Motion Learning of Robot by Reinforcement Learning under POMDPs Environments

Ryunosuke NOBORI and Yuko OSANA

School of Computer Science, Tokyo University of Technology
1404-1 Katakura, Hachioji, Tokyo, 192-0982, Japan Email: osana@stf.teu.ac.jp

Abstract—In this paper, motion learning (maze problem) of bipedal walking robot in POMDPs (Partially Observable Markov Decision Processes) environment is realized by the Profit Sharing that can learn deterministic policy for POMDPs environments. In this research, the Profit Sharing that can learn deterministic policy for POMDPs environments which can obtain the deterministic policy by using the history of observations is employed. We carried out a series of experiments using bipedal walking robot, and confirmed that motion learning (maze problem) can be realized by the Profit Sharing that can learn deterministic policy for POMDPs environments.

1. Introduction

Reinforcement learning is one of the most active research areas in artificial intelligence, and it is a computational approach to learning whereby an agent tries to maximize the total amount of reward[1]. Reinforcement learning algorithms attempt to find a policy that maps states of the world to the actions the agent ought to take in those states. Recently, we have proposed the Profit Sharing that can learn deterministic policy for POMDP (Partially Observable Markov Decision Processes)s environments[2]. This method can obtain the deterministic policy by using the history of observations. Moreover, we applied this learning method to the real robots (bipedal robot and quadrupedal walking robot)[4][5]. However, these problem are not in POMDPs environments.

In this paper, motion learning (maze problem) of bipedal walking robot in POMDPs environment is realized by the Profit Sharing that can learn deterministic policy for POMDPs environments[2].

2. Profit Sharing that can Learn Deterministic Policy for POMDPs Environments

Here, we explain the Profit Sharing that can learn deterministic policy for POMDPs environments[2] which is used in this research.

In the learning method, when an observation is judged as perceptual aliasing, an action is selected based on the history of observations. In the observation judged as perceptual aliasing, if enough observation sequences are not considered when an action is selected, observation sequences including past observation are also considered. Moreover, in this learning method, the deterministic rate of actions of each observation and the progress of learning in order to detect perceptual aliasing.

The flow of the learning method is shown in Fig.1.

2.1. Action Selection

In this learning method, the action is selected based on the ratio of rule values of the current observation by the Boltzmann selection when the state which is not judged as perceptual aliasing. In contrast, the action is selected based on the ratio of rule values of observation sequences by the Boltzmann selection when the state which is judged as perceptual aliasing,

In the observation at the time x (o_x), the action a is selected based on the probability $P(o_x, a, x)$. The probability $P(o_x, a, x)$ is given by

$$P(o_x, a, x) = \begin{cases} \dfrac{\exp(q_n(o_x, a)/T(o_x))}{\displaystyle\sum_{b \in C^A} \exp(q_n(o_x, b)/T(o_x)),} & (o_x \notin C^{PA}) \\ \dfrac{\exp(q_n(o_x, a)/T(o_x)) + Q(o_x, a, x)}{\displaystyle\sum_{b \in C^A} \Big(\exp(q_n(o_x, b)/T(o_x)) + Q(o_x, b, x)\Big),} & \\ & (o_x \in C^{PA}) \end{cases} \quad (1)$$

where $q_n(o_x, a)$ is the normalized value for the rule (o_x, a), $T(o_x)$ is the temperature for the observation o_x, C^{PA} is the set of observations and observation sequences which are judged as perceptual aliasing, and C^A is the set of actions.

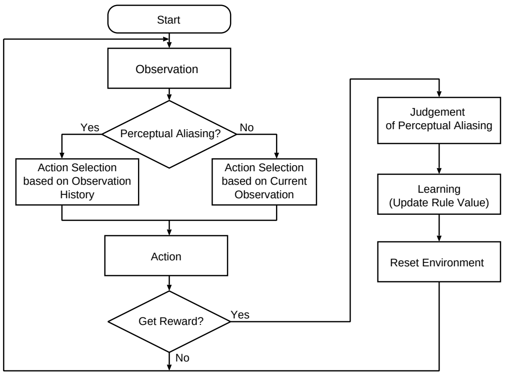
Start
Observation
Perceptual Aliasing?
Yes
No
Action Selection based on Observation History
Action Selection based on Current Observation
Action
Get Reward?
Yes
No
Judgement of Perceptual Aliasing
Learning (Update Rule Value)
Reset Environment

Figure 1: Flow of Profit Sharing that can Learn Deterministic Policy for POMDPs Environments [2].

And, $Q(o_x, a, x)$ is the summation of values for the rules on the observation at the time x (o_x) considering observation sequences, and it is given by

$$Q(o_x, a, x) = \sum_{o_i \to O \in C^{ref}(x)} \exp(q_n(o_i \to O, a)/T(O)) \tag{2}$$

where $o_i \to O$ is the observation sequence which includes the observation o_i before the observation sequence O, and $C^{ref}(x)$ is the set of observations and observation sequences which are used for the action selection at the time x.

2.2. Judgment of Perceptual Aliasing

(1) Judgment for Each Observation in Episode

In this learning method, whether each observation in perceptual aliasing is judged using the deterministic rate of actions in each observation and the progress of learning.

(a) Deterministic Rate of Actions in Each Observation

In the learning process, if the plural actions to be selected in order to obtain the reward in the same observation, the action is selected stochastically. In this learning method, the deterministic rate of actions in each observation which is used in the Extended On-line Profit Sharing with Judgment (EOPSwJ)[3] is used to detect perceptual aliasing.

The deterministic rate of actions in the observation at the time x (o_x), $d(o_x, x)$ ($0 \leq d(o_x, x) \leq 1$) is given by

$$d(o_x, x) = \sum_{a \in C^A} (P(o_x, a, x) - P_{ini})^2 / N \tag{3}$$

where C^A is the set of actions, $P(o_x, a, x)$ is the action selection probability for the action a in the observation at the time x (o_x), P_{ini} is the initial action selection probability, and N is the normalization constant.

(b) Progress of Learning

The progress of learning in the observation at the time x (o_x), $l(o_x, x)$ is given by

$$l(o_x, x) = \begin{cases} I(o_x), & (o_x \notin C^{PA}) \\ \min\{I(O) | O \in C^{ref}(x)\}, & (o_x \in C^{PA}) \end{cases} \tag{4}$$

where $I(o_x)$ is the update times of the the rule value for the observation o_x, $I(O)$ is the update times of the rule value for the observation sequence O. And $C^{ref}(x)$ is the set of observations and observation sequences which are used for the action selection at the time x in the last episode.

(c) Total Judgment

The observation whose progress of learning is high and deterministic rate of actions is low (that is, the action selection is stochastic) is judged as perceptual aliasing. That is, for the observation at the time x (o_x) in the last episode which satisfies the condition which is given by

$$\phi(l(o_x, x))(1 - d(o_x, x)) > \theta^{PA}, \tag{5}$$

the observation o_x is regarded as perceptual aliasing. Here, θ^{PA} is the threshold for judgment of perceptual aliasing. And $\phi(\cdot)$ is given by

$$\phi(u) = 1/(1 + \exp(-(u - \theta^l))) \tag{6}$$

where θ^l is the threshold.

(2) Update of Set of Observations and Observation Sequences Which Are Judged as Perceptual Aliasing C^{PA}

The set of observations and observation sequences which are judged as perceptual aliasing C^{PA} is updated based on the observations which are judged as perceptual aliasing.

(a) Decision of Observation Sequence Which Is Added to C^{PA}

If the condition given by Eq.(7) is satisfied, the observation sequence which is added to the set C^{PA} is determined.

$$o_x \in C^{PA} \cap C^{PA\_E} \tag{7}$$

Here, C^{PA_E} is the set of observations which are judged as perceptual aliasing in the last episode.

The observation sequence which is added to the set C^{PA} for the observation o_x, $O^{PA}(x)$ is given by

$$O^{PA}(x) = \underset{(o \to O) \in \{C^{ref}(x) \cap \overline{C^{PA}}\}}{\text{argmax}} \{d(o \to O)\} \tag{8}$$

where $C^{ref}(x) \cap \overline{C^{PA}}$ is the set union of the set of observation sequences which are considered in the action selection at the time x and the set of the observations which are not judged as perceptual aliasing. $d(o \to O)$ is the deterministic rate of actions for the observation sequence $o \to O$ and it is given by

$$d(o \to O) = \sum_{a \in C^A} (P(o \to O, a) - P_{ini})^2 / N \tag{9}$$

where P_{ini} is the action selection probability and N is the normalized constant. $P(o \to O, a)$ is the probability that the action a is selected for the observation sequence $o \to O$ and it is given by

$$P(o \to O, a) = \frac{\exp(q_n(o \to O, a)/T(O))}{\sum_{b \in C^A} \exp(q_n(o \to O, b)/T(O))} \tag{10}$$

where $q_n(o \to O, a)$ is the normalized rule value ($o \to O, a$) and $T(O)$ is the temperature in the observation sequence O.

(b) Update Set of Observations and Observation Sequences Which Are Judged as Perceptual Aliasing C^{PA}

Then, the observation sequences determined in *(a)* and the observations judged as perceptual aliasing are added to the set C^{PA}.

$$C^{PA} \leftarrow C^{PA} \cup C^{PA\_E} \cup \{O^{PA}(x) | x : o_x \in C^{PA} \cap C^{PA\_E}\} \tag{11}$$

2.3. Learning

When the agent obtains the reward, the rule values are updated after the judgment of perceptual aliasing.

(1) Update of Value of Rule $q(o, a)$

The update times of the rule value $I(o, a)$ and the rule value $q(o, a)$ are updated as

$$q(o,a) \quad \leftarrow \quad \left(1 - \frac{1}{I(o,a)}\right) q(o,a) + \frac{r \cdot F(o)}{I(o,a)} \qquad (12)$$
$$((o,a) \in \{(o_x, a_x) | x = 1, \cdots, W\})$$

where r is the reward, and $F(o)$ is the reinforcement value for the rules in the observation o and it is given by

$$F(o) = \frac{1}{l}(W - x_o) \qquad (13)$$

where W is the episode length, and x_o is first time when the observation o is observed.

(2) Update of Value of Rule $q(O, a)$

If it is considered that the observation sequence $O \rightarrow o_x (\in C^{ref}(x))$ is used for only action selection at the time x, the update time of the rule value $(O \rightarrow o_x, a_x)$ $(I(O \rightarrow o_x, a_x))$ and the rule value $q(O \rightarrow o_x, a_x)$ are updated as

$$q(O \rightarrow o_x, a_x) \quad \leftarrow \quad \left(1 - \frac{1}{I(O \rightarrow o_x, a_x)}\right) q(O \rightarrow o_x, a_x) + \frac{r \cdot F(o_x)}{I(O \rightarrow o_x, a_x)} \qquad (14)$$

where r is the reward. $F(o_x)$ is the reinforcement value for the observation sequence $(O \rightarrow o_x)$ whose last is the observation o_x.

If it is considered that the observation sequence $O \rightarrow o_x (\in C^{ref}(x))$ is used for action selection at plural times, the update time of the rule value $(O \rightarrow o_x, a_y)$ $(I(O \rightarrow o_x, a_y))$ and the rule value $q(O \rightarrow o_x, a_y)$ are updated as

$$q(O \rightarrow o_x, a_y) \quad \leftarrow \quad \left(1 - \frac{1}{I(O \rightarrow o_x, a_y)}\right) q(O \rightarrow o_x, a_y) + \frac{r \cdot F(o_x)}{I(O \rightarrow o_x, a_y)} \qquad (15)$$

where a_y is the action at the time y which satisfies

$$o_x = o_y \qquad (16)$$

and

$$\underset{x=1,\cdots,W}{\operatorname{argmin}}\{(O \rightarrow o_x) \in (C^{ref}(x) \cap C^{ref}(x')) \qquad | x \neq x'\} \leq y. \qquad (17)$$

If the action a_y appears plural times in the episode, the values of all rules for the action are updated one time.

If the observation sequence $O \rightarrow o_x$ appears in plural times in an episode, the values of all rules for the pair of the action and the observation o_x after the observation sequence $O \rightarrow o_x$ appears first in the episode are updated equally.

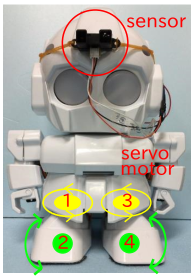

Figure 2: Bipedal Walking Robot

3. Motion Learning of Bipedal Walking Robot

3.1. Bipedal Walking Robot

Figure 2 shows the bipedal walking robot (Rapiro) which has 12 axes. In Fig.2, red circle shows PSD (Position Sensitive Detector) distance sensor. And, 1~4 show servo motors which are used to walk. The bipedal walking robot moves from the start to the goal in a maze, and learns the actions to move on shorter (shortest) route.

3.2. Observations

In this research, PSD distance sensor value is used as the observation. The used PSD sensor can be detected 20~150 cm. So, the quantized value (four states) is used as the observation.

3.3. Actions

In this research, the robot selects one of three actions (a_0 : Go Forward, a_1 : Turn Light, a_2 : Turn Right)

3.4. Reward

In this research, observation and action sequences from the start to the goal are treated as one episode. In each episode, the robot obtained the reward when the robot reaches the goal. The reward r is calculated based on the number of steps.

$$r = \begin{cases} -0.1t + 10 & (t < 100) \\ 0 & (\text{otherwise}) \end{cases} \qquad (18)$$

where t is the number of steps from the start to the goal. If the number of steps until the robot reaches the goal is small, the robot gains more rewards.

4. Experiment Results

Here, we examined motion learning (maze problem) of bipedal walking robot by the Profit Sharing that can learn deterministic policy for POMDPs environments.

4.1. Learning by Robot

Figure 3 shows the transition of steps per episode in the learning of the robot. In this experiment, maximum steps per episode was set to 100, and 100 steps means that the

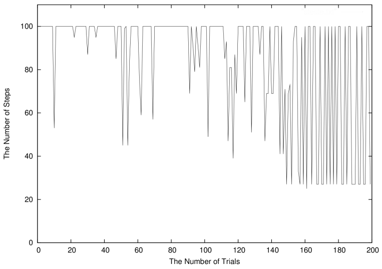

Figure 3: Transition of Steps in Learning of Robot.

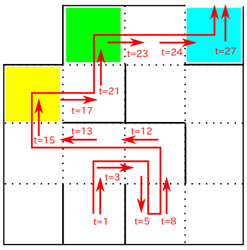

Figure 4: Route in 200th Trial (Robot).

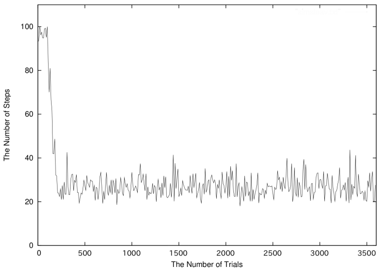

Figure 5: Transition of Steps in Learning (Simulation).

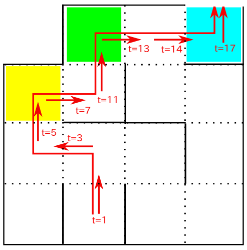

Figure 6: Route in 3600th Trial (Simulation).

robot could not reach the goal. Figure 4 shows the route in 200th trial. In the 200th trial, the robot reaches the goal in 27 steps.

4.2. Learning in Simulation

Here, the learning in the simulation in the same maze problem was carried out. Figure 5 show the transition of steps per episode (average steps per 10 trials) in the learning. Figure 6 shows the route in 3600th trial. In the 3600th trial, the robot reaches the goal in 17 steps.

4.3. Learning by Robot using Result in Simulation

Here, the learning by robot using the result in the simulation in the same maze problem was carried out. Figure 7 show the transition of steps per episode in the learning. Figure 8 shows the route in 39th trial.

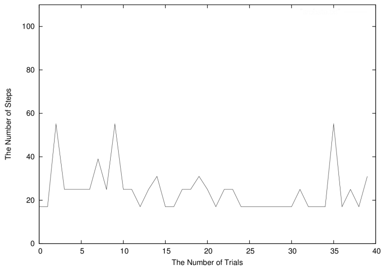

Figure 7: Transition of Steps in Learning of Robot using Result in Simulation.

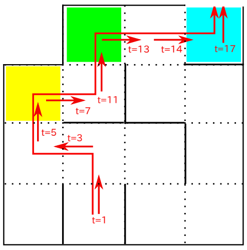

Figure 8: Route in 39th Trial (Robot after Simulation).

5. Conclusions

In this paper, motion learning (maze problem) of bipedal walking robot has been realized by the Profit Sharing that can learn deterministic policy for POMDPs environments[2]. We carried out a series of experiments using bipedal walking robot, and confirmed that the robot can learn the route from the start to the goal by the Profit Sharing that can learn deterministic policy for POMDPs environments.

References

[1] R. S. Sutton and A. G. Barto : Reinforcement Learning : An Introduction, The MIT Press, 1998.

[2] Y. Takamori and Y. Osana : "Profit sharing that can learn deterministic policy for POMDPs environments," Proc. of IEEE SMC, Anchorage, 2011.

[3] K. Saito and S. Masuda : "Profit Sharing Introducing the Judgement of Incomplete Perception," Trans. of the JSAI, Vol.19, No.5, pp.379–388, 2004.

[4] T. Suzuki and Y. Osana : "Fall avoidance of bipedal walking robot by profit sharing that can learn deterministic policy for POMDPs environments," Proc. of NaBIC, Porto, 2014.

[5] Y. Morino and Y. Osana : "Walking motion learning of quadrupedal walking robot by profit sharing that can learn deterministic policy for POMDPs environments," Proc. of SEAL, Dunedin, 2014.