Transportation Research Part B 45 (2011) 1831–1845

Contents lists available at ScienceDirect

Transportation Research Part B

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

Dynamic bus holding strategies for schedule reliability: Optimal linear control and performance analysis

Yiguang Xuan *, Juan Argote, Carlos F. Daganzo

Institute of Transportation Studies, University of California, Berkeley, CA 94720, United States

ARTICLE INFO

Article history:
Received 4 April 2011
Received in revised form 15 July 2011
Accepted 17 July 2011

Keywords:
Bus bunching
Bus schedule reliability
Headway variance
Commercial speed
Dynamic holding
Control method

ABSTRACT

As is well known, bus systems are naturally unstable. Without control, buses on a single line tend to bunch, reducing their punctuality in meeting a schedule. Although conventional schedule-based strategies that hold buses at control points can alleviate this problem these methods require too much slack, which slows buses. This delays on-board passengers and increases operating costs.

It is shown that dynamic holding strategies based on headways alone cannot help buses adhere to a schedule. Therefore, a family of dynamic holding strategies that use bus arrival deviations from a virtual schedule at the control points is proposed. The virtual schedule is introduced whether the system is run with a published schedule or not. It is shown that with this approach, buses can both closely adhere to a published schedule and maintain regular headways without too much slack.

A one-parameter version of the method can be optimized in closed form. This simple method is shown to be near-optimal. To put it in practice, the only data needed in real time are the arrival times of the current bus and the preceding bus at the control point relative to the virtual schedule. The simple method was found to require about 40% less slack than the conventional schedule-based method. When used only to regulate headways it outperforms headway-based methods.

1. Introduction

Bus schedule reliability is an essential attribute of a bus system, and is consistently ranked as one of the major concerns by passengers (Paine et al., 1967; Golob et al., 1972; Wallin and Wright, 1974). Unfortunately, bus systems are naturally unstable in the sense that a small disturbance such as a traffic disruption can start a vicious cycle that causes buses to bunch. The tendency to bunch is due to the fact that the loading time of a bus at a station[1] is a non-decreasing function of the headway between buses. As first explained in Newell and Potts (1964), buses that run early encounter and serve fewer passengers, and tend to catch up with the buses in front of them, while late buses tend to fall further behind. The result is the bus bunching phenomenon, which makes the schedule useless and increases the average waiting time of passengers.

This problem can be alleviated by letting buses skip stops (see e.g., Suh et al., 2002; Fu et al., 2003; Sun and Hickman, 2005; Delgado et al., 2009). The problem can be also alleviated with holding strategies that do not leave passengers stranded; see e.g. the pioneering works in Osuna and Newell (1972), Newell (1974), and Barnett (1974). This paper focuses on holding strategies. They are characterized by embedding slack time in the schedule, and holding buses at each control point for a period of time before their scheduled departure. A bus is generally held longer if it is ahead of schedule and shorter

* Corresponding author.
E-mail address: xuanyg@berkeley.edu (Y. Xuan).

[1] To avoid confusion, we use bus stations to mean locations where buses stop to pick up or drop off passengers. Stop is only used as a verb.

0191-2615/$ - see front matter
doi:10.1016/j.trb.2011.07.009

(or not even held) if it is behind schedule, so that the instability can be neutralized. The most common form of holding is the "schedule-based" method, in which drivers are instructed never to depart a control point ahead of a pre-published schedule.

Among the literature that analytically addresses the bus bunching problem with holding methods, most of the studies (Osuna and Newell, 1972; Newell, 1974; Barnett, 1974; Hickman, 2001; Eberlein et al., 2001; Zhao et al., 2006) try to minimize passenger time (either waiting time at the station only or both waiting at the station and riding time on board). Problems based on this objective are difficult, and most of these studies only discuss problems with one bus line, a single control point, and either one or two buses.[2] Some use "rolling horizon" heuristics (Eberlein et al., 2001). Many more studies resort to simulations (Koffman, 1978; Turnquist and Bowman, 1980; Abkowitz et al., 1986; Vandebona and Richardson, 1986; Senevirante, 1990; Adamski and Turnau, 1998) due to the difficult nature of the problem.

There is also a literature that uses control theory. Daganzo (2009a) approached the bus bunching problem from a different angle: instead of minimizing passenger waiting time, the paper proposed a headway-based dynamic holding strategy to reduce the amount of slack time in the schedule, subject to a headway variability constraint. The idea was to increase the commercial speed of buses[3] while compensating for the effects of small disturbances (e.g., due to traffic). With this new objective, the paper analytically addressed a broader range of problems: systems with many buses, many control points, and stochastic cruising time. The article proposed a general form for this family of dynamic holding strategies by defining a convolution kernel. However, it only studied in depth a particular case: a headway-based control in which buses were held based only on the expected demand and their forward headway (the headway between the current bus and the bus in front) for systems where a schedule is not published. It was shown that headways could be regularized with less slack than required by the schedule-based control method. Unfortunately, as explained in the reference, the method cannot always compensate for large disturbances, such as those due to bus breakdowns.

To alleviate this problem, Daganzo and Pilachowski (2009, 2011) and Daganzo (2009b) proposed a cooperative control method in which bus speed was regulated based on the expected demand and the spacings between the current bus and the preceding and following buses. This method was able to compensate for large disturbances. An Eulerian version of the (Lagrangian) method in Daganzo and Pilachowski (2009, 2011), in which holding times are based on the forward and backward headways (the headway between the current bus and the bus behind), can be easily constructed and will be studied in this paper.[4]

More recently, a holding method in which the holding times are based only on the backward headway, independent of demand, has been proposed (Bartholdi, 2011). The method is appealing because it is very simple, does not require information on demand, and it has been tested successfully with an experiment for a case with low demand. It is claimed that this approach can also compensate for large disturbances (Bartholdi, 2011).

The results about to be described build on Daganzo (2009a). A general control method is proposed that uses both the arrival times of all buses at stations and a virtual schedule. As such, the method includes as special cases all three existing models in the control genre. The virtual schedule is different from the published schedule, and is used even if there is no published schedule, as occurs in bus lines with short headways. Unlike the existing methods, we found that with this general control method, buses cannot only maintain regular headways but also stick to their schedule. This is important because the control method then can be applied to bus lines with both long and short headways. The optimal parameters of the general control method, i.e. those which provide the maximum commercial speed for a given schedule/headway reliability level, are also identified. It turned out that a one-parameter version of the method, which only requires information on the current bus and its leading bus, is near-optimal and outperforms other existing control alternatives.

The paper is organized as follows. Section 2 presents the assumptions and the bus motion laws under the general control method. Section 3 proves that with this method buses are able to adhere to their schedule with only minor random deviations and also have bounded headway variances. Section 4 introduces a mathematical program to obtain the optimal general control and demonstrates that a simple version of the general control method is near-optimal. Section 5 shows that the simple version of the general control outperforms the other existing methods. Finally Section 6 summarizes the main findings.

2. Assumptions and bus motion formulation

2.1. Assumptions

The same assumptions are made as in Daganzo (2009a). (a) The number of bus dispatches and stations in the system can be as high as desired. (b) Buses are always dispatched on time with equal headways from the first station. (c) The bus capacity for passengers is unlimited. (d) Buses do not pass each other. (e) The mean vehicle cruising times between stations are time-independent but location-dependent constants. (f) The passenger arrival processes are stationary but station-dependent. The mean bus loading time is dominantly affected by boarding passengers, and is therefore proportional to the headway. (g) The actual vehicle trip times between stations including both cruising time and loading time are assumed to be mutually independent random variables with a variance that is independent of the headway. (h) Only those passengers

[2] The exception is Eberlein et al. (2001) which addresses an arbitrary number of buses. But the study does not model the cruising time between two stations stochastically.

[3] The commercial speed of buses is the total distance traveled divided by total time taken (including scheduled holding).

[4] Daganzo (2006) demonstrates the close connection between Lagrangian and Eulerian coordinates in the context of automobile motion.

arriving during the inter-arrival time will board the bus, and holding is then applied after the boarding process. (i) There exists enough slack to ensure that the holding time never runs short. (j) Buses stop at all stations and holding is also applied at all stations.

Assumption (a) makes the analysis comparable to a situation with a finite number of buses but enough layover time, so that buses are always dispatched on time. Assumptions (b) and (c) are reasonable if the operational design problem for a bus line (i.e., to choose proper dispatching headways, fleet size, and vehicle size for a given route) has already been addressed. Assumption (d) is reasonable if the system is well managed. But even if passing is allowed, we can renumber the buses and results will not change much. Assumption (e) is reasonable if the environment is approximately time-independent. Assumption (f) is appropriate for bus systems, where boardings and alightings occur simultaneously and the alighting time per passenger is much smaller than the boarding time per passenger. Assumption (g) is a reasonable approximation if buses are operated without severe congestion, or with their own right-of-way, i.e., segregated from car traffic. Assumption (h) greatly simplifies the formulation with only negligible effect, since the loading and holding times are much shorter compared with the inter-arrival time. Assumption (i) makes the formulation linear and the problem tractable, as will soon be shown. Assumption (j) can be relaxed (see Appendix A).

2.2. Bus motion with a control law

Let us use n ($n = 0, 1, 2, \ldots$) to denote the bus number (the buses dispatched first have smaller numbers) and s ($s = 0, 1, 2 \ldots$) to denote the station number (increasing s in the traveling direction). The notation follows Daganzo (2009a):

- $t_{n,s}$ is the scheduled arrival time of bus n at station s. The $t_{n,s}$'s form the virtual schedule for buses; they are not the published schedule to passengers. A published schedule can be obtained by shifting the virtual schedule earlier in time to ensure that buses never depart ahead of the published schedule.
- $a_{n,s}$ is the actual arrival time of bus n at station s.
- $\varepsilon_{n,s} = a_{n,s} - t_{n,s}$ is the deviation from scheduled arrival time of bus n at station s.
- $h_{n,s} = a_{n,s} - a_{n-1,s}$ is the time headway between bus n and its leading bus at station s.
- H is the scheduled headway.
- c_s is the average cruising time from station s to $s + 1$, which includes the time to accelerate and decelerate, but does not include the dwell time to serve passengers.
- $\upsilon_{n,\,s+1}$ is the random noise in the trip time of bus n between station s to $s + 1$, whose mean is zero and variance is $\sigma^2_{n,s+1}$.
- $D_{n,s}$ is the holding time applied to bus n at station s.[5]
- d_s is the amount of slack time in the virtual schedule at station s (i.e., the holding time if the bus arrives when expected).
- β_s is a dimensionless measure for the demand rate at station s, where the demand rate (in passengers/hour) is normalized by the passenger boarding rate (also in passengers/hour). This implies that the passenger loading time at station s increases by β_s if headway increases by one unit of time. Typical values of β_s range from 10^{-2} to 10^{-1}.

With the above notation, the scheduled arrival times can be formulated as:

$$t_{n,s+1} = t_{n,s} + \beta_s H + d_s + c_s, \tag{1a}$$
$$t_{n,s} = t_{n-1,s} + H, \tag{1b}$$

with $t_{n,s} + \beta_s H + d_s$ being the scheduled departure times.[6] The actual arrival times obey:

$$a_{n,s+1} = a_{n,s} + \beta_s h_{n,s} + D_{n,s} + c_s + \upsilon_{n,s+1}. \tag{2}$$

By combining Eqs. (1) and (2), it is possible to express the motion of buses in terms of $\varepsilon_{n,s}$:

$$\varepsilon_{n,s+1} = \varepsilon_{n,s} + \beta_s(\varepsilon_{n,s} - \varepsilon_{n-1,s}) + \upsilon_{n,s+1} + (D_{n,s} - d_s). \tag{3}$$

It is assumed that $D_{n,s}$ is a linear function of the arrival times of all buses at station s, $a_{i,s}$, or equivalently as a function of the deviations from the schedule, the $\varepsilon_{i,s}$.[7]

$$D_{n,s} = d_s + \sum_i \gamma_i \varepsilon_{n-i,s}. \tag{4a}$$

It is convenient to write this function as:

[5] In reality, we recommend holding buses en-route by slowing them down shortly after departing the station and by providing drivers with adequate real-time information. This unnerves passengers less, and releases station capacity in those cases where the station is also used by other bus lines as well.

[6] Passenger boarding time is expressed as $\beta_s H$ per assumption (h). If passengers who arrive after the bus arrival are allowed to board the bus, the expression would be $\beta_s H/(1 - \beta_s)$. The two expressions are similar, however, when demand is low: $\beta_s \ll 1$.

[7] The arrival times of the current bus and the buses in front, $a_{i,s}$'s for $i \leqslant n$, are readily available when bus n arrives at station s. But the arrival times of the buses behind, $a_{i,s}$'s for $i > n$, can only be predicted. For now, we will assume that we have perfect information, i.e., we know all $a_{i,s}$'s. We will demonstrate later in the paper that this type of information is not really needed for a near-optimal control method.

$$D_{n,s} = d_s - [(1+\beta_s)\varepsilon_{n,s} - \beta_s\varepsilon_{n-1,s}] + \sum_i f_i\varepsilon_{n-i,s}, \tag{4b}$$

because plugging (4b) into (3), yields the simple relation:

$$\varepsilon_{n,s+1} = \sum_i f_i\varepsilon_{n-i,s} + \upsilon_{n,s+1}. \tag{4c}$$

Note $f_i = \gamma_i$ for all i except $f_0 = \gamma_0 + 1 + \beta_s$ and $f_1 = \gamma_1 - \beta_s$. Also note that

$$\sum_i f_i = 1 + \sum_i \gamma_i. \tag{4d}$$

To define a specific control method one needs to specify the slack times d_s at each station and all the control coefficients $\{\ldots, f_{-1}, f_0, f_1, \ldots\}$.

All three holding strategies in the control genre mentioned in the introduction are special cases of this general control method. For example, if we set $f_0 = 1 + \beta_s$, $f_1 = -\beta_s$ and $f_i = 0\ \forall i \notin \{0,1\}$, we obtain the case with no control, because then $D_{n,s} = d_s = 0$. In this case, the bus motion is governed by

$$\varepsilon_{n,s+1} = (1+\beta_s)\varepsilon_{n,s} - \beta_s\varepsilon_{n-1,s} + \upsilon_{n,s+1}. \quad (\textit{no control}) \tag{5}$$

The conventional schedule-based control method is the case with $f_i = 0\ \forall i$. In this case, the drivers are instructed not to depart the control point before the scheduled departure time. If there is enough slack time in the schedule as per assumption (i), the buses can always depart the control point on schedule, i.e., $a_{n,s} + \beta_s h_{n,s} + D_{n,s} = t_{n,s} + \beta_s H + d_s$. Therefore, it follows that

$$D_{n,s} = d_s - [\varepsilon_{n,s} + \beta_s(\varepsilon_{n,s} - \varepsilon_{n-1,s})], \quad (\textit{sch. control}) \tag{6a}$$

and

$$\varepsilon_{n,s+1} = \upsilon_{n,s+1}. \quad (\textit{sch. control}) \tag{6b}$$

The method in Daganzo (2009a), which is based on the forward headway is:

$$D_{n,s} = d_s - (\alpha + \beta_s)(h_{n,s} - H). \quad (\textit{forward headway}) \tag{7a}$$

This can be expressed as a function of the deviation from the schedule as:

$$D_{n,s} = d_s - (\alpha + \beta_s)(\varepsilon_{n,s} - \varepsilon_{n-1,s}). \quad (\textit{forward headway}) \tag{7b}$$

This is the special case of (4b) with $f_0 = 1 - \alpha$, $f_1 = \alpha$, $f_i = 0\ \forall i \notin \{0,1\}$, where $0 < \alpha < 1$, and (4c) becomes:

$$\varepsilon_{n,s+1} = (1-\alpha)\varepsilon_{n,s} + \alpha\varepsilon_{n-1,s} + \upsilon_{n,s+1}. \quad (\textit{forward headway}) \tag{7c}$$

Similarly, the Eulerian version of the (Lagrangian) method in Daganzo and Pilachowski (2009, 2011) and Daganzo (2009b), which is based on both the forward and backward headways, is obtained by setting $f_{-1} = f_1 = \alpha$, $f_0 = 1 - 2\alpha$, $f_i = 0\ \forall i \notin \{-1,0,1\}$, where $0 < \alpha < 1/2$. This yields:

$$D_{n,s} = d_s + \alpha(h_{n+1,s} - H) - (\alpha + \beta_s)(h_{n,s} - H), \quad (\textit{two-way headway}) \tag{8a}$$

$$D_{n,s} = d_s + \alpha\varepsilon_{n+1,s} - (\beta_s + 2\alpha)\varepsilon_{n,s} + (\beta_s + \alpha)\varepsilon_{n-1,s}, \quad (\textit{two-way headway}) \tag{8b}$$

and

$$\varepsilon_{n+1,s} = \alpha\varepsilon_{n+1,s} + (1-2\alpha)\varepsilon_{n,s} + \alpha\varepsilon_{n-1,s} + \upsilon_{n,s+1}. \quad (\textit{two-way headway}) \tag{8c}$$

Finally, a demand-independent method based on the backward headway alone has been proposed (Bartholdi, 2011). The control law of this method is: $D_{n,s} = \alpha h_{n+1,s} = \alpha H + \alpha(h_{n+1,s} - H) = \alpha H + \alpha\varepsilon_{n+1,s} - \alpha\varepsilon_{n,s}$. This method uses a slack time αH, which can be quite large for the values of α ($\alpha \approx 0.5$) that are recommended for stability purposes. Thus, to compare the backward headway approach fairly (and more favorably) in our framework, we will study a variant in which both the slack time d_s and the control parameter α can be freely chosen. This method is obtained by setting $f_{-1} = \alpha$, $f_0 = 1 + \beta_s - \alpha$, $f_1 = -\beta_s$ and $f_i = 0\ \forall i \notin \{-1,0,1\}$; then we have:

$$D_{n,s} = d_s + \alpha(h_{n+1,s} - H), \quad (\textit{backward headway}) \tag{9a}$$

$$D_{n,s} = d_s + \alpha\varepsilon_{n+1,s} - \alpha\varepsilon_{n,s}, \quad (\textit{backward headway}) \tag{9b}$$

and

$$\varepsilon_{n,s+1} = \alpha\varepsilon_{n+1,s} + (1 + \beta_s - \alpha)\varepsilon_{n,s} - \beta_s\varepsilon_{n-1,s} + \upsilon_{n,s+1}. \quad (\textit{backward headway}) \tag{9c}$$

3. Stability analysis

This section shows that with the general control method, buses are able to adhere to their schedule with bounded deviations. Of course, this means that they can also maintain regular headways. To illustrate these properties, as in Daganzo

(2009a), we first express the summation term in (4c) as the convolution (denoted with $*$) of two vectors: the bus deviations from the schedule $\boldsymbol{\varepsilon}_s = [\ldots \varepsilon_{n-1,s}, \varepsilon_{n,s}, \varepsilon_{n+1,s} \ldots]^T$ and the kernel of the convolution (the set of control coefficients) $\mathbf{f} = [\ldots, f_{-1}, f_0, f_1 \ldots]^T$. The nth element of the convolution is: $[\mathbf{f} * \boldsymbol{\varepsilon}_s]_n = \sum_k f_{n-k}\varepsilon_{k,s}$. If we also define $\boldsymbol{\upsilon}_s = [\ldots \upsilon_{n-1,s}, \upsilon_{n,s}, \upsilon_{n+1,s} \ldots]^T$ as the vector of disturbances, then the vector form of (4c) is $\boldsymbol{\varepsilon}_{s+1} = \mathbf{f} * \boldsymbol{\varepsilon}_s + \boldsymbol{\upsilon}_{s+1}$.

Now apply the convolution iteratively, so that the $\boldsymbol{\varepsilon}$. terms can be expressed as a function of the control coefficients $\mathbf{f}$ and the noise terms $\boldsymbol{\upsilon}$.. This yields:

$$\boldsymbol{\varepsilon}_{s+1} = \boldsymbol{\upsilon}_{s+1} + \mathbf{f} * \boldsymbol{\varepsilon}_s = \boldsymbol{\upsilon}_{s+1} + \mathbf{f} * (\boldsymbol{\upsilon}_s + \mathbf{f} * \boldsymbol{\varepsilon}_{s-1}) = \boldsymbol{\upsilon}_{s+1} + \mathbf{f} * \boldsymbol{\upsilon}_s + \mathbf{f} * \mathbf{f} * \boldsymbol{\varepsilon}_{s-1} = \ldots.$$

Next define $\mathbf{f}_{|j}$ to be the jth self-convolution of $\mathbf{f}$ (i.e., $\mathbf{f}_{|j} = \mathbf{f} * \mathbf{f}_{|j-1}$, where $\mathbf{f}_{|0} = [\ldots 0, 1, 0 \ldots]^T$). Since assumption (b) states that buses are always dispatched from the first station on time ($\boldsymbol{\varepsilon}_0 = \mathbf{0}$), the above expression becomes:

$$\boldsymbol{\varepsilon}_{s+1} = \sum_{j=0}^{s} \mathbf{f}_{|j} * \boldsymbol{\upsilon}_{s+1-j}, \tag{10a}$$

which expands to

$$\varepsilon_{n,s+1} = \sum_{j=0}^{s} [\mathbf{f}_{|j} * \boldsymbol{\upsilon}_{s+1-j}]_n = \sum_{j=0}^{s} \sum_i f_{i|j} \upsilon_{n-i,s+1-j}. \tag{10b}$$

This expression is fundamental in the analysis of both headway and schedule deviations.

3.1. Deviations from schedule

Let us define $\sigma_\varepsilon^2(\mathbf{f}, n, s)$ as the function that returns $\mathrm{var}(\varepsilon_{n,s})$ given $\mathbf{f}$, n and s. Given Eq. (10), if we assume that the noise terms are independent and identically distributed (i.i.d.) with variance σ^2, then

$$\mathrm{var}(\varepsilon_{n,s}) = \sigma^2 \sum_{j=0}^{s-1} \sum_i (f_{i|j})^2. \tag{11a}$$

Since all the terms in this summation are non-negative, an upper bound to the variance of $\varepsilon_{n,s}$ is:

$$\sigma_\varepsilon^2(\mathbf{f}) \equiv \lim_{\substack{n \to \infty \\ s \to \infty}} \mathrm{var}(\varepsilon_{n,s}) = \sigma^2 \sum_{j=0}^{\infty} \sum_i (f_{i|j})^2 \geqslant \sigma_\varepsilon^2(\mathbf{f}, n, s), \tag{11b}$$

which will be our measure of schedule reliability. It is now possible to see that the following is true.

Lemma 1. *Define $F = \sum_i |f_i|$, then $\sum_i |f_{i|j}| \leqslant F^j, \forall j \geqslant 1$.*

Proof

$$\sum_i |f_{i|j}| = \sum_i \left| \sum_k f_{k|j-1} f_{i-k} \right| \leqslant \sum_i \sum_k |f_{k|j-1}||f_{i-k}| = \sum_k \left(|f_{k|j-1}| \sum_i |f_{i-k}|\right) = F \sum_k |f_{k|j-1}| \leqslant \cdots \leqslant F^2 \sum_k |f_{k|j-2}| \leqslant \cdots \leqslant F^j.$$

□

Theorem 1. (**Sufficient condition for bounded deviation from a schedule**)*If $F < 1$, then $\sigma_\varepsilon^2/\sigma^2 = \sum_{j=0}^{\infty} \sum_i (f_{i|j})^2 \leqslant 1/(1 - F^2)$.*

Proof. An upper bound of $\sum_{j=0}^{\infty} \sum_i (f_{i|j})^2$ is obtained by choosing the $|f_{i|j}|$ that maximizes $\sum_i (f_{i|j})^2$ subject to $\sum_i |f_{i|j}| \leqslant F^j$, as per the previous lemma. Since the objective function $\sum_i (f_{i|j})^2$ is convex, its maximum is reached at the vertices of the feasible region, see Rockafellar (1970). The vertices of the feasible region here are: either $|f_{i|j}| = 0$, $\forall i$; or $|f_{i|j}| = F^j$ for a single i and $|f_{i|j}| = 0$ for all the other i. The maximum is reached in any of the latter cases and therefore $\sum_i (f_{i|j})^2$ is bounded above by F^{2j}. Thus, it now follows that $\sum_{j=0}^{\infty} \sum_i (f_{i|j})^2 \leqslant \sum_{j=0}^{\infty} F^{2j} = 1/(1 - F^2)$. □

The theorem shows that any control method whose coefficients fulfill the condition $F = \sum_i |f_i| < 1$ will be able to adhere to a schedule indefinitely with some degree of reliability.

Note, all headway-based control methods such as (7), (8) or (9) have $F \geqslant 1$; so Theorem 1 does not apply. It is now shown that all headway-based control methods with a finite number of non-zero coefficients have an unbounded measure of schedule reliability, $\sigma_\varepsilon^2(\mathbf{f})$. The methods in question are of the form:

$$D_{n,s} = d_s + \sum_i \alpha_i h_{n-i,s}. \tag{12}$$

This includes (7a), (8a), and (9a) as special cases.

Lemma 2. *Define* $F' = \sum_i f_i$, *then* $\sum_i f_{i|j} = (F')^j, \forall j \geqslant 1.$

Proof

$$\sum_i f_{i|j} = \sum_i \sum_k f_{i-k} f_{k|j-1} = \sum_k f_{k|j-1} \sum_i f_{i-k} = F' \sum_k f_{k|j-1} = F' \sum_i f_{i|j-1}, \quad \forall j \geqslant 1.$$

Applying this result recursively, we get:

$$\sum_i f_{i|j} = (F')^j, \quad \forall j \geqslant 1.$$

□

Lemma 3. (**Sufficient condition for unbounded deviation from a schedule**)*If* **f** *has at most 2n+1 non-zero coefficients centered around i = 0 so that* $F' = \sum_i f_i = \sum_{i=-n}^{n} f_i = 1$ *and* $f_i = 0$ *for all other i then* $\sigma_\varepsilon^2/\sigma^2 = \sum_{j=0}^{\infty}\sum_i (f_{i|j})^2$ *is unbounded.*

Proof. Notice that the index of the non-zero terms of the nth convolution must be in the interval $[-nj, nj]$. Thus $\sum_i (f_{i|j})^2 = \sum_{i=-nj}^{nj} (f_{i|j})^2$, which consists of $2nj+1$ terms. A lower bound to $\sum_{i=-nj}^{nj} (f_{i|j})^2$ is obtained by choosing the $f_{i|j}$ values that minimize $\sum_{i=-nj}^{nj} (f_{i|j})^2$, subject to $\sum_{i=-nj}^{nj} f_{i|j} = 1$ as per Lemma 2. The minimum arises when all the terms in these summations are equal, i.e., when $f_{i-j} = 1/(2nj+1)$, $\forall\ i \in [-nj, nj]$.

$$\sum_i (f_{i|j})^2 = \sum_{i=-nj}^{nj} (f_{i|j})^2 \geqslant \frac{2nj+1}{(2nj+1)^2} = \frac{1}{2nj+1}.$$

Since the sum $\sum_{j=0}^{\infty} \frac{1}{2nj+1}$ diverges because it is a special case of the general harmonic series, so does $\sigma_\varepsilon^2/\sigma^2 = \sum_{j=0}^{\infty}\sum_i (f_{i|j})^2$ when $F' = 1$. □

Theorem 2. *All the headway-based methods with a finite number of non-zero coefficients, including* (7), (8) and (9), *result in an unbounded deviation from the schedule,* $\sigma_\varepsilon^2(\mathbf{f})$.

Proof. Since the number of non-zero coefficients is finite, the indices of these coefficients must be contained in the interval $[-n, n]$ for a sufficiently large n (some of the coefficients in this interval may be zero). Thus in view of Theorem 2, it suffices to show that $F' = 1$. To see this, express (12) in terms of the schedule deviations:

$$D_{n,s} = d_s + \sum_i \alpha_i h_{n-i,s} = \left(d_s + H\sum_i \alpha_i\right) + \sum_i (\alpha_i \varepsilon_{n-i,s} - \alpha_i \varepsilon_{n-i-1,s}).$$

Note from above that $\gamma_i = \alpha_i - \alpha_{i-1}$ and therefore $\sum_i \gamma_i = 0$. Then it follows from (4d) that $F' = \sum_i f_i = 1$. □

In conclusion, any method with $F < 1$ will exhibit bounded deviations from a schedule, while headway-based methods (which have $F' = 1$) cannot provide this type of service. Bounded deviations from a schedule are important for systems with long headways in which passenger arrivals are not uniform in time, but adjust to the schedule. Bowman and Turnquist (1981) showed that if passengers choose their arrival times to minimize their wait, then their average waiting time is proportional to the buses' average deviations from their schedule; see also Daganzo (1997a).

3.2. Headway variance

We can also define $\sigma_h^2(\mathbf{f}, n, s)$ as the function that returns the headway variance $var(h_{n,s})$ given **f**, n and s. We know that the headways $h_{n,s}$ can be also expressed as $h_{n,s} = H + \varepsilon_{n,s} - \varepsilon_{n-1,s}$. Therefore, we see from (10b) that:

$$h_{n,s} = H + \sum_{j=0}^{s-1} \sum_i (f_{i|j} - f_{i-1|j}) \upsilon_{n-i,s-j}. \tag{13a}$$

Since the noise terms are i.i.d., var$(h_{n,s})$ can be expressed as a sum of non-negative terms, and is thus bounded above by the limiting case:

$$\sigma_h^2(\mathbf{f}) \equiv \lim_{\substack{n\to\infty \\ s\to\infty}} \text{var}(h_{n,s}) = \sigma^2 \sum_{j=0}^{\infty} \sum_i (f_{i|j} - f_{i-1|j})^2 \geqslant \sigma_h^2(\mathbf{f}, n, s). \tag{13b}$$

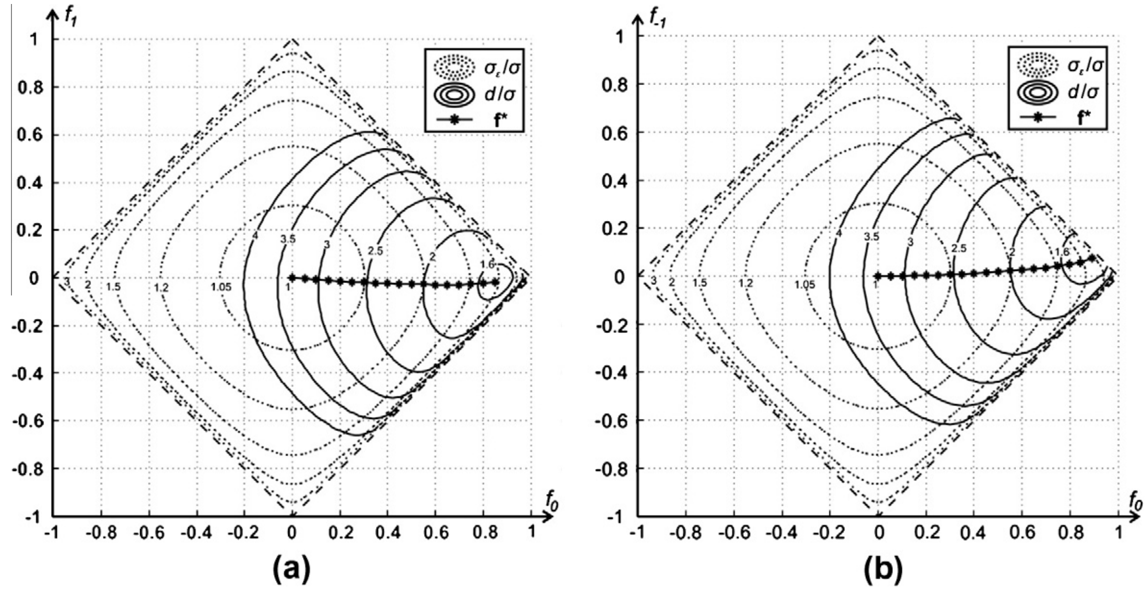

Fig. 1. The iso-$\sigma_\varepsilon/\sigma$ and iso-d/σ contours and the optimal control $\mathbf{f}^*$ with two coefficients, β = 0.1. (a) All $f_i = 0$ except for f_0 and f_1; (b) all $f_i = 0$ except for f_0 and f_{-1}. Dotted lines within the square are the contours with equal $\sigma_\varepsilon/\sigma$-value and solid lines are the contours with equal d/σ-value. The dashed square is the stability region.

Now, it is also possible to demonstrate as a corollary of Theorem 1 that all the control methods which satisfy the conditions in Theorem 1 will also have bounded headway variances.

Corollary 1. *If $\sigma_\varepsilon^2(\mathbf{f})$ is bounded, then $\sigma_h^2(\mathbf{f})$ is also bounded.*

Proof. The headway variance upper bound $\sigma_h^2(\mathbf{f})$ satisfies

$$\begin{aligned}\sigma_h^2(\mathbf{f}) &\equiv \lim_{\substack{n\to\infty\\ s\to\infty}} \mathrm{var}(h_{n,s} - H) = \lim_{\substack{n\to\infty\\ s\to\infty}} \mathrm{var}(\varepsilon_{n,s} - \varepsilon_{n-1,s}) \\ &= \lim_{\substack{n\to\infty\\ s\to\infty}} [\mathrm{var}(\varepsilon_{n,s}) + \mathrm{var}(\varepsilon_{n-1,s}) + 2cov(\varepsilon_{n,s}, \varepsilon_{n-1,s})] \\ &\leqslant 4\sigma_\varepsilon^2(\mathbf{f}).\end{aligned}$$

The last inequality is true because $cov(\varepsilon_{n,s}, \varepsilon_{n-1,s}) \leqslant \sqrt{var(\varepsilon_{n,s})var(\varepsilon_{n-1,s})}$. Thus if $\sigma_\varepsilon^2(\mathbf{f})$ is bounded, $\sigma_h^2(\mathbf{f})$ is bounded. □

In the case of headway-based control methods, both the forward-headway and two-way-headway methods (7) and (8) have been shown to have a bounded headway variance (Daganzo, 2009a,b; Daganzo and Pilachowski, 2009, 2011). Curiously, numerical calculations of (13b) show that the backward-headway method (9) only produces bounded headway variances for low to medium demand levels ($\beta_s < 0.2$). Moreover, when $\beta_s < 0.2$ the control coefficient needs to be carefully chosen ($\alpha \approx 0.5$).

4. Optimal control

It is proposed to choose the control coefficients $\mathbf{f}$ that minimize the slack time d_s required to avoid negative holding times while guaranteeing a maximum standard deviation from the schedule: i.e., a given level of schedule reliability. This proposal is reasonable because slack is inversely related to commercial speed.

To obtain the slack times d_s that avoid negative holding times, we combine Eq. (4b) and (10), so that the holding time is expressed as a function of the control coefficients and noise terms:

$$D_{n,s} = d_s - [(1+\beta_s)\varepsilon_{n,s} - \beta_s\varepsilon_{n-1,s} - \sum_k f_k \varepsilon_{n-k,s}] = d_s - \sum_{j=0}^{s-1}\sum_i [(1+\beta_s)f_{i|j} - \beta_s f_{i-1|j} - f_{i|j+1}]\upsilon_{n-i,s-j}. \tag{14a}$$

Under the assumption of i.i.d. noise, the variance of the holding time $D_{n,s}$ is the sum of many independent random variables, which as before is bounded above by the limiting case; i.e.:

Table 1
Effect of the number of non-zero control coefficients on d^*/σ when $\beta = 0.1$.

d^*/σ	$s_\varepsilon/\sigma = 1$	$s_\varepsilon/\sigma = 1.2$	$s_\varepsilon/\sigma = 1.5$	$s_\varepsilon/\sigma = 2$
$f_i = 0\ \forall i$, except for f_0	3.314	1.989	1.657	1.527
$f_i = 0\ \forall i$, except for f_{-1}, f_0, f_1	3.314	1.978	1.637	1.463
$f_i = 0\ \forall i$, except for $f_{-2}, f_{-1}, f_0, f_1, f_2$	3.314	1.978	1.637	1.463

$$var(D_{n,s}) \leqslant \sigma_D^2(\mathbf{f}, \beta_s) \equiv \lim_{\substack{n \to \infty \\ s \to \infty}} var(D_{n,s}) = \sigma^2 \sum_{j=0}^{\infty} \sum_i [(1+\beta_s) f_{i|j} - \beta_s f_{i-1|j} - f_{i|j+1}]^2. \tag{14b}$$

According to the central limit theorem, $D_{n,s}$ is approximately normal. Therefore, to ensure that the holding time is rarely negative, i.e., $\Pr\{D_{n,s} < 0\} \approx 0$, we shall choose

$$d_s(\mathbf{f}, \beta_s) = 3\sigma_D(\mathbf{f}, \beta_s), \tag{14c}$$

so that the assumption is true 99.87% of the time.

The optimization problem with J control points is then the following mathematical program, where s_ε is a guaranteed standard deviation from the schedule.

$$\text{(MP1)} \quad \min_{\mathbf{f}} \quad \sum_{s=1}^{J} d_s(\mathbf{f}, \beta_s)$$
$$\text{s.t.} \quad \sigma_\varepsilon(\mathbf{f}) \leqslant s_\varepsilon.$$

The functions in (MP1) are given by (11b), (14b) and (14c), and can be calculated numerically. A similar program can be written if the objective is to guarantee a headway variance by using σ_h and (13b), instead of σ_ε and (11b) in the constraint.

4.1. Homogeneous case: sensitivity analysis

Note that in (MP1) the dimensionless demand rates, β_s, at different stations (from 1 to J) can be different; and so can the slacks, d_s. For demonstration purposes, however, it is assumed here that the demand rate is uniform ($\beta_s = \beta$) along the bus line because this is a worst case scenario.[8] A worst case is of interest because if a strategy can prevent buses from bunching in the most challenging situation, it should be able to do the same in other situations.

Now that the demand rate is uniform along the bus line, the slack time will also be the same ($d_1 = d_2 = \ldots$) at all the stations. Thus the subscript s is now dropped, and (MP1) becomes:

$$\text{(MP2)} \quad \min_{\mathbf{f}} \quad d(\mathbf{f}, \beta)$$
$$\text{s.t.} \quad \sigma_\varepsilon(\mathbf{f}) \leqslant s_\varepsilon.$$

Appendix A shows that it is sometimes better to introduce holding times at control points spaced every few stations rather than at every station, and how to choose such spacing.

Fig. 1 shows the contour lines of $\sigma_\varepsilon(\mathbf{f})/\sigma$ and $d(\mathbf{f}, \beta)/\sigma$, when $\beta = 0.1$ for two methods that have two non-zero control coefficients. In Fig. 1a, all $f_i = 0$ except for f_0 and f_1, and in Fig. 1b, only f_0 and f_{-1} are non-zero. The interior of the dashed squares in the figures are the regions where the condition $F = \sum_i |f_i| < 1$ holds. We see that in either case, both $\sigma_\varepsilon(\mathbf{f})$ and $d(\mathbf{f}, \beta)$ are quasi-convex functions of $\mathbf{f}$ within the stability region $\sum_i |f_i| < 1$. Clearly the optimal control coefficient values for any s_ε are at the point where its $\sigma_\varepsilon = s_\varepsilon$ contour is tangent to a d-contour, with the two gradients pointing against each other. This shows that the reliability constraint is always binding. Fig. 1 shows the loci of optimal control coefficients for different $\sigma_\varepsilon/\sigma$ levels by means of dark diamonds.

Since contours are convex, the solutions were obtained with a local gradient search. See Appendix B for the derivation of the gradients. This method works well with up to 11 non-zero control coefficients, which we have tested. Note from Fig. 1 that the schedule-based control method (with $f_i = 0\ \forall i$) is actually among the optimal solutions. Indeed, it provides the best possible schedule reliability (always departing on time), though it requires much slack ($d/\sigma \approx 3.4$).

We also observe that the optimal control coefficients in both cases are very close to the f_0 axis. This indicates that the optimal f_1^* and f_{-1}^* are very small, and that the performance of a control method with a single non-zero coefficient ($f_0 \neq 0$) may be comparable with that with two or more non-zero coefficients. Table 1 confirms this guess. It shows the optimal slack time (in units of σ) for $\beta = 0.1$. Different demand rates yield similar results.

[8] To see this, assume that the total demand is fixed, the trip time is deterministic, and all buses are initially on time. In this case, the motion equation in (3) becomes $\varepsilon_{n,s+1} = (1 + \beta_s)\varepsilon_{n,s}$. Now insert a noise term $\upsilon_{n,1}$ to bus n between stations 0 and 1 so that $\varepsilon_{n,1} = \upsilon_{n,1}$, and see how the noise propagates. It follows from the recursion $\varepsilon_{n,s+1} = (1 + \beta_s)\varepsilon_{n,s}$ that $\varepsilon_{n,s+1} = \upsilon_{n,1}\prod_{i=1}^{s}(1+\beta_i)$. Given that the total demand $\sum_{i=1}^{s}\beta_i$ is fixed, $\varepsilon_{n,s+1}$ is maximized when all β_i's are the same, i.e., when demand is uniform across all stations.

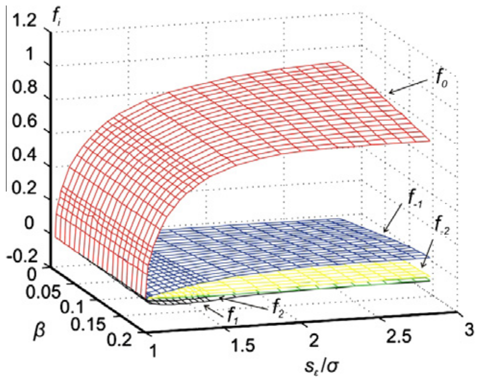

Fig. 2. Optimal values of $f^*_{-2}, f^*_{-1}, f^*_0, f^*_1$, and f^*_2 with different demand rates and schedule reliability values.

Fig. 2 shows the optimal control coefficients $f^*_{-2}, f^*_{-1}, f^*_0, f^*_1$, and f^*_2 for different β and s_ε/σ. Note that the optimal control coefficients are fairly insensitive to changes in the dimensionless demand rate, β. This is good news, because in reality we may not know the demand rate very well. Also note that the values of $f^*_{-2}, f^*_{-1}, f^*_1$ and f^*_2 are roughly negligible (absolute values less than 0.05 in Fig. 2).[9] We did not show the other control coefficients because they are even smaller. So, the performance of a control method with only one non-zero coefficient ($f_0 \neq 0$), which we call the "simple control" method, should be near-optimal. This seems to be the case. Numerical calculations show that for $\beta \in (0.01, 0.1)$ and $s_\varepsilon/\sigma \in (1,2)$, the slack time with the simple control is always within 4% of that of the optimal control obtained with 11 control coefficients.

The strategy is also robust to the selection of the control coefficients. Fig. 3 shows that control coefficients that do not differ much from the optimal control coefficients do not increase the required slack time much. The figure displays the ratio of d/d^* when $\beta = 0.1$ and the general control method has two non-zero coefficients. In this figure, d is the slack time of the non-optimal control coefficients at the indicated point and d^* is the optimal slack time with the same σ_ε value. We see that within a small neighborhood around the optimal coefficients $\mathbf{f}^*$, the ratio is close to 1 but large deviations can result in inefficiency.

The near-optimality of the simple control is nice for both implementation and theoretical analysis. From the implementation point of view, only the arrival times of the current bus and its leading bus, as well as the virtual schedule, are needed to decide the holding time of the current bus at a given station. From an analysis point of view, the simple control method is helpful because formulas simplify and (MP1) can be solved in closed form.

4.2. A simple control law

The control law, bus motion, and metrics of interest are re-derived below, with f_0 being the only decision variable. For the system to be stable: $F = |f_0| < 1$. Note that $f_{i|j} = (f_0)^j \delta(i)$, where $\delta(i)$ is the discrete unit impulse function. In this case, the reader can verify that Eqs. (4b), (4c), (11b), (13b), (14b), and (14c) reduce to:

$$D_{n,s} = d_s - [(1 + \beta_s - f_0)\varepsilon_{n,s} - \beta_s \varepsilon_{n-1,s}], \tag{15a}$$

$$\varepsilon_{n,s+1} = f_0 \varepsilon_{n,s} + \upsilon_{n,s+1}, \tag{15b}$$

$$\sigma^2_\varepsilon(f_0) = \sigma^2/(1 - f_0^2), \tag{15c}$$

$$\sigma^2_h(f_0) = 2\sigma^2/(1 - f_0^2), \tag{15d}$$

$$\sigma^2_D(f_0, \beta) = \frac{\sigma^2[(1 + \beta - f_0)^2 + \beta^2]}{1 - f_0^2}, \tag{15e}$$

and

$$d(f_0, \beta) = 3\sigma_D(f_0, \beta). \tag{15f}$$

These equations only apply if the system is stable (i.e., if $f_0^2 < 1$); then $\sigma^2_\varepsilon(f_0) > \sigma^2$.

The optimal solution of (MP1) is obtained by choosing an $f_0 \in [0, 1)$ that minimizes (15e) such that (15c) is bounded above by s^2_ε, where $s_\varepsilon > \sigma$. Since (15e) declines with f_0 but (15c) increases, it follows that (15c) must be binding. Thus in the optimal solution the actual variance σ^2_ε matches the target s^2_ε. Therefore, the optimal coefficient for the simple control method is:

$$f^*_0 = \sqrt{1 - (\sigma^2/s^2_\varepsilon)}, \text{ where } s^2_\varepsilon > \sigma^2, \tag{15g}$$

[9] With perfect information on future bus arrival times, $|f^*_{-2}|$ and $|f^*_{-1}|$ are already close to zero. In reality, information will be imperfect, leading to even smaller values for $|f^*_{-2}|$ and $|f^*_{-1}|$ (less trustworthy).

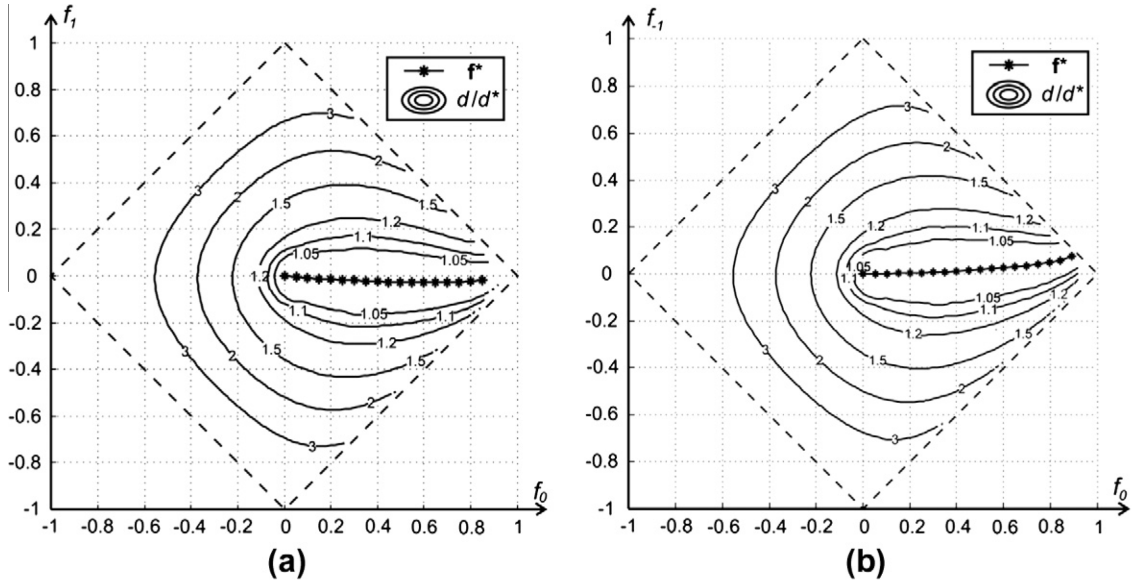

Fig. 3. Sensitivity to control coefficients when $\beta = 0.1$. On both parts of the figure, each point on the plane is associated with a reliability level σ_ε. The contour value at a point is the ratio of the non-optimal slack time d to that point over the optimal slack time d^* with the same σ_ε value of the point. (a) All $f_i = 0$ except for f_0 and f_1; (b) all $f_i = 0$ except for f_0 and f_{-1}.

and the optimal slack is:

$$d^* = 3s_\varepsilon\sqrt{\left(1+\beta-\sqrt{1-(\sigma^2/s_\varepsilon^2)}\right)^2+\beta^2}. \tag{15h}$$

5. Performance analysis

5.1. Comparison with other control methods

Fig. 4 plots expression (15h); it shows how the simple control method performs for different values of β. Note that it requires much less slack time than the schedule-based control method, which is represented by the five points with $s_\varepsilon/\sigma = 1$. Curves like those shown in Fig. 4 cannot be constructed for the headway-based control methods discussed in this paper, because as we have demonstrated, $\sigma_\varepsilon^2(\mathbf{f}) = \infty$ in these cases.

To further compare the simple control method with the headway-based methods, Fig. 5 plots d/σ vs. the dimensionless headway standard deviations σ_h/σ. Fig. 5 shows that the simple control method behaves better than the forward headway method given by (7), the two-way headway method given by (8), and the backward headway method given by (9). Although the two-way method nearly matches the simple control for large σ_h/σ, it underperforms when high reliability (small σ_h/σ) is

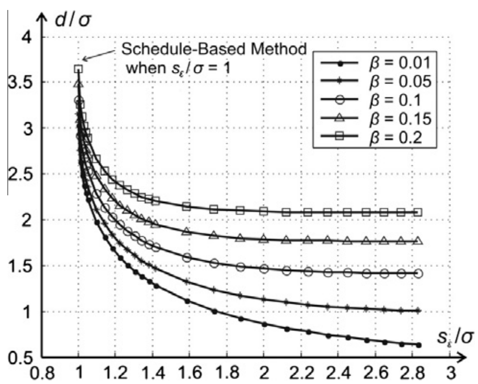

Fig. 4. Slack time d vs. schedule reliability s_ε for the simple control method. Schedule-based control method is represented by the points with $s_\varepsilon/\sigma = 1$.

desired. In all three cases, the reduction in slack time is considerable when high reliability is desired. In fact, none of the headway-based methods can achieve σ_h/σ below 1.5 for any slack whatsoever, while the simple control method can. These results should not be surprising given that headway-based methods are just special cases of the general control method, and that the simple method is near-optimal.

5.2. Balancing headway regularity and slack time

Previously, we have chosen to minimize slack time (maximize commercial speed) subject to a schedule reliability constraint (or similarly a headway regularity constraint). This was appealing because it did not require knowledge of the passenger origin-destination table. Here we assume that the average user trip length l is known and show how to balance the two metrics by minimizing the sum of the average passenger waiting and riding times. We will focus on bus lines operating with short headways and compare the five control methods (schedule-based, forward headway, backward headway, two-way headway and the simple method). For bus lines with long headways, headway-based control methods are not applicable, and the performance comparison of the schedule-based and the simple control method would favor the latter even more.

It is assumed that: (i) the average bus cruising speed is v_c (i.e., including acceleration and deceleration due to the stops); (ii) station spacing S is uniform; (iii) demand β is uniform; (iv) passengers value their waiting time γ times ($\gamma > 1$) as much as their riding time. The passenger waiting time is: $H/2 + \sigma_h^2/(2H)$. The passenger riding time is $l/v_c + (\beta H + d)l/S$. The weighted sum of the two can be expressed as the sum of a fixed component T_0 and a variable component ΔT that depends on the control method:

$$T_0 = \gamma H/2 + l/v_c + \beta Hl/S, \tag{16a}$$

$$\Delta T = \gamma\sigma_h^2/(2H) + dl/S. \tag{16b}$$

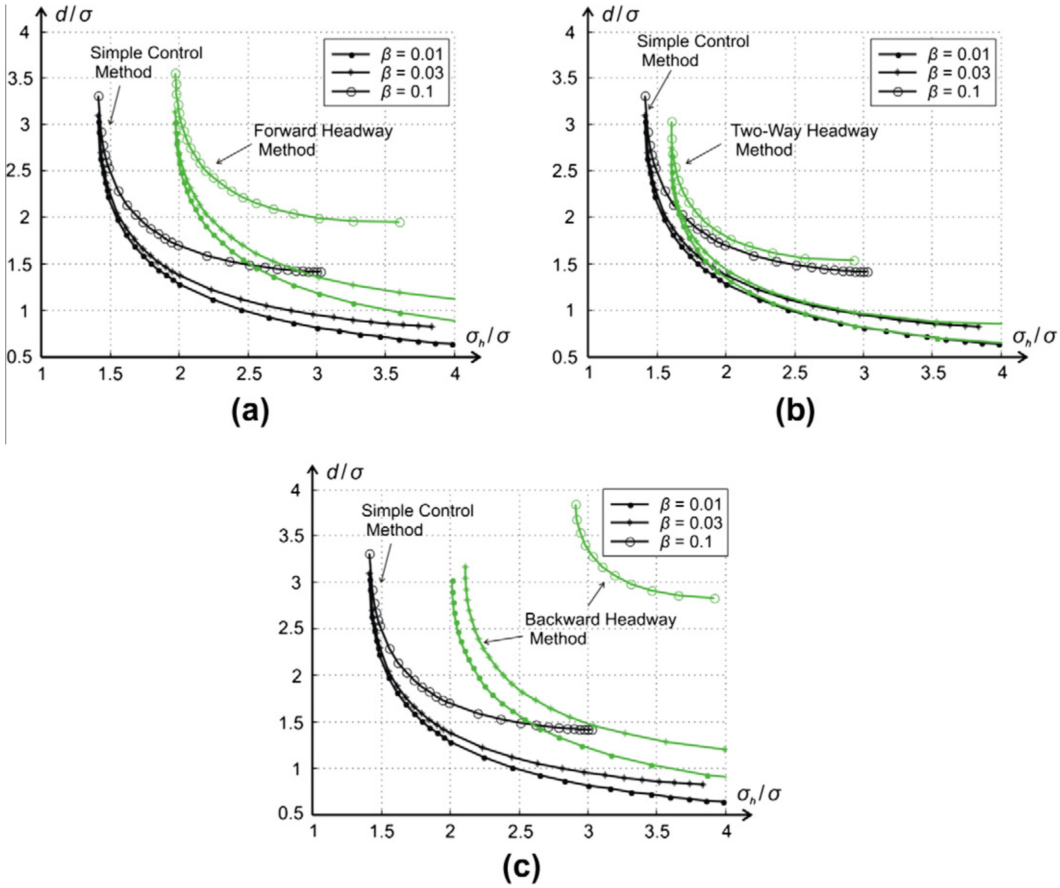

Fig. 5. Slack time d vs. headway reliability σ_h to compare the simple control method with headway-based control methods relying on: (a) forward headway only; (b) forward and backward headways; and (c) backward headway only.

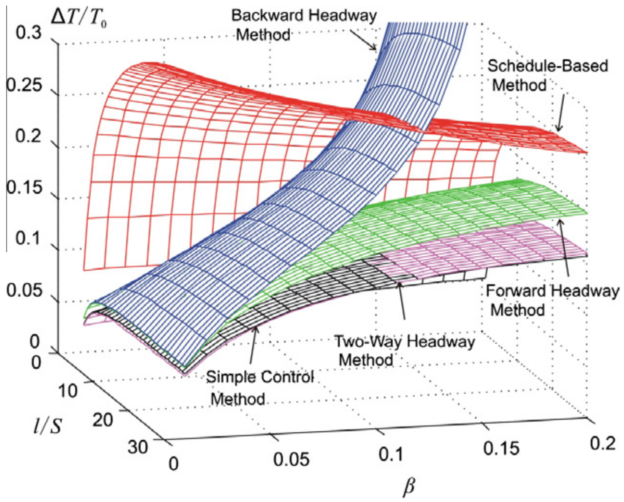

Fig. 6. Ratio of variable trip time over fixed trip time ($\Delta T/T_0$) for different control methods with various demand rates β and dimensionless trip lengths l/S, with control points located at each station.

Note T_0 includes waiting, line-haul riding and loading time under ideal conditions. It is the minimal trip time, and is achievable only if the bus system had no disturbances. The added term ΔT includes the extra time penalty passengers suffer due to non-uniform headways (σ_h^2) and to the slack time (d). The variables σ_h^2 and d can be calculated as a function of the control coefficients with Eqs. (13b), (14b) and (14c) for the four considered methods. For each method the control coefficients that minimize (16b) are obtained.

Fig. 6 shows the ratio $\Delta T/T_0$ for the control methods, as a function of the demand rates β and the dimensionless trip lengths l/S, when the control is applied at all stations. The following parameter values were used: $S = 400$ m, $H = 5$ min, $v_c = 20$ km/h, $\gamma = 2$ and $\sigma = 10$ s. For low demand rates ($0.01 < \beta < 0.05$), all the headway-based methods and the simple method perform much better than the schedule-based method. The $\Delta T/T_0$ ratios in this situation are 4–11% for the backward headway method, 5–10% for the forward headway method and only 3–7% for the two-way headway method and the simple method. As demand increases, the backward headway method becomes unstable and the simple method also outperforms the rest of methods with a maximal $\Delta T/T_0$ ratio of 15% at $\beta = 0.2$ and $l/S = 25$. Note that the two-way headway method and the simple method are practically indistinguishable. Finally, note that the results of the two-way headway method and the backward headway method are optimistic, because in reality the backward headway will not be readily available.

Fig. 7 shows the same information as Fig. 6, assuming that the control points are also optimally located (every N^* stations) as per Appendix A. The value of N^* is different for different methods and different combinations of β and l/S. In the minimi-

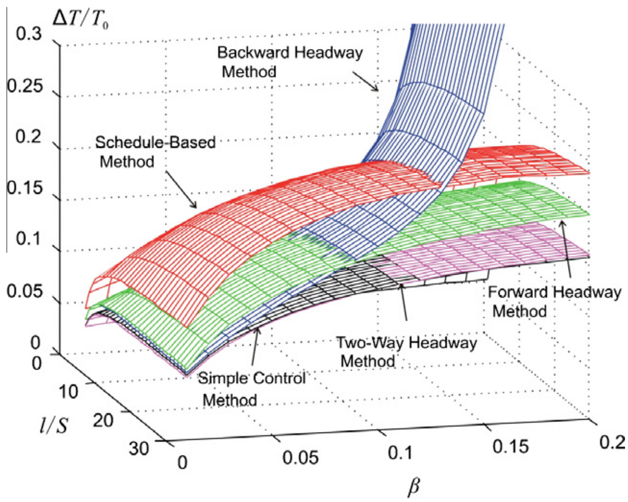

Fig. 7. Ratio of variable trip time over fixed trip time ($\Delta T/T_0$) for different control methods with various demand rates β and dimensionless trip lengths l/S, with control points located optimally.

zation of (16b), the headway variance σ_h^2 is now replaced by its average across all stations $\langle\sigma_h^2\rangle$, calculated as explained in the last paragraph of Appendix A. By allowing values $N > 1$, the slack time needed for the schedule-based method is greatly reduced for low demand as shown in the figure, because in this case $N^* \gg 1$. The backward headway method also saves a little slack time because in this case $N^* \approx 2$. The results for the simple control method, the forward headway method and the two-way headway method remain the same as in Fig. 6, because $N^* = 1$ for these methods.

Note that the simple method reduces slack time by nearly 40% compared with the schedule-based method. The two-way headway method behaves similarly. For moderately heavy demand ($\beta = 0.1$), these two methods reduce the slack time of the schedule-based method by more than 50%, and reduce the commercial speed compared with the ideal by only about 10%. The two-way headway method, however, cannot help the buses keep to a schedule.

6. Conclusion

This paper proposes a family of dynamic holding strategies that can improve bus schedule reliability while maximizing buses' commercial speed. The proposed method allows buses not just to maintain regular headways but also to adhere to their schedule. It is shown that none of the existing dynamic methods in the control genre, which are all headway-based, can maintain a schedule. The proposed method only requires readily available information from the current bus and its leading bus. It is able to maintain quasi-regular headways with a higher commercial speed than other existing methods.

Acknowledgements

This work is jointly supported by the National Science Foundation under Grant No. 0856193, the Volvo Center of Excellence for Future Urban Transport at the University of California, Berkeley, and the ''la Caixa'' Graduate Fellowship Program. Comments from Yanfeng Quyang (University of Illinois at Urbana-Champaign) during the early stage of the work are greatly appreciated.

Appendix A. A method for locating control points

We have assumed during the analysis that the proposed control method is applied at each station, but this is not always desirable. As shown in Daganzo (1997b, 2009a), it is often beneficial to space out the control points more widely. In this spirit, it is assumed here that control points are located every N stations, and N is treated as a decision variable. The demand rate β is assumed to be uniform throughout the bus line and such that $\beta \ll 1$.

We will first transform the equations of bus motion from station to station into similar equations describing the bus motion from control point to control point, as if there were no intermediate stations. It will be shown that when $\beta \ll 1$, one may simply replace β and σ with $\beta' = N\beta$ and $\sigma' = \sigma\sqrt{N + N(N-1)\beta}$ to model the bus motion in this manner.

The variance of the noise between control points $(\sigma')^2$ is simply given by (11a) using the $\{f_{i|j}\}$ of the uncontrolled case and replacing s by N. Recall that for the uncontrolled situation, $\mathbf{f} = [\ldots, 1+\beta, -\beta, \ldots]^T$. It can be seen from the binomial expansion that

$$f_{m|j} = \binom{j}{m}(1+\beta)^m(-\beta)^{j-m}. \tag{A.1}$$

Thus,

$$\mathbf{f}_{|j} = \left[\ldots, (1+\beta)^j, j(1+\beta)^{j-1}(-\beta), \binom{j}{2}(1+\beta)^{j-2}(-\beta)^2, \ldots\right]^T \approx [\ldots, 1+j\beta, -j\beta, 0, \ldots]^T. \tag{A.2}$$

The last approximation works because $\beta \ll 1$ and thus we can neglect terms of order β^2 and higher. Eq. (A2) can now be inserted in (11a) to yield:

$$\begin{aligned}(\sigma')^2 &= \sigma^2 \sum_{j=0}^{N-1} \sum_m (f_{m|j})^2 \\ &= \sigma^2 \sum_{j=0}^{N-1} \sum_m \left[\binom{j}{m}(1+\beta)^m(-\beta)^{j-m}\right]^2 \\ &\approx \sigma^2 \sum_{j=0}^{N-1}(1+2j\beta) \\ &= \sigma^2(N + N(N-1)\beta).\end{aligned} \tag{A.3}$$

It should also be clear from (A2) that when $j = N$ and $\beta \ll 1$, the dimensionless demand between control points is $\beta' = N\beta$.

By setting $\beta' = N\beta$ and $\sigma' = \sigma\sqrt{N + N(N-1)\beta}$, we can treat the bus motion as if there were only stations at the control points and apply (MP2) with these new parameters. The number of stations between control points, N, is however, a decision variable in the new version of (MP2). Consideration of (11b) and (14b) reveals that this new mathematical program is:

$$\text{(MP4)} \quad \min_{\mathbf{f},N} \quad \tfrac{\sigma'}{\sigma} \tfrac{d(\mathbf{f},N\beta)}{N}$$
$$\text{s.t.} \ \tfrac{\sigma'}{\sigma} \sigma_\varepsilon(\mathbf{f}) \leqslant s_\varepsilon$$
$$\tfrac{\sigma\prime}{\sigma} = \sqrt{N + N(N-1)\beta}.$$

This MP can be solved numerically as in Section 4.1, or analytically if we adopt the simple control method. A similar program can be written if the objective is to guarantee a headway variance by using σ_h instead of σ_ε.

In some instances, it may be useful to have an idea of the average of the variances of the schedule deviations across all the stations, $\langle\sigma_\varepsilon^2\rangle$; or the average of the headway variances, $\langle\sigma_h^2\rangle$. To obtain these values, (11b) and (13b) would still be used, but the recursion $\mathbf{f}_{|j} = \mathbf{f} * \mathbf{f}_{|j-1}$ used to evaluate the convolution coefficients, $f_{i|j}$, would use the f_i from (4c) when "j" corresponds to a control point, and $f_0 = 1 + \beta_s$, $f_1 = -\beta_s$ and $f_i = 0\ \forall i \notin \{0,1\}$ from (5) otherwise.

Appendix B. Derivation of the gradients for greedy search

In developing Fig. 1, we solved the following optimization problem, which is equivalent to (MP2):

$$\text{(MP5)} \quad \min_{\mathbf{f}} \ \sigma_D^2(\mathbf{f},\beta)$$
$$\text{s.t.} \ \sigma_\varepsilon^2(\mathbf{f}) \leqslant s_\varepsilon^2.$$

From Eqs. (11b) and (14b), we find the following expressions for $\sigma_\varepsilon^2(\mathbf{f})$ and $\sigma_D^2(\mathbf{f}, \beta)$, whose gradients we seek:

$$\sigma_\varepsilon^2(\mathbf{f}) = \sigma^2 \sum_{j=0}^{\infty} \sum_i (f_{i|j})^2,$$

$$\sigma_D^2(\mathbf{f},\beta) = \sigma^2 \sum_{j=0}^{\infty} \sum_i [(1+\beta)f_{i|j} - \beta f_{i-1|j} - f_{i|j+1}]^2.$$

Using generating functions, one can find the following expression for $f_{i|j}$:

$$f_{i|j} = \sum_{\substack{k_0,k_1,\cdots k_{m-1} \\ \sum_{r=0}^{m-1} k_r = j \\ \sum_{r=0}^{m-1} r k_r = i}} \frac{j!}{k_0! k_1! \cdots k_{m-1}!} f_0^{k_0} f_1^{k_1} \cdots f_{m-1}^{k_{m-1}}. \tag{B.1}$$

Since

$$\partial f_{i|j} / \partial f_r = j f_{i-r|j-1}, \tag{B.2}$$

The partial derivatives of $\sigma_\varepsilon^2(\mathbf{f})$ and σ_D^2 ($\mathbf{f}$, β) with respect to f_r can be expressed as:

$$\frac{\partial \sigma_\varepsilon^2}{\partial f_r} = 2\sigma^2 \sum_{j=0}^{\infty} \sum_i (f_{i|j})(j f_{i-r|j-1}), \tag{B.3}$$

and

$$\frac{\partial \sigma_D^2}{\partial f_r} = 2\sigma^2 \sum_{j=0}^{\infty} \sum_i [(1+\beta)f_{i|j} - \beta f_{i-1|j} - f_{i|j+1}][(1+\beta) j f_{i-r|j-1} - \beta j f_{i-r-1|j-1} - (j+1) f_{i-r|j}]. \tag{B.4}$$

References

Abkowitz, M., Eiger, A., Engelstein, I., 1986. Optimal control of headway variation on transit routes. Journal of Advanced Transportation 20 (1), 73–88.

Adamski, A., Turnau, A., 1998. Simulation support tool for real-time dispatching control in public transport. Transportation Research Part A: Policy and Practice 32 (2), 73–87.

Barnett, A., 1974. On controlling randomness in transit operations. Transportation Science 8 (2), 102–116.

Bartholdi, J., 2011. (private communication).

Bowman, L.A., Turnquist, M.A., 1981. Service frequency, schedule reliability and passengers wait times at transit stops. Transportation Research Part A: General 15 (6), 465–471.

Daganzo, C.F., 1997a. Passenger waiting time: advertised schedules. In: Daganzo, C.F. (Ed.), Fundamentals of Transportation and Traffic Operations. Elsevier, New York, NY, pp. 291–292.

Daganzo, C.F., 1997b. Schedule instability and control. In: Daganzo, C.F. (Ed.), Fundamentals of Transportation and Traffic Operations. Elsevier, New York, NY, pp. 304–309.

Daganzo, C.F., 2006. On the variational theory of the traffic flow: well-posedness, duality and applications. Networks and Heterogeneous Media 1 (4), 601–619.

Daganzo, C.F., 2009a. A headway-based approach to eliminate bus bunching: systematic analysis and comparisons. Transportation Research Part B: Methodological 43 (10), 913–921.

Daganzo, C.F., 2009b. A cheap and resilient way to eliminate bus bunching. In: The 4th International Conference on Future Urban Transport, Gothenburg, Sweden.

Daganzo, C.F., Pilachowski, J., 2009. Reducing bunching with bus-to-bus cooperation. Working Paper UCB-ITS-VWP-2009-12, U.C. Berkeley Center for Future Urban Transport.
Daganzo, C.F., Pilachowski, J., 2011. Reducing bunching with bus-to-bus cooperation. Transportation Research Part B: Methodological 45 (1), 267–277.
Delgado, F., Muñoz, J.C., Giesen, R., Cipriano, A., 2009. Real-time control of buses in a transit corridor based on vehicle holding and boarding limits. Transportation Research Record 2090, 59–67.
Eberlein, X.J., Wilson, N.H.M., Bernstein, D., 2001. The holding problem with real-time information available. Transportation Science 35 (1), 1–18.
Fu, L., Liu, Q., Calamai, P., 2003. Real-time optimization model for dynamic scheduling of transit operations. Transportation Research Record 1857, 48–55.
Golob, T.F., Canty, E.T., Gustafson, R.L., Vitt, J.E., 1972. An analysis of consumer preferences for a public transportation system. Transportation Research 6 (1), 81–102.
Hickman, M.D., 2001. An analytic stochastic model for the transit vehicle holding problem. Transportation Science 35 (3), 215–237.
Koffman, D., 1978. A simulation study of alternative real-time bus headway control strategies. Transportation Research Record 663, 41–46.
Newell, G.F., Potts, R.B., 1964. Maintaining a bus schedule. Proceedings of the 2nd Australian Road Research Board 2, 388–393.
Newell, G.F., 1974. Control of pairing of vehicles on a public transportation route, two vehicles, one control point. Transportation Science 8 (3), 248–264.
Osuna, E.E., Newell, G.F., 1972. Control strategies for an idealized public transportation system. Transportation Science 6 (1), 52–72.
Paine, F.T., Nash, A.N., Hille, S.J., Brunner, G.A., 1967. Consumer Conceived Attributes of Transportation: An Attitude Study. University of Maryland, College Park.
Rockafellar, R.T., 1970. Section 32: The Maximum of a Convex Function. In: Rockafellar, R.T. (Ed.), Convex Analysis. Princeton University Press.
Senevirante, P.N., 1990. Analysis of on-time performance of bus services using simulation. Journal of Transportation Engineering 116 (4), 517–531.
Suh, W., Chon, K., Rhee, S., 2002. Effect of skip-stop policy on a korean subway system. Transportation Research Record 1793, 33–39.
Sun, A., Hickman, M., 2005. The Real-Time Stop-Skipping problem. Journal of Intelligent Transportation Systems 9 (2), 91–109.
Turnquist, M.A., Bowman, L.A., 1980. The effects of network structure on reliability of transit service. Transportation Research Part B: Methodological 14 (1-2), 79–86.
Vandebona, U., Richardson, A.J., 1986. Effect of checkpoint control strategies in a simulated transit operation. Transportation Research Part A: General 20 (6), 429–436.
Wallin, R.J., Wright, P.H., 1974. Factors which influence modal choice. Traffic Quarterly 28 (2), 271–289.
Zhao, J., Dessouky, M., Bukkapatnam, S., 2006. Optimal slack time for schedule-based transit operations. Transportation Science 40 (4), 529–539.