Cooperative Bus Holding and Stop-skipping: A Deep Reinforcement Learning Framework

Joseph Rodriguez[a,*], Haris N. Koutsopoulos[a], Shenhao Wang[b], Jinhua Zhao[b]

[a]*Northeastern University, 360 Huntington Ave, Boston, 02115, MA, United States*
[b]*Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, 02139, MA, United States*

Abstract

The bus control problem that combines holding and stop-skipping strategies is formulated as a multi-agent reinforcement learning (MARL) problem. Traditional MARL methods, designed for settings with joint action-taking, are incompatible with the asynchronous nature of at-stop control tasks. On the other hand, using a fully decentralized approach leads to environment non-stationarity, since the state transition of an individual agent may be distorted by the actions of other agents. To address it, we propose a design of the state and reward function that increases the observability of the impact of agents' actions during training. An event-based mesoscopic simulation model is built to train the agents. We evaluate the proposed approach in a case study with a complex route from the Chicago transit network. The proposed method is compared to a standard headway-based control and a policy trained with MARL but with no cooperative learning. The results show that the proposed method not only improves level of service but it is also more robust towards uncertainties in operations such as travel times and operator compliance with the recommended action.

Keywords: Real-time Bus Control, Multi-agent Reinforcement Learning, Cooperative Agents

*Corresponding author at: 110 Forsyth St, Snell Engineering, Boston, MA 02115
Email address: rodriguez.josep@northeastern.edu (Joseph Rodriguez)

1. Introduction and Background

High-frequency bus operations play a crucial role in increasing economic opportunity and social equality for densely populated areas. However, the lack of adequate control can result in poor quality of service and higher operational costs, especially when buses run in mixed traffic and without traffic signal priority. The control problem consists of addressing, in real-time, the service irregularities that emerge from the combination of stochastic travel times and demand conditions, which cause vehicles to fall behind schedule and distort the frequency of bus arrivals at stops. Some consequences include increased passenger wait times, crowding in buses, and increased risk of delays for subsequent trips (Tirachini et al., 2021).

The literature on the problem is extensive and solutions have been refined as technological improvements have taken place. In the earlier days, schedule-based holding strategies were predominant in the literature and real-world applications due to their simplicity and minimal real-time information requirements (Turnquist, 1974). They consist of holding buses at designated stops so they do not depart before its scheduled departure time. Headway-based control in turn, which corrects for headway regularity instead of schedule adherence, became an increasingly practical alternative for high-frequency services with the advent of Automated Vehicle Location (AVL) technology and wireless communication (Daganzo, 2009; Eberlein et al., 2001; Bartholdi III and Eisenstein, 2012).

Despite the extensive literature on the problem, there is still room to refine the state-of-the-art control methods to make better use of available technologies and frameworks. Current real-time control methods rely on limited information such as a bus trip's headway, which is attractive for practical uses, but fails to consider the complex costs and benefits associated with each alternative control decision. For instance, holding a vehicle at a stop increases the travel time of the passengers on board, increases the wait of certain passengers at downstream stops, and decreases the wait time and crowding of other passengers at downstream stops. Zhao et al. (2003) proposed a real-time approximation of such cost-benefit function to be used as criterion for holding. Quantifying these costs, however, requires simplifications, due to their interdependence and the randomness inherent to the system.

The solution proposed in this study defines the bus control task as a Partially Observable Markov Decision Process (POMDP) and uses multi-agent

reinforcement learning (MARL) to develop a control policy. The motivation to use MARL is that even though the control policy is defined using information that is available in real-time, the action selected depends on prior learning of the alternative actions' cost-benefit values. Having a pre-trained policy is also suitable for practical applications.

However, the difficulty in applying the standard MARL approach, in which the agents' actions are chosen and evaluated jointly, is the decentralized and asynchronous nature of the bus control task. The challenge is therefore to extend the framework to control the agents (buses) independently but evaluate their actions jointly. The information used for action evaluation is discussed and justified, as it is critical to avoid unnecessary information which can add noise to the reward and lead to sub-optimal control policies. The resulting framework, by applying these considerations in the extension of the state and reward definitions, allows for cooperative learning.

Previous MARL approaches to this problem, although effective, have focused on the holding control strategy exclusively. This study adds stop-skipping, a strategy that allows vehicles to skip service at certain stops, as an alternative in the set of possible actions. Such addition allows to develop hybrid strategies that can mitigate bunching more effectively, by skipping and holding the late and early buses respectively. A generalized reward function based on passenger wait time is designed to appropriately account for the impact of both holding and stop-skipping actions.

The paper compares the performance of the proposed cooperative MARL framework against a rule-based and a simplified MARL approach drawn from the literature. We use simulation as the training environment for the MARL-based methods and as the test-bed for performance analysis. The simulation model is calibrated to replicate the rush hour operation of a bus route in Chicago. To analyze transferability and simulation-to-reality gap concerns in the MARL approach, we analyze how performance is impacted by conditions that were not experienced during training; and how performance can be improved by re-training with the updated conditions. In this analysis we evaluate the effect of driver non-compliance with control instructions, given its potential to hinder performance (Phillips et al., 2015) and the unpredictability of such behavior prior to initial field deployment.

In summary, the main contributions of the paper are:

- Proposes a MARL framework to learn a bus control policy based on available real-time data to reduce system-wide passenger wait times

and crowding. Furthermore, the method shows robustness to uncertain driver compliance.

- This methodology addresses the non-stationarity problem inherent to MARL-based bus control by extending the definition of the state and reward to allow for cooperation and awareness among agents during training.

- To the best of our knowledge, it is the first attempt at formulating a generalized deep MARL framework for combined holding and stop-skipping strategies.

The remainder of the paper is organized as follows: section 2 discusses previous work, section 3 formulates the problem, section 4 presents the proposed framework, section 5 describes the case study, section 6 presents and discusses the experiments' results, and section 7 presents concluding remarks.

2. Related Studies

In this section, we review the relevant literature on bus holding and stop-skipping. The approaches range from rule-based, to optimization, to reinforcement learning (RL). An overview of the existing bus control literature is presented in Table 1.

Table 1: Overview of literature in real-time bus control strategies

Author	Strategy			Method			Objective	
	Holding	Skipping	Combined	Rule-based	Optimization	RL	Regularity	Pax cost
Turnquist (1974)	✓			✓			✓	
Barnett (1974)	✓				✓			✓
Turnquist et al. (1980)	✓				✓		✓	✓
Abkowitz et al. (1986)	✓			✓			✓	
Rossetti and Turitto (1998)	✓			✓			✓	
Li et al. (1995)		✓			✓			✓
Eberlein (1995)	✓	✓	✓		✓			✓
Eberlein et al. (2001)	✓				✓			✓
Fu et al. (2003)		✓			✓			✓
Zhao et al. (2003)	✓				✓			✓
Sun and Hickman (2005)		✓			✓			✓
Koutsopoulos and Wang (2007)	✓			✓			✓	
Daganzo (2009)	✓				✓		✓	
Delgado et al. (2009)			✓		✓			✓
Cats et al. (2011)	✓			✓			✓	
Delgado et al. (2012)	✓		✓		✓			✓
Bartholdi III and Eisenstein (2012)	✓				✓		✓	
Sáez et al. (2012)			✓		✓			✓
Chen et al. (2015a)	✓					✓	✓	
Laskaris et al. (2016)	✓			✓			✓	✓
Gao et al. (2016)		✓			✓			✓
Chen et al. (2016)	✓					✓	✓	
Zhang and Lo (2018)	✓				✓		✓	
Alesiani and Gkiotsalitis (2018)	✓					✓	✓	✓
Menda et al. (2019)	✓					✓	✓	
Wang and Sun (2020)	✓					✓	✓	
Saw et al. (2020)			✓			✓	✓	
Wang and Sun (2021)	✓					✓	✓	
Zhang et al. (2021)			✓		✓			✓

2.1. Holding

Bus holding strategies have been studied extensively. They can be categorized primarily as schedule-based, which are most appropriate for low-frequency bus operations (Turnquist, 1974), and headway-based, which are more effective for high-frequency services (Abkowitz and Lepofsky, 1990; Cats et al., 2011). Most of the earlier research focused on threshold-based holding (Turnquist, 1974; Turnquist et al., 1980; Abkowitz et al., 1986; Rossetti and Turitto, 1998). Several studies proposed finding the optimal threshold values to balance wait time savings and imposed delay from holding (Abkowitz and Tozzi, 1986; Abkowitz et al., 1986). The even headway strategy, which regulates departures based on preceding and following vehicles, has shown effectiveness in simulation-based studies focused on train systems (Koutsopoulos and Wang, 2007) and bus routes (Cats et al., 2011). The even headway rule can also be extended to incorporate passenger travel costs associated with holding, as proposed in Laskaris et al. (2016).

Eberlein et al. (2001) first formulated the optimal holding problem with real-time information by using optimization with rolling horizon, assuming deterministic passenger arrivals and travel times. Zhao et al. (2003) proposed a dynamic decision rule, based on passenger time impacts for current and downstream stops. As it is designed for real-time operations, the algorithm relies on simplifications and neglects the downstream stops that have not yet been served by the previous bus. Daganzo (2009); Bartholdi III and Eisenstein (2012) used dynamic programming models to apply holding and bus speed adjustments in response to headway disturbances. Zhang and Lo (2018) extended the approach in Bartholdi III and Eisenstein (2012) to account for the forward headway of the controlled bus in determining holding time. Sánchez-Martínez et al. (2016) proposed a rolling horizon optimization approach.

2.2. Stop-skipping

To reduce its run time, a bus can be allowed to skip one or more stops, either entirely or after allowing for alightings only. Examples include expressing (skip several stops in sequence), deadheading and short-turning (end service mid-route in one direction to begin mid-route in the opposite direction). Stop-skipping strategies have been proposed mostly at the planning level (scheduled) and formulated as optimization problems. Liu et al. (2013) considered strategies for stop-skipping and deadheading as an optimization

problem based on passenger times and operator costs, and evaluated their robustness to random travel time fluctuations. Chen et al. (2015b) proposed a headway optimization problem when a fraction of trips are allowed to express and skip stops. To address the real-time problem, Li et al. (1995) proposed a model for short-turning and skipping based on bus location and evaluated its performance using simulation. Stop-skipping has also been studied as a strategy for incident recovery (Sun and Hickman, 2005; Gao et al., 2016).

2.3. Combined holding and stop-skipping

Although holding and stop-skipping serve the common purpose of increasing headway regularity, the costs associated with each are different. Holding increases the in-vehicle time of passengers on board and wait time of some passengers at downstream stops, while skipping, depending on its implementation, increases wait time for the passengers waiting at the skipped stop. There are limited studies that explore them jointly as a hybrid strategy. Delgado et al. (2009, 2012) proposed an optimization framework with rolling horizon prediction for combined holding and boarding limits. The boarding control, at its limit (0 boarding passengers) can be seen as a version of stop-skipping. The results favor the combined strategy compared to holding-only in terms of average wait time and highlight the positive impact of boarding limits in high-frequency services when the bus behind is sufficiently close. Sáez et al. (2012) formulated the hybrid strategy as an optimization problem that minimizes passenger-related time costs and headway irregularity. The results also support the superiority over the holding-only strategy. Zhang et al. (2021) developed a framework for real-time control combining holding and stop-skipping. The decentralized nature of the formulation, though, limits the framework's ability to develop coordinated control actions. Another study applied RL for combined holding and stop-skipping in the context of a shuttle operating in a loop (Saw et al., 2020).

2.4. Multi-agent Reinforcement Learning for Bus Control

MARL methods are used for learning optimal control tasks with multiple agents. Extending traditional MARL methods to the bus control problem introduces challenges due to the fact that buses are controlled asynchronously upon arrival to a stop. Previous MARL applications with asynchronous control tasks include elevator group control (Crites and Barto, 1998) and maritime logistics (Li et al., 2019). In such cases, the state transition of each agent may be influenced by a neighboring agent's action, causing the

environment to appear non-stationary, from the perspective of each agent (Laurent et al., 2011). Environment non-stationarity, if not addressed, can be detrimental to the learning process.

Previous MARL applications to bus control differ significantly in their approach. Alesiani and Gkiotsalitis (2018) used a fully decentralized learning scheme and neglected the effect of non-stationary conditions. A range of solutions have been proposed to reduce non-stationarity in the bus control task. Chen et al. (2015a, 2016) used sparse-cooperative Q-learning that selectively includes global state information in the learned policy. Although this approach allows for learned cooperative strategies, it increases computational complexity. Wang and Sun (2020) used the actor-critic framework and incorporated a joint action tracker for agents to select actions based also on the expected actions of other agents. Menda et al. (2019) formulated the bus control problem as a discrete-event MARL problem, in which each agent accumulates global rewards in between control events. The rewards are then processed as a single delayed reward at the latter control event. A problem with such solution is that it treats the reward contribution from other agents as equal, no matter how distant they are from the learning agent, which may result in unnecessary noise in the reward. To address this problem, Wang and Sun (2021) introduced a credit assignment framework based on graph learning of bus trajectory plots near the controlled bus during the control event. The approach resulted in reduced variability in headway and load. The results also show potential to generalize to other bus routes without need for re-training.

2.5. Summary

The main observations from the literature review are:

- Studies on holding exhibit trade-offs between making use of real-time information and ease of implementation.
- Most studies on stop-skipping are planning studies where skipping is part of the schedule.
- Limited studies have analyzed stop-skipping and holding as a joint strategy.
- The reinforcement learning frameworks proposed have not addressed adequately the non-stationarity issue, and often incorporate global in-

formation which increases computational expense and potentially adds noise to the reward.

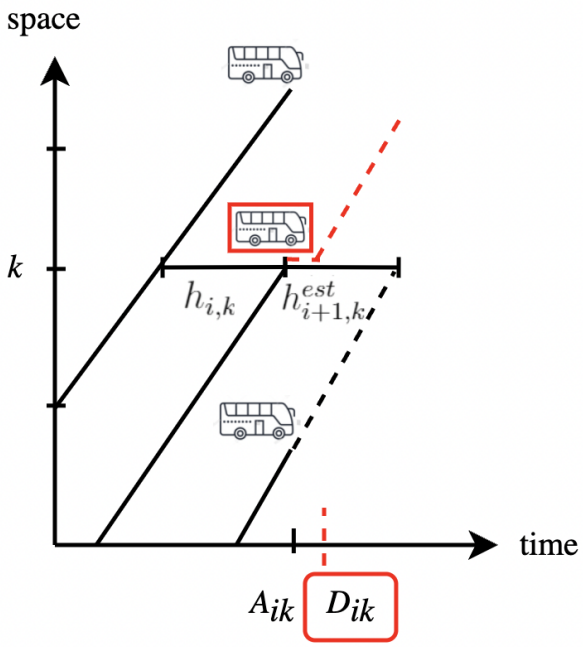

Figure 1: Bus control event in spatio-temporal plot

3. Problem Statement

Consider a bus line with m stops and n trips. The bus line includes c designated control stops, where $c \leq m-1$ (the final stop m is not considered for control). When bus performing trip i arrives at control stop k at time $A_{i,k}$, its departure time $D_{i,k}$ can be delayed, in order to correct for uneven headways between consecutive bus trips. In this paper, we define the forward headway h_{ij} as the inter-arrival time between trip i and the preceding trip $i-1$ at stop j. The backward headway of bus trip i at stop j is the predicted forward headway of bus trip $i+1$, denoted by $h^{est}_{i+1,j}$, based on the predicted arrival time of the following bus at the same stop, $A_{i+1,j}$. The prediction is necessary for control decisions made in real time.

The control actions considered for this problem are holding and stop-skipping. Holding control consists of delaying trip i's departure from stop k by holding it for time T_{ik}^{hold}. The departure time is then determined by:

$$D_{ik} = A_{ik} + \max\{T_{ik}^{hold}, w_{ik}\} \tag{1}$$

w_{ik} is the dwell time at the stop. When stop-skipping, the vehicle departs the stop immediately after the alighting time without boarding of passengers. The departure time is then:

$$D_{ik} = A_{ik} + w_{ik}^{a} \tag{2}$$

In this case, the dwell time w_{ik}^{a} corresponds to alightings only. Skipping is only allowed if the following conditions are met: a) the stop is not the origin terminal and b) the preceding bus did not skip the stop (to avoid passengers being denied twice).

4. Solution Approach

The real-time bus control problem can be defined as a multi-agent system with asynchronous control events, since buses arrive at control stops at irregular times. The complexities of such problem arise from the difficulty in quantifying the impact of a single control action in a system with dynamic conditions and actions taken by other agents. For instance, a holding action impacts the wait time of passengers on board the controlled bus agent and of certain passengers at downstream stops. But this impact is not available until the action has already been taken. It is complex to estimate the impact a priori, due to compounding randomness in passenger demand and travel times, as well as control actions taken by preceding buses. Thus the effectiveness of the action is measured by a proxy variable such as the headway, or by a simplified approximation of the wait time/ride time costs. These simplifications, though leading to more generalizable methods, result in control policies that are less responsive to dynamic conditions and actions from other agents.

To address such challenges, the MARL approach is attractive because it is designed to find approximate solutions to control problems. The approximate solution is a control policy that is a function of observable real-time information at the control event time, but is optimized based on immediate and delayed system feedback received from past experiences. The MARL

approach can efficiently use newly generated experiences to gradually guide the agents to act optimally based on predefined metrics. In this section we model the bus control problem as a Markov Decision Process to be solved with the MARL framework.

4.1. Bus Control as a Markov Decision Process

A Markov Decision Process (MDP) for the single-agent RL case can be formally defined as $(\mathcal{S}, \mathcal{U}, \mathcal{R}, \mathcal{P}, \gamma)$, where $\mathcal{S}$ is the set of states that can be encountered when the agent interacts with the system, $\mathcal{U}$ the set of possible control actions that can be taken, $\mathcal{R}$ the reward function based on action $u \epsilon \mathcal{U}$ taken at state $s \epsilon \mathcal{S}$ and resulting in state s'; $\mathcal{P}$ is the transition probability function which determines the probability that state s' will result, given current state s and action u (i.e. $P(s \rightarrow s'|s, u)$), and γ is the discount factor which discounts the value of future rewards compared to immediate rewards.

The extension of the MDP for a generic MARL problem requires defining a set of agents and a joint state/action set. However, the bus control process is asynchronous and event-driven, triggered by stop arrivals, so the notion of agents observed and controlled jointly is not appropriate. This case can be represented as a Partially Observable Markov Decision Process (POMDP) (Oliehoek and Amato, 2016), in which multiple agents are controlled in a decentralized fashion, acting independently on their own state.

A concern about framing this problem as a POMDP is non-stationarity, given that for agent i, the reward received from action u taken at state s and resulting to state s' may be impacted by another agent's action, taken at some time between s and s'. This makes the environment non-Markovian, which negates the guarantee of convergence to an optimal policy, unless following certain guidelines to mitigate the issue (Laurent et al., 2011). This issue is overlooked even in non-RL bus control methods. For instance, if the method involves real-time information such as forward and backward headways, these are computed assuming that the buses performing the preceding and following trips are not controlled.

Generally, there are two options to address the non-stationarity problem: a) reduce the likelihood of an agent's action impact on another, or b) make the impact more observable to the trained agents (Laurent et al., 2011). The former is not considered, since buses in a high-frequency service inevitably impact each other. The latter, however, can be incorporated by extending

the state definition to include more information about the relevant neighboring agents, and to broaden the scope of the reward function to include the controlled agent's impact on the relevant neighboring agents. As a result, the risk of non-convergence of the policy is reduced and cooperative strategies can emerge.

The basic components of the MARL framework and the training algorithm adapted for the solution of the bus control problem studied in this paper are described in the sections that follow.

State

The state of the controlled bus trip i upon arrival at control stop k is composed of information relevant to make and evaluate control decisions. The choice of state parameters determines how the learning of experiences is categorized in the updating of the control policy. The parameters included are as follows:

- *Spatial*: The impact of an action depends on the location of the control stop and the distance to the next control stop. Therefore the bus location is included, represented by the control stop number normalized by the total number of control stops, i.e. k/c.

- *Regularity*: To account for the bus trip's regularity status, the forward headway h_{ik} and backward headway $h^{est}_{i+1,k}$ are used.

- *Demand*: A count of the passengers that would be immediately impacted by the action is included, namely the passenger load when arriving at stop k reduced by the number of alighting passengers, $\mathcal{L}_{ik}-\mathcal{G}_{ik}$, and the passengers ready to board, $\mathcal{B}_k$. The actual number of passengers waiting to board is typically not known in real time. In theory, it may be computed from the number of fare-box transactions, however, this information is typically not available in real time. Hence, it is estimated from the historical arrival rate and the forward headway.

- *Preceding bus trip information*: As mentioned earlier, to address the non-stationarity problem our approach includes some information of the relevant neighboring agents to make the controlled agents more aware of their surrounding agent's status. We include only the preceding bus trip's forward headway $h_{i-1,k}$. Other bus trips are not considered to avoid increasing the complexity of the state and thus adding

noise to the training process. However, this topic can be explored in future studies.

Lastly, we use the line's end terminal (stop m) as the terminal state to capture the impact of previous actions on the schedule status at the end of the route.

Actions

Upon a bus agent's arrival at a control stop, three high-level control actions can be taken (as described in section 3): departing after boarding and alighting take place (no control), departing after alighting time (stop-skipping), and departing after the holding time or boarding and alighting time, whichever is larger (holding). Specifically, the set of actions $\mathcal{U}$ is composed by:

- The holding time $T_{ik}^{hold} = \omega H_i$ where $\omega \epsilon \Omega$ is a strength parameter from a set Ω with values in $(0, \omega^{max})$ and H_i is the scheduled headway of bus i. The upper limit on holding time is set by ω^{max}. The corresponding holding time is used to determine the departure time from the stop, using Eq. (1).

- A skipping action instructing the driver to depart the stop, after allowing for alightings if needed. The departure time is then determined by Eq. (2).

Reward

The design of the reward function is critical for the agent's learning of optimal control actions. It should be both broad enough to capture the impact of the action, but specific enough to not cause noise during learning. Most RL applications to bus control, which do not include skipping as a possible action, use as reward the resulting impact on the difference between actual and planned headway. To consider the impact on the onboard riders, either a penalty for too much holding is included (Alesiani and Gkiotsalitis, 2018; Wang and Sun, 2020, 2021) or the holding time is constrained. In this way, the benefits and costs are captured in an immediate reward to the agent.

In this study, the proposed reward function $r_{i,k+1}$ is based on a fundamental metric for the effectiveness of control: passenger wait time at stops and the impact of holding on passenger ride time. A similar metric to wait

time was also used in a previous application of MARL to elevator group control (Crites and Barto, 1998). Additionally, by introducing delayed rewards, as used in an application of MARL to resource balancing in logistics (Li et al., 2019) and bus control (Menda et al., 2019), the reward captures more accurately the impact of an action between control events.

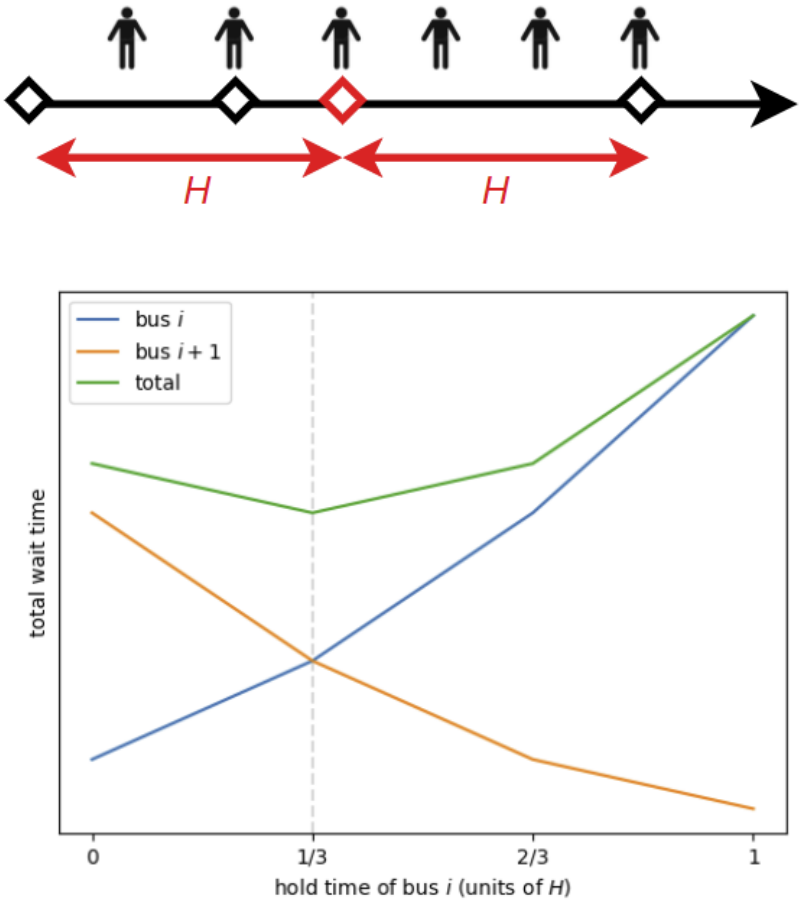

Figure 2: Impact of holding bus i on the wait time of passengers on buses i and $i+1$

To justify the wait time component of the reward function, let us consider an example of a bus service with scheduled headway $\mathcal{H}$ minutes and passenger arrival rate of $\frac{3}{\mathcal{H}}$ passengers per minute. For this example, passenger arrival rates are deterministic. Bus $i-1$ arrives at stop k on schedule at time t_0,

bus i arrives early at time $t_0 + \frac{2\mathcal{H}}{3}$, and bus $i+1$ is predicted to arrive on time at time $t_0 + 2\mathcal{H}$. The total wait time for the passengers boarding bus i and bus $i+1$, assuming bus $i+1$ arrives at the predicted time, as a function of the bus holding time is shown in Figure 2. Delaying the departure of bus i by $\frac{\mathcal{H}}{3}$ (with departure time at $t_0 + \mathcal{H}$, the midpoint between the departures of buses $i+1$ and $i-1$) results in the minimum total wait time. This is the basis for even-headway control methods. However, forcing a departure from control stop k at the midpoint does not guarantee arrival at equal headways at the intermediate stops until the next control stop $k+1$, due to travel time uncertainty. A more appropriate metric for the effectiveness of the action is the total wait times of all boarding passengers at all intermediate stops for buses i and $i+1$. Additionally, this gives proper assessment of actions such as stop-skipping, in which the denied boardings of trip i result in increased wait time for trip $i+1$ boardings.

As such, the reward function is defined as the sum of wait times of the set of passengers boarding buses i and $i+1$ at stops between control stops k (inclusive) and $k+1$ (exclusive), as follows:

$$r_{i,k+1}^{wait} = -\sum_{y=i}^{i+1}\sum_{z=k}^{k+1}\sum_{p\epsilon P_z(D_{y-1,z},A_{yz})} T_p^{wait} \tag{3}$$

$P_z(D_{y-1,z}, A_{yz})$ is the set of passengers arriving at stop z between the departure time of trip $y-1$ and the arrival time of trip y; T_p^{wait} is the wait time for passenger p. The negative sign is added to penalize higher wait times.

The second term in the reward function $r_{i,k+1}^{ride}$ is used to capture the increased ride time experienced by passengers as a result of holding. In addition to the holding time at stop k, passengers experience delay from increased dwell times downstream as a consequence of more passengers arriving in greater headways. Therefore, we add as reward the ride time experienced by passenger $p\epsilon P$ within the time interval $(A_{ik}, A_{i,k+1})$, where A_{ik} and $A_{i,k+1}$ are the arrival times of trip i at stops k and $k+1$, respectively. P is the set of passengers that were on-board trip i at any point in the time interval $(A_{ik}, A_{i,k+1})$.

$$r_{i,k+1}^{ride} = -\sum_{p\epsilon P} T_p^{ride}(A_{ik}, A_{i,k+1}) \tag{4}$$

Where $T_p^{ride}(A_{ik}, A_{i,k+1})$ is the ride time of passenger p measured within the

interval $(A_{ik}, A_{i,k+1})$. The final reward function is thus:

$$r_{i,k+1} = W_{wait} r^{wait}_{i,k+1} + r^{ride}_{i,k+1} \tag{5}$$

The factor W_{wait} is used to weigh the wait time component relative to the ride time component. It should be noted that the reward for bus agent i can only be computed after bus $i+1$ has arrived at control stop $k+1$. The then completed experience tuple $(s_{ik}, u_{ik}, r_{i,k+1}, s'_{i,k+1})$ is used to update the policy.

4.2. Training Algorithm

We describe two approaches for training the RL algorithm, which we combine in our solution to address the discrete-event and high-dimensional characteristics of the problem.

Discrete-Event Q-learning

The agents are trained with a discrete-event version (Bradtke and Duff, 1995) of the Q-learning algorithm (Watkins, 1989). Q-learning is a model-free reinforcement learning method which uses dynamic programming to iteratively update the value of an action u taken at state s, e.g. $Q(s, u)$. Upon trying all actions repeatedly, the action with highest Q-value is judged as the best action overall for that state, considering immediate rewards and discounted long-term rewards. This framework was designed, however, for environments with fixed time-step between control actions, which is not the case for bus control. Therefore, we use the Q-learning update rule adapted to the discrete-event RL proposed in (Bradtke and Duff, 1995):

$$\Delta Q(s, u) = \alpha \cdot [r + e^{-\beta t} Q(s', u') - Q(s, u)] \tag{6}$$

where

$$u' = \arg\max_{u} Q(s', u) \tag{7}$$

α is the learning rate and t is the time between events represented by states s and s'. The only difference from standard Q-learning is the term $e^{-\beta t}$ which acts as a variable discount factor that scales with the duration between events, with β controlling the rate of exponential decay.

Double Deep Q-learning

In order to accommodate the high dimensionality of the state set defined earlier, multi-layered neural networks are used to store and estimate the Q-values. For training, Mnih et al. (2015) formally introduced deep Q-learning with techniques that improve sample efficiency and learning stability. With Deep Q-learning, the state-action value $Q(s, u)$ is estimated by a network with parameters θ that are updated every time-step, referred to as the online network; the value of the next state $Q(s', u')$ is estimated by a network with parameters θ' which are copied periodically from θ, referred to as the target network. In this paper, we use an extension of Deep Q-learning named Double Deep Q-learning (DDQN) (van Hasselt et al., 2015). In this variation, the online network's parameters θ are also used to estimate the next state's action u' in the update rule (see Eq. (7)). The adapted update rule for DDQN is as follows:

$$\theta = \theta + \alpha \cdot (r + e^{-\beta t} Q_{\theta'}(s', u') - Q_\theta(s, u)) \tag{8}$$

where

$$u' = \arg\max_u Q_\theta(s', u) \tag{9}$$

Every τ updates of the online network, its parameters are copied onto the target network, $\theta' := \theta$. This modification helps reduce the overoptimism inherent to standard Q-learning due to using the same network for action selection and evaluation (van Hasselt et al., 2015).

The networks used are fully connected neural networks (FCNN) with N^{layers} hidden layers of size $L^{network}$ and with ReLu activation, while the last layer uses linear activation for the mapping of discrete action values. Each update of the FCNN is made on a batch of L^{batch} experiences, drawn randomly from the experience replay buffer (Lin, 1992) which stores the most recent L^{buffer} experiences. Training follows an ϵ-greedy exploration with ϵ annealed linearly from ϵ_0 to ϵ_{final} for the first N_1^{eps} and fixed on ϵ_{final} for the last N_2^{eps}.

4.3. Training environment: Bus Simulation

The agents must be trained in an environment that captures the dynamic conditions of bus operations. To that effect, a simulation model was developed. Scheduled departures and blocks (sequence of trips assigned to a bus) for both directions are inputs to the simulator. This representation allows to

capture the propagation of delays between consecutive trips by the same bus. Dwell times at stops are a function of the number of boarding and alighting passengers. If a bus is holding, arriving passengers can board the vehicle. Buses have capacities and passengers who are unable to board a bus because of lack of available space must wait for the next vehicle. Passenger arrivals at stops are generated from OD trip demand rates, which follow a Poisson distribution.

4.4. Updating the Q-Network

Algorithm 1 summarizes the computational workflow of the training process. In a new episode the iteration step begins with advancing the simulation until the next control event, $NextControlEvent()$. As mentioned earlier, the reward for bus $i-1$ from action at stop $k-1$ is computed only after bus i arrives at stop k. The completed experience is then added to the replay buffer, from which random experiences are drawn to update the network parameters. Finally, the updated network is used to select the action that bus i takes in the current event.

Algorithm 1 Event-driven MARL Training

for episode 1 to $N_1^{eps} + N_2^{eps}$ **do**
  $ResetEnvironment()$
  while environment is not terminated **do**
    // Advance environment to next control event
    $s_{ik} \leftarrow NextControlEvent()$
    if $i > 1$, $k > 1$ **then**
      Update reward for agent $i-1$ (Eq. (5))
      Add $(s_{i-1,k-1}, u_{i-1,k-1}, r_{i-1,k}, s_{i-1,k})$ to replay buffer
      // Update policy parameters
      Sample a random batch of (s, u, r, s') from replay buffer
      Update Q-Network using update rule (Eq. (8))
    end if
    // Action selection for agent i
    $u_{ik} \leftarrow \epsilon - Greedy(\arg\max_u Q_\theta(s, u))$
    Take control action u_{ik}
  end while
end for

5. Case Study

5.1. Experiment Description

Bus route 20 in Chicago, which connects the western residential region to the city center, including major transit connections (Figure 3), is used for the case study. The line spans 14 km and includes 67 stops in the eastbound direction (ED) and 63 stops in the westbound direction (WD). During the morning peak, the WD service pattern involves deadheading, given the lower demand, and thus most of the stops are served with lower frequency. For this reason, control is only applied to the ED direction. The studied service period is the morning peak between 7 and 9am. The scheduled ED running time during this period is 62 min with 8 min of recovery time. The scheduled headway is 4-6 min. Vehicle capacity is 50 passengers.

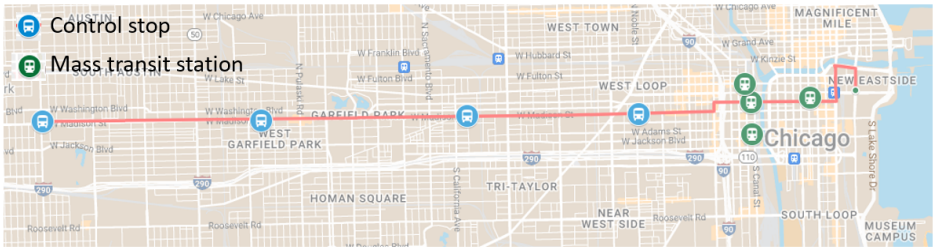

Figure 3: Study route

The criteria used for selecting the control stops are based on findings from previous research efforts (Liu and Wirasinghe, 2001; Turnquist et al., 1980; Hickman, 2001). Studies have concluded that control at the beginning of a high-demand segment can be effective, and stops with a large number of through passengers (staying on-board) should be avoided. Another study found the terminal to be the most effective location for control (Eberlein et al., 2001). Considering those factors, the terminal and three intermediate stops were chosen as control stops, as shown in Figure 3. It should be noted that the stops in the last segment of the route (stops 56 and later), are where business activity and transit connections are located, and have high passenger alighting activity. This is coupled with highly variable run times between the stops, as observed from the empirical data.

5.2. Simulation model inputs

Poisson rates for passenger arrivals are obtained from Automated Fare Collection (AFC) and Automated Passenger Count (APC) data. Historical

travel times from Automated Vehicle Location (AVL) data were used to estimate run times between adjacent stops assumed to follow the log-normal distribution. The demand rates and run time distributions are time-dependent in 30 minute intervals. As an operational constraint, buses are not allowed to overtake each other.

Figure 4 shows the model's consistency with the actual operation by comparing the observed and simulated trip time distribution for the ED.

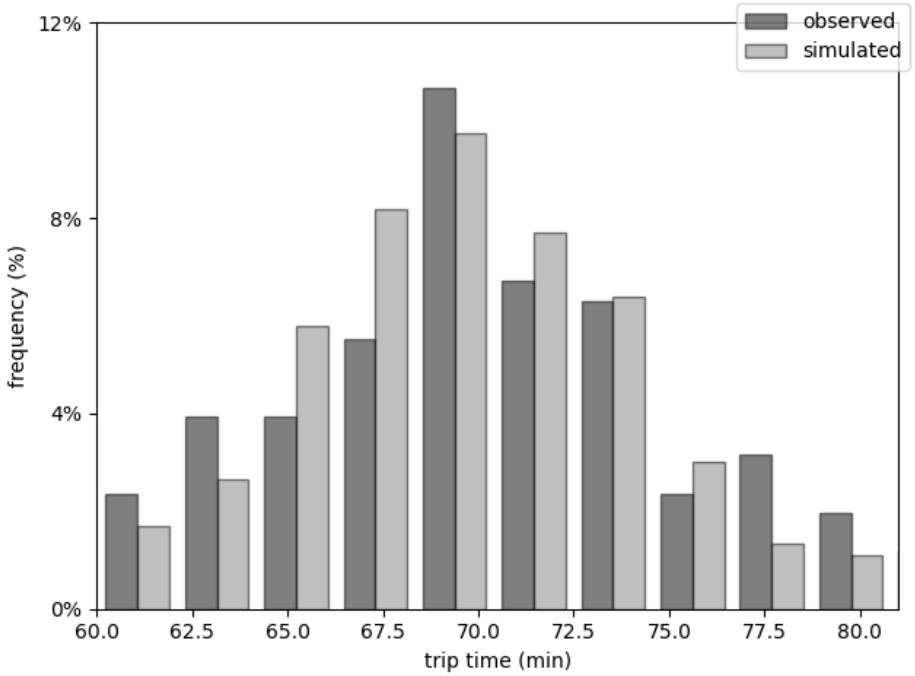

Figure 4: Simulated and observed trip time distribution for the ED. Observed values are obtained from AVL data.

5.3. Scenario design

The study evaluates the performance of two RL-based methods and compares them against a typical control approach (benchmark). The base case is defined as the case with no control (NC). For all methods we use the same set of control stops. The benchmark method refers to a well-established holding method that determines the holding so that forward and backward headways are equal. The two RL-based schemes use the MARL approach. The first

RL method does not apply the techniques for cooperative learning proposed in the previous section. These methods are described in further detail below.

Even Headway Strategy (EH)

The even-headway strategy has proved effective in simulation-based studies (Koutsopoulos and Wang, 2007; Cats et al., 2011). The holding time T_{ik}^{hold} is based on the difference between the trip's forward and backward headways, if the former is smaller than the latter (else holding time is zero). The limit on holding time is $0.4H_i$. The decision rule is thus:

$$T_{ik}^{hold} = \max\left(\min\left(\frac{h_{i+1,k}^{est} - h_{ik}}{2}, 0.4H_i\right), 0\right) \tag{10}$$

Double Deep Q-learning with Low Awareness (DDQN-LA)

To evaluate the improvement of combined control strategies and cooperative learning techniques, this method implements the holding-only strategy with an independent learning scheme. In this variation, the awareness of the learning agent is limited to its local status. The following changes are made to the POMDP definition in 4.1:

- The state parameters consist of all elements listed in 4.1 except for information on the preceding bus $i-1$.
- The skipping action is not included in the action set U. For the holding action, the set of possible strength parameters is $\Omega = \{0.0, 0.1, 0.2, 0.3, 0.4\}$, limiting the holding time to $0.4H_i$.
- We substitute the passenger-based reward in (5), which incorporates global passenger information, with a regularity-based reward used in previous applications (Alesiani and Gkiotsalitis, 2018; Wang and Sun, 2021):

  $$r_{ik}^{LA} = -\left(\frac{h_{ik} - H_i}{H_i}\right)^2 - \omega \tag{11}$$

  The first component punishes larger headway deviations, while the second component punishes excessive holding action.

Double Deep Q-learning with High Awareness (DDQN-HA)

This scenario implements the cooperative learning solution proposed in Section 4. For the holding action, the set of strength parameters used is the same as for DDQN-LA.

5.4. Evaluation metrics

The evaluation metrics used to compare the performance of the above methods capture and measure the overall impact to riders as well as transit operators:

- *Headway variability*: The coefficient of variation $CV = \frac{\sigma(h)}{E(h)}$ is used to measure the variability of the headways recorded.

- *Passenger wait time*: The average passenger wait time in the study period, calculated from the detailed simulation output.

- *Reliability buffer time (RBT)*: As described in Uniman et al. (2010), the RBT measures the difference between the 95th percentile and the median travel time for a trip. A lower value implies less time that a passenger needs to budget into their planned trip to arrive at the destination at the desired time. RBT is an important measure of quality of service. We compute the average RBT across all ODs weighted by the number of passenger trips per OD.

- *Passenger load*: The benefits of service regularity can also be measured by the variability in bus loads across trips, which also indicates the likelihood of crowded and empty buses. As such, we use the average and extreme loads (95th and 10th percentile) per stop.

- *Total trip time variability*: The total trip run time distribution is a relevant measure for transit agencies, as it has ramifications on vehicle and crew scheduling. Holding, for instance, increases travel times, but may reduce the total trip time variability. As such we study the effect of the control strategies on the trip time distribution.

6. Results Analysis

The simulation model outputs detailed vehicle trajectories, as well as passenger journey times from their arrival to the origin stop until the destination stop. The output data is used to extract the metrics described in Section 5.4. The distribution of the primary performance measures, RBT and wait time, are used for comparison.

6.1. Parameter tuning

The reward function of DDQN-HA (Eq. (5)) includes the parameter W_{wait} that weighs the waiting time term relative to the ride time. We fine-tune the value of this weight by comparing its performance in terms of RBT, average wait time, and trip time distribution. The box-and-whisker plots in Figure 5 summarize the results. Based on these results a weight $W_{wait} = 9$ is selected for the training of the DDQN-HA method in the analysis. The training parameters for the RL-based methods (DDQN-LA and DDQN-HA) are listed in Table 2.

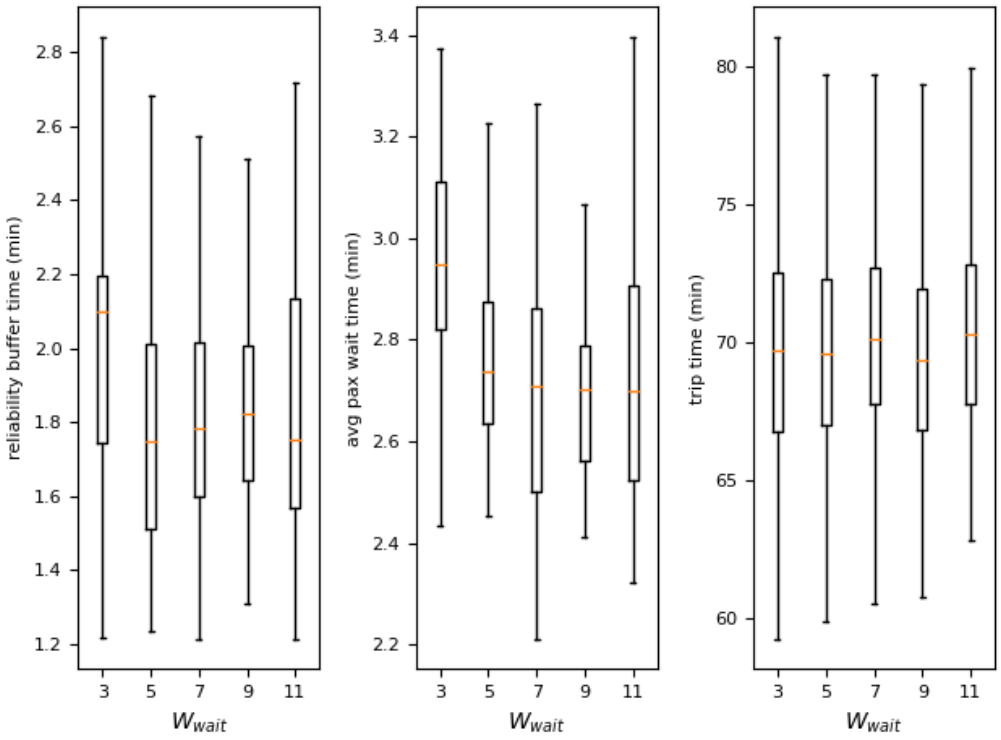

Figure 5: Passenger-centric performance metrics distribution for varying values of W_{wait}

6.2. Performance Evaluation

The trained policy is embedded into the simulation model and called every time a bus is at a control stop. Results from 40 replications are used to compare the three methods against each other and relative to the base case of no control (NC).

Table 2: Training algorithm parameters

τ	Update period	600
$L^{network}$	Hidden layer size	256
L^{batch}	Batch size	32
L^{buffer}	Buffer size	8,000
N^{layers}	Number of hidden layers	2
β	Continuous-time discount rate	0.01
$\epsilon_0, \epsilon_{final}$	Exploration rate	$0.6, 0.01$
N_1^{eps}, N_2^{eps}	Number of episodes	$500, 300$

Table 3: Distribution of passengers (%) by wait time interval

	% Passengers that wait between		
Model	0−2.5 min	2.5−5 min	$>$ 5 min
NC	48.7	31.8	19.8
EH	52.3	32.6	15.1
DDQN-LA	51.7	33.3	15.0
DDQN-HA	52.5	33.6	13.1

Passenger times

Figure 6 shows a box-and-whisker plot for RBT and average wait time. The impact of all control strategies is substantial compared to NC, evidenced by a 12-14% reduction in wait time and a 10-16% improvement in RBT. When comparing mean values for both metrics, EH performs similar or better than DDQN-LA. DDQN-HA outperforms all other methods.

In terms of average wait time, there is a marginal reduction relative to EH when the DDQN-LA method is used (1.5%) and a 3.2% reduction when DDQN-HA is used. The mean RBT value is similar between EH and DDQN-LA, while DDQN-HA results in a 6.5% decrease relative to EH.

Table 3 displays the proportion of passengers who experience wait time in various intervals (0-2.5min, 2.5-5min, and more than 5min). The DDQN-HA approach reduces the number of passengers who experience long delays significantly, not only with respect to the base case of no control, but also the other methods. This is despite the fact that DDQN-HA also includes skipping stops (which increases wait time for some passengers). This is consistent with the findings of Delgado et al. (2012).

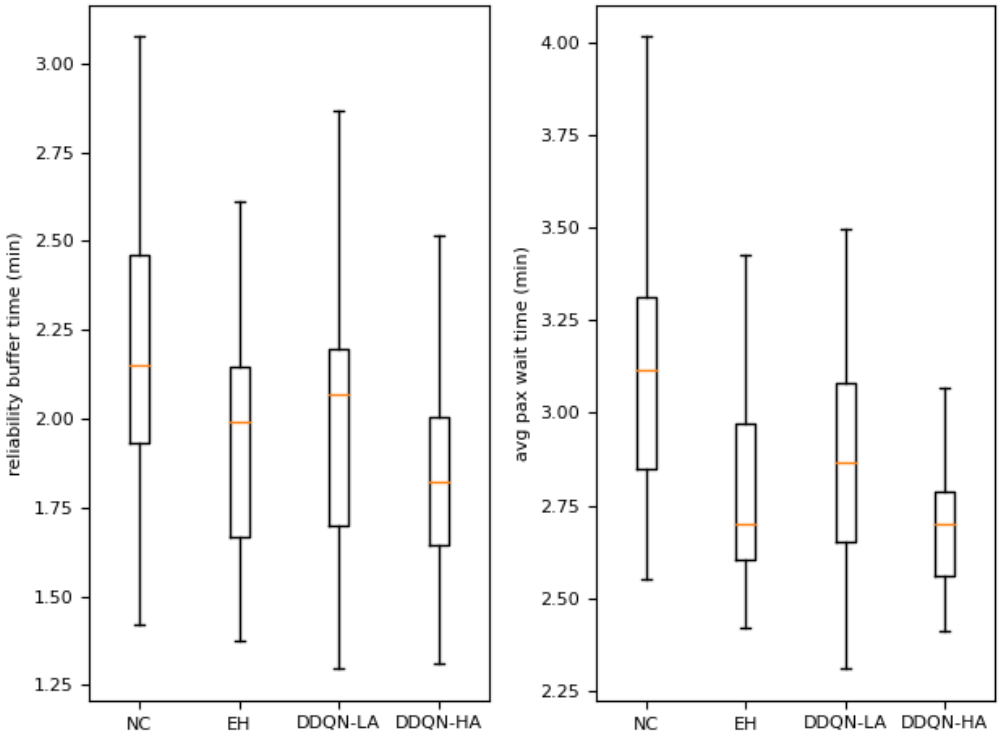

Figure 6: Passenger-centric performance metrics distribution for various methods

Headway and load variability

Figure 7 shows the coefficient of variation of headways at the control stops. In all cases, we note a progressive deterioration of headway regularity along the route, and most noticeably at stop 48. This can be attributed partly to passenger activity and higher travel time uncertainty at the downstream sections of the route. EH is consistently better than DDQN-LA with a greater improvement at stop 48 (12%). DDQN-HA in turn shows comparable performance to EH at the initial stops and better performance at the later stops, with a 8.2% improvement at stop 48.

The effectiveness of the various methods to improve service regularity is measured using two key indicators at critical stops: headway and bus load. Table 4 displays statistics for headways at a key transfer stop to the metro system (stop 61), and for the bus load at the peak point (stop 57). When the DDQN-HA method is used, the standard deviation of headways is reduced

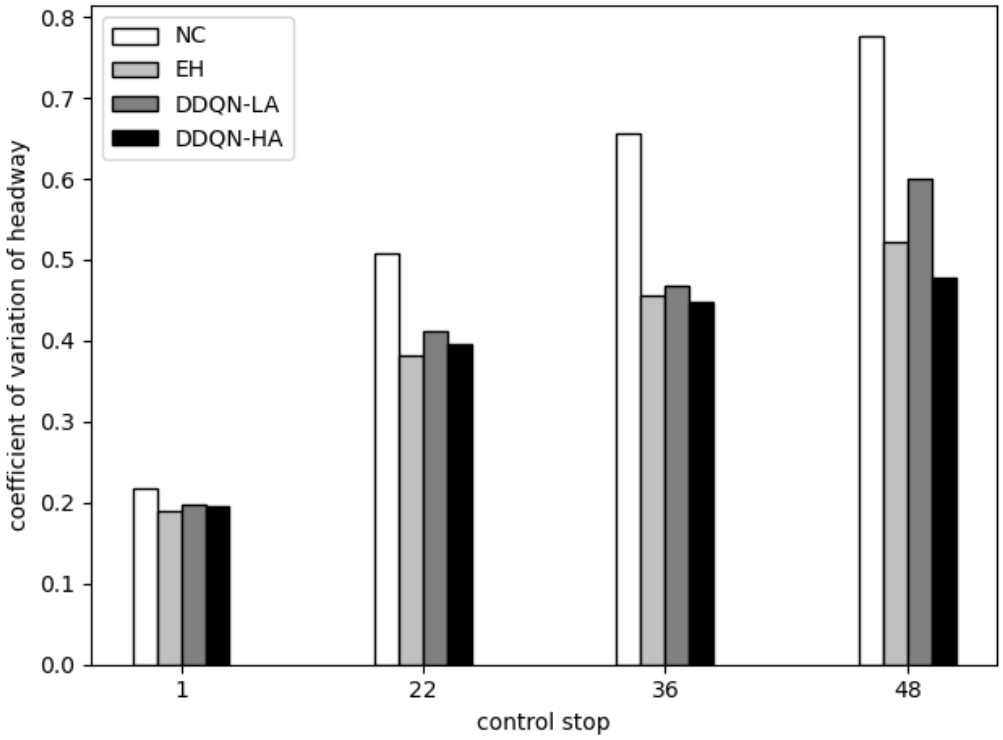

Figure 7: Coefficient of variation of headway at control stops

Table 4: Comparison of performance metrics at critical stops: mean, standard deviation (S.D.) and 95th percentile. Transfer stop and peak load point refer to stops 61 and 57, respectively

	Headway at Transfer Stop			Load At Peak Point		
Model	Average	S.D.	95th	Average	S.D.	95th
NC	4.52	3.83	11.91	20.36	14.72	50.00
EH	4.50	2.74	9.68	19.56	11.58	42.05
DDQN-LA	4.63	3.10	10.44	20.09	12.05	44.02
DDQN-HA	4.66	2.60	9.04	20.39	10.87	40.05

by 32.1% relative to the base case (NC) and 5.1% relative to the second best method (EH). The 95th percentile headway is reduced by 10.7% from NC to DDQN-LA, 9.9% from DDQN-LA to EH and 5.7% from EH to DDQN-HA.

More importantly, the standard deviation of loads is reduced by 18-26%

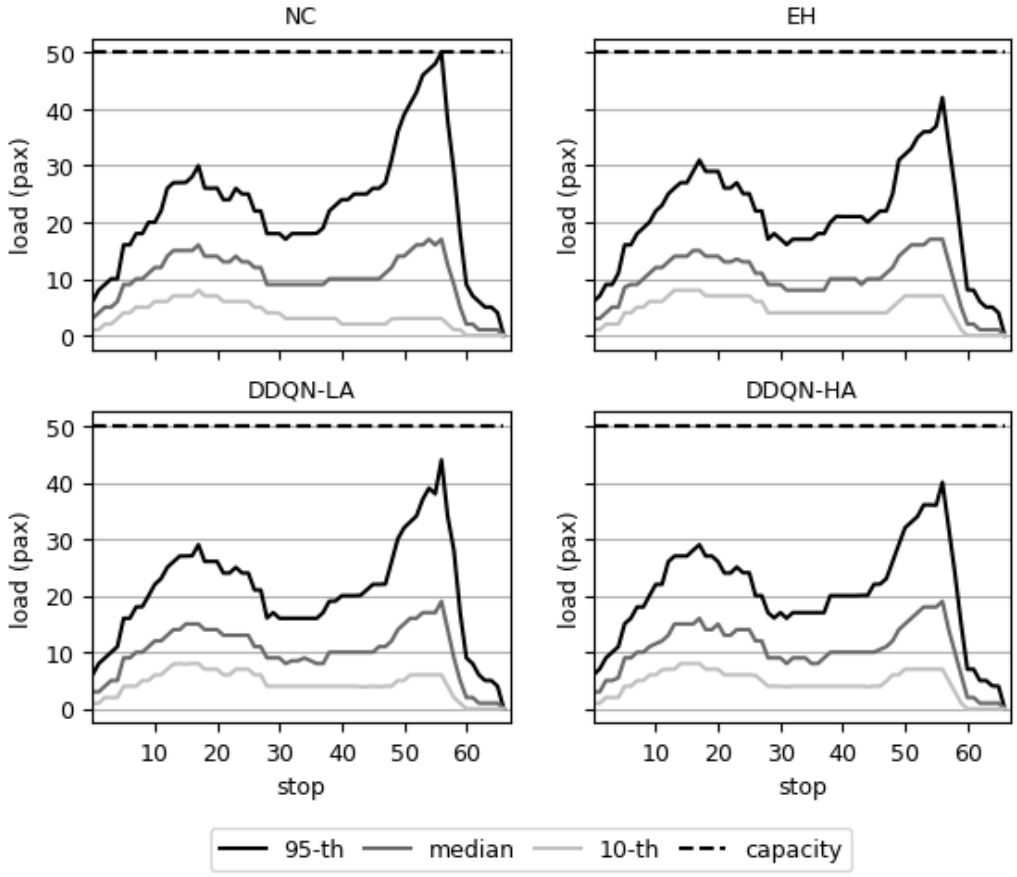

Figure 8: Load profile variability

from the NC case when the various control methods are used. The available capacity is much better utilized under the DDQN-HA strategy (26% load reduction). The 95th percentile load comparison in Figure 8 also supports this observation.

To understand how much of the crowding reduction at the peak load point (stop 57) can be attributed to the skipping control at the previous control stop (stop 48), we examine the headway and load for trips that have a large headway ($\geq$7 minutes) with the bus ahead at the control stop. In these cases, no action is taken by EH and DDQN-LA, as holding can only delay early trips. Table 5 shows the results for all methods. DDQN-HA implements stop-skipping in 28% of the trips following a large headway at the control

Table 5: Measures for trips following a large headway (≥7 min) at stop 48: percent trips skipped, mean headway, and load at peak load point (stop 57) and the previous control stop (stop 48).

		Average Headway		Average Load	
Model	% Skipped	Stop 48	Stop 57	Stop 48	Stop 57
NC	-	9.56	10.02	20.25	39.70
EH	-	8.90	8.46	17.43	32.70
DDQN-LA	-	9.06	9.21	18.11	35.93
DDQN-HA	28	8.94	7.61	17.65	31.90

stop. The resulting reductions in mean headway and load at the peak load point are significant. At stop 57, DDQN-HA reduces the mean headway by 24.1% from NC, 10.0% from EH and 17.4% from DDQN-LA; it also reduces the mean load by 19.6% from NC, 2.5% from EH, 11.2% from DDQN-LA. These results highlight the potential of the combined strategy.

Trip run time distribution

Figure 9 compares the distribution of the total trip time for the various methods. The control strategies show a narrower distribution than NC. In the case of DDQN-HA, the average trip time is not altered compared to NC, and the 95th percentile trip time is reduced by 1.7% (better than the other methods). The results also indicate that DDQN-HA would not increase fleet requirements relative to no control, and may reduce the number of very late arrivals at driver relief points. These results support the findings in Cats et al. (2011).

Control interventions

Table 6 summarises the control interventions at a critical stop before the high-demand segment (stop 48). DDQN-LA yields the most conservative policy in terms of holding times and number of trips held, which in part explains its lower effectiveness. DDQN-HA results in similar mean holding times to EH but with less interventions.

The stop-skipping strategy in DDQN-HA results in denied boardings at the skipped stop. The control policy accounts for the trade-offs between imposed denied boardings and increased regularity benefits. Furthermore, improved regularity might result in lower observed denied boardings that are consequence of overcrowding. Table 7 shows the proportion of passengers

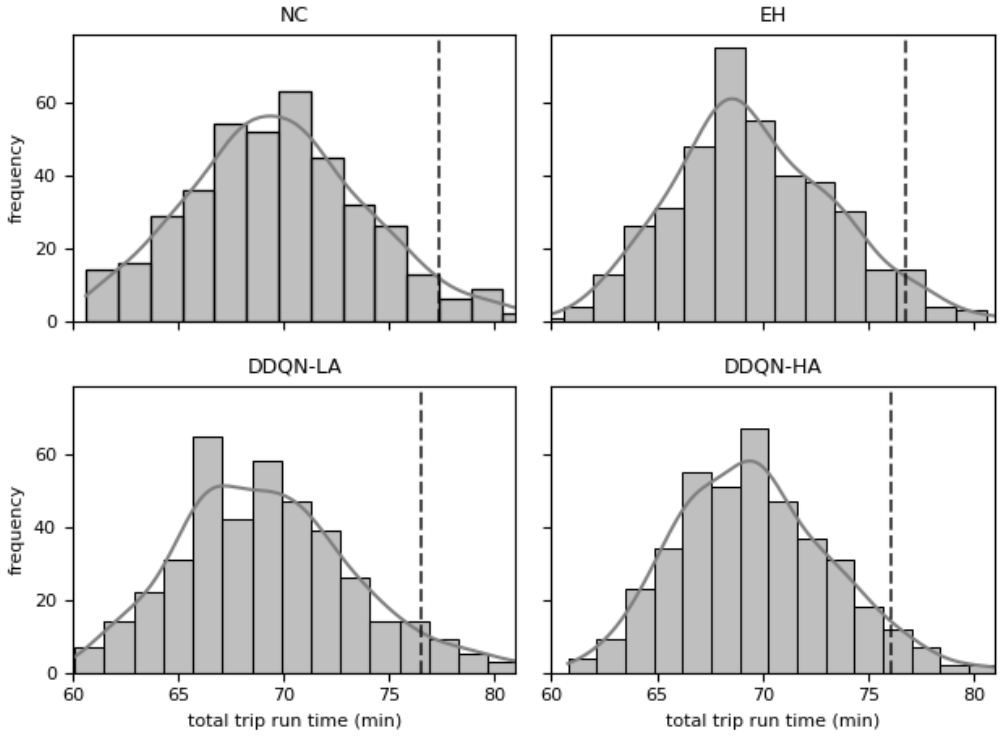

Figure 9: Comparison of total trip run time distribution for compared scenarios. The dashed line marks the 95th percentile trip time.

Table 6: Control interventions at a critical control stop (stop 48)

Model	Average Hold Time (sec)	% Trips Held	% Trips Skipped
EH	61.1	48.3	-
DDQN-LA	50.3	38.8	-
DDQN-HA	59.0	42.8	9.7

that are denied boarding either from overcrowding or from a skipping action, as well as the additional wait time experienced by the denied passengers. All control methods reduce denied boardings from overcrowding, with DDQN-HA having the best performance. DDQN-HA increases denied boardings due to stop-skipping. Overall denied boardings are similar to the other methods. However, DDQN-HA still results in lower overall wait times.

Table 7: Number of denied boardings because of overcrowding and skipping (per 1,000 boardings)

Model	Denied due to Overcrowding	Denied due to Skipping	Total
NC	10.2	-	10.2
EH	4.3	-	4.3
DDQN-LA	3.9	-	3.9
DDQN-HA	0.0	4.1	4.1

6.3. Sensitivity analysis

In order to compare the robustness of the methods, we examine their performance under different levels of travel time variability and the degree to which operators actually apply the recommended control.

For the RL-based methods, we test the policy trained on the base conditions, with no re-training (NR), directly on the altered environment, to analyze how well the policy is able to generalize to unexplored states. We also include the performance of the policy with re-training (R) on the altered environment, to evaluate the improvements.

Run time variability

We introduce two scenarios in which the coefficient of variation of bus run times is changed by ±20% compared to the base conditions (historical service data), while the mean value remains the same.

Figure 10 presents the results. As run time variability increases performance deteriorates. However, DDQN-HA (NR) outperforms DDQN-LA (NR) and EH, exhibiting its ability to generalize the policy to unexplored scenarios. The performance difference is largest in the high-variability scenario, where DDQN-HA (NR) reduces the wait time by 6.6% and 6.3% and RBT by 9.4% and 11.9% compared to DDQN-LA (NR) and EH, respectively.

The benefits from re-training the DDQN-HA method are moderate. In the low variability scenario, for example, DDQN-HA (R) improves performance from its untrained version DDQN-HA (NR) by 2.3 % and 2.5% in wait time and RBT, respectively.

Operator compliance

One assumption in the environment is that bus operators are fully compliant with the recommended control actions. However, in practice, operators may adjust the action based on their experience (Martínez-Estupiñan et al.,

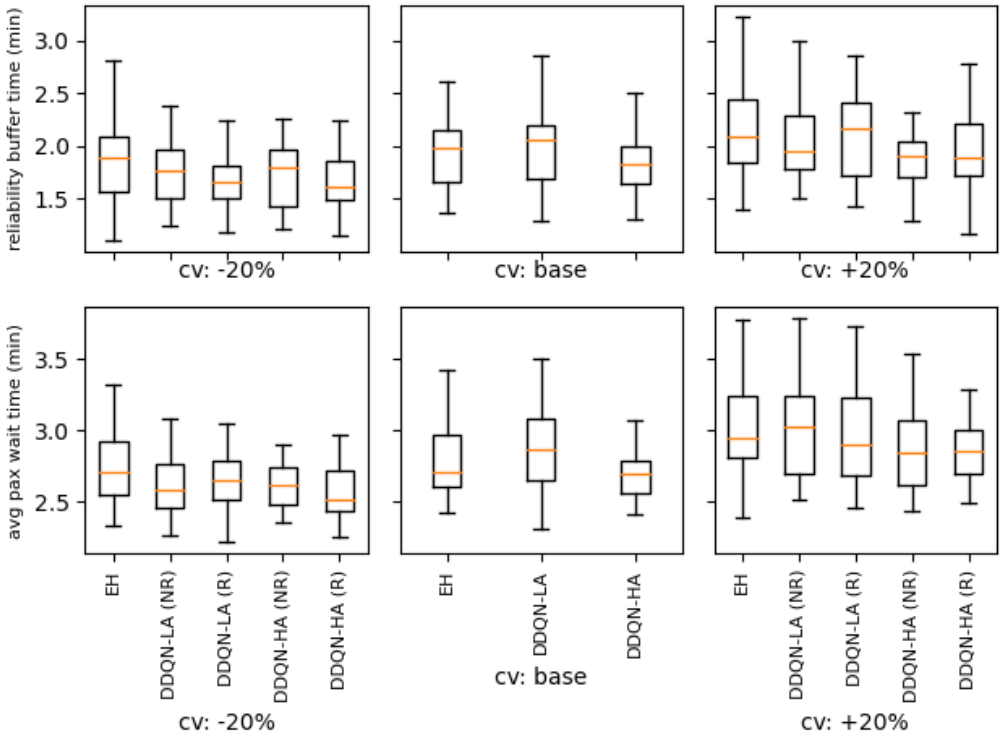

Figure 10: Impact of run time variability on passenger-centric metrics for the various control methods. For RL-based methods, the performance with (R) and without (NR) re-training is shown.

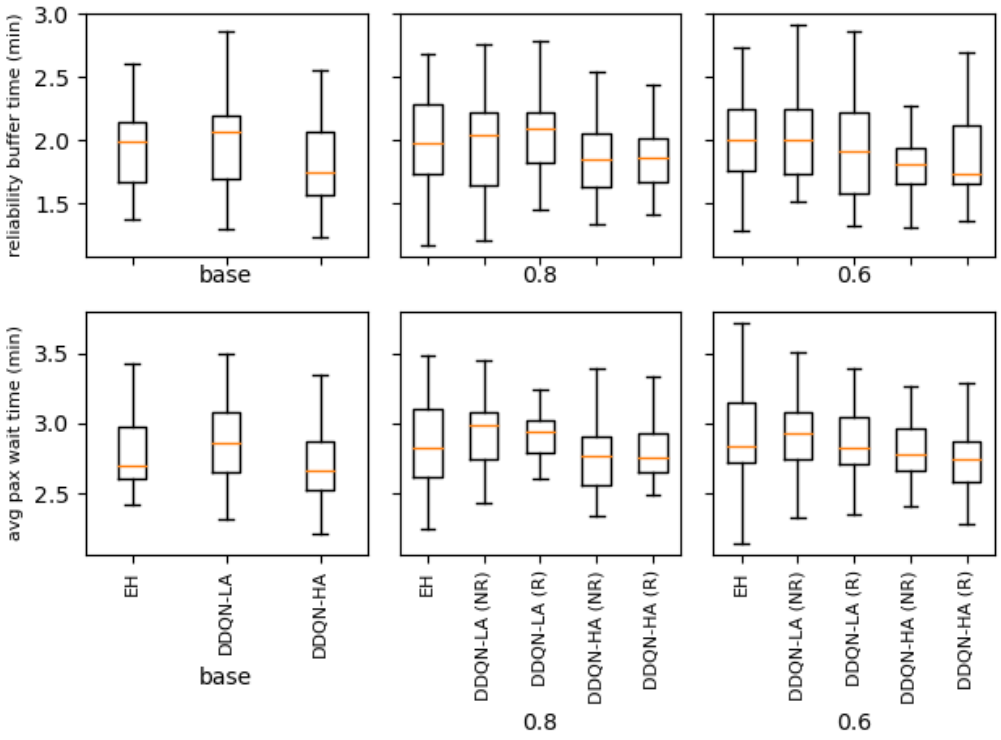

Figure 11: Impact of driver compliance on passenger-centric metrics for the various control methods. For RL-based methods, the performance with (R) and without (NR) re-training is shown.

2022). Phillips et al. (2015) evaluated the impact on wait time if a subset of buses completely ignore holding instructions, showing significant deterioration in the effectiveness of control. They also report that the impact is less significant if the probability of ignoring the instruction is spread across all buses.

To investigate the impact of lower compliance, it is assumed that drivers may depart earlier than instructed but not later. This is reasonable since drivers may depart earlier out of concern for their on-time performance, their break time at the end of their shift, and the on-board passenger discomfort. It is assumed that the executed holding time is a continuous random variable between a certain fraction (0.8 and 0.6) of the recommended holding time and the recommended time. Lastly, it is assumed that this behavior is homogeneous among all drivers. For the RL methods results without (NR) and with re-training (R) are presented. In the NR case the policy trained under full compliance is used.

Figure 11 displays the results. As expected, lower compliance leads to deteriorated performance. DDQN-HA (NR) exhibits robust performance, superior to EH, DDQN-LA (NR) and even DDQN-LA (R). When the compliance lower bound is 0.6, the performance differences between DDQN-HA and the other methods are more substantial. The differences in performance are even greater with DDQN-HA (R).

7. Conclusion

The paper developed and evaluated a multi-agent, deep reinforcement learning approach for learning cooperative bus holding and stop-skipping control strategies. Due to the asynchronous nature of the bus control problem, environment non-stationarity is an important challenge to address. By increasing the awareness of agents through extended information of neighbor agents in the state and reward, the proposed method reduces non-stationarity and facilitates cooperative learning.

The results from an extensive case study support the method's ability to learn policies that combine the advantages of holding and stop-skipping to achieve systematic improvements in passenger wait time, RBT, crowding, and trip time variability. The performance of the proposed method (DDQN-HA) was compared with the even headway strategy and an RL-based method without the techniques employed for cooperative learning. The proposed method outperforms the other methods across all metrics. The 95th

percentile of the trip times is also reduced, a finding that has positive implications for fleet management and crew scheduling. The sensitivity analysis showed that DDQN-HA is robust to changes in running time variability and driver compliance.

Future research could extend skipping to include multiple consecutive stops (expressing). Future work can also focus on more detailed modeling of bus operator responses to control instructions. More complicated state definitions with regard to surrounding agents and methods to deal with the added complexity is another interesting future research direction.

CRediT authorship contribution statement

Joseph Rodriguez: Conceptualization, Methodology, Software, Writing - Original draft preparation, **Haris N. Koutsopoulos:** Conceptualization, Methodology, Writing - Reviewing & Editing, **Shenhao Wang:** Conceptualization, Supervision, Project Administration, **Jinhua Zhao:** Supervision.

Acknowledgements

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Vehicle Technology Program Award Number DE-EE0009211. The views expressed herein do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Declaration of Interests

None.

References

Abkowitz, M., Eiger, A., Engelstein, I., 1986. Optimal control of headway variation on transit routes. Journal of Advanced Transportation 20, 73–88. URL: http://onlinelibrary.wiley.com/doi/abs/10.1002/atr.5670200106, doi:10.1002/atr.5670200106. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/atr.5670200106.

Abkowitz, M., Tozzi, J., 1986. Transit route characteristics and headway-based reliability control. Transportation Research Record 1078, 11–16.

Abkowitz, M.D., Lepofsky, M., 1990. Implementing headway-based reliability control on transit routes. Journal of Transportation Engineering 116, 49–63. Publisher: American Society of Civil Engineers.

Alesiani, F., Gkiotsalitis, K., 2018. Reinforcement Learning-Based Bus Holding for High-Frequency Services, in: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE, Maui, HI. pp. 3162–3168. URL: https://ieeexplore.ieee.org/document/8569473/, doi:10.1109/ITSC.2018.8569473.

Barnett, A., 1974. On Controlling Randomness in Transit Operations. Transportation Science 8, 102–116. URL: https://pubsonline.informs.org/doi/abs/10.1287/trsc.8.2.102, doi:10.1287/trsc.8.2.102. publisher: INFORMS.

Bartholdi III, J.J., Eisenstein, D.D., 2012. A self-coördinating bus route to resist bus bunching. Transportation Research Part B: Methodological 46, 481–491. Publisher: Elsevier.

Bradtke, S., Duff, M.O., 1995. Reinforcement Learning Methods for Continuous-Time Markov Decision Problems, in: Advances in Neural Information Processing Systems, MIT Press. pp. 393–400.

Cats, O., Larijani, A.N., Koutsopoulos, H.N., Burghout, W., 2011. Impacts of Holding Control Strategies on Transit Performance: Bus Simulation Model Analysis. Transportation Research Record 2216, 51–58. URL: https://doi.org/10.3141/2216-06, doi:10.3141/2216-06. publisher: SAGE Publications Inc.

Chen, C., Chen, W., Chen, Z., 2015a. A Multi-Agent Reinforcement Learning approach for bus holding control strategies. Advances in Transportation Studies .

Chen, W., Kunlin Zhou, Chen, C., 2016. Real-time bus holding control on a transit corridor based on multi-agent reinforcement learning, in: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), IEEE, Rio de Janeiro, Brazil. pp. 100–106. URL: http://ieeexplore.ieee.org/document/7795538/, doi:10.1109/ITSC.2016.7795538.

Chen, X., Hellinga, B., Chang, C., Fu, L., 2015b. Optimization of headways with stop-skipping control: a case study of bus rapid transit system. Journal of Advanced Transportation 49, 385–401. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/atr.1278, doi:10.1002/atr.1278. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/atr.1278.

Crites, R.H., Barto, A.G., 1998. Elevator Group Control Using Multiple Reinforcement Learning Agents. Machine Learning 33, 235–262. URL: https://doi.org/10.1023/A:1007518724497, doi:10.1023/A:1007518724497.

Daganzo, C.F., 2009. A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. Transportation Research Part B: Methodological 43, 913–921. URL: https://linkinghub.elsevier.com/retrieve/pii/S0191261509000484, doi:10.1016/j.trb.2009.04.002.

Delgado, F., Munoz, J.C., Giesen, R., 2012. How much can holding and/or limiting boarding improve transit performance? Transportation Research Part B: Methodological 46, 1202–1217. URL: https://linkinghub.elsevier.com/retrieve/pii/S0191261512000653, doi:10.1016/j.trb.2012.04.005.

Delgado, F., Munoz, J.C., Giesen, R., Cipriano, A., 2009. Real-time control of buses in a transit corridor based on vehicle holding and boarding limits. Transportation Research Record 2090, 59–67. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Eberlein, X.J., 1995. Real-time control stategies in transit operations : models and analysis. Thesis. Massachusetts Institute of Technology. URL: https://dspace.mit.edu/handle/1721.1/11387. accepted: 2005-08-17T18:04:20Z.

Eberlein, X.J., Wilson, N.H.M., Bernstein, D., 2001. The Holding Problem with Real–Time Information Available. Transportation Science 35, 1–18. URL: http://pubsonline.informs.org/doi/abs/10.1287/trsc.35.1.1.10143, doi:10.1287/trsc.35.1.1.10143.

Fu, L., Liu, Q., Calamai, P., 2003. Real-Time Optimization Model for Dynamic Scheduling of Transit Operations. Transportation Research Record

1857, 48–55. URL: https://doi.org/10.3141/1857-06, doi:10.3141/1857-06. publisher: SAGE Publications Inc.

Gao, Y., Kroon, L., Schmidt, M., Yang, L., 2016. Rescheduling a metro line in an over-crowded situation after disruptions. Transportation Research Part B: Methodological 93, 425–449. Publisher: Elsevier.

van Hasselt, H., Guez, A., Silver, D., 2015. Deep Reinforcement Learning with Double Q-learning. arXiv:1509.06461 [cs] URL: http://arxiv.org/abs/1509.06461. arXiv: 1509.06461.

Hickman, M.D., 2001. An analytic stochastic model for the transit vehicle holding problem. Transportation Science 35, 215–237. Publisher: INFORMS.

Koutsopoulos, H.N., Wang, Z., 2007. Simulation of Urban Rail Operations: Application Framework. Transportation Research Record 2006, 84–91. URL: https://doi.org/10.3141/2006-10, doi:10.3141/2006-10. publisher: SAGE Publications Inc.

Laskaris, G., Cats, O., Jenelius, E., Viti, F., 2016. A real-time holding decision rule accounting for passenger travel cost, in: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 2410–2415.

Laurent, G.J., Matignon, L., Le Fort-Piat, N., 2011. The world of independent learners is not markovian. International Journal of Knowledge-based and Intelligent Engineering Systems 15, 55–64. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/KES-2010-0206, doi:10.3233/KES-2010-0206.

Li, X., Zhang, J., Bian, J., Tong, Y., Liu, T.Y., 2019. A Cooperative Multi-Agent Reinforcement Learning Framework for Resource Balancing in Complex Logistics Network. arXiv:1903.00714 [cs] URL: http://arxiv.org/abs/1903.00714. arXiv: 1903.00714.

Li, Y., Rousseau, J.M., Gendreau, M., 1995. Real-time dispatching of public transit operations with and without bus location information, in: Computer-Aided Transit Scheduling. Springer, pp. 296–308.

Lin, L.J., 1992. Self-improving reactive agents based on reinforcement learning, planning and teaching. Machine learning 8, 293–321. Publisher: Springer.

Liu, G., Wirasinghe, S., 2001. A simulation model of reliable schedule design for a fixed transit route. Journal of Advanced Transportation 35, 145–174. Publisher: Wiley Online Library.

Liu, Z., Yan, Y., Qu, X., Zhang, Y., 2013. Bus stop-skipping scheme with random travel time. Transportation Research Part C: Emerging Technologies 35, 46–56. URL: https://www.sciencedirect.com/science/article/pii/S0968090X13001265, doi:10.1016/j.trc.2013.06.004.

Martínez-Estupiñan, Y., Delgado, F., Muñoz, J.C., Watkins, K.E., 2022. Understanding what elements influence a bus driver to use headway regularity tools: case study of Santiago public transit system. Transportmetrica A: Transport Science , 1–34URL: https://www.tandfonline.com/doi/full/10.1080/23249935.2022.2025950, doi:10.1080/23249935.2022.2025950.

Menda, K., Chen, Y.C., Grana, J., Bono, J.W., Tracey, B.D., Kochenderfer, M.J., Wolpert, D., 2019. Deep Reinforcement Learning for Event-Driven Multi-Agent Decision Processes. IEEE Transactions on Intelligent Transportation Systems 20, 1259–1268. doi:10.1109/TITS.2018.2848264. conference Name: IEEE Transactions on Intelligent Transportation Systems.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–533. URL: http://www.nature.com/articles/nature14236, doi:10.1038/nature14236.

Oliehoek, F.A., Amato, C., 2016. A concise introduction to decentralized POMDPs. Springer.

Phillips, W., del Rio, A., Muñoz, J.C., Delgado, F., Giesen, R., 2015. Quantifying the effects of driver non-compliance and communication

system failure in the performance of real-time bus control strategies. Transportation Research Part A: Policy and Practice 78, 463–472. URL: https://www.sciencedirect.com/science/article/pii/S0965856415001639, doi:10.1016/j.tra.2015.06.005.

Rossetti, M.D., Turitto, T., 1998. Comparing static and dynamic threshold based control strategies. Transportation Research Part A: Policy and Practice 32, 607–620. Publisher: Elsevier.

Saw, V.L., Vismara, L., Chew, L.Y., 2020. Intelligent buses in a loop service: Emergence of no-boarding and holding strategies. Complexity 2020. Publisher: Hindawi.

Sun, A., Hickman, M., 2005. The Real–Time Stop–Skipping Problem. Journal of Intelligent Transportation Systems 9, 91–109. URL: https://www.tandfonline.com/doi/full/10.1080/15472450590934642, doi:10.1080/15472450590934642.

Sáez, D., Cortés, C.E., Milla, F., Núñez, A., Tirachini, A., Riquelme, M., 2012. Hybrid predictive control strategy for a public transport system with uncertain demand. Transportmetrica 8, 61–86. URL: http://www.tandfonline.com/doi/abs/10.1080/18128601003615535, doi:10.1080/18128601003615535.

Sánchez-Martínez, G., Koutsopoulos, H., Wilson, N., 2016. Real-time holding control for high-frequency transit with dynamics. Transportation Research Part B: Methodological 83, 1–19. URL: https://linkinghub.elsevier.com/retrieve/pii/S019126151500257X, doi:10.1016/j.trb.2015.11.013.

Tirachini, A., Godachevich, J., Cats, O., Muñoz, J.C., Soza-Parra, J., 2021. Headway variability in public transport: a review of metrics, determinants, effects for quality of service and control strategies. Transport Reviews , 1–25URL: https://www.tandfonline.com/doi/full/10.1080/01441647.2021.1977415, doi:10.1080/01441647.2021.1977415.

Turnquist, M.A., 1974. Strategies for improving reliability of bus transit service. Computers and Operations Research 1, 201–211.

Turnquist, M.A., Blume, S.W., others, 1980. Evaluating potential effectiveness of headway control strategies for transit systems. Transportation Research Record 746, 25–29.

Uniman, D.L., Attanucci, J., Mishalani, R.G., Wilson, N.H.M., 2010. Service Reliability Measurement Using Automated Fare Card Data: Application to the London Underground. Transportation Research Record 2143, 92–99. URL: https://doi.org/10.3141/2143-12, doi:10.3141/2143-12. publisher: SAGE Publications Inc.

Wang, J., Sun, L., 2020. Dynamic holding control to avoid bus bunching: A multi-agent deep reinforcement learning framework. Transportation Research Part C: Emerging Technologies 116, 102661. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20305763, doi:10.1016/j.trc.2020.102661.

Wang, J., Sun, L., 2021. Reducing Bus Bunching with Asynchronous Multi-Agent Reinforcement Learning. arXiv:2105.00376 [cs] URL: http://arxiv.org/abs/2105.00376. arXiv: 2105.00376.

Watkins, C., 1989. Learning from delayed rewards. 1989. University of Cambridge .

Zhang, L., Huang, J., Liu, Z., Vu, H.L., 2021. An agent-based model for real-time bus stop-skipping and holding schemes. Transportmetrica A: Transport Science 17, 615–647. URL: https://www.tandfonline.com/doi/full/10.1080/23249935.2020.1802363, doi:10.1080/23249935.2020.1802363.

Zhang, S., Lo, H.K., 2018. Two-way-looking self-equalizing headway control for bus operations. Transportation research part B: methodological 110, 280–301. Publisher: Elsevier.

Zhao, J., Bukkapatnam, S., Dessouky, M., 2003. Distributed architecture for real-time coordination of bus holding in transit networks. IEEE Transactions on Intelligent Transportation Systems 4, 43–51. doi:10.1109/TITS.2003.809769. conference Name: IEEE Transactions on Intelligent Transportation Systems.