SWAPPING THE ORDER OF SCHEDULED SERVICES TO
MINIMIZE EXPECTED COSTS OF DELAYS

Malachy Carey
Andrzej Kwieciński*
Faculty of Business and Management, University of Ulster, Northern Ireland, BT37 0QB

(Received 10 September 1993; in revised form 28 February 1994)

Abstract—In operating scheduled public transport services, by bus, train, or airline, any delay in
the start time of one activity may cause “knock-on” delays to the next activity: for example, a
train departure delay may delay the next train. In view of this, dispatchers and operators often
have to decide whether delays and costs may be reduced by swapping the order of the delayed
activities. We consider this decision problem here, with trains as an example. We analyse a
minimal information model in which the only information about departure delays is the probabil-
ity distribution of the “ready to depart” times. We show that in this case the optimal (cost
minimizing) swapping policy usually reduces to one of two “bang-bang” policies: swap immedi-
ately or never swap. We develop simple heuristics for deciding which of these two policies is best.
We also consider the effect of new or updated information about “ready to start” times becoming
available as that time approaches, and extend this to a full information model in which the exact
delays are known in advance. We discuss the application of these methods developed for pairs of
trains to multiple trains and stations.

1. INTRODUCTION

In a variety of contexts particularly in public transportation, activities are timetabled in
advance. However, when the timetable is in operation, there are usually random varia-
tions in activity times, so that some activities are not ready to start at their scheduled
times. To reduce the knock-on effects and costs from this, operators may “swap” the
order of such activities and start on the next activity while waiting for a delayed activity
to become “ready to start”. We investigate the question of how long the operator should
wait before swapping activities or whether to swap at all. How long to wait before
swapping (the optimal swap time s) depends on the probabilities of delays and on the
various costs of delays. For specificity, and because it is a particularly interesting general
case, we use the example of a scheduled train service to illustrate the discussion
throughout.

The above problem arises frequently each day for operators of scheduled transporta-
tion. For example, British Rail operated 15,000 scheduled train services per day (2 million
passenger journeys per day) in 1992 and the above problem arose for several percent of
these, hence, at least several hundred times per day. The same problem occurs on the
extensive passenger train systems throughout Europe. In these, each line is normally
dedicated to trains in one direction and trains usually can not pass on the line so that any
delay to one train can substantially delay the next train, especially if a fast train is
following a slow train.

A similar problem arises for scheduled departures of airplanes, buses, etc., particu-
larly if headways are required between departure times or arrival times. Such headways
may be required for reasons of safety or because they share the same track or runway or
passenger boarding facilities or for other policy reasons or technical operating reasons.
A similar problem also arises in nontransportation contexts, for example, in dealing with
scheduled waiting lists for hospital beds or operations or scheduled production or mainte-
nance work, etc.

*Present address: Mathematical Institute, University of Wroclaw, pl. Grunwaldzki 2/4, 50-384 Wroclaw,
Poland.
The above decision as to whether or when to swap the departure order of trains, when some are not ready to depart on schedule is one of the tasks of controllers or dispatchers. In practice, they usually apply some simple prespecified policy rule (or else make a rather arbitrary decision). For example, let the second train depart first (swap the departure order) if and only if the first train is not ready to depart and is more than 10 minutes late. "More than 10 minutes" might be extended to say, "more than 10 minutes late and will not be ready within the next 3 minutes". We are concerned with deriving such swapping rules so as to minimize expected costs of delays and exploring the properties of such optimal swapping rules.

The construction or revision of timetables has been discussed relatively little as compared to the very extensive literature on the scheduling of crews, vehicles, locomotives, etc., to operate already prespecified timetabled services. Also, discussion of timetabling has mainly assumed deterministic times, hence, is not concerned with random delays nor concerned (as we are here) with whether or when to change the departure order in response to such delays. For example, the construction of bus timetables is discussed in Ceder (1986) and construction of airline timetables in Teodorovic (1989) Ch. 3. Construction of train timetables for a single line is discussed in, for example, Kikuchi (1985) and for larger scale rail systems is discussed in, for example, Jovanovic & Harker (1991), Carey & Lockwood (1992), and Carey (19992b, 1993c). There has also recently been much work on (vehicle) scheduling with time-windows (e.g., Kolen, Kan, & Trienekens, 1987, Desrochers, Desrosiers, & Solomon, 1992, Koskosidis, Powell, & Solomon, 1992), and because the times are chosen from a window rather than prespecified, this work includes constructing a timetable.

In general, the above work is deterministic, hence, is not concerned with random delays nor with whether or when to switch the departure orders in response to such delays. However, some other authors have introduced random variation in trip times, wait times, etc. Powell and Sheffi (1983) and Carey (1994a) introduced random variation in link trip times, wait times, etc., and consider the resulting delays in arrival and departure times, etc. and the effects on measures of cost and reliability. However, they assume that the scheduled order is adhered to. Carey (1992) allowed deviation from the scheduled order based on prespecified rules but does not consider whether or when it is desirable to so deviate. Carey and Kwiecinski (1994a) used simulation methods to test and calibrate approximations to knock-on delays when there are random initial delays. Carey and Kwiecinski (1994b) considered convexity of expected cost and reliability measures when there are various kinds of random variations in a schedule. However, neither paper discusses whether or when to deviate from the scheduled order.

All of the aforementioned work is concerned with constructing a fixed timetable in advance and not (as here) with responding to the inevitable deviations from the schedule that occur on the day, due to behaviour of equipment, operators or customers, or due to breakdowns, failures, or other incidents. Some authors (e.g., Newell, * 1974, Barnett, 1978, Turnquist & Blume, 1980, Akkowitz, Figer, & Engelstein, 1986) have considered such real-time variation in trip times of buses or rapid transit, and have proposed real-time holding or control strategies so as to maintain desirable headways or frequencies. However, our concern here is with a service having a specified timetable rather than only a specified frequency.

We focus on the problem of deciding whether or when to swap the scheduled departure order of trains when one or more trains are delayed. In Section 2, we formulate a basic two-train scenario, which we call the minimal information model. In this, the only information available about the delays of trains is their probability distributions. In Section 3, we analyse this model for a slow train scheduled to depart first, and in Section 4 for a fast train scheduled first. We derive simple heuristics for approximating the optimal swapping policies. The optimal policies often turn out to be of the "bang-bang" type, i.e., always swap (let the second train depart first if and when it is ready) or never swap.

In Section 5, we extend the minimal information model to allow for various ways in which the information about the delays may be improved or updated as the ready-to-depart time approaches. This discussion is continued in an Appendix 1, where we consider
a full information model in which an accurate estimate of delay is available in advance. In Section 6, we relate the basic two-train model to the context of multiple trains and multiple stations.

2. A BASIC TWO TRAIN SCENARIO: MINIMAL INFORMATION MODEL

Consider the following scenario. Two trains travel from Station A to Station B. The scheduled departure headway (i.e., the difference between the scheduled departure times from Station A) equals $H$, e.g., Train 1 is due to depart at time $T$ and Train 2 at $T + H$, but to avoid expanding notation we will assume that $T = 0$. The first train may not be ready to depart at its scheduled time. We are concerned with the cost of lateness of arrival of the two trains at Station B and whether the cost can be reduced by introducing a “swap time” $s$. The swap time (or the swapping policy or criterion) $s$ is a time such that if the initial delay of the 1st train causes it to depart later than $s$ then the trains are swapped, that is, Train 2 departs first.

Naturally, train operators would be interested in determining the optimal swapping policy $s$, that is, the policy under which the cost of delays at the terminal station is minimal. The choice of optimal swapping criterion $s$ depends, of course, on the assumed measure of costs or reliability. We assume the following measure of cost of variations in a train’s arrival at its destination.

$$\text{[arrival lateness]} = (\text{[actual arrival time]} - \text{[reference arrival time]})^+, \quad \text{[cost]} = E r ([\text{arrival lateness}]). \quad (1)$$

Here, $E$ denotes the expected value w.r.t. probability distributions of all random variables present, $r$ is a cost function (in the simplest case linear, e.g., $r(d) = c \times d$), and $(\cdot)^+$ indicates that only positive lateness is counted, so that if lateness is negative then it is treated as zero. The actual arrival time is the (random) arrival time of the train in each instance. The reference arrival time is the time with respect to which lateness is measured.

The reference arrival time in (1) will usually be the train’s scheduled arrival time. The arrival lateness in equation (1) would then be the lateness as perceived by users. Indeed, this is the measure we apply in the bulk of this article. However, operators often plan services to run to an unpublished “working” timetable which may be somewhat earlier (or later) than the published scheduled timetable. We can measure lateness costs with respect to this working timetable by using the working arrival time as the reference arrival time in (1). A third alternative, which may be most relevant when there is random variation in trip times, is to take the reference arrival time as the “free running” or “natural” arrival time (we do this in Section 4.3). The “free running” arrival time can be taken (Section 4.3) as including random disturbances and delays en route but excluding initial delays, knock-on effect from a preceding train (hence excluding any swapping policy), and excluding any slack or recovery time built into the timetable. The latter exclusion is important since such slack absorbs, and hence may obscure, the effects of delays.

We assume that the ready-to-depart time of Train 1 is not known exactly in advance. The only information available in advance about the ready-to-depart time of Train 1 is its probability density function.

We assume here (until Section 4.4) that Train 2 becomes ready to depart at its scheduled departure time $H$. We also assume that any delay to either train may incur a cost. In view of this, an obvious dispatching policy to pursue is as follows.

1. If Train 1 becomes ready to depart before some prespecified time $s$, then let it depart first (i.e., maintain the scheduled order of departures), and let it depart immediately. If $s > H$, then letting Train 1 depart first implies delaying the departure time of Train 2 until after Train 1 has departed.

2. If at time $s$ Train 1 is still not ready to depart, then let Train 2 depart first (i.e., swap the scheduled order of departures). Note that $s$ may be less than $H$, in which case Train 2, and hence, Train 1 have to wait until time $H$ to depart.
The swap time \( s \) can be anything from 0 to \( 1 \), and the optimal value to choose for \( s \) will depend on the probability function of the ready-to-depart time of Train 1 and on the respective costs of delaying Trains 1 and 2.

Let \( L \) denote the (random) lateness of the ready-to-depart time of Train 1, and \( l \) or \( L = l \) indicate an instance when \( L \) assumes the value \( l \). We assume that \( L \) is given by a density function \( f(l) \), that is the probability that Train 1 is ready to depart in a time interval \([l, l + dl]\) is equal to \( f(l) \times dl \). Let \( F(l) = \int_{-\infty}^{l} f(x) dx \) denote the cumulative probability function corresponding to the density function \( f \).

For Trains \( i = 1, 2 \) let \( t_i \) and \( T_i \) be respectively the free running trip time and scheduled trip time of Train \( i \). Let \( H \) be the scheduled departure headway between the two trains. For each train, set the reference arrival time in (1) to its scheduled arrival time at Station B, i.e., \( T_1 \) for Train 1 and \( H + T_2 \) for Train 2. Let \( d_i(l, s) \) be the arrival lateness of Train \( i \) at Station B, given the initial delay \( L = l \) of Train 1 and the swap time \( s \). The expected total cost of lateness at Station B given the swap time \( s \) is,

\[
C(s) = Er_1(d_1(L, s)) + Er_2(d_2(L, s)),
\]

where \( r_i \) is the cost of lateness of arrival at Station B for Train \( i \). The optimal or least-cost swap time can be defined as the swap time \( s \) which yields a minimum value of the expected total cost in (2). If the density \( f \) and the cost functions \( r_1 \) and \( r_2 \) are known then the cost \( C(s) \) can be computed numerically from (2) for a range of values of \( s \), and the optimal swap time can be obtained this way. However, it is also valuable and may be more practical to investigate the nature and properties of optimal policies. This allows us to form some simple heuristics for approximating optimal swap times, thus reducing or eliminating the need for repetitive and context dependent calculations. We pursue this in Sections 3 and 4. In Section 3, the slower train is scheduled to depart first, and in Section 4, the faster train is scheduled to depart first.

To take account of interaction between trains travelling on the same line and separated by short headways, we introduce “knock-on” trip times \( K_{ij}(h) \) and \( K_{ji}(h) \), where \( K_{ij}(h) \) is the actual trip time of Train \( j \) if Train \( i \) departs \( h \) units of time before it. These knock-on trip times may have the form of (or may be approximated by) functions of the actual departure headway \( h \) and the free running times \( t_i \) and \( t_j \), that is, \( K_{ij}(h) = K_{ij}(t_i, t_j, h) \), see Carey and Kwiecinski (1992a).

To simplify the discussion, we assume that these knock-on effects on trip times are as follows. One of the trains is a fast train, the other train is a slow train. If the faster train departs first then the slower train is not affected. This means \( K_{ij}(h) = t_j \). If the slower train departs first then the faster train travels to B at its normal speed, unless it catches up with the slower train in which case it arrives at B immediately after the slower train. Equivalently, \( K_{ji}(h) = \max\{t_j, t_i - h\} \).

3. TRAIN SWAPPING WHEN SLOWER TRAIN IS SCHEDULED TO START FIRST

Throughout this section we assume that Train 1 is slower than Train 2. The reverse order of trains is discussed in Section 4. We also assume that the scheduled headway \( H \) is sufficiently large, so that if Train 1 departs on time it would not interfere with Train 2, that is, \( t_2 + H > t_1 \). However, if Train 1 does not depart on time and does not depart until “too near” to the scheduled departure time \( H \) of Train 2 then Train 2, being faster, may catch up with and hence get delayed by Train 1 in front of it. (Specifically, if \( t_1 + l > t_2 + H \) then Train 2 will catch up with and be delayed by Train 1). In view of this, if the (slower) Train 1 is not ready to depart by a certain time \( s \), then it may be best to hold it until after the departure of the next (faster) Train 2. Also, note that because Train 2 is faster than Train 1, letting it go before Train 1 will not delay Train 1. Hence, Train 2 should always be let go when it is ready, which we are here assuming is at its scheduled time \( H \). This means that the train order need never be swapped at a time later than \( H \), that is, \( s \leq H \).

For \( 0 \leq s \leq H \) the lateness \( d_i(l, s) \) for trains \( i = 1, 2 \) can be expressed as,
Minimizing expected delays

\[ d_1(l, s) = \begin{cases} 
(H + t_1 - T_1) & \text{if } s < l \leq H, \\
(l + t_1 - T_1) & \text{otherwise},
\end{cases} \]

\[ d_2(l, s) = \begin{cases} 
(l + t_1 - T_2 - H) & \text{if } H - (t_1 - t_2) < l \leq s, \\
(t_2 - T_2) & \text{otherwise}.
\end{cases} \]

For instance, the first alternative in (4) corresponds to the situation when the trains are not swapped, and Train 1 departs late enough for Train 2 to be affected by knock-on delay from Train 1. In this case, Train 2 arrives at B at time \( l + t_1 \), immediately after Train 1. This alternative is void if \( s \leq H - (t_1 - t_2) \), i.e., the trains are swapped before Train 1 starts affecting Train 2. In the other alternative, either the trains are swapped (\( l > s \)) or the headway is large enough to absorb knock-on effects from Train 1 (\( l \leq H - (t_1 - t_2) \)). Equation (3) can be explained in a similar way.

To reduce the number of parameters in the model we assume for the present that the free running trip times equal the scheduled trip times, that is \( t_1 = T_1 \) and \( t_2 = T_2 \). This can be justified on the grounds that there may be little or no random variation in the free running trip times (that is, they are deterministic). Setting \( t_1 = T_1 \), \( t_2 = T_2 \) and \( k = t_2 - t_1 \) in (3)-(4) yields,

\[ d_1(l, s) = \begin{cases} 
H & \text{if } s < l \leq H, \\
0 & \text{otherwise},
\end{cases} \]

\[ d_2(l, s) = \begin{cases} 
(l + k - H) & \text{if } H - k < l \leq s, \\
0 & \text{otherwise}.
\end{cases} \]

Recall that we need only consider swap times \( s \) smaller than \( H \). Also, for similar reason, once we assume that \( k = t_1 - t_2 \) is constant, we need only consider \( s \) larger than \( H - k \). For \( H - k \leq s \leq H \), taking expectations of (5)-(6) expands (2) to,

\[ C(s) = \int_0^s r_1(l) f(l) dl + \left[F(H) - F(s)\right] r_1(H) + \int_{H-k}^H r_2(l + k - H) f(l) dl. \]

Consequently,

\[ C'(s) = f(s) [r_1(s) - r_1(H) + r_2(s + k - H)], \]

assuming that the functions \( r_1(\cdot) \) and \( f(\cdot) \) are continuous at the point \( s \), and \( r_2(\cdot) \) at the point \( s + k - H \).

The following proposition establishes an important property of the cost function (7).

**Proposition 3.1.** Suppose that (a) \( r_1(\cdot) \) and \( r_2(\cdot) \) are nonnegative and (strictly) increasing, (b) \( r_2(0) = 0 \), and (c) \( f(\cdot) > 0 \) in the interval \([H - k, H]\).

Then the cost function \( C(s) \) in (7) is first (strictly) decreasing and then (strictly) increasing in the interval \([H - k, H]\), hence (strictly) quasiconvex, so that \( C(s) \) has a unique minimum in the interval \([H - k, H]\).

**Proof:** From assumption (a), \([r_1(s) - r_1(H)] + r_2(s + k - H)\) in (8) is (strictly) increasing in \( s \) in the interval \([H - k, H]\), and is nonnegative (positive if the \( r \)'s are strictly increasing) in the right end of the interval. From (b), it is nonpositive (negative) in the left end of the interval. Also, \( f(\cdot) > 0 \) hence \( C'(s) \) in (8) is first negative and then positive. The proposition follows.

**Corollary 3.2.** Under the assumptions of Proposition 3.1, the value of the swap time \( s \)
which yields a minimum of the cost $C(s)$ does not depend on the form of the pdf $f$.

**Proof:** Note that the sign of $C'(s)$ in (8), and hence the optimal value of $s$, does not depend on $f$, as long as $f(\cdot)$ is positive in $[H-k, H]$.

Corollary 3.2 greatly simplifies the decision problem, because it indicates that knowledge of the exact form of the pdf is not required to choose an optimal swap time $s$. Such as $s$ can be obtained simply from (7) by solving $[r_1(s) - r_1(H) + r_2(s + k - H)] = 0$, either analytically or numerically. The solution of this equation has a particularly simple form when the cost functions $r_1$ and $r_2$ are linear, i.e., $r_i(d) = c_i \times d$, for $i = 1, 2$. Then the optimal $s$ is,

$$s = H - \frac{k}{1 + c_1/c_2}. \quad (9)$$

Equation (9), though exact only when the cost functions are linear, may also be used as a heuristic in more complex cases. The ratio $c_1/c_2$ may be interpreted as a measure of relative importance of (the slower) Train 1 w.r.t. (the faster) Train 2. Such a measure may be estimated even if the costs are not linear. The constant $k$ is a measure of knock-on effects, and if the trip times are not deterministic then it may be replaced by another measure of knock-on effects, say, the expected knock-on delay to the faster train if it departs from the station shortly after the slower train.

4. Train Swapping When Faster Train is Scheduled to Start First

We consider the basic model from Section 2 but with the scheduled order of departures reversed that of in Section 3. That is, Train 1 is now faster than Train 2.

Note that if the costs $r_i$ are nondecreasing then the range of acceptable swap times $s$ can be restricted to $s \geq H$. This can be justified in a similar way to the analogous restriction in Section 3. For $s \geq H$ the lateness $d_i(l,s)$ for trains $i = 1, 2$ can be expressed as,

$$d_i(l,s) = \begin{cases} (s + t_2 - T_1) + & \text{if } s < l \leq s + t_2 - t_1, \\ (l + t_1 - T_1) + & \text{otherwise}, \end{cases} \quad (10)$$

$$d_2(l,s) = \begin{cases} (t_2 - T_2)^+ & \text{if } l \leq H, \\ (l + t_2 - H - T_2)^+ & \text{if } H < l \leq s, \\ (s + t_2 - H - T_1)^+ & \text{if } s < l, \end{cases} \quad (11)$$

where these equations can be explained in a similar way to (3)-(4).

As in Section 3, we assume that $t_1 = T_1$ and $t_2 = T_2$, and that there is little random variation in the free running trip times or that they are deterministic. This assumption is relaxed in Section 4.3 below. Setting $t_1 = T_1$, $t_2 = T_2$ and $k = t_2 - t_1$ in (10)-(11) yields,

$$d_1(l,s) = \begin{cases} s + k & \text{if } s < l \leq s + k, \\ l & \text{otherwise}, \end{cases} \quad (12)$$

$$d_2(l,s) = \begin{cases} 0 & \text{if } l \leq H, \\ l - H & \text{if } H < l \leq s, \\ s - H & \text{if } s < l. \end{cases} \quad (13)$$

Taking expectations of (12)-(13) expands (2) to,

$$C(s) = \int_0^s r_1(l)f(l)dl + [F(s + k) - F(s)]r_1(s + k) + \int_{s+k}^{\infty} r_1(l)f(l)dl$$
Minimizing expected delays

\[ + \int_{-\infty}^s r_i(l - H)f(l)dl + [1 - F(s)]r_2(s - H). \] (14)

Throughout the rest of this section, until explicitly stated otherwise in Section 4.5, we assume that the cost functions \( r_i \) are linear, i.e., of the form \( r_i(d) = c_i \times d \), for \( i = 1,2 \). We do this to draw some general conclusions as to properties of expected costs in Section 4.1 and form a simple heuristic for approximating the optimal swapping criterion based on a set of examples in Section 4.2. This heuristic, like the heuristic from Section 3.1, does not depend on the particular form of the pdf \( f \). In Sections 4.3-4.5 we discuss properties of models obtained by relaxing one or more restrictions imposed here.

4.1. Properties and implications of the cost function

To find properties of the swap time \( s \) which will minimize expected total cost (2) we write out the derivatives of (14). Recall that for the moment we assume that the cost functions \( r_i(d) = c_i \times d \), for \( i = 1,2 \), are linear. Thus,

\[ c'(s) = c_1[F(s + k) - F(s) - kf'(s)] + c_2[1 - F(s)] = -c_1A_1 + c_2A_2, \] (15)

where \( A_1 \) and \( A_2 \) are as in Fig. 1, assuming that \( f(a) \) is continuous at the point \( s \), and

\[ c''(s) = c_1[f(s + k) - f(s) - kf'(s)] - c_2f(s), \] (16)

assuming that \( f'(\cdot) \) exists at the point \( s \).

We can now make the following observations about the value of \( s \) that will yield a minimum cost \( C(s) \). The derivative (16) can be negative or positive in which case \( C(s) \) is concave or convex, respectively. However, \( C(s) \) will very often be concave, for the following reasons.

1. If for given \( s \), \( f(\cdot) \) is concave on the interval \( s \) to \( s + k \), then \([f(s + k) - f(s) - kf'(s)] < 0\), which implies that (16) is negative, hence \( C(s) \) is concave at \( s \). Further, if \( f(\cdot) \) is concave on the interval \( s \) to \( s + k + \delta \) then \( C(\cdot) \) is concave on the interval \( s \) to \( s + \delta \).

2. If \( f \) is convex over some regions (e.g., in the right tail) then \([f(s + k) - f(s) - kf'(s)] > 0\) in this region, but (16) may still be negative due to the negative term \(-c_2f(s)\), hence \( C(s) \) may still be concave.

3. The relevant portion of \( f \) is the right tail (the relevant part is the part beyond the scheduled departure time of Train 2). Suppose this tail is approximately linearly downward sloping, then \([f(s + k) - f(s) - kf'(s)] = 0\) so that (16) reduces to \(-c_2f(s)\) which is negative, hence \( C(s) \) is concave.

If a cost function \( C(s) \) is concave then the values of \( s \) which yield minimum cost are the boundary values of \( s \). That is \( s = H \) or \( s = +\infty \), or, in the latter case, \( s = t_m \) where \( t_m \) is the very latest time by which Train 1 will be ready to depart (this is \( s = +\infty \) yield

![Fig. 1. Illustration of \( C'(s) = -c_1A_1 + c_2A_2 \).](TR(B) 25:6-8)
equivalent swapping policies). To find which of these is the optimal \( s \) we simply evaluate \( C(H) \) and \( C(+\infty) \) or \( C(t_m) \) to see which is smallest.

Setting \( s = H \) means, always let Train 2 depart at its scheduled time if Train 1 is not yet ready to depart (i.e., swap immediately). Setting \( s = +\infty \) (or \( s = t_m \)) means, do not let Train 2 depart until Train 1 has departed, no matter how long this requires waiting (i.e., never swap). We thus have a “bang-bang” policy: swap immediately or never swap.

Even if \( C(s) \) is not concave, it is quite likely that it is upward sloping [(15) positive] for small values of \( s \), and/or downward sloping [(15) negative] for large values of \( s \). This can be seen by examining (15) and Fig. 1. In this case, the optimal (minimum) cost \( C(s) \) is again likely to occur at \( s = H \) or \( s = +\infty \) (or \( s = t_m \)). This again yields extremal or bang-bang dispatching policies (swap immediately or never swap).

4.2. Examples and heuristics

The above discussion suggests that the optimal swapping policy depends significantly on the shape of the probability density function \( f \). However, it also suggests that, in many cases, the optimal policy reduces to a bang-bang swapping policy—swap immediately or not at all. To explore and illustrate the optimal swapping policy further we consider in turn three different families of probability density functions, namely the density functions with a uniform, a triangular and a negative exponential right tail, respectively. For each of these, we derive an optimal swapping decision rule. We find that (a) the optimal policy turns out to be a bang-bang policy in each case, and (b) the numerical criterion as to which bang-bang policy is optimal turns out to be much the same in each case, irrespective of the form of the pdf. This has important practical implications, because in practice information as to the form of the pdf is often poor. It also suggests using a simple heuristic decision rule that does not require information on the form of the pdf. This heuristic decision rule can be extended to allow stochastic trip times (Section 4.3) and letting both trains experience random delays in their ready-to-depart time (Section 4.4).

The parameters traditionally used to define the density \( f \) of a random variable \( L \) have different meanings in each of the three examples following. We, therefore, define the following characteristic of the delay \( L \) and form all conclusions in terms of this characteristic. Consider the right tail of \( f \) beginning at point \( H \). Rescale this tail to create a full probability density function and rename \( H \) as the new origin. Let \( \mu \) be the expected value of this residual probability distribution. In other words, \( \mu \) is the mean lateness of the ready-to-depart time of Train 1 in excess of the scheduled departure time of Train 2, that is, lateness that can cause any interference between the two trains. More formally,

\[
\mu \overset{\text{def}}{=} E(L - H|L \geq H) = \frac{E(L - H)^+}{P(L \geq H)} = \frac{1}{1 - F(H)} \int_H^\infty (l - H)f(l)dl.
\]

Example 1 (Uniform tail): Suppose that the probability distribution of the ready-to-depart time delay \( l \) of Train 1 consists of a mass \( p \) of "early arrivals" distributed arbitrarily up to time \( H \), and a uniform tail for delays greater than \( H \). That is, let \( F(H) = p \) and, for \( H \leq l \leq H + \Delta \), let \( F(l) = 1 - \frac{q}{\Delta} (H + \Delta - l) \), where \( \Delta > 0 \) and \( q = 1 - p \).

Then from (15),

\[
C'(s) = \begin{cases} 
\frac{1}{\Delta} c_2(H + \Delta - s) & \text{if } s + k \leq H + \Delta, \\
\frac{q}{\Delta} [(c_1 + c_2)(H + \Delta - s) - c_1 k] & \text{otherwise.}
\end{cases}
\]

The derivative \( C'(s) \) is everywhere decreasing, so that \( C(s) \) is concave. Hence, the optimal policy is one of the two extremal policies: either never to swap trains or to swap at time \( H \). The density \( f(\cdot) \) and cost \( C(\cdot) \) are sketched in Fig. 2.
We can easily determine which of the two policies is the optimal one by calculating the difference \( C(H + \Delta) - C(H) \), which can be obtained by integrating (17) over the interval \([H, H + \Delta]\). Integrating and simplifying, it turns out that the “never swap” policy is optimal (i.e., \( C(H + \Delta) < C(H) \)) if and only if
\[
\frac{c_2}{c_1} < \begin{cases} \frac{k^2}{\Delta^2} & \text{if } k \leq \Delta, \\ 2\frac{k}{\Delta} - 1 & \text{if } k > \Delta, \end{cases}
\]  

(18)

Suppose that the ratio of costs \( c_2/c_1 \) is fixed in the problem, while other parameters in (18) may not be. Then, by reversing (18), we find that the trains should never be swapped if and only if
\[
\frac{k}{\mu} > \varphi \left( \frac{c_2}{c_1} \right) \triangleq \begin{cases} 2 \sqrt{\frac{c_2}{c_1}} & \text{if } \frac{c_2}{c_1} \leq 1, \\ \frac{c_2}{c_1} + 1 & \text{if } \frac{c_2}{c_1} > 1, \end{cases}
\]

where \( \mu = \Delta/2 \) is the mean delay of Train 1 excess to the scheduled departure time \( H \) of Train 2. The “swap immediately at \( H \)” policy is optimal if the reverse inequality holds, and in the case of equality the two policies yield the same cost.

We use the symbol \( \varphi \) from the above example to express similar relationships in the remaining examples, even though the exact formulas will change, thus treating \( \varphi(\cdot) \) as a collective term for a boundary separating the ranges of parameters for which one or the other extreme swapping policy is optimal. The function \( \varphi \) is illustrated in Fig. 3.

**Example 2 (Triangular tail):** Suppose that the probability distribution of the ready-to-depart time delay \( I \) of Train 1 consists of a mass \( p \) of early arrivals distributed arbitrarily up to time \( H \), and a triangular tail for delays greater than \( H \). That is, let \( F(I) = p \) and,
\[
\text{for } H < I < H + A, \text{ let } F(I) = 1 - 3\frac{I - H}{A}, \quad \text{where } A > 0 \text{ and } 4 = 1 - p.
\]

Then from (15),
\[
C'(s) = \begin{cases} \frac{q}{\Delta^2}[c_s(H + \Delta - s)^2 - c_kk^2] & \text{if } s + k \leq H + \Delta, \\ \frac{q}{\Delta^2}(H + \Delta - s)[(c_s + c_k)(H + \Delta - s) - 2c_kk] & \text{otherwise.} \end{cases}
\]  

(19)
As can be easily seen, the derivative $dC/ds$ is decreasing when the first of the above alternatives holds, hence $C(s)$ is concave for $s + k \leq H + \Delta$ (this range may be empty if $k > \Delta$). For $s + k > H + \Delta$ the derivative is not necessarily decreasing (and the function concave). Note, however, that for $s \in (H + \Delta - k, H + \Delta)$ we have

$$\frac{dC}{ds} \leq 0 \iff (c_1 + c_2)(H + \Delta - s) - 2c_1k \leq 0. \tag{20}$$

The latter expression in (20) is linear with a negative value in the right end of the interval. Thus, it is either negative for all $s$ in the interval or first positive and then negative. For the function $C$ this means that it is either decreasing in $(H + \Delta - k, H + \Delta)$, or first increasing and then decreasing. Moreover, it can be easily checked that $C$ has an increasing part in this interval if and only if it is increasing for $s \leq H + \Delta - k$. Hence, on the entire interval $[H, H + \Delta]$ the function $C(s)$ either (a) is everywhere decreasing or (b) is first increasing and then decreasing. Thus, $C(s)$ therefore has a global minimum at either of the two ends of its domain.

As in Example 1, by integrating (19) over $[H, H + \Delta]$ the difference $C(H + \Delta) - C(H)$ can be calculated to determine the criterion for one or the other of the two policies to be the optimal one. Omitting the details, the criterion indicates the “never swap” policy to be the best if and only if $k/\mu > \varphi(c_1/c_2)$, where $\varphi$ is an increasing function (though not the same function as in Example 1) and $\mu = \Delta/3$ is the mean delay of Train 1 excess to the departure time $H$ of Train 2.

**Example 3 (Exponential tail):** Suppose that the probability distribution of the ready-to-depart time delay $I$ of Train 1 consists of a mass $\rho$ of “early arrivals” distributed arbitrarily up to time $H$, and an exponential tail for delays greater than $H$. That is, let $F(H) = \rho$ and, for $I \geq H$, let $F(I) = 1 - qe^{-\lambda(I-H)}$, where $\lambda > 0$ and $q = 1 - \rho$. Then from (15),

$$C'(s) = c_i q e^{-\lambda s - H}(1 - \lambda k - e^{-\lambda s}) + c_2 q e^{-\lambda s}.$$  

It can be readily seen that,

$$C'(s) \leq 0 \iff c_2 \leq c_1 \lambda k - 1 + e^{-\lambda s}. \tag{21}$$
Also, \( C''(s) = -\lambda C'(s) \). Because \( \lambda k - 1 + e^{-\lambda k} \) is a constant not dependent on \( s \), the cost function \( C(s) \) is everywhere either (a) strictly decreasing and convex, (b) constant, or (c) strictly increasing and concave, depending on the values of the constants \( c_2/c_1 \) and \( \lambda k \). This means that the corresponding optimal (cost minimizing) swapping policy is either (a) \( s = +\infty \) hence never swap, (b) swap at any time, as it does not affect costs \( C(s) \), or (c) \( s = H \) hence swap immediately.

Note that \( \mu = 1/\lambda \) because \( f \) is negative exponential), hence \( k/\mu = \lambda k \). Also, because the function \( \lambda k - 1 + e^{-\lambda k} \) in (21) strictly increases from 0 to \( \infty \) w.r.t. \( \lambda k \), it can be reversed. Denote the inverse function by \( \varphi \). Then (21) states that the trains should never be swapped if \( k/\mu > \varphi(c_2/c_1) \), should be swapped at \( H \) if the inequality is reversed, and the two policies yield the same costs if the equality holds. This is the same result as in Examples 1 and 2, though the form of \( \varphi \) is different.

Note that due to the “memoryless” property of the negative exponential distribution the expected value of the right tail would remain the same if the departure time \( H \) of Train 2 was now to be changed, provided that \( f(l) \) was still exponential for \( l \geq H \). Hence, changing the scheduled departure times would not affect the choice of one or the other of the bang-bang policies.

In each of the examples above we found a function \( \varphi \) such that the optimal decision rule is:

\[ \text{Never swap trains if } (k/\mu) > \varphi(c_2/c_1), \text{ otherwise, swap immediately,} \]

where “swap immediately” means, if Train 1 is not ready to depart at the departure time of Train 2 then let Train 2 go first. Though the form of the decision function \( \varphi \) looks different for each of these examples, we found that the numerical result yielded by these forms are very similar. This is shown by plotting the three forms of Fig. 3. As Fig. 3 shows, the graphs are very close to each other, indicating a strong tendency for the parameters \( (c_2/c_1) \) to cause one or other of the bang-bang swapping policies to be the optimal one, irrespective of which of the three different forms of the pdf we assume.

This suggests that the bang-bang decision rule implicit in Fig. 3 will apply as a good approximation even for other forms of the pdf or when, as is often the case, we have little information on the form of the pdf.

We therefore suggest the following simple heuristic decision rule to be used whether or not the pdf is known: calculate \( (c_2/c_1) \) and \( (k/\mu) \) and apply these to Fig. 3 to choose the optimal swapping policy. Note that the required data \( (c_1, c_2, \text{ and } k) \) are assumed known constants, and \( \mu \) may be known at least approximately even if the exact form of the pdf is not known. Furthermore, unless the data yields a point in Fig. 3 close to the dividing line, only a rough estimate of \( \mu \) is needed to decide in favour of one or other of the two extremal policies (swap or not swap). If, in a more complex model (e.g., in Section 4.3), the trip times are stochastic then \( k \) may be replaced by another measure of knock-on effects, say the expected knock-on delay of Train 1 if it departs from Station A shortly after Train 2.

If the ratios define a point within one of the two disjoint areas on the graph then choose the corresponding policy. If the point lies in or close to the boundary area, roughly the area between the three plots in Fig. 3, assume that the two policies yield the same or similar costs and choose either, or choose the one more desirable from other operational point of view.

If a heuristic similar to the above is to be applied in more complex models then some grounds must exist for the assumption that a bang-bang policy yields (approximately) optimal costs. We provide this in Sections 4.3–4.5, by extending the discussion from Section 4.1 to more complex more flexible models.

4.3. Introducing stochastic trip times

Let us now allow \( k = t_2 - t_1 \) to be stochastic (a random variable). Some conclusions can be reached about this case by reducing it to the deterministic case. To distinguish the two cases let \( K \) and \( k \) denote the difference \( (t_2 - t_1) \) in the stochastic and deterministic
case, respectively. If the random variables $K$ and $L$ (random initial delay of Train 1) are not independent, the density function $f(l)$ of $L$ has to be replaced with a conditional density $f(l|k)$ of the random variable $L$ given that $K = k$. If these variables are independent then the conditional density is the same as the unconditional one.

In the deterministic case already discussed, it is clear from the cost equation (14) that expected costs depend on the knock-on delay $k$ as well as on the swap time $s$. However, because we were assuming throughout that the knock-on time $k$ was a fixed constant, we simply denoted costs by $C(s)$, rather than $C(s,k)$. But as we now wish to consider $k$ varying randomly, we denote expected costs, from (14), by $C(s,k)$, and similarly write $C(s,k)$ instead of $C(s)$ in equations (15)-(16), etc.

If the random variables $K$ and $L$ (random initial delay of Train 1) are not independent, the density function $f(l)$ of $L$ has to be replaced with a conditional density $f(l|k)$ of the random variable $L$ given that $K = k$. If these variables are independent then the conditional density is the same as the unconditional one.

In the deterministic case already discussed, it is clear from the cost equation (14) that expected costs depend on the knock-on delay $k$ as well as on the swap time $s$. However, because we were assuming throughout that the knock-on time $k$ was a fixed constant, we simply denoted costs by $C(s)$, rather than $C(s,k)$. But as we now wish to consider $k$ varying randomly, we denote expected costs, from (14), by $C(s,k)$, and similarly write $C(s,k)$ instead of $C(s)$ in equations (15)-(16), etc.

However, to maintain equation (14) based on equations (12)-(13) as the definition of the cost yielded by a swapping criterion $s$ and knock-on effect $k$, we have to change the interpretation we associated with functions $d(l,s)$ given by (12)-(13). This is because equations (12)-(13) were derived from (10)-(11) by making the free running trip times $t_i$ and the scheduled trip times $T_i$ equal on the grounds that they are deterministic.

We can, nevertheless, use (14) as a measure of costs, if we change the reference arrival times in (10)-(11) from the scheduled arrival times to the “free running” arrival times, as discussed in the comment following equation (1). The free running arrival time is $l_i$ for Train 1 and $H + t_i$ for Train 2. Substituting these for the scheduled arrival times $T_i$ and $H + T_i$ in (10)-(11) leads to equations (12)-(13), as before.

We may now define the expected total cost for the stochastic case, say $\bar{C}(s)$, as,

$$\bar{C}(s) = E[C(s,K)],$$

where $C(s,k)$ is given by (14). The following simple lemma shows that if the cost function for $C(s,k)$ for a given value of $k$ is concave (convex, monotone) then so is the expected cost $\bar{C}(s)$.

**Lemma 4.1** Suppose that for each $k \in [a, b]$ a function $\psi(\cdot, k)$ is concave (convex, decreasing, increasing, respectively) in an interval $[c, d]$ w.r.t. its first argument. Let $\Psi(s) = E[\psi(s,K)]$, where $K$ is a random variable in $[a, b]$. Then $\Psi(\cdot)$ is concave (convex, decreasing, increasing, respectively) in $[c, d]$.

**Proof:** The standard proof is to write out an inequality defining a given property (e.g., concavity) for a fixed $k$, and take expectations of both sides of the inequality. Due to linear properties of expectation the inequality will be preserved.

To illustrate the practical importance of this lemma, let us return to Example 1 (uniformly distributed initial delay). In this situation the cost $C(s) = C(s,k)$ was proved to be concave in $s$ for any assumed fixed value of $k$. This concavity yields the bang-bang nature of the optimal swapping policy. Lemma 4.1 shows that even if we let the knock-on delay $k$ be a random variable $K$, the expected cost $\bar{C}(s)$ is still concave. This again, as in Example 1, implies a bang-bang optimal swapping policy. Note, of course, that we would have to rework Example 1 to find the new expression for the bang-bang decision rule $\phi$.

### 4.4. Letting both trains be ready-to-depart late

Suppose that Train 2 may be late and, as for Train 1, its initial delay cannot be estimated in advance. Denote the random delay of Train $i$ by $L_i$, and a particular instance of this by $l_i$.

It turns out that we can construct the optimal swapping policy as though $l_i$ (and hence the departure time of Train 2) were known in advance, so that the same results as before apply but with a new “scheduled” departure time for Train 2. To see this, note that no decision about swapping or not swapping need be made or implemented until at least one of the trains is ready to depart. If Train 1 is ready to depart before Train 2 then, because of the simplified knock-on effects assumed in this section, the optimal policy is to let the trains depart in the scheduled order, irrespective of the lateness $l_1$ and $l_2$ of the two trains, and no question of swapping arises. Only if Train 2 is ready to depart and Train 1 is not yet ready (i.e., $H + l_2 > l_1$) does the question of swapping arise. Thus we
can assume that the lateness $l_2$ of Train 2 is known when deciding whether or when to swap (and can assume that the swap time is $s \geq l_2$).

To decide on a swap time $s$ we consider a cost function similar to (2), but because it now depends on how late is Train 2, we denote it by $C(s|l_2)$. Let $d_i(l_i,s|l_2)$ be the lateness of Train $i$ at Station B, given the initial delays $l_i$ and $l_2$ of Trains 1 and 2 and the swap time $s$. An expression for $d_i(l_i,s|l_2)$ can be obtained by revising equations (12)-(13):

$$d_i(l_i,s|l_2) = \begin{cases} 
  s + k & \text{if } s < l_i \leq s + k, \\
  l_i & \text{otherwise}, 
\end{cases}$$

$$d_i(l_i,s|l_2) = \begin{cases} 
  l_i - H & \text{if } H + l_i < l_i \leq s, \\
  s - H & \text{if } s < l_i.
\end{cases}$$

Let $C(s|l_2) = E[c_1 \times d_1(l_1,s|l_2)] + E[c_2 \times d_2(l_1,s|l_2)]$. If $L_1$ and $L_2$ are independent then it can be easily checked that the derivatives $\frac{\partial C_i(s|l_2)}{\partial s}$ and $\frac{\partial^2 C_i(s|l_2)}{\partial s^2}$ are given by the same formulas as in equations (15)-(16). In particular, if $C(\cdot)$ is concave (quasiconcave, increasing, decreasing) then $C(\cdot|l_2)$ is also concave (quasiconcave, increasing, decreasing), so that if any of these properties results in the bang-bang optimal swapping policy for $C(s)$, this effect carries over to $C(s|l_2)$.

### 4.5. Introducing nonlinear cost functions

In Section 3 (which is concerned with the case in which the slower train is scheduled to depart first), it is assumed that the lateness costs $r_1$ and $r_2$ may be nonlinear functions. However, in Sections 4.1-4.4 we have assumed that the cost functions are linear, $r_i(d) = c_i \times d$. We relax this assumption here but will continue to assume that $r_i$ is increasing or at least nondecreasing with lateness as this is the most natural case.

To determine the properties of the optimal swapping criteria $s$, when the delay costs $r_i(\cdot)$ are nonlinear, we need to determine whether the cost function $C(s)$ is everywhere increasing, decreasing, concave, convex, etc. For example, as already discussed in Section 4.1, if $C(s)$ is everywhere increasing or decreasing or concave in $s$, then the optimal swapping policy $s$ is of the bang-bang (swap immediately or never swap) type. For convenience we will assume that $C(s)$ is differentiable, so that the above properties of $C(s)$ are determined by whether $C'(s)$ is positive or negative, and decreasing or increasing.

It is convenient here to split the expected total cost $C(s)$, from (14), into two components, $C_1(s)$ and $C_2(s)$, representing costs associated with Trains 1 and 2, respectively. Then,

$$C(s) = C_1(s) + C_2(s),$$

where,

$$C_1(s) = [F(s + k) - F(s)]r_1(s + k) - [r_1(s + k) - r_1(s)](s),$$

$$C_2(s) = r_2(s - H)[1 - F(s)].$$

To consider properties of $C'(s)$ we consider properties of $C_1(s)$ and $C_2(s)$ separately.

**Proposition 4.2 (A):** If $r_2(\cdot)$ is increasing then $C_2(s)$ is increasing w.r.t. $s$. (B) If $r_2(\cdot)$ is increasing concave then $C_2(s)$ is increasing concave in $s$.

**Proof:** Factor $[1 - F(s)]$ in (23) is nonnegative and decreasing, and $r_2(s - H)$ is nonnegative [and in part (B) is also decreasing]. Hence, $C_2(s)$ is also nonnegative (and decreasing), and $C_2(s)$ is increasing (and concave).

Now consider $C_1(s)$. Rearranging (22) and multiplying through by $k$ gives,

$$C_1(s) \equiv 0 \iff \frac{kr_1(s + k) - r_1(s)}{r_1(s) - r_1(s)} \leq \frac{kr_1(s + k) - r_1(s)}{r_1(s) - r_1(s)} \leq \frac{kF(s)}{F(s + k) - F(s)}.$$
In this equation the ratio \( R_1 = r_1(s + k)/[(r_1(s + k) - r_1(s))/k] \) is the ratio of the slope of \( r_1(\cdot) \) at \( s + k \) to its average slope over the range \([s, s + k]\). Hence this ratio measures the rate of increase of the slope \( r_1' \): the larger is \( R_1 \) the greater the rate of increase of \( r_1' \).

To see this, it is easy to check (from simple diagrams) that if \( r_1 \) is strictly convex, \( R_1 = 1 \) if \( r_1 \) is linear, and \( 0 < R_1 < 1 \) if \( r_1 \) is strictly concave.

Similarly, the ratio \( R_2 = f(s)/[(F(s + k) - F(s))/k] \) measures \( f(s) \), the slope of \( F(\cdot) \) at \( s \), relative to its average value over the range \([s, s + k]\). However, the larger is \( R_2 \) the smaller is the rate of increase, or larger the rate of decrease, of the slope \( f \); hence, \( R_2 \) measures the rate of decrease of \( f \). This dissimilarity of \( R_2 \) and \( R_1 \) is due to \( R_2 \) being measured at the left end of the interval \([s, s + k]\) whereas \( R_1 \) is measured at the right hand end.

Using this interpretation, (24) can be restated in the following informal proposition.

**Proposition 4.3** (informal). Suppose \( r_1'(\cdot) \) is positive. Then \( C_1'(s) \) is positive (or zero or negative, respectively) if the rate of increase of \( r_1'(\cdot) \) is greater than (or equal to, or less than, respectively) the rate of decrease of \( f(\cdot) \).

Propositions 4.2 and 4.3 give conditions under which \( C_1'(s) > 0 \) and \( C_1'(s) < 0 \). If these conditions hold they imply \( C'(s) > 0 \) hence \( C(s) \) strictly increasing. As already noted earlier, an increasing \( C(s) \) implies the optimal \( s \) is \( s = H \), i.e., swap the train immediately the second train is ready to depart.

However, Proposition 3.3 also gives conditions under which \( C_1'(s) < 0 \). If \( C_1'(s) \) is sufficiently negative then \( C(s) = C_1'(s) + C_2'(s) \) may be negative, even though \( C_2'(s) \) is positive (from Proposition 3.2). If \( C'(s) \) is everywhere negative then the optimal policy is never swap the trains.

Finally, if \( C_1'(s) \) is initially sufficiently negative then \( C'(s) = C_1'(s) + C_2'(s) \) can be initially negative. For larger \( s \), \( C_1'(s) \) may become less negative or positive, so that \( C'(s) \) is positive. In this case we have a \( C(s) \) initially decreasing and then increasing, hence having a minimum which is not at either extreme of \( s \), and hence is not of the bang-bang type. As a simple example of this last case (nonbang-bang) we give the following.

**Example 4** (Nonlinear costs may cause a nonbang-bang optimal policy): Suppose that the cost function \( r_1(\cdot) \) for Train 1 remains linear, that is \( r_1(d) = c_1 \times d \). Let the cost function \( r_2(\cdot) \) for Train 2 be as follows. Small delays, say \( d_1(l, s) \leq \delta_0 \), yield no cost, that is, \( r_2(d) = 0 \) for \( d \leq \delta_0 \). For all other delays let the cost be constant, say \( r_2(d) = \gamma \). Consider the density function from Example 1 (a uniform tail).

As can be easily seen, e.g., from (22), \( C_1(s) \) is constant on the interval \( H \to H + \Delta - k \) and then decreasing \( C_2(s) \) is equal to zero for \( s < H + \delta_0 \) and equal to \( \gamma \times \Pr \{ L > H + \delta_0 \} \) for larger \( s \). Suppose that \( \delta_0 > \Delta - k \). The expected total cost of lateness is then illustrated in Fig. 4. It is clear from Fig. 4 that if \( \gamma \) is small then the optimal (cost minimizing) policy is set \( s = H + \Delta \), i.e., never swap trains. If, however, \( \gamma \) is large enough then the optimal value of \( s \) is \( H + \delta_0 \) (nonbang-bang).

The above example assumes for simplicity a two-piece piecewise constant cost function \( r_2(\cdot) \). This can be extended to a continuous increasing cost function, in such a way as

![Graph](image-url)
to retain the above effect—that is, an optimal swap time \( s \) which is not of the bang-bang type. For instance, if \( r_2(\cdot) \) is initially zero (or very small) but eventually rises sharply as the delay of Train 2 increases then the cost function \( C(s) \) is first decreasing and then increasing, hence, not of the bang-bang type.

5. EXTENSIONS TO THE MINIMAL INFORMATION MODEL

The assumption that the only information available on the delay \( L \) of Train 1 is its pdf \( f \) may be rather restrictive. In practice it would usually be possible to say something more about the nature or the cause of the delay, and hence estimate roughly or exactly the ready-to-depart time. This may be possible, for example, because of information forwarded from the previous station regarding the actual departure times and sequence order, or from intermediate timings of train's journey, or reports from station crew servicing the train, etc. In this section we comment on such variations of the minimal information model from Section 2.

5.1. Using new or updated information on delays

Available information may be at least sufficient to significantly modify the pdf \( f \) of the delay \( L \). We may assume, for instance, that there is a discrete choice of pdfs \( f_v \), for \( v = 1, \ldots, V \), and that the information provided by operators from previous stations and other control points is sufficient to determine which \( f_v \) applies. If not, or if the information is not available at the time when the decision on swapping must be taken, then some general pdf \( f \) has to be used, as in Section 2. However, if \( f_v \) is known sufficiently early then the model from Section 2 and results from Sections 3 and 4 hold, with \( f_v \) in place of \( f \).

A variant of the above choice of pdf's is as follows. Suppose that a rough estimate of the delay \( L \) is known. That is, it is estimated that the ready-to-depart time of Train 1 will be say \( l_0 \), with a possible random deviation from this estimate. This deviation may be given by a pdf (e.g., normal) say \( f_{l_0} \) centered around \( l_0 \). If the standard deviation of \( f_{l_0} \) is very small then the estimate \( l_0 \) may be considered as accurate (hence the delay as deterministic) and decisions may be based on this, see Appendix 1. If the random component of the estimate is not negligible then the pdf \( f_{l_0} \) may be treated as a modified density \( f_v \), and the comment above applies.

5.2. Limited advance notice

Suppose now that Train 1 is not ready to depart at time \( l \), but at time \( l \) an accurate estimate \( w \) of the remaining delay becomes available. That is, at a (random) time \( l \) it becomes known that Train 1 will be ready to depart at exactly time \( l + w \), where \( w \) is a constant.

We will adhere to our previous assumption that \( T_i = t_i \) for \( i = 1, 2 \), and see how introducing the estimate \( w \) changes equations (12) (13). To do that, we first have to adjust the meaning of a swapping policy based on a time \( s \). For a given \( s \), we assume that the trains are always swapped if \( s < l \), and are not swapped if \( l + w < s \). This leaves us with the case of \( l \leq s \leq l + w \). For this case we temporarily assume that the trains are not swapped. Finally, to avoid discussing many possible subcases, we assume that \( w < k \) (as before, \( k = t_2 - t_1 \)). Then, making appropriate changes in (12)-(13) gives,

\[
\begin{align*}
\quad d_1 (l,s) &= \begin{cases} 
  l + w & \text{if } -w \leq l \leq s, \\
  s + k & \text{if } s < l \leq s + k - w, \\
  l + w & \text{if } s + k - w < l,
\end{cases} \\

\quad d_2 (l,s) &= \begin{cases} 
  0 & \text{if } l \leq H, \\
  l + w - H & \text{if } H < l \leq s, \\
  s - H & \text{if } s < l.
\end{cases}
\end{align*}
\]
The case \( l \leq s \leq l + w \) needs additional discussion. If the trains are not swapped by time \( l \), the information available changes radically. In fact, at this point we are placed in a "full information" position (the delay of Train 1 becomes known), and hence we are able to calculate the exact costs of swapping or not swapping trains, compare these costs and make decision on this ground. Moreover, if the trains are to be swapped then there is no need for Train 2 to wait until time \( l + w \) (the ready to depart time of Train 1) and it should depart at time \( l \). Suppose that costs yielded by \( s \) are defined by (2) but based constant unit costs \( c_i \), as in Sections 4.1-4.4. To take advantage of the "full information" position we must adjust (2) as follows.

If \( l < s < l + w \) then the cost calculated from (25)-(26), assuming that the trains are not swapped, equals \( c_1(l + w) + c_2(l + w - H) \). However, if we swap the trains (i.e., at time \( l \)), the cost equals \( c_1(l + k) + c_2(l - H) \). The trains should be swapped if the latter cost is less than the former. Note that the difference between the latter and the former, for \( l < s < l + w \), equals \( c_1(k - w) - c_2 w \) and does not depend on \( l \). Depending whether it is negative or positive, the trains should always be swapped or not swapped, respectively. Therefore, (2) should be adjusted by adding the component \( c_\omega \times \Pr\{L < s < L + w\} \), where \( c_\omega = \min\{0, c_1(k - w) - c_2 w\} \). Thus, the expected total cost of lateness given a swap time \( s \) is,

\[
C(s) = c_1 \times \left[ \int_{s-w}^{s} (l + w)f(l)dl + [F(s + k - w) - F(s)](s + k) \right]
+ \int_{s+k-w}^{+\infty} (l + w)f(l)dl + c_2 \times \left[ \int_{s}^{s+k} (l + w)f(l)dl + (1 - F(s))s \right.
- H(1 - F(H)) \left. \right] + c_\omega \times [F(s) - F(s - w)].
\]

The derivative of (27) is,

\[
C'(s) = c_1[F(s + k - w) - F(s) - (k - w)f(s)] + c_2[1 - F(s) + wf(s)] + c_\omega[f(s) - f(s - w)].
\]

A discussion similar to Section 4.1 applies to the properties of the cost (27). Note that the first part of (28) corresponds to the first part of (15) with the constant \( k - w \) substituted for \( k \), and the second part has an additional component \( wf(s) \). The new terms in (28) are \( c_2 wf(s) + c_\omega[f(s) - f(s - w)] \). By definition \( c_\omega = 0 \) or \( c_\omega < 0 \). If \( c_\omega = 0 \) then the new terms in (28) reduce to \( c_2 wf(s) \), which is decreasing if \( f(s) \) is decreasing, hence yielding a concave component in \( C(s) \). If \( c_\omega < 0 \) then the new terms in (28) reduce to

\[
c_2 wf(s) + c_\omega[f(s) - f(s - w)] = c_1(k - w)[f(s) - f(s - w)] + c_2 wf(s - w),
\]

which yields concavity of the corresponding components of the cost \( C(\cdot) \) at a point \( s \) if \( f(\cdot) \) is concave and decreasing on the interval \( s - w \) to \( s \). Thus decreasing concave \( f \) is sufficient (though not necessary) to ensure that the new terms in the cost \( C(s) \) are concave in \( s \). Hence, the likelihood that \( C(s) \) is concave is even stronger than under minimal information (Section 4). And as in Section 4, a concave \( C(s) \) implies that a simple bang-bang (swap immediately or never swap) policy is optimal. The above discussion can also be extended, as in Section 4.3, to allow the parameters \( k \) and \( w \) to be stochastic.

6. MULTIPLE TRAINS AND STATIONS

Consider a situation when several trains travel through a sequence of several stations. Some of the trains may be late at some of the stations. If these delays are significant, it may be that to at least partially recover from them the scheduled order of departure should be changed for some trains at least at some stations.
Multiple trains or multiple stations pose similar difficulties. The optimal procedure would be to set up an “all-in-one” model that would include, as variables to be optimized, the swapping criteria $s_{ij}$ for each pair of trains $i$ and station $j$, thus including all possible knock-on effects of the decision concerning a particular pair of trains at a particular station. However, such a model might prove impossible to solve. Also, collecting and updating information on the constantly evolving pattern of delays and interdependencies on a network would be difficult and costly.

The results and heuristics stated earlier in this paper can be applied to this multiple train multiple station situation if it can be decomposed to a sequence of generic subproblems, each consisting of two neighbouring trains between two neighbouring stations. Multiple stations seem to be easier to handle than multiple trains. Whether trains were or were not swapped at a previous station, their scheduled departure times, sequence order and (random) delays can be considered again at the current station to minimize total expected (costs of) delays at this station and the next. Of course, decisions taken at each station may modify the information available on the initial delays at the next station—comments from Section 5.1 apply. Thus, earlier decisions have an effect on later decisions.

Multiple trains scheduled to depart from a station raise additional questions. If we optimize separately for successive pairs of trains this may yield a suboptimal solution. Fortunately we observe that in practice the question of whether to swap trains tends to arise for individual pairs of trains, where knock-on effects on other trains are small or are of only secondary importance. This gives us confidence that the two-train analysis in the bulk of this paper is widely applicable. If swapping trains would clearly have substantial effects on further trains then the approach taken in this paper would be less useful.

Some problems involving more complex multiple train interactions can be modelled or approximated by a two-train analysis. For instance, one of the situations with strong knock-on effects is a delayed slow train followed by a batch of fast trains. This situation can be dealt with as follows. If the headways separating the fast trains are large then it is likely that the knock-on effects are small enough to be neglected. If, on the other hand, the headways are small, the batch of fast trains can be approximately represented in the analysis as one “very long” fast train with a cost function comprising costs for all trains in the batch.

Also, as explained in Section 5.1, in practice the information currently available about the next pair of trains is usually much better than the information currently available about later trains. This suggests concentrating the analysis on the next pair of trains, as in most of this paper.

7. CONCLUDING REMARKS

If a scheduled service is delayed it may delay following services, and net delay costs can often be reduced by swapping the sequence order of the services. This paper deals with this widespread problem, and with deciding whether and when to let a service start if the service scheduled to start before it is delayed. We discuss this swapping problem in the context of scheduled passenger train services.

We obtain some interesting and sometimes surprising results. For example:

(a) We show that the optimal (delay cost minimizing) swapping policy depends on the amount of information available. Hence, we consider three different typical levels of information available (in Sections 2-5, 5.2, and the Appendix 1).

(b) For the “minimal information” case (Section 2), if the slower service is scheduled to depart first, then the optimal (delay cost minimizing) time $s$ at which to swap services is independent of the form of the probability distribution of the delays. This is a very useful and important result as there is often only poor information available on the form of the probability distribution of delays (see Section 3).

(c) For the “minimal information” case we find that, if the faster service is scheduled to depart first, then the optimal swapping policy is likely to be a bang-bang type. That
is, either “swap immediately” or “never swap.” Swap immediately means, always let the second train depart when it is ready and do not delay it to allow the first train more time to depart or to become ready to depart. Never swap means never let the second train depart before the first train, no matter how long the first train is delayed (see Section 4).

In this case, which of the bang-bang policies (swap or no-swap) is optimal does depend on the probability distribution of delays. However, by analysing the effects of three quite different families of probability distributions on the choice of optimal swap time s, we find that the decision to swap or not swap is almost independent of the form of the probability distributions: it depends only on the “mean of the tail” of the distribution, which can be known or estimated without knowing the shape of the distribution. We are thus able to derive a quite robust heuristic decision rule which requires little information about the form of the probability distribution of delays.

Acknowledgements—This research was supported by U.K. Science and Engineering Research Council (SERC) Grants GR/F/91407 and GR/H/50432. We also thank British Rail and the UK Fellowship of Engineering for their support.

REFERENCES


Carey M. (1992). Reliability of interconnected scheduled services with resequencing allowed. N. Ireland: Research Report, Faculty of Business and Management, University of Ulster.


APPENDIX 1: TRAIN SWAPPING WITH FULL INFORMATION

Suppose, in contrast to Section 2, that the ready-to-depart time of Train 1, and hence the lateness l of this, is known prior to the scheduled departure time of Train 1. However, suppose also that the free running trip times t from A to B are subject to random variations, which are not known in advance. As in Section 7, let $K_{ij}(h)$ be the actual trip time of train $j$ including any knock-on delays from train $i$ if the latter departs $h$ units of time before it. In this section, we allow the knock-on trip times $K_{ij}$ to have an arbitrary form and depend on any random variables. The only assumption we make is that, for any particular values of the random variables on which they depend, the $K_{ij}(h)$ are nonincreasing w.r.t. $h$. 

In view of these assumptions, the operator can make a fully informed decision at the scheduled departure
time of Train 1 as to when to let the trains depart: waiting any longer to decide will not bring any further
information. Thus, by the scheduled departure time of Train 1 the operator should specify whether (a) the
original order of trains is to be preserved, that is, Train 1 is to depart as soon as it is ready (at time $l$) and Train
2 is to depart immediately after it (but not before its scheduled departure time), or (b) the trains are to be
swapped, that is, Train 2 is to depart on time and Train 1 after it, when it is ready.

Let $d_{i\text{swap}}^1(l)$ be the actual lateness of train $i$ at station B if the initial delay of Train 1 is $l$ and the trains are
swapped: lateness is measured w.r.t. the scheduled arrival time. Similarly, let $d_{i\text{no swap}}^0(l)$ be the actual lateness
of train $i$ if the trains are not swapped. These can be stated explicitly in terms of the notation introduced above,

\[
\begin{align*}
d_{1\text{swap}}^1(l) &= (l + t_1 - T_1)^+, \\
d_{2\text{swap}}^1(l) &= ((l - H)^+ + K_{12}((H - l)^+) - T_1)^+, \\
d_{1\text{no swap}}^0(l) &= \max(l,H) + K_{21}((l - H)^+) - T_1)^+, \\
d_{2\text{no swap}}^0(l) &= (t_2 - T_2)^+,
\end{align*}
\]

where $(x)^+ = \max\{0,x\}$. The expected total costs of lateness is then,

\[
\begin{align*}
C_{\text{no swap}}^0(l) &= Er_1(d_{1\text{no swap}}^0(l)) + Er_2(d_{2\text{no swap}}^0(l)), \\
C_{\text{swap}}^1(l) &= Er_1(d_{1\text{swap}}^1(l)) + Er_2(d_{2\text{swap}}^1(l)),
\end{align*}
\]

where the expectations have of course been taken over all values of the random variables $t_1$ and $t_2$.

The optimal (cost minimizing) swapping policy is therefore: allow the trains to depart in their original order
(no swap) if $C_{\text{no swap}}^0(l) < C_{\text{swap}}^1(l)$, and swap the trains if this inequality is reversed. Note that the expected
costs $C_{\text{no swap}}^0(l)$ and $C_{\text{swap}}^1(l)$ depend on the lateness $l$ of Train 1, which is assumed here (under "full informa-
tion") to be known in advance. It follows that the sign of $[C_{\text{no swap}}^0(l) - C_{\text{swap}}^1(l)]$ and hence the decision as to
whether to swap or not depends on $l$. Intuitively, it would seem likely that if it is known that Train 1 will be only
a small amount late ($l$ small) then we should not let Train 2 go ahead of it (i.e., no swap). Conversely, it would
seem likely that if it is known that Train 1 will be very late ($l$ large) then we should let Train 2 depart when it is
ready and not wait for Train 1 (i.e., swap immediately). The following proposition shows that, under mild
assumptions, this intuition is correct, and that there is a cut-off point $s$ (see Fig. 5) such that if Train 1 will be
later than $s$ ($l > s$) then swap immediately and if it will not be as late as $s$ ($l < s$) then do not swap.

**Proposition A.1** (A) Suppose that (a) $t_1 + H \geq T_1$, and that the cost functions $r_i$ are linear and nondecreasing,
for $i = 1,2$. Then,

(i) $[C_{\text{no swap}}^0(l) - C_{\text{swap}}^1(l)]$ is nondecreasing in $l$, and

(ii) there exists a finite $s \geq 0$ such that $C_{\text{no swap}}^0(l) \leq C_{\text{swap}}^1(l)$ for all $0 \leq l < s$, and $C_{\text{no swap}}^0(l) \geq C_{\text{swap}}^1(l)$ for
all $l > s$.

(B) Suppose that (a) $t_1 + H \geq T_1$, (b) $t_2 \leq K_{21}(h)$, (c) $r_1(\cdot)$ is nondecreasing, and (d) $r_2(\cdot)$ is nondecreasing
concave. Then (i)-(ii) from (A) again hold but in part (ii) may not be finite.

**Proof:** For short, denote $d_1^i = d_{1\text{swap}}^i(l)$ and $d_2^i = d_{2\text{swap}}^i(l)$. For part (i), it follows from (29) that it is
sufficient to prove that for $i = 1,2$ the difference $[r_i(d_2^i) - r_i(d_1^i)]$ is nondecreasing w.r.t. $l$, for any fixed values
of $t_1$ and $t_2$. For $i = 2$ this is straightforward, because $d_2^i$ is constant, and $d_1^i$ is nondecreasing since $K_{21}(h)$
is assumed nonincreasing in $h$. Hence, from linearity of $r_2$ in part (A) or assumption (c) in part (B), $[r_2(d_2^i) -
r_2(d_1^i)]$ is nondecreasing.

For $i = 1$, if $l \leq H$ then $d_1^i$ is nondecreasing and $d_2^i$ constant. If $l > H$ then from the assumption (a) it
follows that $[d_1^i - d_1^i] = [t_1 - K_{21}((l - H)^+)]$, which is nondecreasing because $K_{21}(h)$ is assumed nonin-
creasing in $h$. In part (A) this is sufficient for $[r_1(d_2^i) - r_1(d_1^i)]$ to be nondecreasing. For part (B), note that both $d_2^i$
and $d_1^i$ are also nondecreasing in $l$, and from (b) that $d_2^i \leq d_1^i$. This together with the assumption (d) implies
that $[r_2(d_2^i) - r_2(d_1^i)]$ is nondecreasing.

Part (ii) follows easily from part (i): because $[C_{\text{no swap}}^0(l) - C_{\text{swap}}^1(l)]$ is nondecreasing in $l$, it intercepts the

![Fig. 5. Critical lateness s. If l > s, swapping reduces costs.](image-url)
horizontal axis at most once, hence (ii). The finiteness of $s$ in part (A(ii)) follows from the observation that when $l \to +\infty$ then for $i = 1$, $[d_i^* - d_i]$ converges to a constant, while for $i = 2$ it tends to infinity. This implies $[C^{s\to\infty'}(l) - C^{s\to\infty'}(d)] \to +\infty$ as $l \to +\infty$, hence it is optimal to swap before $l = +\infty$.

The proposition shows that the cost functions $C^{s\to\infty'}(l)$ and $C^{s\to\infty}(l)$ intersect only once, as in Fig. 5. Hence, to decide which train to let go first, operators only need to determine whether the ready-to-depart time of Train 1 is going to be later than this single critical lateness $s$.

Remarks on the assumptions in Proposition A.1. Part (B) of the proposition allows the cost functions $r_i$ to be nonlinear but introduces assumption (b). However, assumption (b) is trivial—it merely states that any knock-on effects from a train can only delay rather than speed up the following train.

The proposition assumes (in (a)) that $t_i + H \geq T_i$ must hold for all values of random trip time $t_i$. This can be rewritten as $H \geq (T_i - \min(t_i))$. The term $(T_i - \min(t_i))$ is the amount by which the scheduled trip time exceeds the minimum free running trip time. This is sometimes referred to as the "slack" or "recovery" time in the timetable, and timetables are often constructed (or can be constructed) so that the headway $H$ exceeds this so that the above assumption is likely to be widely satisfied in practice. Also note that the above assumption is sufficient but not always necessary for the proposition to hold. Moreover, the property actually shown in the proof of the proposition is much stronger than in the statement of the proposition.

The monotonicity of the cost functions $r_1$ and $r_2$, assumed in (c) and (d) of (B), is a natural assumption, because larger delays usually imply larger costs. The concavity of $r_1$ may reflect decreasing sensitivity of the system to increments of delay. For example, an increase in the delay from 5 to 10 minutes may result in a significant increase in costs, while an increase from 55 to 60 minutes may cause only a small or negligible increase in costs.

APPENDIX 2: NOTATION USED IN THE PAPER

This appendix lists the notation used and already defined earlier in this paper.

$E = \text{the symbol for mathematical expectation.}$

$i = \text{train number.}$

$T_i = \text{scheduled trip time of train } i \text{ on link AB.}$

$H = \text{scheduled departure headway between two trains.}$

$t_i = \text{free running trip time of train } i \text{ on link AB.}$

$K_i(h) = \text{actual trip time of train } j \text{ (including knock-on effects from the previous train) if train } i \text{ departs } h \text{ minutes (or other units of time) before it } (K_i(h) = t_i \text{ if there are no knock-on effects).}$

$s = \text{a decision variable representing a prespecified (optimal) "swap" time for a pair of trains.}$

$W = \text{an accurate estimate of the remaining delay in the ready-to-depart time of Train 1; used only in Section 5.2.$

$d_i(l,s) = \text{lateness of arrival of train } i \text{ at the destination B, given the delay } l \text{ in the ready-to-depart time of Train 1 and swap time } s.$

$d_i(l,s) = \text{same as } d_i(l,s) \text{ but both trains may be late.}$

$d_i(l,s) = \text{similar to } d_i(l,s) \text{ but both trains may be late.}$

$d_i(l,s) = \text{an accurate estimate of the remaining delay in the ready-to-depart time of Train 1; used only in Section 5.2.$

$C(s) = \text{expected total cost of lateness at the destination B, given the departure order of trains.}$

$C(s) = \text{same as } C(s), \text{ if the dependence on } k \text{ needs to be emphasized.}$

$C(s) = \text{same as } C(s), \text{ given delay } l_1 \text{ in the ready-to-depart time of Train 2.$

The following notation is used in Section 4.2.

$\Delta = \text{the range on which } f(l) \text{ is defined (in Examples 1 and 2).}$

$p = \text{probability density function for the ready-to-depart time of a train.}$

$q = \text{cumulative probability function corresponding to } f(l).$