Rokhlin D.B. (Southern Federal University, Rostov-on-Don, Russia). **Q-learning in a stochastic Stackelberg game.**

We consider a game between a leader and a follower, where the players' actions affect the stochastic evolution of the state process x_t, $t \in \mathbb{Z}_+$. The players observe their rewards and the system state x_t. They do not know the transition kernel of the process x and the reward function of the other player. At each stage of the game the leader selects his action a_t first. This action is known to the follower before he selects b_t. Follower's actions are unknown to the leader (uniformed leader). Each player tries to maximize his discounted reward, applying the Q-learning algorithm [1]. A special feature of the algorithm under consideration is that, when updating his Q-function, the follower believes that the action of the leader in the next state will be the same as in the current one (naive follower). Under other assumptions the Q-learning algorithm for the stochastic Stackelberg game was considered in [2].

Let X be a finite state space. Denote by A and B the sets of admissible actions of the leader and the follower. Assume that the evolution of the system is characterized by the transition kernel $p(y|x,a,b)$. This means that if the system is at the state $x \in X$, and the leader and follower select $a \in A$, $b \in B$, then the probability of transition to the state $y \in X$ equals to $p(y|x,a,b)$. At each stage of the game

- the leader observes the system state $x_t \in X$ and selects an action $a_t \in A$,
- the follower observes the system state and leader's action and selects an action $b_t \in B$,
- the system moves to a state x_{t+1} with probability $p(x_{t+1}|x_t,a_t,b_t)$,
- the leader and the follower get the rewards $r_1(x_t,a_t,b_t,x_{t+1})$ and $r_2(x_t,a_t,b_t,x_{t+1})$ respectively.

These rules imply the inequality of players which is typical for a Stackelberg game.

Randomized strategies of players are defined as Boltzmann distributions, depending on the Q-functions Q^l, Q^f of the leader and follower that are updated in the course of learning. So, at each stage of the game the leader and follower sequentially select their actions $a \in A$, $b \in B$ with probabilities

$$\frac{\exp(Q^l(x,a)/\tau_1)}{\sum_{a' \in A} \exp(Q^l(x,a')/\tau_1)}, \qquad \frac{\exp(Q^f(x,a,b)/\tau_2)}{\sum_{b' \in B} \exp(Q^f(x,a,b')/\tau_2)}.$$

It is shown that the existence of deterministic stationary strategies generating an irreducible Markov chain is sufficient for the convergence of the algorithm. The proof is based on the known results [3], related to the development of the idea of stochastic approximation.

The limiting large time behaviour of Q-functions is described in terms of controlled Markov processes, which for clarity are related to the virtual leader and the virtual follower. The distributions of the players' actions converge to the Boltzmann distributions, depending on the limiting Q-functions. We also consider the behaviour of these limiting distributions for small "temperature" parameters τ_i.

REFERENCES

1. *Watkins C.J.C.H., Dayan P.* Q-learning, Machine learning, 1992, vol. 8, № 3, pp. 279-292.
2. *Könönen, V.* Asymmetric multiagent reinforcement learning, Web intelligence and agent systems: an international journal, 2004, vol. 2, № 2, pp. 105-121.
3. *Bertsekas D.P., Tsitsiklis, J.N.* Neuro-dynamic programming, Athena Scientific: Belmont, MA, 1996.

The research is supported by the Russian Science Foundation, project 17-19-01038.