Multi-agent Hierarchcial Reinforcement Learning with Dynamic Termination

Dongge Han, Wendelin Boehmer, Michael Wooldridge, Alex Rogers

Department of Computer Science, University of Oxford

Introduction

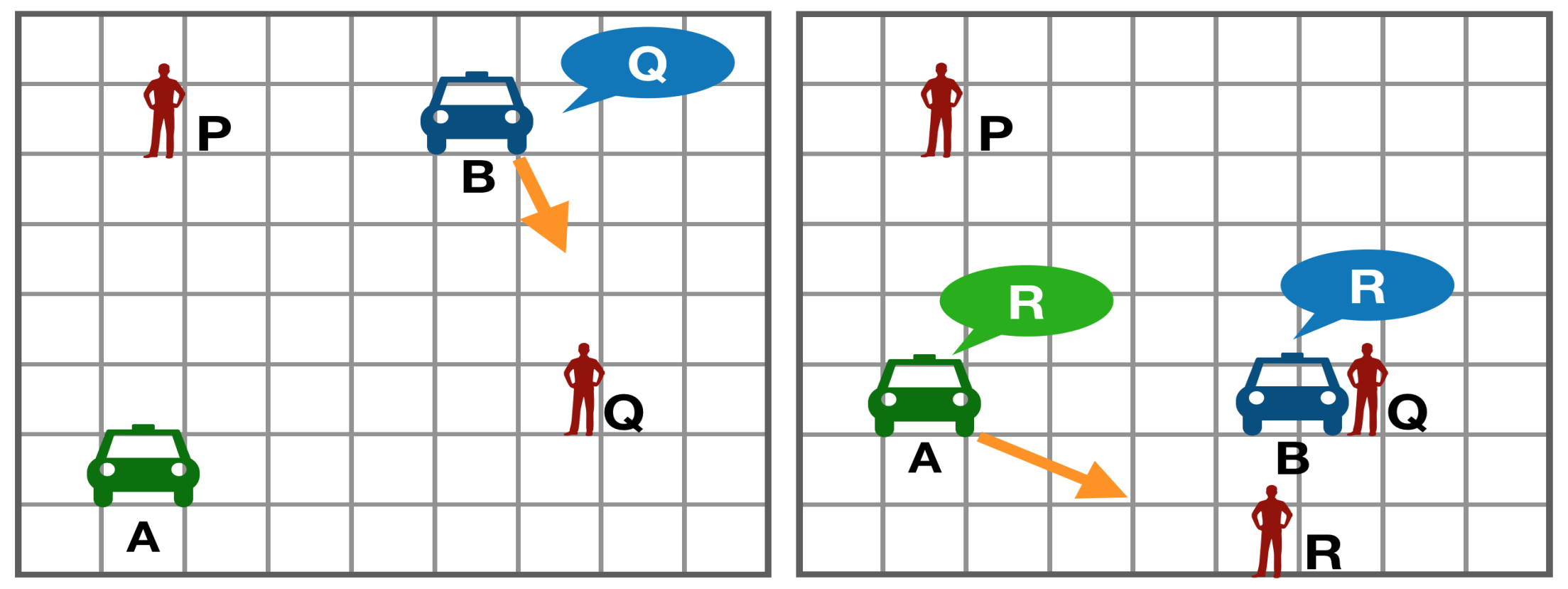
In a multi-agent system, an agent's optimal policy will typically depend on the policies of other agents. The options framework allows agents in a multi-agent system to take into account others' behaviours when making their own decisions, by broadcasting their current options. We identified the flexibility vs. predictability dilemma in static option termination schemes, and propose dynamic termination of options to achieve both flexibility towards changes and predictability which improves coordination.

Multi-agent Options Framework

Figure 2 shows the Multi-agent Options Framework.

- Decentralized: each agent keeps its own Q-values
- Each agent has two levels: *Manager* and *Worker*.
- Manager **(SMDP Policy)** chooses *options* o_t, eg. chop a tomato, and when complete, it chooses a new option.
- Worker **(Option Policy)** chooses an action at each step to complete the option chosen by manager.
- Agents **broadcast** their current options, and condition on others' options when choosing a new option.

Flexibility vs. Predictability



(a) Taxi A choosing a target (b) Taxi B switching target

Figure 1: Taxi Scenario Examples

Option broadcast benefits coordination: In Figure 1(a), *B*'s option broadcast helps *A* to choose its target as passenger *P*.

- **Flexibility**: Agent can terminate its current option early, and switch to a better option. In Figure 1(b), after *B* pickuped *Q* and heads for *R*, *A* can be more flexible if it immediately stops going towards *R* and switch target.
- **Predictability**: In a multi-agent system, if agent changes options frequently, its behaviour becomes unpredictable, as its broadcast option does not represent its real behaviours.

Dynamic Termination

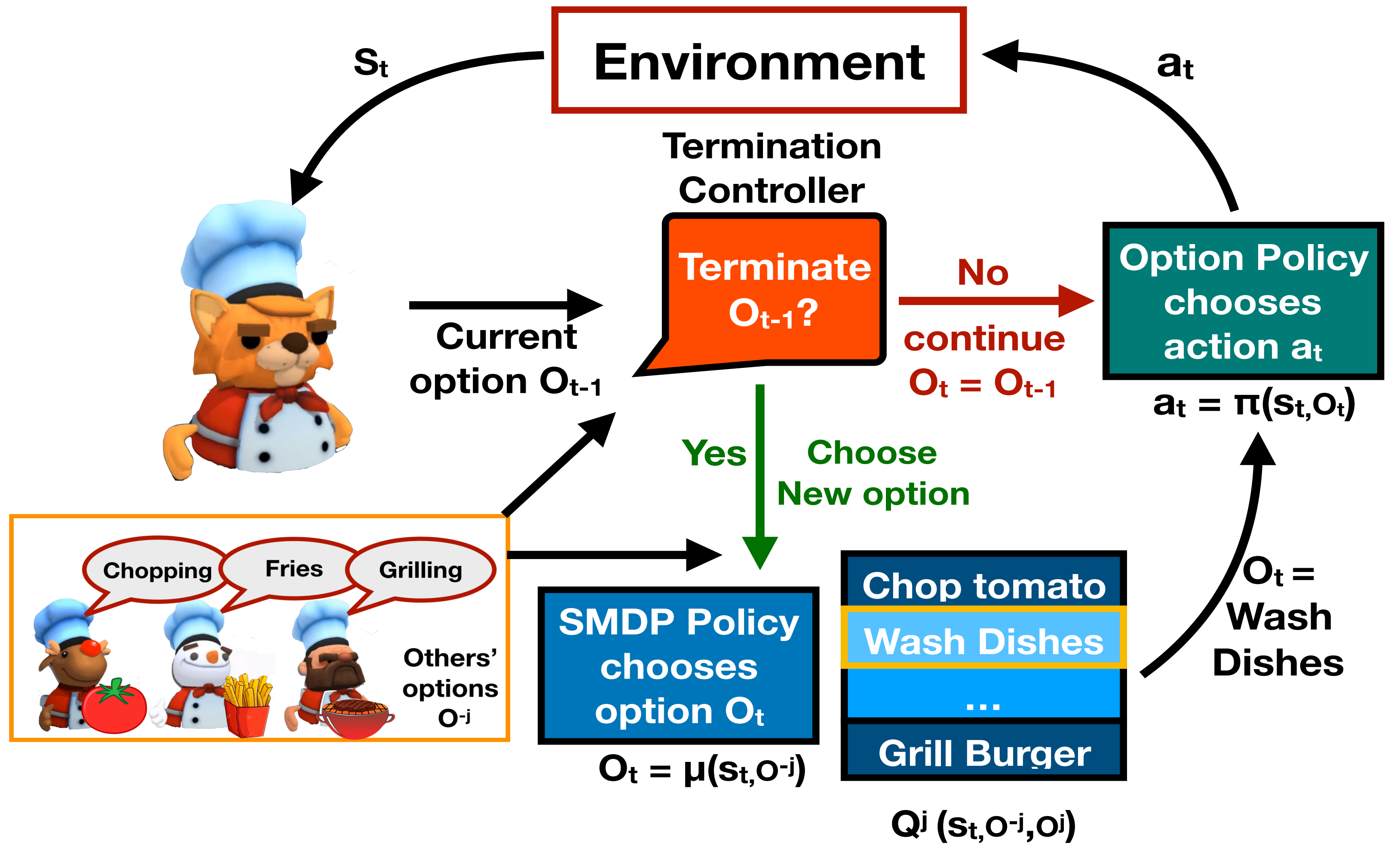


Figure 2: Multi-agent Options Framework

Termination Controller was previously a static condition, eg. terminate when option complete, terminate after each step, etc. We introduce **dynamic termination**, which learns to choose whether to terminate the current option.

Model Architecture

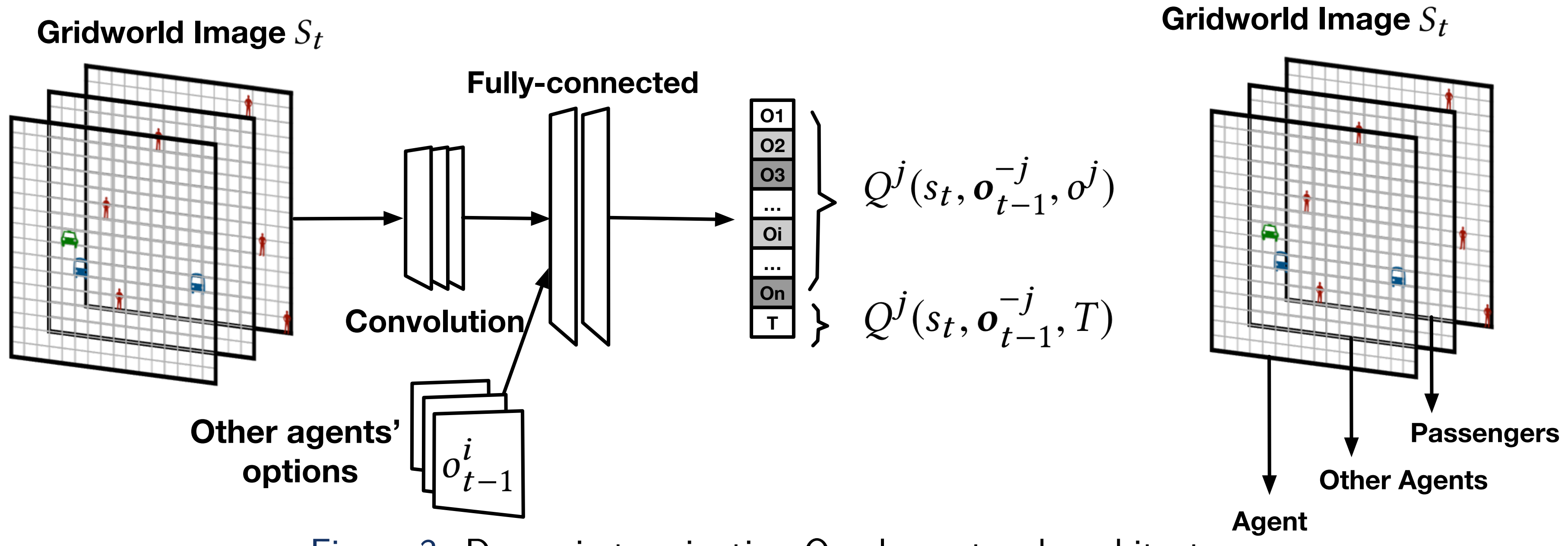


Figure 3: Dynamic termination Q-value network architecture.

We model dynamic termination **as an option**, by introducing a *Termination Option* T besides all other options. At each step, choose whether to terminate by comparing the value of the previous option $Q^j(s_t,\vec{o}^{\,-j}_{t-1},o^j_{t-1})$ with the value of termination $Q^j(s_t,\vec{o}^{\,-j}_{t-1},T)$. Therefore the novel Bellman equation is:

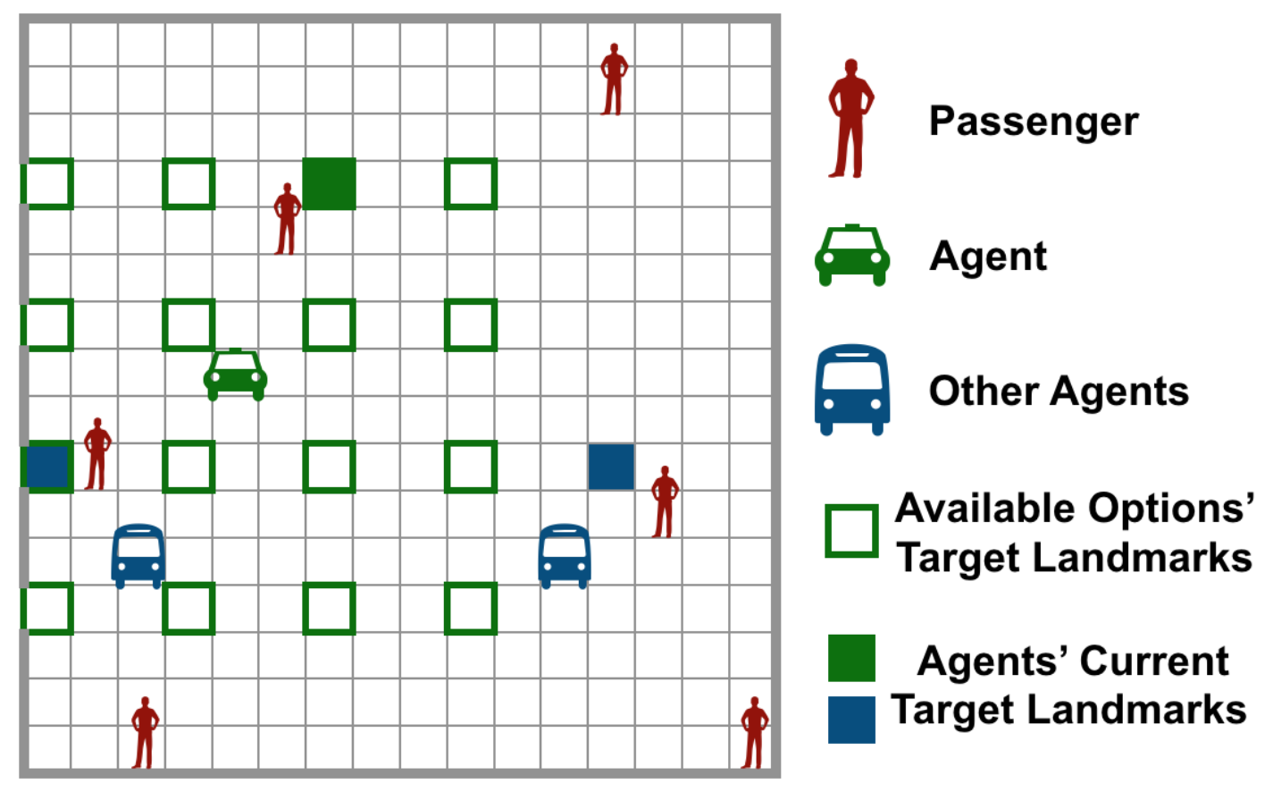
$$Q^j(s_t,\vec{o}^{\,-j}_{t-1},o^j \neq T) := \mathbb{E}\left[r_t+\gamma \max_{o'^j\in\{o^j,T\}} Q^j(s_{t+1},\vec{o}^{\,-j}_t,o'^j)\right]$$

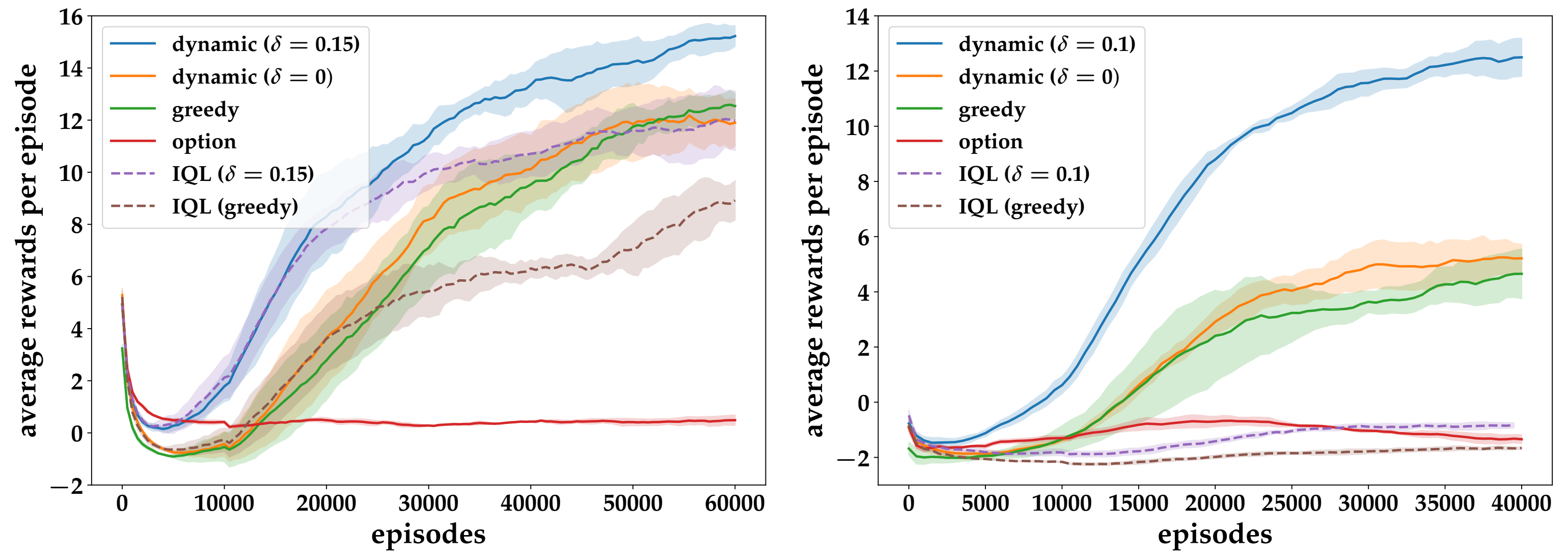
$$Q^j(s_t,\vec{o}^{\,-j}_{t-1},o^j = T) := \max_{o'^j\in\mathcal{O}^j} Q^j(s_t,\vec{o}^{\,-j}_{t-1},o'^j) - \delta$$

The Q-value of T means switching to an optimal option after termination, with a small penalty δ. The values are learned off-policy and via Intra-option Learning.

Experiments

We test our framework on multi-agent taxi pickup and pursuit. The equidistant landmarks are destinations of options currently visible to the agent, who may switch between options to reach a non-landmark location.


Passenger
Agent
Other Agents
Available Options' Target Landmarks
Agents' Current Target Landmarks



(a) 19 × 19 taxi with 10 agents (b) 19 × 19 pursuit with 10 agents

Figure 4: Results from Taxi Pickup and Pursuit Tasks

We compare our dynamic termination agent (dynamic) with three types baselines using multi-agent options framework: greedy (terminate each step, trained off-policy with improved exploration via a behavioural policy), option (terminate when option is complete) and independent Q learning (IQL, where option broadcasts are disabled).

		taxi			2 agent pursuit			3 agent pursuit	
		n=5, m=10		n=10, m=20	n=3, m=5		n=10, m=10	n=3, m=5	
Agents		*19x19*	*25x25*	*19x19*	*16x16 (r=1)*	*19x19 (r=1)*	*19x19 (r=1)*	*10x10 (r=1)*	*16x16 (r=2)*
Dynamic Termination	$\delta = 0.1$	**7.89**	**5.75**	**15.29**	**10.24**	**9.30**	**12.50**	**6.71**	**10.38**
	$\delta = 0$	6.58	3.28	11.81	6.73	4.07	5.38	5.53	6.54
Greedy Termination		6.62	3.23	12.39	7.36	3.74	4.65	5.89	6.40
Option Termination		-0.32	-0.94	0.52	5.47	-1.82	-1.42	-3.77	5.20
IQL	$\delta = 0.1$	7.11	5.09	12.02	-1.57	-2.29	-0.84	-1.62	-0.59
	greedy	6.08	2.79	9.06	-2.12	-2.49	-1.64	-2.13	-0.42

Table 1: Average reward after training for Taxi and Pursuit. n is the number of agents and m is the number of passengers (preys). NxN denotes grid-world size, k agent pursuit denotes the required number of agents for capture, and r is the capture range.

Contact Information

- Email: dongge.han@cs.ox.ac.uk
- Website: https://www.cs.ox.ac.uk/people/dongge.han/