2018 21st International Conference on Intelligent Transportation Systems (ITSC)
Maui, Hawaii, USA, November 4-7, 2018

Reinforcement Learning-based Bus Holding for high-frequency services

Francesco Alesiani *Member IEEE*[1], Konstantinos Gkiotsalitis *Member IEEE*[2]

***Abstract*— Since the bus holding problem is an operational control problem, bus holding decisions should be made in real-time. For this reason, common bus holding approaches, such as the one-headway-based holding, focus on computationally inexpensive, rule-based techniques that try to minimize the deviation of the actual headways from the planned ones. Nevertheless, rule-based methods optimize the system locally without considering the full effect of the bus holding decisions to future trips or other performance indicators. For this reason, this work introduces a Reinforcement Learning approach which is capable of making holistic bus holding decisions in real-time after the completion of a training period. The proposed approach is trained in a circular bus line in Singapore using 400 episodes (where an episode is one day of operations) and evaluated using 200 episodes demonstrating a significant improvement in scenarios with strong travel time disturbances and a slight improvement in scenarios with low travel time variations.**

Keywords: Bus Holding; Operational Control; Reinforcement Learning; Service Regularity.

I. Introduction

The design of bus routes is generally addressed at the strategic planning phase [1]. After the strategic planning, the frequencies of the bus lines, the daily timetables and the crew and vehicle schedules are defined at the tactical planning stage [2]. Nevertheless, even if the tactical planning is based on sophisticated statistical methods, the actual performance of the bus services deviates significantly from the expectations of the bus operators due to the travel time and passenger demand variations throughout the day [3], [4].

The unstable nature of bus operations impacts significantly the daily planning and can increase the in-vehicle travel times and the waiting times of passengers at regular and transfer stops [5]. Public transport authorities (PTAs) are aware of this problem and provide monetary incentives to bus operators for improving the reliability of their services. For instance, the Land Transport Authority (LTA) in Singapore provides a monetary benefit of 2,000 Singaporean dollars for each 0.1-minute improvement of the service regularity measured in terms of operational headway adherence to the planned headways [6].

These incentives have increased the pressure to introduce corrective measures during the daily operations for improving the service reliability under the presence of travel time and passenger demand variations. For this reason, several studies have investigated the effect of operational control actions such as bus holding [7], [8], stop-skipping [9] and short-turning [10] to the improvement of the service reliability.

This work focuses specifically on the bus holding problem. In the bus holding problem, automated vehicle location (AVL) data provides the current positions of the running buses and the bus operator decides whether to hold a bus or not when it arrives at a bus stops. This decision is typically made based on the actual headway deviation from the planned headway value. This is a local optimization approach, known as one-headway-based optimization [11], which considers only the position of two buses for making a decision.

Such local-based decision-making approaches have prevailed in practice because their decision-making process is trivial and allows the computation of bus holding times in real time. Nevertheless, local decisions that do not consider the implications of the bus holding times to future bus trips and to other operational constraints, such as the limitations of the total trip travel times, cannot address the bus holding optimization problem in a holistic manner.

Alternatively, formulating the bus holding optimization problem as a mathematical program that considers the impact of the bus holding decisions to all other running trips and future operations leads to a combinatorial problem that cannot be solved in real time [12]. For this reason, this study investigates the potential of using a reinforcement learning-based bus holding scheme where, after a training period, decisions about bus holding times can be made in real time while considering the impact of bus holding decisions to other running buses and future trips instead of deciding by following local-based rules.

II. Related Studies

Most studies on the bus holding problem of high-frequency services (which operate based on a regularity scheme) focus on holding buses at stops for reducing the deviation of the actual headways from the planned ones; thus, improving the service reliability [13]–[16]. One exception is the work of [17] that proposed a self-coordination scheme within which bus holding times try to reduce the headway variability without adhering to the planned headway. In the above-mentioned works, buses are not held at any bus stop, but at specific intermediate time point (ITP) stops for mitigating the passenger inconvenience caused by multiple holdings [18].

Most of the above-mentioned works employ local-based, heuristic optimization approaches for deciding the bus holding times and do not consider the impact of the holding times to the prolongation of the trip travel times that can

[1]NEC Laboratories Europe, Kurfuersten-Anlage 36, Heidelberg, Germany, 69121 francesco.alesiani@neclab.eu
[2]University of Twente, De Horst 2, Enschede, The Netherlands, 7522LW k.gkiotsalitis@utwente.nl

delay the dispatching of future trips. [11] tested two of the most common headway-based bus holding strategies that are based on local rules in the bus network of Waterloo, Ontario: (i) the one-headway-based control method where a bus is held at one stop if it is closer to its preceding bus than its planned headway; and (ii) the two-headway-based method where the position of both its preceding and following bus are considered when making a holding time decision. [19] and [20] have also proposed distributed control models where buses act as agents that communicate in real-time to achieve local-level coordination.

Even if most works on bus holding consider deterministic trip travel times when holding a bus at a stop, a distinct number of works have considered the potential variation of the future trip travel times when computing bus holding times [13], [21]. For instance, [15] introduced stochasticity to the trip travel times when deciding about the holding time of a bus at the first bus stop of the trip. The work of [15] and the works of [22]–[24] focused on the single-holding problem where bus trips are held only at the first bus stop and not at any other intermediate stop. Therefore, such works cannot exploit the full potential of holding times on correcting the headway deviations.

This work contributes to the prior art by examining the bus holding problem in a more holistic manner by modeling the implications of bus holding times to the dispatching times of future trips and the headways of the running buses. Such a holistic mathematical formulation of the bus holding problem cannot be solved with classical exact optimization approaches [25], [26]. Therefore, the second contribution of this study is the introduction of a reinforcement learning approach for making real-time bus holding decisions.

Since the bus trips can be seen as independent entities, this study perceives the bus holding problem as a multi-agent decentralized policy optimization problem. Multi-agent decentralized policy optimization problems are complex, especially if the coordination among agents is limited. Distributed multi-agent systems have been studied as distributed optimization problems [27], as game theory problems [28], [29] or as decentralized partially observable Markov decision processes (Dec-POMDPs) [30]. Our proposed method follows the Dec-POMDPs approach and assumes a common policy for all the agents as it will be described in the following sections.

III. Problem Formulation

A. Bus Movement Model

Each bus trip has to serve all stops along its path. Let $N = \{1, ..., n, ..., |N|\}$ be the set of daily bus trips of one bus line and $S = \{1, ..., s, ..., |S|\}$ the set of bus stops. If δ_n is the dispatching time of a bus trip $n \in N$, then its departure time from any other stop $s \in S \setminus \{1\}$ is:

$$d_{n,s}(X) = \begin{cases} \delta_n + x_{n,1}, & s = 1 \\ d_{n,s-1}(X) + t_{n,s-1} + k_{n,s} + x_{n,s}, & s \in S \setminus \{1\} \end{cases} \quad (1)$$

where $d_{n,s}$ is the departure time of trip n from stop s, $x_{n,s}$ the holding time of trip n at stop s, X an $|N| \times |S|$-dimensional matrix of all bus holding times at stops, $d_{n,s-1}(X)$ the departure time of the same trip from its previous stop $s-1$, $t_{n,s-1}$ the travel time of trip n from stop $s-1$ to stop s, and $k_{n,s}$ the required dwell time at stop s for the passenger boardings/alightings.

The recursive formula of eq.1 for determining the departure time of a bus n from any stop $s \in S \setminus \{1\}$ can be written as:

$$d_{n,s}(X) = \delta_n + x_{n,1} + \sum_{j=2}^{s} (t_{n,j-1} + k_{n,j} + x_{n,j}) \quad (2)$$

The arrival time of a trip n at any stop $s \in S \setminus \{1, 2\}$ is:

$$a_{n,s}(X) = d_{n,s-1}(X) + t_{n,s-1} = \left(\delta_n + x_{n,1} + \sum_{j=2}^{s-1} (t_{n,j-1} + k_{n,j} + x_{n,j})\right) + t_{n,s-1} \quad (3)$$

and the arrival time at stop $s = 2$ is $a_{n,s}(X) = \delta_n + x_{n,1} + t_{n,s-1}$.

The headway between a bus trip n and its preceding trip $n-1$ at stop s is:

$$h_{n,s}(X) = a_{n,s}(X) - a_{n-1,s}(X) \quad (4)$$

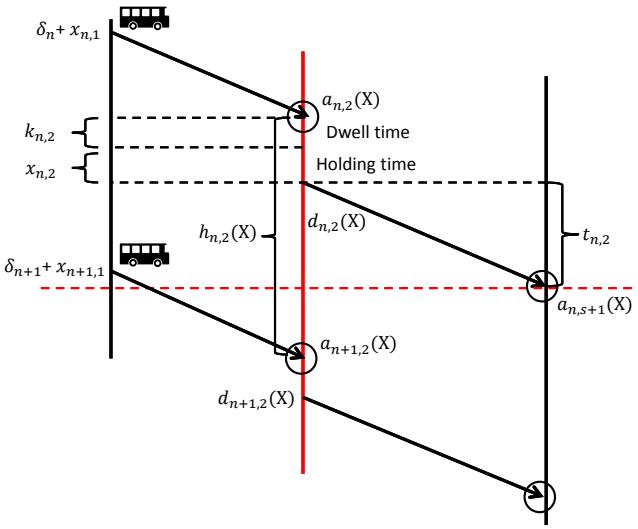

Fig. 1. Vehicle movements and decision points

At each time instance, the positions of the running buses are updated and a bus holding decision is made every time a bus arrives at one bus stop. An example of the movements of running buses is provided in Fig.1.

B. Bus Trip Model

This work focuses on circular bus lines and makes the following assumptions regarding the daily operations:

- each bus can be assigned to multiple daily trips;

- a bus can be held at specific ITP stops;
- a bus trip can start only when the previous trip that was operated by the same bus is finished;
- buses that serve the same line do not overtake each other [21], [31].

Following the 3^{rd} assumption, a bus trip n is allowed to be dispatched only after the previous bus trip n' which was operated by the same bus has been completed. This yields the following expression:

$$d_{n,1}(X) = \max(\delta_n + x_{n,1}; a_{n',|S|}(X) + k_{n',|S|}) \quad (5)$$

where $d_{n,1}(X)$ is the actual dispatching time of trip n, $|S|$ the last stop of the bus line and $a_{n',|S|}(X)$ the arrival time of trip n' at the last stop.

C. Cost Functions

In this work we introduce two metrics that will be used for measuring the performance of the bus holdings.

The first metric is based on the deviation of the actual headways from the planned ones. This metric measures the reliability of high-frequency services and is the key performance indicator of several transport authorities [32]. This metric can be expressed as:

$$f_1(X) = \sqrt{\frac{1}{|S|(|N|-1)} \sum_{s \in ITP, n \in N \setminus \{1\}} |h_{n,s}(X) - w_{n,s}|^2} \quad (6)$$

where $f_1(X)$ measures the average daily deviation of the actual headways from their planned values which are denoted as $w_{n,s}$. In addition, ITP is the set of intermediate time point stops where bus holdings are allowed.

The planned headway, $w_{n,s}$, depends on the time of the day when bus n arrives at stop s because the bus services might operate in higher frequencies during peak hours and in lower frequencies during offpeaks. For instance, if w_s^ρ is the planned headway at stop s during the time period ρ, then $w_{n,s} = w_s^\rho$ if the bus trip n arrives at bus stop s during the ρ^{th} time period of the day.

In this work, we also monitor the travel time of each trip since holding a bus at several stops might prolong significantly its total trip travel time. The total trip travel time is the time difference between the departure from the first station (Eq. 5) and the arrival at the last station (Eq. 3),

$$T_n(X) = a_{n,|S|}(X) - d_{n,1}(X) \quad (7)$$

We can then introduce the second performance metric that measures the excess trip travel time (ETT). The ETT is the exceeding trip travel time from a pre-defined total trip travel time threshold $T^{\max}$ and the second metric can be formulated as:

$$f_2(X) = \sqrt{\frac{1}{|N|} \sum_{n \in N} \max\left(T_n(X) - T^{\max}, 0\right)^2} \quad (8)$$

where $f_2(X)$ is the travel time prolongation of the average bus trip from the pre-defined total trip travel time threshold which is used for penalizing excessive trip delays that can affect the dispatching times of future trips and the crew/vehicle schedules.

Each trip travel time, $T_n(X)$, is the sum of the travel times between bus stops plus the dwell and the holding times at stops. The quantity $\max\left(T_n(X) - T^{\max}, 0\right)^2$ is squared for increasing the penalization of trip travel times that exceed significantly the travel time threshold. At the same time, the term $\max\left(T_n(X) - T^{\max}, 0\right)$ ensures that travel times which are lower than or equal to the trip travel time threshold, $T_n(X) \leq T^{\max}$, do not penalize the performance of the operations since $\max\left(T_n(X) - T^{\max}, 0\right) = 0$ in such case.

IV. Reinforcement Learning

A. General Introduction

In this section, we form the bus holding problem that aims at minimizing the performance metrics $f_1(X), f_2(X)$, as a Reinforcement Learning (RL) problem. In general, a single agent RL problem [33] is defined by a Markov Decision Process (MDP) as (Y, U, P, R, γ) where Y is the set of states of the process, U the set of actions that can be preformed on the system, P is the transition probability between a state y and the following state y' after action u has been applied (i.e. $P(y \to y'|y, u)$), and R the reward at state y when selecting action u, (i.e. $r(y, u)$). If the system is not stationary, the state and the action are denoted with a time index, k, thus y_k, u_k denote the state and action at instant k, while r_k the reward between state y_k and y_{k+1} under the action u_k.

A MDP also considers a discount factor γ that is used to define the expected total reward at a state y as $V(y) = E_P\{\sum_{k=0}^{+\infty} \gamma^k r_k | y_0 = y\}$, while the Q function is the expected reward on the joint state and action $Q(y, u) = E_P\{\sum_{k=0}^{+\infty} \gamma^k r_k | y_0 = s, u_0 = u\}$. On an MDP, a deterministic policy is defined as a deterministic mapping from the state to the action $u_k = \pi(y_k)$, while a probabilistic policy defines the probability of an action u_k, given the current state y_k, i.e. $\pi(u_k|y_k)$.

Under the presence of multiple agents with decentralized and partial observable MDPs, one can use the Decentralize Partial Observable MDP (Dec-POMDP) method [30]. In this work, we model the bus holding (BH) problem in the form of a Dec-POMDP where the observation of the single agent (bus trip) depends only on its own state. The single agent is the single bus trip that decides about its holding time (action) when it arrives at an ITP stop.

B. States and Observation space of the BH problem

The state of the BH problem comprises of the states of all bus trips (agents). However, the observation space when deciding the holding time of a trip n at a stop $s \in ITP$ depends only on the following information:

- $d_{n-1,s}$: is the departure time of the preceding bus trip, $n-1$, from stop $s \in ITP$
- $a_{n,s}$: is the arrival time of the examined trip at stop s
- $w_{n,s}$: is the target headway time of bus trip n at stop s

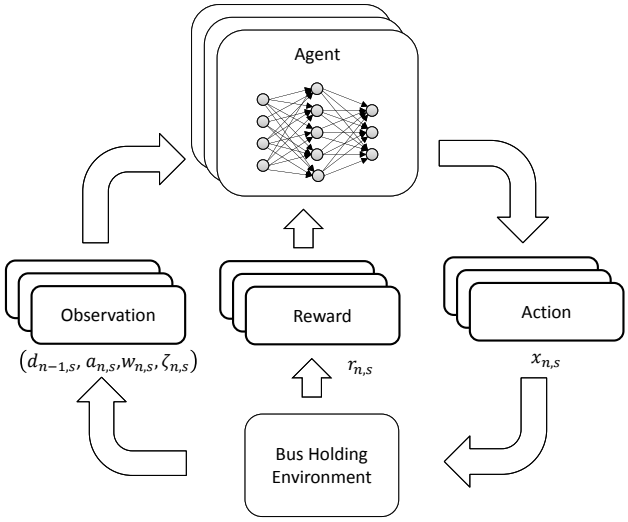

Fig. 2. Reinforcement Learning Architecture

In this way, each agent (bus trip) is only informed about its status, target headway and the departure time of its preceding trip at stop $s \in ITP$. This mimics the information which is typically used in bus holding control [11].

C. Reward Function

The performance of the daily operations measured by Eq.6 and Eq.8 is only available at the end of the episode (e.g. end of the day), while the reward function of the RL must be evaluated for every action that takes place in the system. Because of that, the reward function plays a critical role for the convergence of the Dec-POMDP and it is a critical design choice.

In this work, we shape the reward function in accordance with the elements of the cost functions. Therefore, we include a component r_w which is inversely proportional to the difference from the target headway

$$r_w^n(s) = -\rho_\mu \left(\| h_{n,s}(X) - w_{n,s} \| \right) \tag{9}$$

As discussed before, the planned headway, $w_{n,s}$ does not depend on the bus holdings, but on the time of arrival of trip n at stop $s \in ITP$. The penalization function is a non-decreasing positive function and we use the Huber function that penalizes small deviations more than large deviations because this can provide a more robust learning of the expected reward on the joint state and action [34]:

$$\rho_\mu(r) = \begin{cases} \frac{1}{2} r^2 & \text{for} |r| \leq \mu \\ \mu(|r| - \frac{1}{2}\mu) & \text{otherwise} \end{cases} \tag{10}$$

where μ is the threshold value of the Huber function.

An additional term penalizes actions that lead to excessive trip travel times

$$r_t^n(s) = -\rho_\mu \left(\max \left(T_n - T^{\max}, 0 \right) \right) \tag{11}$$

Furthermore, a prolonged bus holding at an ITP stop has a negative effect to the travel time of the trip and the passengers' convenience. Therefore, we add a reward term that penalizes solutions with longer holding times

$$r_h^n(s) = -\rho_\mu \left(|x_{n,s}| \right) \tag{12}$$

The final reward function that evaluates the effect of an action u, where this action is the bus holding time of trip n at ITP stop s, is defined as the weighted sum of the reward contributions:

$$r_s^n = \beta_w r_w^n(s) + \beta_t r_t^n(s) + \beta_h r_h^n(s) \tag{13}$$

where β_w, β_t and β_h are the respective weight factors of the reward function, where we choose β_w=β_t=β_h=1, since all components have time dimension.

D. Action space

The action of the RL algorithm is the holding of a bus trip when it arrives at an ITP stop. The bus holding time can take a value from a set of pre-determined quantities that can be encoded in a Human Machine Interface (HMI) and can be interpreted by the bus driver. This set can have a granularity of 10 seconds as suggested in other works [18], [35] and can be limited to 30 seconds for avoiding prolonged holdings. Therefore, in this work the set of actions is defined as $Z = \{0, 10/60, 20/60, 30/60\}$ minutes.

V. Proposed Solution Method

To solve the previously described bus holding MDP problem, we use an adapted version of the Prioritized Double Deep Q-Learning (PDDQN) method [36]–[38]. Even if the problem is a Multi-Agent MDP (as presented in Fig.2), we define a common network for all agents, so that the learned policy is applied to all of them. Each agent (i.e., bus trip) generates a new experience defined by $(y_{n,s}, u_{n,s}, r_s^n, y_{n,s+1})$ every time a holding decision is made. The Q-Learning algorithm [39] estimates the Q function independently of the policy that has been applied. Since the structure of the Q function is unknown, we use a Neural Network representation based on the Deep Q-Network (DQN) algorithm [36]. A typical training of a Neural Network with the stochastic gradient descent (SGD) method requires a large number of samples. For this reason, an experience replay buffer of size B^{buffer} is used [36] in order to store the previous experiences. The Q function is updated using supervised learning, where the loss function is defined based on the Bellman equation:

$$\| r_s^n + \gamma_s Q_{\theta'}(y_{n,s+1}, u'(y_{n,s+1})) - Q_\theta(y_{n,s}, u_{n,s}) \|^2 \tag{14}$$

where

$$u'(y_{n,s+1}) = \arg\max_u Q_{\theta'}(y_{n,s+1}, u) \tag{15}$$

Eq.14 uses a second target network that is updated every T^{update} steps to improve stability. θ and θ' are the parameters of the Neural Network and the update is performed with a batch of size B^{batch} from the replay buffer. Along with the basic observation variables defined in Sec.IV-B, we include the additional feature:

$$\zeta_{n,s} = d_{n-1,s} + w_{n,s} - x_{n,s} \tag{16}$$

TABLE I

PARAMETER VALUES

T^{update} (Update period)	$10,000$
B^{buffer} (buffer size)	$500,000$
B^{batch} (batch size)	64
$\epsilon^{expl.}$ (Exploration percentage)	$1.0 \rightarrow 0.02$
Exploration iterations (total)	$400,000$
ϵ^{Adam} (Adam learning rate)	$5e^{-4}$
H (size of the hidden layer)	2048
γ_s (discount factor)	1
c (OHHR threshold variable)	1
μ (Threshold of the Huber function)	100

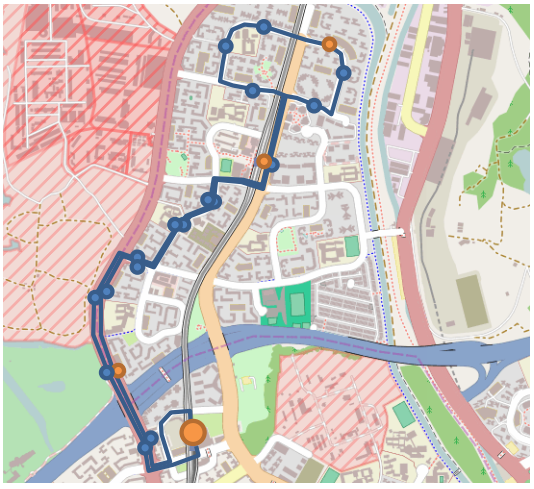

Fig. 3. Evaluation line overview where the ITP stops are in orange and the terminal is the biggest stop of the line [source: OpenStreetMaps]

for facilitating the convergence of the Q function. The used neural network is a Full Connected Neural Network (FCNN) with ReLu activation [40], while the last layer has a linear activation for enabling the mapping of the different discrete actions. The size of the hidden layer is denoted as H and for the training we use the Adam optimizer [41] with Adam learning rate ϵ^{Adam}. Due to the previous choices, we name the employed solution method Single-Agent Prioritized Double Deep Q-Network (SA-PDDQN).

VI. EXPERIMENTAL EVALUATION

A. Experimental scenario

Our case study is a main bus line (Fig.3) in Singapore with a total length of ca 7.5km. Its total travel time ranges from 30 to 45 minutes, depending on the hour of the day. The desired headway at each bus stop is reported in Tab.II. The

TABLE II

DESIRED BUS LINE HEADWAYS IN MINUTES

Time Interval	Min. allowed Headway	Max. allowed Headway	Desired Headway
Before 06:30	6	10	8
06:30-08:30	3	5	4
08:31-19:00	4	6	5
After 19:00	6	10	8

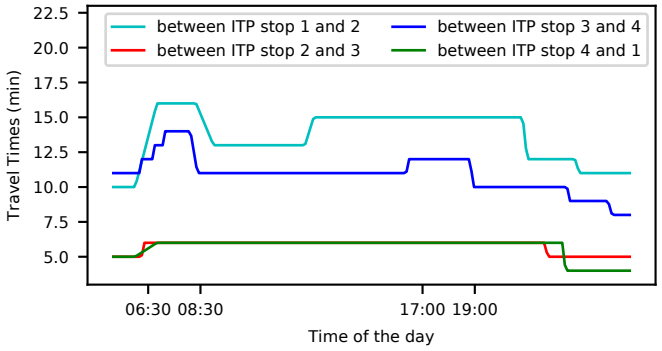

Fig. 4. Mean travel times between ITP stops at different times of the day

travel times between ITP stops are reported in Fig.4 where the mean travel times are approximated with a piecewise linear function. The two peak periods (early morning and afternoon) are from 6:30 to 8:30 and from 17:00 to 19:00.

The total number of daily trips is 220 and all of them perform a circular service. The total number of bus stops is $|S| = 22$ and 4 of them are ITP stops. The 4 ITP stops are selected from the bus operator because they exhibit a higher number of boardings/alightings. This results in a total of $|ITP| \times |N| = 4 \times 220 = 880$ bus holding variables and a total number of $|Z|^{|ITP|\times|N|} = 4^{880}$ possible solutions.

B. One-Headway-based Holding

In this study we use as comparison the well-known One-Headway-based Holding Rule (OHHR) which determines the holding time of a bus trip n at a bus stop $s \in ITP$ based on (a) its arrival time $a_{n,s}$; (b) the departure time of the preceding bus trip at the same stop $d_{n-1,s}$; and (c) the target headway $w_{n,s}$. The OHHR considers the time after which trip n can be held at stop s, $t_0 = a_{n,s} + k_{n,s}$, and a threshold variable $H_0 = d_{n-1,s} + cw_{n,s}$, with $c \in [0, 1]$. If $t_0 < H_0$, then the vehicle can leave the ITP stop immediately, $x_{n,s} = 0$; otherwise, the departure time of the current vehicle is $d_{n,s} = d_{n-1,s} + w_{n,s}$ (i.e. $x_{n,s} = d_{n-1,s} + w_{n,s} - t_0$). This policy is a local rule that seeks to minimize the deviation of the actual headway from the planned one [11].

C. Results and Evaluation

In this work, we implemented the Bus Holding environment into the *Gym* framework [42]. While this approach allows the application of state-of-the-art methods, the special multi-agent system which is proposed in this study requires the development of specific solutions. For this reason, we use the Priority Experience Replay and the implementation of the Double Deep Q-Network of *Baselines*[1] [43] to implement the SA-PDDQN method.

Fig. 5 shows the behavior of the learning process during the training stage. The exploration percentage[2], $\epsilon^{expl.}$, decreases linearly. The reward is increased along with the

[1]Baseline v0.1.5 uses tensorflow and python3

[2]the exploration percentage is the probability of selecting a random action during the training phase

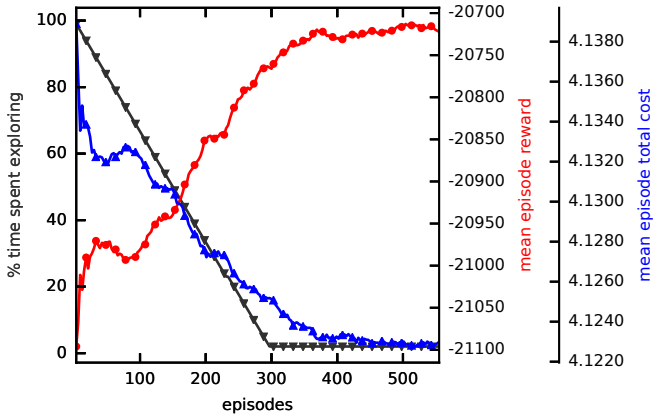

Fig. 5. Reward and total cost evolution along epochs with full a connected neural network (FCNN)

TABLE III
SA-PDDQN PERFORMANCE VS DO NOTHING VS ONE HEADWAY HOLDING RULE

Travel time variation	DN	OHHR	SA-PDDQN
1 %	4.140	4.125	**4.123**
5 %	4.185	4.197	**4.180**
10 %	4.430	4.479	**4.399**
20 %	5.294	5.474	**4.726**

increase of iterations, while the total episode[3] cost of the sum of the cost functions $f_1(X)+f_2(X)$ is reduced. At the early stage, the cost and reward oscillate due to the exploration. After the early stage, they are more stable.

Table III reports the performance of the metrics of Eq.6 and Eq.8 (i.e., sum of functions $f_1(X)$ and $f_2(X)$) during the testing phase where the bus holding times based on the OHHR method and the SA-PDDQN method are tested for 200 different episodes. The network is initially trained using 400 episodes and is tested using 200 more episodes. To account for different travel time conditions, we generated different scenarios and the travel times between stops were allowed to vary from their expected values by 1%, 5%, 10% and 20%.

From table III, one can notice that the proposed approach, SA-PDDQN, outperforms the do-nothing[4] (DN) and the OHHR. One can also notice that the OHHR method becomes counterproductive when the level of travel time variation is increased. This is due to the cost function $f_2(X)$ which is not considered by the OHHR because it focuses only on the minimization of the headway deviations from the planned values. As a result, the value of the cost function $f_2(X)$ is higher when using the OHHR method instead of the SA-PDDQN in scenarios with significant travel time variations as presented in Fig.6.

The proposed SA-PDDQN method is able to mimic the OHHR for low levels of travel time variation, while considering other cost terms, such as the total trip travel time

[3]one episode is a full day of operations

[4]in the do-nothing scenario bus holdings are not allowed

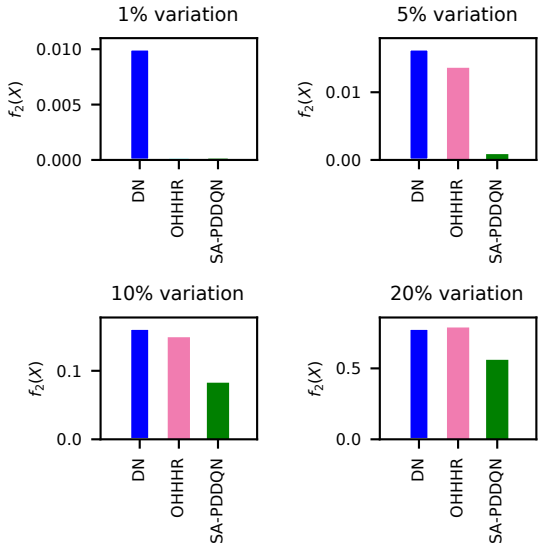

Fig. 6. Cost function $f_2(X)$ for different travel time variation levels

limits that can delay the dispatching of future trips, in case of large travel time variations. These results confirm that the proposed SA-PDDQN method is able to address more issues, such as the delay of future trips, when making a bus holding decision.

The proposed approach learns a Q-function in the form of a NN for all the trips in the same line. The method requires to learn only a new Q-function for each new line, so scales linearly with the number of lines. The training complexity though depends on the number of control stations, number of actions and on the rate of convergence of the learning process.

D. Concluding Remarks

This work introduced a mathematical program for the bus holding problem which is based on the bus movement model and the cost functions that measure the service regularity and the trip delays due to excessive trip travel times. Given the computational intractability of the bus holding problem, we introduced a RL model for solving the bus holding multi-agent problem where the agents share the same Q-function. For implementing the RL, this work introduced a reward function and designed a Neural Network architecture.

After the training phase of the experimentation, the RL method was compared against the OHHR method and its performance was evaluated. During the testing phase, the proposed method outperformed the OHHR method and the do-nothing scenario proving its stability in both low and high travel time variation scenarios.

The proposed approach can be an important step for the introduction of data-driven methods in order to obtain more holistic solutions for the computationally intractable bus holding problem in real-time. In future research, more advanced architectures of the Neural Network and their effect to the RL performance can be investigated. The extension of this approach for addressing more performance indicators, such as the occupancy rates of buses and the synchronization

levels among multiple bus services, can also be a potential topic for future research.

REFERENCES

[1] A. Ceder and N. H. Wilson, "Bus network design," *Transportation Research Part B: Methodological*, vol. 20, no. 4, pp. 331–344, 1986.

[2] O. Ibarra-Rojas, F. Delgado, R. Giesen, and J. Muñoz, "Planning, operation, and control of bus transport systems: A literature review," *Transportation Research Part B: Methodological*, vol. 77, pp. 38–75, 2015.

[3] K. Gkiotsalitis and R. Kumar, "Bus operations scheduling subject to resource constraints using evolutionary optimization," in *Informatics*, vol. 5, no. 1. Multidisciplinary Digital Publishing Institute, 2018, p. 9.

[4] K. Gkiotsalitis and O. Cats, "Reliable frequency determination: Incorporating information on service uncertainty when setting dispatching headways," *Transportation Research Part C: Emerging Technologies*, vol. 88, pp. 187–207, 2018.

[5] K. Gkiotsalitis and N. Maslekar, "Towards transfer synchronization of regularity-based bus operations with sequential hill-climbing," *Public transport*, vol. 10, no. 2, pp. 335–361, 2018.

[6] W. Leong, K. Goh, S. Hess, and P. Murphy, "Improving bus service reliability: The singapore experience," *Research in Transportation Economics*, vol. 59, pp. 40–49, 2016.

[7] S. Zhang and H. K. Lo, "Two-way-looking self-equalizing headway control for bus operations," *Transportation Research Part B: Methodological*, vol. 110, pp. 280–301, 2018.

[8] G. Sánchez-Martínez, H. Koutsopoulos, and N. Wilson, "Real-time holding control for high-frequency transit with dynamics," *Transportation Research Part B: Methodological*, vol. 83, pp. 1–19, 2016.

[9] X. Chen, B. Hellinga, C. Chang, and L. Fu, "Optimization of headways with stop-skipping control: a case study of bus rapid transit system," *Journal of advanced transportation*, vol. 49, no. 3, pp. 385–401, 2015.

[10] H. Zhang, S. Zhao, Y. Cao, H. Liu, and S. Liang, "Real-time integrated limited-stop and short-turning bus control with stochastic travel time," *Journal of Advanced Transportation*, vol. 2017, 2017.

[11] L. Fu and X. Yang, "Design and implementation of bus-holding control strategies with real-time information," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1791, pp. 6–12, 2002.

[12] M. Asgharzadeh and Y. Shafahi, "Real-time bus-holding control strategy to reduce passenger waiting time," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2647, pp. 9–16, 2017.

[13] C. F. Daganzo, "A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons," *Transportation Research Part B: Methodological*, vol. 43, no. 10, pp. 913–921, 2009.

[14] L. A. Koehler, W. Kraus, and E. Camponogara, "Iterative quadratic optimization for the bus holding control problem," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1568–1575, 2011.

[15] M. D. Hickman, "An analytic stochastic model for the transit vehicle holding problem," *Transportation Science*, vol. 35, no. 3, pp. 215–237, 2001.

[16] X. J. Eberlein, N. H. Wilson, and D. Bernstein, "The holding problem with real–time information available," *Transportation science*, vol. 35, no. 1, pp. 1–18, 2001.

[17] J. J. Bartholdi and D. D. Eisenstein, "A self-coördinating bus route to resist bus bunching," *Transportation Research Part B: Methodological*, vol. 46, no. 4, pp. 481–491, 2012.

[18] O. Cats, A. Larijani, Á. Ólafsdóttir, W. Burghout, I. Andreasson, and H. Koutsopoulos, "Bus-holding control strategies: simulation-based evaluation and guidelines for implementation," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2274, pp. 100–108, 2012.

[19] C. F. Daganzo and J. Pilachowski, "Reducing bunching with bus-to-bus cooperation," *Transportation Research Part B: Methodological*, vol. 45, no. 1, pp. 267–277, 2011.

[20] J. Zhao, S. Bukkapatnam, and M. M. Dessouky, "Distributed architecture for real-time coordination of bus holding in transit networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 4, no. 1, pp. 43–51, 2003.

[21] Q. Chen, E. Adida, and J. Lin, "Implementation of an iterative headway-based bus holding strategy with real-time information," *Public Transport*, vol. 4, no. 3, pp. 165–186, 2013.

[22] M. Dessouky, R. Hall, A. Nowroozi, and K. Mourikas, "Bus dispatching at timed transfer transit stations using bus tracking technology," *Transportation Research Part C: Emerging Technologies*, vol. 7, no. 4, pp. 187–208, 1999.

[23] J. Strathman, K. Dueker, T. Kimpel, R. Gerhart, K. Turner, P. Taylor, S. Callas, D. Griffin, and J. Hopper, "Automated bus dispatching, operations control, and service reliability: Baseline analysis," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1666, pp. 28–36, 1999.

[24] K. Gkiotsalitis and N. Maslekar, "Multiconstrained timetable optimization and performance evaluation in the presence of travel time noise," *Journal of Transportation Engineering, Part A: Systems*, vol. 144, no. 9, p. 04018058, 2018.

[25] K. Gkiotsalitis and N. Maslekar, "Improving bus service reliability with stochastic optimization," in *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*. IEEE, 2015, pp. 2794–2799.

[26] K. Gkiotsalitis and N. Maslekar, "Sequential evolutionary optimization for improving regularity-based bus services," in *Transportation Research Board 96th Annual Meeting*, no. 17-00463, 2016.

[27] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[28] R. J. Aumann and J. H. Dreze, "Cooperative games with coalition structures," *International Journal of game theory*, vol. 3, no. 4, pp. 217–237, 1974.

[29] G. Chalkiadakis, E. Elkind, E. Markakis, M. Polukarov, and N. R. Jennings, "Cooperative games with overlapping coalitions," *Journal of Artificial Intelligence Research*, vol. 39, no. 1, pp. 179–216, 2010.

[30] F. A. Oliehoek, C. Amato, *et al.*, *A concise introduction to decentralized POMDPs*. Springer, 2016, vol. 1.

[31] Y. Xuan, J. Argote, and C. F. Daganzo, "Dynamic bus holding strategies for schedule reliability: Optimal linear control and performance analysis," *Transportation Research Part B: Methodological*, vol. 45, no. 10, pp. 1831–1845, 2011.

[32] M. Trompet, X. Liu, and D. Graham, "Development of key performance indicator to compare regularity of service between urban bus operators," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2216, pp. 33–41, 2011.

[33] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[34] P. J. Huber, "Robust estimation of a location parameter," *The annals of mathematical statistics*, pp. 73–101, 1964.

[35] J. Cortés, L. Jaillet, and T. Siméon, "Molecular disassembly with rrt-like algorithms," in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 3301–3306.

[36] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[37] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized Experience Replay," *arXiv:1511.05952 [cs]*, Nov. 2015, arXiv: 1511.05952. [Online]. Available: http://arxiv.org/abs/1511.05952

[38] H. van Hasselt, A. Guez, and D. Silver, "Deep Reinforcement Learning with Double Q-learning," *arXiv:1509.06461 [cs]*, Sept. 2015, arXiv: 1509.06461. [Online]. Available: http://arxiv.org/abs/1509.06461

[39] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[40] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[41] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Dec. 2014, arXiv: 1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[42] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016.

[43] P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, and Y. Wu, "Openai baselines," https://github.com/openai/baselines, 2017.