following choice for the actions: $\boldsymbol{a}^i(t)=\boldsymbol{\mu}^i(t)$ for all $i\in[N]$. Thus, under such setting and according to Theorem 5, the players' joint action $\boldsymbol{a}(t) = [\boldsymbol{a}^1(t),\dots,\boldsymbol{a}^N(t)]$ in long run of the payoff-based algorithm converges to a Nash equilibrium almost surely.

These convergences take place under an appropriate choice of the parameters $\gamma(t)$ and $\sigma(t)$. Note that, analogously to optimization methods based on the gradient descent iterations, the condition $\sum_{t=0}^{\infty}\gamma(t)\sigma^2(t)=\infty$ guarantees sufficient energy for the time-step parameter $\gamma(t)\sigma^2(t)$ to let the algorithm (3) get to a neighborhood of a desired stationary point, whereas the condition $\sum_{t=0}^{\infty}\gamma^2(t)<\infty$ does not allow the iteration under the projection operator to be unbounded as time goes to infinity.

3.2 Proof of Main Result (Theorem 5)

Our approach is as follows. Firstly, we estimate the distance between the mean vector $\boldsymbol{\mu}(t+1)$ in the run of the algorithm and some other $\boldsymbol{\mu} \in \boldsymbol{A}$ by this distance on the previous step, namely by $\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}\|$. After that, we analyse each term in this estimation to demonstrate applicability of Theorem 4 to the sequence $\{\boldsymbol{\mu}(t)\}_t$. Finally, we use the properties of Nash equilibria in games satisfying Assumptions 1-3 (see Corollary 3) to demonstrate that the almost sure limit of the sequence $\{\boldsymbol{\mu}(t)\}_t$ is a Nash equilibrium.

Let $\beta(t)=\gamma(t)\sigma^2(t)$. Then [2]

$$\boldsymbol{\mu}(t+1) = \mathrm{Proj}_{\boldsymbol{A}}[\boldsymbol{\mu}(t) - \beta(t+1) \times (\boldsymbol{M}(\boldsymbol{\mu}(t)) + \boldsymbol{Q}(\boldsymbol{\mu}(t)) + \boldsymbol{R}(\boldsymbol{\mu}(t)))]. \quad (4)$$

Let $\boldsymbol{\mu} \in \boldsymbol{A}$ be any point from the joint action set of the game Γ. Then, taking into account the iterative procedure for the update of $\boldsymbol{\mu}(t)$ above and the non-expansion property of the projection operator, we get

$$\begin{aligned}
&\|\boldsymbol{\mu}(t+1)-\boldsymbol{\mu}\|^2 \\
&= \|\mathrm{Proj}_{\boldsymbol{A}}[\boldsymbol{\mu}(t) - \beta(t+1) \times (\boldsymbol{M}(\boldsymbol{\mu}(t)) + \boldsymbol{Q}(\boldsymbol{\mu}(t)) + \boldsymbol{R}(\boldsymbol{\mu}(t)))] - \boldsymbol{\mu}\|^2 \\
&\le \|\boldsymbol{\mu}(t) - \beta(t+1) \times (\boldsymbol{M}(\boldsymbol{\mu}(t)) + \boldsymbol{Q}(\boldsymbol{\mu}(t)) + \boldsymbol{R}(\boldsymbol{\mu}(t))) - \boldsymbol{\mu}\|^2 \\
&= \|\boldsymbol{\mu}(t)-\boldsymbol{\mu}\|^2 - 2\beta(t+1)(\boldsymbol{M}(\boldsymbol{\mu}(t)), \boldsymbol{\mu}(t)-\boldsymbol{\mu}) \\
&\quad - 2\beta(t+1)(\boldsymbol{Q}(\boldsymbol{\mu}(t)) + \boldsymbol{R}(\boldsymbol{\mu}(t)), \boldsymbol{\mu}(t)-\boldsymbol{\mu}) \\
&\quad + \beta^2(t+1)\|g(\boldsymbol{\mu}(t))\|^2, \qquad (5)
\end{aligned}$$

where $g(\boldsymbol{\mu}(t)) = \boldsymbol{M}(\boldsymbol{\mu}(t)) + \boldsymbol{Q}(\boldsymbol{\mu}(t)) + \boldsymbol{R}(\boldsymbol{\mu}(t))$.

Let $\mathcal{F}_T$ be the σ-algebra generated by the random variables $\{\boldsymbol{\mu}(k), k \le T\}$. By taking the conditional expectation with respect to $\mathcal{F}_T$ of the both sides in the inequality above, we obtain that for any $T>0$, almost surely

$$\begin{aligned}
&2\sum_{t=0}^{T}\beta(t+1)(\boldsymbol{M}(\boldsymbol{\mu}(t)), \boldsymbol{\mu}(t)-\boldsymbol{\mu}) \\
&\le \|\boldsymbol{\mu}(0)-\boldsymbol{\mu}\|^2 - \mathrm{E}\{\|\boldsymbol{\mu}(T+1)-\boldsymbol{\mu}\|^2|\mathcal{F}_T\} \\
&\quad + 2\sum_{t=0}^{T}\beta(t+1)\|\boldsymbol{Q}(\boldsymbol{\mu}(t))\|\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}\| \\
&\quad + \sum_{t=0}^{T}\beta^2(t+1)\mathrm{E}_{\mathbf{x}(t)}\|g(\boldsymbol{\mu}(t))\|^2. \qquad (6)
\end{aligned}$$

In inequality (6) we used the property of the conditional expectation, namely $\mathrm{E}\{\boldsymbol{\mu}(t_1)|\mathcal{F}_{t_2}\} = \boldsymbol{\mu}(t_1)$ almost surely for any $t_1 \le t_2$, as well as the fact that $\mathrm{E}\{\boldsymbol{R}(\boldsymbol{\mu}(t))|\mathcal{F}_T\} = 0$ for all $t \le T$, which is implied by (2).

According to Assumption 3, we can show that

$$\begin{aligned}
\tilde{M}_{i,k}(\boldsymbol{\mu}) &= \frac{1}{\sigma^2}\int_{\mathbb{R}^{Nd}} J_i(\boldsymbol{x})(x_k^i - \mu_k^i)p(\boldsymbol{\mu},\boldsymbol{x})d\boldsymbol{x} \\
&= \int_{\mathbb{R}^{Nd}} \frac{\partial J_i(\boldsymbol{x})}{\partial x_k^i} p(\boldsymbol{\mu},\boldsymbol{x})d\boldsymbol{x}. \qquad (7)
\end{aligned}$$

Thus,

$$\tilde{\boldsymbol{M}}(\boldsymbol{\mu}(t)) = \int_{\mathbb{R}^{Nd}} M(\boldsymbol{x})p(\boldsymbol{\mu}(t),\boldsymbol{x})d\boldsymbol{x}. \qquad (8)$$

[2] We omit further the argument $\sigma(t)$ in terms $\boldsymbol{Q}$ and $\boldsymbol{R}$ for the sake of notation simplicity.

Since $\boldsymbol{Q}(\boldsymbol{\mu}(t)) = \tilde{\boldsymbol{M}}(\boldsymbol{\mu}(t)) - \boldsymbol{M}(\boldsymbol{\mu}(t))$ and due to Assumption 2 and equation (8), we can write the following:

$$\begin{aligned}
\|\boldsymbol{Q}(\boldsymbol{\mu}(t))\| &\le \int_{\mathbb{R}^{Nd}} \|\boldsymbol{M}(\boldsymbol{\mu}(t)) - \boldsymbol{M}(\boldsymbol{x})\|p(\boldsymbol{\mu}(t),\boldsymbol{x})d\boldsymbol{x} \\
&\le \int_{\mathbb{R}^{Nd}} L\|\boldsymbol{\mu}(t) - \boldsymbol{x}\|p(\boldsymbol{\mu}(t),\boldsymbol{x})d\boldsymbol{x} \\
&\le \int_{\mathbb{R}^{Nd}} L\left(\sum_{i=1}^{N}\sum_{k=1}^{d}|\mu_k^i(t) - x_k^i|\right)p(\boldsymbol{\mu}(t),\boldsymbol{x})d\boldsymbol{x} \\
&= O(\sigma(t)), \qquad (9)
\end{aligned}$$

where L is the Lipschitz constant defined in Assumption 2. The last equality in (9) is due to the fact that the first central absolute moment of a random variable with a normal distribution $\mathcal{N}(\mu,\sigma)$ is $O(\sigma)$.

Obviously, $\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}\|$ is bounded for any t, since $\boldsymbol{\mu}(t) \in \boldsymbol{A}$ for any t and $\boldsymbol{\mu} \in \boldsymbol{A}$. Now we proceed with estimating the terms $\mathrm{E}_{\mathbf{x}(t)}\|g(\boldsymbol{\mu}(t))\|^2$ in (6):

$$\begin{aligned}
\mathrm{E}_{\mathbf{x}(t)}\|g(\boldsymbol{\mu}(t))\|^2 &\le \|\boldsymbol{M}(\boldsymbol{\mu}(t))\|^2 + \|\boldsymbol{Q}(\boldsymbol{\mu}(t))\|^2 \\
&\quad + 2\|\boldsymbol{M}(\boldsymbol{\mu}(t))\|\|\boldsymbol{Q}(\boldsymbol{\mu}(t))\| + \mathrm{E}_{\mathbf{x}(t)}\|\boldsymbol{R}(\boldsymbol{\mu}(t))\|^2. \qquad (10)
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathrm{E}_{\mathbf{x}(t)}\|\boldsymbol{R}(\boldsymbol{\mu}(t))\|^2 \\
&\le \sum_{i=1}^{N}\sum_{k=1}^{d}\int_{\mathbb{R}^{Nd}} J_i^2(\boldsymbol{x})\frac{(x_k^i - \mu_k^i(t))^2}{\sigma^4(t)}p(\boldsymbol{\mu}(t),\boldsymbol{x})d\boldsymbol{x}.
\end{aligned}$$

Thus, we can use Assumption 3 to conclude that

$$\mathrm{E}_{\mathbf{x}(t)}\|\boldsymbol{R}(\boldsymbol{\mu}(t))\|^2 \le \frac{1}{\sigma^4(t)}f(\boldsymbol{\mu}(t),\sigma(t)),$$

where $f(\boldsymbol{\mu}(t),\sigma(t))$ is a polynomial of $\boldsymbol{\mu}(t)$ and $\sigma(t)$. Hence, taking into account boundedness of $\boldsymbol{\mu}(t)$ for all t, we conclude that

$$\beta^2(t+1)\mathrm{E}_{\mathbf{x}(t)}\|\boldsymbol{R}(\boldsymbol{\mu}(t))\|^2 \le k_1\gamma^2(t), \qquad (11)$$

for some constant k_1. Moreover, according to boundedness of $\boldsymbol{\mu}(t)$ for all t, we can conclude that the first term on the right hand side of (10) is bounded.

Bringing (9) - (11) together and taking into account conditions on the parameters $\gamma(t)$, $\sigma(t)$, we conclude that the right hand side of inequality (6) stays finite almost surely, if $T \to \infty$ and, thus, almost surely

$$\sum_{t=0}^{\infty}\beta(t+1)(\boldsymbol{M}(\boldsymbol{\mu}(t)), \boldsymbol{\mu}(t)-\boldsymbol{\mu}) < \infty. \qquad (12)$$

Next, we demonstrate that almost surely

$$\varliminf_{t\to\infty}(\boldsymbol{M}(\boldsymbol{\mu}(t)), \boldsymbol{\mu}(t)-\boldsymbol{\mu}) \le 0. \qquad (13)$$

Indeed, let us assume that, on the contrary, there exists such $\epsilon > 0$ and $t_0 > 0$ that almost surely

$$(\boldsymbol{M}(\boldsymbol{\mu}(t)), \boldsymbol{\mu}(t)-\boldsymbol{\mu}) \ge \epsilon$$

for any $t \ge t_0$. In this case, taking into account that $\sum_{t=0}^{\infty}\beta(t+1) = \infty$, we obtain

$$\begin{aligned}
&\sum_{t=0}^{\infty}\beta(t+1)(\boldsymbol{M}(\boldsymbol{\mu}(t)), \boldsymbol{\mu}(t)-\boldsymbol{\mu}) \\
&\ge \sum_{t=0}^{t_0}\beta(t+1)(\boldsymbol{M}(\boldsymbol{\mu}(t)), \boldsymbol{\mu}(t)-\boldsymbol{\mu}) + \epsilon\sum_{t=t_0}^{\infty}\beta(t+1) = \infty
\end{aligned}$$

almost surely, which contradicts (12). Thus, (13) holds. Since $\boldsymbol{\mu}(t)$ is bounded for any t, there exists such a limit point $\boldsymbol{\mu}^* \in \boldsymbol{A}$ that $\overline{\lim}_{t\to\infty} \boldsymbol{\mu}(t) = \boldsymbol{\mu}^*$ and, according to (13),

$$(\boldsymbol{M}(\boldsymbol{\mu}^*), \boldsymbol{\mu} - \boldsymbol{\mu}^*) \geq 0. \tag{14}$$

Since we did not specify the choice of $\boldsymbol{\mu} \in \boldsymbol{A}$, the inequality above holds for any $\boldsymbol{\mu} \in \boldsymbol{A}$. Thus, according to Corollary 3, $\boldsymbol{\mu}^*$ is a Nash equilibrium in the game Γ.

Next, we notice that, if $\boldsymbol{\mu} = \boldsymbol{\mu}^*$ in (5), this inequality (5) together with (9) - (11) imply that

$$\begin{aligned}\mathrm{E}\{\|\boldsymbol{\mu}(t+1) - \boldsymbol{\mu}^*\|^2 | \mathcal{F}_t\} \leq \|\boldsymbol{\mu}(t) - \boldsymbol{\mu}^*\|^2 \\ - 2\beta(t+1)(\boldsymbol{M}(\boldsymbol{\mu}(t)), \boldsymbol{\mu}(t) - \boldsymbol{\mu}^*) \\ + h(t),\end{aligned} \tag{15}$$

where $h(t) = k_2\beta(t+1)\sigma(t) + k_3\beta^2(t+1) + k_4\beta^2(t+1)\sigma(t) + (\beta(t+1)\sigma(t))^2 + k_1\gamma^2(t)$. According to the properties of $\sigma(t)$ and $\gamma(t)$,

$$\sum_{t=0}^{\infty} h(t) < \infty.$$

Moreover, since M is pseudo-monotone, (14) implies $(\boldsymbol{M}(\boldsymbol{\mu}(t)), \boldsymbol{\mu}(t) - \boldsymbol{\mu}^*) \geq 0$ for any t. Thus, we can apply Theorem 4 to conclude that

$$\|\boldsymbol{\mu}(t) - \boldsymbol{\mu}^*\| \text{ converges almost surely as } t \to \infty.$$

Since $\overline{\lim}_{t\to\infty} \boldsymbol{\mu}(t) = \boldsymbol{\mu}^*$ almost surely,

$$\lim_{t\to\infty} \boldsymbol{\mu}(t) = \boldsymbol{\mu}^* \text{ almost surely.}$$

Since $\sum_{t=0}^{\infty} \gamma(t)\sigma^2(t) = \infty$ and $\sum_{t=0}^{\infty} \gamma(t)\sigma^3(t) < \infty$, $\lim_{t\to\infty} \sigma(t) = 0$. Taking into account that $\mathbf{x}(t) \sim \mathcal{N}(\boldsymbol{\mu}(t), \sigma(t))$, we conclude that $\mathbf{x}(t)$ converges weakly to a Nash equilibrium $\boldsymbol{a}^* = \boldsymbol{\mu}^* \in \boldsymbol{A}$ as time runs. Moreover, according to Portmanteau Lemma (Klenke, 2008), this convergence is also in probability.

4. NUMERICAL CASE STUDY

We illustrate the proposed approach of payoff-based learning to a game arising from electricity market. The problem setup is motivated by the game theoretic formulation of plug-in-electric vehicle (PEV) charging considered in several previous work including (Ma et al., 2010; Grammatico et al., 2016; Couillet et al., 2012; Gan et al., 2013). The agents optimize their power consumption in response to a price signal. In contrast to most past work, we consider the case in which the form of the price function is unknown to agents and the agents do not communicate with each other. They can only observe their cost function for every strategy they play.

There are N market participants, also referred to as players or agents. Let $\boldsymbol{a}^i = [a_1^i, \ldots, a_d^i]^\top \in \mathbb{R}^d$ denote the decision variable of the player i, $i \in [N]$, that is the vector corresponding to its consumption profile over d periods. The constraints for each player i are

$$\begin{aligned}0 \leq & a_k^i \leq \bar{a} \quad \text{for } k = 1, \ldots, d, \\ & \sum_{k=1}^{d} a_k^i = \bar{a}^i.\end{aligned} \tag{16}$$

These constraints indicate that for each player the electricity consumption at each time instance is limited and the total electricity consumption over the considered period of time needs to match a desired amount. The convex and compact set defined by the constraints in (16) is considered the action set A_i for the corresponding player i.

The cost function is the price paid for electricity consumption by each agent (Paccagnan et al., 2016)

$$J_i(\boldsymbol{a}^i, \boldsymbol{a}^{-i}) = \boldsymbol{a}^{i\top} Q^i \boldsymbol{a}^i + 2\left(C^i \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{a}^j + \mathbf{c}^i\right)^\top \boldsymbol{a}^i \tag{17}$$

with $Q^i, C^i \in \mathbb{R}^{d\times d}$, $\mathbf{c}^i \in \mathbb{R}^d$ for all $i \in [N]$. In the above, the first term presents each agent's private value, while the second term corresponds to price of electricity and its functional form may not be known to the agents.

Consider the following setup. At iteration t, each player submits its proposed consumption profile over time horizon of d units, $\mathbf{x}^i(t) = [x_1^i(t), \ldots, x_d^i(t)]^\top$. How should players update their profiles, using only values of the function $J_i(\mathbf{x})$, in order to make the sequence of the joint profiles convergent to a Nash equilibrium? Note that Q^i can in general be known by individual agents, while the second term (17) is assumed unknown. Furthermore, Q^i can also equal zero.

We assume the matrices Q^i and C^i to be such that $Q^i + \frac{C^i}{N} \succeq 0$ on $\mathbb{R}^{Nd}$ and $\hat{M} \succeq 0$ on $\boldsymbol{A}$. It can readily be verified that the resulting game mapping $\hat{M}\boldsymbol{a} + \boldsymbol{m}$ (see (1)) is affine on $\mathbb{R}^{Nd}$ and, hence, Lipschitz on $\mathbb{R}^{Nd}$. Moreover, the positive semidefinite matrix $\hat{M}$ on $\boldsymbol{A}$ implies that $\hat{M}\boldsymbol{a}+\boldsymbol{m}$ is pseudo-monotone on $\mathbb{R}^{Nd}$ (Gowda, 1990). Thus, under such setting for the matrices Q^i and C^i, $i \in [N]$, Assumptions 1-3 hold in the game under consideration.

We consider strategies are the consumption profiles of the agents for $d = 4$ periods, the matrices Q^i and C^i, $i \in [N]$, in their cost functions (17) are the identity matrices of the size 4×4, and the vector $\mathbf{c}^i$, $i \in [N]$, is a 4-dimensional vector, whose coordinates are some random variables taking values in the interval $(0, 5)$. We assume that the action set A_i for each user $i \in [N]$ is defined by (16), where $\bar{a} = 6$ and $\bar{a}_i$ is a random variable taking values in the interval $(0.5, 10)$. The initial mean vector $\boldsymbol{\mu}(0)$ is a random vector with the uniform distribution on $\boldsymbol{A} = A_1 \times \ldots \times A_N$. Let the agents follow the payoff-based algorithm described by (3).

Figures 1 presents the relative error $\frac{\|\boldsymbol{\mu}(t) - \boldsymbol{a}^*\|}{\|\boldsymbol{a}^*\|}$ during the algorithm's run for $\gamma(t) = \frac{1}{t^{0.51}}$, $\sigma(t) = \frac{0.1}{t^{0.2}}$, $N = 10$ and $N = 100$ respectively, where $\boldsymbol{a}^*$ is the unique Nash equilibrium of the corresponding game. The uniqueness follows from the fact that the game mapping is strongly monotone in this example (Pang and Facchinei, 2003). We can see that after the first iteration the algorithm gives an approximation for the Nash equilibrium in the game, irrespectively to the initial vector $\boldsymbol{\mu}(0)$. However, convergence of the error to zero is slow. The slow decrease of the relative error after the first iteration can be explained by the choice of the rapidly decreasing parameter $\sigma(t)$ and $\gamma(t)$ as well as by the projection step. These factor prevent a significant change of the projected mean vectors and of the states' values chosen according to the normal distribution with the variance $\sigma(t)$.