TECHNISCHE UNIVERSITÄT DRESDEN

ComNets
Deutsche Telekom Chair of Communication Networks

Johannes Valentin Stanislaus Busch
Deutsche Telekom Chair of Communication Networks
TU Dresden

Deep Reinforcement Learning for Traffic Control

Defense of Diploma Thesis

Monday, 02/09/2019

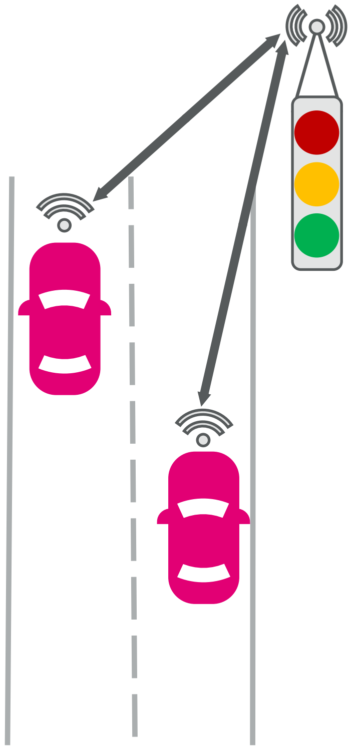
Contents

Traffic Congestion

- Costs of congestion in the EU amount to 1 % of its GDP [1]
- In large cities, commuters spend up 200 hours yearly in congested traffic [2]
- Average driving velocities in major cities go as low as 11 km/h [2]

Vehicle to Infrastructure (V2I) Communication

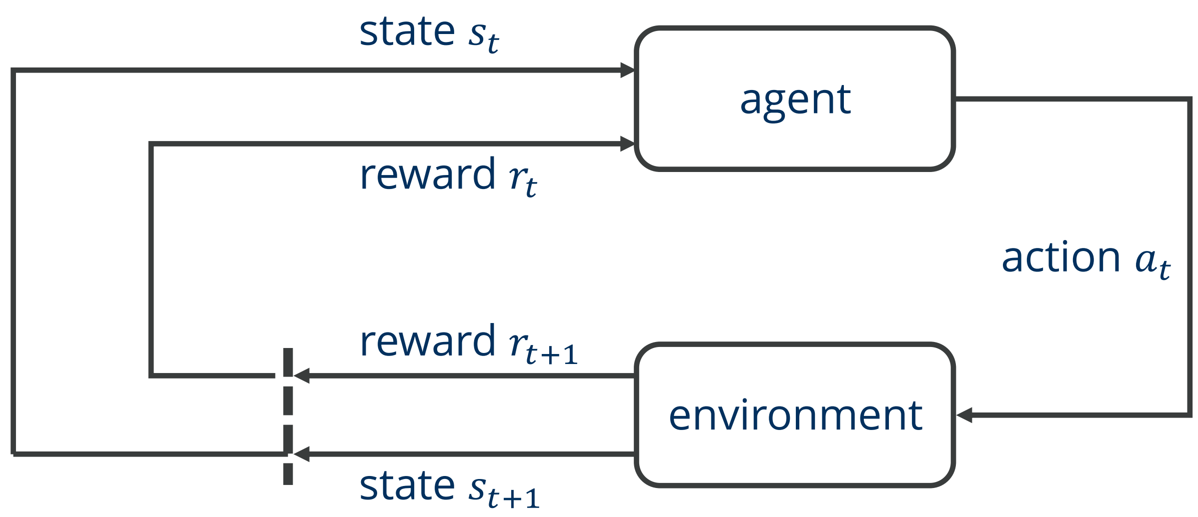
- Emerging V2I communication technology enables fast, bilateral exchange of information between vehicles and the infrastructure
- The information about the state of individual vehicles might enable better traffic control decision
- But, translating large amounts of data into a control decision is difficult and asks for novel optimisation techniques



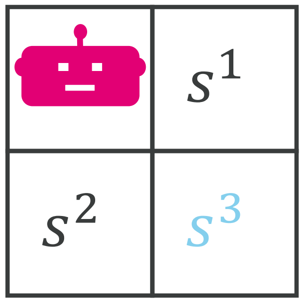
Objectives

- Design and implementation of a Reinforcement Learning (RL) algorithm that can be used to learn the control of traffic lights
- Design and implementation of a traffic simulation environment
- Investigation of the benefit of V2I communication on the efficacy of traffic control and assessment of the ability of RL to learn a good strategy

Markov Decision Processes

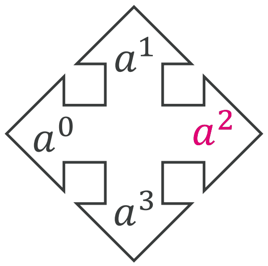


intial state



terminal state

states



actions

s_{t+1}	s^0	s^1	s^2	s^3
r_{t+1}	-1	-1	-1	3

rewards

Q-Learning

Policy $\boldsymbol{\pi}(\boldsymbol{s})$: "Which action $\boldsymbol{a}$ should I take when I am in state $\boldsymbol{s}$?"

Q-Function $\boldsymbol{Q^{\pi}}(\boldsymbol{s},\boldsymbol{a})$: "How much (discounted) reward can I expect in the future if I am in state $\boldsymbol{s}$, take action $\boldsymbol{a}$ and follow the policy $\boldsymbol{\pi}$ afterwards?"

Q^{π}

Policy Improvement

$$\pi(s_t) \leftarrow \operatorname*{argmax}_{a'} Q^{\pi}(s_t, a')$$

π

Policy Evaluation

$$Q^{\pi}(s_t, a_t) \leftarrow (1-\lambda)Q^{\pi}(s_t, a_t) - \lambda[r_{t+1} + \gamma \max_{a'} Q^{\pi}(s_{t+1}, a')]$$

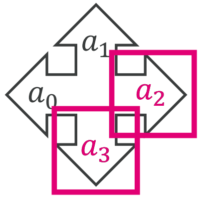
learning rate

discount factor

Q-Learning

initial state	s^0	s^1	
	s^2	s^3	terminal state

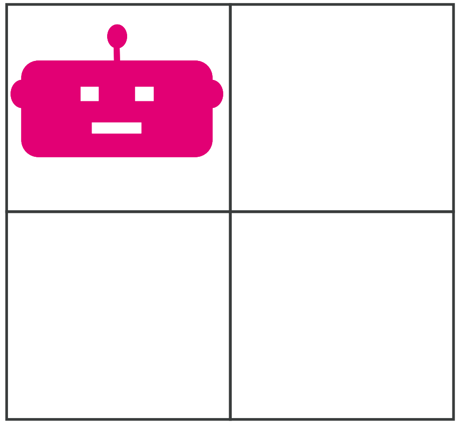
states



actions

s_{t+1}	s^0	s^1	s^2	s^3
r_{t+1}	-1	-1	-1	3

rewards



Q	a^0	a^1	a^2	a^3
s^0			2	2
s^1	1			3
s^2		1	3	

$Q^\pi(s^0, a^3) \leftarrow -1 + \max(0,0)$

$Q^\pi(s^2, a^2) \leftarrow 3 + 0$

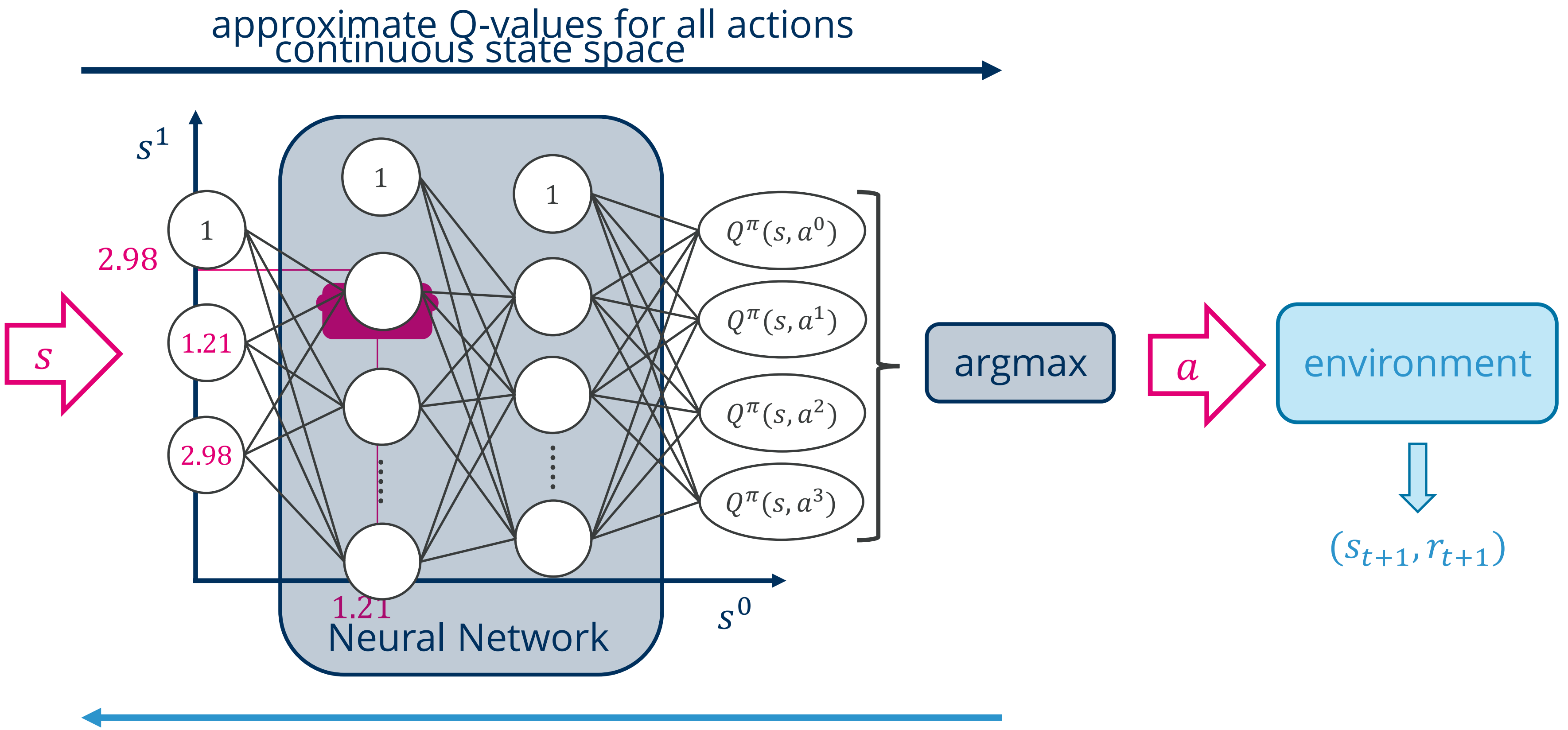
Policy Improvement

$$\pi(s_t) \leftarrow \underset{a'}{\operatorname{argmax}}\, Q^\pi(s_t, a')$$

Policy Evaluation

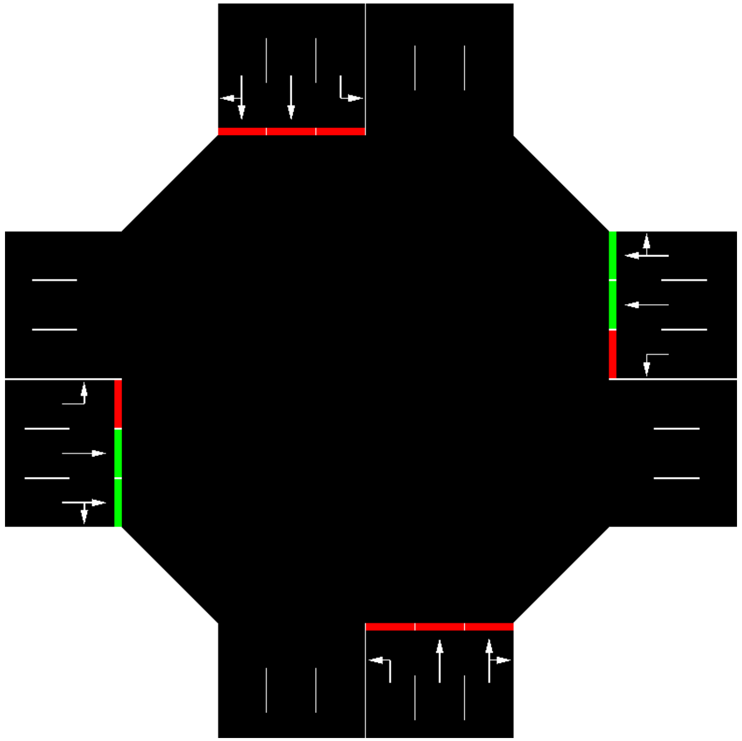
$$Q^\pi(s_t, a_t) \leftarrow (1-\lambda) Q^\pi(s_t, a_t) + \lambda [r_{t+1} + \gamma \max_{a'} Q^\pi(s_{t+1}, a')]$$

Deep Q-Learning [3]

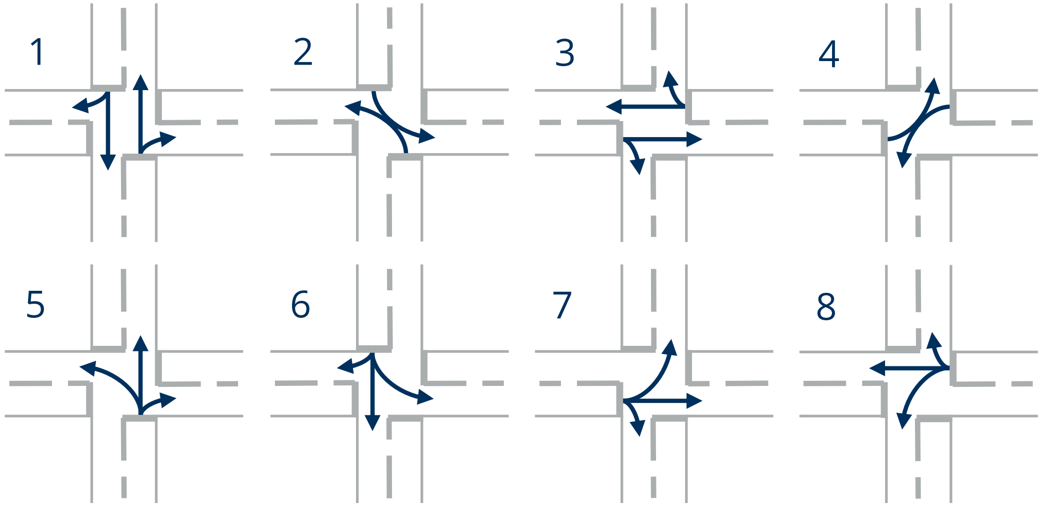


minimise: $MSE[Q^\pi(s_t, a_t)\ ,\ r_{t+1} + \gamma \max_{a'} Q^\pi(s_{t+1}, a')]$

A Traffic Control MDP: Action Space



- next phase



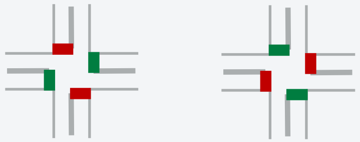
discrete phase options

- phase time of current phase

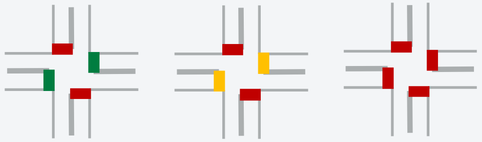
continuous phase length

A Traffic Control MDP: State Space

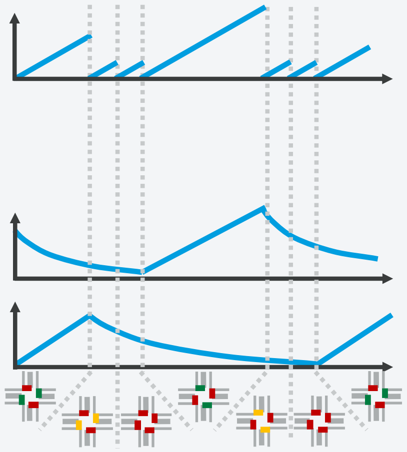
- phase



- period

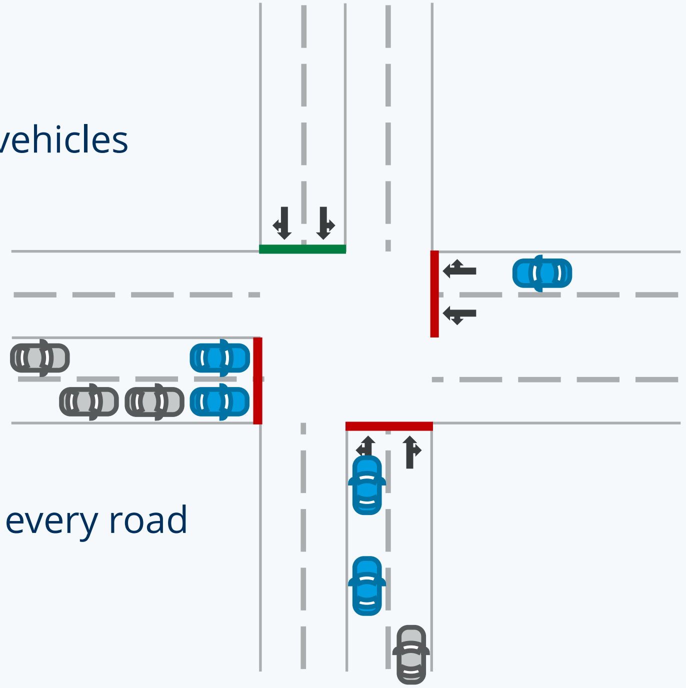


- time since last change
- phase traces



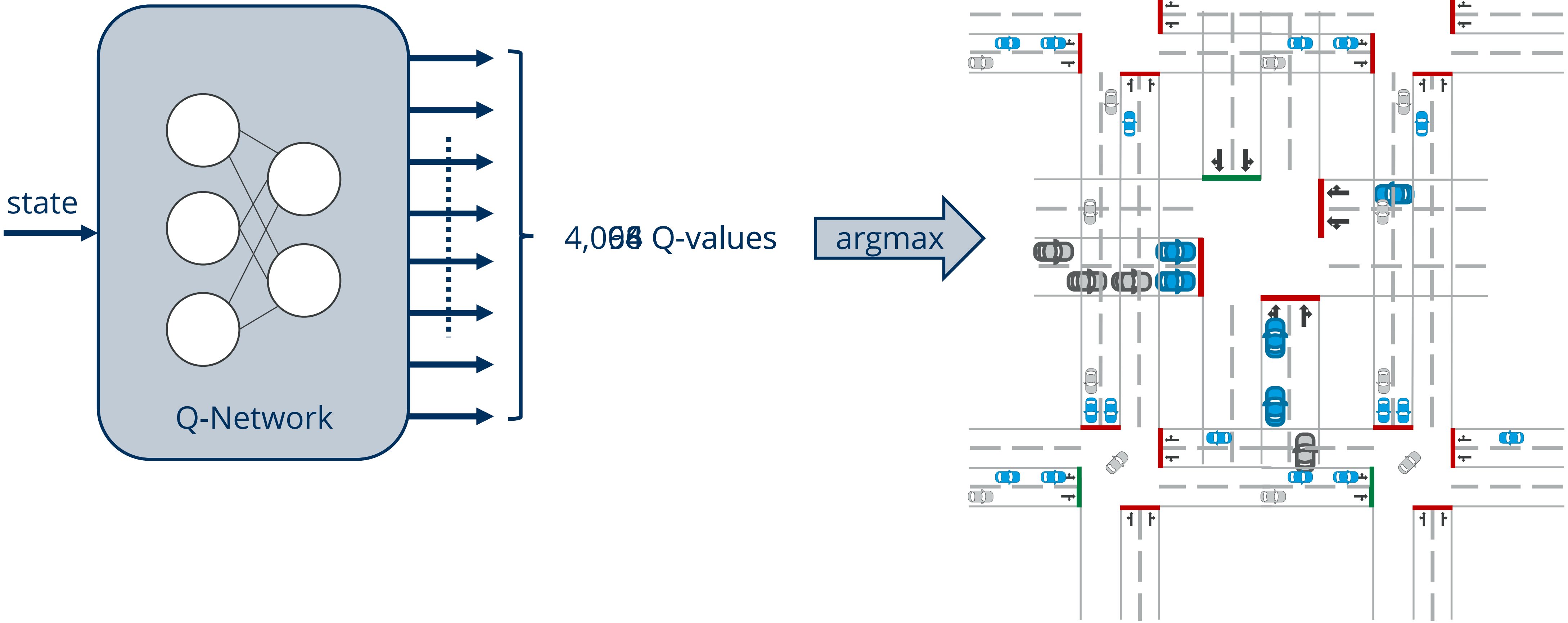
solitary agent

- all elements of the solitary agent
- positions of observed vehicles
- velocities of observed vehicles
- number of vehicles on every road
- average velocity on every road



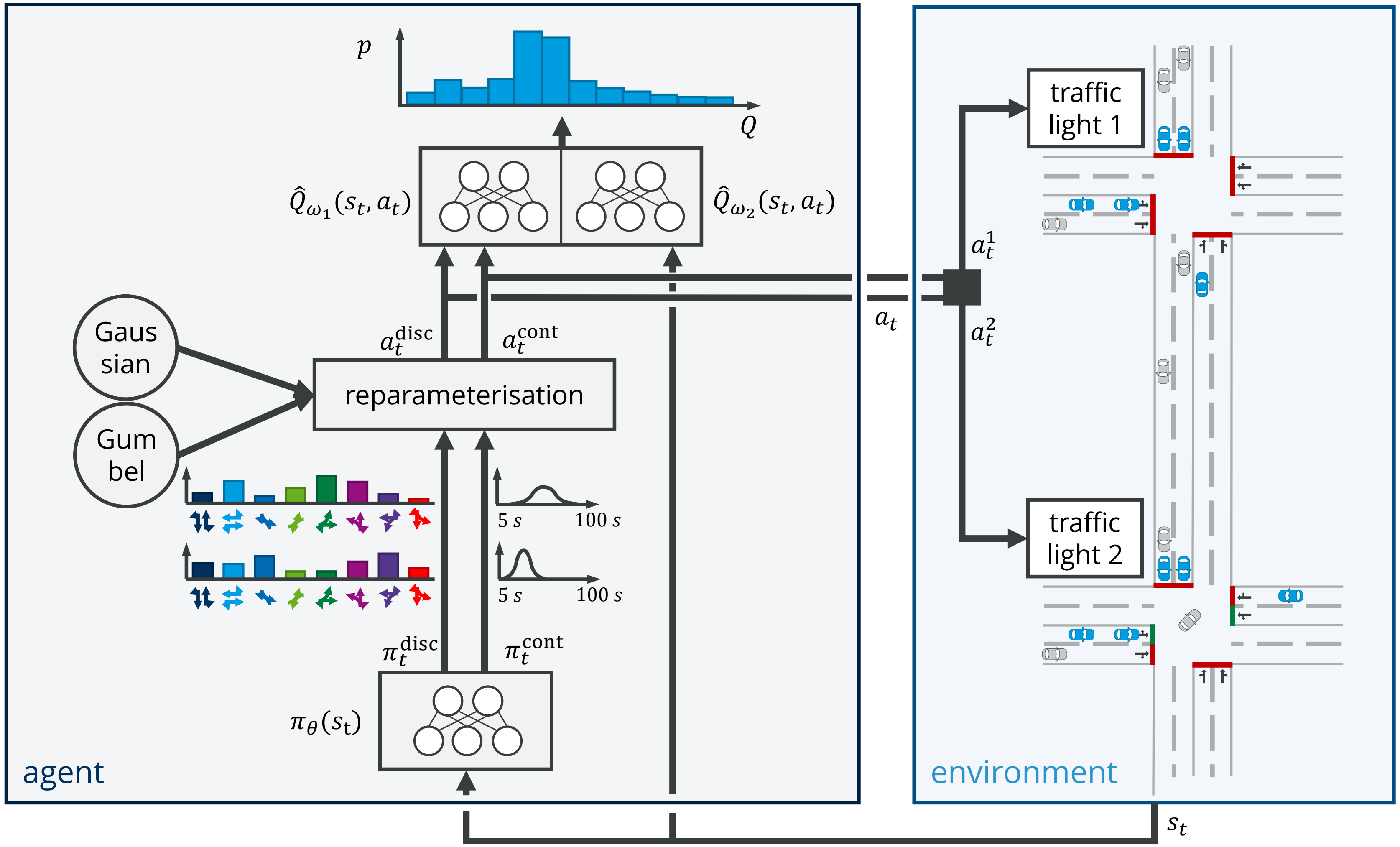
communicative agent

A Combinatorial Action Space

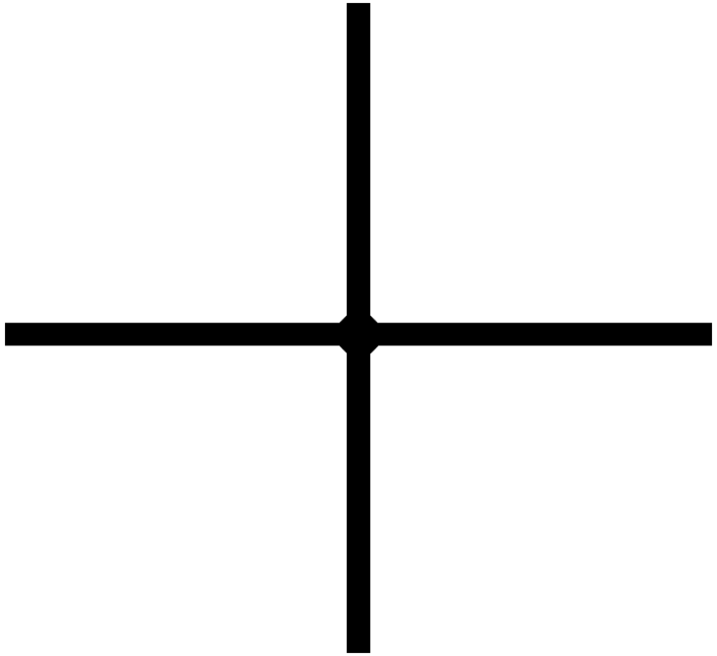


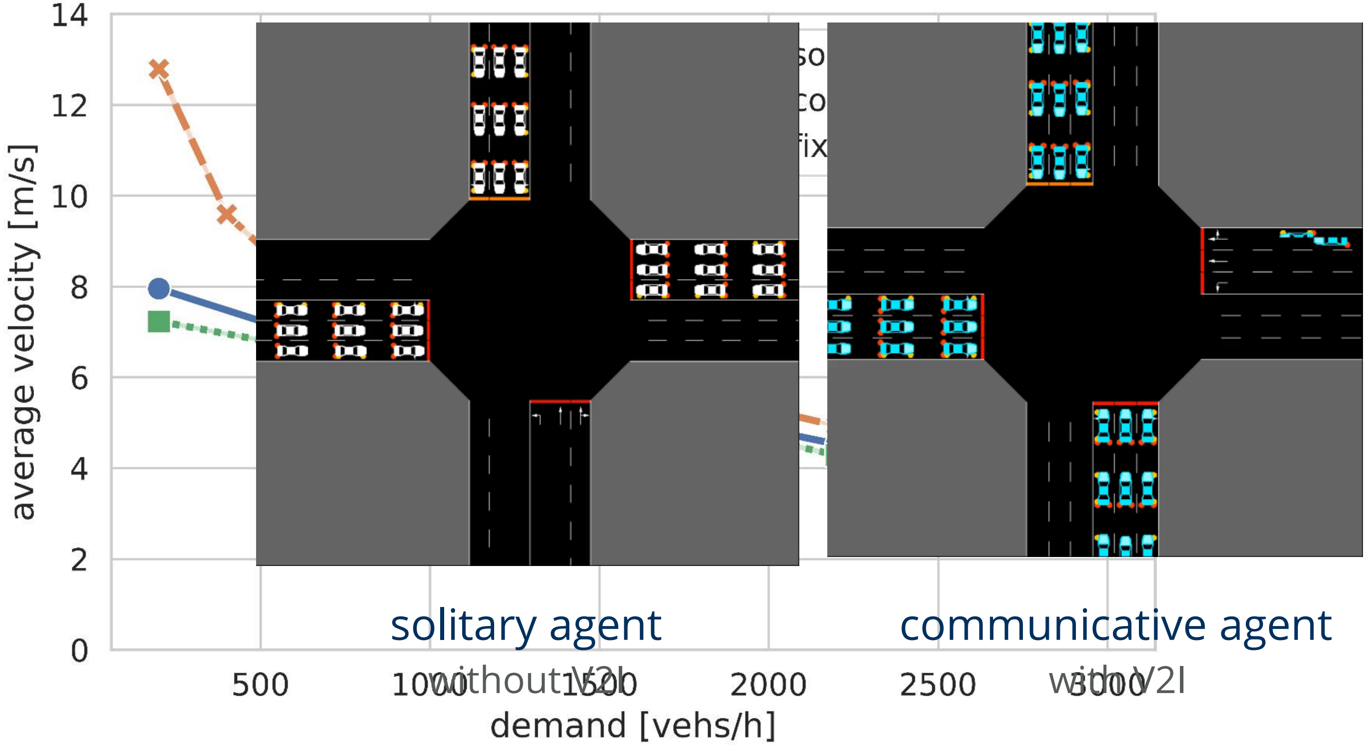
Agent 4D7

- Learn two Q-distributions and use the smaller one to mitigate the overestimation of the Q-function [4, 5]
- Learn a NN that approximates the Bolzmann-distribution over the Q-values [6]
- Reparameterise the sampling operation of actions so that we can compute the gradients of the [7, 8]



Single Intersection

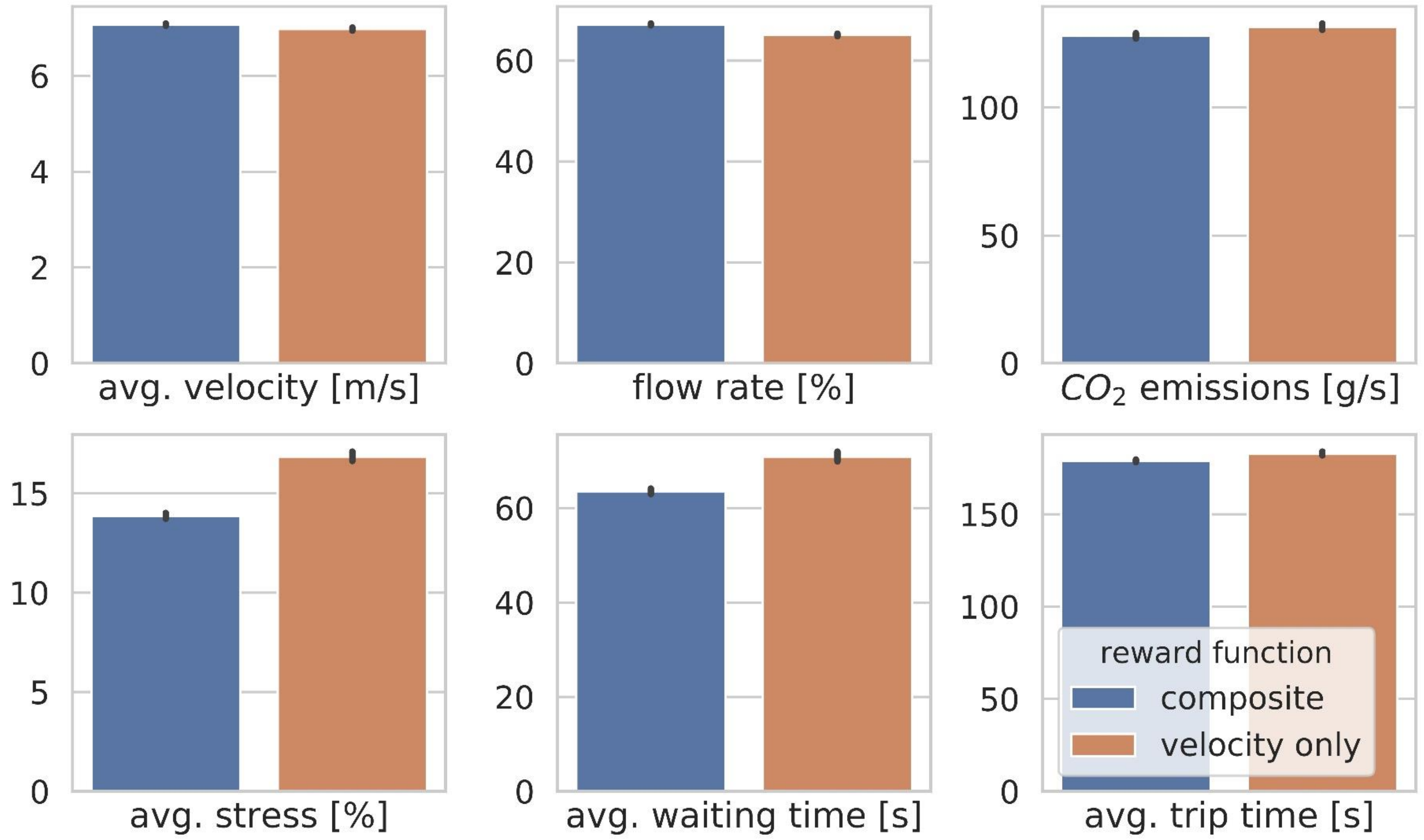




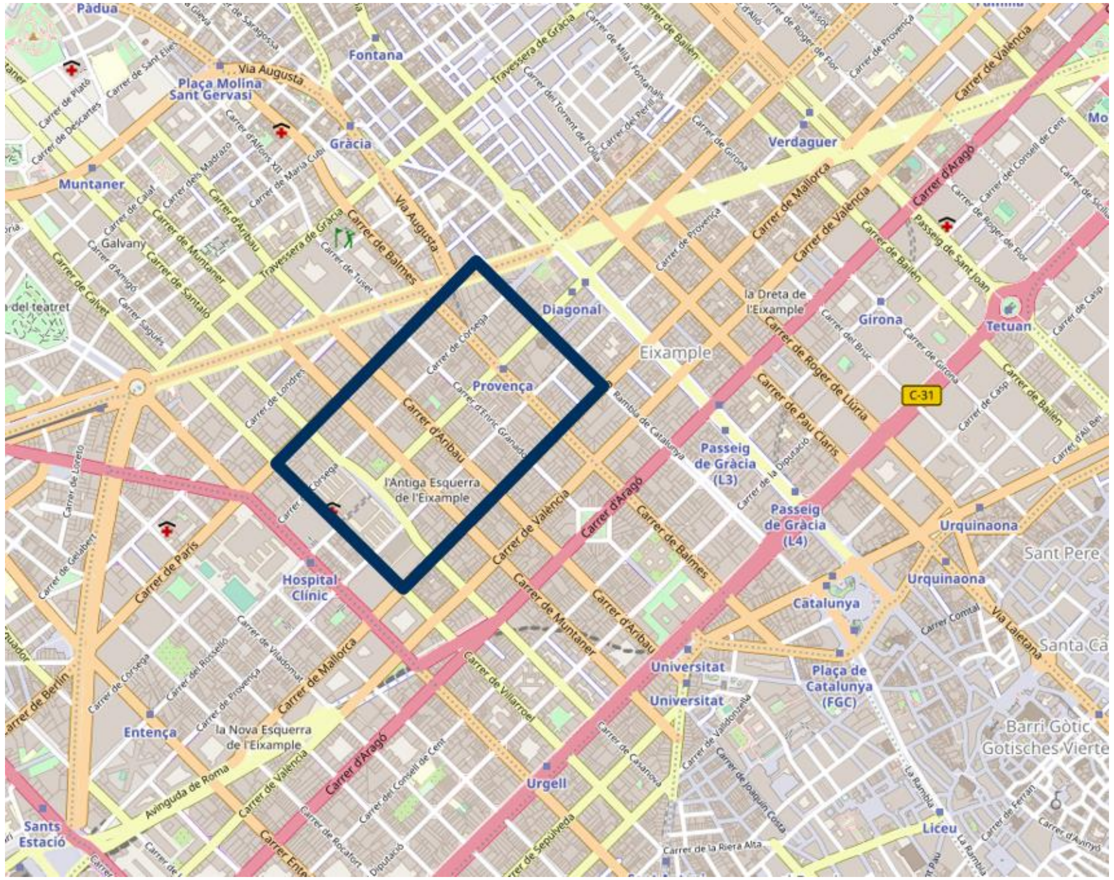
Composite Reward

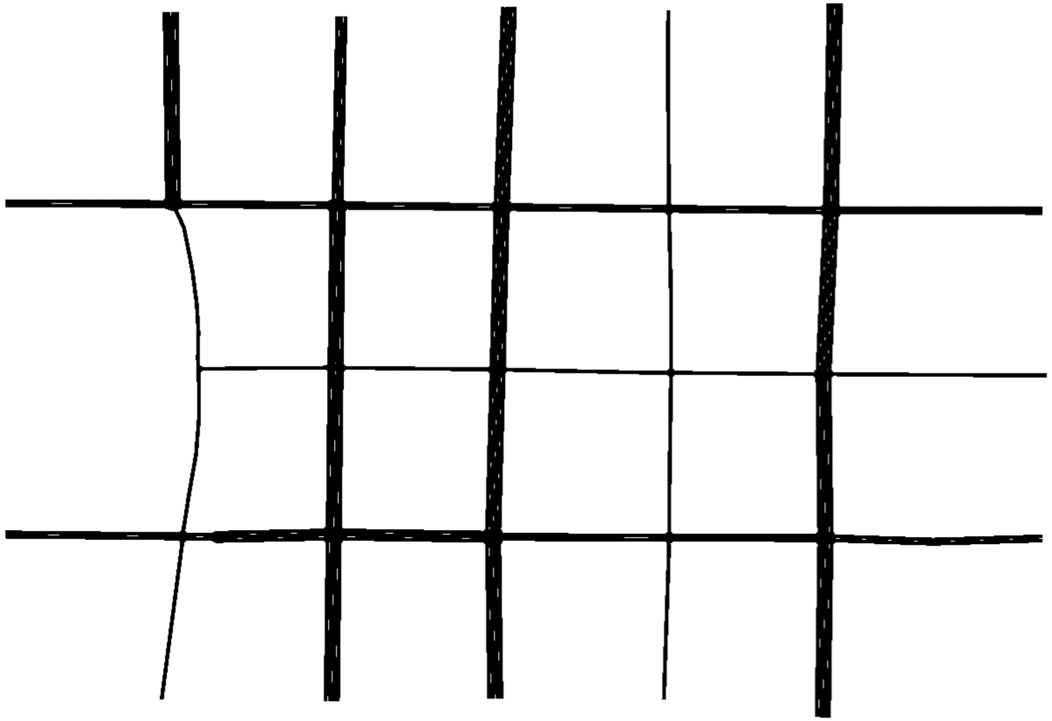
Joint optimisation of:

- average velocity of vehicles
- flow rate (percentage of vehicles that are not moving
- CO_2 emissions
- average stress of drivers (quadratic in the time spent waiting lately [9])

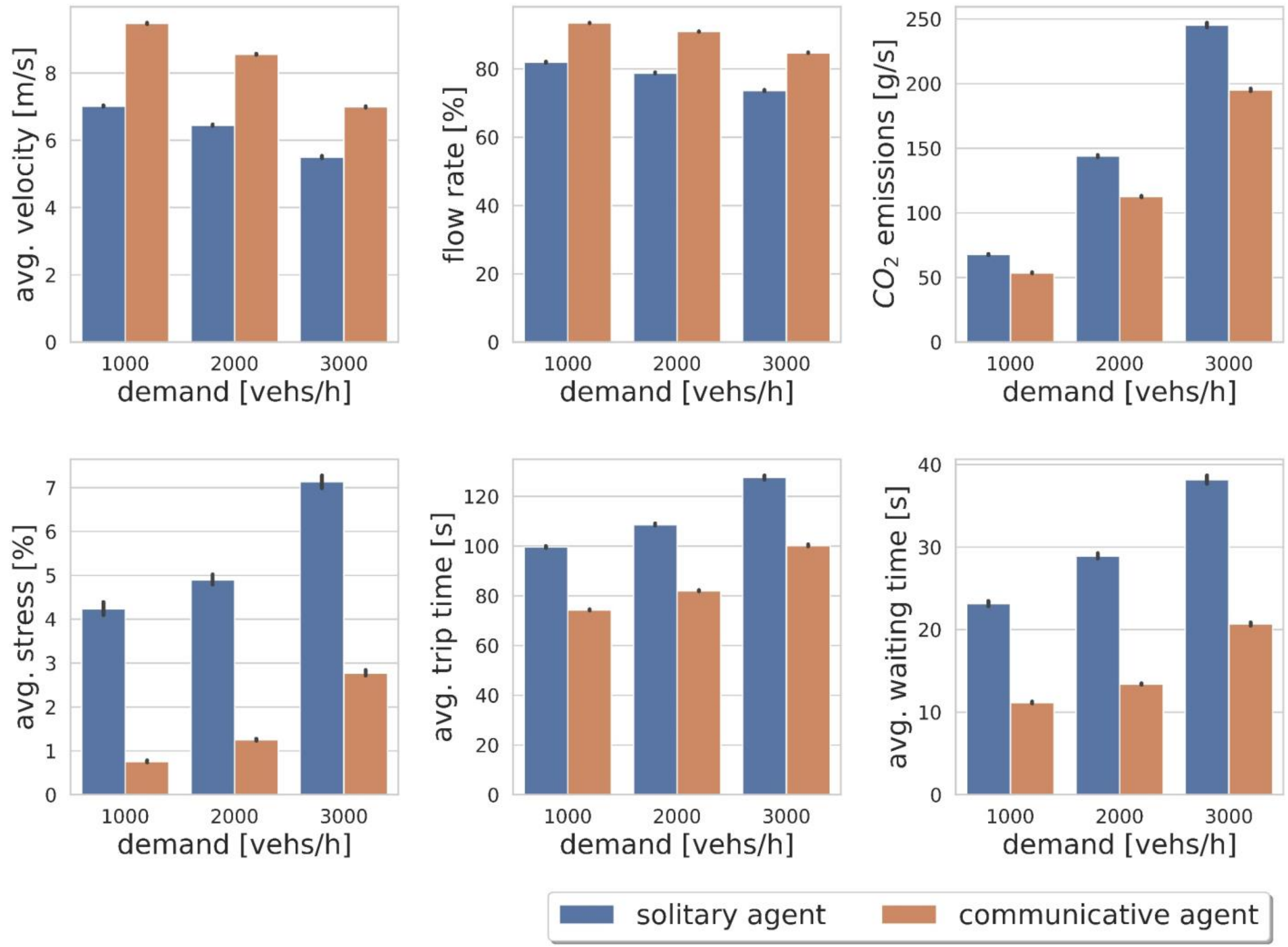


L’Antiga Esquerra de l’Eixample





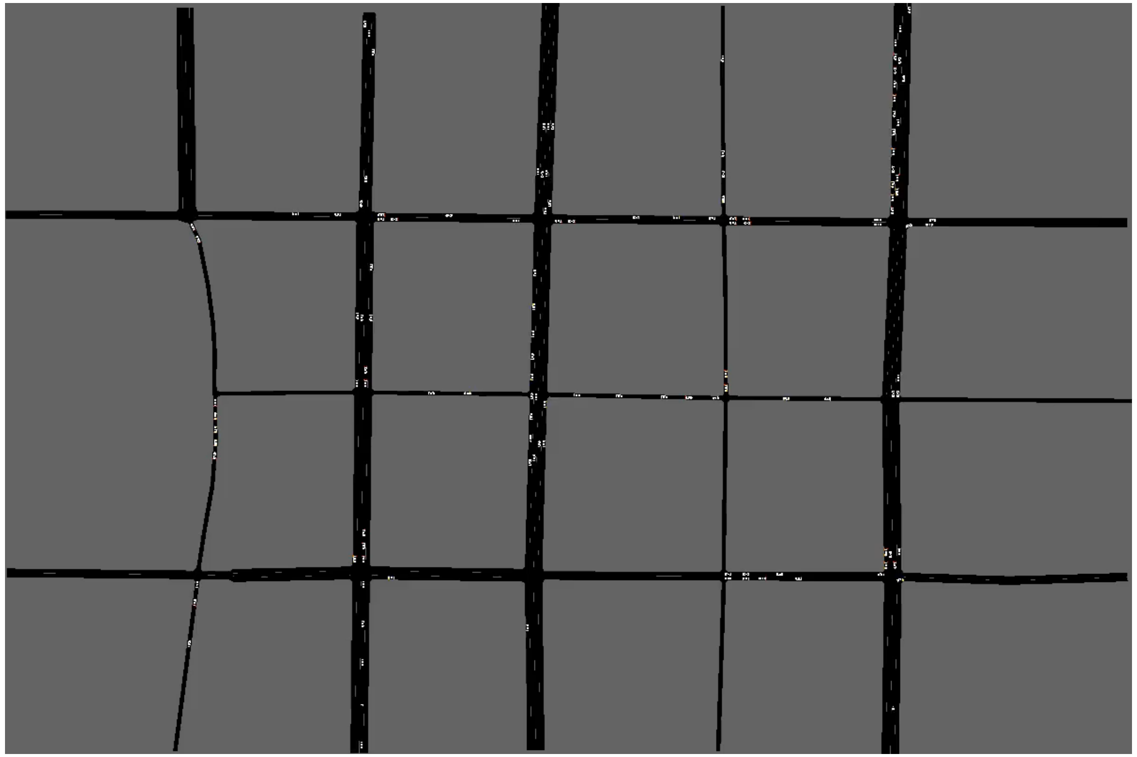
L'Antiga Esquerra de l'Eixample



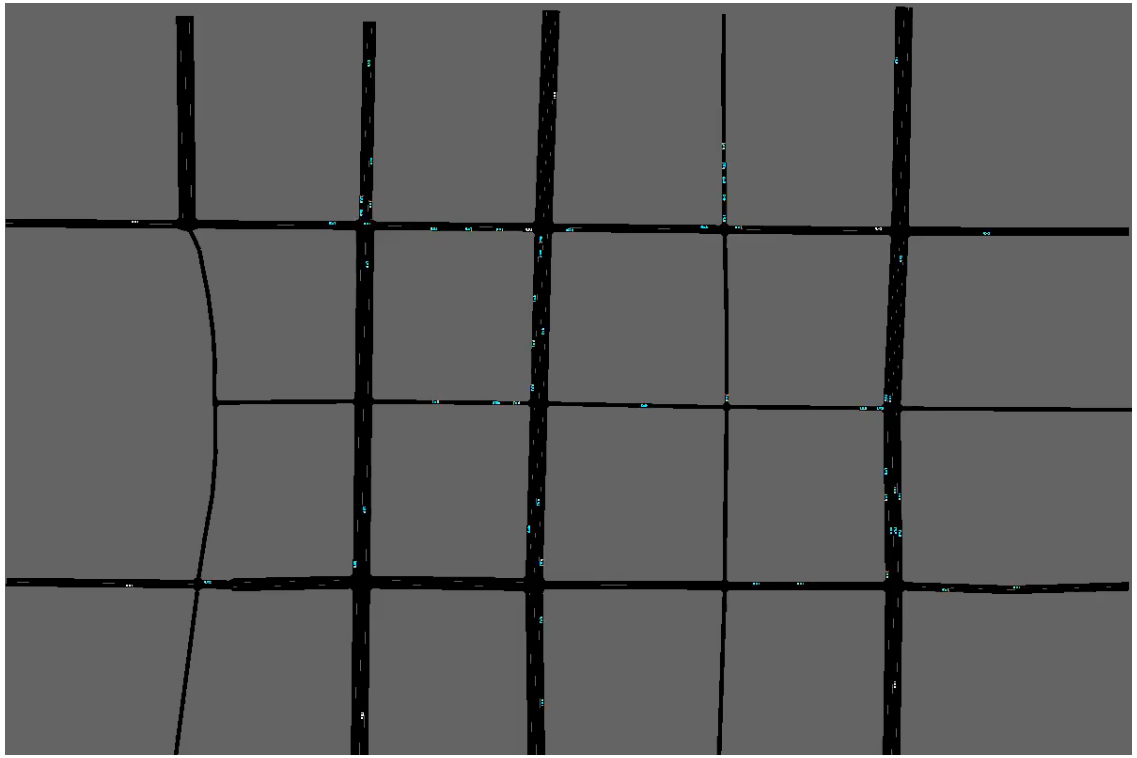
Through the availability of V2I communication, CO_2 emissions were lowered by ~20%.

The average time that vehicles spend waiting at traffic lights was reduced by ~50%.

L’Antiga Esquerra de l’Eixample



solitary agent
without V2I



communicative agent
with V2I

Conclusions

- The availability of Vehicle to Infrastructure communication has the potential to mitigate the problem of traffic congestion.
- Reinforcement Learning can be used to leverage the massive amounts of vehicle data in order to make more informed control decisions
- The model-free nature of Reinforcement Learning allows us to optimise arbitrary objective functions and not to rely on questionable model assumptions.

Outlook

- Comparison of DRL for traffic control with V2I communication against other approaches
- Improvement of the traffic simulation: more realistic driving behaviour, adaptive routing choices, other vehicle types, pedestrians…
- Incorporation of the ability of the traffic infrastructure to send messages to individual vehicles

Sources

[1] European Comission (2017). “European Urban Mobility - Policy Context”. Brussels, Belgium.

[2] Inrix (2018). “Inrix Global Traffic Scoreboard”.

[3] Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis (2015). “Human-level control through deep reinforcement learning”. In: *Nature* 518, p. 529.

[4] Barth-Maron, G., M. W. Hoffman, D. Budden, W. Dabney, D. Horgan, D. TB, A. Muldal, N. Heess, and T. P. Lillicrap (2018). “Distributed Distributional Deterministic Policy Gradients”. In: *arXiv e-prints* arXiv:1804.08617.

[5] Fujimoto, S., H. van Hoof, and D. Meger (2018). “Addressing Function Approximation Error in Actor-Critic Methods”. In: *arXiv e-prints* arXiv:1802.09477.

[6] Haarnoja, T., A. Zhou, P. Abbeel, and S. Levine (2018a). “Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor”. In: *arXiv e-prints* arXiv:1801.01290.

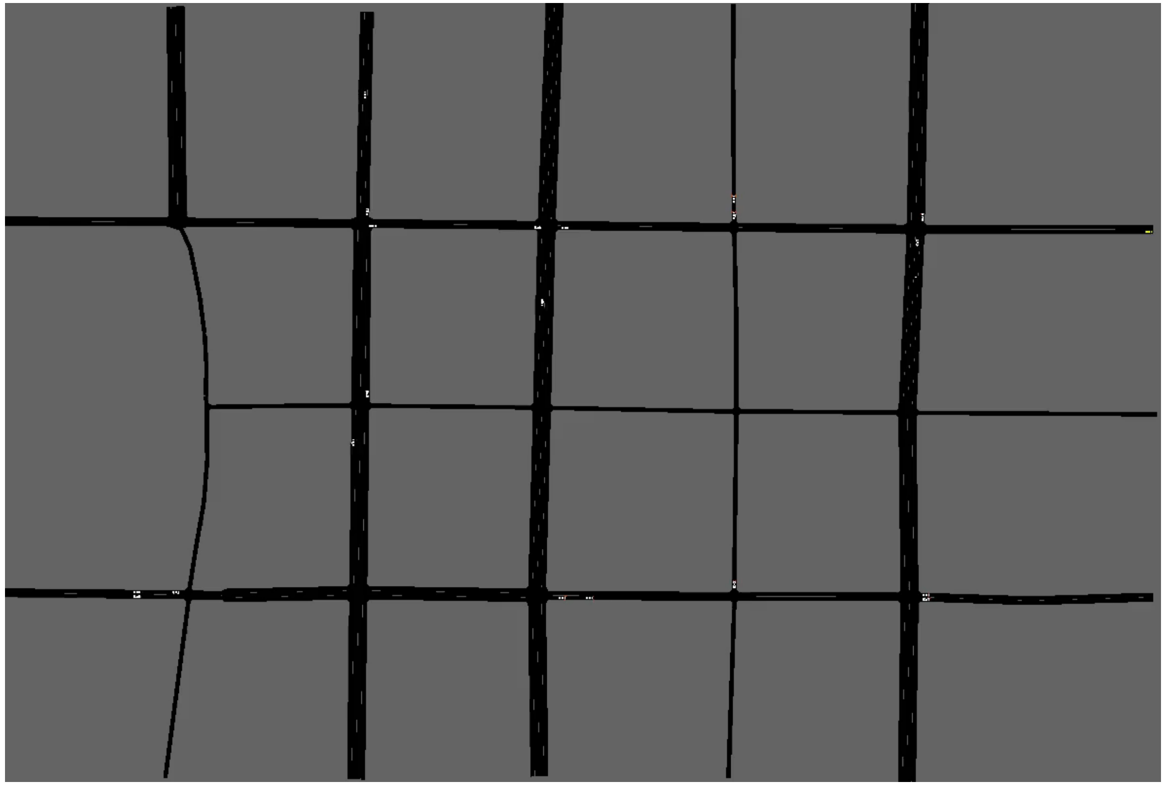
[7] Maddison, C. J., A. Mnih, and Y.W. Teh (2016). “The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables”. In: *arXiv e-prints* arXiv:1611.00712.

[8] Jang, E., S. Gu, and B. Poole (2016). “Categorical Reparameterization with Gumbel-Softmax”. In: *arXiv e-prints* arXiv:1611.01144.

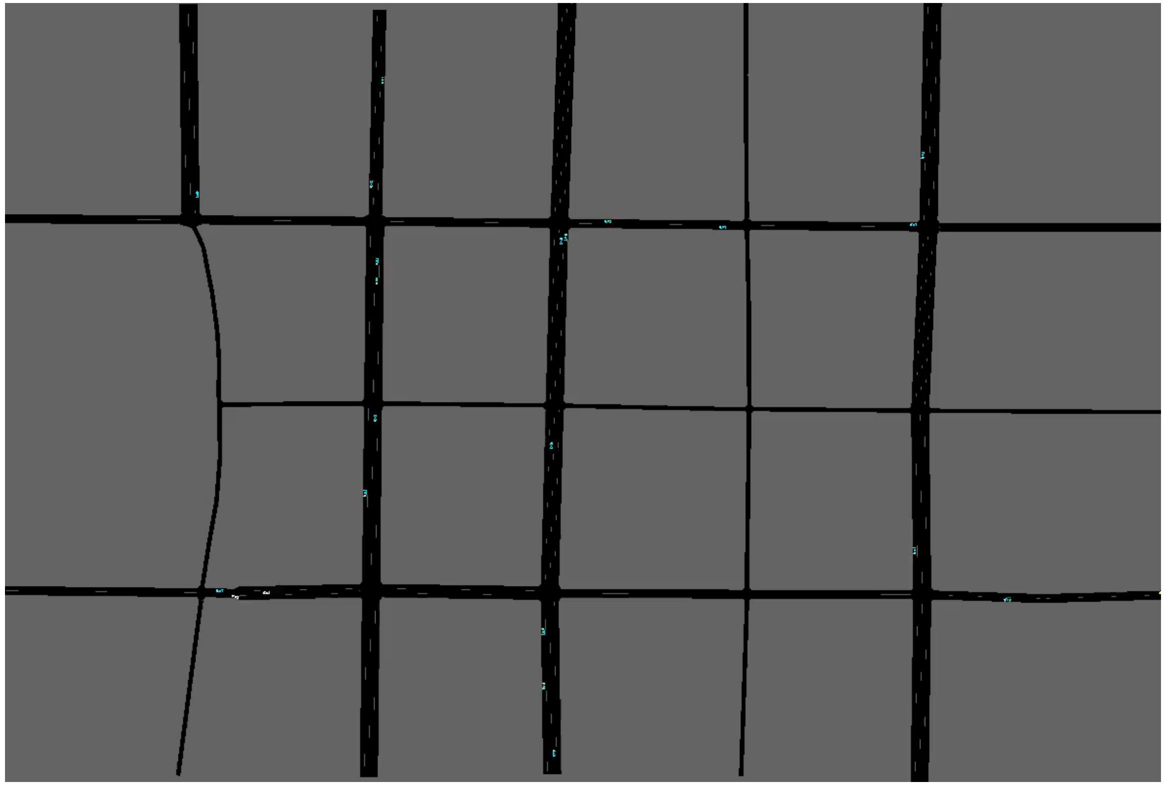
[9] Liu, M., J. Deng, X. Ming, Xianbo Zhang, and W. Wang (2017). “Cooperative Deep Reinforcement Learning for Traffic Signal Control”. In: *Proceedings of the 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.

Thank you for your attention.

L'Antiga Esquerra de l'Eixample

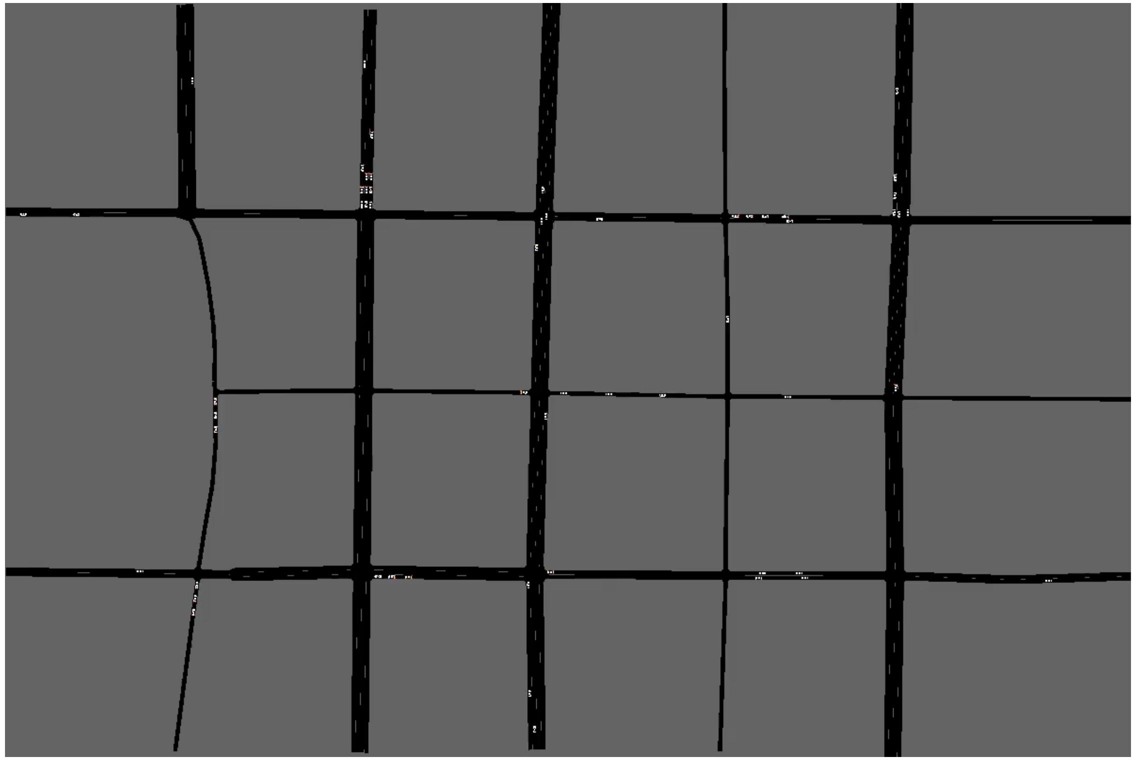


solitary agent

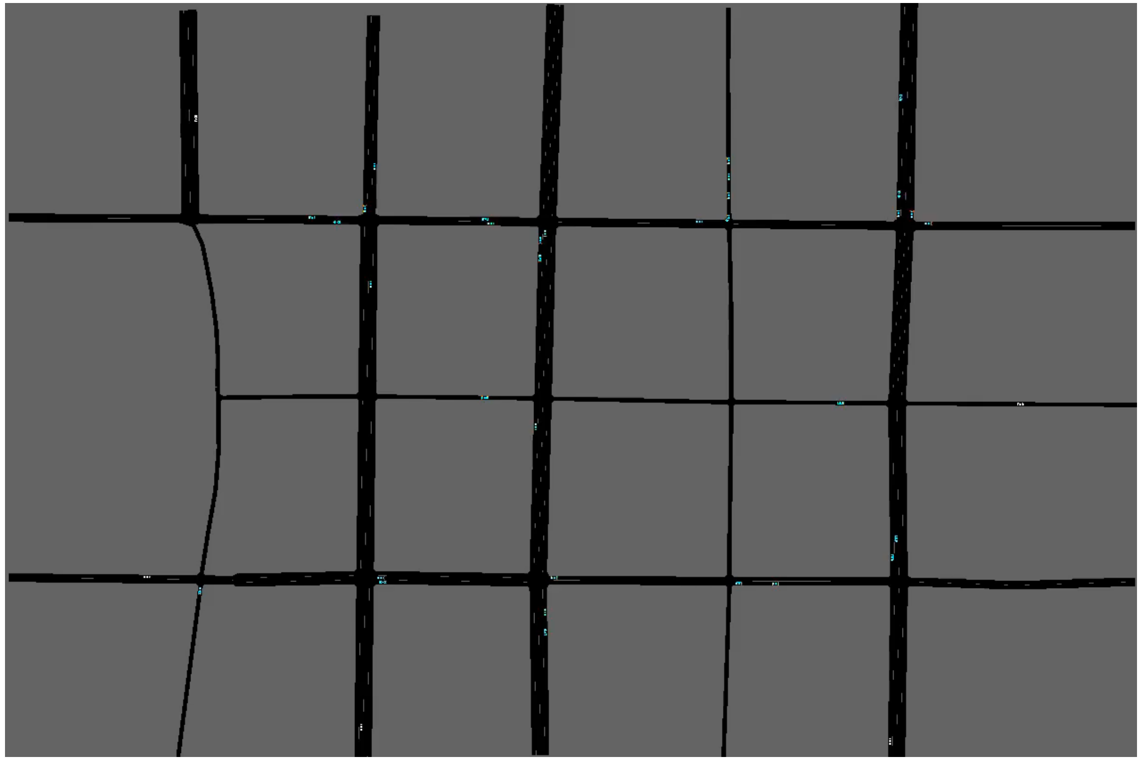


communicative agent

L'Antiga Esquerra de l'Eixample



solitary agent



communicative agent