Transportation Research Part C 121 (2020) 102815

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

An analytic solution for real-time bus holding subject to vehicle capacity limits

K. Gkiotsalitis *, E.C. van Berkum

University of Twente, Center for Transport Studies, Horst - Ring Z-222, P.O. Box 217, 7500 AE Enschede, the Netherlands

ARTICLE INFO

Keywords:
Bus holding
Operational control
Public transit
Capacity
Even headways

ABSTRACT

This study focuses on single variable optimization approaches which determine the holding time of a vehicle when it is ready to depart from a bus stop. Up to now, single variable optimization methods resort to rule-based control logics to equalize the inter-departure headways or adhere to the target headway values. One of them is the two-headway-based control logic which determines the holding time of a bus based on its headway with its preceding and following bus without addressing other implications, such as overcrowding. To rectify this, we introduce a new model for the single variable bus holding problem that considers the passenger demand and vehicle capacity limits. Then, we reformulate this problem to an easier-to-solve program with the use of slack variables and introduce an analytic solution that can determine the holding time of a vehicle at the respective bus stop. Our analytic solution does not add a computational burden to the two-headway-based control logic and can be applied in real time. The operational benefit of our bus holding approach compared to other analytic solutions that do not consider the vehicle capacity is investigated using actual data from bus line 302 in Singapore.

1. Introduction

Decisions regarding the operations of bus services are made at different planning stages. At the tactical planning stage, one has to determine the frequency (Yu et al., 2009; Gkiotsalitis and Cats, 2018), the timetable (Sun et al., 2015; Wu et al., 2016), and the crew and vehicle schedules (Wren and Rousseau, 1995; Gintner et al., 2005; Kliewer et al., 2006) of every bus line. Tactical plans are communicated well in advance, and all stakeholders (i.e., public transport authorities/operators, bus drivers, passengers) are aware of them prior to the start of the daily operations (Ceder, 2007).

The fixed service interval (time headway) of every bus line is determined from the tactical planning stage and is equal to the inverse of the service frequency (Ceder, 2007). The time headway of two trips, which is the time difference between the time instances they were at the same location, will henceforth be simply called “headway”. The main challenge in high-frequency services with more than 5 trips per hour is to maintain the planned headways among buses at every bus stop (Trompet et al., 2011). If the demand and the travel times of all bus trips operating in a service line are equal and stable, bus trips will maintain their even headways at all downstream stops. This will result in a regular service where the actual passenger waiting times at stops meet the passengers’ expectations. Nevertheless, travel time and passenger demand variations during the actual operations result in unreliable services (Chen et al., 2009; Daganzo, 2009). Knoppers and Muller (1995), Berrebi et al. (2018), Gkiotsalitis (2020a) and Knoppers and Muller (1995) have shown

* Corresponding author.
E-mail addresses: k.gkiotsalitis@utwente.nl (K. Gkiotsalitis), e.c.vanberkum@utwente.nl (E.C. van Berkum).

https://doi.org/10.1016/j.trc.2020.102815
Received 2 December 2019; Received in revised form 18 August 2020; Accepted 19 September 2020
Available online 27 October 2020

that the fixed dispatching intervals cannot be maintained at all stops. Indeed, even if buses are dispatched according to their planned headways, their headways are expected to deviate from their scheduled values as they are moving towards downstream stops (Hans et al., 2015). This leads to irregular services where buses are too close or too far away from each other, and thus fail to maintain their scheduled headway(s).

To address the adverse effects of the demand and travel time variability, several flexible scheduling approaches have emerged over the past 40 years. Such flexible approaches have a shifted focus towards operational control that reacts to changes in quasi-real-time. Operational control includes a variety of options, such as bus holding (Bartholdi and Eisenstein, 2012; Delgado et al., 2012), stop-skipping or refused boardings (Delgado et al., 2009; Liu et al., 2013; Chen et al., 2015), short-turning (Cortés et al., 2011), inter-lining (Gkiotsalitis et al., 2019), re-scheduling (Gkiotsalitis, 2020b), and speed control (Daganzo and Pilachowski, 2011; Muñoz et al., 2013). All options aim at improving the reliability of services during the actual operations.

In this study, we specifically focus on the problem of real-time bus holding that holds buses at specific bus stops to reduce the deviation between the actual and the planned headways. In its simplest form, bus holding holds a trip n at a stop s for a time period $x \geqslant 0$ if its actual headway with its preceding trip, $n-1$, is lower than the planned headway, H_s. This is the well-known naïve one-headway-based holding method which strives to maintain the planned headway between a trip n and its preceding one, $n-1$ (Fu and Yang, 2002). Other approaches do not consider only the headway between one bus trip, n, and its preceding trip, $n-1$, but also the headway with the following trip, $n+1$. Such approaches are known as two-headway-based methods (Fu and Yang, 2002).

An entirely different line of research determines the holding times of multiple bus trips, instead of only trip n, following a periodic optimization approach (Gkiotsalitis and Cats, 2019). Periodic optimization approaches consider multiple decision variables and are based on iterative, finite-horizon optimization(s) of a bus holding model. At time t, the current state of each bus (i.e., current positions of running trips) is used as input and, together with the expected travel times within a relatively short time horizon $t+T$, the holding times of multiple running trips are determined. This is equivalent to scheduling the bus holding times of all running trips within a short time horizon with the use of travel time expectations (Eberlein et al., 2001). The holding times can be recomputed every time new information becomes available. This will result in receding horizon control, or "rolled" rolling horizon optimization (Eberlein et al., 2001).

In "rolled" rolling horizon optimization, most of the computed holding times might not be implemented in practice by the time new information becomes available if one implements holding control at isolated control point stops (see Eberlein et al. (2001)). The reason behind this is that the values of those holding times can be updated in a new "rolled" rolling horizon if we receive new information regarding the travel conditions within a very short time. We note, however, that if we adopt more recent approaches that apply holding control at every bus stop the probability of not implementing a holding suggestion before new information becomes available reduces significantly. Rolling horizon optimization has similarities to model predictive control (MDP), where multiple decisions are made but only some of them have the chance to be implemented by the time new information becomes available triggering a repeat of the optimization process (Nikolaou, 2001). In contrast to periodic optimization in rolling horizons, in this study we propose an analytic solution for the bus holding problem under capacity limitations that can determine (immediately) the holding time of a single bus trip upon its arrival at a bus stop. Our analytic solution differs from other analytic solutions or rule-based holding approaches because it considers the bus load variations and vehicle capacity limitations. Our analytic solution is a closed-form expression of arrival times, boardings, alightings, passenger arrival rates, and vehicle capacity limits.

The remainder of this paper is structured as follows: in Section 2, we provide the literature review in periodic optimization and analytic approaches for the bus holding problem. In Section 3, we model the bus holding problem with the objective of maintaining the service regularity while meeting the vehicle capacity limits. This problem is proved to be nonlinear and non-smooth; thus, it cannot be solved to global optimality because its functions are not differentiable at every point in their domain. In Section 4, we reformulate the aforementioned bus holding problem by introducing slack variables that are commonly used in mathematical modeling to transform inequality constraints into equality constraints. Then, we prove that its reformulated version has a globally optimal solution. In Section 5, we develop an analytic solution for the reformulated program. In Section 6, we compare our approach against the two-headway-based and the self-equalizing bus holding methods - which are also based on closed-form expressions and do not have any computation costs. We also explore the performance sensitivity of our holding solution to demand and travel time variations using real data from the high-frequency bus line 302 in Singapore. The main findings and the limitations of this study are discussed in Sections 7 and 8.

2. Literature review

Control methods for bus holding have been studied since the early 1970s (see Osuna and Newell (1972), Newell (1974)). Nevertheless, the bus holding problem remains a prominent research topic because of its inherent complexity. Newell (1974) considered only one control point at which buses can be intentionally delayed, and devised a strategy for holding a bus to minimize the average waiting time of the passengers. The strategy envisioned to correct the random fluctuations in trip travel times so that the headways will not become unequal and lead to bunching.

Typical objectives of bus holding methods are headway adherence Rossetti and Turitto (1998), Gkiotsalitis and Cats (2019), headway regularity Bartholdi and Eisenstein (2012), Daganzo (2009), and the minimization of passenger waiting and in-vehicle times Delgado et al. (2009), Delgado et al. (2012), Sáez et al. (2012). As previously stated, two different directions of research have emerged. One research direction models the bus holding problem as a multivariable, periodic optimization problem where decisions about the holding times concern the entirety of trips that will operate in a short horizon, $t+T$. To achieve that, information about the current trajectories of bus trips and their predicted values in the short future is incorporated in the respective mathematical programs (Eberlein et al., 2001; Gkiotsalitis, 2019b). The second direction of research determines the holding time of a single trip when it arrives at a

control point stop (event-based control). Such single variable optimization problems lead to closed-form expressions that can determine the holding time of a trip based on its headway with its preceding/following trips (Hickman, 2001; Fu and Yang, 2002; Van Oort et al., 2010).

In the remainder of the literature review, we discuss the multivariable and the single variable bus holding optimization approaches. The former approaches are typically used for periodic optimization, while the latter for event-based, real-time control. In Sections 2.1 and 2.2, a distinction is made between approaches that consider the vehicle capacity in the optimization process and the ones that do not.

2.1. Bus Holding without considering the vehicle capacity

Although bus holding methods that do not consider the vehicle capacity are not the primary focus of our work, we hereby discuss the main multivariable and single variable bus holding optimization methods that belong to this category. Multivariable bus holding approaches that try to determine the holding times of multiple bus trips within a time period $t+T$ might not have an analytic solution due to the complexity of the respective mathematical programs. For this reason, we report past works that do not offer an analytic solution and works that offer an analytic solution in the separate Sections 2.1.1 and 2.1.2.

2.1.1. Mathematical Programs of the bus holding problem without Analytic Solution

Examples of multivariable bus holding optimization methods are the periodic optimization mathematical programs of Eberlein (1995), Eberlein et al. (2001), Shen and Wilson (2001), Sánchez-Martínez et al. (2016). Such mathematical programs determine simultaneously the holding times of all buses that are expected to operate within a rolling horizon. The optimized holding times are updated in rolled rolling horizons when new information becomes available.

Eberlein et al. (2001) considered real-time information and assumed that travel times and passenger arrival rates remain constant in rolling horizons with short time duration. The holding problem of all running buses was modeled as a quadratic program with the objective to minimize the total passenger waiting times. Sáez et al. (2012) utilized a dynamic objective function and a predictive model of the bus system to make decisions on bus holding and stop-skipping (known also as expressing). The uncertain passenger demand was included in the model as a disturbance. The resulting optimization problem was NP-hard and was solved using an ad hoc implementation of a Genetic Algorithm. Gkiotsalitis (2019b) used also a metaheuristic from the area of evolutionary optimization to solve an NP-Hard program that returns holding times at the first stop of the line which minimize the waiting times of passengers under regulatory constraints.

Zolfaghari et al. (2004) developed a mathematical control model for bus holding using real-time information regarding the locations of buses along a specified route. Their resulting mathematical program was solved with metaheuristics (specifically, simulated annealing). Hickman (2001) used the stochastic model developed by Marguier (1985) for deriving the trajectories of buses on a single route. Using Marguier's model, Hickman (2001) developed a bus holding algorithm that is applied each time a bus arrives at the control point stop. To this end, Marguier's model was used to approximate the trajectories of all "upstream" buses. The bus holding time was selected using a line search method because obtaining an analytic solution was not possible given the complexity of deriving the first-order conditions of the optimization problem.

2.1.2. Models of the bus holding problem with Analytic Solutions

In this sub-section, we report works that proposed closed-form expressions for the determination of the bus holding time(s). The closed-form expressions can be a result of analytic solutions of optimization problems or rule-based approaches that determine the holding times based on pre-defined threshold values.

Fu and Yang (2002) tested two of the most common rule-based bus holding strategies: (i) the one-headway-based control where a bus is held at a control point stop if its time headway with its preceding bus is lower than a pre-defined threshold; and (ii) the two-headway-based control that considers the time headway of a bus with its preceding and following bus. Similarly, Sun and Hickman (2004) set the holding time of a bus trip to zero if its predicted headway with its following bus is less than or equal to the scheduled headway. When the actual vehicle headway is less than the prescribed minimum headway, the following vehicle will be delayed until the minimum headway requirement can be satisfied.

Even if its focus was on speed control, we also report the work of Daganzo and Pilachowski (2011). Daganzo and Pilachowski (2011) proposed an adaptive control scheme that adjusts a bus cruising speed in real-time based on both its front and rear spacings. In line with other closed-form approaches, it had a simple and decentralized logic enabling to correct the effect of traffic disruptions in real-time. Bartholdi and Eisenstein (2012) proposed an analytic bus holding solution which changes the headway of each newly arrived bus to the weighted average of its former headway and that of the trailing bus. This approach tends to re-equalize the headways after any disturbance. Thus, its objective is to maintain the headway regularity and not to adhere to a scheduled (target) headway. In Bartholdi and Eisenstein (2012) the holding decisions constantly adjust and re-equalize the headways.

Berrebi et al. (2015), Berrebi et al. (2018) proposed a method consisting of identifying probabilistically the bus that will arrive the latest to a particular point. Then, each preceding bus is held to prevent the lagging bus from departing with a big gap. Van Oort et al. (2010) also tested schedule-based and headway-based holding strategies where the solution was expressed as a closed-form expression of arrival times and scheduled headways. They tested the importance of setting a maximum holding time and a reliability buffer time in tram line 9 in The Hague.

Wu et al. (2017) incorporated the passenger demand into the estimation of bus trajectories and addressed the single variable bus holding problem with the use of the one-headway-based holding logic. In the one-headway-based holding of Wu et al. (2017), a bus is

held if the headway with its preceding bus is less than the scheduled headway - otherwise, it is dispatched immediately. Although Wu et al. (2017) incorporates the demand and the capacity of vehicles in the calculation of bus dwell times, the objective of their one-headway-based control logic does not consider the improvement of bus loads and focuses on the service regularity. For this reason, the study of Wu et al. (2017) is assigned to the category of studies that do not use the violation of capacity limits as an optimization objective.

2.2. Bus Holding Methods that consider the Vehicle Capacity limits

Previously, we reviewed bus holding methods that do not account for the passenger demand and vehicle capacity limitations. In this sub-section, we review past works that, similar to our approach, consider the capacity limitations in the bus holding optimization process.

Sánchez-Martínez et al. (2016) formulated a mathematical model to produce a plan of holding times for all running vehicles in a rolling horizon that caters for the passenger demand. Its effectiveness was evaluated within a simulation environment. The objective function in that model was not convex and did not allow the derivation of an analytic solution. Instead, Sánchez-Martínez et al. (2016) employed the optimization algorithm of Powell (2009) to derive local minima of the nonlinear objective function.

Delgado et al. (2009) developed a mathematical program that incorporates vehicle-capacity constraints. As in Sánchez-Martínez et al. (2016), they calculated the holding times of all vehicles in a rolling horizon resulting in a multivariable decision problem. In a later work, Delgado et al. (2012) also addressed the problem of determining the holding times of all running buses on a rolling horizon. Their objective was to minimize the total times experienced by all passengers in the system resulting in a non-convex, nonlinear objective function. Then, they performed a simulation-based evaluation of two control policies applied within a rolling horizon framework: (i) vehicle holding that does not consider boarding limits, and (ii) holding combined with boarding limits, in which the number of boarding passengers at any stop can be limited. The respective mathematical programs were solved using MINOS as an optimization solver.

Luo et al. (2017) proposed a nonlinear optimization model to improve the headway adherence considering bus capacity. In Luo et al. (2017) the bus capacity was not explicitly modeled as a problem constraint. Instead, Luo et al. (2017) aimed at maintaining a stable passenger load within the buses. Li et al. (2019) also considered the demand uncertainty that can affect the vehicle loads when applying bus holding. However, the vehicle capacity was not explicitly considered in their problem formulation. The same holds true in the target-headway-based holding approach of He et al. (2020). In He et al. (2020), the capacity limit is not explicitly considered in the problem formulation but it is used to stop loading passengers until unoccupied space is available at a later stop. Finally, Koehler et al. (2018) proposed an integrated holding and priority control model for bus rapid transit services. Koehler et al. (2018) considered bus capacity in the model formulation. This resulted in a mixed integer nonlinear program that cannot be solved effectively in large scale problem instances. In addition, the model of Koehler et al. (2018) had a non-convex objective function which cannot guarantee the convergence to a globally optimal solution. Thus, Koehler et al. (2018) simplified their original problem by replacing variables with approximated constant values resulting in a convex objective function, and their simplified problem was solved iteratively to mitigate the variables' approximation errors.

2.3. Contribution of our work

From the above studies, bus holding control methods with analytic solutions (e.g., methods that do not require the solution of a mathematical program every time a decision needs to be made) focus on improving the regularity of bus operations and do not consider vehicle capacity limits. Additionally, bus holding works that consider capacity limitations result in multivariable optimization problems that do not have analytic solutions. Our study contributes in this area by proposing a mathematical formulation for the single variable bus holding problem that, after several reformulations, is proven to have an analytic solution. Hence, our approach can determine immediately the holding time of a vehicle when it arrives at a control point stop without increasing the computational burden of past analytic solutions that did not consider the vehicle capacity limits (e.g., one-headway-based, two-headway-based, or self-equalizing-based bus holding methods).

The incremental contributions of this work to the state-of-the-art are:

- the introduction of a nonlinear model for the single variable bus holding problem under capacity limitations and the analysis of its mathematical properties;
- the reformulation of the nonlinear, non-smooth mathematical program to a program with a quadratic objective function and linear (in) equality constraints that can be solved to global optimality;
- the introduction of an analytic solution that determines the holding time based on the arrival times, boardings, alightings, passenger arrival rates, vehicle capacity limits, and scheduled headways;
- the investigation of its operational performance compared to other analytic solutions using operational data from bus line 302 in Singapore.

3. Problem definition and mathematical program

Proceeding to the introduction of our method, we present the main assumptions of our work, which are also commonly used in past literature related to the bus holding problem of high-frequency services:

Table 1
Notation.

Sets/Indices	
$S = \langle 1,2,\ldots \rangle$	ordered set of bus stops.
n	index of the bus trip for which a holding decision needs to be made at the current time instance.
$n-1$	index of the preceding bus trip of trip n.
$n+1$	index of the following bus trip of trip n.
s	specific bus stop at which a holding decision for trip n needs to be made. Note that $s \in S \backslash \{1, \lvert S \rvert\}$.
Parameters	
t	time when bus trip n has completed its boardings/alightings at stop s and is ready to depart if there is no further holding.
$d_{n-1,s}$	departure time of trip $n-1$ from stop s.
λ_s	arrival rate of passengers at stop s (i.e., passengers per sec).
c_j	capacity of bus trip j, where $j \in \{n-1, n, n+1\}$.
ϕ_n	observed bus load of trip n at time t including the number of passengers who are refused to board trip n at stop s due to overcrowding. By definition, ϕ_n can be greater than c_n.
$\widetilde{l}_{n+1}$	expected bus load of trip $n+1$ at the time of its arrival at stop s.
$\widetilde{\beta}_{n+1}$	expected passenger alightings of bus trip $n+1$ at stop s.
$\widetilde{a}_{n+1,s}$	expected arrival time of trip $n+1$ at stop s.
H_s	planned inter-departure headway of adjacent trips at stop s. Note that H_s might have the same value at all stops if the planned headway is not stop-dependent.
t_b	required time for each passenger boarding.
t_a	required time for each passenger alighting.
ζ	maximum allowed holding time due to the inconvenience caused to on-board passengers.
Decision Variable	
x	holding time of trip n at stop s. Note that $\{x \in \mathbb{R} \lvert 0 \leqslant x \leqslant \zeta\}$.
Variables	
$d_{n,s}$	departure time of trip n from stop s. Note that $d_{n,s} \triangleq t + x$.
$\widetilde{d}_{n+1,s}$	expected departure time of trip $n+1$ from stop s.
l_n	stranded passengers by bus trip n at stop s.

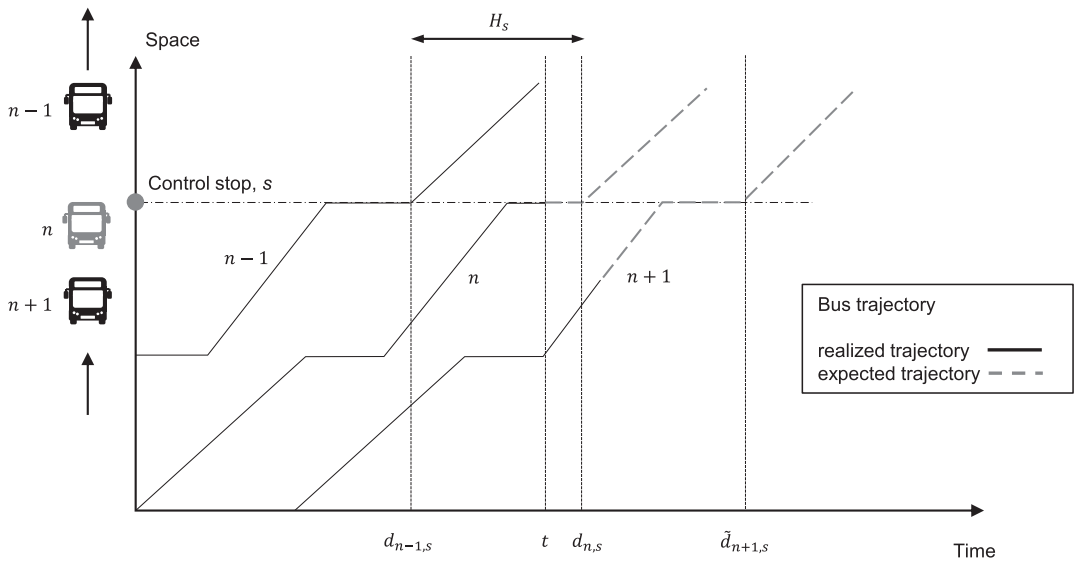

Fig. 1. Realized and expected trajectories of the preceding, $n-1$, and following, $n+1$, bus trips of trip n. The holding decision of trip n at stop s is made at time t when trip n has completed all its boardings/alightings.

(1) In high-frequency services, passengers who cannot board a bus will wait for the next trip of the same bus line because their waiting times are relatively small (Delgado et al., 2009; Delgado et al., 2012; Muñoz et al., 2013).
(2) Passengers cannot coordinate their arrivals at stops to the arrival times of buses at high-frequency services (Berrebi et al., 2015). Thus, we assume a demand-based passenger arrival rate, λ_s, at any stop s (Fu and Yang, 2002; Delgado et al., 2012).

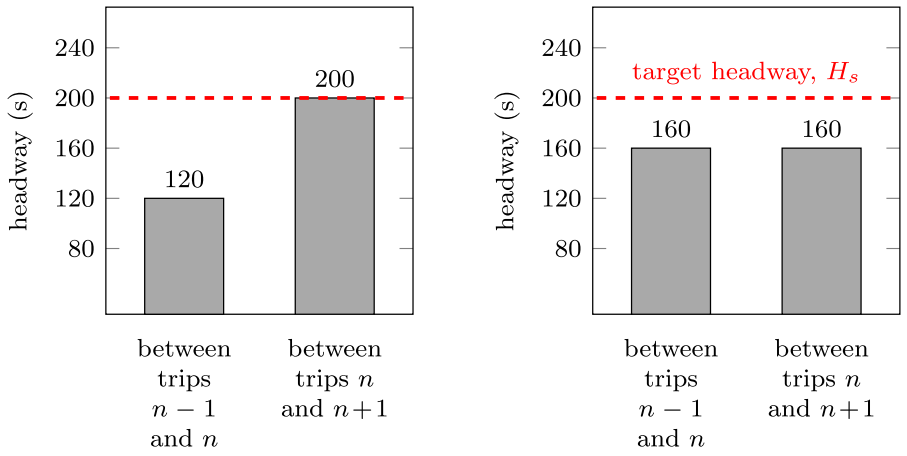

Fig. 2. Example of headways between trips $n-1, n$ and $n, n+1$ that yield the same absolute headway deviation, but a different squared headway deviation.

(3) Overtaking among buses is permitted while they travel from stop to stop. When two buses are dwelling at the same stop, we assume that they are not allowed to overtake each other for simplifying the modeling of passenger boardings (Luo et al., 2017; Koehler et al., 2018).

To formulate our bus holding problem that considers vehicle capacity limits, we introduce the notation of Table 1.

In our event-based bus holding problem, we decide about the holding time of any trip n when it is at stop $s \in S \backslash \{1, |S|\}$ and has completed all its boardings/alightings (Fig. 1). Thus, an event is defined as the occasion when a bus is at a control point stop and has completed all its boardings/alightings. Note that the arrival time of trip $n+1$ at stop s can be estimated with the use of prediction methods. Past attempts on predicting the arrival times of buses have included artificial neural networks (ANN) or support vector machines (SVM) (see Chen et al. (2004), Van Lint et al. (2005), Vlahogianni et al. (2005), Bin et al. (2006)), non-parametric regression models (NPR), (see Chang et al. (2010)) and Kalman filters (see Chien and Kuchipudi (2003), Shalaby and Farhan (2004)). Other approaches include the use of a Bayesian committee of neural networks to predict travel times with confidence intervals (van Hinsbergen et al., 2009).

3.1. Problem objective

The objective of the bus holding problem in high-frequency services is to adhere to the target (scheduled) headways. When we determine the holding time of trip n at stop s, we strive to minimize the *squared deviation* between the realized/expected headways with its adjacent trips, $n-1, n+1$, and the ideal headway, H_s. This is expressed in Eq. (1) where $(t+x)$ is the determined departure time of trip n from stop s. Note that $\widetilde{d}_{n+1,s}$ is an expected value because trip $n+1$ has not arrived at stop s when the holding decision of trip n is made.

$$f\left(x\right) \triangleq \left(\left(t+x\right) - d_{n-1,s} - H_s\right)^2 + \left(\widetilde{d}_{n+1,s} - \left(t+x\right) - H_s\right)^2 \tag{1}$$

We should note here that Eq. (1) uses the *squared deviation* between the expected/realized headways and their target values. The reason behind this is that if we use the *absolute deviation*, $\left|(t+x) - d_{n-1,s} - H_s\right| + \left|\widetilde{d}_{n+1,s} - (t+x) - H_s\right|$, then we do not balance the headway deviation among trips. For instance, consider the two cases in Fig. 2. If the objective is to minimize $\left|(t+x) - d_{n-1,s} - H_s\right| + \left|\widetilde{d}_{n+1,s} - (t+x) - H_s\right|$, then the solutions in the left and the right sub-figures are equivalent and yield an absolute headway deviation of 80 s. In contrast, if we use the squared deviation of headways, the solution in the right sub-figure, which distributes the headways more evenly, will be the selected option with a performance of $(160-200)^2 + (160-200)^2$ s^2 $< (120-200)^2 + (200-200)^2$ s^2. This is in line with the key performance indicators used to monitor the regularity of bus services (Newell, 1974; Trompet et al., 2011).

3.2. Constraints and infeasibility

A first constraint when we consider the vehicle capacity limits is that trip n cannot serve more passengers than its capacity, c_n. This can be expressed as:

$$\phi_n + x\lambda_s \leqslant c_n \tag{2}$$

where $x\lambda_s$ is the number of additional passengers that are willing to board bus trip n if it is held at stop s for time x after it completes its boardings/alightings. Additionally, ϕ_n is the sum of the bus load of trip n and the number of (potentially) stranded passengers when it has completed its boardings/alightings at stop s.

In some problem instances where $\phi_n > c_n$, constraint (2) cannot be satisfied even if holding time x is equal to zero. This will result in refused boardings. Hence, the number of stranded passengers, l_n, by bus trip n at stop s can be expressed as:

$$l_n \triangleq \max(0, \phi_n + x\lambda_s - c_n) \tag{3}$$

Since constraint $\phi_n + x\lambda_s \leqslant c_n$ cannot be always satisfied, it can be perceived as a *soft* constraint which is allowed to be violated if, and only if, our holding time x cannot ensure that there are no stranded passengers by bus trip n at stop s. This soft constraint is added to the objective function as a penalty term $M_1 \max(0, \phi_n + x\lambda_s - c_n)$, where M_1 is a very large positive number:

$$f\left(x\right) \triangleq \left((t+x) - d_{n-1,s} - H_s\right)^2 + \left(\widetilde{d}_{n+1,s} - \left(t+x\right) - H_s\right)^2 + M_1 \max\left(0, \phi_n + x\lambda_s - c_n\right) \tag{4}$$

Note that the very large positive number M_1 in the penalty term $M_1 \max(0, \phi_n + x\lambda_s - c_n)$ ensures that the satisfaction of constraint $\phi_n + x\lambda_s \leqslant c_n$ is prioritized over $\left((t+x) - d_{n-1,s} - H_s\right)^2 + \left(\widetilde{d}_{n+1,s} - (t+x) - H_s\right)^2$. Indeed, if $\phi_n + x\lambda_s \leqslant c_n$, then this solution does not add any penalty to the objective function since $M_1 \max(0, \phi_n + x\lambda_s - c_n) = 0$. In reverse, when $\phi_n + x\lambda_s > c_n$, the penalty term penalizes the objective function by a very large number $M_1(\phi_n + x\lambda_s - c_n)$ and directs the program towards another solution x that reduces the value of $M_1 \max(0, \phi_n + x\lambda_s - c_n)$ as much as possible. Consequently, a solution x that minimizes the objective function would be such that the number of stranded passengers by bus trip n at stop s is reduced to the greatest extent possible. That is to say, avoiding refused passenger boardings has a higher priority than meeting the target headway.

A second constraint is related to the vehicle capacity limit of the following trip, $n+1$. Note that the vehicle capacity limit of the preceding trip, $n-1$, is not considered because our decision variable, x, cannot affect its value since it has already served stop s. When trip $n+1$ arrives at stop s it has a bus load $\widetilde{l}_{n+1}$ and is expected to alight $\widetilde{\beta}_{n+1}$ passengers. Because of the time needed for the alightings, $\widetilde{\beta}_{n+1} t_a$, we get $\widetilde{\beta}_{n+1} t_a \lambda_s$ more passenger boardings assuming that passengers use the same door channel for boardings and alightings. In addition, the stranded passengers by trip n, l_n, are willing to board trip $n+1$. Furthermore, given the passenger arrival rate λ_s, $(\widetilde{d}_{n+1,s} - (t+x))\lambda_s$ more passengers will be willing to board trip $n+1$, where $(\widetilde{d}_{n+1,s} - (t+x))$ is the inter-departure headway between trips n and $n+1$. $\widetilde{d}_{n+1,s}$ is a variable that depends on the expected arrival time of trip $n+1$ at stop s and its expected boardings and alightings.

Variable $\widetilde{d}_{n+1,s}$ is calculated in Eq. (7) based on the following considerations. By the time its previous trip n departs from stop s, $(t + x)$, until trip $n+1$ arrives at stop s, $(\widetilde{a}_{n+1,s})$, we have $(\widetilde{a}_{n+1,s} - (t+x))\lambda_s$ more passengers willing to board trip $n+1$. Thus, the expected bus load of trip $n+1$ when it departs from stop s is $\widetilde{l}_{n+1} - \widetilde{\beta}_{n+1} + \widetilde{\beta}_{n+1} t_a \lambda_s + l_n + (\widetilde{a}_{n+1,s} - (t+x))\lambda_s$. Note that this is the lowest possible bus load of trip $n+1$ when it departs from stop s because the holding time of trip $n+1$ at stop s is not considered at the time we make a holding decision for trip n.

Remark 1. In our study, we consider only the passengers that will arrive while boarding passengers $\widetilde{\beta}_{n+1} t_a \lambda_s + l_n + (\widetilde{a}_{n+1,s} - (t+x))\lambda_s$ and we assume that the number of passenger arrivals during subsequent boardings is negligibly small. That is to say, while boarding passengers $(\widetilde{\beta}_{n+1} t_a \lambda_s + l_n + (\widetilde{a}_{n+1,s} - (t+x))\lambda_s) t_b \lambda_s$ the number of new passengers arriving at the stop is insignificant because the time duration of $(\widetilde{\beta}_{n+1} t_a \lambda_s + l_n + (\widetilde{a}_{n+1,s} - (t+x))\lambda_s) t_b^2 \lambda_s$ is infinitesimal and $(\widetilde{\beta}_{n+1} t_a \lambda_s + l_n + (\widetilde{a}_{n+1,s} - (t+x))\lambda_s) t_b^2 \lambda_s^2 \approx 0$. This formulation offers a more accurate representation of the potential passenger boardings compared to past works that oversimplify the problem by ingoring all passenger arrivals at a stop while the bus is dwelling (see Marguier (1985), Hickman (2001), Fu and Yang (2002)).

The assumption in Remark 1 allows us to determine a closed-form expression of the expected bus load of trip $n+1$ from stop s. This bus load should be lower or equal to the capacity of the bus that operates trip $n+1$. This is expressed in the inequality constraint of Eq. (5).

$$\widetilde{l}_{n+1} - \widetilde{\beta}_{n+1} + \left(\widetilde{\beta}_{n+1} t_a \lambda_s + l_n + \left(\widetilde{a}_{n+1,s} - \left(t+x\right)\right)\lambda_s\right)\left(1 + t_b \lambda_s\right) \leqslant c_{n+1} \tag{5}$$

Note that, unlike Eq. (2), in Eq. (5) we do not consider the additional passenger boardings caused by the holding time of trip $n+1$ at stop s because we will have to decide about that holding time when trip $n+1$ arrives at stop s. Considering the capacity limit of trip $n+1$, it is conjectured that the inequality constraint of Eq. (5) cannot be always satisfied for $x \in \mathbb{R} | 0 \leqslant x \leqslant \zeta$. This is proved in Lemma Appendix 1.

Similarly to the capacity constraint of trip n, the capacity constraint of trip $n+1$ expressed in Eq. (5) can be perceived as a *soft* constraint which is allowed to be violated if, and only if, our holding time x cannot ensure that there are no stranded passengers by bus trip $n+1$ at stop s. This soft constraint is added to the objective function as a penalty term $M_2 \max\left[0, \widetilde{l}_{n+1} - \widetilde{\beta}_{n+1} + \left(\widetilde{\beta}_{n+1} t_a \lambda_s + l_n + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b \lambda_s) - c_{n+1}\right]$:

$$f\left(x\right) \triangleq \left(\left(t+x\right)-d_{n-1,s}-H_s\right)^2+\left(\widetilde{d}_{n+1,s}-\left(t+x\right)-H_s\right)^2+M_1\max\left(0,\phi_n+x\lambda_s-c_n\right) \\ +M_2\max\left[0,\widetilde{l}_{n+1}-\widetilde{\beta}_{n+1}+\left(\widetilde{\beta}_{n+1}t_a\lambda_s+l_n+\left(\widetilde{a}_{n+1,s}-t-x\right)\lambda_s\right)\left(1+t_b\lambda_s\right)-c_{n+1}\right] \tag{6}$$

Remark 2. Note that we use very large numbers M_1, M_2 to penalize the soft constraints related to the stranded passengers from bus trips n and $n+1$, respectively. Additionally, we set $M_1 \gg M_2$. $M_1 \gg M_2$ indicates that if trip n reaches its capacity limit, it will depart immediately from stop s even if this is expected to lead to the overcrowding of trip $n+1$. That is to say, we cannot hold an overcrowded bus trip, n, even if this has a positive effect to its following trip, $n+1$. This is realistic in practice because if bus trip n is held after reaching its capacity limit, it will cause inconvenience to the passengers who are refused to board while the bus is held at the stop (Trompet et al., 2011).

The expected departure time of trip $n+1$ from stop s, $\widetilde{d}_{n+1,s}$, is equal to the expected arrival time at stop s, $\widetilde{a}_{n+1,s}$, plus the required time for boardings/alightings (dwell time). The required time for boardings/alightings is $\widetilde{\beta}_{n+1}t_a$ for passenger alightings and $\left(\widetilde{\beta}_{n+1}t_a\lambda_s+l_n+(\widetilde{a}_{n+1,s}-(t+x))\lambda_s\right)(1+t_b\lambda_s)t_b$ for passenger boardings. Note that all $\left(\widetilde{\beta}_{n+1}t_a\lambda_s+l_n+(\widetilde{a}_{n+1,s}-(t+x))\lambda_s\right)(1+t_b\lambda_s)$ passengers might not be able to board trip $n+1$ at stop s if its capacity limit is reached. Hence, the required time for passenger boardings is $\min\left[\left(\widetilde{\beta}_{n+1}t_a\lambda_s+l_n+(\widetilde{a}_{n+1,s}-(t+x))\lambda_s\right)(1+t_b\lambda_s)t_b, (c_{n+1}+\widetilde{\beta}_{n+1}-\widetilde{l}_{n+1})t_b\right]$.

This results to the *expected* departure time of trip $n+1$ from stop s:

$$\widetilde{d}_{n+1,s} \triangleq \widetilde{a}_{n+1,s}+\widetilde{\beta}_{n+1}t_a+\min\left[\left(\widetilde{\beta}_{n+1}t_a\lambda_s+l_n+\left(\widetilde{a}_{n+1,s}-t-x\right)\lambda_s\right)\left(1+t_b\lambda_s\right)t_b, \\ \left(c_{n+1}+\widetilde{\beta}_{n+1}-\widetilde{l}_{n+1}\right)t_b\right] \tag{7}$$

3.3. Mathematical program

The above-mentioned constraints form the following bus holding program, (Q), that determines the holding time x of trip n at time instance t. This program can be solved every time a bus trip n is ready to depart from a bus stop s resulting in an event-based bus holding scheme.

$$\left(Q\right) \quad \min_{x} \quad f(x) \\ \text{s.t.} \quad \left(l_n, f, \widetilde{d}_{n+1,s}\right) | \left(l_n, f, \widetilde{d}_{n+1,s}\right) \text{ satisfy Eq. } \left(3\right), \left(6\right), \left(7\right) \\ 0 \leqslant x \leqslant \zeta \tag{8}$$

4. Reformulation to a quadratic program

4.1. Reformulation

Let us consider the nonlinear term $\max(0, \phi_n+x\lambda_s-c_n)$ of our objective function that appears also in the equality constraint $l_n = \max(0, \phi_n+x\lambda_s-c_n)$ expressed in Eq. (3). Note that the "max" term introduces non-smoothness to our objective function and our equality constraint. To rectify this, we introduce a slack variable ν_1 that, due to its bounds and the direction of optimization, will take the value $\max(0, \phi_n+x\lambda_s-c_n)$ at the solution of the program. With the introduction of this slack variable ν_1 that replaces $\max(0, \phi_n + x\lambda_s-c_n)$, the objective function becomes

$$f(x,\nu_1) \triangleq \left(t+x-d_{n-1,s}-H_s\right)^2+\left(\widetilde{d}_{n+1,s}-t-x-H_s\right)^2+M_1\nu_1 \\ +M_2\max\left[0,\widetilde{l}_{n+1}-\widetilde{\beta}_{n+1}+\left(\widetilde{\beta}_{n+1}t_a\lambda_s+\nu_1+\left(\widetilde{a}_{n+1,s}-t-x\right)\lambda_s\right)\left(1+t_b\lambda_s\right)-c_{n+1}\right] \tag{9}$$

and the expected departure time of trip $n+1$ from stop s:

$$\widetilde{d}_{n+1,s} \triangleq \widetilde{a}_{n+1,s}+\widetilde{\beta}_{n+1}t_a+\min\left[\left(\widetilde{\beta}_{n+1}t_a\lambda_s+\nu_1+\left(\widetilde{a}_{n+1,s}-t-x\right)\lambda_s\right)\left(1+t_b\lambda_s\right)t_b, \\ \left(c_{n+1}+\widetilde{\beta}_{n+1}-\widetilde{l}_{n+1}\right)t_b\right] \tag{10}$$

Hence, we reformulate program (Q) to

$$\left(\overline{Q}\right) \quad \min_{x,\nu_1} \quad f(x,\nu_1) \\ \text{s.t.} \quad \left(f, \widetilde{d}_{n+1,s}\right) | \left(f, \widetilde{d}_{n+1,s}\right) \text{ satisfy Eq. } \left(9\right), \left(10\right) \\ \nu_1 \geqslant 0 \\ \nu_1 \geqslant \phi_n + x\lambda_s - c_n \\ 0 \leqslant x \leqslant \zeta \tag{11}$$

Note that the term $M_1\nu_1$ in the reformulated objective function $f(x,\nu_1)$ forces ν_1 to receive its lowest possible value which is always greater than or equal to zero and has the equivalent effect of term $M_1\max(0,\phi_n + x\lambda_s - c_n)$.

The objective function of program $(\overline{Q})$ has another non-smooth term: $M_2\max\left[0, \widetilde{l}_{n+1} - \widetilde{\beta}_{n+1} + \left(\widetilde{\beta}_{n+1}t_a\lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b\lambda_s) - c_{n+1}\right]$. With the introduction of another slack variable ν_2 that takes the value of the above term at the solution of the program, the objective function becomes

$$f\left(x,\nu_1,\nu_2\right) \triangleq \left(t + x - d_{n-1,s} - H_s\right)^2 + \left(\widetilde{d}_{n+1,s} - t - x - H_s\right)^2 + M_1\nu_1 + M_2\nu_2 \tag{12}$$

and program $(\overline{Q})$ is reformulated to

$$\begin{aligned}
\left(\widehat{Q}\right) \quad & \min_{x,\nu_1,\nu_2} \quad f(x,\nu_1,\nu_2) \\
& \text{s.t.} \quad \left(f,\widetilde{d}_{n+1,s}\right)|\left(f,\widetilde{d}_{n+1,s}\right) \text{ satisfy Eq. } \left(10\right),\ \left(12\right) \\
& \quad\ \nu_1 \geqslant 0 \\
& \quad\ \nu_1 \geqslant \phi_n + x\lambda_s - c_n \\
& \quad\ \nu_2 \geqslant 0 \\
& \quad\ \nu_2 \geqslant \widetilde{l}_{n+1} - \widetilde{\beta}_{n+1} - c_{n+1} + \\
& \quad\ \left(\widetilde{\beta}_{n+1}t_a\lambda_s + \nu_1 + \left(\widetilde{a}_{n+1,s} - t - x\right)\lambda_s\right)\left(1 + t_b\lambda_s\right) \\
& \quad\ 0 \leqslant x \leqslant \zeta
\end{aligned} \tag{13}$$

The equality constraint of Eq. (10) that defines the value of variable $\widetilde{d}_{n+1,s}$ is the last non-smooth term in our reformulated program, $\widehat{Q}$, due to the nonlinear term $\min\left[\left(\widetilde{\beta}_{n+1}t_a\lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b\lambda_s)t_b, (c_{n+1} + \widetilde{\beta}_{n+1} - \widetilde{l}_{n+1})t_b\right]$. To avoid this nonlinearity, we re-write $\widetilde{d}_{n+1,s}$ as

$$\widetilde{d}_{n+1,s} = \widetilde{a}_{n+1,s} + \widetilde{\beta}_{n+1}t_a + \left(\widetilde{\beta}_{n+1}t_a\lambda_s + \nu_1 + \left(\widetilde{a}_{n+1,s} - t - x\right)\lambda_s\right)\left(1 + t_b\lambda_s\right)t_b - \nu_2 t_b \tag{14}$$

In our theorem presented in Theorem Appendix 1 we prove that the values of $\widetilde{d}_{n+1,s}$ derived by Eq. (10) and the reformulated Eq. (14) are equivalent at the solution of mathematical program $(\widehat{Q})$.

To simplify the notation, let $k \triangleq 1 + t_b\lambda_s$, where $k \in \mathbb{R}_{\geqslant 0}$ because $t_b, \lambda_s \geqslant 0$. Then, the objective function can be re-written as

$$\begin{aligned}
f(x,\nu_1,\nu_2) \triangleq\ & \left(t + x - d_{n-1,s} - H_s\right)^2 + \left[\widetilde{a}_{n+1,s} + \widetilde{\beta}_{n+1}t_a + \left(\widetilde{\beta}_{n+1}t_a\lambda_s + \nu_1 + \left(\widetilde{a}_{n+1,s} - t - x\right)\lambda_s\right)kt_b\right. \\
& \left. - \nu_2 t_b - t - x - H_s\right]^2 + M_1\nu_1 + M_2\nu_2
\end{aligned} \tag{15}$$

and this leads to the reformulation of program $(\widetilde{Q})$ to

$$\begin{aligned}
\left(\widetilde{Q}\right) \quad & \min_{x,\nu_1,\nu_2} \quad f(x,\nu_1,\nu_2) \\
& \text{s.t.} \quad (f)|(f) \text{ satisfies Eq. (15)} \\
& \quad\ \nu_1 \geqslant 0 \\
& \quad\ \nu_1 \geqslant \phi_n + x\lambda_s - c_n \\
& \quad\ \nu_2 \geqslant 0 \\
& \quad\ \nu_2 \geqslant \widetilde{l}_{n+1} - \widetilde{\beta}_{n+1} - c_{n+1} + \left(\widetilde{\beta}_{n+1}t_a\lambda_s + \nu_1 + \left(\widetilde{a}_{n+1,s} - t - x\right)\lambda_s\right)k \\
& \quad\ 0 \leqslant x \leqslant \zeta
\end{aligned} \tag{16}$$

This reformulation has introduced two slack variables (ν_1,ν_2) to transform the non-smooth, nonlinear program (Q) to a program $(\widetilde{Q})$ with a quadratic objective function and linear inequality constraints that attains an equivalent solution to (Q). As it is shown in our theorem presented in Theorem Appendix 2, a locally optimal solution of program $(\widetilde{Q})$ is also a globally optimal one because $(\widetilde{Q})$ is convex.

5. Analytic solution of bus holding considering capacity limits

In Theorem Appendix 2 we proved that our reformulated program $(\widetilde{Q})$ is convex and any local minimizer is also a globally optimal solution. In this section, we present an analytic solution for the bus holding problem under capacity limits. This analytic solution is provided in Theorem 5.1. The analytic solution allows the service operator to determine immediately the holding time of any trip, n, by using a closed-form expression instead of solving a mathematical program. This is a major advantage of our approach because we can compute the holding time of a bus in real-time without requiring any computational costs.

To simplify the notation, we set $\eta \triangleq kt_b$ and $\theta \triangleq \widetilde{a}_{n+1,s} + \widetilde{\beta}_{n+1}t_a + \left(\widetilde{\beta}_{n+1}t_a\lambda_s + (\widetilde{a}_{n+1,s} - t)\lambda_s\right)kt_b - t - H_s$. Note that η and θ are parameters with pre-computed values. Then, we proceed to Theorem 5.1.

Theorem 5.1. *The analytic solution of program $(\widetilde{Q})$, which is the optimal holding time of bus trip n at stop s, is:*

Table 2
Potential cases for $\nu_1, \nu_2 = 0$.

Case	ρ_1	ρ_2	ρ_3	ρ_4	x
1	>0	$=0$	>0	$=0$	$\frac{(\lambda_s\eta+1)\theta-(t-d_{n-1,s}-H_s)}{1+(\lambda_s\eta+1)^2}$
2	>0	>0	>0	$=0$	$\frac{c_n-\phi_n}{\lambda_s}$
3	>0	$=0$	>0	>0	$\frac{(\lambda_s\eta+1)\theta-(t-d_{n-1,s}-H_s)}{1+(\lambda_s\eta+1)^2}$
4	>0	>0	>0	>0	$\frac{c_n-\phi_n}{\lambda_s}$

$$x^* = \left\{\max\left(0, \min\left[\zeta, \frac{c_n-\phi_n}{\lambda_s}, \frac{(\lambda_s\eta+1)\theta-(t-d_{n-1,s}-H_s}{1+(\lambda_s\eta+1)^2}\right]\right)\right\} \tag{17}$$

Proof. Using the simplified notation (η,θ), program $(\widetilde{Q})$ is re-written as:

$$\begin{aligned}
\min_{x} \quad & \left(t+x-d_{n-1,s}-H_s\right)^2 \\
& +(\theta+\nu_1\eta-x\lambda_s\eta-\nu_2 t_b-x)^2+M_1\nu_1+M_2\nu_2 \\
\text{s.t.}: \quad & -\nu_1 \leqslant 0 \\
& \phi_n+x\lambda_s-c_n-\nu_1 \leqslant 0 \\
& -\nu_2 \leqslant 0 \\
& \widetilde{l}_{n+1}-\widetilde{\beta}_{n+1}-c_{n+1}+\left(\widetilde{\beta}_{n+1}t_a\lambda_s+\nu_1+\left(\widetilde{a}_{n+1,s}-t-x\right)\lambda_s\right)k-\nu_2 \leqslant 0 \\
& 0 \leqslant x \leqslant \zeta
\end{aligned} \tag{18}$$

Let us introduce constraint functions $g_1(\nu_1) \triangleq -\nu_1$, $g_2(x,\ \nu_1) \triangleq \phi_n + x\lambda_s - c_n - \nu_1$, $g_3(\nu_2) \triangleq -\nu_2$, and $g_4(x,\ \nu_1,\ \nu_2) \triangleq \widetilde{l}_{n+1} - \widetilde{\beta}_{n+1} - c_{n+1} + \left(\widetilde{\beta}_{n+1}t_a\lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)k - \nu_2$. Then, our program is equivalent to

$$\begin{aligned}
\min_{x} \quad & \left(t+x-d_{n-1,s}-H_s\right)^2 \\
& +(\theta+\nu_1\eta-x\lambda_s\eta-\nu_2 t_b-x)^2+M_1\nu_1+M_2\nu_2 \\
\text{s.t.}: \quad & g_1(\nu_1), g_2(x,\nu_1), g_3(\nu_2), g_4(x,\nu_1,\nu_2) \leqslant 0 \\
& x \in [0,\zeta]
\end{aligned} \tag{19}$$

From the Karush–Kuhn–Tucker (KKT) conditions, x^* minimizes the above program, if, and only if, there exist dual variables (KKT multipliers $\rho_1,\rho_2,\rho_3,\rho_4$) such that

(1) $\nabla\mathscr{L}(x^*,\nu_1^*,\nu_2^*,\rho_1,\rho_2,\rho_3,\rho_4) = 0$
(2) $\rho_j g_j = 0, \forall j \in \{1,2,3,4\}$ (complementary slackness)
(3) $\rho_j \geqslant 0, \forall j \in \{1,2,3,4\}$
(4) (x^*,ν_1^*,ν_2^*) satisfy the inequality constraints of the program in Eq. (19).where

$$\begin{aligned}
\mathscr{L}(x,\nu_1,\nu_2,\rho_1,\rho_2,\rho_3,\rho_4) \triangleq\ & \left(t+x-d_{n-1,s}-H_s\right)^2+(\theta+\nu_1\eta-x\lambda_s\eta-\nu_2 t_b-x)^2 \\
& +M_1\nu_1+M_2\nu_2+\rho_1 g_1(\nu_1)+\rho_2 g_2(x,\nu_1)+\rho_3 g_3(\nu_2)+\rho_4 g_4(x,\nu_1,\nu_2)
\end{aligned} \tag{20}$$

Each constraint g_j is active (binding) if $\rho_j > 0$, because in that case $\rho_j g_j = 0 \Rightarrow g_j = 0$. From the KKT conditions, we get the following system of equations

(1) $\partial\mathscr{L}/\partial x = 0 \Rightarrow 2x + 2(t - d_{i-1,s} - H_s) + 2x(\lambda_s\eta+1)^2 - 2(\lambda_s\eta+1)(\theta+\nu_1\eta-\nu_2 t_b) - \rho_2\lambda_s - \rho_4\lambda_s k = 0.$
(2) $\partial\mathscr{L}/\partial\nu_1 = 0 \Rightarrow 2\eta^2\nu_1 + 2\eta(\theta - x\lambda_s\eta - x - \nu_2 t_b) + M_1 - \rho_1 - \rho_2 + \rho_4 k = 0.$
(3) $\partial\mathscr{L}/\partial\nu_2 = 0 \Rightarrow 2t_b^2\nu_2 - 2t_b(\theta+\nu_1\eta - x\lambda_s\eta - x) + M_2 - \rho_3 - \rho_4 = 0.$
(4) $-\rho_1\nu_1 = 0.$
(5) $\rho_2(\phi_n + x\lambda_s - c_n - \nu_1) = 0.$
(6) $-\rho_3\nu_2 = 0.$
(7) $\rho_4\left(\widetilde{l}_{n+1} - \widetilde{\beta}_{n+1} - c_{n+1} + \left(\widetilde{\beta}_{n+1}t_a\lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)k - \nu_2\right) = 0.$
(8) $\rho_1 \geqslant 0, \rho_2 \geqslant 0, \rho_3 \geqslant 0, \rho_4 \geqslant 0.$
(9) (x^*,ν_1^*,ν_2^*) satisfy the inequality constraints of the program in Eq. (19).

The above system of equations can be solved for $2^4 = 16$ different cases, given the potential active/inactive combinations of KKT multipliers $\rho_1,\ldots,\rho_4$.

For the general case where $\nu_1,\nu_2 = 0$, we have the four sub-cases with their respective solutions, x, expressed in Table 2.

Table 3
Potential cases for $\nu_1 = 0, \nu_2 > 0$.

Case	ρ_1	ρ_2	ρ_3	ρ_4	x
5	>0	$=0$	$=0$	$=0$	$\left(t-d_{n-1,s}-H_s\right)+M_2\frac{(1+\eta\lambda_s)}{2\eta\lambda_s t_b}$
6	>0	>0	$=0$	$=0$	$\frac{c_n-\phi_n}{\lambda_s}$
7	>0	$=0$	$=0$	>0	$\left(t-d_{n-1,s}-H_s\right)+M_2\frac{(1+\eta\lambda_s)}{2\eta\lambda_s t_b}$
8	>0	>0	$=0$	>0	$\frac{c_n-\phi_n}{\lambda_s}$

Table 4
Potential cases for $\nu_1 > 0, \nu_2 = 0$.

Case	ρ_1	ρ_2	ρ_3	ρ_4	x
9	$=0$	$=0$	>0	$=0$	$\left(H_s+d_{n-1,s}-t\right)-M_1\frac{\eta\lambda_s+1}{2\eta}$
10	$=0$	>0	>0	$=0$	$\left(H_s+d_{n-1,s}-t\right)-M_1\frac{\eta\lambda_s+1}{2\eta}$
11	$=0$	$=0$	>0	>0	$\left(H_s+d_{n-1,s}-t\right)-M_1\frac{\eta\lambda_s+1}{2\eta}$
12	$=0$	>0	>0	>0	$\left(H_s+d_{n-1,s}-t\right)-M_1\frac{\eta\lambda_s+1}{2\eta}$

For cases 1 and 3 the KKT system of equations yields solution $x = \frac{(\lambda_s\eta+1)\theta-(t-d_{n-1,s}-H_s)}{1+(\lambda_s\eta+1)^2}$. For cases 2, 4 the solution is $x = \frac{c_n-\phi_n}{\lambda_s}$. Hence, when $\nu_1, \nu_2 = 0$, which means that the capacity limits of both trips n and $n+1$ are not exceeded, the holding time is:

- $\frac{(\lambda_s\eta+1)\theta-(t-d_{n-1,s}-H_s)}{1+(\lambda_s\eta+1)^2}$ when the capacity limit of trip n is not reached,
- $\frac{c_n-\phi_n}{\lambda_s}$ when the capacity limit of trip n is just reached.

Note also that x has an upper and lower bound, $x \in [0, \zeta]$. Consequently, the holding time solution in case of $\nu_1, \nu_2 = 0$ can be succinctly written as:

$$\max\left(0, \min\left(\frac{(\lambda_s\eta+1)\theta-\left(t-d_{n-1,s}-H_s\right)}{1+(\lambda_s\eta+1)^2}, \frac{c_n-\phi_n}{\lambda_s}, \zeta\right)\right)$$

For the general case where $\nu_1 = 0$ and $\nu_2 > 0$ the capacity limit of trip n is not reached, but the capacity limit of its following trip $n+1$ is exceeded. Intuitively, in that case trip n will be held at stop *s as much as possible* to reduce its headway with trip $n+1$ and the number of passengers willing to board that trip. Indeed, for $\nu_1 = 0$ and $\nu_2 > 0$ we have the four sub-cases with their respective solutions expressed in Table 3.

Note that in cases 6 and 8 the holding time is $\frac{c_n-\phi_n}{\lambda_s}$ because in both cases the capacity limit of bus trip n is just reached. This triggers the immediate release of trip n despite the potentially high value of ν_2. If the capacity of trip n is not reached, bus trip n will be held as much as possible to reduce the value of ν_2 (see cases 5, 7). Note that in cases 5 and 7 the proposed holding is a very large number way beyond the maximum allowed holding, ζ, because it includes the term $M_2\frac{(1+\eta\lambda_s)}{2\eta\lambda_s t_b}$ where $M_2 \gg 0$. Thus, even if holding bus trip n for $\left(t-d_{n-1,s}-H_s\right)+M_2\frac{(1+\eta\lambda_s)}{2\eta\lambda_s t_b}$ has the optimal effect to the overcrowding of trip $n+1$, trip n cannot be held for so long and will be released after time ζ. Consequently, the solution for $\nu_1 = 0, \nu_2 > 0$ is succinctly written as $\max\left(0, \min\left(\frac{c_n-\phi_n}{\lambda_s}, \zeta\right)\right)$.

For the general case where $\nu_1 > 0$ and $\nu_2 = 0$ the capacity limit of trip n is exceeded and passengers are refused to board. Intuitively, in that case trip n will be released from stop *s as soon as possible*. For $\nu_1 > 0$ and $\nu_2 = 0$, we have the four sub-cases with their respective solutions expressed in Table 4.

The solution $\left(H_s+d_{n-1,s}-t\right)-M_1\frac{\eta\lambda_s+1}{2\eta}$ is a negative number given that M_1 is a very large number. That is to say, ideally trip n should have a "negative" holding time if at time t its capacity limit is already reached and there are stranded passengers. Since a negative holding is not possible, trip n will depart immediately. Thus, for $\nu_1 > 0, \nu_2 = 0$ the holding solution is equal to zero. For $\nu_1 > 0 \Leftrightarrow \phi_n + x\lambda_s - c_n > 0$, and since $x = 0, \phi_n > c_n$. Consequently, the holding time solution in case of $\nu_1 > 0$ and $\nu_2 = 0$ can be succinctly written as $\max\left(0, \min\left(\frac{(\lambda_s\eta+1)\theta-(t-d_{n-1,s}-H_s)}{1+(\lambda_s\eta+1)^2}, \frac{c_n-\phi_n}{\lambda_s}, \zeta\right)\right)$ which is always equal to zero given that $\frac{c_n-\phi_n}{\lambda_s} < 0$ for $\nu_1 > 0$, which means that $\frac{c_n-\phi_n}{\lambda_s} < 0 \Rightarrow \min\left(\frac{(\lambda_s\eta+1)\theta-(t-d_{n-1,s}-H_s)}{1+(\lambda_s\eta+1)^2}, \frac{c_n-\phi_n}{\lambda_s}, \zeta\right) < 0$.

Table 5
Parameter values of the idealized scenario.

Parameter	Value	Unit	Parameter	Value	Unit
$d_{i-1,s}$	1000	s	t_a	1.5	s
t	1500	s	t_b	4	s
H_s	600	s	$a_{n+1,s}$	2500	s
ϕ_n	40	passengers	ζ	300	s
c_n, c_{n+1}	60	passengers	M_1	10E+14	-
$\widetilde{\beta}_{n+1}$	10	passengers	M_2	10E+12	-
$\widetilde{l}_{n+1}$	50	passengers	λ_s	0.02	passengers/ s

Table 6
Optimal Holding decisions for different values of (λ_s, ϕ_n).

scenarios			Analytic Solution with capacity			Fu and Yang (2002)
	λ_s	ϕ_n	$\frac{c_n-\phi_n}{\lambda_s}$	$\mathscr{Z}$	x^*	x^*
I	0.02	58	100 s	296 s	**100 s**	199 s
II	0.05	58	40 s	361 s	**40 s**	229 s
III	0.02	59	50 s	296 s	**50 s**	199 s
IV	0.02	62	−100 s	296 s	**0 s**	199 s

Summarizing the solutions from all potential cases of slack variable values, (ν_1, ν_2), we get:

$$x^* = \left\{ \max\left(0, \min\left[\zeta, \frac{c_n-\phi_n}{\lambda_s}, \frac{(\lambda_s\eta+1)\theta-(t-d_{n-1,s}-H_s)}{1+(\lambda_s\eta+1)^2}\right]\right)\right\}$$

that returns the optimal solution despite the values of ν_1, ν_2 and completes our proof. □ In Appendix A.1, we illustrate the equivalency between the solution of program $(\widetilde{Q})$ and our analytic solution, x^*, which is clearly non-negative.

6. Case study and numerical experiments

6.1. Demonstration

In this sub-section, we perform numerical experiments which manifest the holding decisions of our analytic solution in different scenarios compared to the holding decisions of the two-headway-based control logic of Fu and Yang (2002) that does not consider vehicle capacity limitations. The parameter values of the idealized scenarios are presented in Table 5.

To cover multiple cases, we modify the parameter values of λ_s (passenger arrival rate) and ϕ_n and compute the respective solutions of the following idealized scenarios. The data and source code of each scenario is publicly released in Gkiotsalitis (2019a), and the respective solutions are presented in Table 6. To simplify the notation in Table 6, we set $\mathscr{Z} \triangleq \frac{(\lambda_s\eta+1)\theta-(t-d_{n-1,s}-H_s)}{1+(\lambda_s\eta+1)^2}$. This allows us to re-write our analytic solution expressed in Theorem 5.1 as

$$x^* = \left\{ \max\left(0, \min\left[\zeta, \frac{c_n-\phi_n}{\lambda_s}, \mathscr{Z}\right]\right)\right\}$$

The bus holding solutions when applying the analytic solution of Fu and Yang (2002) are presented in the last column of Table 6. Note that in all four scenarios bus trip n is ahead of schedule and needs to be held at the control point stop to normalize the headways. Our scenarios are selected in such a way that bus trip n is running close to capacity in order to demonstrate the difference between our analytic solution and analytic solutions that do not consider the vehicle capacity.

As demonstrated in Table 6, the solution of Fu and Yang (2002) is not sensitive to the value changes of parameter ϕ_n since it does not cater for overcrowding but merely balances the headways between the preceding and following trip(s) using an estimate of $\widetilde{d}_{n+1,s} \approx \widetilde{a}_{n+1,s} + \widetilde{\beta}_{n+1}t_a + (\widetilde{a}_{n+1,s} - t)\lambda_s t_b$.

The results of the comparative analysis between our approach and the classic two-headway-based approach of Fu and Yang (2002) are presented in Fig. 3. Fig. 3 demonstrates the potential benefit of our control method in comparison to similar approaches that ignore the overcrowding of buses in the optimization process. In Fig. 3 we plot the sum of the bus load and the number of (potentially) stranded passengers that are refused to board trip n until it departs from stop s. Note that if this value is higher than the vehicle capacity, $c_n = 60$ passengers, this results in refused boardings. The implementation of our analytic solution leads to stranded passengers only in scenario VIII, in which 2 passengers were already waiting for trip n when it arrived at stop s. In all other cases, our analytic solution held the bus until it reached its capacity without leading to refused passenger boardings. In contrast, the control logic of Fu

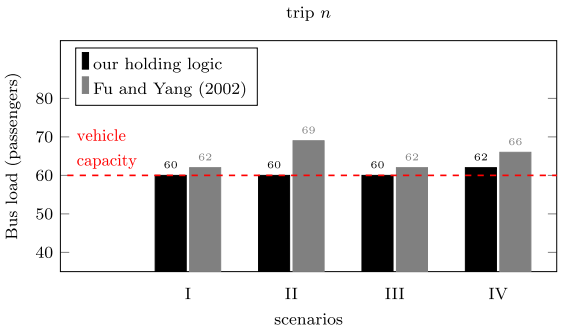

Fig. 3. Bus load plus stranded passengers of trip n when it departs from stop s in every scenario with the implementation of our analytic solution and the one of Fu and Yang (2002).

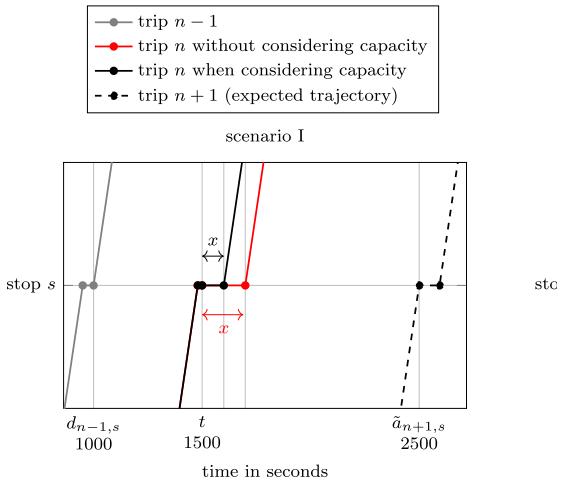

Fig. 4. Illustration of the trajectories of trips $n-1, n$ and $n+1$ in scenario I when applying our holding method and the method of Fu and Yang (2002) that does not consider the vehicle capacity constraint.

Table 7
In-vehicle and out-of-vehicle passenger waiting times in seconds in the four scenarios.

	In-vehicle waiting times of all passengers due to the holding of trip n		Out-of-vehicle waiting times of all passengers	
Scenario	Our control logic	Fu and Yang (2002)	Our control logic	Fu and Yang (2002)
I	6000	12334	13481	14669
II	2400	15904	35115	40218
III	3000	12533	13945	15584
IV	0	13130	16680	18272

and Yang (2002) results in refused boardings in all 4 cases.

In addition to that, in Fig. 4 we report the trajectories of bus trips $n-1, n$ and $n+1$ when applying our approach and the approach of (Fu and Yang, 2002) in scenario I. In this illustration, the trajectory of bus trip $n-1$ remains unchanged because it does not depend on our holding decision. The expected trajectory of trip $n+1$ is slightly modified based on our holding decision. Finally, the trajectory of trip n differs significantly when considering the vehicle capacity in the holding control.

From Fig. 4 one can note that the inter-departure headways of bus trips when departing from control point stop s are more evenly distributed when applying the holding logic of Fu and Yang (2002). However, many passengers are refused to board trip n when the vehicle capacity is not considered. As we will later see, the inability to board additional passengers after reaching the vehicle capacity means that the out-of-vehicle passenger waiting times are not improved despite achieving a more even distribution of headways. In addition, continuing to hold the bus after reaching its capacity will result in increased travel times for the in-vehicle passengers without providing any tangible benefits.

To investigate the effect of the holding decisions on the passenger waiting times, in Table 7 we report the values of the following

Table 8
Scheduled headways of bus line 302 at different times of the day.

Period	Target Headway
05:30–06:30	–
06:30–08:30	≈ 4 min
08:30–19:00	≈ 5 min
After 19:00	≈ 8 min

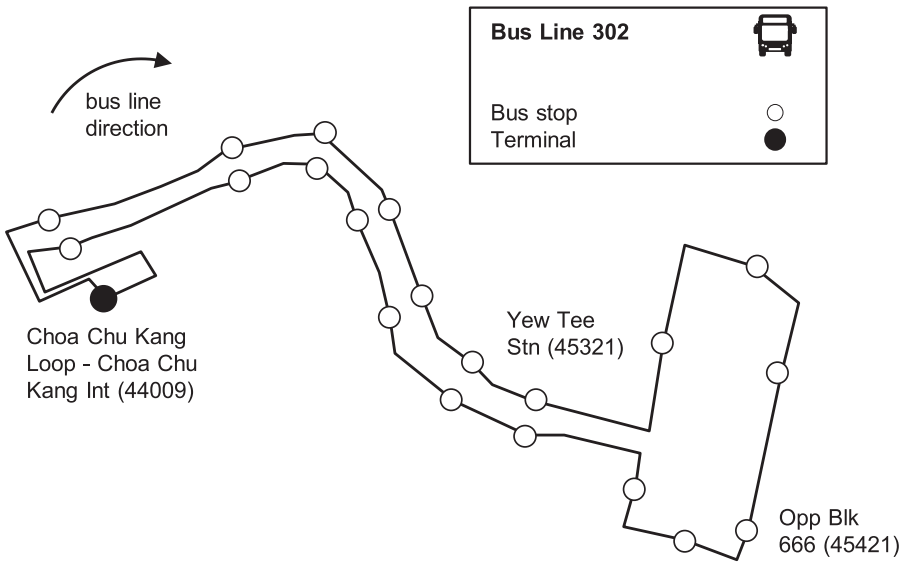

Fig. 5. Topology of bus line 302 in Singapore.

two key performance indicators:

- the additional in-vehicle waiting times of all passengers onboard bus n incurred due to its holding at stop s;
- the out-of-vehicle waiting times of all passengers that board trips n and $n + 1$. Note that if a passenger cannot board trip n, he/she will have to wait for the following trip $n + 1$.

There are two interesting observations from the results in Table 7. The first observation is that the out-of-vehicle waiting times of all passengers do not differ significantly when using different control logics. This is an interesting observation because one might have expected that if we distribute the vehicle headways more evenly in the expense of increasing the number of stranded passengers (as in Fig. 4), we can improve significantly the out-of-vehicle passenger waiting times. This is not the case though because even if the headways are more evenly distributed, passengers will not be able to board a bus if its capacity limit is reached and will have to wait for the next one. That is, holding a bus at a stop after reaching its capacity does not have an effect on reducing the out-of-vehicle passenger waiting times (even if one might have expected a positive influence due to the more even distribution of headways). The second observation is that we have significant in-vehicle waiting time gains when we consider the vehicle capacity in our holding control logic. The reason is that we allow the bus to depart from the control stop when its capacity is reached instead of holding it further without being able to reduce the waiting times of out-of-vehicle passengers because they have to wait for the next bus.

6.2. *Case study*

Our numerical experiments in our case study aim to investigate (i) the potential effect of our bus holding method compared to other analytic solutions that do not consider the capacity of vehicles, and (ii) the sensitivity of our bus holding decisions to the passenger demand and travel time (i.e., arrival time) variations. The sensitivity analysis is performed with Monte Carlo simulations and underlines the importance of estimating the travel times of upstream trips with high accuracy.

6.2.1. *Operational performance of our analytic solution against analytic solutions that do not consider vehicle capacity limits*

In this sub-section we investigate the operational performance of our analytic bus holding solution compared to state-of-the-art analytic solutions that do not consider vehicle capacity limits. Our case study is the high-frequency, circular bus line 302 in

Table 9
Parameter values when determining the holding times of trips.

Parameter	Value	Parameter	Value
H_s	4 min → 240 s	t_a	1 s
c_n	75 passengers	t_b	2 s
H_s	4 min → 240 s	ζ	90 s
M_1	10E+14	M_2	10E+12

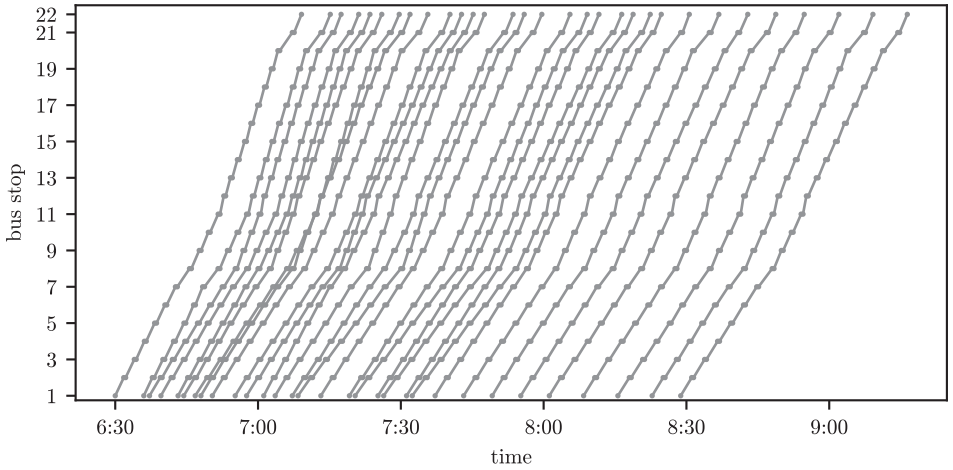

Fig. 6. Trajectories of buses operating from 6:30 until 8:30. Note that the 7th and the 8th bus overtake each other twice: when traveling from stop 10 to 11 and from stop 12 to 13.

Singapore. Bus line 302 has 22 stops departing from Choa Chu Kang Loop - Choa Chu Kang Int (44009) and ending at the same stop. It is operated by SMRT and its regularity is monitored by the Land Transport Authority (LTA). Normally starts operating at 05:30 and ends at 00:55. Its route length is 8.1 km and its total travel time typically ranges from 35 to 40 min. Bus line 302 is selected because it is one of the seven high-frequency bus lines in Singapore that are monitored in terms of service regularity and are placed under the Bus Service Reliability Framework (BSRF) from the LTA (Leong et al., 2016). Under the BSRF framework, bus lines that do not maintain their scheduled headways are penalized, whereas well-performing lines receive monetary incentives (up to 3000$ for every 0.1 min improvement in regularity at the end of each month, as of May 2014).

Bus line 302 is a feeder service that serves residential blocks, schools, and public amenities, connecting them to Choa Chu Kang Town Centre and Yew Tee Mass Rapid Transit (MRT) station. Its primary area of service is Choa Chu Kang neighborhoods 5 and 6. Typically, in this bus line operate 12-meter single-decker buses with a seated capacity of 42 passengers and a standing capacity of 33 passengers (75 passengers in total). High capacity, articulated buses have also been deployed due to high demand from residents. The total number of operating trips per day is 245, and the scheduled (target) headways differ among peak/off-peak hours, as it is presented in Table 8. Note that from 05:30 until 6:30, the service is not headway-based due to the low passenger demand; thus, the scheduled headways in Table 8 start after 6:30.

Our experiments focus on the time period 06:30–08:30, which exhibits the highest frequency with 31 trips and a scheduled headway of 4 min. The topology of bus line 302 is presented in Fig. 5.

All trips are operated by single decker buses with a total capacity of 75 passengers (including standees). We assume uniformly distributed passenger arrivals at any stop s because passengers are not able to coordinate their arrival times at stops with the arrival times of buses in high-frequency services (Ibarra-Rojas et al., 2015).

Based on historical data, the observed (average) time for an extra passenger boarding and alighting is 2 and 1 s, respectively. Our historical data observations are in line with the findings of Meng and Qu (2013) that observed an extra time of 1.36 s for each boarding/alighting in bus lines in Singapore. To summarize, the parameter values of our case study are presented in Table 9. Note that, as in Cortés et al. (2010), we do not allow a holding time of more than 90 s due to the inconvenience caused to on-board passengers.

In this experimentation, we demonstrate the application of our control logic in the 31 trips dispatched from 06:30 until 08:30 compared to the applications of the two-headway-based control logic of Fu and Yang (2002), and the self-equalizing headway method of Bartholdi and Eisenstein (2012). Those two methods are selected because they have analytic solutions. Therefore, similarly to our approach, they can be applied in real-time without requiring any computational costs. We note that we do not compare our control logic against approaches that solve mathematical programs, because such approaches do not have an analytic solution and their computational burden increases significantly with the number of decision variables given their usually non-polynomial time complexity. The two-headway-based control logic in Fu and Yang (2002) decides about the holding time of a trip based on its headways

Table 10
Results of the simulation-based evaluation using our control logic and the control logic(s) of Fu and Yang (2002) and Bartholdi and Eisenstein (2012) in the time period 6:30–8:30.

	Fu and Yang (2002)	Bartholdi III and Eisenstein (2012)	Our control logic
Mean squared headway deviation (min^2)	27.2	26.7	27.3
Average out-of-vehicle waiting times (min)	2.17	2.12	2.09
Refused passenger boardings	69	74	17
Total vehicle capacity violations	19	21	6

with its preceding/following trips, whereas the self-equalizing headway method determines the holding time by considering the headways among multiple trips. That is, the self-equalizing headway method strives to equalize the headways among multiple buses instead of adhering to the target headway(s).

In our simulation-based evaluation, we simulate the operations of 31 trips using realistic inter-station travel time and passenger arrival data from one day of operations (see the vehicle trajectories of Fig. 6).

In the simulation, the inter-station travel times of trips are equal to the observed inter-station travel times from the actual data. The same applies for the passenger arrival rates at stops. First, we simulate the operations of the 31 trips using our control logic yielding a holding time of $x = \left\{ \max\left(0, \min\left[\zeta, \frac{c_n - \phi_n}{\lambda_s}, \frac{(\lambda_s\eta+1)\theta-(t-d_{n-1,s}-H_s)}{1+(\lambda_s\eta+1)^2}\right]\right)\right\}$ every time a bus n arrives at stop s. Then, we simulate the operations using the two-headway-based control logic in Fu and Yang (2002). Every time we compute a bus holding, we use the Kalman filter predictor of Cathey and Dailey (2003) to predict the arrival time of the following trip $n+1$ at stop s. Last, we simulate the operations using the control logic of Bartholdi and Eisenstein (2012) that updates the holding times of multiple trips to equalize their headways. All control logics are event-based and compute a holding time every time a bus is ready to depart from a stop.

The results of the simulation-based evaluation are presented in Table 10. The passenger arrival rates and the inter-station travel times are based on operational data from bus line 302. Additionally, overtaking among buses is permitted in our simulation, even when buses are dwelling at the same stop. This adds realism to our simulation-based evaluation that investigates the performance potential of our approach. In Table 10 we report the results of the following key performance indicators: (i) the average squared headway deviation from the target headway (indicating the service regularity); (ii) the average out-of-vehicle waiting time per passenger; (iii) the total number of passengers who were refused to board onto the first arriving trip; and (iv) the total vehicle capacity violations. Note that we do not report the computation costs because the analytic solutions do not have a computational cost.

The performance of the three control logics show that the self-equalizing headway approach of Bartholdi and Eisenstein (2012) results in the most improved service regularity with a mean squared headway deviation from the target headway of 26.7 min^2. This is achieved because the self-equalizing headway approach considers the equalizing of headways among multiple buses, and not only between the following and the preceding bus. Because of the improved regularity, the control logic of Bartholdi and Eisenstein (2012) resulted also in lower passenger waiting times compared to the two-headway-based control logic in Fu and Yang (2002). The implementation of our control logic reduced further the average passenger waiting times because we had less refused passenger boardings, and thus passengers did not have to wait for prolonged time periods. Overall, our control logic reduced considerably the number of stranded passengers, while resulting in a slight improvement in the average waiting time per passenger. In terms of service regularity, our approach performed comparatively to the control logic of Fu and Yang (2002) which also considers the headway between the preceding and following trip when making a holding decision. Summarizing, our control logic reduces the number of stranded passengers and the average passenger waiting times in the expense of a slight service regularity deterioration.

6.2.2. Sensitivity of our control logic to demand variations

Herein, we investigate the sensitivity of the potential gain when applying our control logic in the case of demand (boarding and alighting) variations. This is important because every time we make a holding decision about a bus trip n we assume that the realized bus load and the alightings of its future trip will be in line with our expectations (an assumption that is rarely true in practice). In this investigation, we perform a simulation-based evaluation of the performance of the service regularity under different demand variation levels compared to our demand expectations. To reduce the sampling bias in our simulation, we perform a large number of 1,000 Monte Carlo simulations for each demand variation level.

In the first demand variation scenario, we assume that the passenger alightings and the passenger arrival rates at stops are random variables with coefficient of variation ($CV=\frac{\text{standard deviation}}{\text{mean}}$) equal to 20%. For validation purposes, their realized values at each one of the $i = \{1, 2, \ldots, 1000\}$ Monte Carlo simulations are sampled from restricted normal distributions:

- $\widetilde{\beta}^i_{n+1,s} = \max\{0, \widetilde{\sim}\mathcal{N}(\widetilde{\beta}_{n+1,s}, CV \cdot \widetilde{\beta}_{n+1,s})\}$
- $\lambda^i_{n,s} = \max\{0, \widetilde{\sim}\mathcal{N}(\lambda_{n,s}, CV \cdot \lambda_{n,s})\}$

where $\widetilde{\beta}_{n+1,s}$ are the expected passenger alightings per trip $n+1$ at stop s, and $\lambda_{n,s}$ the expected passenger arrival rate per trip n at the time of its arrival at stop s. Note that the sampled (realized) values $\widetilde{\beta}^i_{n+1,s}$ and $\lambda^i_{n,s}$ are restricted to receive only positive values because

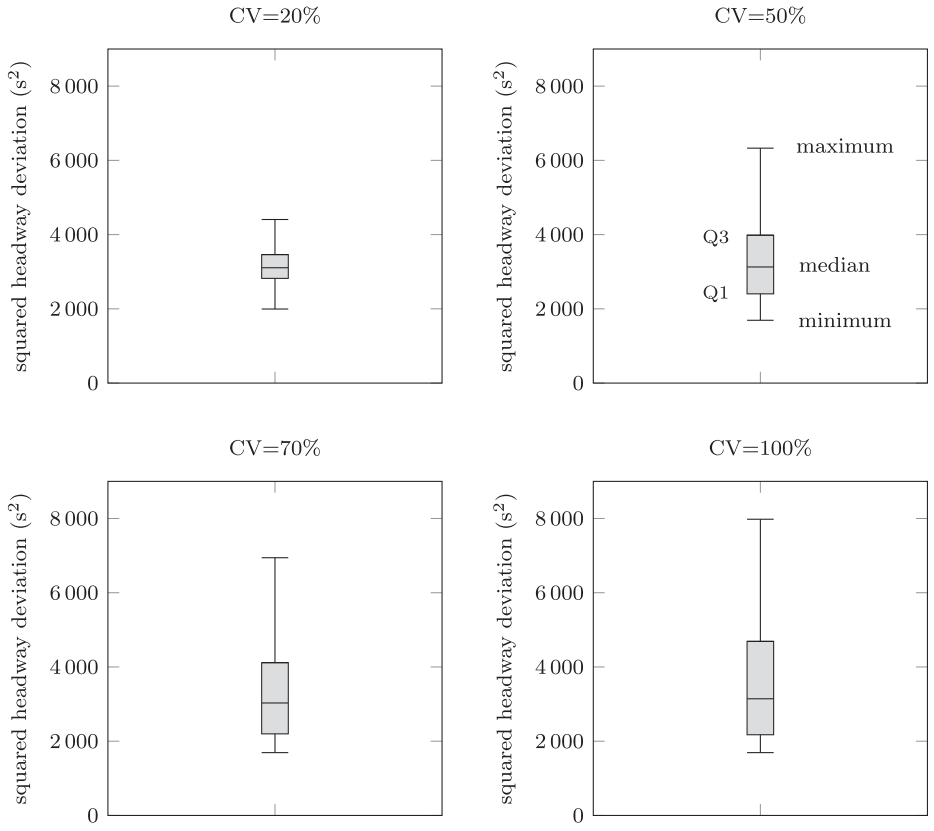

Fig. 7. Squared headway deviation performance when using our control logic in 1,000 simulations for CV = 0.2, 0.5, 0.7, and 1.0, respectively.

they refer to passenger alightings and arrival rates. The performance of our control logic that considers the expected values $(\widetilde{\beta}_{n+1,s}, \lambda_{n,s})$ when making holding decisions is evaluated at every one of the 1,000 simulation scenarios. The same simulation-based evaluation is performed for (progressively) stronger disruptions with CV = 50%, CV = 70%, and CV = 100% by evaluating the performance of our holding solution in 1,000 simulation scenarios at each time.

The performance of our control logic when applied in the aforementioned scenarios is presented in Fig. 7 following the Tukey boxplot convention (McGill et al., 1978). The upper and lower boundaries of the boxes indicate the upper and lower quartiles (i.e., 75th and 25th percentiles denoted as Q3 and Q1, respectively). The black lines vertical to the boxes (whiskers) show the maximum and minimum values that are not outliers. The whiskers are determined by plotting the lowest datum still within 1.5 the interquartile range (IQR = Q3-Q1) of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile.

Interestingly, from the results of Fig. 7, one can conclude that the median performance of our control logic in terms of service regularity has little sensitivity to changes in passenger demand expressed in alightings and passenger arrival rates (the median value is relatively stable and lies in the range of 3107–3143 s^2 regardless the CV level). Although the median performance of our control logic remains stable for different CV levels, both the upper whisker (maximum) and the IQR increase for stronger demand disruptions (e.g., for higher values of CV). That is, even if our control logic is expected to perform similarly in scenarios with mild demand variations (CV = 20%) or strong ones (CV = 100%) in terms of the median performance, stronger demand variations might result in more extreme scenarios with squared headway deviations of up to 7980 s^2. A supplementary sensitivity analysis with respect to travel time variations can be found in the Appendix A.2.

7. Discussion

7.1. Results

The findings of this study show that our analytic solution can be used to determine the holding time of a bus while catering for the capacity limitations of vehicles, instead of resorting to rule-based solutions that focus only on improving the service regularity (i.e., Fu and Yang (2002), Wu et al. (2017)). Clear benefits of our approach are its easy implementation, the reduction of passenger waiting times, and the reduced numbers of stranded passengers - as demonstrated in the experiments of Tables 7 and 10. Because our analytic solution does not require any computational costs, it can be continuously applied every time a bus is ready to depart from a stop using updated information with respect to the inter-station travel times and passenger demand.

With respect to the performance of our holding approach, simulation-based experiments with data from bus line 302 in Singapore showed that our holding solution results in reduced passenger waiting times compared to other bus holding approaches because it avoids holding the bus at a stop after reaching its capacity. This has the benefit of reducing the in-vehicle passenger waiting time without increasing the out-of-vehicle passenger waiting times because passengers who are refused boarding will anyway have to wait for the next available bus.

When applying our control logic to trip n at stop s, we use the expected values of the arrival time, the number of alightings, and the boardings of its following trip, $n + 1$. Obviously, those travel time and demand estimations for trip $n+1$ might result in a reduced performance gain when our expectations differ from the realized values. This was tested with sensitivity analysis tests presented in 6.2.2 and Appendix A.2. The sensitivity analysis tests showed that:

- the median performance (in terms of regularity) when applying our holding solution is rather insensitive to demand variations (i.e., alightings or passenger arrival rates of future trips);
- the performance (in terms of regularity) when applying our holding solution is very sensitive to unexpected travel time variations of future trips.

The latter finding deserves further elaboration. Past works have also reported the limited effect of bus holding measures when the realized travel times deviate significantly from their expected values (see Hickman (2001), Fu and Yang (2002), Berrebi et al. (2018)). Reckon that in real-time bus holding one makes a decision based on the currently available information and historical expectations. Similarly to other real-time problems that are based on currently available information, such as the fastest path problem, a decision which is made now might not be optimal in the short future. Thus, it is imperative to re-optimize the problem when new information is available and this underlines importance of our analytic solution that can be applied in real-time.

7.2. Limitations

Herein, we make explicit the main limitations of our model and its respective analytic solution. Those limitations are:

- it must be applied to high-frequency bus lines (more than 6 buses per hour) that operate under regularity-based schemes and consider the headway regularity as the main objective;
- it cannot provide appropriate benefits in the case of severe disruptions to the daily operations. I.e., if the demand is too high, refusing boardings cannot be avoided even if our analytic solution considers the vehicle capacity. In that case, bus operators should consider more radical measures, such as changes in the planned service provision and resource allocation (i.e., rescheduling, increase of service frequency);
- similarly to other event-based or periodic optimization models, the performance of our approach depends on the accuracy of the estimated inter-station travel times of future trips.

8. Conclusion

This work provided a model, (Q), for performing real-time bus holding under capacity limitations. The consideration of the bus load and the vehicle capacity limits added another dimension to the traditional bus holding problem, and this resulted in a nonlinear, non-smooth model. With the use of slack variables, the nonlinear, non-smooth model was transformed into a quadratic program $(\widetilde{Q})$ with linear (in) equality constraints that can be solved to global optimality. The reformulated program was proven to have an analytic solution.

Our analytic bus holding solution was applied in several idealized scenarios demonstrating the improvement potential compared to two-headway-based and self-equalizing multiple-headway-based methods that do not consider the capacity and the bus loads in the optimization process. In the case study of bus line 302 in Singapore, we show that our control logic can reduce the number of stranded passengers and the average passenger waiting times by deteriorating slightly the service regularity compared to regularity-based holding approaches. Because a real-time holding decision is made based on future expectations that might differ from reality, we investigated the sensitivity of our solution to demand and travel time variations with the use of Monte Carlo simulations. After carrying out sensitivity analysis tests, it was established that the median performance of our solution is relatively insensitive to demand variations, but it can deteriorate significantly in case of inter-station travel time variations from their expected values.

In future research, our approach can be expanded in a wide range of problems involving rail operations. For instance, with the proper modifications, future research can expand our method to railway operations that operate under regularity-based schemes. Other advances could be an expansion of our model to incorporate additional constraints related to the timetables and the recommended total trip travel times. We finally note that our approach can be used even if bus operators prioritize the reduction of bus bunching over refused boardings caused by over-crowding. This can be achieved by reducing the values of the penalty terms M_1, M_2 in our program $(\widetilde{Q})$ enabling the search of holding times that aim at adhering to the planned headways.

CRediT authorship contribution statement

K. Gkiotsalitis: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software,

Validation, Visualization, Writing - original draft. **E.C. van Berkum:** Conceptualization, Investigation, Methodology, Project administration, Supervision, Writing - review & editing.

Appendix A

Lemma Appendix 1. *There is no holding decision, x, for trip n that can satisfy the inequality constraint of Eq. (5) in all cases.*

Proof. Let us consider the extreme case where bus trip $n+1$ is full when it arrives at stop s. Then, $\widetilde{l}_{n+1} = c_{n+1}$. Let also that in this particular case there are no passenger alightings at stop $s, \widetilde{\beta}_{n+1} = 0$, there are no stranded passengers by trip $n, l_n = 0$, the expected arrival time of bus trip $n+1$ is $\widetilde{a}_{n+1,s} > t + \zeta$, and the passenger arrival rate is $\lambda_s > 0$. Then, the inequality constraint of Eq. (5) becomes $((\widetilde{a}_{n+1,s} - (t + x))\lambda_s)(1 + t_b\lambda_s) \leqslant 0$. Since $t_b \geqslant 0 \wedge \lambda_s > 0$, this means that we should satisfy $\widetilde{a}_{n+1,s} \leqslant (t + x)$. Let us write $\widetilde{a}_{n+1,s} > t + \zeta$ as $\widetilde{a}_{n+1,s} = t + \zeta + \epsilon$ where $\epsilon > 0$. Then, $\widetilde{a}_{n+1,s} \leqslant (t+x) \Leftrightarrow t + \zeta + \epsilon \leqslant t + x$ for some $x^* \in \mathbb{R} | 0 \leqslant x^* \leqslant \zeta$. However, $\neq$ $xists x^* \in \mathbb{R} | 0 \leqslant x^* \leqslant \zeta$ such that $t + \zeta + \epsilon \leqslant t + x^*$, and this completes our proof. □

Theorem Appendix 1. $\widetilde{a}_{n+1,s} + \widetilde{\beta}_{n+1} t_a + \left(\widetilde{\beta}_{n+1} t_a \lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)\left(1 + t_b\lambda_s\right)t_b - \nu_2 t_b = \widetilde{a}_{n+1,s} + \widetilde{\beta}_{n+1} t_a + \min\left[\left(\widetilde{\beta}_{n+1} t_a \lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)\left(1 + t_b\lambda_s\right)t_b, \left(c_{n+1} + \widetilde{\beta}_{n+1} - \widetilde{l}_{n+1}\right)t_b\right]$

at the solution of program $(\widehat{Q})$.

Proof. The term $M_2\nu_2$ in the objective function $f(x, \nu_1, \nu_2)$ and the inequality constraints $\nu_2 \geqslant 0$ and $\nu_2 \geqslant \widetilde{l}_{n+1} - \widetilde{\beta}_{n+1} - c_{n+1} + \left(\widetilde{\beta}_{n+1} t_a \lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b\lambda_s)$ force the slack variable ν_2 to attain one of the following two values at the solution of program $(\widehat{Q})$:

- $\nu_2 = 0$, or
- $\nu_2 = \widetilde{l}_{n+1} - \widetilde{\beta}_{n+1} - c_{n+1} + \left(\widetilde{\beta}_{n+1} t_a \lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b\lambda_s)$.

If the solution of program $(\widehat{Q})$ is $\nu_2 = 0$, then from the second inequality constraint of ν_2 in program $(\widehat{Q})$ we get $\widetilde{l}_{n+1} - \widetilde{\beta}_{n+1} - c_{n+1} + \left(\widetilde{\beta}_{n+1} t_a \lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b\lambda_s) \leqslant 0$. We multiply the previous inequality by the positive number t_b which yields $\left[\left(\widetilde{\beta}_{n+1} t_a \lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b\lambda_s)\right]t_b \leqslant \left[c_{n+1} + \widetilde{\beta}_{n+1} - \widetilde{l}_{n+1}\right]t_b$. Hence, for $\nu_2 = 0$, Eq. (10) yields $\widetilde{d}_{n+1,s} = \widetilde{a}_{n+1,s} + \widetilde{\beta}_{n+1} t_a + \left[\left(\widetilde{\beta}_{n+1} t_a \lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b\lambda_s)t_b\right]$. Similarly, if $\nu_2 = 0$ then Eq. (14) yields $\widetilde{d}_{n+1,s} = \widetilde{a}_{n+1,s} + \widetilde{\beta}_{n+1} t_a + \left(\widetilde{\beta}_{n+1} t_a \lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b\lambda_s)t_b$ and this completes the first part of our proof.

Now, if the solution of program $(\widehat{Q})$ is $\nu_2 = \widetilde{l}_{n+1} - \widetilde{\beta}_{n+1} - c_{n+1} + \left(\widetilde{\beta}_{n+1} t_a \lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b\lambda_s)$, then from the first inequality constraint of ν_2 in program $(\widehat{Q})$ we get $\widetilde{l}_{n+1} - \widetilde{\beta}_{n+1} - c_{n+1} + \left(\widetilde{\beta}_{n+1} t_a \lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b\lambda_s) \geqslant 0$. Hence, Eq. (10) yields $\widetilde{d}_{n+1,s} = \widetilde{a}_{n+1,s} + \widetilde{\beta}_{n+1} t_a + \left[\left(c_{n+1} + \widetilde{\beta}_{n+1} - \widetilde{l}_{n+1}\right)t_b\right]$. At the same time, Eq. (14) yields $\widetilde{d}_{n+1,s} = \widetilde{a}_{n+1,s} + \widetilde{\beta}_{n+1} t_a + \left(\widetilde{\beta}_{n+1} t_a \lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b\lambda_s)t_b$ -
$-\left[\widetilde{l}_{n+1} - \widetilde{\beta}_{n+1} - c_{n+1} + \left(\widetilde{\beta}_{n+1} t_a \lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b\lambda_s)\right]t_b = \widetilde{a}_{n+1,s} + \widetilde{\beta}_{n+1} t_a + \left[\left(c_{n+1} + \widetilde{\beta}_{n+1} - \widetilde{l}_{n+1}\right)t_b\right]$ which completes the second part of our proof. □

Theorem Appendix 2. *A local minimizer of* $(\widetilde{Q})$ *is a globally optimal solution.*

Proof. A local minimizer of $(\widetilde{Q})$ is a global minimizer of $(\widetilde{Q})$ if the objective function is convex and the feasible region is a convex set. The feasible region is defined by linear inequalities and is a polyhedron (thus, it is also a *convex set*). Further, we prove that the objective function $f(x, \nu_1, \nu_2)$ is convex with respect to x, ν_1, ν_2.

The first-order partial derivatives of $f(x, \nu_1, \nu_2)$ are

$$\frac{\partial f}{\partial x} = 2x + 2\left(t - d_{n-1,s} - H_s\right) + 2x(\lambda_s k t_b + 1)^2 - 2\left(\lambda_s k t_b + 1\right)\left[\widetilde{a}_{n+1,s} + \widetilde{\beta}_{n+1} t_a + \left(\widetilde{\beta}_{n+1} t_a \lambda_s + \nu_1 + \left(\widetilde{a}_{n+1,s} - t\right)\lambda_s\right)k t_b - \nu_2 t_b - t - H_s\right]$$

$$\frac{\partial f}{\partial \nu_1} = 2k^2 t_b^2 \nu_1 + 2k t_b \left[\widetilde{a}_{n+1,s} + \widetilde{\beta}_{n+1} t_a + \left(\widetilde{\beta}_{n+1} t_a \lambda_s + \left(\widetilde{a}_{n+1,s} - t - x\right)\lambda_s\right)k t_b - \nu_2 t_b - t - x - H_s\right] + M_1$$

Table 11
Optimal Holding decisions for different values of (λ_s, ϕ_n).

Scenarios			Analytic Solution			Solving $\widetilde{Q}$ with CPLEX		
	λ_s	ϕ_n	$\frac{c_n-\phi_n}{\lambda_s}$	$\mathscr{Z}$	x^*	ν_1	ν_2	x^*
I	0.02	40	1000 s	296 s	**296 s**	0	0	**296 s**
II	0.002	40	10000 s	261 s	**261 s**	0	0	**261 s**
III	0.02	58	100 s	296 s	**100 s**	0	0	**100 s**
IV	0.02	55	250 s	296 s	**250 s**	0	0	**250 s**
V	0.05	58	40 s	361 s	**40 s**	0	38.5	**40 s**
VI	0.02	59	50 s	296 s	**50 s**	0	0.84	**50 s**
VII	0.05	40	400 s	361 s	**300 s**	0	16.9	**300 s**
VIII	0.02	62	−100 s	296 s	**0 s**	2.00	1.92	**0 s**

$$\frac{\partial f}{\partial \nu_2} = 2t_b^2\nu_2 - 2t_b \quad \big[\widetilde{a}_{n+1,s} + \widetilde{\beta}_{n+1}t_a + \big(\widetilde{\beta}_{n+1}t_a\lambda_s + \nu_1 + (\widetilde{a}_{n+1,s} - t - x)\lambda_s\big)kt_b - t - x - H_s\big] + M_2$$

Therefore, the *Hessian* matrix of f reads:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial \nu_1} & \frac{\partial^2 f}{\partial x \partial \nu_2} \\ \frac{\partial^2 f}{\partial \nu_1 \partial x} & \frac{\partial^2 f}{\partial \nu_1^2} & \frac{\partial^2 f}{\partial \nu_1 \partial \nu_2} \\ \frac{\partial^2 f}{\partial \nu_2 \partial x} & \frac{\partial^2 f}{\partial \nu_2 \partial \nu_1} & \frac{\partial^2 f}{\partial \nu_2^2} \end{bmatrix} = \begin{bmatrix} 2+2(\lambda_s k t_b+1)^2 & -2(\lambda_s k t_b+1)kt_b & 2(\lambda_s k t_b+1)t_b \\ -2(\lambda_s k t_b+1)kt_b & 2k^2t_b^2 & -2kt_b^2 \\ 2(\lambda_s k t_b+1)t_b & -2kt_b^2 & 2t_b^2 \end{bmatrix}$$

To prove the convexity of f, we should prove that the Hessian matrix, $\mathbf{H}$, with elements $H_{ij} \in \mathbf{H}$, is positive semi-definite (P.S.D.). That is, all the leading principal minors are non-negative:

$$\mathbf{H} \text{is P.S.D.} \Leftrightarrow H_{11} \geqslant 0, \begin{vmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{vmatrix} \geqslant 0, \det\Big(\mathbf{H}\Big) \geqslant 0.$$

In our case, we have $H_{11} = 2 + 2(\lambda_s k t_b + 1)^2 > 0$.

In addition, $\begin{vmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{vmatrix} = \Big(2 + 2(\lambda_s k t_b + 1)^2\Big)2k^2t_b^2 - 4(\lambda_s k t_b + 1)^2k^2t_b^2 = 4k^2t_b^2 > 0.$

Furthermore,

$$\begin{aligned} \det(\mathbf{H}) = & \Big(2 + 2(\lambda_s k t_b + 1)^2\Big)\begin{vmatrix} H_{22} & H_{23} \\ H_{32} & H_{33} \end{vmatrix} \\ & + 2\Big(\lambda_s k t_b + 1\Big)kt_b\begin{vmatrix} H_{21} & H_{23} \\ H_{31} & H_{33} \end{vmatrix} \\ & + 2\Big(\lambda_s k t_{b+1}\Big)t_b\begin{vmatrix} H_{21} & H_{22} \\ H_{31} & H_{32} \end{vmatrix} \\ & = \big(2 + 2(\lambda_s k t_{b+1})^2\big) \cdot 0 + 2\big(\lambda_s k t_{b+1}\big)kt_b \cdot 0 \\ & + 2(\lambda_s k t_b + 1)t_b \cdot 0 = 0. \end{aligned}$$

Thus, f is convex and this completes our proof. We finally note that for *strict* convexity, $\det(\mathbf{H})$ should have been greater than zero. Since this is not the case, we might have more than one globally optimal solutions. □

A.1. Illustrationof equivalency between the solution of $(\widetilde{Q})$ and our analytic solution

Herein we perform numerical experiments which manifest that our proposed analytic solution, x^*, is equivalent to the solution of $(\widetilde{Q})$. The parameter values of the idealized scenarios are presented in Table 5.

To cover multiple cases, we modify the parameter values of λ_s and ϕ_n and compute the respective solutions of the following idealized scenarios. As expected, the analytic solution proposed in Theorem 5.1 is always equivalent to the solution of program $(\widetilde{Q})$, which is solved with the use of CPLEX. The data and source code of each scenario is publicly released in Gkiotsalitis (2019a), and the respective solutions are presented in Table 11. To simplify the notation in Table 11, we set $\mathscr{Z} \triangleq \frac{(\lambda_s\eta+1)\theta-(t-d_{n-1,s}-H_s)}{1+(\lambda_s\eta+1)^2}$. This allows us to

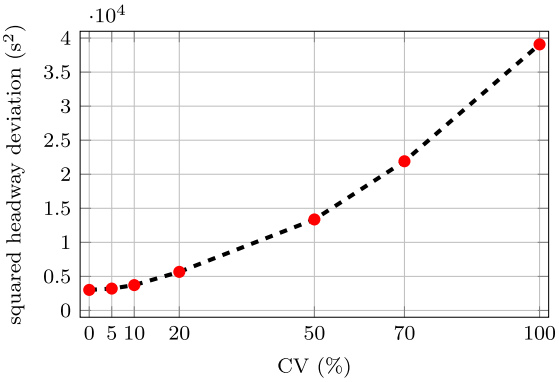

Fig. 8. Mean performance when running 1,000 simulations for different inter-station travel time variation levels.

re-write our analytic solution as

$$x^* = \left\{ \max\left(0, \min\left[\zeta, \frac{c_n - \phi_n}{\lambda_s}, \mathscr{Z}\right]\right)\right\}$$

As shown in Table 11, our analytic solution is equivalent to the solution of CPLEX when solving the mathematical program $(\widetilde{Q})$, regardless the scenario and the values of ν_1, ν_2.

A.2. Sensitivity of our control logic to travel time variations

To investigate the effect of the performance of our control logic to travel time variations from their expected values, we treat the inter-station travel times as uncertain variables with mean $t_{n+1,s}$ and standard deviation CV $\cdot t_{n+1,s}$. As in the previous sub-section, we perform 1,000 simulations for each travel time variation level using six different coefficients of variation: CV = 0.05, 0.1, 0.2, 0.5, 0.7, and 1.0, respectively.

At each one of the $i = \{1, 2, \ldots, 1000\}$ simulations for a CV level, the simulated travel time, $t^i_{n+1,s}$, differs from its expected value, $t_{n+1,s}$, resulting in a realized arrival time, $a^i_{n+1,s}$, that differs from its estimated value, $\widetilde{a}_{n+1,s}$, based on which we determined the holding time of trip n at stop s. The simulated travel time $t^i_{n+1,s}$ is sampled from a restricted normal distribution:

$$t^i_{n+1,s} = \max\left\{t^{free}_{n+1,s}, \sim\mathscr{N}\left(t_{n+1,s}, \text{CV} \cdot t_{n+1,s}\right)\right\} \tag{A.1}$$

where $t^{free}_{n+1,s}$ is the minimum possible travel time based on the historical data observations. Note that sampling values of $t^i_{n+1,s}$ with Eq. (A.1) ensures that a sampled travel time from the normal distribution will not be lower than the smallest possible inter-station travel time, $t^{free}_{n+1,s}$. This logical restriction avoids negative travel time realizations.

The average performances from the 1,000 simulations for each CV level are summarized in Fig. 8.

The mean performance of our control logic in terms of service regularity remains relatively stable for CV values of up to 10%. Then, it deteriorates following a nonlinear pattern until reaching CV = 100%. The nonlinear deterioration can be explained because every time our control logic makes a holding decision, it considers the expected arrival time of the future trip at the same stop. If the realized and the expected arrival times deviate significantly, then our holding decision is based on inaccurate information which impacts its potential gain. This observation is in line with other bus holding works that highlight the importance of accurate travel time information when making a holding decision (see Berrebi et al. (2018)) and demonstrates that a solution can be effective if the CV remains below 10%.

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.trc.2020.102815.

References

Bartholdi III, J.J., Eisenstein, D.D., 2012. A self-coördinating bus route to resist bus bunching. Transport. Res. Part B: Methodol. 46 (4), 481–491.
Berrebi, S.J., Hans, E., Chiabaut, N., Laval, J.A., Leclercq, L., Watkins, K.E., 2018. Comparing bus holding methods with and without real-time predictions. Transport. Res. Part C: Emerg. Technol. 87, 197–211.

Berrebi, S.J., Watkins, K.E., Laval, J.A., 2015. A real-time bus dispatching policy to minimize passenger wait on a high frequency route. Transport. Res. Part B: Methodol. 81, 377–389.
Bin, Y., Zhongzhen, Y., Baozhen, Y., 2006. Bus arrival time prediction using support vector machines. J. Intell. Transport. Syst. 10 (4), 151–158.
Cathey, F., Dailey, D.J., 2003. A prescription for transit arrival/departure prediction using automatic vehicle location data. Transport. Res. Part C: Emerg. Technol. 11 (3–4), 241–264.
Ceder, A., 2007. Public Transit Planning and Operation: Modeling, Practice and Behavior. CRC Press.
Chang, H., Park, D., Lee, S., Lee, H., Baek, S., 2010. Dynamic multi-interval bus travel time prediction using bus transit data. Transportmetrica 6 (1), 19–38.
Chen, M., Liu, X., Xia, J., Chien, S.I., 2004. A dynamic bus-arrival time prediction model based on apc data. Comput.-Aided Civil Infrastruct. Eng. 19 (5), 364–376.
Chen, X., Hellinga, B., Chang, C., Fu, L., 2015. Optimization of headways with stop-skipping control: a case study of bus rapid transit system. J. Adv. Transport. 49 (3), 385–401.
Chen, X., Yu, L., Zhang, Y., Guo, J., 2009. Analyzing urban bus service reliability at the stop, route, and network levels. Transport. Res. Part A: Policy Practice 43 (8), 722–734.
Chien, S.I.-J., Kuchipudi, C.M., 2003. Dynamic travel time prediction with real-time and historic data. J. Transport. Eng. 129 (6), 608–616.
Cortés, C.E., Jara-Díaz, S., Tirachini, A., 2011. Integrating short turning and deadheading in the optimization of transit services. Transport. Res. Part A: Policy Practice 45 (5), 419–434.
Cortés, C.E., Sáez, D., Milla, F., Núñez, A., Riquelme, M., 2010. Hybrid predictive control for real-time optimization of public transport systems' operations based on evolutionary multi-objective optimization. Transport. Res. Part C: Emerg. Technol. 18 (5), 757–769.
Daganzo, C.F., 2009. A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. Transport. Res. Part B: Methodol. 43 (10), 913–921.
Daganzo, C.F., Pilachowski, J., 2011. Reducing bunching with bus-to-bus cooperation. Transport. Res. Part B: Methodol. 45 (1), 267–277.
Delgado, F., Munoz, J.C., Giesen, R., 2012. How much can holding and/or limiting boarding improve transit performance? Transport. Res. Part B: Methodol. 46 (9), 1202–1217.
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. Transp. Res. Rec. 2090 (1), 59–67.
Eberlein, X.J., 1995. Real-time control strategies in transit operations: Models and analysis. Ph.D. thesis, Massachusetts Institute of Technology, Department of Civil and Environmental Engineering.
Eberlein, X.J., Wilson, N.H., Bernstein, D., 2001. The holding problem with real–time information available. Transport. Sci. 35 (1), 1–18.
Fu, L., Yang, X., 2002. Design and implementation of bus-holding control strategies with real-time information. Transport. Res. Rec.: J. Transport. Res. Board 1791, 6–12.
Gintner, V., Kliewer, N., Suhl, L., 2005. Solving large multiple-depot multiple-vehicle-type bus scheduling problems in practice. OR Spectrum 27 (4), 507–523.
Gkiotsalitis, K., 2019a. Bus Holding under Capacity Limitations. https://github.com/KGkiotsalitis/bus-holding-model-under-capacity-limitations.
Gkiotsalitis, K., 2019b. Bus rescheduling in rolling horizons for regularity-based services. J. Intell. Transport. Syst. 1–20.
Gkiotsalitis, K., 2020a. Bus holding of electric buses with scheduled charging times. IEEE Trans. Intell. Transp. Syst. 1–12.
Gkiotsalitis, K., 2020b. A model for the periodic optimization of bus dispatching times. Appl. Math. Model. 82, 785–801.
Gkiotsalitis, K., Cats, O., 2018. Reliable frequency determination: Incorporating information on service uncertainty when setting dispatching headways. Transport. Res. Part C: Emerg. Technol. 88, 187–207.
Gkiotsalitis, K., Cats, O., 2019. Multi-constrained bus holding control in time windows with branch and bound and alternating minimization. Transportmet. B: Transport Dyn. 7, 1258–1285.
Gkiotsalitis, K., Wu, Z., Cats, O., 2019. A cost-minimization model for bus fleet allocation featuring the tactical generation of short-turning and interlining options. Transport. Res. Part C: Emerg. Technol. 98, 14–36.
Hans, E., Chiabaut, N., Leclercq, L., Bertini, R.L., 2015. Real-time bus route state forecasting using particle filter and mesoscopic modeling. Transport. Res. Part C: Emerg. Technol. 61, 121–140.
He, S.-X., Liang, S.-D., Dong, J., Zhang, D., He, J.-J., Yuan, P.-C., 2020. A holding strategy to resist bus bunching with dynamic target headway. Comput. Ind. Eng. 140, 106237.
Hickman, M.D., 2001. An analytic stochastic model for the transit vehicle holding problem. Transport. Sci. 35 (3), 215–237.
Ibarra-Rojas, O., Delgado, F., Giesen, R., Muñoz, J., 2015. Planning, operation, and control of bus transport systems: A literature review. Transport. Res. Part B: Methodol. 77, 38–75.
Kliewer, N., Mellouli, T., Suhl, L., 2006. A time–space network based exact optimization model for multi-depot bus scheduling. Eur. J. Oper. Res. 175 (3), 1616–1627.
Knoppers, P., Muller, T., 1995. Optimized transfer opportunities in public transport. Transport. Sci. 29 (1), 101–105.
Koehler, L.A., Seman, L.O., Kraus, W., Camponogara, E., 2018. Real-time integrated holding and priority control strategy for transit systems. IEEE Trans. Intell. Transp. Syst. 20 (9), 3459–3469.
Leong, W., Goh, K., Hess, S., Murphy, P., 2016. Improving bus service reliability: The singapore experience. Res. Transport. Econ. 59, 40–49.
Li, S., Liu, R., Yang, L., Gao, Z., 2019. Robust dynamic bus controls considering delay disturbances and passenger demand uncertainty. Transport. Res. Part B: Methodol. 123, 88–109.
Liu, Z., Yan, Y., Qu, X., Zhang, Y., 2013. Bus stop-skipping scheme with random travel time. Transport. Res. Part C: Emerg. Technol. 35, 46–56.
Luo, X., Liu, S., Jin, P.J., Jiang, X., Ding, H., 2017. A connected-vehicle-based dynamic control model for managing the bus bunching problem with capacity constraints. Transport. Plann. Technol.l 40 (6), 722–740.
Marguier, P., 1985. Bus route performance evaluation under stochastic conditions. Ph.D. thesis. Massachusetts Institute of Technology, Cambridge, MA.
McGill, R., Tukey, J.W., Larsen, W.A., 1978. Variations of box plots. Am. Stat. 32 (1), 12–16.
Meng, Q., Qu, X., 2013. Bus dwell time estimation at bus bays: A probabilistic approach. Transport. Res. Part C: Emerg. Technol. 36, 61–71.
Muñoz, J.C., Cortés, C.E., Giesen, R., Sáez, D., Delgado, F., Valencia, F., Cipriano, A., 2013. Comparison of dynamic control strategies for transit operations. Transport. Res. Part C: Emerg. Technol. 28, 101–113.
Newell, G.F., 1974. Control of pairing of vehicles on a public transportation route, two vehicles, one control point. Transport. Sci. 8 (3), 248–264.
Nikolaou, M., 2001. Model predictive controllers: A critical synthesis of theory and industrial needs. Adv. Chem. Eng. 26, 131–204.
Osuna, E., Newell, G., 1972. Control strategies for an idealized public transportation system. Transport. Sci. 6 (1), 52–72.
Powell, M.J., 2009. The bobyqa algorithm for bound constrained optimization without derivatives. Cambridge NA Report NA2009/06, University of Cambridge, Cambridge, pp. 26–46.
Rossetti, M.D., Turitto, T., 1998. Comparing static and dynamic threshold based control strategies. Transport. Res. Part A: Policy Practice 32 (8), 607–620.
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 (1), 61–86.
Sánchez-Martínez, G., Koutsopoulos, H., Wilson, N., 2016. Real-time holding control for high-frequency transit with dynamics. Transport. Res. Part B: Methodol. 83, 1–19.
Shalaby, A., Farhan, A., 2004. Prediction model of bus arrival and departure times using avl and apc data. J. Public Transport. 7 (1), 3.
Shen, S., Wilson, N.H., 2001. An optimal integrated real-time disruption control model for rail transit systems. In: Computer-aided Scheduling of Public Transport. Springer, pp. 335–363.
Sun, A., Hickman, M., 2004. Scheduling considerations for a branching transit route. J. Adv. Transport. 38 (3), 243–290.
Sun, D.J., Xu, Y., Peng, Z.-R., 2015. Timetable optimization for single bus line based on hybrid vehicle size model. J. Traffic Transport. Eng. (English Ed.) 2 (3), 179–186.

Trompet, M., Liu, X., Graham, D., 2011. Development of key performance indicator to compare regularity of service between urban bus operators. Transport. Res. Rec.: J. Transport. Res. Board 2216, 33–41.
van Hinsbergen, C.I., Van Lint, J., Van Zuylen, H., 2009. Bayesian committee of neural networks to predict travel times with confidence intervals. Transport. Res. Part C: Emerg. Technol. 17 (5), 498–509.
Van Lint, J., Hoogendoorn, S., van Zuylen, H.J., 2005. Accurate freeway travel time prediction with state-space neural networks under missing data. Transport. Res. Part C: Emerg. Technol. 13 (5–6), 347–369.
Van Oort, N., Wilson, N., Van Nes, R., 2010. Reliability improvement in short headway transit services: Schedule-and headway-based holding strategies. Transport. Res. Rec.: J. Transport. Res. Board 2143, 67–76.
Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2005. Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach. Transport. Res. Part C: Emerg. Technol. 13 (3), 211–234.
Wren, A., Rousseau, J.-M., 1995. Bus driver scheduling—an overview. In: Computer-aided Transit Scheduling. Springer, pp. 173–187.
Wu, W., Liu, R., Jin, W., 2017. Modelling bus bunching and holding control with vehicle overtaking and distributed passenger boarding behaviour. Transport. Res. Part B: Methodol. 104, 175–197.
Wu, Y., Yang, H., Tang, J., Yu, Y., 2016. Multi-objective re-synchronizing of bus timetable: Model, complexity and solution. Transport. Res. Part C: Emerg. Technol. 67, 149–168.
Yu, B., Yang, Z., Yao, J., 2009. Genetic algorithm for bus frequency optimization. J. Transport. Eng. 136 (6), 576–583.
Zolfaghari, S., Azizi, N., Jaber, M.Y., 2004. A model for holding strategy in public transit systems with real-time information. Int. J. Transport Manage. 2 (2), 99–110.