Title

A Dynamic Holding Strategy to Improve Bus Schedule Reliability and Commercial Speed

Authors

Xuan, Yiguang
Argote, Juan
Daganzo, Carlos F.

Publication Date

2011-03-01

INSTITUTE OF TRANSPORTATION STUDIES
UNIVERSITY OF CALIFORNIA, BERKELEY

A Dynamic Holding Strategy to Improve Bus Schedule Reliability and Commercial Speed

Yiguang Xuan, Juan Argote, and Carlos F. Daganzo

WORKING PAPER
UCB-ITS-VWP-2011-1

UC Berkeley Center for Future Urban Transport
A VOLVO Center of Excellence

March 2011

A Dynamic Holding Strategy to Improve Bus Schedule Reliability and Commercial Speed

Yiguang Xuan, Juan Argote, and Carlos F. Daganzo
Institute of Transportation Studies
University of California, Berkeley, CA 94720

(March 18, 2011)

Abstract

Bus systems are naturally unstable. Without control, the slightest disturbance to bus motion can cause buses to bunch, reducing schedule reliability. Holding strategies can eliminate this instability. However, the conventional schedule-based holding method requires too much slack time, which slows buses. This delays on-board passengers and increases operating costs. This paper studies a family of dynamic holding strategies that use the current state of all buses, as well as a virtual schedule. The virtual schedule is introduced whether the system is run with a published schedule or not. We found that with this control method, which we term general control method, buses can both closely adhere to schedule and maintain regular headways without too much slack. Thus the general control idea is applicable to bus lines with both long and short headways. Although the optimal set of control parameters can be found numerically, a one-parameter version of the control method can be optimized in closed form. This simple method was shown to be near-optimal. To put it in practice, one only needs the arrival times of the current bus and the preceding bus relative to the virtual schedule. This simple method was found to outperform alternative control methods (i.e., require less slack for the same headway variance). While the paper mostly focuses on recurrent small disturbances under quasi-deterministic demand, it also shows that the proposed control method can deal with large disturbances.

Keywords: bus bunching, bus schedule reliability, headway variance, commercial speed, dynamic holding, control method.

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 et al., 1974). Unfortunately, bus systems are naturally unstable, and buses tend to fall off schedule without intervention. This instability 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 by Newell & Potts (1964), early buses 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 tendencies result in the so-called

[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.

"bus bunching" phenomenon, which annuls the schedule and increases the average waiting time of passengers.

Bus holding strategies may eliminate this problem. They are characterized by embedding slack time in the schedule, and holding buses at each control station for a period of time before their scheduled departure. A bus is generally held longer if it is ahead of schedule and shorter (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 station 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 station, and either one or two buses.[2] Some use "rolling horizon" heuristics. Many more studies resort to simulations (Koffman, 1978; Turnquist & Bowman, 1980; Abkowitz et al., 1986; Vandebona & Richardson, 1986; Senevirante, 1990; Adamski & Turnau, 1998) due to the difficult nature of the problem.

There is also a literature that uses control theory. Daganzo (2009b) 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 much broader range of problems: systems with many buses, many control stations, 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 breakdown.

To alleviate this problem, Daganzo and Pilachowski (2009, 2011) and Daganzo (2009a) 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.[4] More recently, a holding method in

[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] The Eulerian version of the method in Daganzo and Pilachowski (2009, 2011) is used here, in which holding times would be based on the forward and backward headways (the headway between the current bus and the bus behind).

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 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 (2009b). 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 can not 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. While this paper focuses on small disturbances to the bus operation, the case of large disturbances is also addressed.

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. Section 4 demonstrates that a simple version of the general control method is near-optimal, and that it outperforms the other existing methods. Sections 5 and 6 discuss how to handle large disturbances and highly stochastic demand. Finally Section 7 summarizes the main findings.

2 Assumptions and Bus Motion Formulation

2.1 Assumptions

The same assumptions are made as in Daganzo (2009b): (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) vehicle running times between adjacent stations are modeled as independent random variables; (f) demand is stochastic but its contribution to the variance of the running time between stations is assumed to be negligible compared with the variance due to traffic (we call this quasi-deterministic demand); (g) the bus loading time is dominantly affected by boarding passengers; (h) only those passengers 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 (f) is relaxed in Section 6 to account for highly stochastic demand. Assumption (g) 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 (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, …) to denote the bus number (the buses dispatched first have smaller numbers) and s (s = 0, 1, 2 …) to denote the station number (increasing s in the traveling direction). The notation follows Daganzo (2009b):

- $t_{n,s}$ 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.
- $h_{n,s} = a_{n,s} - a_{n\text{-}1,s}$ 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 cruising time of bus n between station s to s+1, whose mean is zero and variance is $\sigma_{n,s+1}^2$.
- $D_{n,s}$ 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:

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

$$t_{n,s} = t_{n-1,s} + H, \tag{1b}$$

[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.

with $t_{n,s} + \beta_s H + d_s$ being the scheduled departure times. 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 equations (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 $\varepsilon_{i,s}$.[6] It is convenient to write this function as:

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

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

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

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

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

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 station 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 station 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

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

and

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

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

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

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

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

[6] The arrival times of the current bus and the buses in front, $a_{i,s}$'s for $i \le n$, 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 that this assumption is acceptable.

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

$$\varepsilon_{n,s+1} = (1-\alpha)\varepsilon_{n,s} + \alpha\varepsilon_{n-1,s} + \upsilon_{n,s+1}. \qquad \begin{matrix}(\textit{forward}\\ \textit{headway})\end{matrix} \qquad (7c)$$

Similarly, the Eulerian version[7] of the (Lagrangian) method in Daganzo and Pilachowski (2009, 2011) and Daganzo (2009a), 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:

$$D_{n,s} = d_s + \alpha\left(h_{n+1,s} - H\right) - \left(\alpha + \beta_s\right)\left(h_{n,s} - H\right), \qquad \begin{matrix}(\textit{two-way}\\ \textit{headway})\end{matrix} \qquad (8a)$$

$$D_{n,s} = d_s + \alpha\varepsilon_{n+1,s} - \left(\beta_s + 2\alpha\right)\varepsilon_{n,s} + \left(\beta_s + \alpha\right)\varepsilon_{n-1,s}, \qquad \begin{matrix}(\textit{two-way}\\ \textit{headway})\end{matrix} \qquad (8b)$$

and

$$\varepsilon_{n+1,s} = \alpha\varepsilon_{n+1,s} + \left(1 - 2\alpha\right)\varepsilon_{n,s} + \alpha\varepsilon_{n-1,s} + \upsilon_{n,s+1}. \qquad \begin{matrix}(\textit{two-way}\\ \textit{headway})\end{matrix} \qquad (8c)$$

Finally, a demand-independent method based on the backward headway alone (Bartholdi, 2011) is obtained by setting $f_{-1} = \alpha, f_0 = 1 + \beta_s - \alpha, f_1 = -\beta_s$ and $f_i = 0\ \forall i \notin \{-1, 0, 1\}$. This results in:

$$D_{n,s} = \alpha h_{n+1,s} = \alpha H + \alpha\left(h_{n+1,s} - H\right), \qquad \begin{matrix}(\textit{backward}\\ \textit{headway})\end{matrix} \qquad (9a)$$

$$D_{n,s} = \alpha\left(H + \varepsilon_{n+1,s} - \varepsilon_{n,s}\right) = d_s - \alpha\varepsilon_{n,s} + \alpha\varepsilon_{n+1,s}, \qquad \begin{matrix}(\textit{backward}\\ \textit{headway})\end{matrix} \qquad (9b)$$

and

$$\varepsilon_{n,s+1} = \alpha\varepsilon_{n+1,s} + \left(1 + \beta_s - \alpha\right)\varepsilon_{n,s} - \beta_s\varepsilon_{n-1,s} + \upsilon_{n,s+1}. \qquad \begin{matrix}(\textit{backward}\\ \textit{headway})\end{matrix} \qquad (9c)$$

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

[7] The control method introduced in Daganzo and Pilachowski (2009, 2011) and Daganzo (2009a) is consistent with a Lagrangian specification of the buses system, while the proposed general control in this paper is based on an Eulerian specification. Daganzo (2006) demonstrates the close connection between Lagrangian and Eulerian coordinates in the context of automobile motion.

$\upsilon_{n\text{-}1,s}\quad \upsilon_{n,s}\quad \upsilon_{n+1,s}\ \ldots]^{\mathrm{T}}$ as the vector of disturbances, then the vector form of (4b) 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 j$^{\text{th}}$ self-convolution of $\mathbf{f}$ (i.e., $\mathbf{f}_{|j} = \mathbf{f} * \mathbf{f}_{|j-1}$, where $\mathbf{f}_{|0}$ = [… 0 1 0 …]$^{\mathrm{T}}$). 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

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

It is now possible to see that the following is true.

Lemma: Define $F = \sum_i |f_i|$, then $\sum_i |f_{i|j}| \le F^j$.

Proof:

$$\sum_i |f_{i|j}| = \sum_i \left|\sum_k f_{k|j-1} f_{i-k}\right| \le \sum_i \sum_k |f_{k|j-1}||f_{i-k}| = \sum_k \left(|f_{k|j-1}| \sum_i |f_{i-k}|\right) = F \sum_k |f_{k|j-1}|$$
$$\le \ldots \le F^2 \sum_k |f_{k|j-2}| \le \ldots \le F^j. \qquad \square$$

Proposition: If $F < 1$ and $|\upsilon_{n,s+1}| \le M$, then $|\varepsilon_{n,s+1}|$ is bounded above by $M/(1-F)$

Proof:

$$|\varepsilon_{n,s+1}| = \left|\sum_{j=0}^{s} \sum_i f_{i|j}\, \upsilon_{n-i,s+1-j}\right| \le \sum_{j=0}^{s} \sum_i |f_{i|j}||\upsilon_{n-i,s+1-j}|$$
$$\le M \sum_{j=0}^{s} \sum_i |f_{i|j}| \qquad (\text{since } |\upsilon_{n,s+1}| \le M)$$
$$\le M \sum_{j=0}^{s} F^j \qquad (\text{as per Lemma})$$
$$= M \frac{1-F^{s+1}}{1-F} \to \frac{M}{1-F} \qquad (\text{since } F < 1). \qquad \square$$

Corollary: If $|\varepsilon_{n,s+1}|$ is bounded, then $|h_{n,s+1} - H|$ is also bounded.

Proof: $|h_{n,s+1} - H| = |\varepsilon_{n,s+1} - \varepsilon_{n-1,s+1}| \le |\varepsilon_{n,s+1}| + |\varepsilon_{n-1,s+1}|$. $\square$

Any method with $F < 1$ will exhibit bounded deviations from a schedule, $\varepsilon_{n,s+1}$, and from the average headway, $h_{n,s+1} - H$. 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).

Unfortunately, none of the headway-based methods, (7), (8) and (9), satisfy the condition $F < 1$. It will be shown in Section 4 that the variances of their deviations from a schedule are unbounded.

4 Optimal control

It is proposed to choose the control coefficients **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.

Let us define $\sigma_\varepsilon^{\,2}(\mathbf{f},n,s)$ as the function that returns $\mathrm{var}(\varepsilon_{n,s})$ given **f**, n and s. From equation (10), we have $\varepsilon_{n,s} = \sum_{j=0}^{s-1}\sum_i f_{i|j}\upsilon_{n-i,s-j}$. 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:

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

which will be our measure of schedule reliability.

Theorem: If $\sum_{i=-n}^{n} f_i = 1$ and $f_i = 0$ for all other i then the right-hand side of equation (11b) is unbounded.

Proof:

Let us first show that if $\sum_i f_i = 1$ then

$$\sum_i f_{i|j} = 1, \quad \forall j. \tag{12a}$$

If we assume that $\sum_i f_i = 1$, the following is true:

$$\sum_i f_{i|j} = \sum_i \sum_k f_{i-k} f_{k|j-1} = \sum_k f_{k|j-1} \sum_i f_{i-k} = \sum_k f_{k|j-1} = \sum_i f_{i|j-1}, \quad \forall j > 0.$$

The first equality is the definition of convolution; the second is obtained by interchanging the order of the summation; the third follows from the assumption; and the fourth by relabeling the dummy subscript index. When $j = 1$, $\sum_i f_{i|j} \equiv \sum_i f_i$ by definition. Therefore $\sum_i f_{i|j} = 1$ if $j = 1$. Thus by induction, $\sum_i f_{i|j} = 1, \forall j$.

Using this result, it is possible to show that $\sum_{j=0}^{\infty} \sum_i (f_{i|j})^2$ is unbounded. Notice that the index of the nonzero terms of the n^{th} convolution must be in the interval $[-nj, nj]$. Thus $\sum_i \left(f_{i|j}\right)^2 = \sum_{i=-nj}^{nj} \left(f_{i|j}\right)^2$, which consists of $2nj + 1$ terms. A lower bound to $\sum_{i=-nj}^{nj} \left(f_{i|j}\right)^2$ is obtained by choosing the $f_{i|j}$ values that minimize $\sum_{i=-nj}^{nj} \left(f_{i|j}\right)^2$, subject to $\sum_{i=-nj}^{nj} f_{i|j} = 1$ as per (12a). The minimum arises when all the terms in these summations are equal, i.e., when $f_{i|j} = 1/(2nj+1), \forall -nj \le i \le nj$.

$$\sum_i \left(f_{i|j}\right)^2 = \sum_{i=-nj}^{nj} \left(f_{i|j}\right)^2 \ge \frac{2nj+1}{(2nj+1)^2} = \frac{1}{2nj+1}. \tag{12b}$$

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 (11b). □

Note that all the headway-based control methods discussed in this paper, (7), (8) and (9), satisfy the conditions of this theorem. Therefore as mentioned in Section 3, $\sigma_\varepsilon^2(\mathbf{f}) = \infty$ and the methods cannot maintain a schedule.

We also define $\sigma_h^{\,2}(\mathbf{f}, n, s)$ as the function that returns the headway variance $\text{var}(h_{n,s})$ given $\mathbf{f}$, n and s. Because the headway $h_{n,s}$ can be expressed as $h_{n,s} = H + \varepsilon_{n,s} - \varepsilon_{n-1,s}$, we have

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

The headway variance again can be expressed as a sum of non-negative terms, and is thus bounded above by the limiting case:

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

Methods (7) and (8) have been shown to have a bounded headway variance (Daganzo, 2009a, 2009b; Daganzo and Pilachowski, 2009, 2011). Numerical calculations of (13b) show that the demand-independent method (9) only produces bounded headway variances for low to medium demand levels, if the control coefficient is carefully chosen ($\alpha \approx 0.5$). However, these high values of α result in long slack times, and therefore low commercial speeds.[8]

[8] Refer to (9a) and note that the slack time is αH.

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

$$\begin{aligned} D_{n,s} &= d_s - [(1+\beta_s)\varepsilon_{n,s} - \beta_s \varepsilon_{n-1,s} - \textstyle\sum_k f_k \varepsilon_{n-k,s}] \\ &= d_s - \textstyle\sum_{j=0}^{s-1} \sum_i \left[(1+\beta_s) f_{i|j} - \beta_s f_{i-1|j} - f_{i|j+1}\right] v_{n-i,s-j}. \end{aligned} \tag{14a}$$

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.:

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

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 stations is then the following mathematical program, where s_ε is the guaranteed standard deviation from the schedule.

$$\begin{aligned} \text{(MP1)} \quad \min_{\mathbf{f}} \quad & \textstyle\sum_{s=1}^{J} d_s(\mathbf{f}, \beta_s) \\ \text{s.t.} \quad & \sigma_\varepsilon(\mathbf{f}) \le s_\varepsilon. \end{aligned}$$

The functions in (MP1) are given by (11b), (14b) and (14c), and can be calculated numerically.

4.1 Homogeneous Case

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. But for demonstration purposes, it is assumed here that the demand rate is uniform ($\beta_s = \beta$) along the bus line. Now the slack time will also be the same ($d_1 = d_2 = ...$) at all the stations. Thus the subscript s is now dropped, and (MP1) becomes:

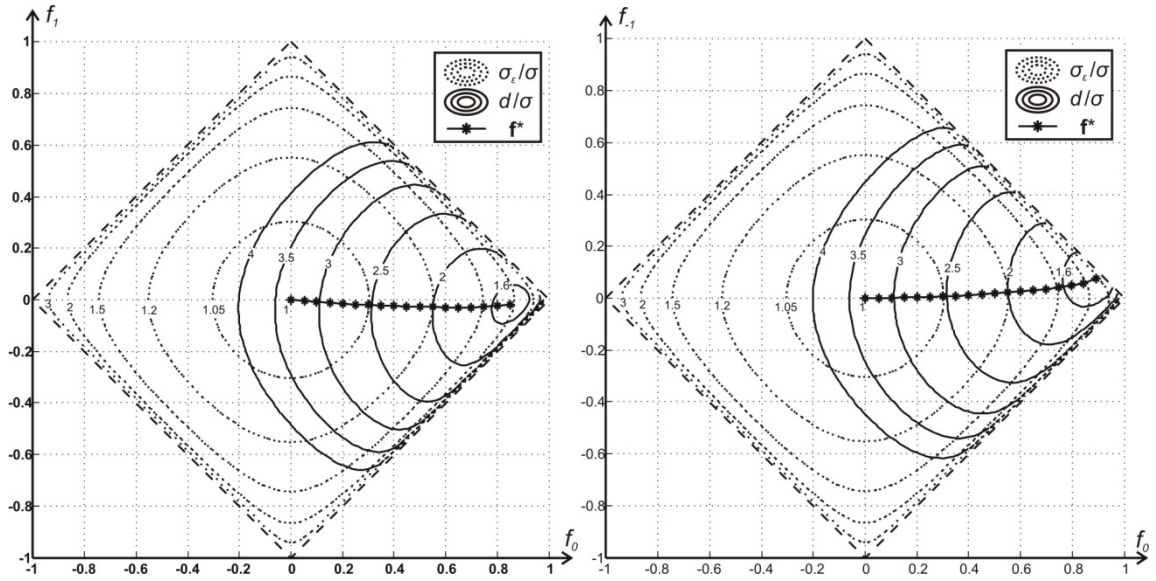
$$\begin{aligned} \text{(MP2)} \quad \min_{\mathbf{f}} \quad & d(\mathbf{f}, \beta) \\ \text{s.t.} \quad & \sigma_\varepsilon(\mathbf{f}) \le s_\varepsilon. \end{aligned}$$

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

Figure 1 shows the contour lines of $\sigma_\varepsilon(\mathbf{f}) = s_\varepsilon$ and $d(\mathbf{f}, \beta) = d$, when $\beta = 0.1$ for two methods that have two nonzero control coefficients. In Figure 1a, all $f_i = 0$ except for f_0 and f_1, and in Figure 1b, only f_0 and f_{-1} are nonzero. 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. Figure 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 7 nonzero control coefficients, which we have tested. Note from Figure 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 ($d/\sigma = 3.4$).



(a) (b)

Figure 1. The iso-$\sigma_\varepsilon/\sigma$ and iso-d/σ contours and the optimal control $\mathbf{f}^*$ with two coefficients, $\beta = 0.1$. (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.

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 nonzero coefficient ($f_0 \neq 0$) may be comparable with that with two or more nonzero 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.

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

d^* / σ	$\sigma_\varepsilon / \sigma = 1$	$\sigma_\varepsilon / \sigma = 1.2$	$\sigma_\varepsilon / \sigma = 1.5$	$\sigma_\varepsilon / \sigma = 2$
$f_i = 0\ \forall i$, 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

Figure 2 shows the optimal control coefficients $f_{-2}^*, f_{-1}^*, f_0^*, f_1^*$, and f_2^* for different β and $\sigma_\varepsilon/\sigma$. Note that the optimal control coefficients are almost 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 Figure 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 nonzero coefficient ($f_0 \neq 0$), which we call the "simple control" method, should be near-optimal.

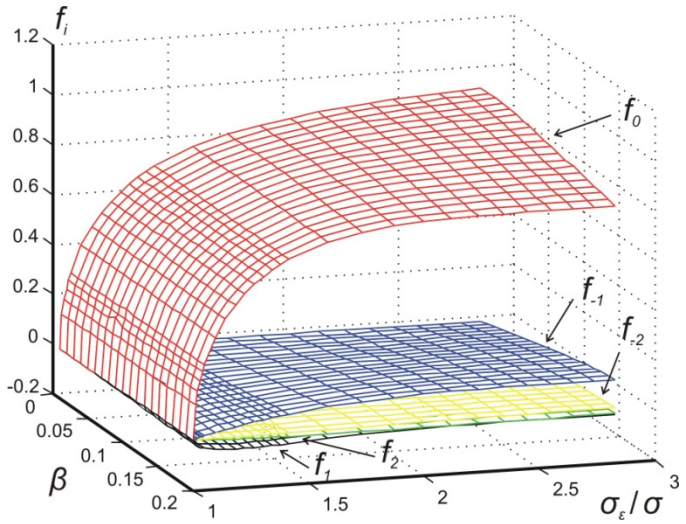


Figure 2. Optimal values of f_{-2}^*, f_{-1}^*, f_0^*, f_1^*, and f_2^* with different demand rate and schedule reliability.

This result 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.

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 equations (4a), (4b), (11b), (13b), (14b), and (14c) reduce to:

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

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

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

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

[9] With perfect information on future bus arrival times, the elements f_{-2}^* and f_{-1}^* are already close to zero. In reality, information will be imperfect and transit agencies will probably rely less on the future bus arrival times, leading to the use of even smaller values for f_{-2}^* and f_{-1}^*.

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

and

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

Note that $\sigma_\varepsilon^2(f_0) > \sigma^2$ if the system is stable ($f_0^2 < 1$).

The optimal solution of (MP1) is obtained by minimizing (15e) such that (15c) is bounded above by s_ε^2, where $s_\varepsilon > \sigma$. Clearly, (15c) must be binding. Thus in the optimal solution the actual variance σ_ε^2 matches the target s_ε^2. Therefore, the optimal coefficient for the simple control method is:

$$f_0{}^* = \sqrt{1-(\sigma^2/\sigma_\varepsilon{}^2)}, \text{ where } \sigma_\varepsilon{}^2 > \sigma^2, \tag{15g}$$

and

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

4.2 Comparison with Other Control Methods

Figure 3 plots (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 $\sigma_\varepsilon/\sigma = 1$. Curves like those shown in Figure 3 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; i.e., because their deviations from schedule $\varepsilon_{n,s}$ grow unbounded as s grows.

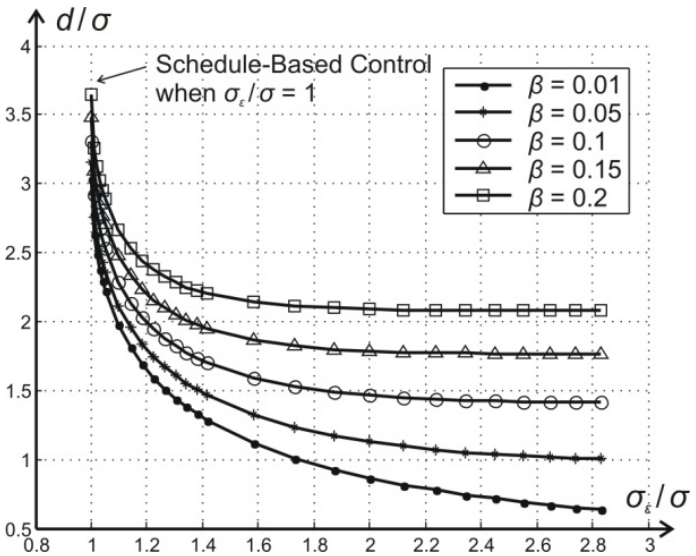


Figure 3. Slack time d vs. schedule reliability σ_ε for the simple control method. Schedule-based control method is represented by the points with $\sigma_\varepsilon/\sigma = 1$.

To further compare the simple control method with the headway-based methods, Figure 4 plots d/σ vs. the dimensionless headway standard deviations σ_h/σ that are allowed. Figure 4 shows that the simple control method behaves better than the forward-headway-based method given by (7), the two-way-headway-based method given by (8), and the backward-headway-based method given by (9). In all three cases, the reduction in slack time is considerable. Note that the headway-based methods cannot achieve σ_h/σ below 1.5 for any slack whatsoever, while the simple control method can. The improved 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.

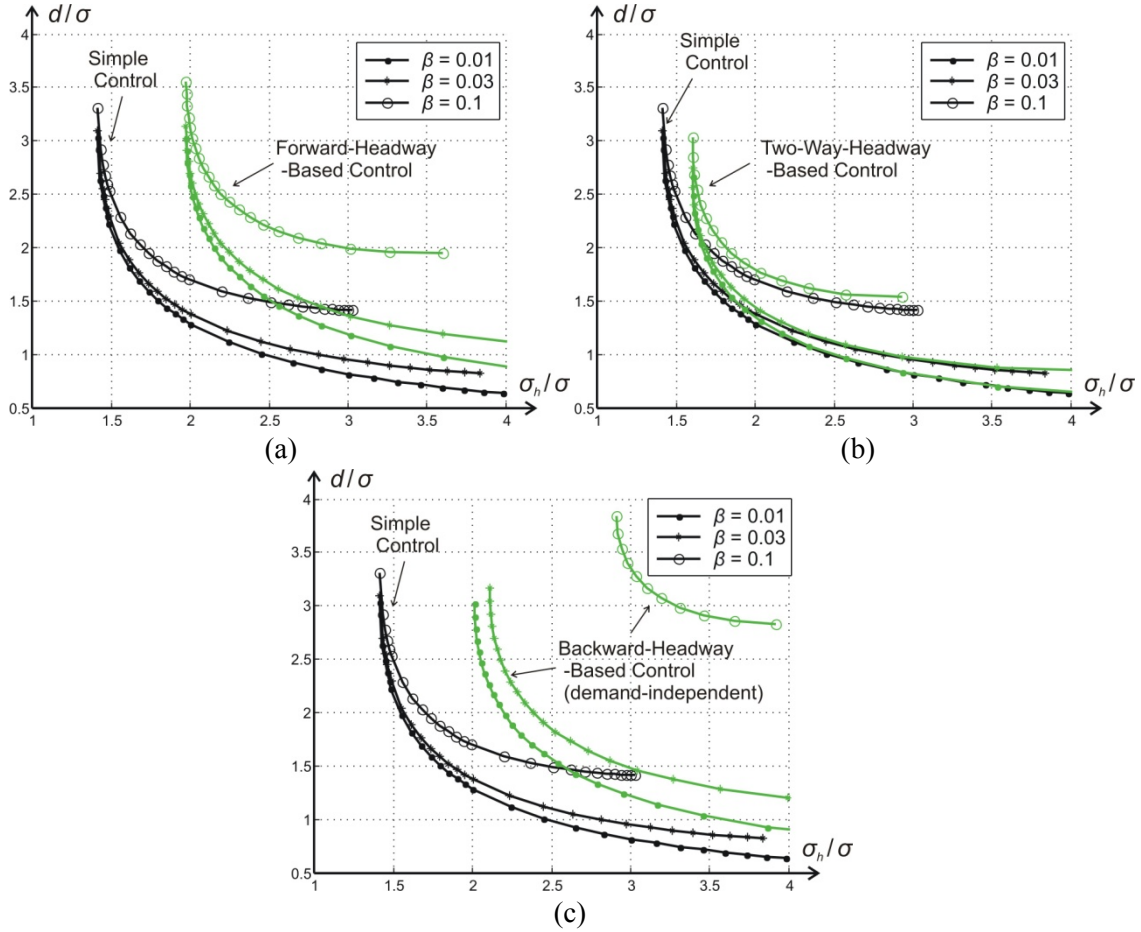


Figure 4. Slack time *d* vs. headway reliability σ_h to compare the simple control method with headway-based control methods relying on: (a) forward headway only; (b) forward and backward headways; and (c) backward headway only.

4.3 Balancing Schedule Reliability and Slack Time

Previously, we have chosen to minimize slack time (maximize commercial speed) subject to a schedule reliability 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-based, backward-headway-based,

two-way-headway-based 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 (γ> 1) times as much as their riding time; and (v) control is applied at all stations. 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:

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

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

Note T_0 includes waiting, line-haul riding and loading time under ideal conditions. It is the minimal possible travel 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 and the slack time. The parameters σ_h^2 and d can be calculated as a function of their control coefficients with equations (13b), (14b) and (14c) for the four considered methods. For each method the control coefficients that minimize (16b) are obtained.

Figure 5 shows the ratio $\Delta T/T_0$ for the optimized control methods, as a function of the demand rates β and the dimensionless trip lengths l/S. The following parameter values were used: S = 400 meters, H = 5 minutes, v_c = 20 km/hr, γ= 2 and σ= 10 seconds. 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-based method, 5-10% for the forward-headway-based method and only 3-7% for the two-way-headway-based method and the simple method. As demand increases, the backward-headway-based method becomes unstable and the simple method also outperforms the rest of methods with a maximal $\Delta T/T_0$ ratio of 15% at β = 0.2 and l/S = 25. Finally, note that the two-way-headway-based method and the simple method are practically indistinguishable. Both methods achieve such similar results because the optimization procedure leads to values of α (the two-way-headway-based control coefficient) which are practically zero.

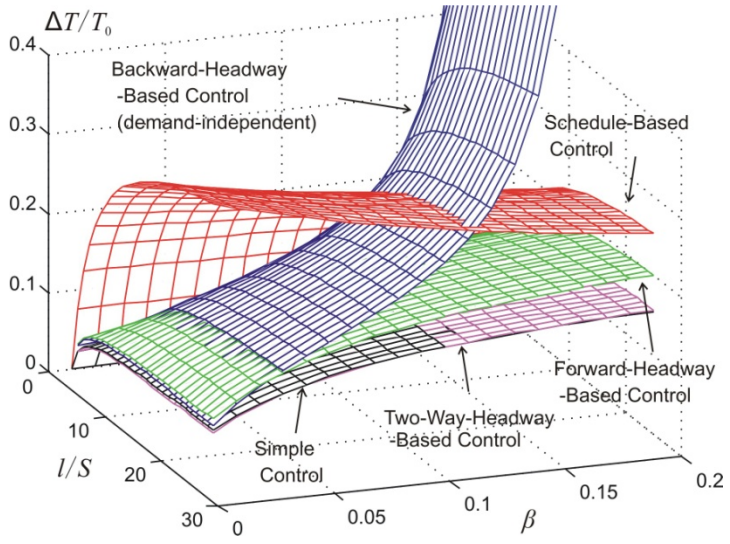
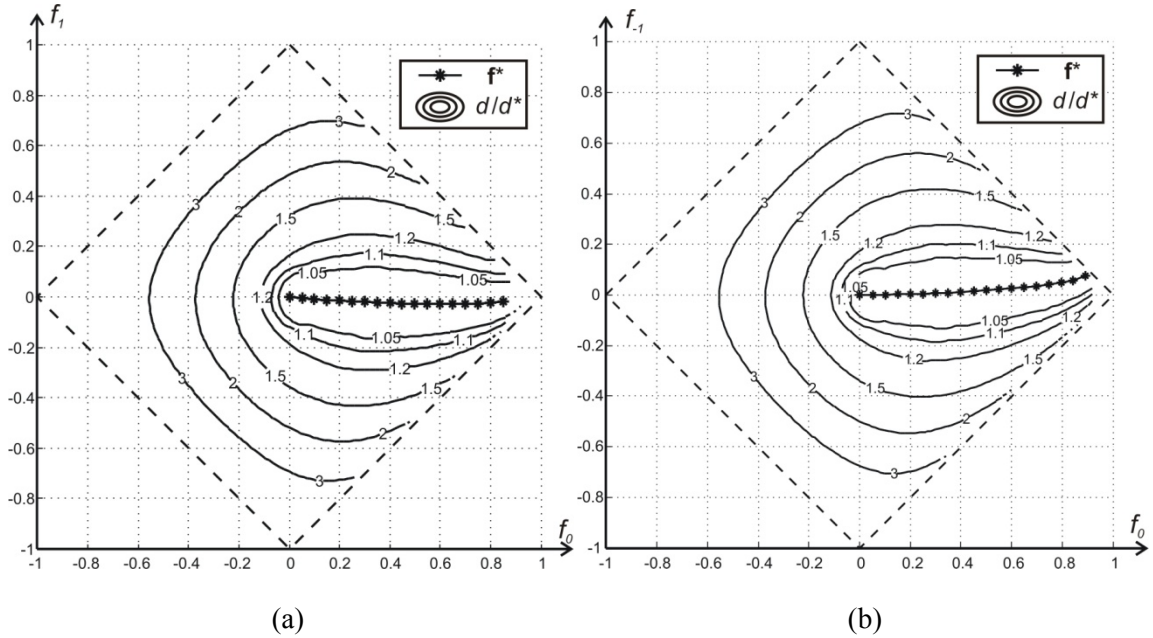


Figure 5. Ratio of variable travel time over fixed travel time ($\Delta T/T_0$) for different control methods with various demand rates β and dimensionless trip lengths l/S.

4.4 Sensitivity to the Control Coefficients

Numerical calculation shows that control coefficients that do not differ much from the optimal control coefficients do not increase the required slack time much. Figure 6 shows the ratio of d/d^* when $\beta = 0.1$ and the general control method has two nonzero 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.



(a) (b)

Figure 6. Sensitivity to control coefficients when $\beta = 0.1$. (a) All $f_i = 0$ except for f_0 and f_1; (b) all $f_i = 0$ except for f_0 and f_{-1}. The contour lines show the ratio of the non-optimal slack time d^* over the optimal slack time d with the same σ_ε value.

5 Robustness to Large Disturbances

When the system is subject to a large disturbance, such as a bus breakdown, the holding times as in (4a) may become negative and buses may bunch if nothing is done. It is proposed in this case to abandon the published schedule (if one is provided) and to focus on regularizing headways. To do this, the virtual schedule is reformulated so it can absorb the effect of the disturbance. This will enable buses to recover from the effect of the large disruption and still provide service with regular headways.

To demonstrate the idea, we next show how to modify the virtual schedule in two situations: (i) when a bus is so far behind schedule that it cannot catch up by itself and (ii) when a bus in the system goes out of service unexpectedly. Other situations can be handled similarly. Only the simple control method is studied.

5.1 Bus Far Behind Schedule

According to assumption (i), the holding time $D_{n,s}$ as in (15a) will almost always be non-negative. But when there is an unexpected large disturbance, the assumption can be violated, and $D_{n,s}$ as calculated from (15a) may turn out to be negative. A negative $D_{n,s}$ means that the control, instead of holding the bus back, would want to push it forward. This is an indication that the bus cannot keep up with the schedule by itself.

To remedy this situation, the current virtual schedule can be shifted forward in time by the smallest possible Δt, so that the critical bus n_0 at s_0 (and therefore all other buses) will be able to keep up with the new schedule by themselves. If we use primes to denote variables under the new virtual schedule, the changes are:

$$t'_{n,s} = t_{n,s} + \Delta t, \quad \forall n, s, \tag{17a}$$

$$\varepsilon'_{n,s} = a_{n,s} - t'_{n,s} = \varepsilon_{n,s} - \Delta t, \quad \forall n, s, \tag{17b}$$

$$D'_{n,s} = d_s - [(1+\beta_s - f_0)\varepsilon'_{n,s} - \beta_s \varepsilon'_{n-1,s}] = D_{n,s} + (1-f_0)\Delta t, \quad \forall n, s. \tag{17c}$$

However, adding too much slack time decreases the commercial speed. Therefore, only the necessary amount of Δt would be added in order to keep the holding time of the critical bus D'_{n_0,s_0} close to but greater than zero, i.e.:

$$\Delta t = (-D_{n_0,s_0})/(1-f_0), \tag{17d}$$

where D_{n_0,s_0} is given by (15a). In reality, (17d) can be modified to include an extra buffer time $\delta > 0$: $\Delta t = (-D_{n_0,s_0})/(1-f_0) + \delta$. Doing this ensures that the procedure does not have to be repeated too often, but delays all the buses a bit more. Note, this process ensures that the system returns to stability and that the process is repeatable, i.e., whenever we find the holding time $D_{n,s}$ for a bus as in (15a) to be negative, the virtual schedule can be shifted forward again using the same method.

5.2 Bus Breakdown

Now imagine there are N buses running on a loop and one bus suddenly breaks down. We will define a new virtual schedule so that the remaining N-1 buses can adhere to it and maintain regular headways, while keeping their commercial speed as fast as possible.

As shown in Figure 7, we number the buses in a way that bus N breaks down, with bus 1 behind it. The headway in the original schedule is H, and the original schedule satisfies $t_{n+1,s} = t_{n,s} + H$ (for $\forall s$, n = 1 to N-2). Under the new schedule, the loss of one bus results in a slightly larger equilibrium headway $H' \approx NH/(N\text{-}1)$.[10] Thus the new schedule satisfies $t'_{n+1,s} = t'_{n,s} + H'$ (for $\forall s$, n = 1 to N-2). We assume that the schedule of bus 1 is shifted backward by a time Δt to be found ($\Delta t < 0$ if shifted forward), i.e., $t'_{1,s} = t_{1,s} - \Delta t$. The schedules of the other buses are shifted to maintain the new headway H':

$$t'_{n,s} = t_{n,s} - \Delta t + (n-1)(H'-H) = t_{n,s} - \Delta t + \frac{n-1}{N-1}H, \quad \forall s, n = 1 \text{ to } N-1, \tag{18a}$$

$$\varepsilon'_{n,s} = a_{n,s} - t'_{n,s} = \varepsilon_{n,s} + \Delta t - \frac{n-1}{N-1}H \quad \forall s, n = 1 \text{ to } N-1. \tag{18b}$$

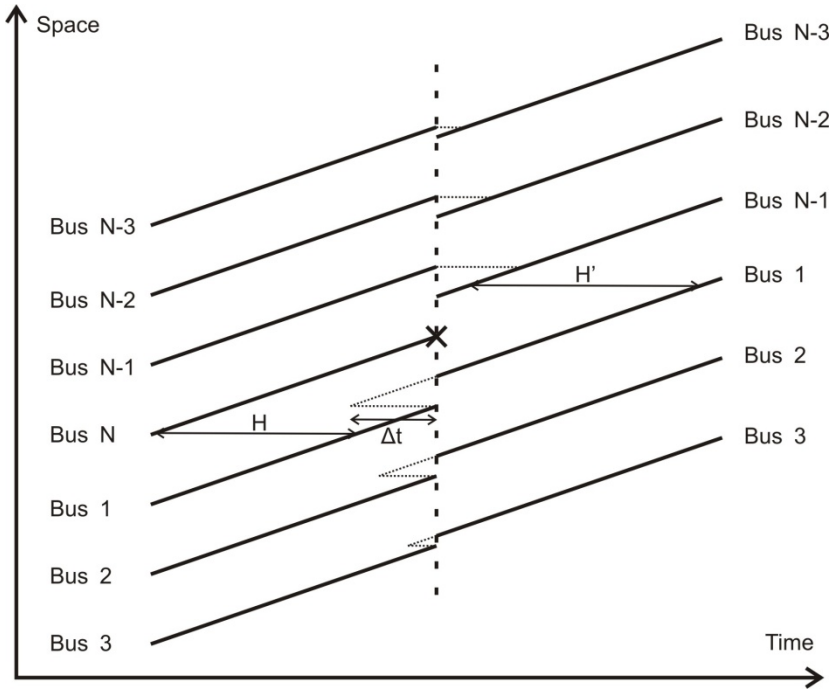


Figure 7. Illustration of virtual schedule change when bus *N* is suddenly out of service.

We assume that at the time of bus breakdown, the current (or closest upstream) station of bus n is $s(n)$. Under the new virtual schedule, the holding times for the given Δt would be:

[10] This approximation is good if the cycle time of buses does not change much after the breakdown. The exact expression for the new headway is $H' = H[(c+d)N]/[(c+d)(N-1) - \beta H]$, where c is the average cruising time between stations, d is the slack time per station, and β is the dimensionless demand rate, in the homogeneous case.

$$D'_{n,s(n)} = d_{s(n)} - [(1+\beta_{s(n)} - f_0)\varepsilon'_{n,s(n)} - \beta_{s(n)}\varepsilon'_{n-1,s(n)}] \qquad (\text{for } n = 2 \text{ to } N\text{-}1)$$

$$= d_{s(n)} - [(1+\beta_{s(n)} - f_0)\varepsilon_{n,s(n)} - \beta_{s(n)}\varepsilon_{n-1,s(n)}]$$

$$-(1-f_0)(\Delta t) + (1-f_0)\frac{n-1}{N-1}H + \beta_{s(n)}\frac{H}{N-1}, \qquad (19a)$$

$$D'_{1,s(1)} = d_{s(1)} - [(1+\beta_{s(1)} - f_0)\varepsilon'_{1,s(1)} - \beta_{s(1)}\varepsilon'_{N-1,s(1)}] \qquad (\text{for } n = 1)$$

$$= d_{s(1)} - [(1+\beta_{s(1)} - f_0)\varepsilon_{1,s(1)} - \beta_{s(1)}\varepsilon_{N-1,s(1)}]$$

$$-(1-f_0)(\Delta t) - \beta_{s(1)}\frac{N-2}{N-1}H. \qquad (19b)$$

As in the last subsection, negative holding times indicate that the corresponding buses cannot keep up with the schedule. Thus Δt is found by solving:

$$(\text{MP3}) \quad \Delta t = \arg\min\left\{\min_n D'_{n,s(n)}(\Delta t)\right\}$$

$$\text{s.t.} \quad D'_{n,s(n)}(\Delta t) \geq 0 \quad \forall n = 1 \text{ to } N-1.$$

Note that each non-negativity constraint imposes an upper bound on Δt, and the solution for Δt is the minimum of these upper bounds. Thus the solution always exists. In reality, the transit agency may add a little buffer δ to Δt to ensure that the remaining N-1 buses do not experience frequent disturbances. The system can then be run with the simple control method, and the process in Section 5.1 can be used to account for other disturbances.

6 Control Law with Real-Time Information on Demand

So far, we have assumed that demand is quasi-deterministic and known, but this may not always be realistic. In this section, we discuss how to relax our assumption on quasi-deterministic demand by exploiting real-time information on passenger boarding.

We assume that the arrival of passengers follows a Poisson process, with the average demand at station s, β_s, known. We also assume that the actual number of passengers to board bus n at station s, $X_{n,s}$, is stochastic. Thus if the headway is given,

$$E[X_{n,s} \mid h_{n,s}] = \beta_s h_{n,s} / t_b, \qquad (20a)$$

$$\text{var}(X_{n,s} \mid h_{n,s}) = \beta_s h_{n,s} / t_b, \qquad (20b)$$

where t_b is the average time needed to board a passenger.

Although $X_{n,s}$ is a random variable, we assume that its realization is observable when bus n finishes boarding at station s.[11] We will show that with this real-time information, the control law can be slightly modified to achieve similar performance as in the quasi-deterministic case.

[11] This is possible because the number of boarding passengers can be measured either by a person at the station, or by an automatic passenger counter system.

The scheduled arrival times satisfy the same equation (1a). Average demand β_s is used for the schedule, since real demand is unknown yet.

$$t_{n,s+1} = t_{n,s} + \beta_s H + d_s + c_s. \tag{21a}$$

The actual motion of buses is affected by the real demand:

$$a_{n,s+1} = a_{n,s} + t_b X_{n,s} + D_{n,s} + c_s + \upsilon_{n,s+1}. \tag{21b}$$

Then, the deviations from schedule can be obtained by subtracting the previous two equations, (21a) from (21b):

$$\varepsilon_{n,s+1} = \varepsilon_{n,s} + t_b X_{n,s} - \beta_s H + (D_{n,s} - d_s) + \upsilon_{n,s+1}. \tag{22a}$$

Real-time demand information is used to decide the holding time. This is feasible because boarding occurs before holding, and the realization of $X_{n,s}$ is known when holding time is calculated. We propose the following instead of (4a):

$$D_{n,s} = d_s - \left(t_b X_{n,s} - \beta_s H + \varepsilon_{n,s} - \sum_i f_i \varepsilon_{n-i,s} \right). \tag{22b}$$

Now if we combine (22a) and (22b), the evolution equation for the deviations from schedule as shown in (23) is the same as equation (4b):

$$\varepsilon_{n,s+1} = \sum_i f_i \varepsilon_{n-i,s} + \upsilon_{n,s+1}. \tag{23}$$

Thus with real-time demand information, the motion of buses under the new control law will not change. The stability analysis and the expression for σ_ε and σ_h still hold. Only the expression for the slack time d_s is slightly different, and is derived in Appendix C. The final result is that:

$$d_s = 3\sqrt{{\sigma_D}^2 + \beta_s t_b H}. \tag{24}$$

where σ_D^2 is calculated using (14b). Figure 8 compares the slack time required vs. schedule reliability under various demand rates, with parameters H = 5 minutes, t_b = 4 seconds, and σ = 10 seconds. Slack times under quasi-deterministic and stochastic situations are calculated by (14c) and (24) respectively. Clearly, the slack time has to be increased, but only significantly if $\beta_s t_b H$ is large compared with σ_D^2.

Therefore, the only changes needed to rigorously account for random demand fluctuations are to use (22b) instead of (4a) for the holding time, and (24) instead of (14c) for the slack time d_s.

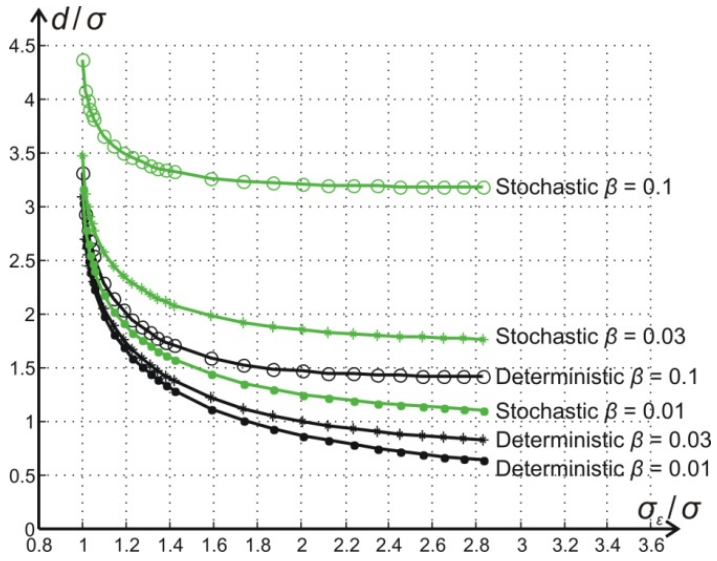


Figure 8. Comparison of slack time *d* vs. schedule reliability σ_ε under various quasi-deterministic and stochastic demand rates for the simple control method.

7 Conclusion

In this paper, we studied a general control method, which is a family of dynamic holding strategies, to improve bus schedule reliability while providing the service with the fastest possible commercial speed. We have four main findings:

- First, the general control method allows buses not only to maintain regular headways but also to adhere to their schedule. None of the existing adaptive methods can achieve this feat. Thus the proposed method is applicable to bus lines with both long and short headways.
- Second, a simple control method, which is a one-parameter version of the general control method, is found to be near-optimal. It outperforms alternative control methods, but only requires information of the current bus and its leading bus.
- Third, in case of large disturbances, buses can still maintain regular headways, by adapting the virtual schedule. This schedule adaptation method guarantees the robustness of the control method under any circumstance.
- Finally, highly stochastic demand can be taken into account by slightly modifying the proposed general control method to incorporate real-time demand information.

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, 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, 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 cheap and resilient way to eliminate bus bunching. *The 4th International Conference on Future Urban Transport*, Gothenburg, Sweden.

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

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.

Eberlein, X. J., Wilson, N. H. M., & Bernstein, D. (2001). The holding problem with real-time information available. *Transportation Science, 35*(1), 1-18.

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*. College Park: University of Maryland.

Senevirante, P. N. (1990). Analysis of on-time performance of bus services using simulation. *Journal of Transportation Engineering, 116*(4), 517-531.

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.

Appendix A: A Method for Locating Control Stations

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, 2009), it is often beneficial to space out the control points more widely. In this spirit, it is assumed here that control stations 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 $\beta << 1$.

We will first transform the equations of bus motion from station to station into similar equations describing the bus motion from control station to control station, as if there were no intermediate stations. It will be shown that when $\beta << 1$, 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 stations $(\sigma')^2$ 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, $\mathbf{f} = [\cdots, \quad 1+\beta, \quad -\beta, \quad \cdots]^T$. It can be seen from the binomial formulas that

$$f_{m|j} = \binom{j}{m}(1+\beta)^m(-\beta)^{j-m}. \tag{A1}$$

Thus,

$$\mathbf{f}_{|j} = \left[\cdots, \quad (1+\beta)^j, \quad j(1+\beta)^{j-1}(-\beta), \quad \binom{j}{2}(1+\beta)^{j-2}(-\beta)^2, \quad \cdots\right]^T$$
$$\approx \left[\cdots, \quad 1+j\beta, \quad -j\beta, \quad 0, \quad \cdots\right]^T. \tag{A2}$$

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

$$\begin{aligned}(\sigma')^2 &= \sigma^2 \sum_{j=0}^{N-1} \sum_m (f_{m|j})^2 \\ &= \sigma^2 \sum_{j=0}^{N-1} \sum_m \left[\binom{j}{m}(1+\beta)^m(-\beta)^{j-m}\right]^2 \\ &\approx \sigma^2 \sum_{j=0}^{N-1}(1+2j\beta) \\ &= \sigma^2(N + N(N-1)\beta).\end{aligned} \tag{A3}$$

It should also be clear from (A2) that when $j = N$ and $\beta << 1$, the dimensionless demand between control stations 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 stations and apply (MP2) with these new parameters. The number of stations between control stations, N, is however, a decision variable in the new version of (MP2). Consideration of (11b) and (14b) reveals that this new mathematical program is:

$$\begin{aligned}\text{(MP4)} \quad \min_{\mathbf{f},N} \quad & \frac{\sigma'}{\sigma}\frac{d(\mathbf{f}, N\beta)}{N} \\ \text{s.t.} \quad & \frac{\sigma'}{\sigma}\sigma_\varepsilon(\mathbf{f}) \le s_\varepsilon \\ & \frac{\sigma'}{\sigma} = \sqrt{N + N(N-1)\beta}.\end{aligned}$$

This MP can be solved numerically as in Section 4.1, or analytically if we adopt the simple control method.

Appendix B: Derivation of the Gradients for Greedy Search

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

$$\begin{aligned}\text{(MP5)} \quad \min_{\mathbf{f}} \quad & \sigma_D{}^2(\mathbf{f}, \beta) \\ \text{s.t.} \quad & \sigma_\varepsilon{}^2(\mathbf{f}) \le s_\varepsilon{}^2.\end{aligned}$$

From equations (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:

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

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

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

$$f_{i|j} = \sum_{\substack{k_0, k_1, \cdots k_{m-1} \\ \sum_{r=0}^{m-1} k_r = j \\ \sum_{r=0}^{m-1} r k_r = i}} \frac{j!}{k_0! k_1! \cdots k_{m-1}!} f_0{}^{k_0} f_1{}^{k_1} \cdots f_{m-1}{}^{k_{m-1}}. \tag{B1}$$

Since

$$\partial f_{i|j} / \partial f_r = j f_{i-r|j-1}, \tag{B2}$$

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

$$\frac{\partial \sigma_\varepsilon{}^2}{\partial f_r} = 2\sigma^2 \sum_{j=0}^{\infty} \sum_i (f_{i|j})(j f_{i-r|j-1}), \tag{B3}$$

and

$$\begin{aligned} \frac{\partial \sigma_D{}^2}{\partial f_r} = 2\sigma^2 \sum_{j=0}^{\infty} \sum_i & \left[(1+\beta) f_{i|j} - \beta f_{i-1|j} - f_{i|j+1}\right] \\ & \left[(1+\beta) j f_{i-r|j-1} - \beta j f_{i-r-1|j-1} - (j+1) f_{i-r|j}\right]. \end{aligned} \tag{B4}$$

Appendix C: Derivation of the Slack Time Formula for the Control Law with Real-Time Demand Information

Recall that it is assumed that $\Pr\{D_{n,s} < 0\} \approx 0$, i.e.,

$$E[D_{n,s}] = 3\sqrt{\text{var}(D_{n,s})}. \tag{C1}$$

It is first shown that the expectation of $D_{n,s}$ is d_s:

$$\begin{aligned} E[D_{n,s}] &= \underset{\upsilon}{E}[\underset{X}{E}(D_{n,s} \mid \upsilon_{\cdot,\cdot})] \\ &= \underset{\upsilon}{E}\left[d_s - \left(\beta_s h_{n,s} - \beta_s H + \varepsilon_{n,s} - \sum_i f_i \varepsilon_{n-i,s}\right)\right] \\ &= d_s - \underset{\upsilon}{E}\left[(1+\beta_s)\varepsilon_{n,s} - \beta_s \varepsilon_{n-1,s} - \sum_i f_i \varepsilon_{n-i,s}\right] \\ &= d_s - \underset{\upsilon}{E}\left[\sum_{j=0}^{s-1} \sum_i \left((1+\beta_s) f_{i|j} - \beta_s f_{i-1|j} - f_{i|j+1}\right) \upsilon_{n-i,s-j}\right] \\ &= d_s. \end{aligned} \tag{C2}$$

Now, consider the law of total variance:

$$\mathrm{var}(D_{n,s}) = \underset{\upsilon}{E}[\underset{X}{\mathrm{var}}(D_{n,s} \mid \upsilon.,.)] + \underset{\upsilon}{\mathrm{var}}(\underset{X}{E}[D_{n,s} \mid \upsilon.,.]), \tag{C3}$$

where

$$\underset{\upsilon}{E}[\underset{X}{\mathrm{var}}(D_{n,s} \mid \upsilon.,.)] = \underset{\upsilon}{E}[\underset{X}{\mathrm{var}}(-t_b X_{n,s} \mid \upsilon.,.)] = \underset{\upsilon}{E}[\beta_s t_b h_{n,s}] = \beta_s t_b H, \tag{C4}$$

and

$$\begin{aligned}
\underset{\upsilon}{\mathrm{var}}[\underset{X}{E}(D_{n,s} \mid \upsilon.,.)] &= \underset{\upsilon}{\mathrm{var}}\left(d_s - \left\lfloor \beta_s h_{n,s} - \beta_s H + \varepsilon_{n,s} - \textstyle\sum_i f_i \varepsilon_{n-i,s} \right\rfloor\right) \\
&= \underset{\upsilon}{\mathrm{var}}\left(d_s - \left[\beta_s (\varepsilon_{n,s} - \varepsilon_{n-1,s}) + \varepsilon_{n,s} - \textstyle\sum_i f_i \varepsilon_{n-i,s} \right]\right) \\
&= \underset{\upsilon}{\mathrm{var}}\left(\beta_s (\varepsilon_{n,s} - \varepsilon_{n-1,s}) + \varepsilon_{n,s} - \textstyle\sum_i f_i \varepsilon_{n-i,s} \right) \\
&= \underset{\upsilon}{\mathrm{var}}\left\{\textstyle\sum_{j=0}^{s-1} \sum_i \left((1+\beta_s) f_{i|j} - \beta_s f_{i-1|j} - f_{i|j+1}\right) \upsilon_{n-i,s-j} \right\} \\
&= \sigma_D{}^2.
\end{aligned} \tag{C5}$$

Thus it follows that $\mathrm{var}(D_{n,s}) = \underset{\upsilon}{E}[\underset{X}{\mathrm{var}}(D_{n,s} \mid \upsilon.,.)] + \underset{\upsilon}{\mathrm{var}}(\underset{X}{E}[D_{n,s} \mid \upsilon.,.]) = \sigma_D{}^2 + \beta_s t_b H,$ where $\sigma_D{}^2$ is calculated using (14b). Now according to (C1):

$$d_s = 3\sqrt{\sigma_D{}^2 + \beta_s t_b H}. \tag{C6}$$