Multiagent Reinforcement Learning: Algorithm Converging to Nash Equilibrium in General-Sum Discounted Stochastic Games

Natalia Akchurina
International Graduate School of Dynamic Intelligent Systems
University of Paderborn
100 Warburger Str.
Paderborn, Germany
anatalia@mail.uni-paderborn.de

ABSTRACT

This paper introduces a multiagent reinforcement learning algorithm that converges with a given accuracy to stationary Nash equilibria in general-sum discounted stochastic games. Under some assumptions we formally prove its convergence to Nash equilibrium in self-play. We claim that it is the first algorithm that converges to stationary Nash equilibrium in the general case.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent systems*

General Terms

Algorithms, Theory

Keywords

algorithmic game theory, stochastic games, computation of equilibria, multiagent reinforcement learning

1. INTRODUCTION

Reinforcement learning turned out a technique that allowed robots to ride a bicycle, computers to play backgammon on the level of human world masters and solve such complicated tasks of high dimensionality as elevator dispatching. Can it come to rescue in the next generation of challenging problems like playing football or bidding on virtual markets? Reinforcement learning that provides a way of programming agents without specifying how the task is to be achieved could be again of use here but the convergence of reinforcement learning algorithms to optimal policies is only guaranteed under the conditions of stationarity of the environment that is violated in multiagent systems. For reinforcement learning in multiagent environments general-sum discounted stochastic games become a formal framework instead of Markov decision processes. Also the optimal policy concept in multiagent systems is different — we can't speak anymore about optimal policy (policy that provides the maximum cumulative reward) without taking into account the policies of other agents that influence our payoffs. In the environment where every agent tries to maximize its cumulative reward it is the most natural to accept Nash equilibrium as the optimal solution concept. In Nash equilibrium each agent's policy is the best response to the other agents' policies. Thus no agent can gain from unilateral deviation.

A number of algorithms [1, 2, 4, 5, 6, 7, 8, 9] were proposed to extend reinforcement learning approach to multiagent systems. The convergence to Nash equilibria was proved for very restricted class of environments: strictly competitive [6], strictly cooperative [2, 7] and 2-agent 2-action iterative game [1]. In [5] convergence to Nash equilibrium has been achieved in self-play for strictly competitive and strictly cooperative games under additional very restrictive condition that all equilibria encountered during learning stage are unique [7].

In this paper we propose a reinforcement learning algorithm that converges to Nash equilibria with some given accuracy in general-sum discounted stochastic games and prove it formally under some assumptions. We claim that it is the first algorithm that finds Nash equilibrium for the general case.

The paper is organized as follows. In section 2 we present formal definitions of stochastic game, Nash equilibrium, as well as prove some theorems that we will need for equilibrium approximation theorem in section 3. Section 3 is devoted to discussion and necessary experimental estimations of the conditions of the equilibrium approximation theorem. In sections 5 and 6 the developed algorithm Nash-DE[1] and analysis of the results of experiments are presented correspondingly.

2. PRELIMINARIES

In this section we recall some definitions from game theory.

Definition 1. A pair of matrices (M^1, M^2) constitute a bimatrix game G, where M^1 and M^2 are of the same size. The rows of M^k correspond to actions of player 1, $a^1 \in A^1$.

[1]Maintaining the tradition the name of the algorithm reflects the result — the approximation of Nash-equilibrium as well as the approach — differential equations.

Cite as: Multiagent Reinforcement Learning: Algorithm Converging to Nash Equilibrium in General-Sum Discounted Stochastic Games, Natalia Akchurina, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009),* Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. 725–732

The columns of M^k correspond to actions of player 2, $a^2 \in A^2$. A^1 and A^2 are the sets of discrete actions of players 1 and 2 respectively. The payoff $r^k(a^1, a^2)$ to player k can be found in the corresponding entry of the matrix M^k, $k = 1, 2$.

Definition 2. A pure ε-equilibrium of bimatrix game G is a pair of actions (a^1_*, a^2_*) such that

$$r^1(a^1_*, a^2_*) \geq r^1(a^1, a^2_*) - \varepsilon \; for \; all \; a^1 \in A^1$$

$$r^2(a^1_*, a^2_*) \geq r^2(a^1_*, a^2) - \varepsilon \; for \; all \; a^2 \in A^2$$

Definition 3. A mixed ε-equilibrium of bimatrix game G is a pair of vectors (ρ^1_*, ρ^2_*), such that

$$\rho^1_* M^1 \rho^2_* \geq \rho^1 M^1 \rho^2_* - \varepsilon \; for \; all \; \rho^1 \in \sigma(A^1)$$

$$\rho^1_* M^2 \rho^2_* \geq \rho^1_* M^2 \rho^2 - \varepsilon \; for \; all \; \rho^2 \in \sigma(A^2)$$

where $\sigma(A^k)$ is the set of probability distributions over action space A^k, such that for any $\rho^k \in \sigma(A^k)$, $\sum_{a \in A^k} \rho^k_a = 1$

$$\begin{aligned} \rho^1 M^k \rho^2 &= \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \rho^1_{a^1} r^k(a^1, a^2) \rho^2_{a^2} = \\ &= \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} r^k(a^1, a^2) \prod_{i=1}^{2} \rho^i_{a^i} \end{aligned}$$

is the expected reward of agent k induced by (ρ^1, ρ^2).

Definition 4. Nash equilibrium of bimatrix game G is ε-equilibrium with $\varepsilon = 0$.

Obviously definitions 1, 2, 3 and 4 can be generalized for arbitrary number of players.

Definition 5. A 2-player discounted stochastic game Γ is a 7-tuple $\langle S, A^1, A^2, \gamma, r^1, r^2, p \rangle$, where S is discrete finite set of states ($|S| = N$), A^k is the discrete action space of player k for $k = 1, 2$ ($|A^k| = m^k$), $\gamma \in [0, 1)$ is the discount factor, $r^k : S \times A^1 \times A^2 \to \mathbb{R}$ is the reward function for player k bounded in absolute value by $R_{\max}$, $p : S \times A^1 \times A^2 \to \Delta$ is the transition probability map, where Δ is the set of probability distributions over state space S.

Discount factor γ reflects the notion that a reward at time $t + 1$ is worth only $\gamma < 1$ of what it is worth at time t.

Every state of a 2-player stochastic game can be regarded as a bimatrix game.

It is assumed that for every $s, s' \in S$ and for every action $a^1 \in A^1$ and $a^2 \in A^2$, transition probabilities $p(s'|s, a^1, a^2)$ are stationary for all $t = 0, 1, 2, \ldots$ and $\sum_{s' \in S} p(s'|s, a^1, a^2) = 1$.

Policy of agent $k = 1, 2$ is a vector $x^k = (x^k_{s_1}, x^k_{s_2} \ldots, x^k_{s_N})$, where $x^k_s = (x^k_{sa^k_1}, x^k_{sa^k_2} \ldots, x^k_{sa^k_{m^k}})$, $x^k_{sh} \in \mathbb{R}$ being the probability assigned by agent k to its action $h \in A^k$ in state s. A policy x^k is called a *stationary policy* if it is fixed over time. Since all probabilities are nonnegative and sum up to one, the vector $x^k_s \in \mathbb{R}^{m^k}$ belongs to the unit simplex Δ^k in m^k-space defined as

$$\Delta^k = \{x^k_s \in \mathbb{R}^{m^k}_+ : \sum_{a^k \in A^k} x^k_{sa^k} = 1\}$$

.

The policy x^k will belong then to policy space of agent k:

$$\Theta^k = \times_{s \in S} \Delta^k$$

Each player k ($k = 1, 2$) strives to learn policy by immediate rewards so as to maximize its expected discounted cumulative reward (players don't know state transition probabilities and payoff functions):

$$v^k(s, x^1, x^2) = \sum_{t=0}^{\infty} \gamma^t E(r^k_t | x^1, x^2, s_0 = s)$$

where x^1 and x^2 are the policies of players 1 and 2 respectively and s is the initial state.

$v^k(s, x^1, x^2)$ is called the discounted value of policies (x^1, x^2) in state s to player k.

Definition 6. A 2-player discounted stochastic game Γ is called zero-sum when $r^1(s, a^1, a^2) + r^2(s, a^1, a^2) = 0$ for all $s \in S$, $a^1 \in A^1$ and $a^2 \in A^2$, otherwise general-sum.

Definition 7. A ε-equilibrium of 2-player discounted stochastic game Γ is a pair of policies (x^1_*, x^2_*) such that for all $s \in S$ and for all policies $x^1 \in \Theta^1$ and $x^2 \in \Theta^2$:

$$v^1(s, x^1_*, x^2_*) \geq v^1(s, x^1, x^2_*) - \varepsilon$$

$$v^2(s, x^1_*, x^2_*) \geq v^2(s, x^1_*, x^2) - \varepsilon$$

Definition 8. Nash equilibrium of 2-player discounted stochastic game Γ is ε-equilibrium with $\varepsilon = 0$.

Definition 9. A n-player discounted stochastic game is a tuple $\langle K, S, A^1, \ldots, A^n, \gamma, r^1, \ldots, r^n, p \rangle$, where $K = \{1, 2, \ldots, n\}$ is the player set, S is the discrete state space ($|S| = N$), A^k is the discrete action space of player k for $k \in K$ ($|A^k| = m^k$), $\gamma \in [0, 1)$ is the discount factor, $r^k : S \times A^1 \times A^2 \times \ldots \times A^n \to \mathbb{R}$ is the reward function for player k bounded in absolute value by $R_{\max}$, $p : S \times A^1 \times A^2 \times \ldots \times A^n \to \Delta$ is the transition probability map, where Δ is the set of probability distributions over state space S.

Definitions 6, 7 and 8 can be generalized for n-player stochastic game.

Definition 10. A profile is a vector $x = (x^1, x^2, \ldots, x^n)$, where each component x^k is a policy for player $k \in K$. The space of all profiles $\Phi = \times_{k \in K} \Theta^k$.

Let's define the probability transition matrix induced by x:

$$p(s'|s, x) = \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \ldots \sum_{a^n \in A^n} p(s'|s, a^1, a^2, \ldots, a^n) \prod_{i=1}^{n} x^i_{sa^i}$$

$$P(x) = (p(s'|s, x))_{s, s' \in S}$$

The immediate expected reward of player k in state s induced by x will be:

$$r^k(s, x) = \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \ldots \sum_{a^n \in A^n} r^k(s, a^1, a^2, \ldots, a^n) \prod_{i=1}^{n} x^i_{sa^i}$$

Then the immediate expected reward matrix induced by profile x will be:

$$r(x) = (r^k(s,x))_{s\in S, k\in K}$$

The discounted value matrix of x will be [3]:

$$v(x) = [I - \gamma P(x)]^{-1} r(x)$$

where I is $N \times N$ identity matrix.

Note that the following recursive formula will hold for the discounted value matrix [3]:

$$v(x) = r(x) + \gamma P(x) v(x)$$

The kth columns of $r(x)$ and $v(x)$ (the immediate expected reward of player k induced by profile x and the discounted value of x to agent k) let us respectively denote $r^k(x)$ and $v^k(x)$.

2.1 Useful Theorems

In this section we will prove a lemma and a theorem for an arbitrary n-player discounted stochastic game $\Gamma = \langle K, S, A^1, \ldots, A^n, \gamma, r^1, \ldots, r^n, p\rangle$.

LEMMA 1. *If $k \in K$, $x \in \Phi$ and $v, \epsilon \in \mathbb{R}^N$ are such that*

$$v \geq r^k(x) + \gamma P(x) v - \epsilon$$

then $v \geq v^k(x) - \sum_{t=0}^{\infty} \gamma^t P^t(x)\epsilon$.

PROOF.

$$\begin{aligned} v &\geq r^k(x) + \gamma P(x)[r^k(x) + \gamma P(x)v - \epsilon] - \epsilon \\ &= r^k(x) + \gamma P(x) r^k(x) + \gamma^2 P^2(x) v \\ &- \epsilon - \gamma P(x)\epsilon \end{aligned}$$

Upon substituting the above inequality into itself i times we obtain:

$$\begin{aligned} v &\geq r^k(x) + \gamma P(x) r^k(x) + \gamma^2 P^2(x) r^k(x) + \\ &+ \ldots + \\ &+ \gamma^{i-1} P^{i-1}(x) r^k(x) + \gamma^i P^i(x) v \\ &- \epsilon - \gamma P(x)\epsilon - \gamma^2 P^2(x)\epsilon - \ldots - \gamma^{i-1} P^{i-1}(x)\epsilon \end{aligned}$$

which upon taking the limit as $i \to \infty$, yields $v \geq v^k(x) - \sum_{t=0}^{\infty} \gamma^t P^t(x)\epsilon$. □

THEOREM 1. *From 1 ⇒ 2*

1. *For each $s \in S$, the vector $(x_s^1, x_s^2, \ldots, x_s^n)$ constitutes an ϵ-equilibrium in the n-matrix game $(B_s^1, B_s^2, \ldots, B_s^n)$ with equilibrium payoffs $(v_s^1, v_s^2, \ldots, v_s^n)$, where for $k \in K$ and $(a^1, a^2, \ldots, a^n) \in A^1 \times A^2 \times \ldots \times A^n$ entry $(a^1, a^2, \ldots, a^n)$ of B_s^k equals*

$$\begin{aligned} b^k(s, a^1, a^2, \ldots, a^n) &= r^k(s, a^1, a^2, \ldots, a^n) \\ &+ \gamma \sum_{s'\in S} (p(s'|s, a^1, a^2, \ldots, a^n) \\ &+ \varsigma(s'|s, a^1, a^2, \ldots, a^n)) \\ &\cdot (v_{s'}^k + \sigma_{s'}^k) \end{aligned}$$

where $-\sigma < \sigma_s^k < \sigma$, $-\varsigma < \varsigma(s'|s, a^1, a^2, \ldots, a^n) < \varsigma$.

2. *x is an ε-equilibrium in the discounted stochastic game Γ where $\varepsilon = [2\gamma(\sigma + \varsigma N \max_{k,s} |v_s^k| + N\varsigma\sigma) + \epsilon] \sum_{t=0}^{\infty} \gamma^t$.*

PROOF.

$$\begin{aligned} b^k(s, a^1, a^2, \ldots, a^n) &= r^k(s, a^1, a^2, \ldots, a^n) + \\ &+ \gamma \sum_{s'\in S} p(s'|s, a^1, a^2, \ldots, a^n) v_{s'}^k + \\ &+ \gamma \sum_{s'\in S} \varsigma(s'|s, a^1, a^2, \ldots, a^n) v_{s'}^k + \\ &+ \gamma \sum_{s'\in S} p(s'|s, a^1, a^2, \ldots, a^n) \sigma_{s'}^k + \\ &+ \gamma \sum_{s'\in S} \varsigma(s'|s, a^1, a^2, \ldots, a^n) \sigma_{s'}^k \\ &= r^k(s, a^1, a^2, \ldots, a^n) + \xi^k(s, a^1, a^2, \ldots, a^n) + \\ &+ \gamma \sum_{s'\in S} p(s'|s, a^1, a^2, \ldots, a^n) v_{s'}^k \end{aligned}$$

where

$$\begin{aligned} \xi^k(s, a^1, a^2, \ldots, a^n) &= \gamma \sum_{s'\in S} p(s'|s, a^1, a^2, \ldots, a^n) \sigma_{s'}^k \\ &+ \gamma \sum_{s'\in S} \varsigma(s'|s, a^1, a^2, \ldots, a^n) v_{s'}^k + \\ &+ \gamma \sum_{s'\in S} \varsigma(s'|s, a^1, a^2, \ldots, a^n) \sigma_{s'}^k \end{aligned}$$

Let's estimate the worst case:

$$\begin{aligned} &- \gamma \sum_{s'\in S} p(s'|s, a^1, a^2, \ldots, a^n)\sigma \\ &- \gamma \sum_{s'\in S} \varsigma \max_{s'} |v_{s'}^k| - \gamma \sum_{s'\in S} \varsigma\sigma \\ &< \gamma \sum_{s'\in S} p(s'|s, a^1, a^2, \ldots, a^n) \sigma_{s'}^k \\ &+ \gamma \sum_{s'\in S} \varsigma(s'|s, a^1, a^2, \ldots, a^n) v_{s'}^k \\ &+ \gamma \sum_{s'\in S} \varsigma(s'|s, a^1, a^2, \ldots, a^n) \sigma_{s'}^k \\ &< \gamma \sum_{s'\in S} p(s'|s, a^1, a^2, \ldots, a^n)\sigma \\ &+ \gamma \sum_{s'\in S} \varsigma \max_{s'} |v_{s'}^k| + \gamma \sum_{s'\in S} \varsigma\sigma \end{aligned}$$

Let's denote $\omega = \gamma\sigma + \gamma\varsigma N \max_{k,s} |v_s^k| + \gamma N \varsigma\sigma$

$$-\omega < \xi^k(s, a^1, a^2, \ldots, a^n) < \omega$$

Let's take some arbitrary $f \in \Theta^1$. If (1) is true, then for each state $s \in S$ by definition of ϵ-equilibrium:

$$\begin{aligned} r^1(s, f, x^2, \ldots, x^n) + \zeta^1(s, f, x^2, \ldots, x^n) \\ + \gamma \sum_{s'\in S} p(s'|s, f, x^2, \ldots, x^n) v_{s'}^1 \leq v_s^1 + \epsilon \end{aligned}$$

where

$$\zeta^k(s,x) = \sum_{a^1\in A^1} \sum_{a^2\in A^2} \ldots \sum_{a^n\in A^n} \xi^k(s, a^1, a^2, \ldots, a^n) \prod_{i=1}^{n} x_{sa^i}^i$$

In the worst case:

$$-\sum_{a^1\in A^1}\sum_{a^2\in A^2}\cdots\sum_{a^n\in A^n}\omega\prod_{i=1}^{n}x^i_{sa^i} <$$

$$< \sum_{a^1\in A^1}\sum_{a^2\in A^2}\cdots\sum_{a^n\in A^n}\xi^k(s,a^1,a^2,\ldots,a^n)\prod_{i=1}^{n}x^i_{sa^i} <$$

$$< \sum_{a^1\in A^1}\sum_{a^2\in A^2}\cdots\sum_{a^n\in A^n}\omega\prod_{i=1}^{n}x^i_{sa^i}$$

$$-\omega < \zeta^k(s,x) < \omega$$

$$r^1(s,f,x^2,\ldots,x^n)+\gamma\sum_{s'\in S}p(s'|s,f,x^2,\ldots,x^n)v^1_{s'}$$

$$\leq \quad v^1_s+\omega+\epsilon$$

Applying the lemma 1 we get

$$v^1(f,x^2,\ldots,x^n)-(\omega+\epsilon)\sum_{t=0}^{\infty}\gamma^t\leq v^1$$

for all $f\in\Theta^1$, and by symmetrical arguments it follows that

$$v^k(x^1,\ldots,x^{k-1},f,x^{k+1},\ldots,x^n)-(\omega+\epsilon)\sum_{t=0}^{\infty}\gamma^t\leq v^k$$

for all $k\in K$ and for all $f\in\Theta^k$.

But

$$v^k=r^k(x)+\gamma P(x)v^k+\zeta^k(x)$$

$$\begin{aligned}v^k &= r^k(x)+\gamma P(x)[r^k(x)+\gamma P(x)v^k+\zeta^k(x)]+\zeta^k(x)\\ &= r^k(x)+\gamma P(x)r^k(x)+\gamma^2P^2(x)v^k\\ &+ \zeta^k(x)+\gamma P(x)\zeta^k(x)\end{aligned}$$

Upon substituting the above inequality into itself and taking the limit, we get:

$$v^k=v^k(x)+\sum_{t=0}^{\infty}\gamma^tP^t(x)\zeta^k(x)$$

$$v^k(x)-\omega\sum_{t=0}^{\infty}\gamma^t<v^k<v^k(x)+\omega\sum_{t=0}^{\infty}\gamma^t$$

$$v^k(x^1,\ldots,x^{k-1},f,x^{k+1},\ldots,x^n)-(\omega+\epsilon)\sum_{t=0}^{\infty}\gamma^t$$

$$\leq \quad v^k(x)+\omega\sum_{t=0}^{\infty}\gamma^t$$

for all $k\in K$ and for all $f\in\Theta^k$.

$$v^k(x^1,\ldots,x^{k-1},f,x^{k+1},\ldots,x^n)\leq v^k(x)+(2\omega+\epsilon)\sum_{t=0}^{\infty}\gamma^t$$

for all $k\in K$ and for all $f\in\Theta^k$.

So the condition of theorem 1 is satisfied with $\varepsilon=[2(\gamma\sigma+\gamma\varsigma N\max_{k,s}|v^k_s|+\gamma N\varsigma\sigma)+\epsilon]\sum_{t=0}^{\infty}\gamma^t$ and we get (2). □

3. ε-EQUILIBRIUM THEOREM

Let

$$\Gamma=\langle K,S,A^1,\ldots,A^n,\gamma,r^1,\ldots,r^n,p\rangle$$

and

$$\widetilde{\Gamma}=\langle K,S,A^1,\ldots,A^n,\gamma,r^1,\ldots,r^n,\widetilde{p}\rangle$$

be n-player discounted stochastic games such that:

for all $s'\in S$ and $(s,a^1,a^2,\ldots,a^n)\in S\times A^1\times A^2\times\ldots\times A^n$:

$$\begin{aligned}\widetilde{p}(s'|s,a^1,a^2,\ldots,a^n) &= p(s'|s,a^1,a^2,\ldots,a^n)\\ &+ \varsigma(s'|s,a^1,a^2,\ldots,a^n)\end{aligned}$$

where $-\varsigma<\varsigma(s'|s,a^1,a^2,\ldots,a^n)<\varsigma$.

Since the agents in reinforcement learning don't know transition probabilities they have only an approximation $\widetilde{\Gamma}$ of the actual game Γ at each learning stage. We suppose though that the agents have already learned the reward functions.

Further on~will indicate that we are using an approximation of transitions $\widetilde{p}$ instead of the actual transitions p for calculation of the corresponding values.

Let's consider the following system of differential equations:

$$\frac{dx^k_{sh}}{dt}=[\vartheta^k_{sh}(x)-\widetilde{v}^k_s(x)]x^k_{sh}\qquad k\in K,s\in S,h\in A^k \tag{1}$$

where

$$\vartheta^k_{sh}(x)=r^k(s,\chi)+\gamma\sum_{s'\in S}\widetilde{p}(s'|s,\chi)\widetilde{v}^k_{s'}(x)$$

and χ denotes the profile equal to x but where player k plays pure strategy h in state s.

Let $x^k_{sh}(t)$ for $k\in K,s\in S$ and $h\in A^k$ be the solution of the system of differential equations 1 satisfying some initial conditions:

$$x^k_{sh}(0)=x^{k(0)}_{sh}\in(0,1),\ \sum_{h\in A^k}x^k_{sh}(0)=1$$

Let y denote the profile where $y^k_{sh}=\frac{1}{T}\int_0^T x^k_{sh}(t)dt$ for some T.

And let $\upsilon=(v^k_s)_{s\in S,k\in K}$ denote the matrix where $\upsilon^k_s=\frac{1}{T}\int_0^T\widetilde{v}^k_s(x(t))dt$.

Let

$$\nu^k_{sh}(y)=r^k(s,\varphi)+\gamma\sum_{s'\in S}\widetilde{p}(s'|s,\varphi)v^k_{s'}$$

where φ stands for the profile equal to y but where player k plays pure strategy h in state s.

THEOREM 2. *If $T,\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4\in\mathbb{R}$ are such that:*

1. *for all $k\in K$ and $s\in S$ $\exists C1^k_s$ and $C2^k_s: C1^k_s\cup C2^k_s=A^k$ and $C1^k_s\cap C2^k_s=\emptyset$ and such that*
   - *(a) for all $h\in C1^k_s$: $\left|\frac{1}{T}\ln x^k_{sh}(T)-\frac{1}{T}\ln x^k_{sh}(0)\right|<\epsilon_1$*
   - *(b) for all $h\in C2^k_s$: $\frac{1}{T}\ln x^k_{sh}(T)-\frac{1}{T}\ln x^k_{sh}(0)\leq-\epsilon_1$*
   - *(c) $\sum_{h\in C2^k_s}y^k_{sh}(\max_{i\in A^k}\nu^k_{si}-\nu^k_{sh})<\epsilon_2$*

2. for all $k \in K$, $s, s' \in S$ and $(a^1, a^2, \ldots, a^n) \in A^1 \times A^2 \times \ldots \times A^n$ the following holds:

(a) $\left| \frac{1}{T} \int_0^T \prod_{i=1, i\neq k}^n x^i_{sa^i}(t)dt - \prod_{i=1, i\neq k}^n y^i_{sa^i} \right| < \epsilon_3$

(b) $\left| \frac{1}{T} \int_0^T \widetilde{v}^k_{s'} \prod_{i=1, i\neq k}^n x^i_{sa^i} dt - \upsilon^k_{s'} \prod_{i=1, i\neq k}^n y^i_{sa^i} \right| < \epsilon_4$

then
$y^k_{sh} = \frac{1}{T} \int_0^T x^k_{sh}(t)dt$ *constitute ε-equilibrium of discounted stochastic game Γ where* $\varepsilon = [2\gamma(\sigma + \varsigma N \max_{k,s} |v^k_s| + N\varsigma\sigma) + \epsilon] \sum_{t=0}^{\infty} \gamma^t$ *and* $\epsilon = 2\epsilon_1 + 2R_{\max}\epsilon_3 \prod_{k=1}^n m^k + 2\gamma\epsilon_4 \prod_{k=1}^n m^k + \epsilon_2$ *and* $\sigma = 3\epsilon_1 + 3R_{\max}\epsilon_3 \prod_{k=1}^n m^k + 3\gamma\epsilon_4 \prod_{k=1}^n m^k + \epsilon_2$.

Proof. Let

$$b^k_{sh} = \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \cdots \sum_{a^{k-1} \in A^{k-1}} \sum_{a^{k+1} \in A^{k+1}} \cdots \sum_{a^n \in A^n} r^k(s, a^1, a^2, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^n) \prod_{i=1, i\neq k}^n y^i_{sa^i} + \gamma \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \cdots \sum_{a^{k-1} \in A^{k-1}} \sum_{a^{k+1} \in A^{k+1}} \cdots \sum_{a^n \in A^n} \sum_{s' \in S} \widetilde{p}(s'|s, a^1, a^2, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^n) \upsilon^k_{s'} \prod_{i=1, i\neq k}^n y^i_{sa^i}$$

and
$b^k_s = \sum_{h \in A^k} b^k_{sh} y^k_{sh}$
If we could show that for some ϵ and σ and for all $k \in K$, $s \in S$ and $h \in A^k$:

1. $b^k_{sh} \leq b^k_s + \epsilon$ (in other words that $y^k_{sh} = \frac{1}{T} \int_0^T x^k_{sh}(t)dt$ constitute ϵ-equilibrium of corresponding games with equilibrium payoffs $(b^1_s, b^2_s, \ldots, b^n_s)$)
2. $\left| b^k_s - \upsilon^k_s \right| < \sigma$

then by applying the theorem 1 we could get the implication in question.

Let's consider arbitrary agent $k \in K$ and state $s \in S$.

Without losing generality let $C1^k_s = \{a^k_1, \ldots, a^k_l\}$ and $C2^k_s = \{a^k_{l+1}, \ldots, a^k_{m^k}\}$.

Let's consider the first case:

Since $T < \infty$ $x^k_{sh}(t) > 0$ on $[0, T]$ and

$$(\ln x^k_{sh})' = \frac{x^{k\,\prime}_{sh}}{x^k_{sh}} = \vartheta^k_{sh}(x) - \widetilde{v}^k_s(x)$$

which gives, if one integrates from 0 to T and then divides by T:

$$\frac{1}{T} \ln x^k_{sh}(T) - \frac{1}{T} \ln x^k_{sh}(0) = \frac{1}{T} \int_0^T \vartheta^k_{sh}(x(t))dt - \frac{1}{T} \int_0^T \widetilde{v}^k_s(x(t))dt$$

Thus

$$\left| \frac{1}{T} \int_0^T \vartheta^k_{sa^k_1}(x(t))dt - \frac{1}{T} \int_0^T \widetilde{v}^k_s(x(t))dt \right| < \epsilon_1$$

$$\vdots$$

$$\left| \frac{1}{T} \int_0^T \vartheta^k_{sa^k_l}(x(t))dt - \frac{1}{T} \int_0^T \widetilde{v}^k_s(x(t))dt \right| < \epsilon_1$$

Upon using the properties of integral we get:

$$\begin{aligned}
&\frac{1}{T} \int_0^T \vartheta^k_{sh}(x(t))dt \\
= &\sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \cdots \sum_{a^{k-1} \in A^{k-1}} \sum_{a^{k+1} \in A^{k+1}} \cdots \sum_{a^n \in A^n} r^k(s, a^1, a^2, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^n) \\
&\cdot \frac{1}{T} \int_0^T \prod_{i=1, i\neq k}^n x^i_{sa^i}(t)dt \\
+ &\gamma \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \cdots \sum_{a^{k-1} \in A^{k-1}} \sum_{a^{k+1} \in A^{k+1}} \cdots \sum_{a^n \in A^n} \sum_{s' \in S} \widetilde{p}(s'|s, a^1, a^2, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^n) \\
&\cdot \frac{1}{T} \int_0^T \widetilde{v}^k_{s'}(x(t)) \prod_{i=1, i\neq k}^n x^i_{sa^i}(t)dt
\end{aligned}$$

Let

$$\frac{1}{T} \int_0^T \prod_{i=1, i\neq k}^n x^i_{sa^i}(t)dt - \prod_{i=1, i\neq k}^n \frac{1}{T} \int_0^T x^i_{sa^i}(t)dt = \epsilon_{3_{sq}}$$

$$q = 1, \ldots, m^1 m^2 \ldots m^n$$

and

$$\frac{1}{T} \int_0^T \widetilde{v}^k_{s'}(x(t)) \prod_{i=1, i\neq k}^n x^i_{sa^i}(t)dt - \frac{1}{T} \int_0^T \widetilde{v}^k_{s'}(x(t))dt \cdot \prod_{i=1, i\neq k}^n \frac{1}{T} \int_0^T x^i_{sa^i}(t)dt = \epsilon_{4_{ss'j}}$$

$$j = 1, \ldots, m^1 m^2 \ldots m^n$$

And according to prerequisite 2:
$-\epsilon_3 < \epsilon_{3_{sq}} < \epsilon_3$ and $-\epsilon_4 < \epsilon_{4_{ss'j}} < \epsilon_4$.
Thus we get:

$$\begin{aligned}
&\frac{1}{T} \int_0^T \vartheta^k_{sh}(x(t))dt \\
= &\sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \cdots \sum_{a^{k-1} \in A^{k-1}} \sum_{a^{k+1} \in A^{k+1}} \cdots \sum_{a^n \in A^n} r^k(s, a^1, a^2, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^n) \\
&\cdot \left(\prod_{i=1, i\neq k}^n \frac{1}{T} \int_0^T x^i_{sa^i}(t)dt + \epsilon_{3_{sq}} \right) \\
+ &\gamma \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \cdots \sum_{a^{k-1} \in A^{k-1}} \sum_{a^{k+1} \in A^{k+1}} \cdots \sum_{a^n \in A^n} \sum_{s' \in S} \widetilde{p}(s'|s, a^1, a^2, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^n) \\
&\cdot \left(\frac{1}{T} \int_0^T \widetilde{v}^k_{s'}(x(t))dt \cdot \prod_{i=1, i\neq k}^n \frac{1}{T} \int_0^T x^i_{sa^i}(t)dt + \epsilon_{4_{ss'j}} \right)
\end{aligned}$$

$$
\begin{aligned}
&\frac{1}{T}\int_0^T \vartheta^k_{sh}(x(t))dt \\
= &\sum_{a^1\in A^1}\sum_{a^2\in A^2}\cdots\sum_{a^{k-1}\in A^{k-1}}\sum_{a^{k+1}\in A^{k+1}}\cdots\sum_{a^n\in A^n} \\
&r^k(s,a^1,a^2,\ldots,a^{k-1},h,a^{k+1},\ldots,a^n) \\
\cdot &\prod_{i=1,i\neq k}^{n}\frac{1}{T}\int_0^T x^i_{sa^i}(t)dt \\
+ &\gamma\sum_{a^1\in A^1}\sum_{a^2\in A^2}\cdots\sum_{a^{k-1}\in A^{k-1}}\sum_{a^{k+1}\in A^{k+1}}\cdots\sum_{a^n\in A^n}\sum_{s'\in S} \\
&\widetilde{p}(s'|s,a^1,a^2,\ldots,a^{k-1},h,a^{k+1},\ldots,a^n) \\
\cdot &\frac{1}{T}\int_0^T \widetilde{v}^k_{s'}(x(t))dt\cdot\prod_{i=1,i\neq k}^{n}\frac{1}{T}\int_0^T x^i_{sa^i}(t)dt+\epsilon_{5^k_{sh}}
\end{aligned}
$$

where $-\epsilon_5 < \epsilon_{5^k_{sh}} < \epsilon_5$ and $\epsilon_5 = m^1m^2\ldots m^n(R_{\max}\epsilon_3 + \gamma\epsilon_4) = (R_{\max}\epsilon_3+\gamma\epsilon_4)\prod_{k=1}^n m^k$

Apparently $\frac{1}{T}\int_0^T \vartheta^k_{sh}(x(t))dt = b^k_{sh} + \epsilon_{5^k_{sh}}$.

Hence we will have the following inequalities:

$$-\epsilon_6 < b^k_{sa^k_1} - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt < \epsilon_6$$

$$\vdots$$

$$-\epsilon_6 < b^k_{sa^k_l} - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt < \epsilon_6$$

where $\epsilon_6 = \epsilon_1 + \epsilon_5$

So the difference between b^k_{sh} for $h \in C1^k_s$ and $\frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt$ won't exceed ϵ_6. Hence the difference between any two $b^k_{sh_1}$ and $b^k_{sh_2}$, $h_1, h_2 \in C1^k_s$ won't be more than $2\epsilon_6$.

For all $h \in C2^k_s$ the following condition holds:

$$\frac{1}{T}\ln x^k_{sh}(T) - \frac{1}{T}\ln x^k_{sh}(0) \leq -\epsilon_1$$

$$
\begin{aligned}
&\frac{1}{T}\int_0^T \vartheta^k_{sh}(x(t))dt - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt \\
= &\frac{1}{T}\ln x^k_{sh}(T) - \frac{1}{T}\ln x^k_{sh}(0) \leq -\epsilon_1
\end{aligned}
$$

$$b^k_{sh} + \epsilon_{5^k_{sh}} - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt \leq -\epsilon_1$$

$$b^k_{sh} - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt < \epsilon_5 - \epsilon_1 < \epsilon_5 + \epsilon_1 = \epsilon_6$$

Let $b^k_{sh_*} = \max_{h\in A^k} b^k_{sh}$.

If $h_* \in C1^k_s$ then for any $h \in C1^k_s$ the difference between corresponding b^k_{sh} and $b^k_{sh_*}$ won't exceed $2\epsilon_6$ (as we have already demonstrated the difference between any two $b^k_{sh_1}$ and $b^k_{sh_2}$, $h_1, h_2 \in C1^k_s$ won't be more than $2\epsilon_6$). If $h_* \in C2^k_s$ then for any $h \in C1^k_s$ the difference between corresponding b^k_{sh} and $b^k_{sh_*}$ also won't exceed $2\epsilon_6$ (b^k_{sh} from $C2^k_s$ that deviates from $\frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt$ on more than ϵ_6 can't for sure be the maximal for the whole A^k because it will be less than any b^k_{sh} for $h \in C1^k_s$).

The condition

$$\sum_{h\in C2^k_s} y^k_{sh}(\max_{i\in A^k}\nu^k_{si} - \nu^k_{sh}) < \epsilon_2$$

we can rewrite as

$$\sum_{h\in C2^k_s} b^k_{sh_*}y^k_{sh} - \sum_{h\in C2^k_s} b^k_{sh}y^k_{sh} < \epsilon_2$$

$$\sum_{h\in C2^k_s} b^k_{sh}y^k_{sh} > \sum_{h\in C2^k_s} b^k_{sh_*}y^k_{sh} - \epsilon_2$$

$b^k_s = \sum_{h\in A^k} b^k_{sh}y^k_{sh} = \sum_{h\in C1^k_s} b^k_{sh}y^k_{sh} + \sum_{h\in C2^k_s} b^k_{sh}y^k_{sh} > \sum_{h\in C1^k_s}(b^k_{sh_*} - 2\epsilon_6)y^k_{sh} + \sum_{h\in C2^k_s} b^k_{sh_*}y^k_{sh} - \epsilon_2 = b^k_{sh_*}\sum_{h\in A^k} y^k_{sh} - 2\epsilon_6\sum_{h\in C1^k_s} y^k_{sh} - \epsilon_2 > b^k_{sh_*} - 2\epsilon_6 - \epsilon_2$

Thus the first inequality $b^k_{sh} \leq b^k_s + \epsilon$ that we must prove will hold with $\epsilon = 2\epsilon_6 + \epsilon_2$ for all $h \in A^k$.

As we have just demonstrated for all $h \in A^k$:

$$b^k_{sh} - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt < \epsilon_6$$

If we multiply each inequality by y^k_{sh} accordingly and sum up we will get:

$$\sum_{h\in A^k} b^k_{sh}y^k_{sh} - \sum_{h\in A^k}\frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt\cdot y^k_{sh} < \sum_{h\in A^k}\epsilon_6 y^k_{sh}$$

$$b^k_s - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt < \epsilon_6$$

From the first inequality that we have already proved and the estimations of b^k_{sh} for $h \in C1^k_s$ we can derive:

$$-\epsilon_6 < b^k_{sh} - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt \leq b^k_s + \epsilon - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt$$

$$-\epsilon_6 - \epsilon < b^k_s - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt$$

The second inequality:
$|b^k_s - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt| < \sigma$ will hold with $\sigma = 3\epsilon_6 + \epsilon_2$.
Let's calculate ϵ and σ.

$$
\begin{aligned}
\epsilon &= 2\epsilon_6 + \epsilon_2 = 2(\epsilon_1+\epsilon_5) + \epsilon_2 = \\
&= 2(\epsilon_1 + (R_{\max}\epsilon_3 + \gamma\epsilon_4)\prod_{k=1}^n m^k) + \epsilon_2 = \\
&= 2\epsilon_1 + 2R_{\max}\epsilon_3\prod_{k=1}^n m^k + 2\gamma\epsilon_4\prod_{k=1}^n m^k + \epsilon_2
\end{aligned}
$$

$$
\begin{aligned}
\sigma &= 3\epsilon_6 + \epsilon_2 = 3(\epsilon_1+\epsilon_5) + \epsilon_2 = \\
&= 3(\epsilon_1 + (R_{\max}\epsilon_3 + \gamma\epsilon_4)\prod_{k=1}^n m^k) + \epsilon_2 = \\
&= 3\epsilon_1 + 3R_{\max}\epsilon_3\prod_{k=1}^n m^k + 3\gamma\epsilon_4\prod_{k=1}^n m^k + \epsilon_2
\end{aligned}
$$

Applying the theorem 1 we get the the implication in question. □

4. DISCUSSION AND EXPERIMENTAL ESTIMATIONS

Let's consider the conditions of the theorem 2 in detail. For each orbit $x^k_{sh}(t)$ there are only two possibilities:

1. for any $t \in [0, \infty)$ the orbit $x^k_{sh}(t)$ remains bounded from 0 on some value $\delta > 0$
2. $x^k_{sh}(t)$ comes arbitrarily close to 0

In the first case we can reduce ϵ_1 arbitrarily by increasing T (k belongs to $C1^k_s$ in this case).

In the second case if the condition on ϵ_1 for class $C1^k_s$ holds — k belongs to $C1^k_s$ otherwise to $C2^k_s$ ($\frac{1}{T}\ln x^k_{sh}(T) - \frac{1}{T}\ln x^k_{sh}(0) > 0$ will never be true for big enough T).

We can arbitrarily decrease ϵ_2 by increasing T in the second case since ν^k_{sh} is a bounded function.

ϵ_3 and ϵ_4 are much more difficult to deal with...

In general the systems of differential equation can be solved:

1. analytically (solution in explicit form)
2. qualitatively (with the use of vector fields)
3. numerically (numerical methods, e.g., Runge-Kutta)

It is hopeless to try to solve the system of such complexity as 1 by the first two approaches and therefore a proof that its solutions satisfy the prerequisites of the theorem 2 seems to us non-trivial. Till now we have managed to find ϵ_3 and ϵ_4 estimations only experimentally.

In table 1 estimations of average relative $\overline{\epsilon_{5^k_{sh}}}$ and average relative $\overline{\epsilon_5}$ are presented for different game classes (with different number of states, agents and actions). The averages are calculated for 100 games of each class and $T = 1000$. The initial conditions for the system of differential equations 1 were chosen quite randomly. The games are generated with uniformly distributed payoffs. Transition probabilities were also derived from uniform distribution. As we can see the preconditions of the theorem 2 hold with a quite acceptable accuracy for all the classes.

5. NASH-DE ALGORITHM

To propose an algorithm we have to make one more assumption that we have managed to confirm only experimentally till now, namely:

The more accurate approximation of Nash-equilibrium we choose as an initial condition for our system 1 the more precisely the prerequisites of the theorem 2 hold.

So now we can propose an iterative algorithm for calculating ϵ-equilibria of discounted stochastic games with some given accuracy ϵ (see algorithm 1).

An example of its work on a 2-state 2-agent 2-action discounted stochastic game is presented on the figure 1 (because of the space restrictions we are illustrating convergence only for state s_1 but no state can be examined in isolation for analysis). On each figure the probabilities assigned to the first actions of the first and the second agents are presented as xy-plot (it is quite descriptive since the probabilities of the second actions are equal to one minus probabilities of the first ones). The solutions are lighter at the end of $[0, T]$ interval. The precise Nash-equilibrium is designated by a star and the average $\frac{1}{T}\int_0^T x^k_{sh}(t)dt$ for each iteration — by a cross. Since the agents in reinforcement learning don't know either transition probabilities or reward functions and they learn them online the first policies are quite random. The algorithm converges in self-play to Nash-equilibrium with the given relative accuracy $\epsilon = 1\%$ in two iterations.

Algorithm 1 Nash-DE algorithm for the player k

Input: accuracy ϵ, T
for all $s \in S$, $k \in K$ and $h \in A^k$ **do**
  $x^k_{sh}(0) \leftarrow 1/|A^k|$
end for
while $x(0)$ doesn't constitute ϵ-equilibrium **do**
  Find solution of the system 1 through the point $x(0)$ on the interval $[0, T]$ (updating model in parallel)
  Let the initial point be $x^k_{sh}(0) = \frac{1}{T}\int_0^T x^k_{sh}(t)dt$
end while

6. EXPERIMENTAL RESULTS

Since the agents in reinforcement learning don't know either transition probabilities or reward functions they have to approximate the model somehow. We tested our algorithm as an off-policy version (the agents pursue the best learned policy so far in the most of cases (we chose — 90% of cases) and explore the environment in 10% of cases). The results of the experiments are presented in table 1. The number of independent transitions to be learned can be calculated by the formula $Tr = N(N-1)\prod_{k=1}^{n} m^k$ and is presented in the corresponding column for each game class. In column "Iterations" the average number of iterations necessary to find Nash-equilibrium with relative accuracy $\epsilon = 1\%$ is presented. And in the last column — the percentage of games for which we managed to find Nash-equilibrium with the given relative accuracy $\epsilon = 1\%$ after 500 iterations.

In general one can see the following trend: the larger is the model the more iterations the agents require to find a 1%-equilibrium, and the oftener they fail to come to this equilibrium during 500 iterations.

The main reason for it is their incapability to approximate large models to the necessary accuracy (their approximations of transition probabilities are too imprecise — they explore the environment only in 10% of cases each and the transition probabilities of some combinations of actions remain very poorly estimated) and as a result they can't find an equilibrium or converge to it more slowly (let us not forget that the accuracy of transition probabilities acts as a relative factor and comes to ε estimation of theorem 2 multiplied by the maximal discounted value). In order to decrease the average number of iterations and to increase the percentage of solved games it appears promising to test a version of the algorithm with a more intensive exploration stage (first learn the model to some given precision and only then act according to the policy found by the algorthm and keep on learning in parallel). For instance, it can be achieved by setting ϵ to larger values at the beginning.

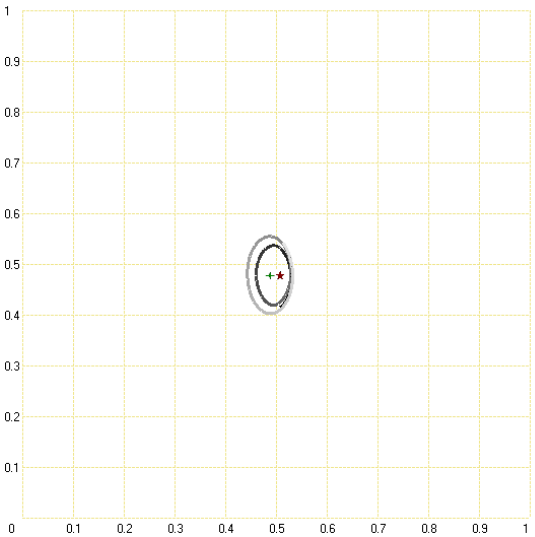
7. CONCLUSION

This paper is devoted to an actual topic of extending reinforcement learning approach for multiagent systems. An

Table 1: Estimations and Results of Experiments

States	*Agents*	*Actions*	Tr	$\overline{\epsilon_{5^k_{sh}}}$	$\overline{\epsilon_5}$	*Iterations*	*%*
2	2	2	8	0.08%	0.24%	11.23	98%
2	2	3	18	0.20%	0.36%	9.43	95%
2	2	5	50	0.16%	0.25%	18.60	90%
2	2	10	200	0.48%	0.73%	38.39	94%
2	3	2	16	0.18%	0.85%	16.03	87%
2	3	3	54	0.68%	1.74%	30.64	91%
2	5	2	64	1.80%	4.36%	27.79	87%
5	2	2	80	0.00%	0.04%	31.60	83%
5	2	3	180	0.14%	0.22%	52.26	93%
5	2	5	500	0.10%	0.14%	62.74	91%
5	3	3	540	0.35%	1.58%	85.83	75%
10	2	2	360	0.02%	0.06%	69.68	82%

(a) Iteration 1 State 1



(b) Iteration 2 State 1

Figure 1: Convergence of Algorithm 1

algorithm based on system of differential equations of special type is developed. A formal proof of its convergence with a given accuracy to Nash equilibrium for environments represented as general-sum discounted stochastic games is given under some assumptions. We claim that it is the first algorithm that converges to Nash equilibrium in general case. Thorough testing showed that the assumptions necessary for the formal convergence hold with quite a good accuracy that allowed the proposed algorithm to find Nash equilibrium with relative accuracy of 1% in approximately 90% of randomly generated games.

8. REFERENCES

[1] M. H. Bowling and M. M. Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136(2):215–250, 2002.

[2] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *AAAI '98/IAAI '98: Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, pages 746–752, Menlo Park, CA, USA, 1998. American Association for Artificial Intelligence.

[3] J. Filar and K. Vrieze. *Competitive Markov decision processes*. Springer-Verlag New York, Inc., New York, NY, USA, 1996.

[4] A. Greenwald and K. Hall. Correlated-q learning. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 242–249, 2003.

[5] J. Hu and M. P. Wellman. Multiagent reinforcement learning: theoretical framework and an algorithm. In *Proc. 15th International Conf. on Machine Learning*, pages 242–250. Morgan Kaufmann, San Francisco, CA, 1998.

[6] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *ICML*, pages 157–163, 1994.

[7] M. L. Littman. Friend-or-foe q-learning in general-sum games. In C. E. Brodley and A. P. Danyluk, editors, *ICML*, pages 322–328. Morgan Kaufmann, 2001.

[8] G. Tesauro. Extending q-learning to general adaptive multi-agent systems. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *NIPS*. MIT Press, 2003.

[9] M. Zinkevich, A. R. Greenwald, and M. L. Littman. Cyclic equilibria in markov games. In *NIPS*, 2005.