Online Learning of Shaping Reward with Subgoal Knowledge

Takato Okudo*1, Seiji Yamada*2*1
*1 The Graduate University for Advanced Studies(SOKENDAI), *2 National Institute of Informatics(NII)

Introduction

☞ To accelerate learning in Reinforcement Learning
☞ SARSA-RS needs aggregation of states to an abstract state.

? Aggregation of the SARSA-RS is often very difficult because the designer access to all the state is required.

✓ Our method accesses to only subgoals.

Background

Potential-based Reward Shaping

A reward transformation using the potential-based reward shaping remains the optimal policy. Formally,

$$F(s,s') = \gamma\Phi(s') - \Phi(s)$$

An agent update its policy with a environmental reward r and shaping reward F.

SARSA-RS

SARSA-RS used a state value V as the potential Φ. The state value was defined over an abstract state Z.

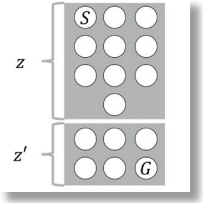
$$F(z,z') = \gamma \mathrm{V}(z') - \mathrm{V}(z)$$

The aggregation $S \mapsto Z$ was required.

Mehod

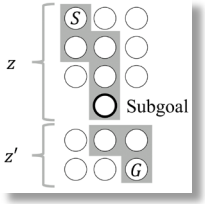
■ Traditional Aggregation

The aggregation explicitly maps all the states S into abstract states Z.
In particular, this aggregation is difficult for a designer to be provided in an environment with continuous states such as pinball domain.



■ Dynamic Trajectory Aggregation

This method aggregates the visited states until a subgoal achievement, and shaping them by a value over an abstract state Z_i.
This aggregation is generated from only subgoal.



Dynamic trajectory aggregation makes it easy to expand applications of SARSA-RS.

Experiment

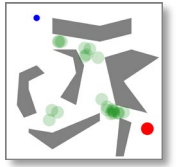
User study
We recruited 10 participants.
Two subgoals are provided by each participant and the total number of subgoals is 20.

Evaluation
The navigation task in pinball, which has a reward on the goal. A state is continuous and an action is discrete.

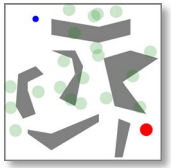
Result

HRS: Our method with participants' subgoals
RRS: Our method with random subgoals
AC: Actor-critic as basis RL method.
NRS: Naïve reward shaping which generate positive potential only when subgoal achievement.

HRS outperforms the other four methods.



Participants'



Random

* Two subgoals are used in each learning.