Andhill-98: A RoboCup Team which Reinforces Positioning with Observation *

Tomohito Andou

C&C Media Research Laboratories, NEC Corporation
Miyazaki 4-1-1, Miyamae-ku, Kawasaki, 216-8555, Japan
E-mail: tandou@ccm.cl.nec.co.jp

Abstract. On reinforcement learning with limited exploration, an agent's policy tends to fall into a worthless local optimum. This paper proposes Observational Reinforcement Learning method with which the learning agent evaluates inexperienced policies and reinforces it. This method provides the agent more chances to escape from a local optimum without exploration. Moreover, this paper shows the effectiveness of the method from experiments in the RoboCup positioning problem. They are advanced experiments described in our RoboCup-97 paper[1].

1 Introduction

Andhill [2] won the second prize of RoboCup-97 simulator league and the championship of RoboCup Japan Open 98 simulator league. In these competitions, Andhill's on-line learning mechanism was only used in a few games because of the insufficiency of the advantage from fixed-positioning strategies. In RoboCup-98, the offside rule was introduced and it was expected that positioning strategies of many teams would be changed. A learning mechanism would be effective if strategies of the opponent were unimaginable. Therefore, we used an on-line learning mechanism in many games at RoboCup-98.

The on-line learning mechanism of Andhill-98 was implemented with consideration of the following three points. First, the agents should behave on their best policies during a whole game because the game is not a practice game. So agents are set to explore very little policies. Second, the learning results must be reflected within a very short game. The learning agent's policy can be started from a randomly initialized policy, a zero initialized policy, or an already learned policy. We chose a zero initialized policy because it is the most sensitive way for short-term learning. Third, learning from zero is too dangerous especially at the first period of the game. To avoid the danger, Andhill-98 started with the fixed-positioning strategy while learning, and switched to the learned-positioning strategy when the team scores.

* This work was mainly done when the author was in Dept. of Mathematical and Computing Sciences, Tokyo Institute of Technology.

[2] The team name Andhill is a parody of "anthill" in which ants live. This comes from its character of a multi-agent system.

M. Asada and H. Kitano (Eds.): RoboCup-98, LNAI 1604, pp. 338-345, 1999.

The first point above, namely, limited exploration is an essential difficulty on reinforcement learning. To deal with this difficulty, Andhill-98 used Observational Reinforcement Learning method. This paper describes the RoboCup positioning problem in section 2, Observational Reinforcement Learning method in section 3, experiments in section 4, and conclusions in section 5.

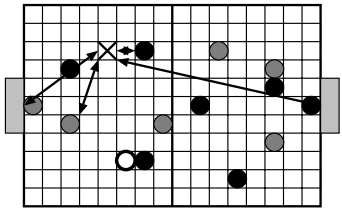
2 The RoboCup Positioning Problem

A RoboCup agent has mainly the following three routines:

1. If the ball is within kickable range: Where will it kick the ball to?
2. If the ball is close from it: How will it catch the ball?
3. If the ball is far from it: Where will it run to?

Andhill-98 is designed as (1) and (2) are already set by human, and (3) can be obtained by learning.

About (3), most of the RoboCup teams applied fixed-positioning strategies while dynamic positioning strategies are general in real-world soccer games. Dynamic positioning can be determined by something like a positioning function which inputs an agent's environment and outputs its suitable position. We attempted to learn such a dynamic positioning function by on-line reinforcement learning. We designed the function as a three layer neural network $(5-6-1)$ which inputs the following elements for each place and outputs the suitability of the place for positioning.



1. Distance to the goal of the agent's side
2. Distance to the goal of the opponents' side
3. Distance to the closest team-mate
4. Distance to the closest opponent
5. Which team possesses the ball

Fig. 1. Field information

3 Observational Reinforcement Learning

3.1 Difficulties in an ordinary reinforcement learning method

In an ordinary reinforcement learning method, an agent can only reinforce its experienced policies[2]. This restriction causes the following two difficulties in the RoboCup positioning problem. The first difficulty is the fact that positioning is

a combined action. While previous reinforcement learning techniques had only dealt with primitive actions, we had to also apply reinforcement learning to combined actions. In RoboCup problems, positioning consists of two kinds of primitive actions, dash and turn. A combined action usually costs much time and/or something else. The RoboCup positioning costs much time and great stamina. This means exploration can not be executed sufficiently in on-line learning.

The second difficulty is the fact that positioning is a cooperative action. In the case of on-line learning, all agents have to learn simultaneously in a cooperative multi-agent environment. The learning of an agent requires the other agents behave stably in their best policies. It means that in order to learn better, they can not explore. This dilemma limits trials and errors. Both of the difficulties limit exploration of an agent. In an ordinary reinforcement learning method with limited exploration, an agent's policy tends to fall into a worthless local optimum. We proposed Observational Reinforcement Learning method which offers more chances to escape from worthless local optimal policies without exploration.

3.2 Observational Reinforcement Learning method

Observational Reinforcement Learning method is a reinforcement learning method in which an agent can also reinforce an inexperienced policy which is evaluated as good from its observation. In the RoboCup positioning problem, an agent can evaluate some positions as good just only from its observation. One evaluation is like: A place where the ball comes frequently will suit for positioning. In this method, an agent can reinforce suitable places immediately by this evaluation. The agent needs no actual experience of positioning to reinforce the policy. Therefore, an agent can reinforce various positioning independently of the cost of the positioning or the agent's actual behaviors. Observational Reinforcement Learning method enables an agent to reinforce not only low-costed policies but various policies. Moreover, an agent can reinforce various policies while behaving in its best policies. Consequently it offers more chances to escape from worthless local optimal policies.

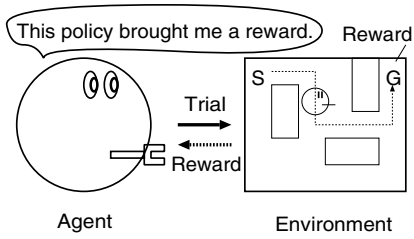


Fig. 2. Ordinary reinforcement learning

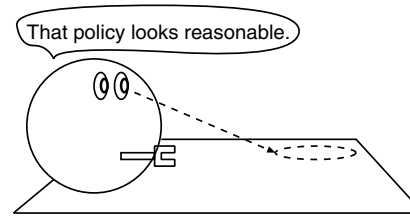


Fig. 3. Observational Reinforcement Learning

Observational Reinforcement Learning method could be regarded as a combination method of reinforcement learning and supervised learning. Therefore, supervised learning techniques like a complementary error back propagation algorithm can be applied in this method.

3.3 Comparison with ordinary reinforcement learning

Ordinary reinforcement learning can be regarded as an imitation of the process of gaining "confidence", because the learning agent gets rewards when it acts something good and becomes sure that the action is good. Observational Reinforcement Learning is, however, an imitation of "regret", because the learning agent judges its recent action was worse than another action and becomes sure that the unexecuted action would be good.

As mentioned above, Observational Reinforcement Learning offers more chance to escape from a local optimal policy and this advantage is remarkable in on-line learning. However, it also has a weakness especially in multi-agent learning. Generally, a multi-agent system works efficiently by diversity among the agents. Diversity among the agents is important on a cooperative behavior. Ordinary reinforcement learning agents have some diversity because experience is unique for each agent, but Observational Reinforcement Learning agents have little diversity because observation tends to similar among all agents. Thus both can make up the weakness of the other.

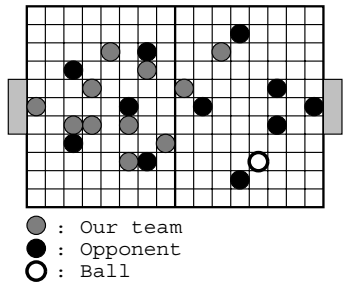
	Type of Learning	On-line Learning	Diversity in a Multi-Agent System
Ordinary RL	Confidence	Difficult	Various
Observational RL	Regret	Easier	Monotonous

4 Experiments

We attempted the RoboCup positioning problem with Observational Reinforcement Learning method. We compared the following three experiments:

- **Experiment 1 – Ordinary reinforcement learning**: An agent gets a reward when it kicks the ball, and then it reinforces its recent policies.
- **Experiment 2 – Observational Reinforcement Learning**: An agent reinforces the ball location when all team-mates are far from the ball. See Fig. 4.
- **Experiment 3 – Combination of ordinary reinforcement learning and Observational Reinforcement Learning**: Both reinforcements in Experiment 1 and Experiment 2 are to be done.

Experiments were executed in an unending on-line game between a learning team and a fixed position team on the RoboCup-97 rule. Ten players of the learning team excepting a goal-keeper were learning simultaneously. Exploration is not used explicitly. That is, all learning agents choose policies which are they



The ball location is far from all team-mates. Then, they will reinforce a policy of staying at the place.

Fig. 4. A situation in which agents regret

think the best. In the experiments, the fixed position team was Andhill-97, and the learning team was a position learning team who has the same faculties of Andhill-97 excepting the positioning. The learning was started with randomly initialized policies. Each experiment needed about 20 hours until policies of all the learning agents would converge.

The next table shows the experimental results. Values of the table represent averaged scores per a game of over 60 games. In the case that both of the teams are the same fixed position teams, the game will end in score 6.0 to 6.0 on the average. The learning team won against Andhill-97 only in Experiment 3.

	Experiment 1	Experiment 2	Experiment 3
Learned team score	2.6	4.5	5.4
Fixed position team score	3.0	9.3	5.0
Margin	-0.4	-4.8	0.4

Details of each experimental results are described in Section 4.1, Section 4.2, and Section 4.3.

4.1 Results of Experiment 1 (Ordinary RL)

In Experiment 1, we attempted the RoboCup positioning problem with an ordinary reinforcement learning method. In this method, an agent will get a reward when the agent is doing something good, and will reinforce the recent behaviors. The reward was defined as: When an agent kicks the ball, a reward is given to it.

Fig. 5. shows the policy transitions of 10 agents of the learning team. The horizontal axis means the game time count, and the vertical axis means the distance from their goal. The values of the vertical axis are regularized into $(-1.0, 1.0)$ and averaged from the behaviors of 10 games. This figure shows that the agents' policies branched into two types. There were seven agents who has defensive policies of positioning, and four agents who has offensive policies. This

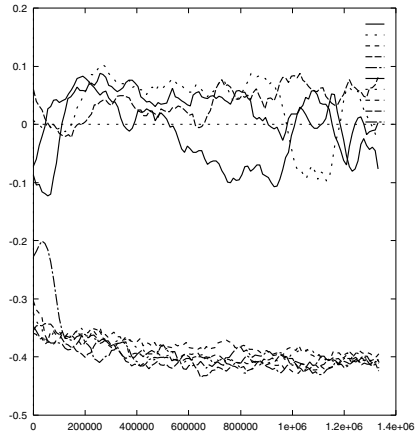


Fig. 5. Policy transitions of 10 agents in Experiment 1: The vertical axis means the distances from their goal.

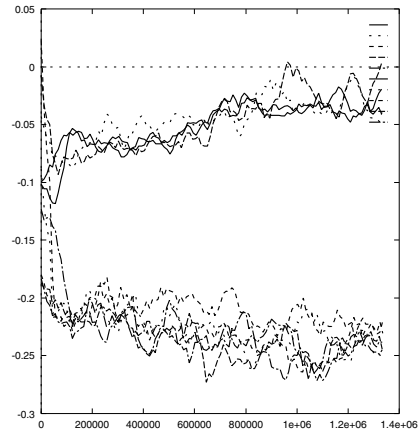


Fig. 6. Policy transitions of 10 agents in Experiment 1: The vertical axis means the distances from their closest team-mates.

strategy is a little too defensive, so the score rate was slightly less than that of the fixed positioning.

The vertical axis of Fig. 6. means the distance from a learning agent to the closest team-mate. The values are also regularized into $(-1.0, 1.0)$ and averaged from the behaviors of 10 games. There were seven agents which gather into a lump, and four agents which keep away from other agents. Two types of policies never crossed nor joined. This means that it had little possibilities of escaping from local optimal policies. This was the most important difference from the other two experiments.

4.2 Results of Experiment 2 (Observational RL)

In Experiment 2, Observational Reinforcement Learning was used independently. A learning agent would reinforce good looking policies. The good looking policy was defined as: When all team-mates are far from the ball, the ball location would be a good place for positioning.

Fig. 7. corresponds to Fig. 5. of Experiment 1. The horizontal axis means the game count, and the vertical axis means the distance from their goal. This figure shows that the all agents' policies were similar. As mentioned in section 3, Observational Reinforcement Learning tended to make the diversity monotonous. Fig. 8. corresponds to Fig. 6. of Experiment 1. This figure shows the same tendency that the diversity was monotonous. In this strategy, all the agents were keeping around the ball. This looked good positioning, but not so good in practice because of little diversity of agents' policies and factors of stamina and so on.

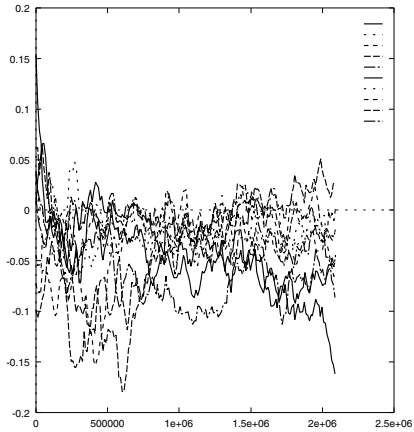


Fig. 7. Policy transitions of 10 agents in Experiment 2: The vertical axis means the distances from their goal.

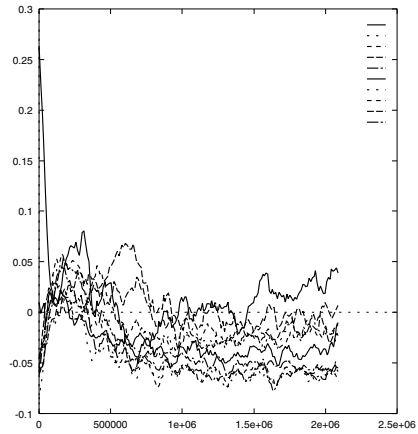


Fig. 8. Policy transitions of 10 agents in Experiment 2: The vertical axis means the distances from their closest team-mates.

4.3 Results of Experiment 3 (Ordinary RL and Observational RL)

In Experiment 3, a combination method of ordinary reinforcement learning and Observational Reinforcement Learning was used. A learning agent would get a reward when the agent kicked the ball. This was the same way of Experiment 1. A learning agent would also reinforce good looking positioning which was the ball location when all team-mates were far from it. This was the same way of Experiment 2. It was expected that there are more possibilities of escaping from local optimal policies than those of Experiment 1, and that there is more diversity of agents' policies than that of Experiment 2.

Fig. 9. corresponds to Fig. 5. of Experiment 1 and Fig. 7. of Experiment 2. There were two types of policies. Three agents were defensive and eight agents were offensive. This means that the learning agents had more diversity than that of Experiment 2. It is the same tendency of Experiment 1, but this has an important difference. The agents' policies crossed or joined occasionally. This means that there were more possibilities of escaping from local optimal policies than those of Experiment 1.

Fig. 10. corresponds to Fig. 6. of Experiment 1 and Fig. 8. of Experiment 2. This figure shows the most complicated process of learning. There is a great change of policies before time count 600000. We believe that this is just an important process of getting cooperative policies in multi-agent learning.

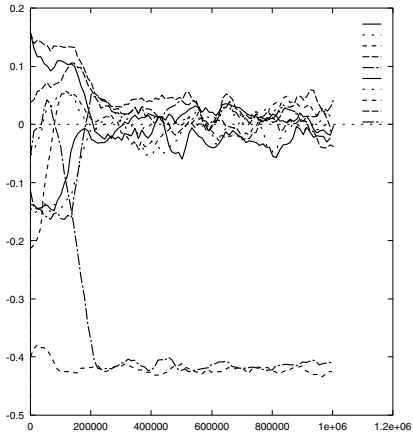


Fig. 9. Policy transitions of 10 agents in Experiment 3: The vertical axis means the distances from their goal.

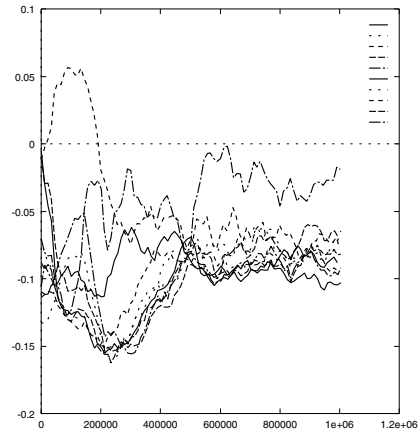


Fig. 10. Policy transitions of 10 agents in Experiment 3: The vertical axis means the distances from their closest team-mates.

5 Conclusions

This paper proposed the Observational Reinforcement Learning method and compared it with an ordinary reinforcement learning method. Experiments showed that both of the methods have important roles of learning. The followings are the comparison among the three methods.

- **Ordinary reinforcement learning independently**: This is effective on developing diversity of the learning agents. However, it tends to fall into a worthless local optimal policy.
- **Observational Reinforcement Learning independently**: This is effective on escaping from worthless local optimal policies. However, it tends to lose diversity of the learning agents.
- **Combined method of ordinary reinforcement learning and Observational Reinforcement Learning**: This can fill up the each other method's weak point.

References

1. Andou, T.: "Refinement of Soccer Agents' Positions Using Reinforcement Learning", In *RoboCup-97: Robot Soccer World Cup I, pp.373 – 388* (1998).
2. Kaelbling, L. P.: "Reinforcement Learning: A Survey", *Journal of Artificial Intelligence Research 4, pp.237 – 285* (1996).