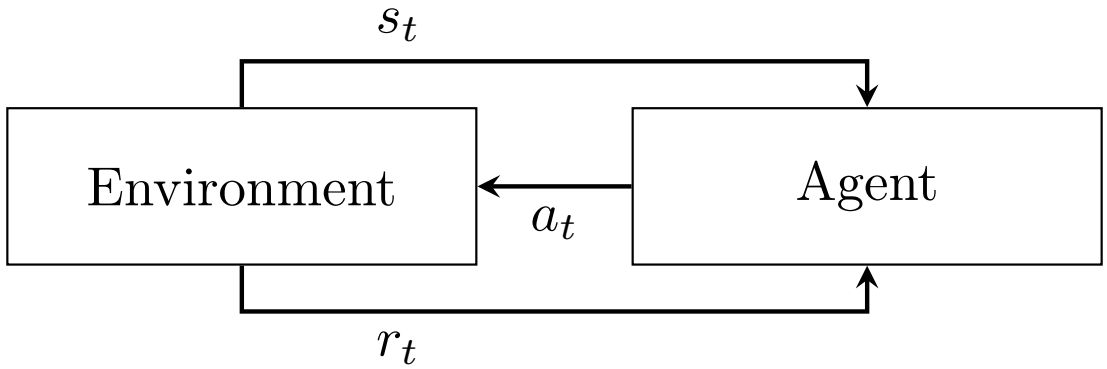
Guidelines for Action Space Definition in Reinforcement Learning-based Traffic Signal Control Systems

Maxime Tréca, Julian Garbiso, Dominique Barth

October 15, 2020

Outline

I - Basics of Reinforcement Learning



Reinforcement Learning methods aim at learning from the feeback of the state-action-reward loop:

- By testing all possible state / action combinations
- By storing the resulting rewards of these combinations
- By establishing a *policy* using this stored data

I - Q-Learning

Q-learning is a Reinforcement Learning algorithm developed by Watkins [2].

- The agent records and updates an estimation of the payoff of each state/action pair it encounters in a Q-table.

	a_1	$\dots$	a_n
s_1	v_{s_1,a_1}	$\dots$	v_{s_1,a_n}
$\dots$	$\dots$	$\dots$	$\dots$
s_m	v_{s_m,a_1}	$\dots$	v_{s_m,a_n}

- An iterative formula is used to update this estimation:

$$Q(s_t, a_t) \leftarrow (1 - \alpha_t)Q(s_t, a_t) + \alpha_t(r_t + \gamma \max Q(s_{t+1}, a_t))$$

I - Reinforcement Learning applied to Traffic Signal Control

Reinforcement Learning (RL) algorithms have been applied to Traffic Signal Control (TSC) since the early 2000s:

- Wiering [3]: first use of Q-learning at the intersection level to decrease vehicle waiting time on a road network.
- El-Tantawy [1]: `MARLIN` algorithm, which coordinates multiple RL-based agents using real traffic data.

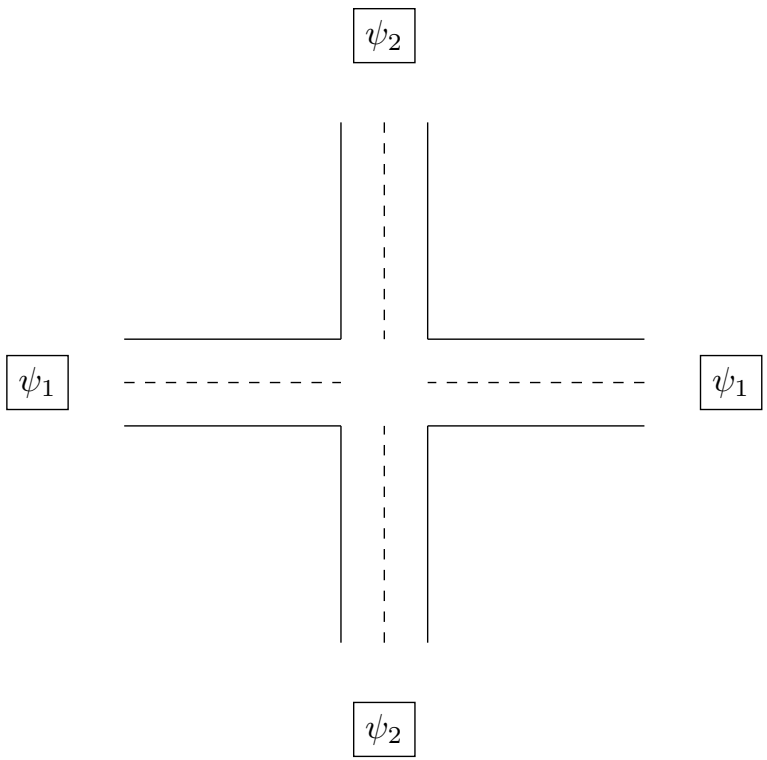
I - Problem Statement

In the paper cited above, agent actions are either:

- *Phase-based*: the agent sets the entire length of the green phase.
- *Step-based*: the agent decides whether to extend the current phase every k steps.

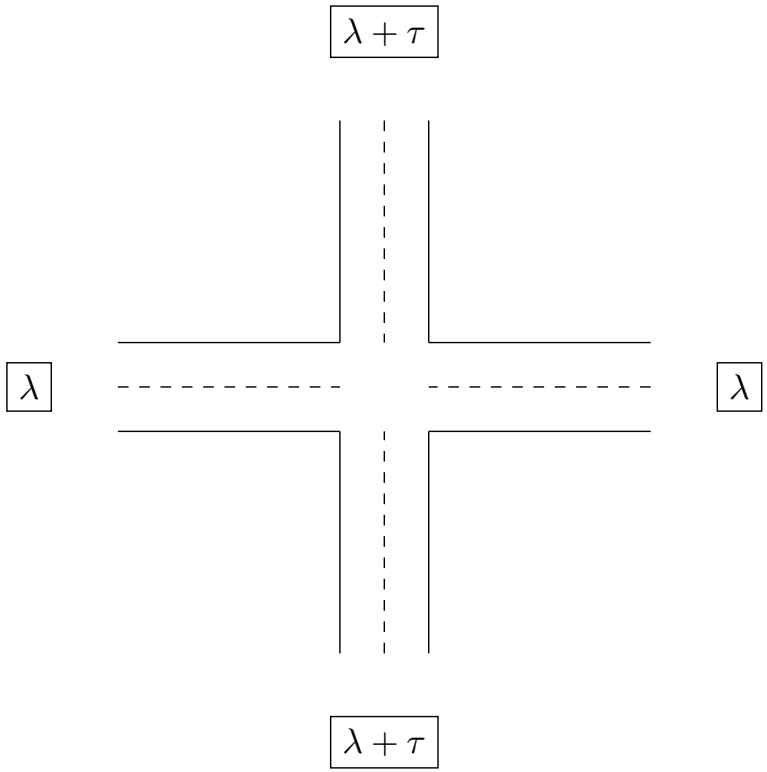
$\rightarrow$ No action space definition comparison in the litterature.

II - Experimental Framework



- We consider a single intersection
- State : $\langle \psi_i, d_i, n_1, n_2 \rangle$
- Action is either:
  - The length of ψ_i
  - Extend ψ_i by k steps
- Reward: $\sum \omega_{t+a} - \sum \omega t$

II - Traffic Generation



- Two Poisson processes
- $\lambda + \tau = 0.5$
- τ measures the un-evenness of traffic

II - Simulation Settings

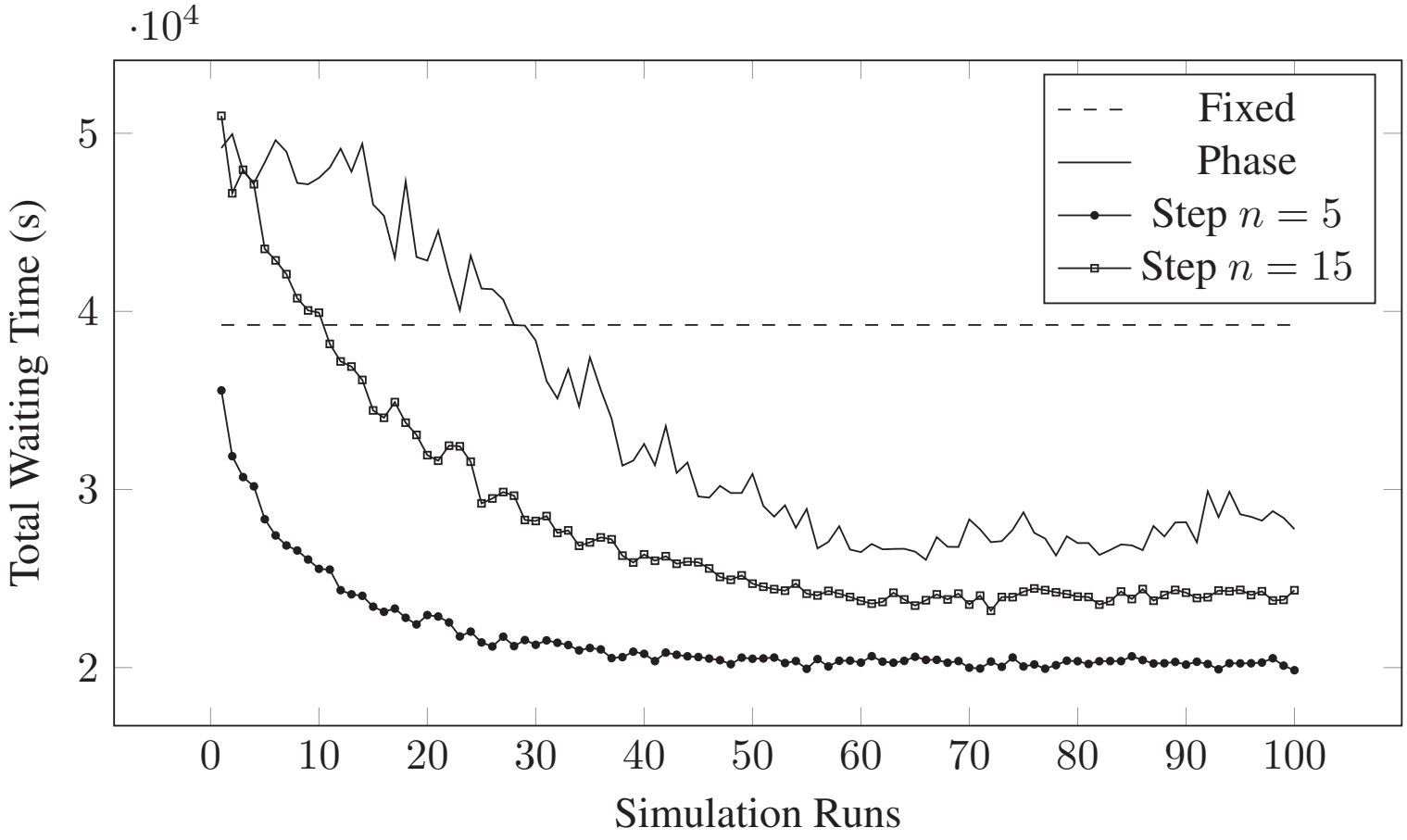
- SUMO microscopic traffic simulator.
- 100 successive iterations of 10 000 steps each
- We mesure total vehicular delay over each iteration.
- Results normalized over 50 distinct runs.

III - Guideline #1 - Step-based v. Phase-based Methods

τ	Fixed	Phase	Step (Best)	Step (Worst)
0.0	3.617	2.672	2.053	2.473
0.1	4.070	2.746	1.956	2.595
0.2	4.603	3.070	1.977	2.570
0.3	7.773	4.582	2.032	2.531
0.4	6.807	5.773	2.088	2.216
0.5	18.329	3.240	1.994	2.473

Table: Average vehicle waiting time after convergence per agent type and traffic parameter τ (in 10^3 seconds).

III - Guideline #1 - Step-based v. Phase-based Methods



Guideline #1 Step-based methods are always preferable to phase-based ones.

III - Guideline #2 - Decision Interval Length

τ	$k = 1$	$k = 5$	$k = 10$	$k = 15$	$k = 20$
0.0	0	4.86	10.29	22.37	27.81
0.1	0	4.17	7.20	23.96	30.15
0.2	0	0.45	5.49	22.68	31.03
0.3	3.04	0	4.38	11.02	26.39
0.4	9.53	0	2.74	11.09	7.84
0.5	22.12	0.56	0	13.60	3.33

Table: Percentage difference with respect to the optimum average vehicle waiting time (marked as 0) for step-based methods by action interval value k and traffic scenario τ

Guideline #2 Very short intervals between decision points are preferable for uniform traffic while slightly longer intervals are preferable for skewed traffic demand.

III - Optimal Interval Length

τ	$k = 1$	$k = 5$	$k = 10$	$k = 15$	$k = 20$
0.0	0	4.86	10.29	22.37	27.81
0.1	0	4.17	7.20	23.96	30.15
0.2	0	0.45	5.49	22.68	31.03
0.3	3.04	0	4.38	11.02	26.39
0.4	9.53	0	2.74	11.09	7.84
0.5	22.12	0.56	0	13.60	3.33

Table: Percentage difference with respect to the optimum average vehicle waiting time (marked as 0) for step-based methods by action interval value k and traffic scenario τ

Guideline #3 Defining longer intervals between successive decision points (from 5 to 10 seconds) yields satisfactory to optimal results for step-based agents.

V - Conclusion

Issue: No in-depth comparison of step-based and phase-based action spaces fo RL-TSC.
Conclusions:

- Step-based is always preferable
- Shorter action interval for uniform traffic demand
- Optimal and realistic step interval between 5 and 10 seconds.

V - Conclusion

- Results only on a simple 4-street intersection
- However, guidelines validated on a NEMA-type intersection.

References I


Samah El-Tantawy, Baher Abdulhai, and Hossam Abdelgawad.
Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (marlin-atsc): methodology and large-scale application on downtown toronto.
IEEE Transactions on Intelligent Transportation Systems, 14(3):1140–1150, 2013.


Christopher JCH Watkins and Peter Dayan.
Q-learning.
Machine learning, 8(3-4):279–292, 1992.


MA Wiering.
Multi-agent reinforcement learning for traffic light control.
In *Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000)*, pages 1151–1158, 2000.