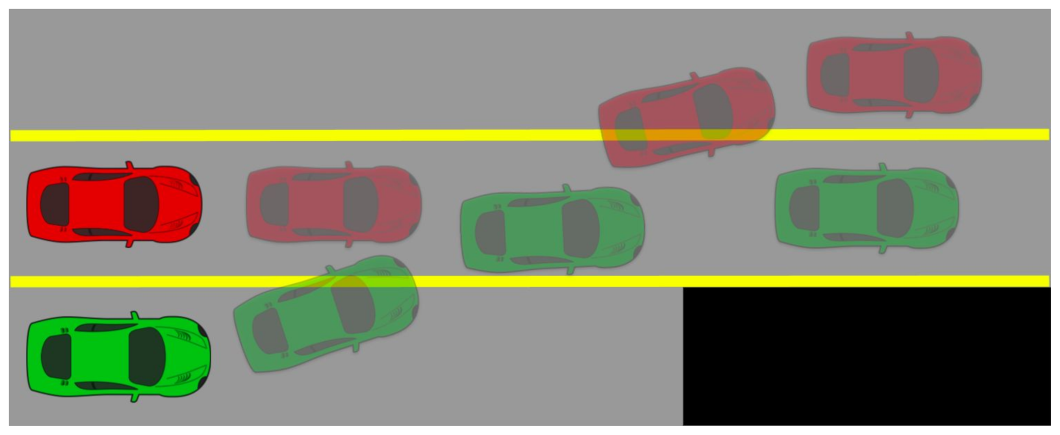
MARL-PPS: Multi-agent Reinforcement Learning with Periodic Parameter Sharing

UCLAVISIONLAB

Honda Research Institute US

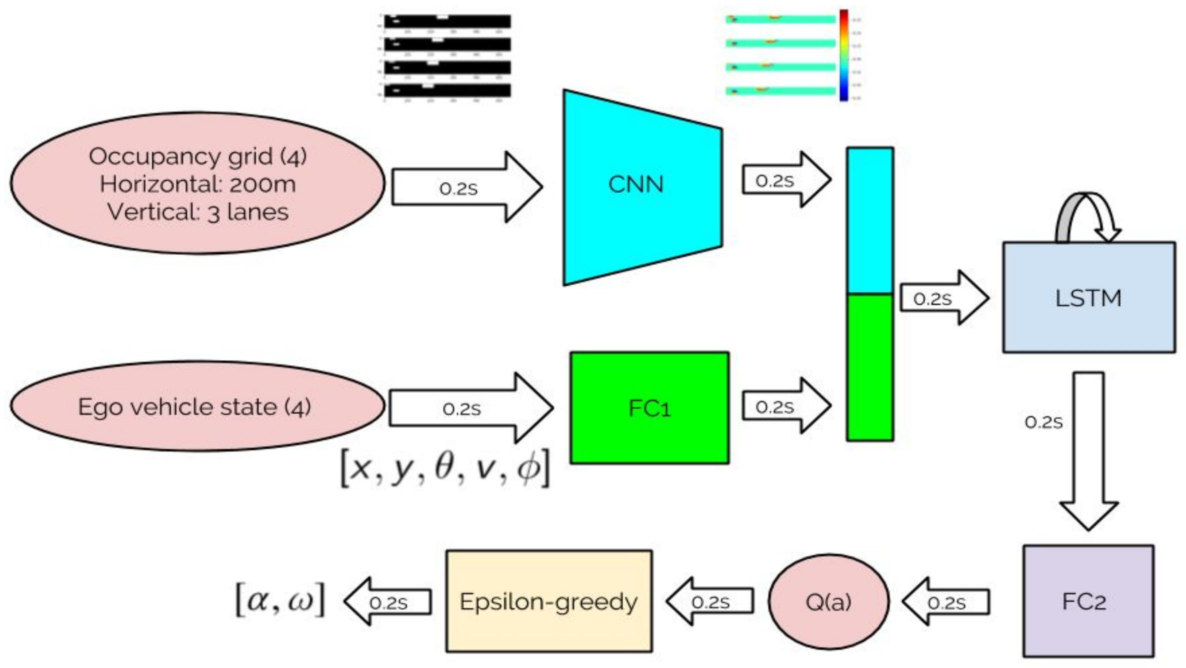
Safa Cicek, Alireza Nakhaei, Stefano Soatto, Kikuo Fujimura

Motion planning on highways



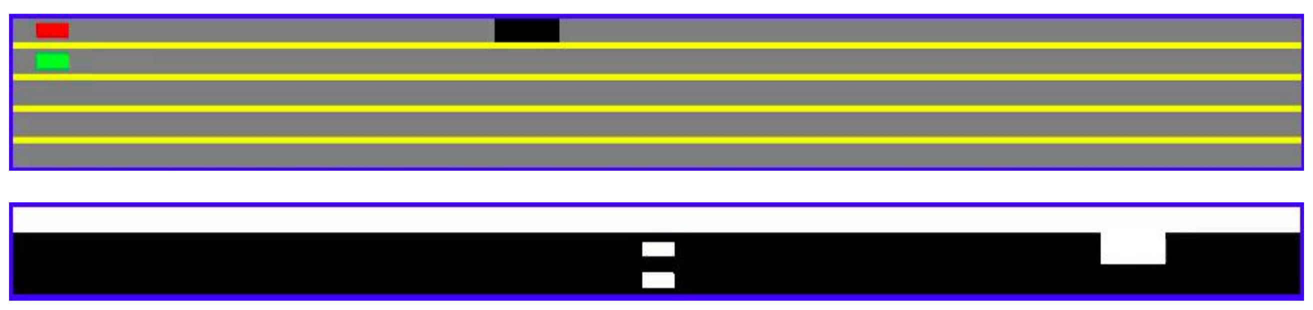
- An interaction-aware planning algorithm is expected to exhibit cooperative behavior.
- The red vehicle must maintain a predictive model of the green one for cooperative behaviour.

Flow of an RL agent



- The RL algorithm gets two modes of inputs at every 0.2 sec. It takes the last 4 occupancy grids in its observation range in the form of binary matrices. All these grids are fed to the CNN.
- The last 4 ego-motion states of the vehicle are also given as input and fed to the fully connected layer for preprocessing.
- Outputs of the fully connected layer are concatenated with CNN outputs to be sent to the LSTM.
- The LSTM output is fed to another fully connected layer to get the Q-value estimates.
- Finally, the epsilon-greedy block chooses an action with 0.2 sec resolution.

The highway simulator



- Screenshots are from the highway simulator RL agents are trained on.
- The top panel is the scene with two vehicles (red and green rectangles) and a static obstacle to be avoided in the top lane (black rectangle).
- The bottom panel shows the observation of the red vehicle.

Reward function

- Intent-aware (interaction-unaware) reward for any agent:

$$r = -\lambda_{\text{collision}} I_{\text{collision}} + \lambda_v \frac{{v_0}^2}{(v - v^*)^2 + {v_0}^2} + \lambda_\theta \frac{{\theta_0}^2}{(\theta - \theta^*)^2 + {\theta_0}^2} + \lambda_{\text{jerk}}\left[\frac{\dot{\alpha}_0^2}{\dot{\alpha}^2 + \dot{\alpha}_0^2} + \frac{\dot{\omega}_0^2}{\dot{\omega}^2 + \dot{\omega}_0^2}\right]$$

- Interaction-aware reward for agent 1 when agents j=2,...,J are in the observation range of agent 1:

$$\mathbf{r}_{1,t} = r_{1,t} + \lambda_{\text{coop}} \sum_{j=2}^{J} r_{j,t}$$

DQN

- Moving target problem in fitted Q-learning:

$$L(w_t) = \mathbb{E}_{(s,a,r,s')\sim U(D)}[(\underbrace{r + \gamma \max_{a'} Q(s', a'; w_t^-)}_{\text{Target}} - Q(s, a; w_t))^2]$$

Use replay buffer for reducing the correlations in between samples

Update target network periodically to stabilize the target

DQN in MARL

- Moving environment problem in MARL:

$$L(w_t) = \mathbb{E}_{(s,a,r,s')\sim U(D)}[(r + \gamma \max_{a'} Q(s', a'; w_t^-) - Q(s, a; w_t))^2]$$

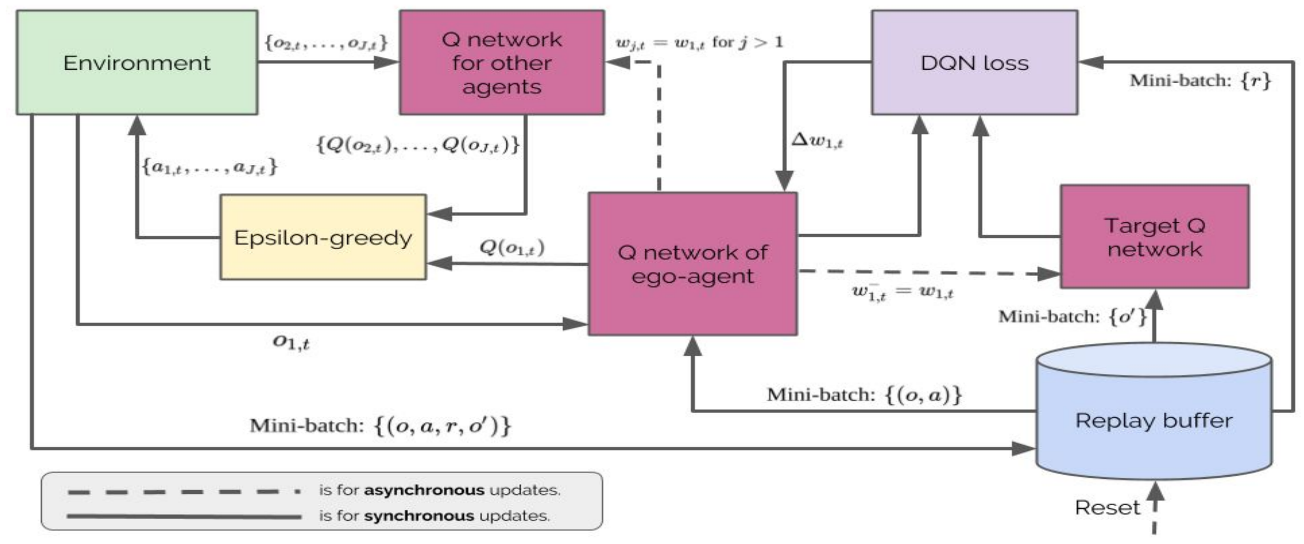
- The next state (s') is function of the actions of other agents. Thus, environment dynamics change as policies of other agents updated.
  - Policy of other agents: epsilon-greedy selection from their Q estimates.
  - Proposed solution: Update Q function of other agents with large periods.

MARL in POMDP setting:

- In POMDP case,
  - Sample observation from the buffer
  - Feed hidden layer of LSTM and observation to the Q networks.

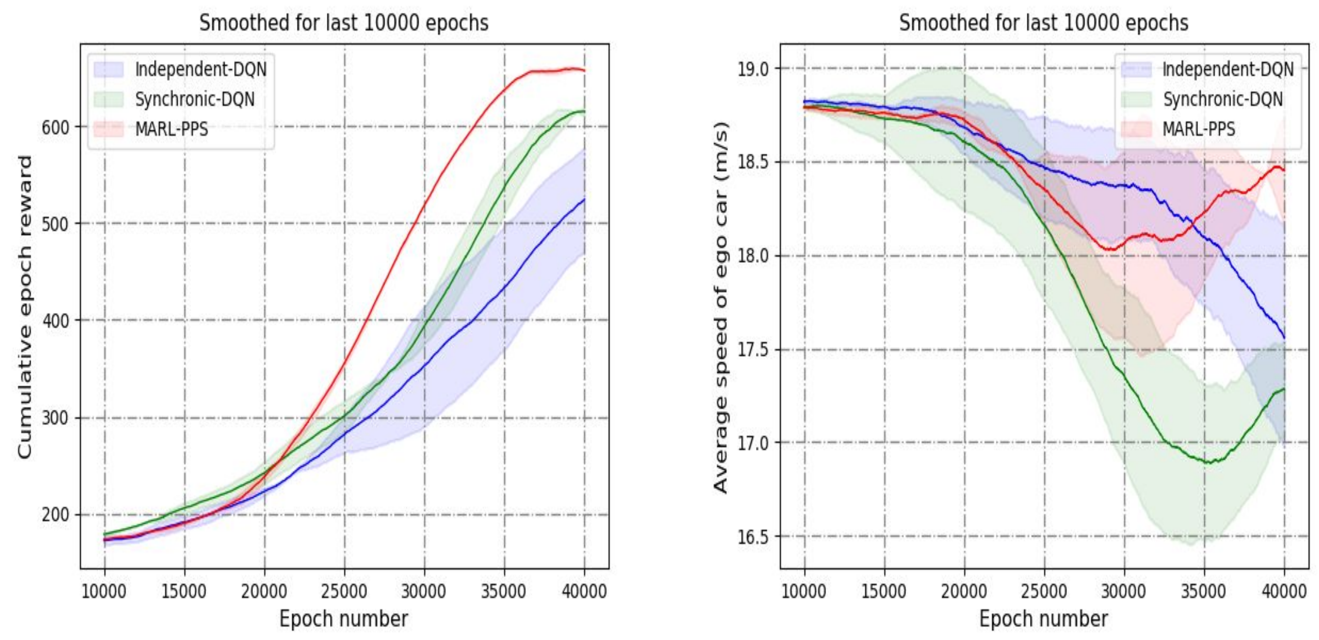
$$L(w_t) = \mathbb{E}_{(o,a,r,o')\sim U(D)}[(r + \gamma \max_{a'} Q(o', h, a'; w_t^-) - Q(o, h, a; w_t))^2]$$

MARL-PPS



- Key differences of the proposed algorithm from other DQN based algorithms are
  - The large periodic updates of the parameters of other agents.
  - Resetting of the replay buffer with the same period.

Results



- Training curves for baselines and the proposed algorithm. The left plot shows the mean epoch rewards for different methods. The right plot is for the average speed of the agents.
- Baselines: Independent-DQN [Tampuu, 2017] and Synchronic-DQN [Gupta, 2017].
- In Independent-DQN, each agent updates its DQN policy with its own observations concurrently.
- In synchronic-DQN, one ego-agent updates its policy with its own observations and shares its parameters at every time step with others.
- MARL-PPS converges to a better solution benefiting from the stability of the training

References

[1] Jayesh K Gupta, Maxim Egorov, andMykel Kochenderfer. 2017. Cooperative multiagent control using deep reinforcement learning. In International Conference on Autonomous Agents and Multiagent Systems. Springer, 66–83.

[2] Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus,Juhan Aru, Jaan Aru, and Raul Vicente. 2017. Multiagent cooperation and competition with deep reinforcement learning. PloS one 12, 4 (2017), e0172395.