Adaptive Action Selection Strategy Of Reinforcement Learning Approach For Intelligent Traffic Signal Control

Lin Lyu, Dongqin Zhou, Hao Liu, S. Ilgin Guler, Vikash V. Gayah

Department of Civil and Environmental Engineering, The Pennsylvania State University

PennState College of Engineering

LARSON TRANSPORTATION INSTITUTE

Paper No. 23-04427

Overview

How to address the issue of adaptive exploration and exploitation when using Reinforcement Learning (RL) for traffic signal control (TSC) ?

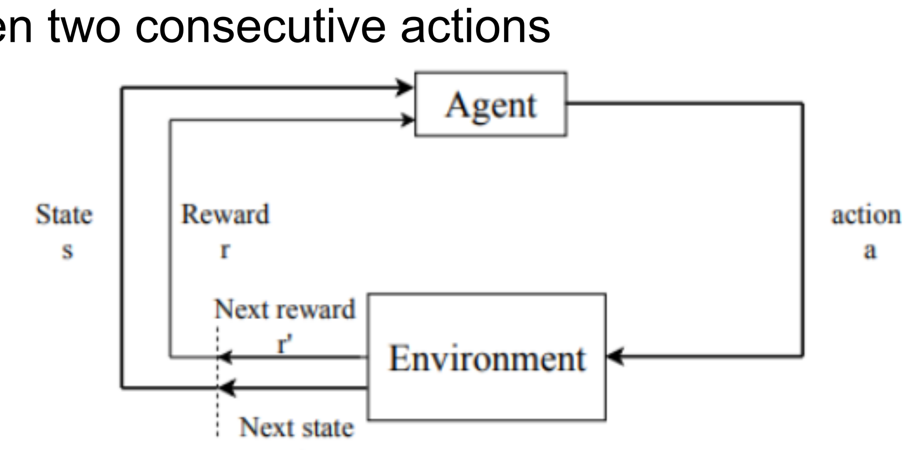
- Traffic signals are the major bottlenecks in urban areas and hence are major contributors to congestion
- Fixed time signal, actuated and adaptive signals are hard to be generalized in different scenarios
- TSC can be improved by adaptive learning algorithm in RL
- Balancing exploration & exploitation in RL models is important:
  - Too much exploitation: learning a sub-optimal policy finally
  - Too much exploration: leading to congestion currently
- Traditional exploration & exploitation methods, such as epsilon greedy (ε-greedy) in RL cannot handle this problem well
  - Will introduce hyper-parameters needed to be tuned
  - Hard to be generalized
- Challenges solved by our adaptive action selection strategy in RL, where agent adjust the policy based on different states

Methodology

Double Q learning Agent Design

- RL agent works as the central controller to update traffic signal timing at a single, isolated intersection
- State s: current green phase p; Discretized density K_i for each upstream lane i; Phase duration t, where $t = \begin{cases} t & t \le 60s \\ 60 & t > 60s \end{cases}$
- Action a: action 1 is green phase for NS, action 2 is green phase for EW.
- Reward r: total delay between two consecutive actions

Intersection



Interactions between agent and environment

Exploration/Exploitation Strategies

Baselines

- Fixed time signal
- ε-greedy: with a fixed ε to perform random actions
- Time-decayed ε-greedy: with a decayed ε to perform random actions
- Upper confidence bound (UCB): select the action that has the highest estimated action-value plus the upper-confidence bound term

Our Adaptive ε-greedy Strategy

- Extend the ε-greedy strategy to a state-dependent exploration rate $\varepsilon(s) = 1/\sqrt{n(s)}$, n(s) is the number of times state s has been visited

Experiments results

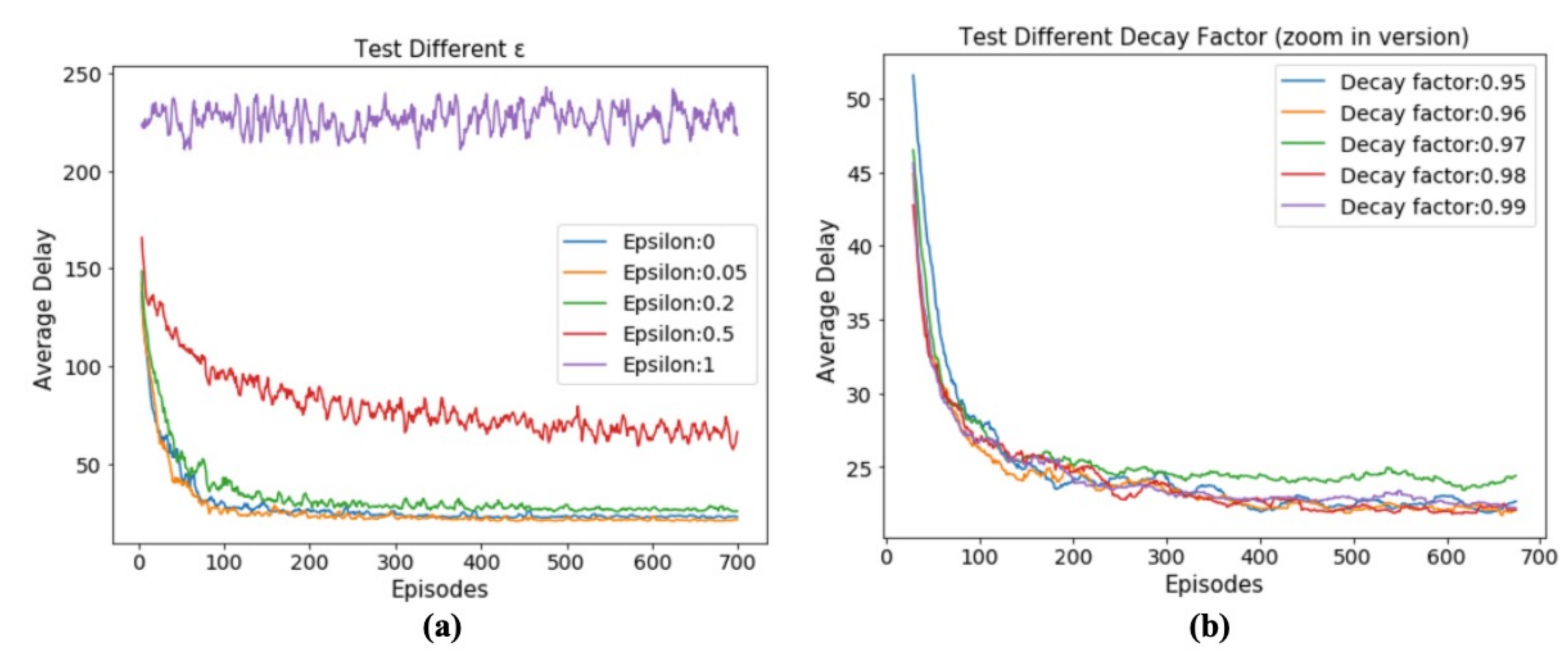
Model Overall Performance

- Performance is measured by the average delay on different flow configuration, where ε-greedy with ε of 0.05

Flow N/S (veh/h/lane/ direction)	Flow E/W (veh/h/lane/ direction)	Fixed-Time(s)	Adaptive ϵ-greedy(s)	UCB(s)	ϵ-greedy(s)
700	700	26.44 ± 0.046	**19.77 ± 0.0539**	20.53 ± 0.0324	21.11 ± 0.0856
500	900	24.21 ± 0.0364	**19.56 ± 0.0385**	22.23 ± 0.1627	22.81 ± 0.0280
400	1000	22.87 ± 0.0487	**21.41 ± 0.0694**	20.04 ± 0.1095	22.83 ± 0.1050
600	900	31.07 ± 0.0459	**22.17 ± 0.0289**	23.51 ± 0.3866	23.75 ± 0.5279
700	900	28.34 ± 0.0596	**26.77 ± 0.2352**	28.12 ± 0.2592	28.21 ± 0.1992
800	900	47.67 ± 0.0725	**28.60 ± 0.0793**	31.42 ± 0.2222	29.15 ± 0.1510
700	1000	46.13 ± 0.0653	**29.48 ± 0.1449**	35.75 ± 0.4246	30.61 ± 0.3029

Exploration/Exploitation Comparisons

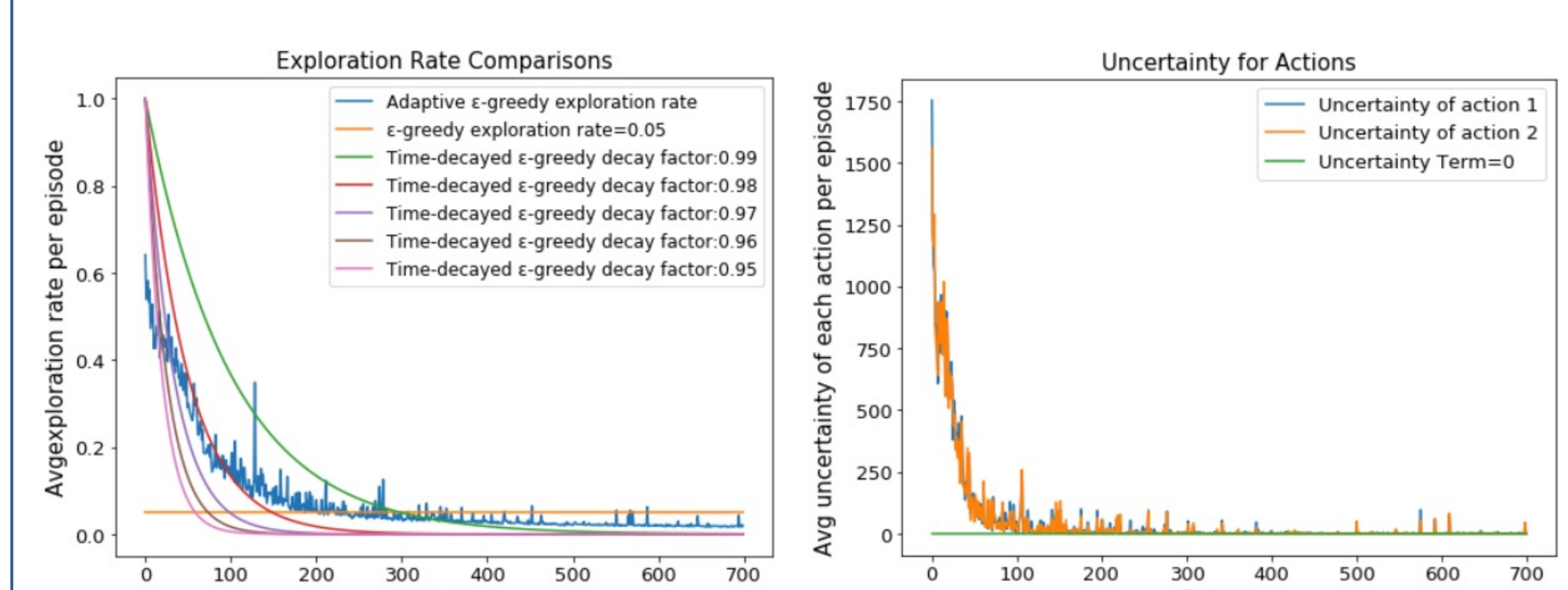
- A case study
  - Flows of 500 vehicles/h/lane for NS direction and 900 vehicles/h/lane for EW direction
  - Comparing the exploration/exploitation process of different models
- Tune hyper-parameters for ε-greedy (a) & time-decayed ε-greedy (b)



- After finding the best hyper-parameters, testing results - average delay for different models:

Model	Average Delay (s)
Adaptive ε-greedy	**19.562 ± 0.0385**
UCB	22.2274 ± 0.16269
Best ε-greedy	22.1138 ± 0.0280
Best Time-decayed ε-greedy	21.7625 ± 0.0646
Fixed time	24.21 ± 0.0364

- Average exploration rate comparisons and uncertainty during training:



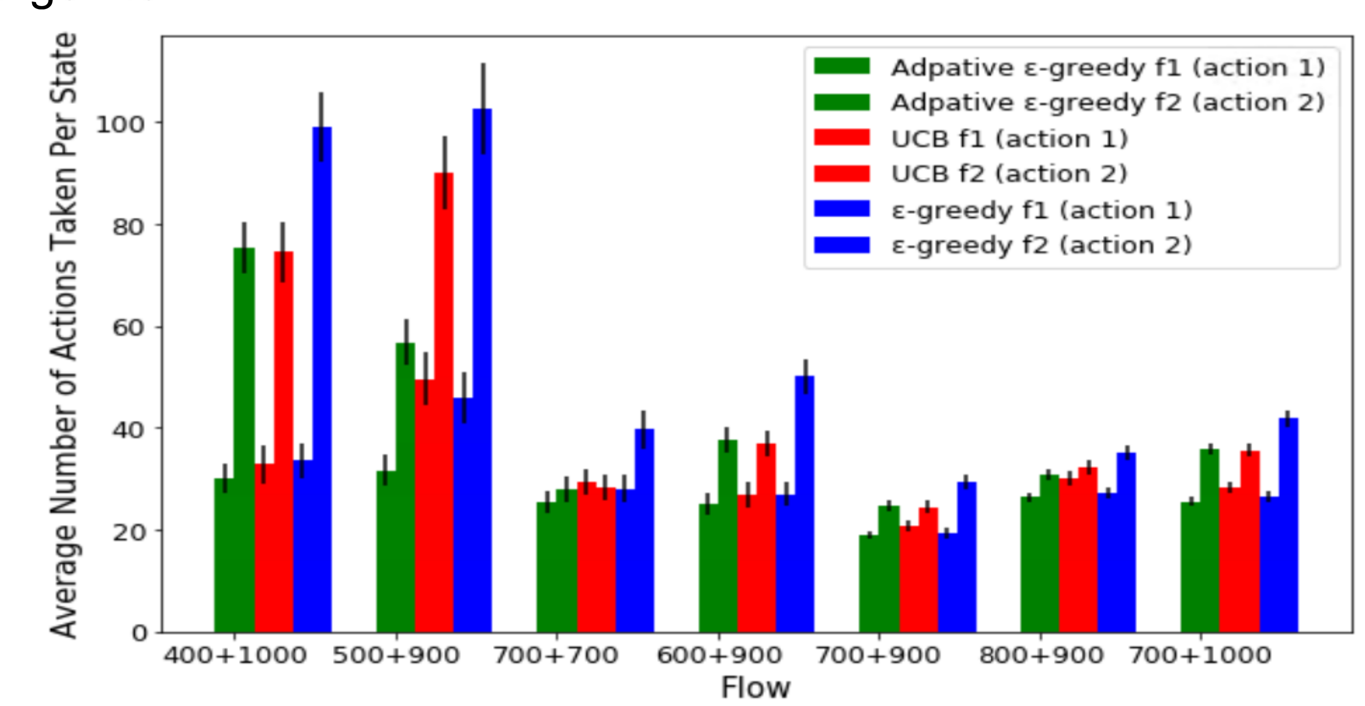
Experiments results

Numerical Comparisons among Different Strategies

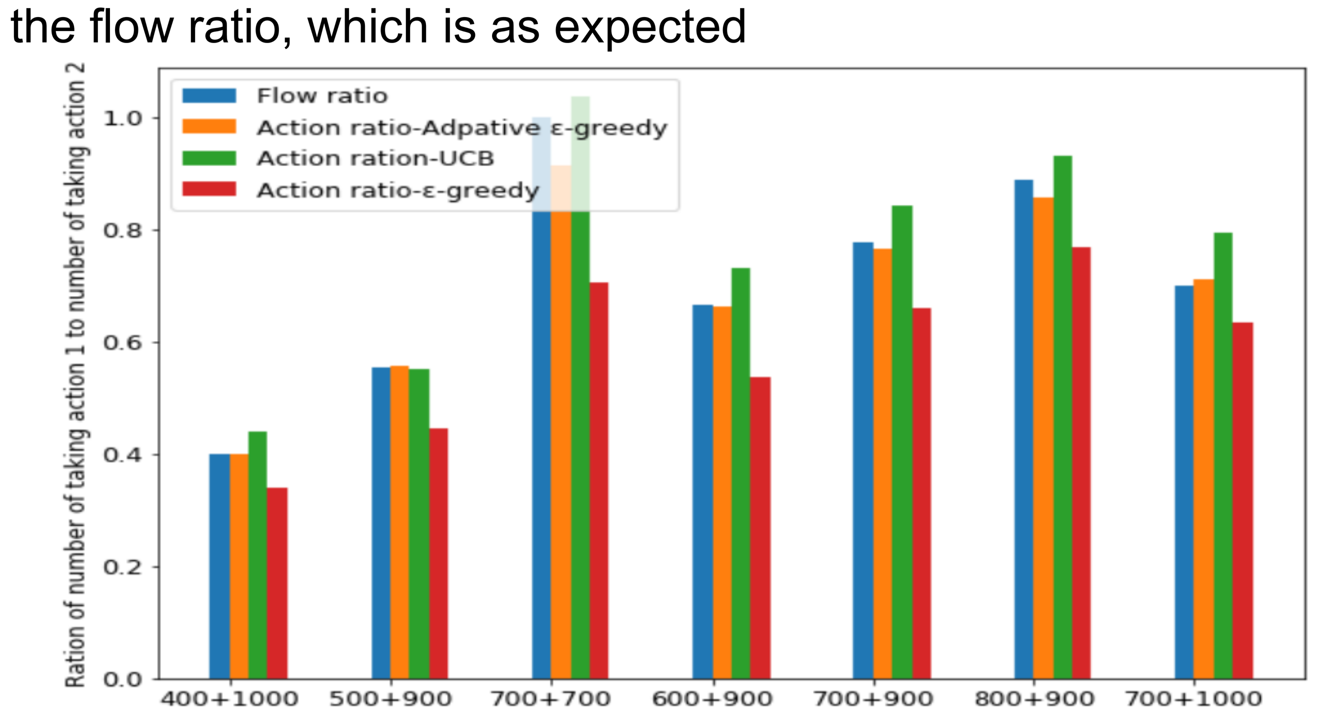
- Our adaptive strategy could explore more unique state under different scenarios:

Model	Flow (Veh/h/lane/direction)						
	400×1000	500×900	700×700	600×900	70×900	800×900	700×1000
Adaptive ε-greedy	16897	28342	46041	39127	53960	62578	58918
UCB	15949	24206	41205	37104	50449	57037	54903
ε-greedy	13258	16630	36595	32198	48342	56283	52873

- Agent takes action 2 (EW green) more frequently compared to taking action 1 (NS green) for all unbalanced flow setting where EW has larger flow



- Action ratio of the adaptive ε-greedy strategy is always the closest one to the flow ratio, which is as expected



Concluding Remarks

- Addresses the issue of adaptive exploration when using RL for traffic signal control, considering the agent's uncertainty for the environment
- Advantages of the adaptive ε-greedy strategy with double Q-learning:
  - Outperforms baselines and results in lower delays and highly compatible under different scenarios this approach
  - Explore more states with no hyper-parameters needed to be tuned
- The policy learned by the agent keep the action ratios consistent with the flow ratio, which leads to low delays
- Work only tested on a simple intersection setting; future work could apply the model to an intersection with right turns and lest turns