Supplementary Materials: Rule-Based Reinforcement Learning for Efficient Robot Navigation with Space Reduction

Yuanyang Zhu, Zhi Wang, Chunlin Chen, Daoyi Dong

Appendix A: Supplementary experimental results from the final path in Env 1, 2 and 3.

This appendix shows the full final planned paths of A*, ACO and RuRL methods, which are listed in Figs. S1(a)-(c).

As shown in Fig. S1(a), A*, ACO and RuRL methods can find the optimal path with 23 steps in Env 1. When executing the optimal path, A* and ACO algorithms need to switch directions for 11 and 13 times, respectively, while RuRL only requires 3 times. Frequently switching directions will slow down the speed of mobile robots with more consumed energy and may be critical of the performance of the robot motion kinematics [46].

From Fig. S1(b), we find that A*, ACO and RuRL can obtain the optimal path with 23 steps. Furthermore, the planned path of A*, ACO and RuRL methods need to switch directions for 11, 9 and 9 times, respectively.

In the multi-room environment, the A* and ACO approaches obtain sub-optimal paths with 73 and 74 steps and need to switch directions for 29 and 27 times, respectively, as shown in Fig. S1-(c). In contrast, RuRL requires only 14 times of switching directions while learning the optimal path with 72 steps, where RuRL is implemented by the Q-learning algorithm with ε-greedy exploration strategy. Compared with A* and ACO classic algorithms, we can find that RuRL is capable of finding smoother paths with a better optimality guarantee.

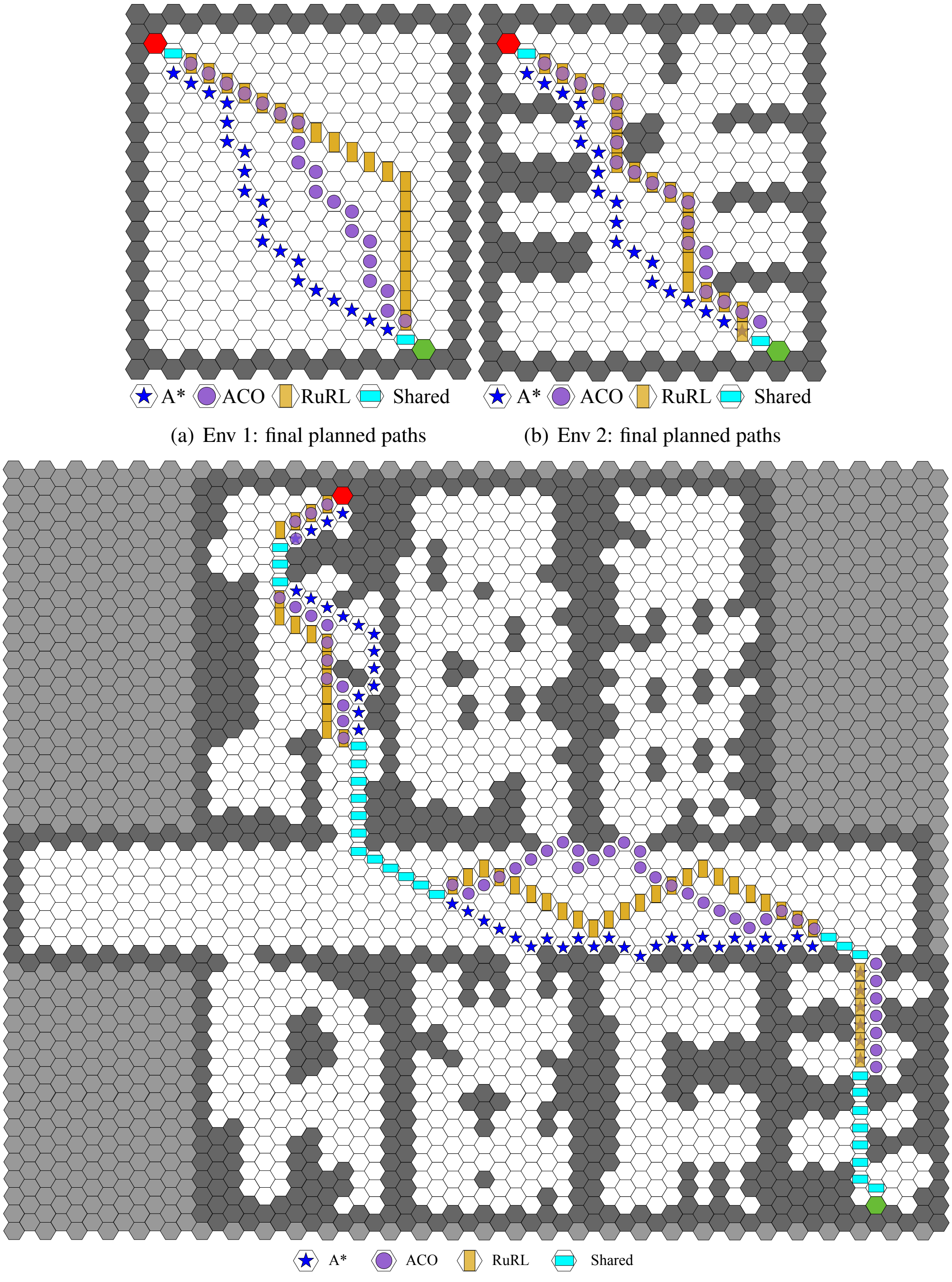
Appendix B: Supplementary experimental results with different optimization step K.

This appendix shows the improved learning performance with the optimization step K increasing.

From Table SI, as the optimization step K increases, the learning performance is improved slightly. On the other hand, a larger optimization step will lead to large computational cost. Generally, a moderate optimization step (e.g., $K = 2$) is enough for an appropriate trade-off between the performance improvement and computational cost. As shown in Fig. S2, when the Pledge rule is withdrawn, the exploratory strategy requires more exploration steps to ensure optimal path and the number of learning steps will increase suddenly, while the steps of RuRL are still much smaller than those of the other methods. It indicates that the rule can effectively guide the early exploration strategy in the reduced space.

Table. S I: Numerical results in terms of total learning steps of all tested algorithms in multi-room tasks.

	Env 3				
Implementation algorithm	RL without Rules ($\times 10^7$)	RuRL ($K = 1$) ($\times 10^6$)	RuRL ($K = 2$) ($\times 10^6$)	RuRL ($K = 3$) ($\times 10^6$)	Maximal reduction (%)
Q-learning (ε-greedy)	1.69	5.18	4.96	4.77	**71.79%**
Q-learning (Softmax)	2.36	7.86	7.29	7.08	**69.93%**
SARSA (ε-greedy)	1.23	5.12	4.85	4.72	**61.69%**
SARSA (Softmax)	2.78	8.45	7.89	7.58	**72.78%**

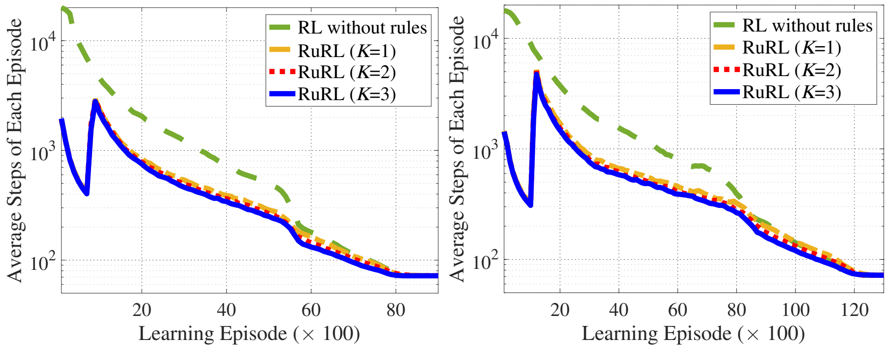


(a) Env 1: final planned paths

(b) Env 2: final planned paths

(c) Env 3: final planned paths

Fig. S 1: The paths planned by A*, ACO and RuRL are denoted with blue stars, purple circles and yellow vertical bars, respectively. The grids marked by the dark blue horizontal bars represent the shared part of their paths.



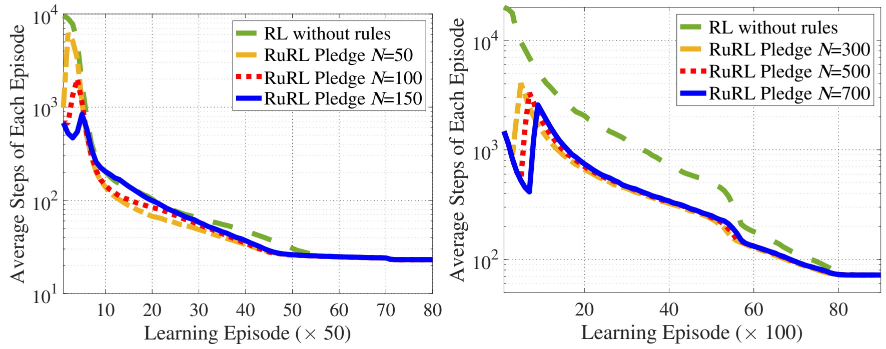
(a) Q-learning with ε-greedy (b) SARSA with Softmax

Fig. S 2: The average steps per episode of RL and RuRL with $K = 1, 2, 3$ in multi-room navigation tasks.

Appendix C: Supplementary experimental results with different numbers of episodes N with the Pledge rule.

This appendix shows the full results of the performance of RuRL with different numbers of episodes N with the Pledge rule, which are listed in Fig. S3 and Table SII, respectively.

In addition, we analyze the relationship between the number of episodes N with the Pledge rule and the performance of RuRL. We set different values of N with optimization step $K = 3$ for tasks in the single-room environment with obstacles and the multi-room environment, respectively. The learning curves and corresponding numerical results are presented in Fig. S3 and Table SII, respectively. It can be observed that the learning steps decrease as the Pledge rule guides the early exploration strategy for more episodes. A potential drawback is that a too large N might lead to sub-optimal policies. Generally, using the Pledge rule for an appropriate number of episodes can accelerate the navigation process at the beginning of the learning process, without influencing the final optimal policies.



(a) Env 2 with different N (b) Env 3 with different N

Fig. S 3: The performance of RuRL using different N episodes with the Pledge rule.

Table. S II: Numerical results in terms of total learning steps of RL and RuRL, which are implemented by Q-learning with different N episodes using the Pledge rule under the ε-greedy exploration strategy in Env 2 and Env 3.

Algorithm Name	Env 2					Env 3				
	RL without Rules ($\times 10^6$)	RuRL $N = 100$ ($\times 10^6$)	RuRL $N = 150$ ($\times 10^5$)	RuRL $N = 200$ ($\times 10^5$)	Maximal Reduction (%)	RL without Rules ($\times 10^7$)	RuRL $N = 100$ ($\times 10^6$)	RuRL $N = 300$ ($\times 10^6$)	RuRL $N = 500$ ($\times 10^6$)	Maximal Reduction (%)
Q-learning(ε-greedy)	1.91	1.10	7.14	5.38	**71.78%**	1.69	5.48	5.13	4.77	**71.79%**

References

[46] A. Ravankar, A. A. Ravankar, Y. Kobayashi, Y. Hoshino, and C.-C. Peng, “Path smoothing techniques in robot navigation: State-of-the-art, current and future challenges,” *Sensors*, vol. 18, no. 9, p. 3170, 2018.