Guiding Robot Exploration in Reinforcement Learning via Automated Planning

Paper accepted to ICAPS 2021

Yohei Hayamizu[†*], Saeid Amiri[‡], Kishan Chandan[‡], Keiki Takadama[†], and Shiqi Zhang [‡]
†: University of Electro-Communications, ‡: SUNY Binghamton, *Work done while visiting SUNY Binghamton
Contact E-mail: hayamizu@cas.lab.uec.ac.jp

Background

Robots in human-inhabited environments are able to conduct various service and interaction tasks.

Requirements for these tasks

- Fulfilling requests from humans such as navigation and delivery
- Learning efficiency in the real world

A Segway-based mobile robot platform used for various task execution

Reserch fields for these tasks

Model-Based Reinforcement Learning (RL) and Automated Planning have been used to meet these two requirements.

- **Model-Based RL**

Learn a world model while learning an action policy to achieve long-term goals from both of real and simulated experiences

- **Automated Planning**

Reason with declarative domain knowledge, including commonsense knowledge, that is provided a priori

Aim

Efficient task learning to fulfill diverse service requests

Main Contributions

Efficient exploration strategy for RL agents in navigation domains
Avoid less-relevant states by reasoning with contextual knowledge while using trial-and-error experiences

Exploiting complementary features of model-based RL and automated planning
Aim at improving sample-efficiency in a real robot domain

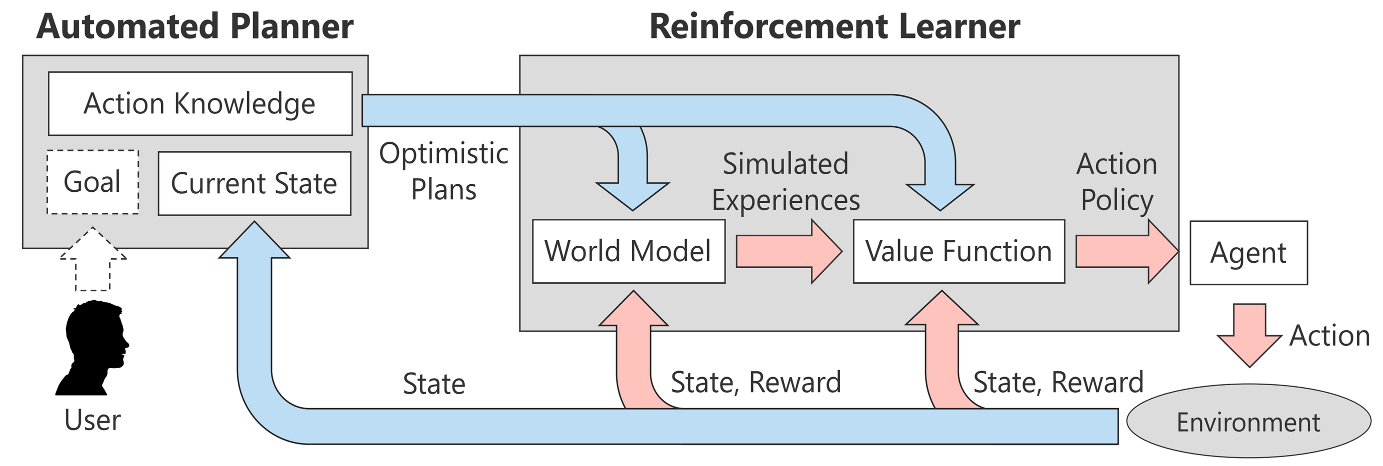
Guided Dyna-Q (GDQ):

Bridging the gap between model-based RL and automated planning

GDQ integrates the two sub-procedures for **optimistic initialization** and repeatedly conducting runtime **policy update**.

1. **Optimistic Initialization**
   Help the agent avoid the near-random exploration behaviors through a "warm start" enabled by our automated planner

2. **Policy Update**
   Guide the agent to only try the actions that can potentially lead to the goal states



An overview of Guided Dyna-Q (GDQ)

Red-color loop: Standard control loop of Dyna-Q
Blue-color loop: Incorporation of an automated planner into the learning process

Experiments using Navigation Tasks

Hypotheses about GDQ:

1) Perform better than existing RL methods from the literature in cumulative reward.
2) Help the robot avoid visiting "irrelevant" areas. (A navigation task are achieved via relevant areas.)
3) Is more robust to goal changes

Experiment Settings

Indoor office environment settings:

- All states are categorized into 7 areas.
- There 6 doors that a robot can use to enter rooms.

$Easy \leftarrow D3 < D1 < D4 < D5 < D0 < D2 \rightarrow Difficult$

An Occupancy-grid map of the experiment domain (Indoor office environment)

Action sets: 4 types of actions for navigational purposes
$\{goto, gothrough, approach, opendoor\}$

Action knowledge (designed by a human expert)
Answer Set Programming (ASP) [Lifschitz, 2002]:

- Performing well in knowledge intensive domain
- Computing plans with reasoning paradigm

Example of action knowledge representation by ASP
$at(Z, I+1) :- gothrough(Y, I), at(X, I), acc(X, Y, Z), I < n.$
$hasdoor(s1, d0).\ acc(s0, s1).$
$:- approach(D, I), facing(D, I), door(D), I = 0..n-1.$

Results

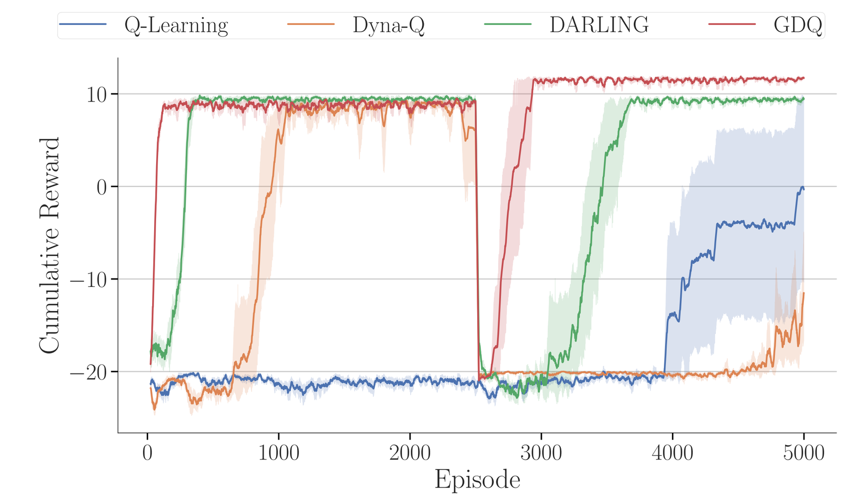
A) Simulation Experiment

Tasks:

- Task 1: $(P1 \rightarrow P2 \rightarrow P3)$

Baseline Methods:

- Q-Learning [Sutton, 2018]
- Dyna-Q [Sutton, 2018]
- DARLING [Leonetti, 2016]



Task 1: $P1 \rightarrow P2 \rightarrow P3$

Average cumulative rewards over ten runs: GDQ learned the practical policy faster and is more robust to task changes.

B) Real Robot Experiment

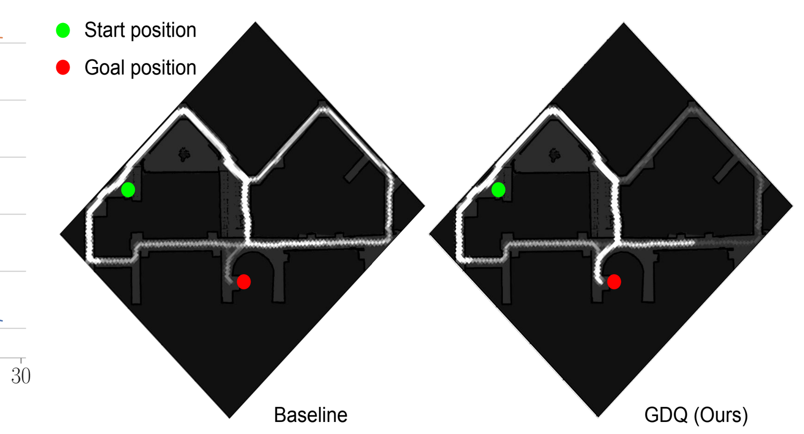
Task:

- Task 2: $(P4 \rightarrow P2)$

Baseline Method:

- Dyna-Q

Cumulative rewards on a real robot. GDQ enabled robot to find the practical path in 27 trials.



Heatmaps of our office domain for visualizing where the robot visited using the Dyna-Q Baseline (Left) and GDQ (Right).

Conclusion

Aim

- Efficient task learning to fulfill diverse service requests

Approach

- **Guided Dyna-Q**: Optimistic Initialization & Policy Update

Results

- GDQ improves the quality of a learned policy
- GDQ reduces the effort in exploration

References

Sutton, R. S.; and A. G. Barto (2018). "Reinforcement learning: An introduction." MIT Press.

Lifschitz, V. (2002). "Answer set programming and plan generation." In: Artificial Intelligence 138(1-2), pp. 39–54.

Leonetti, M.; Iocchi, L.; and Stone, P. (2016). "A synthesis of automated planning and reinforcement learning for efficient." In: Artificial Intelligence 241(Dec), pp. 103–130.

Acknowledgement

A portion of this work has taken place in the Autonomous Intelligent Robotics (AIR) Group at SUNY Binghamton. AIR research is supported in part by grants from the National Science Foundation (IIS-1925044 and REU Supplement), Ford Motor Company (two URP Awards), OPPO (Faculty Research Award), and SUNY Research Foundation.