Proceedings of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

A Modified Q-Learning Algorithm for Potential Games [★]

Yatao Wang [*] **Lacra Pavel** [*]

[*] *Edward S. Rogers Department of Electrical and Computer Engineering, University of Toronto,*
(e-mail: yatao.wang@utoronto.ca, pavel@control.utoronto.ca).

Abstract: This paper presents a modified Q-learning algorithm and provides conditions for convergence to a pure Nash equilibrium in potential games. In general Q-learning schemes, convergence to a Nash equilibrium may require decreasing step-sizes and long learning time. In this paper, we consider a modified Q-learning algorithm based on constant step-sizes, inspired by JSFP. When compared to JSFP, the Q-learning with constant step-sizes requires less information aggregation, but only reaches an approximation of a Nash equilibrium. We show that by appropriately choosing frequency dependent step-sizes, sufficient exploration of all actions is ensured and the estimated equilibrium approaches a Nash equilibrium.

1. INTRODUCTION

Potential games, Monderer and Shapley (1996), are an important class of games that can be used as template for the design of decentralized algorithms in large-scale problems, Scutari et al. (2006), Arslan et al. (2007). One of the most known examples is the problem of distributed traffic routing modelled as a large-scale congestion game, Rosenthal (1973). In such a game, a large number of vehicles or agents makes daily routing decisions to optimize their own objectives in response to their own observations.

The above setup can be modelled as a *large-scale repeated* game. In a *repeated* game, agents or players update and adapt their strategies depending on the opponents' actions in the previous stage(s) of the game. The dynamics of learning in repeated games is an area of active interest, Fudenberg and Levine (1998), Young (2005). A variety of algorithms have been proposed as well as analysis of their long-term behaviour and convergence to Nash equilibria. Guaranteed convergence to a Nash equilibrium in potential games has been shown for adaptive play and the broad class of finite-memory better-reply processes Young (1993), Young (2005). However, the assumption is that the agents' rewards (utilities) for different joint action profiles is a-priori known. In many large-scale games this assumption on the utilities is not realistic.

One of the most studied learning algorithms is the well-known fictitious-play (FP), Brown (1951). While known to be convergent to a Nash equilibrium in potential games, FP requires that each player can observe the (actions) decisions of all other players. Each player computes the empirical frequencies (i.e. running averages) of these observed decisions. Then, each player updates its strategy in a best-response manner to the empirical frequencies of all the other players.

In a *large-scale* game players are inherently faced with limited observational and computational capabilities. An open research problem is how to design learning algorithms with minimum information requirements for each player.

Recently, an elegant variant of Fictitious Play called Joint Strategy Fictitious Play (JSFP) has been proposed in Marden et al. (2009a), as a plausible decision making model in large-scale potential games. In JSFP, each player tracks the empirical frequencies of the *joint actions* of all other players. In contrast to FP, the action of a player is based on the (still incorrect) presumption that other players are playing randomly but jointly according to their joint empirical frequencies, i.e., each player views all other players as a collective group. The authors showed its beneficial features when applied to the large-scale congestion game. However, while reducing the information requirement, players still have to monitor the joint actions and need to know their own utility so as to find their optimal actions.

Players often have also limited information on the analytical structure of their own utility function, an even more challenging problem. Such problem setting of games with unknown utilities (rewards) and unobserved opponent actions, is a natural setup for Reinforcement Learning (RL) algorithms, or payoff-based dynamics, Sutton and Barto (1998). Agents observe only the actual utilities received as a result of the joint actions of their opponents, and use these actual utilities to choose future actions. In Marden et al. (2009b), the authors investigate payoff-based dynamics that converge to pure-strategy Nash equilibria in weakly acyclic games, one of which, sample experimentation dynamics, can admit perturbations in agents' rewards. The algorithm alternates between two phases, exploration and exploitation, and requires that several parameters are set in advance, such as the exploration phase length, exploration rates, and tolerances on payoff difference and switching rates for deciding when to change strategies.

Q-learning is a useful approach for learning Nash equilibria in games with unknown noisy utilities or rewards. Players'

[★] This work was supported by NSERC-CRD and Bell Labs/ Alcatel-Lucent.

rewards are initially unknown and must be learned or estimated online from actual observations.

Although Q-learning and the action adaptation processes are well understood independently, the combined problem of learning Nash equilibria in games with unknown reward functions is less well understood. In Claus and Boutilier (1998), the authors specify a joint action learner (JAL), in which each player keeps track of the frequency of other players' actions, while updating the utility (reward) estimate for the joint action played. However, the authors do not provide convergence conditions for their algorithm. Their investigation relies on experimental evidence of convergence, and, furthermore, it is restricted to team games, i.e., with a common utility function. Convergence is proved for identical interest games in Leslie and Collins (2003) assuming that learning takes place at multiple time scales. In Leslie and Collins (2005), players use variants of the Q-learning procedure independent of each other, oblivious of the effects of changes in other players' actions on their own payoffs. Convergence results to a Nash distribution were developed in Leslie and Collins (2005) under the assumption of multiple time-scales. Specifically, almost sure convergence to a Nash distribution was shown in two-player zero-sum games and in N-player partnership games. In K.Tuyls et al. (2006), Kaisers and K.Tuyls (2010) a continuous approximation of the discrete time system was introduced to study the dynamics of Q-learning in several examples.

Most prior results in Q-learning primarily focused on diminishing step-sizes. This allows the application of powerful results in stochastic approximation for studying the long-term stochastic processes' behaviour via their ODE approximations, Kushner and Yin (2003). Diminishing step-sizes are also used in the recent Q-learning schemes in Chapman et al. (2013), where convergence results are obtained without the ODE approximation. However the size of the learning problem faced by the agents grows exponentially with the number of players, thereby reducing the usefulness in large games. Moreover, diminishing step-sizes result in a long learning time and slow convergence.

This paper represents an effort in this direction. Specifically, our contribution is an algorithm that combines the strengths of Q-learning in terms of minimal information requirements, while at the same time achieving faster convergence, albeit to a near-optimal (approximate) Nash equilibrium. Our standing assumption is that players do not have information about the actions of the other players, and, moreover, they do not have complete information of their own payoff structure. We consider a modified Q-learning algorithm with constant step-size and develop some convergence results for potential games. The trade-off is convergence to near-optimal Nash equilibrium.

We achieve a faster convergence for the modified Q-learning algorithm by introducing non-negligible, constant step-sizes to reach a sub-optimal state in a reasonably short time, and then approach a Nash equilibrium via a slightly modified perturbation function as in Chasparis et al. (2011). The main challenge is proving convergence to Nash equilibria without the averaging effect of stochastic approximation, as this results in long convergence time. Our analysis techniques are similar to those used in the JSFP case Marden et al. (2009a). However, when compared to Fictitious Play algorithms, the setup here is complicated by players' lack of information on the analytical structure of their own utility.

The paper is organized as follows. Section 2 reviews background material. Section 3 introduces a modified Q-learning scheme with a state-based perturbation function. Section 4 establishes convergence to a pure Nash equilibrium, while Section 5 presents numerical results for a traffic congestion game. Section 6 presents concluding remarks.

Notation

The following notations are used.

Notation	**Meaning**
$\mathcal{A}_i$	finite action set of player i
$\mathcal{M}_i$	index set of actions player i $\mathcal{M}_i = \{1, \dots, \lvert\mathcal{A}_i\rvert\}$
$a_i(k) \in \mathcal{A}_i$	action of player i at time k
$\mathbf{e}_{ij} \in \Delta_i$	pure strategy of player i, j-th unit vector, $j \in \mathcal{M}_i$
$a_i(k) = \mathbf{e}_{ij}$	player i's j-th action, index notation
$\mathbf{1}_{ij}/\lvert\mathcal{A}_i\rvert \in \Delta_i$	mixed strategy of player i based on uniform distribution on each action
$x_i(k) \in \Delta_i$	mixed strategy of player i at time k
a	joint action (pure-strategy profile)
x	mixed-strategy profile
$a^* \in \mathcal{A}$	pure Nash equilibrium
$x^* \in \Delta$	mixed-strategy Nash equilibrium
$\mathcal{U}_i(a) \in \mathbb{R}$	utility of player i for joint action a
$\mathcal{U}_i(x) \in \mathbb{R}$	expected utility of player i for joint mixed-strategy (profile) x
$P(a) \in \mathbb{R}$	potential value for joint action a
$Q_i(a) \in \mathbb{R}^{\lvert\mathcal{A}_i\rvert}$	Q-vector of player i for joint action a

2. BACKGROUND

In this section we present a brief review of the background material.

2.1 Finite strategic-form games

A finite strategic-form game $\mathcal{G}$ involves a set $\mathcal{I}$ of N players, $\mathcal{I} = \{1, \dots, N\}$, where each player $i \in \mathcal{I}$ has a finite action set $\mathcal{A}_i$ and a utility function $\mathcal{U}_i : \mathcal{A} \to \mathbb{R}$, with $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_N$ denoting the joint-action set.

Let $a_i \in \mathcal{A}_i$ be an action of player (agent) $i \in \mathcal{I}$, and $a = (a_1, a_2, \dots, a_N) \in \mathcal{A}$ the joint action profile of all players. Let $|\mathcal{A}_i|$ denote the cardinality of the set $\mathcal{A}_i$ and $\mathcal{M}_i = \{1, \dots, |\mathcal{A}_i|\}$, the index set of player's i actions. Sometimes we use index notation and write $a_i = \mathbf{e}_{ij}$, $j \in \mathcal{M}_i$, to indicate the j^{th} action selected by player i. Let a_{-i} denote the profile of actions for players other than player i, i.e., $a_{-i} = (a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_N)$. With this notation, we often write a joint action profile a as (a_i, a_{-i}), or as $(\mathbf{e}_{ij}, a_{-i})$, if index notation is used. Similarly, we may write $\mathcal{U}_i(a)$ as $\mathcal{U}_i(a_i, a_{-i})$ or $\mathcal{U}_i(\mathbf{e}_{ij}, a_{-i})$. A player's goal is to maximize its own utility conditional on the choices of its opponents. A best response correspondence $\beta_i(a_{-i})$ is the set of optimal strategies for player i, given the strategy profile of its opponents,

$$\beta_i(a_{-i}) = \{a_i^* \in \mathcal{A}_i : \mathcal{U}_i(a_i^*, a_{-i}) = \max_{a_i \in \mathcal{A}_i} \mathcal{U}_i(a_i, a_{-i})\}, \quad (1)$$

or, in index notation,

$$\beta_i(a_{-i}) = \{\mathbf{e}_{ij^* \in \mathcal{M}_i} : \mathcal{U}_i(\mathbf{e}_{ij^*}, a_{-i}) = \max_{j \in \mathcal{M}_i} \mathcal{U}_i(\mathbf{e}_{ij}, a_{-i})\}.$$

In a Nash equilibrium, each agent plays a best response, $a_i^* \in \beta_i(a_{-i}^*)$, for all $i \in \mathcal{I}$, hence no individual player has an incentive to unilaterally change its strategy. A Nash equilibrium (N-tuple) profile, $a^* \in \mathcal{A}$, is a fixed-point $a^* \in \beta(a^*)$ of the overall N-tuple best-response correspondence, where $\beta := (\beta_1, \dots, \beta_N)$. Thus $a^* \in \mathcal{A}$ is called a pure Nash Equilibrium, if for all players $i \in \mathcal{I}$, $\mathcal{U}_i(a_i^*, a_{-i}^*) = \max_{a_i \in \mathcal{A}_i} \mathcal{U}_i(a_i, a_{-i}^*)$.

Sometimes, a mixed strategy is used where each player randomly chooses between several actions. Let x_{ia_i} or $x_{ij} \in [0,1]$ denote the probability that player $i \in \mathcal{I}$ selects action $a_i = \mathbf{e}_{ij}$ in the action set $\mathcal{A}_i$. Then $x_i = (x_{i1}, \dots, x_{i|\mathcal{A}_i|})$ is a probability distribution on its action set $\mathcal{A}_i$, or a mixed-strategy for player $i \in \mathcal{I}$. Hence, a mixed strategy x_i is an element of Δ_i, where

$$\Delta_i = \{x_i \in \mathbb{R}^{|\mathcal{A}_i|} | \mathbf{1}_{|\mathcal{A}_i|}^T x_i = 1, x_{ij} \geq 0, \forall j \in \mathcal{M}_i\},$$

with $\mathbf{1}$ denoting the all ones vector, is the set of probability distributions over $\mathcal{A}_i$. The vertices of Δ_i are the unit vectors $\mathbf{e}_{ij}$. When $x_i = \mathbf{e}_{ij}$, player i chooses the j^{th} action with probability one and such a strategy is called a pure strategy. Hence, using index notation, a_i can simply be identified by unit vectors $\mathbf{e}_{ij}$ in Δ_i, and the action set $\mathcal{A}_i$ by the vertices of the simplex Δ_i.

Likewise, we denote by $x = (x_1, x_2, \dots, x_N) \in \Delta$ the mixed-strategy profile of all players, where $\Delta = \Delta_1 \times \Delta_2 \times \dots \times \Delta_N$, and $x = (x_i, x_{-i})$. Given a mixed-strategy profile $x \in \Delta$, the expected utility of player i, also denoted $\mathcal{U}_i$, is the multi-linear extension from $\mathcal{A}$ to Δ, given as

$$\mathcal{U}_i(x) = \sum_{a \in \mathcal{A}} (\prod_{s \in \mathcal{I}} x_{sa_s}) \mathcal{U}_i(a_i, a_{-i}),$$

where each element x_{sa_s} is the probability of player s playing a_s. Equivalently, $\mathcal{U}_i(x) = \sum_{j \in \mathcal{M}_i} x_{ij} \mathcal{U}_i(\mathbf{e}_{ij}, x_{-i})$ where $\mathcal{U}_i(\mathbf{e}_{ij}, x_{-i}) = \sum_{a_{-i} \in \mathcal{A}_{-i}} (\prod_{s \in \mathcal{I}/\{i\}} x_{sa_s}) \mathcal{U}_i(\mathbf{e}_{ij}, a_{-i})$ and $\mathcal{I}/\{i\}$ denotes the set of players other than player i.

Similarly, a mixed-strategy profile $x^* \in \Delta$ is called a mixed-strategy Nash Equilibrium, if for all players $i \in \mathcal{I}$

$$\mathcal{U}_i(x_i^*, x_{-i}^*) \geq \mathcal{U}_i(x_i, x_{-i}^*), \qquad \forall x_i \in \Delta_i, x_i \neq x_i^*.$$

In the case when the inequality is strict, the Nash equilibrium is called a strict Nash equilibrium. Such a Nash equilibrium is a fixed-point of the mixed-strategy best-response extension, i.e., $x_i^* \in \beta_i(x_{-i}^*)$, for all $i \in \mathcal{I}$, where

$$\beta_i(x_{-i}) = \{x_i^* \in \Delta_i : \mathcal{U}_i(x_i^*, x_{-i}) \geq \mathcal{U}_i(x_i, x_{-i}), \forall x_i \in \Delta_i\}$$

where $\beta_i(x_{-i})$ is the best-response set. To avoid set-valued mappings (multiple elements in the best-response set), a smooth best-response can be used, e.g.

$$\sigma_i(x_{-i}) = \arg\max_{x_i \in \Delta_i} \{ \sum_{j \in \mathcal{M}_i} x_{ia_i} \mathcal{U}_i(\mathbf{e}_{ij}, x_{-i}) + \tau\, v_i(x_i)\},$$

where $\tau > 0$ is a temperature parameter and $v_i : \Delta_i \to \mathbb{R}$ is a player-dependent smoothing function. One example is Boltzmann selection, $v_i(x_i) = -\sum_{j \in \mathcal{M}_i} x_{ij} \log x_{ij}$. This results in the smooth best-response function

$$\sigma_i(x_{-i}) = \frac{e^{\tau^{-1} \mathcal{U}_i(\mathbf{e}_{ij}, x_{-i})}}{\sum_{j \in \mathcal{M}_i} e^{\tau^{-1} \mathcal{U}_i(\mathbf{e}_{ij}, x_{-i})}}. \quad (2)$$

As the temperature parameter $\tau \to 0$, this set approaches the set of best responses.

2.2 Potential games

A potential game is characterized by a single function, called potential function, that specifies the players' joint preference over outcomes Monderer and Shapley (1996).

Definition 1. A function $P : \mathcal{A} \to \mathbb{R}$ is a potential for the game $\mathcal{G}$ if for all $i \in \mathcal{I}$, for all $a_{-i} \in \mathcal{A}_{-i}$,

$$P(a_i, a_{-i}) - P(a_i', a_{-i}) = \mathcal{U}_i(a_i, a_{-i}) - \mathcal{U}_i(a_i', a_{-i}), \quad (3)$$

for all $a_i, a_i' \in \mathcal{A}_i$, or, equivalently,

$$P(\mathbf{e}_{ij}, a_{-i}) - P(\mathbf{e}_{ij'}, a_{-i}) = \mathcal{U}_i(\mathbf{e}_{ij}, a_{-i}) - \mathcal{U}_i(\mathbf{e}_{ij'}, a_{-i}),$$

for all $j, j' \in \mathcal{M}_i$. A game that admits such a potential function is called a potential game.

The local optima of the potential function are Nash equilibria of the game. Intuitively, a potential is a function of action profiles such that the difference induced by a unilateral deviation equals the change in the deviator's payoff. The existence of a potential function for a game implies a strict joint preference ordering over game outcomes, which in turn, ensures that the game has a number of useful properties.

Theorem 2. Monderer and Shapley (1996) Every finite potential game possesses at least one pure-strategy equilibrium.

A step in a game $\mathcal{G}$ is a change in one player's strategy. An improvement step in $\mathcal{G}$ is a change in one player's strategy such that its utility is improved. A path in $\mathcal{G}$ is a sequence of steps, $(a(0), a(1), \dots, a(T))$ in which exactly one player changes their strategy at each step. A path has an initial point, $a(0)$, and if it is of finite length T, a terminal point $a(T)$. A path is an improvement path in $\mathcal{G}$ if for all k, $\mathcal{U}_i(a(k-1)) < \mathcal{U}_i(a(k))$ for the deviating player i at step k. A game $\mathcal{G}$ is said to have the finite improvement property if every improvement path is finite.

Theorem 3. Monderer and Shapley (1996) Every improvement path in an ordinal potential game is finite.

The finite improvement property ensures that the behaviour of players who play "better responses" in each period of the repeated game converges to a Nash equilibrium in finite time. These results ensure that a number of simple adaptive processes converge to a pure-strategy Nash equilibrium in the game.

2.3 Repeated games

In a repeated version of the game $\mathcal{G}$, at every iteration $k \in \{0, 1, 2, \dots\}$, each player $i \in \mathcal{I}$ selects an action $a_i(k) \in \mathcal{A}_i$ and receives the utility $\mathcal{U}_i(a(k))$, function of the joint actions $a(k) = (a_1(k), \dots, a_N(k))$ of all players. Each player $i \in \mathcal{I}$ chooses action $a_i(k)$ according to the probability distribution $x_i(k)$. This selection is a function of the information and observations available to player $i \in \mathcal{I}$ up to iteration k. Both the action selection function and the available information depend on the learning process.

For example if the player knows the functional form (analytical structure) of its utility function $\mathcal{U}_i$ and can observe the actions of each of the other players at each step, the well-known fictitious play (FP) algorithm can be used. In a fictitious-play (FP) algorithm, each player computes the empirical frequency vectors, $q_{-i}(k)$, of its opponents and chooses its action as the best-response against this vector, hence

$$\beta_i(q_{-i}(k)) := \{a_i^* \in \mathcal{A}_i : \mathcal{U}_i(a_i^*, q_{-i}(k)) = \max_{a_i \in \mathcal{A}_i} \mathcal{U}_i(a_i, q_{-i})\}.$$

In case of a non-unique best-response, a random selection from the set $\beta_i(q_{-i}(k))$ is made as action selection. Alternatively, the use of a smooth best-response leads to a smooth fictitious-play (sFP) algorithm. The empirical frequencies generated by the FP converge to a Nash equilibrium in potential games, Monderer and Shapley (1996), but requires observations of the individual actions of all other players. This is relaxed in joint strategy FP (JSFP), Marden et al. (2009a), where each player tracks the empirical frequencies of the *joint actions* of all other players. In JSFP, the action of a player is based on the assumption that other players are playing randomly but jointly according to their joint empirical frequencies, i.e., each player views all other players as a collective group.

In cases in which the utility function is not known, a useful approach is to use a Q-learning algorithm in order to estimate the reward (utility) function in a recursive manner, Sutton and Barto (1998). The action selection can be based this time on estimated utilities, or Q-values, which characterize the relative utility of a particular action. Either a best-response or a smooth best-response can be used as the action selection mechanism. The task of a Q-learning agent is to learn a mapping from environment to actions so as to maximize a numerical utility or reward signal. One of the attractive features of Q-learning is the fact that it does not assume knowledge about utility or reward functions, rather these must be learned from the environment. In each step, the player receives a signal from the environment indicating its state and chooses an action. Once the action is performed, it changes the environment, generating a reinforcement signal that is then used to evaluate the quality of the decision by updating the corresponding Q values. The Q-values are estimations of the actual reward and the optimal policy is then followed by selecting the actions where the Q-values are maximum.

In the single-agent case, assuming a stationary environment, i.e., that the probabilities of receiving specific reinforcement signals do not change over time, if each action is executed in each state an infinite number of times and the learning rate is decayed appropriately, the Q-values will converge with probability 1 to the optimal ones, Sutton and Barto (1998).

In the multi-agent setting, of N players playing a game repeatedly, the process of learning Q values by observing actual utilities presents a significantly more complex problem, since the utilities (rewards) available to each player depend on the joint-actions or strategies of all the other players, and hence are not stationary. In Leslie and Collins (2005), each of the players are equipped with a standard Q-learning algorithm and learn independently without considering the presence of each other in the environment. Each player i selects an action $a_i(k) = \mathbf{e}_{ij}$ based on strategy x_i, receives a utility (reward) $\mathcal{U}_i(a(k)$ and then updates $Q_i(k)$, such as for example,

$$Q_{ij}(k+1) = (1 - \mu_i(k))Q_{ij}(k) + \mu_i(k)\mathcal{U}_i(a(k)), \qquad (4)$$

where $\mu_i(k)$ are learning rates or step-sizes, assumed to be diminishing. In Leslie and Collins (2005) an extra normalization was used in (4), where the utility error prediction term is divided by the probability with which $a_i(k)$ was selected. In (4), $a(k) = (a_i(k), a_{-i}(k)))$ denotes the joint-action. For the action selection, each player plays according to a Q-value based smooth best-response, (2), based on Boltzman selection,

$$x_{ij}(k) = \frac{e^{\tau^{-1} Q_{ij}(k)}}{\sum_{j' \in \mathcal{M}_i} e^{\tau^{-1} Q_{ij'}(k)}}, \qquad (5)$$

closely related to the soft-max exploration method of reinforcement learning Sutton and Barto (1998).

Based on standard theorems of stochastic approximation Benaim (1999), the behavior of these learners in 2-player games is analyzed by the corresponding ODE, Leslie and Collins (2005). The strategy evolution is closely related to the smooth best response dynamics, the same dynamical system that characterizes stochastic fictitious play (FP), Benaim and Hirsch (1999) despite the fact that individual Q-learning uses significantly less information. Using techniques from Leslie and Collins (2003), extension to N-player partnership games is studied for player-dependent learning rates. The use of diminishing learning rates is beneficial and allows one to use stochastic approximation results, but in general leads to slow convergence time. In the next section we introduce a modified Q-learning algorithm.

3. A MODIFIED Q-LEARNING ALGORITHM

In this section we present the Q-learning algorithm we consider. The two components of such an algorithm are the action selection and the Q-value updating rule.

At each time-step $k > 0$, each player $i \in \mathcal{I}$ chooses an action $a_i(k) = \mathbf{e}_{ij}$ based on its mixed-strategy $x_i(k)$ and its Q-value. Its probability vector $x_i(k)$ is updated according to the recursion

$$x_i(k+1) = (1 - \alpha_i(k))x_i(k) + \alpha_i(k)\widehat{\beta}_i(Q_i(k)), \qquad (6)$$

where $Q_i(k)$ is the Q-value vector, $\widehat{\beta}_i$ is defined as

$$\widehat{\beta}_i(Q_i(k)) = \{\mathbf{e}_{i\widehat{j^*}}, \widehat{j}^* \in \mathcal{M}_i : Q_{i\widehat{j^*}} = \max_{j \in \mathcal{M}_i} Q_{ij}(k)\}, \qquad (7)$$

hence maximizing the Q-value or numerical utility, and $\alpha_i(k)$ is the player step-size.

The Q-value of each player $Q_i(k)$ acts as the estimation of $\mathcal{U}_i(a(k))$ following the joint action $a(k) = (a_i(k), a_{-i}(k)) = (\mathbf{e}_{ij}(k), a_{-i}(k))$, providing the key information in the decision making for each player. This $Q_i(k)$ is a $|\mathcal{A}_i|$-dimensional vector with components $Q_{ij}(k+1)$, $j \in \mathcal{M}_i$. Each of its components is updated similar to (4) as follows: for the j-th component corresponding to the played action at time-step k, $a_i(k) = \mathbf{e}_{ij}$,

$$Q_{ij}(k+1) = (1 - \mu_{ij}(k))Q_{ij}(k) + \mu_{ij}(k)\mathcal{U}_i(a(k)), \qquad (8)$$

where $0 < \mu_{ij}(k) < 1$ is the learning rate (step size), while for the other components $j' \in \mathcal{M}_i$, $j' \neq j$ not played at time-step k,

$$Q_{ij'}(k+1) = Q_{ij'}(k), \qquad (9)$$

Thus in the algorithm we consider, for each player $i \in \mathcal{I}$, at time-step $k > 0$, a player chooses action $a_i(k) = \mathbf{e}_{ij}$ with probability $x_{ij}(k)$ based on the mixed-strategy $x_i(k)$, and updates $x_i(k)$ and $Q_i(k)$ as in (6), (7) and (8), (9).

We assume that:

Assumption 3.1. Player step-sizes are constant if not specified, i.e. for all players $i \in \mathcal{I}$, actions $j \in \mathcal{M}_i$,

$$\mu_{ij}(k) = \mu, \quad \alpha_i(k) = \alpha,$$

where $0 < \alpha < \mu < 1$.

An important component of Q-learning is the action selection mechanism, responsible for selecting the actions that the agent will perform during the learning process.

Our proposed action selection (6), (7) is based on greedy selection, when the action with the highest Q-value is selected with some inertia ($\alpha \neq 1$). This is a slightly modified Q-learning algorithm: instead of a Q-value based smooth best-response as in (5), inspired by the JSFP in Marden et al. (2009a), the action selection (6), (7) uses a Q-value based best-response with inertia.

Comparing the Q-value based best-response $\widehat{\beta}_i$ in (7) to (1), it can be seen that $\widehat{\beta}_i$ acts as an estimated best response. Let $\widehat{a}_i^*(k) := \mathbf{e}_{i\widehat{j}^*} = \widehat{\beta}_i(Q_i(k))$, $i \in \mathcal{I}$, consider the overall N-tuple $\widehat{a}^*(k) = (\widehat{a}_1^*(k), \ldots, \widehat{a}_N^*(k))$. Then denoting $\widehat{\beta}(Q(k)) := (\widehat{\beta}_1(Q_1(k)), \ldots, \widehat{\beta}_N(Q_N(k)))$, it follows that $\widehat{a}^*(k) = \widehat{\beta}(Q(k))$ acts as an estimated equilibrium at time-step k. Since it results from a process that lacks utility information, this estimated equilibrium would most likely be suboptimal. We can observe that, for the JSFP to converge to a Nash equilibrium, we don't need to estimate the exact value of utilities for each action, but a correct order of utilities of all actions. The Q-value of each action are incomplete estimations of joint strategy utility as in JSFP, while the action that has the largest probability to be updated is the estimated equilibrium. We prove in the following section that the corresponding Q-values will eventually converge to the actual utility, if the estimated equilibrium is no longer changing. Therefore, if the action space of the game is well explored, the estimated equilibrium will most likely to converge to an actual (true) Nash equilibrium.

In fact the action selection mechanism should allow for a trade-off between exploitation and exploration such that the agent can reinforce the evaluation of the actions it already knows to be good but also explore new actions. The Boltzmann action selection (5) used in a standard Q-learning algorithm, Leslie and Collins (2005), K.Tuyls et al. (2006) incorporates this trade-off. The greedy action selection, with constant step-sizes in (6), (7) might generally lead to suboptimal solutions. In order to incorporate a means of exploring less-optimal strategies, in the second part of the paper we introduce a perturbation in the Q-learning algorithm.

Specifically, inspired by the perturbing scheme in Chasparis et al. (2011), we assume that each player i selects the j-th action, $j \in \mathcal{M}_i$ according to a modified strategy with probability

$$\chi_{ij} = (1 - \rho_i(x_i, \xi))x_{ij} + \rho_i(x_i, \xi)\mathbf{1}_{ij}/|\mathcal{A}_i|, \tag{10}$$

where $\rho_i(x_i, \xi)$ is a perturbation function defined next.

Assumption 3.2. The perturbation function $\rho_i : \Delta_i \times [\bar{\epsilon}, 1] \to [0, 1]$ is continuously differentiable. Furthermore, for some $\zeta \in (0, 1)$ sufficiently close to one, ρ_i satisfies the following properties:

- $\rho_i(x_i, \xi) = 0$, $\forall x_i$ such that $|x_i|_\infty < \zeta$ $\forall \xi \geq \bar{\epsilon}$;
- $\lim_{|x_i|_\infty \to 1} \rho_i(x_i, \xi) = \xi$;
- $\lim_{|x_i|_\infty \to 1} \frac{\partial \rho_i(x_i, \xi)}{x_{ij}}|_{\xi=0} = 0$, $\forall j \in \mathcal{A}_i$.

This perturbation function is slightly modified from the one in Chasparis et al. (2011) and ensures mutation and exploration of all actions. Note that this mechanism is similar to the ϵ-greedy exploration, where it selects a random action with small probability ρ_i and the best action, i.e. the one that has the highest Q-value at the moment, with probability $(1 - \rho_i)$.

4. CONVERGENCE ANALYSIS

In this section, we give conditions under which the modified Q-learning algorithm (6) - (9) converges to a pure strategy Nash equilibrium almost surely. In the first part we consider constant learning rates (step-sizes), while in the second part we consider frequency dependent step-sizes based on perturbation (10), as a mechanism of exploration.

4.1 Convergence to estimated equilibria

In the following we show that in the absence of a mechanism of exploring all actions, the modified Q-learning algorithm (6) - (9) converges to an estimated equilibrium. This is part due to the lack of full utility function information.

Proposition 4. Under Assumption 3.1, if for some $K > 0$, an action profile $a(k) = (\mathbf{e}_{ij}, a_{-i})$ is repeatedly played in the consequent K iterations, i.e., $a(k + \kappa) = a(k)$, for all $1 \leq \kappa < K$, then

$$\begin{aligned} Q_{ij}(k + K) = &(1 - \mu)^K Q_{ij}(k) \\ &+ (1 - (1 - \mu)^K)\mathcal{U}_i(a(k)). \end{aligned} \tag{11}$$

Proof. Assume as in the statement that an action profile $a(k) = (\mathbf{e}_{ij}, a_{-i})$ is repeatedly played, i.e., the j^{th} action is played by player i. Then from (6), (8) and (9), it follows that if the j^{th} action is played, the other actions $j' \in \mathcal{M}_i$, $j' \neq j$ are never played during the following K iterations, so that each $Q_{ij'}$ stays unchanged. Moreover, for the played j^{th} action and Q_{ij}, from recursively using (8) it follows that

$$\begin{aligned} Q_{ij}(k + K) = &(1 - \mu)^K Q_{ij}(k) \\ &+ \mu \frac{1 - (1 - \mu)^K}{1 - (1 - \mu)} \mathcal{U}_i(a(k)), \end{aligned}$$

which yields (11).

□

Corollary 4.1. If for some sufficiently large $K > 0$, conditions in Proposition 4 hold, then

$$\lim_{K \to \infty} Q_{ij}(k + K) = \mathcal{U}_i(a(k)).$$

Corollary 4.1 can be verified simply by taking the limit $K \to \infty$ in (11) in Proposition 4, and using the Assumption 3.1 that $\mu < 1$.

Remark 4.1. If conditions of Proposition 4 are satisfied, from the description of the Q-learning algorithm (6), (8) and (9), other actions $j' \in \mathcal{A}_i$ are never played during K iterations, so that each $Q_{ij'}$ stays unchanged.

The following result shows an absorption property of estimated equilibria in Q-learning with constant step-sizes. The proof is similar to the proof of Theorem 3.1 in Marden et al. (2009a).

Proposition 5. Assume that at some time-step k, $\widehat{a}^* = \widehat{a}^*(k)$ is played, where $\widehat{a}^* = (\widehat{a}_i^*, \widehat{a}_{-i}^*)$, $\widehat{a}_i^* = \mathbf{e}_{\widehat{ij}^*}$ and that for every player $i \in \mathcal{I}$, $\forall j' \in \mathcal{M}_i$ and $j' \neq \widehat{j}^*$, $Q_{ij'}(k) < Q_{\widehat{ij}^*}(k)$, $Q_{ij'}(k) < \mathcal{U}_i(\widehat{a}^*)$. Then at any consequent K-th iteration, with probability of at least $\prod_{\kappa=1}^{K}(1-(1-\alpha)^\kappa)^N$, the following holds

$$\begin{aligned} Q_{\widehat{ij}^*}(k+K+1) &= (1-\mu)^{K+1} Q_{\widehat{ij}^*}(k) \\ &\quad + (1-(1-\mu)^{K+1})\mathcal{U}_i(\widehat{a}^*), \\ x_i(k+K+1) &= (1-\alpha)^{K+1} x_i(k) \\ &\quad + (1-(1-\alpha)^{K+1})\widehat{a}_i^*. \end{aligned}$$

Proof. We prove the result by induction. For $K = 1$, based on the conditions in the statement, since for every player i, $\forall j' \in \mathcal{A}_i$ and $j' \neq \widehat{j}^*$, $Q_{ij'}(k) < Q_{\widehat{ij}^*}(k)$, from (7) it follows that $\widehat{a}^*$ is the estimated best-response at time-step k, and $\widehat{a}_i^* = \mathbf{e}_{\widehat{ij}^*}$ is the component corresponding to player i. Therefore using (6) it follows that at time-step $(k+1)$,

$$x_i(k+1) = (1-\alpha)\, x_i(k) + \alpha\, \mathbf{e}_{\widehat{ij}^*}.$$

Since $\mathbf{e}_{\widehat{ij}^*}$ is the unit vector, this indicates $x_{\widehat{ij}^*}(k+1) \geq \alpha$. This holds for every player, and therefore at time-step $(k+1)$, $\widehat{a}^*$ is played with probability of at least α^N. When $\widehat{a}^*$ is played, hence $\widehat{a}_i^* = \mathbf{e}_{\widehat{ij}^*}$, i.e., the $\widehat{j}^*$-th action is played by player i at time-step $(k+1)$. Then from (8) it follows that the corresponding Q-value is updated as

$$\begin{aligned} Q_{\widehat{ij}^*}(k+1) &= (1-\mu) Q_{\widehat{ij}^*}(k) + \mu\, \mathcal{U}_i(\widehat{a}^*) \\ &> (1-\mu) Q_{ij'}(k) + \mu\, Q_{ij'}(k), \end{aligned} \tag{12}$$

where the middle inequality follows from the conditions given in the statement. Thus, $Q_{\widehat{ij}^*}(k+1) > Q_{ij'}(k)$, for all $j' \neq j$. Since any other j'-th action, $j' \neq \widehat{j}^*$ is not played at time-step $(k+1)$, from (9) it also follows that $Q_{ij'}(k+1) = Q_{ij'}(k)$. Therefore, $Q_{\widehat{ij}^*}(k+1) > Q_{ij'}(k+1)$, $\forall j' \in \mathcal{M}_i$, $j' \neq \widehat{j}^*$. This shows that, at time-step $(k+1)$, $\widehat{a}_i^* = \mathbf{e}_{\widehat{ij}^*}$ remains the estimated best-response component for player i. Repeating the above argument for all players, it follows that at time-step $(k+1)$, $\widehat{a}^*$ remains the estimated best-response, and the claim follows for $K = 1$.

As the next step of induction, suppose now that at every consequent κ-th iterations, $1 \leq \kappa \leq K-1$, $\widehat{a}^*$ is played with probability $\prod_{\kappa=1}^{K-1}(1-(1-\alpha)^\kappa)^N$. When $\widehat{a}^*$ is played, hence $\widehat{a}_i^* = \mathbf{e}_{\widehat{ij}^*}$ is played by player i, it follows that at time-step $(k+K)$, x_i is updated as

$$x_i(k+K) = (1-\alpha)^K x_i(k) + (1-(1-\alpha)^K)\, \mathbf{e}_{\widehat{ij}^*}, \tag{13}$$

where $\mathbf{e}_{\widehat{ij}^*} = \widehat{a}_i^*$. Moreover, by Proposition 4,

$$Q_{\widehat{ij}^*}(k+K) = (1-\mu)^K Q_{\widehat{ij}^*}(k) + (1-(1-\mu)^K)\mathcal{U}_i(\widehat{a}^*). \tag{14}$$

Since, for player i, any other j'-th action other than the $\widehat{j}^*$-th, $j' \neq \widehat{j}^*$ is not played, it also follows that $Q_{ij'}(k+K) = Q_{ij'}(k)$. Moreover, $Q_{ij'}(k+K) < Q_{\widehat{ij}^*}(k+K)$, and $Q_{ij'}(k+K) < \mathcal{U}_i(\widehat{a}^*)$, $\forall j' \in \mathcal{A}_i$, $j' \neq \widehat{j}^*$.

Following the same argument as for $K = 1$, from (13) it follows that at time-step $(k+K+1)$, $\widehat{a}^*$ is played with probability of at least $(1-(1-\alpha)^K)^N$. Hence from (8) and the foregoing two inequalities it follows that

$$\begin{aligned} Q_{\widehat{ij}^*}(k+K+1) &= (1-\mu) Q_{\widehat{ij}^*}(k+K) + \mu \mathcal{U}_i(\widehat{a}^*) \\ &> (1-\mu) Q_{ij'}(k+K) + \mu Q_{ij'}(k+K) \\ &= Q_{ij'}(k+K). \end{aligned}$$

Also, note that $Q_{ij'}(k+K+1) = Q_{ij'}(k+K)$, so that $Q_{\widehat{ij}^*}(k+K+1) > Q_{ij'}(k+K+1)$, $\forall j' \in \mathcal{A}_i$, $j' \neq \widehat{j}^*$, hence the estimated best response for player i is unchanged and is given as $\widehat{a}_i^* = \mathbf{e}_{\widehat{ij}^*}$. The same can be shown for all players $i \in \mathcal{I}$, and therefore,

$$\widehat{\beta}(Q(k+K+1)) = \widehat{\beta}(Q(k)) = \widehat{a}^*,$$

which indicates that the estimated equilibrium is unchanged. Substituting the above and (13), with $\widehat{a}_i^* = \mathbf{e}_{\widehat{ij}^*}$, into (6) yields that, at time-step $(k+K+1)$,

$$\begin{aligned} x_i(k+K+1) &= (1-\alpha)^{K+1} x_i(k) \\ &\quad + (1-(1-\alpha)^{K+1})\widehat{a}_i^*. \end{aligned}$$

Substituting (14) into (8) gives at time-step $(k+K+1)$,

$$\begin{aligned} Q_{\widehat{ij}^*}(k+K+1) &= (1-\mu) Q_{\widehat{ij}^*}(k+K) + \mu\, \mathcal{U}_i(\widehat{a}^*) \\ &= (1-\mu)^{K+1} Q_{\widehat{ij}^*}(k) \\ &\quad + (1-(1-\mu)^{K+1})\mathcal{U}_i(\widehat{a}^*), \end{aligned}$$

which completes the induction argument.

□

The next corollary follows immediately by using $0 < \alpha < \mu < 1$.

Corollary 4.2. If conditions in Proposition 5 hold for some sufficiently large $K > 0$, then with probability $\prod_{\kappa=1}^{\infty}(1-(1-\alpha)^\kappa)^N$ for every player i

$$\lim_{K\to\infty} Q_{\widehat{ij}^*}(k+K) = \mathcal{U}_i(\widehat{a}^*), \tag{15}$$

$$\lim_{K\to\infty} x_i(k+K) = \widehat{a}_i^*, \tag{16}$$

where $\widehat{a}^* = (\widehat{a}_i^*, \widehat{a}_{-i}^*)$, $\widehat{a}_i^* = \mathbf{e}_{\widehat{ij}^*}$.

Remark 4.2. Proposition 5 showed that if for every player, the actual (true) utility of the estimated equilibrium is greater than the numerical (estimated) utility of other actions, the Q-value of the estimated equilibrium is likely to converge to the true utility. The sufficient conditions in Proposition 5 mean that actions other than $\widehat{a}_i^*$ are sub-optimal not only for the estimated utility (Q-value) (hence not played under the algorithm), but also sub-optimal for the actual (true) utility. While intuitively these are reasonable conditions, we do not have yet an argument to show how strict they are. Under these conditions we can give a more precise characterization as in the next result.

Theorem 6. If the conditions in Proposition 5 hold then for sufficiently large K, $x(k+K)$ converges to a neighbourhood of $\widehat{a}^*$, almost surely.

Proof. From Proposition 5 and Corollary 4.2, (16) holds for every player with the same probability $\prod_{\kappa=1}^{\infty}(1-(1-\alpha)^{\kappa})^N$. We show that this probability is strictly positive, using an approach as used in Proposition 6.1 in Chasparis et al. (2011). The product $\prod_{\kappa=1}^{\infty}(1-(1-\alpha)^{\kappa})$ is non-zero if and only if $\sum_{\kappa=1}^{\infty}\log(1-(1-\alpha)^{\kappa}) > -\infty$, i.e.,

$$-\sum_{\kappa=1}^{\infty}\log(1-(1-\alpha)^{\kappa}) < \infty. \tag{17}$$

Note that,

$$\lim_{\kappa\to\infty}\frac{-\log(1-(1-\alpha)^{\kappa})}{(1-\alpha)^{\kappa}} = \lim_{\kappa\to\infty}\frac{1}{1-(1-\alpha)^{\kappa}} = 1,$$

since $0 < (1-\alpha) < 1$. Thus, from the limit comparison test, (17) holds if and only if $\sum_{\kappa=1}^{\infty}(1-\alpha)^{\kappa} < \infty$. This obviously holds for $0 < (1-\alpha) < 1$, since

$$\sum_{\kappa=1}^{\infty}(1-\alpha)^{\kappa} = \frac{1}{1-(1-\alpha)} = \frac{1}{\alpha} < \infty.$$

Therefore, indeed

$$\lim_{K\to\infty}\prod_{\kappa=1}^{K}(1-(1-\alpha)^{\kappa}) > 0,$$

i.e., as $K \to \infty$, (16) holds and $x(k+K)$ converges to $\widehat{a}^*$ with non-zero probability, $\lim_{K\to\infty}\prod_{\kappa=1}^{K}(1-(1-\alpha)^{\kappa})^N > 0$. In fact, from the above strictly positive limit we can conclude that $\forall\zeta$, such that $0 < \zeta < 1$, $\exists K_0$ sufficiently large, such that $(1-(1-\alpha)^{\kappa}) \geq \zeta$, for all $\kappa \geq K_0$, or $(1-\alpha)^{\kappa} \leq (1-\zeta)$, for all $\kappa \geq K_0$. From Proposition 5, this implies that after K_0 iterations, $x(k+K_0)$ enters a neighbourhood of radius $(1-\zeta)$, $B_{(1-\zeta)}(\widehat{a}^*)$, with probability $\prod_{\kappa=1}^{K_0}(1-(1-\alpha)^{\kappa})^N$. Following a similar argument as in the proof of Theorem 3.1 in Marden et al. (2009a) and Theorem 6.2 in Young (2005), we conclude that there exists constants $p_0 = \prod_{\kappa=1}^{\infty}(1-(1-\alpha)^{\kappa})^N > 0$, $K_0 \geq \log_{1-\alpha}(1-\zeta) > 0$, both independent of k, such that $x(k+K)$ enters the neighbourhood of $\widehat{a}^*$ with probability p_0, hence converges to $\widehat{a}^*$ almost surely.

□

Remark 4.3. In the case when players use the JSFP algorithm (assuming own utility information and observation of opponents' joint-actions), players reach a Nash equilibrium and stay there with probability $p > 0$ over a finite time $T > 0$, cf. Theorem 3.1 in Marden et al. (2009a). As shown in Theorem 6 when players use the Q-learning algorithm and estimate their own utility function, players are only guaranteed to reach the estimated equilibrium with some probability.

4.2 Equilibrium under perturbation

In the previous section, we proved that the Q-learning algorithm converges to an estimated equilibrium almost surely. Next we impose some additional assumptions that help to show that the estimated equilibrium can reach an actual Nash equilibrium. Based on the perturbation function in (10), we adjust the learning rates of Q_{ij} depending on the probability x_{ij} of action $a_i = \mathbf{e}_{ij}$ being played.

Assumption 4.1. Player step-sizes of Q are adjusted based on the frequency of a particular action, i.e., for all players $i \in \mathcal{I}$, actions $j \in \mathcal{A}_i$,

$$\mu_{ij}(k) = (1-\chi_{ij}(k)),$$

where χ_{ij} is defined in (10) and Assumption 3.2.

Assumption 4.2. Player utilities satisfy the following: for all players $i \in \mathcal{I}$, actions $j, j' \in \mathcal{M}_i$, $j \neq j'$, and joint actions $a_{-i} \in \mathcal{A}_{-i}$

$$\mathcal{U}_i(\mathbf{e}_{ij}, a_{-i}) \neq \mathcal{U}_i(\mathbf{e}_{ij'}, a_{-i}).$$

Assumption 4.2 means that no player is indifferent between distinct strategies; alternatively we could assume that all pure equilibria are strict.

Assumption 4.3. After $|x_i|_\infty > \zeta$, $\forall i \in \mathcal{I}$, i.e., when every player has entered the perturbation zone, no more than one player choose action j' other than $\widehat{j}^*$ at each iteration.

In order to satisfy Assumption 4.3, we can either force the perturbation to be asynchronous, so that it affects one player at a time, or we can choose ζ in Assumption 3.2 to be sufficiently large and ξ to be sufficiently small, so that $(1-\zeta(1-\xi))^2$ is sufficiently close to 0.

Next we show that under the perturbation function in Assumption 3.2, the estimated equilibrium would converge to a Nash equilibrium almost surely.

Theorem 7. If for some sufficiently large $K > 0$, conditions in Proposition 5 hold, and in addition Assumption 4.1, 4.2, 4.3 hold, then the estimated equilibrium $\widehat{a}^*(k)$ would converge to a Nash equilibrium a^* almost surely.

Proof. From Proposition 5 and Corollary 4.2, it follows that (15) and (16) holds for player m, that is

$$\lim_{K\to\infty} Q_{m\widehat{j}^*}(k+K) = \mathcal{U}_m(\widehat{a}^*(k)),$$
$$\lim_{K\to\infty} x_m(k+K) = \mathbf{e}_{m\widehat{j}^*}.$$

Suppose that the perturbation becomes active at some large enough time-step $\bar{k}$, and player m choose a different j'-th action other than the $\widehat{j}^*$-th one, i.e., chooses $a'_m(\bar{k}) = \mathbf{e}_{mj'}$ other than $\widehat{a}^*_m(\bar{k}) = \mathbf{e}_{m\widehat{j}^*}$. By Assumption 4.2, this is the only player to do so at time $\bar{k}$. From (10), such perturbation happens with probability of at least $\xi/|\mathcal{A}_i|$ for player m. $\widehat{a}^*_m(\bar{k})$ was the component of player m in the estimated equilibrium $\widehat{a}^* = (\widehat{a}^*_m, \widehat{a}^*_{-m})$. Let $a'(\bar{k}) = (a'_m(\bar{k}), \widehat{a}^*_{-m}(\bar{k}))$ the new action profile (joint-action) at time-step $\bar{k}$. Since player m choose the j'-th action at $\bar{k}$, from (8) and Assumption 4.1, at $(\bar{k}+1)$, he would update its $Q_{mj'}$ to be

$$\begin{aligned} Q_{mj'}(\bar{k}+1) &= \chi_{mj'}(\bar{k})Q_{mj'}(\bar{k}) \\ &\quad + (1-\chi_{mj'}(\bar{k}))\mathcal{U}_m(a'(\bar{k})). \end{aligned} \tag{18}$$

By assumption, conditions of Proposition 5 hold, so that for player m, $Q_{m\widehat{j}^*}(\bar{k}) > Q_{mj'}(\bar{k})$. Now consider the following two cases:

- If $\mathcal{U}_m(a'(\bar{k})) < \mathcal{U}_m(\widehat{a}^*(\bar{k}))$, then player m does not find any response that is better.

- If $\mathcal{U}_m(a'(\bar{k})) > \mathcal{U}_m(\widehat{a}^*(\bar{k}))$, then player m finds an action $a'_m(\bar{k})$ that is a better response than $\widehat{a}^*_m(\bar{k})$, i.e. the joint-action $a'(\bar{k})$ becomes the new estimated best response and this is denoted by $\widehat{a}^*(\bar{k}+1)$,

$$\begin{aligned} \widehat{a}^*(\bar{k}+1) := a'(\bar{k}) &= (a'_m(\bar{k}), \widehat{a}^*_{-m}(\bar{k})) \\ &\neq (\widehat{a}^*_m(\bar{k}), \widehat{a}^*_{-m}(\bar{k})) = \widehat{a}^*(\bar{k}). \end{aligned} \tag{19}$$

In the first case, i.e. $\mathcal{U}_m(a'(\bar{k})) < \mathcal{U}_m(\widehat{a}^*(\bar{k})))$, i.e., at a failed attempt to improve the utility, players would stay at the estimated equilibrium $\widehat{a}^*$ almost surely, and the potential stays unchanged.

The second case is a successful attempt to improve the utility, i.e., when $\mathcal{U}_m(a'(\bar{k})) > \mathcal{U}_m(\widehat{a}^*(\bar{k}))$. Consider (18) and note that, since actions other than estimated best response $\widehat{j}^*$ have sufficiently small probability, i.e. $\chi_{mj'}(\bar{k})$ is sufficiently close to 0, $(1 - \chi_{mj'}(\bar{k}))$ is sufficiently close to 1. Thus, from (18), $Q_{mj'}$ is updated to be sufficiently close to $\mathcal{U}_m(a'(\bar{k}))$, and therefore the j'-th action, $a'_m(\bar{k}) = \mathbf{e}_{mj'}$ becomes its new estimated best response. From (3) it follows that when the utility of player m is improved, the potential of the game is improved also. Thus a successful attempt to improve the utility of player m, i.e., $\mathcal{U}_m(a'_m(\bar{k}), \widehat{a}^*_{-m}(\bar{k})) > \mathcal{U}_m(\widehat{a}^*_m(\bar{k}), \widehat{a}^*_{-m}(\bar{k}))$, would result in an improved potential of the game $\mathcal{G}$ and a new estimated best response profile $\widehat{a}^*(\bar{k}+1)$ as in (19).

In summary, the first case leads to an estimated best response that is unchanged and results in an unchanged potential of the game $\mathcal{G}$, while the second case leads to a new estimated best response that improves the potential of the game $\mathcal{G}$. Whenever a player m takes a successful attempt and shifts to the new estimated equilibrium, the utility of this player would improve by $\mathcal{U}_m(\widehat{a}^*(\bar{k}+1)) - \mathcal{U}_m(\widehat{a}^*(\bar{k}))$. Hence, by (3),

$$P(\widehat{a}^*(\bar{k}+1)) - P(\widehat{a}^*(\bar{k})) = \mathcal{U}_m(\widehat{a}^*(\bar{k}+1)) - \mathcal{U}_m(\widehat{a}^*(\bar{k}))$$

hence, the potential value of the whole profile would also increase by the same amount. By the finite improvement property (Theorem 3, Monderer and Shapley (1996) Lemma 2.3), the estimated equilibrium $\widehat{a}^*(k)$ converges to an actual Nash equilibrium a^* almost surely.

□

Remark 4.4. Theorem 7 and Corollary 4.2, show that in a potential game, the Q-learning scheme with a perturbation function as in Assumption 3.2 and Assumption 4.1 will converge to a Nash equilibrium almost surely, while using less information than JSFP. Instrumental for this is the finite improvement property of potential games. Our analysis techniques are similar to those used in the JSFP case Marden et al. (2009a). However, the setup here is complicated by players' lack of information on the analytical structure of their own utility.

5. SIMULATIONS

In this section, we present simulation results of the Q-learning algorithm (6) and (8), for an example of a congestion game in a similar setup as in Marden et al. (2009a). A typical congestion game consists of a set $\mathcal{I}$ of N players and a set $\mathcal{R}$ of resources. For each player i, let the set of pure strategies $\mathcal{A}_i$ be the set of resources. An action $a_i \in \mathcal{A}_i$ reflects a selection of (multiple) resources, $a_i \in \mathcal{R}$. A player i is "using" resource r if $r \in a_i$. For an action profile a, let $q_r(a)$ be the number of drivers using road r, i.e., $\{i \in \mathcal{I} : r \in a_i\}$. For each resource $r \in \mathcal{R}$ an associated cost function c_r is defined that reflects the cost of using the resource as a function of the number of players using that resource. In a congestion game, the utility of player i using resources indicated by a_i depends only on the total number of players using the same resources, i.e.,

$$\mathcal{U}_i(a) = -\sum_{r \in a_i} c_r(q_r(a)),$$

where the negative sign reflects the cost of using a resource and its effect on a utility function. Any such congestion game is a potential game Rosenthal (1973).

In the case of distributed routing, consider the simple case of $N = 100$ players seeking to traverse from node A to node B along 10 different parallel roads, Marden et al. (2009a). Each driver can select any road as a possible route, so that the set of resources is the set of roads, $\mathcal{R}$, and each player can select one road. Each road has a quadratic cost function with positive (randomly chosen) coefficients,

$$c_r(q) = a_r q^2 + b_r q + c_r, \quad r = 1, \ldots, 10,$$

where q represent the number of vehicles on that particular road. The parameter α are chosen as 0.5 for all days and all players, and μ is chosen as 0.97.

Fig.2, shows results obtained by implementing the JSFP algorithm with a similar setup as in Marden et al. (2009a), while Fig. 1 shows corresponding results for the Q-learning algorithm. Comparing the two cases, it can be seen that the two algorithms have similar convergence time, while transient fluctuations in strategies and utilities over time are smaller for the Q-learning algorithm. These advantages are obtained even though Q-learning has less information requirements. However, the lack of the utility structure information results in a sub-optimal solution, compared to the Nash equilibrium in JSFP. On the other hand, when compared to other Q-learning algorithms, this is a reasonably good sub-optimal point to stay on while the perturbation function is still trying to optimize the solution to reach a Nash equilibrium.

6. CONCLUSIONS

We considered a Q-learning scheme for distributed convergence to Nash equilibria in potential games. The main difference from prior schemes lies in the choice of step-sizes and perturbation function. The non-negligible constant step-sizes result in faster convergence to an estimated equilibrium. This helps reduce the learning cost to a sub-optimal point while searching for the Nash equilibria. When compared to JSFP, the Q-learning with constant step-sizes requires less information aggregation, but only reaches a sub-optimal state that can be considered an approximation of a Nash equilibrium. We showed that by appropriately choosing frequency dependent step-sizes, sufficient exploration of all actions is ensured and the estimated equilibrium approaches the Nash equilibria. Future work will consider relaxing these conditions, as well as possible extensions to other classes of games.

REFERENCES

Arslan, G., Marden, J., and Shamma, J. (2007). Autonomous vehicle target assignment: A game-theoretical formulation. *ASME Journal of Dynamic Systems, Measurement and Control*, (129), 584–596.

Benaim, M. (1999). Dynamics of stochastic approximation algorithms. *Le Seminaire de probabilites, Lecture Notes in Math. 1709*, 1–68.

Benaim, M. and Hirsch, M.W. (1999). Mixed equilibria and dynamical systems arising from fictitious play in

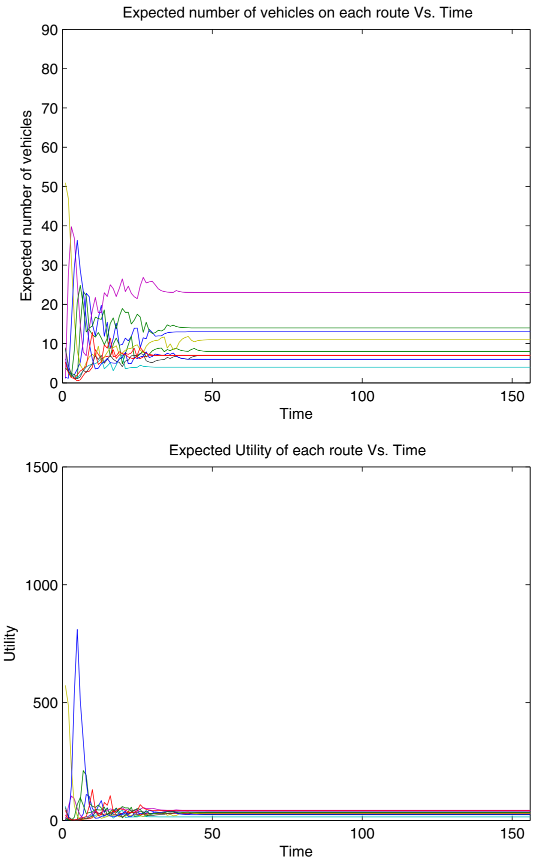

Fig. 1. Q-learning with $\mu = 0.97$

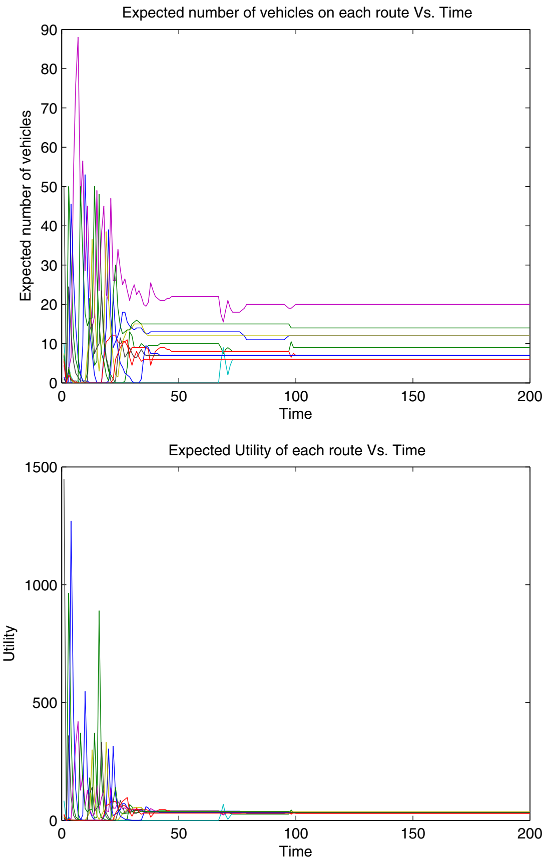

Fig. 2. JSFP with $\rho = 0.03$

perturbed games. *Games Econom. Behaviour*, 29, 36–72.

Brown, G.W. (1951). Iterative solutions of games by fictitious play. *in Koopmans, T. C. et al., editors, Activity Analysis of Production and Allocation*, Wiley, New York, 374–376.

Chapman, A.C., Leslie, D., Rogers, A., and Jennings, N.R. (2013). Convergent learning algorithms for unknown reward games. *SIAM Journal on Control and Optimization*, 51(4), 3154 – 3180.

Chasparis, G.C., Shamma, J.S., and Rantzer, A. (2011). Perturbed learning automata in potential games. In *Proceedings of the IEEE Conference on Decision and Control*, 2453–2458.

Claus, C. and Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multi agent systems. In *Proceedings of the 15th AAAI Nat. Conf. on Artificial IntelligenceTheoretical Population Biology*, 746–752.

Fudenberg, D. and Levine, D.K. (1998). *The Theory of Learning in Games*. The MIT Press, Cambridge.

Kaisers, M. and K.Tuyls (2010). Frequency adjusted multi-agent Q-learning. In *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems*, 309–315.

K.Tuyls, P., H., and Vanschoenwinkel, B. (2006). An evolutionary dynamical analysis of multi-agent learning in iterated game. *Autonomous Agents and Multi-Agent Systems*, 12(1), 115–153.

Kushner, H. and Yin, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer.

Leslie, D. and Collins, E. (2003). Convergent multiple-timescales reinforcement learning algorithms in normal form games. *Annals of Applied Probability*, 13, 1231–1251.

Leslie, D. and Collins, E. (2005). Individual Q-learning in normal form games. *SIAM Journal on Control and Optimization*, 1, 495–514.

Marden, J.R., Arslan, G., and Shamma, J.S. (2009a). Joint strategy fictitious play with inertia for potential games. *IEEE Transactions on Automatic Control*, 54, 208–220.

Marden, J.R., Young, H.P., Arslan, G., and Shamma, J.S. (2009b). Payoff-based dynamics for multiplayer weakly acyclic games. *SIAM J. Control and Optimization*, 48(1), 373 – 396.

Monderer, D. and Shapley, L.S. (1996). Potential games. *Games and Economic Behavior*, 14, 124–143.

Rosenthal, R.W. (1973). A class of games possessing pure-strategy Nash equilibria. *Int. J. Game Theory*, 2, 65–67.

Scutari, G., Barbarossa, S., and Palomar, D.P. (2006). Potential games: a framework for vector power control problems with coupled constraints. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Sutton, R.S. and Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Young, H.P. (1993). The evolution of conventions. *Econometrica*, 61(1), 57–84.

Young, H.P. (2005). *Strategic Learning and Its Limits*. Arne Ryde Memorial Lectures. Oxford University Press.