A Distributed Deep Reinforcement Learning Based Integrated Dynamic Bus Control System in a Connected Environment TRBAM-22-00882
Haotian Shi, Qinghui Nie, Sicheng Fu, Xin Wang, Yang Zhou*, Bin Ran

INTRODUCTION

Background
- Bus bunching problem refers to a group of two or more buses, initially planned to be regularly spaced along the same route, arrive at the same bus stop simultaneously due to travel time and passenger demand uncertainty.

- Efficiently controlling buses to solve bus bunching.

Research Gap
- Most previous methods utilize a single method to control buses (e.g., bus holding; speed operation; signal priority).
- Considers a single type of uncertainties and treating the uncertainties with an analytical distribution.
- Suffers heavy computation load by exactly solve a constrained stochastic optimization

Deep reinforcement learning
- Suitable for capturing complex and stochastic system characteristics.
- Rapidly implemented in online process.

Equilibrium and consensus concept
- Regulate bus close to the pre-defined equilibrium point.
- Gives DRL an exploration direction in the training process to improve the convergence.
- Consensus as a crucial property of multi-agent networks can effectively impede the accumulation of disturbances.

Objectives
- Inverse multiple continuous-variable-based control methods to form a generic bus control strategy.
- Utilize adequate historical and real-time traffic information to make the control algorithm more efficient.
- Incorporate the merits of DRL equilibrium concept, and consensus concept to enhance system robustness.

SYSTEM DESCRIPTION

DRL-based algorithm design (MDP Scheme)
- State representation $s_t = [e^t, d^t, b^t, h^t]$
  - $e^t$: schedule deviation; $d^t$: weighted headway deviation
  - $b^t$: actual dwell time
- Action representation $u_t = \pi_\theta(s^t)$, representing time adjustment at each station.
- Policy representation $\pi_\theta$ $\pi_\theta$ depends on the neural network parameter $\theta$, which assigns probabilities to every possible action $a^t_i$ given each state $s^t_i$. $\pi_\theta$ is improved to $\pi_\theta^*$ after training.
- Reward (opposite of stage cost) representation $r^t_i$
  - Multiple objectives: (i) schedule deviation $e^t$; (ii) weighted headway deviation $d^t_i$; (iii) control force $u^t_i$
  - Stage cost: $c_i = c_1 \cdot \text{Q} \cdot c_1 = [e_i^t, d_i^t, u^t_i]$.
  - Reward: $r^t_i = \exp(-c_i^t)$
- $Q_i$ is a diagonal matrix with three objective coefficients.

Distributed Proximal Policy Optimization
- DDPPO algorithm for policy iteration in the training process, which incorporates an actor network and critic network.
- Parallel version of the PPO algorithm
- Actor network
  - Responsible for selecting actions based on its policy parameter $\theta$. Objective function:
    $L_{\text{Actor}}(\theta) = E_i \cdot (\ln(p_i(\theta, x_i) \cdot c_i \cdot \text{clip}(p_i(\theta, 1 - x_i), 1 + e, 1 + e) \cdot A_i))$
  - Use the clipping method to make the policy gradient less sensitive to the step size
- Critic network
  - Responsible for evaluating the current policy $\pi_\theta$ with parameter $\Phi$. Objective function:
    $L_{\text{Critic}}(\phi) = E_i \cdot (R_i - V_\phi(s_i))^2$, estimate advantage function
    $A_i = R_i - V_\phi(s_i)$ for the actor network.

Training procedure and results

CONTRIBUTIONS
- Developing a distributed DRL-based integrated dynamic bus control method to efficiently solve the bus bunching problem.
- Better utilizing the bus historical and real-time traffic information by incorporating these stochastic characteristics into the DRL environment.
- Incorporating the merits of the equilibrium concept and consensus control theory in the DRL framework, which effectively prevents errors amplifying.