OPTIMISATION TECHNIQUES
FOR REDUCING DELAY TO TRAFFIC
IN SIGNALISED ROAD NETWORKS

by

Richard Edward Allsop

A thesis submitted for the Degree of Doctor of Philosophy

Research Group in Traffic Studies
University College London

June 1970
ABSTRACT

Traffic signals are commonly used to control traffic at intersections on busy roads, and their operation is controlled wholly or partly by preset schedules. Signals at neighbouring intersections can be linked to co-ordinate their operation. This thesis surveys previous methods of calculating settings both for signals at one intersection and for linked signals, and describes new methods.

For signals at one intersection, a survey is made of previous theoretical studies of the relation between delay to traffic passing through the intersection, the rates at which traffic on each approaching road arrives and can depart, and the signal settings. Previous methods of calculating settings are shown to be unsatisfactory in some respects. A new method is developed in which settings are calculated, subject to certain practical constraints, to minimise the estimated delay to all traffic passing through the intersection. Such settings are shown to exist, provided that the intersection is not overloaded with traffic, and to be unique. A computer program for calculating the settings is described.

Previous methods of calculating settings for linked signals in networks of roads comprise two parts: a model of traffic movement is used to estimate the delay to traffic in the network in terms of the signal settings, and an optimisation technique is used to choose settings to minimise the estimated delay. The traffic model from a previous method is used, together with
the new method of calculating settings at a single intersection and a new optimisation technique, to provide an improved method of calculating settings for signals in all but exceptionally large networks. The necessary computer programs are briefly described.

Further research is suggested, and a classified bibliography of related work is appended.
ACKNOWLEDGEMENTS

The work described in this thesis was carried out in the Research Group in Traffic Studies at University College London under the supervision of Professor R.J. Smeed. The author is most grateful to Professor Smeed for his constant guidance and advice, and to his colleagues, especially Mr. J.G. Wardrop and Dr. A.W. Evans, for their help and encouragement. He is also grateful to K.W. Huddart and E.D. Turner, of the Greater London Council Department of Highways and Transportation, for making available a computer program, to A.P. Howes of the same Department, J. Taylor of the London Borough of Camden Engineer and Surveyor's Department, and D.G. Abraham, formerly of International Computers Limited, each of whom assisted in the implementation of one of the author's computer programs, and to Mrs. G. Lunt and Mrs. D.E.E. Turner for typing and the preparation of diagrams.

Thanks are due to International Computers Limited for supporting the work financially.
CONTENTS

Abstract

Acknowledgements

1. Introduction definitions and notation
   1.1 Introduction
   1.2 Definitions and notation
      1.21 Rule of the road and sequence of signal aspects
      1.22 Probability notation
      1.23 Traffic signals at a single intersection
      1.24 Notation for studies of a single intersection
      1.25 Systems of linked signals
      1.26 Notation for linked signals

2. Theoretical studies of fixed-time signals at a single intersection
   2.1 Some aims of the studies
   2.2 The arrival and departure of traffic
      2.21 Regular arrivals
      2.22 Binomial arrivals
      2.23 Poisson arrivals
      2.24 Some other arrival models
      2.25 Departure models
      2.26 The overflow
2.3 Expressions for delay: regular arrivals 37
2.4 Expressions for delay: binomial arrivals 41
2.5 Expressions for delay: Poisson arrivals 44
   2.51 Webster's expression 45
   2.52 Other work related to Poisson arrivals 50
2.6 Expressions for delay: more general arrival models 55

2.7 Calculation of signal settings 62
   2.71 The stage matrix 62
   2.72 The cycle time 63
   2.73 Allocation of green time 66
   2.74 Comment 68

3. Calculation of delay-minimising settings for a single fixed-time traffic signal 69

3.1 Preliminaries 69
   3.11 Choice of expression for delay 69
   3.12 Lost time and effective green times 70
3.2 Mathematical statement of the problem 73
   3.21 Rate of delay 74
   3.22 Capacity constraints 75
   3.23 Cycle time constraint 75
   3.24 Green time constraints 76
   3.25 Example 79
   3.26 Summary 80

3.3 Existence and location of feasible solutions 83
   3.31 A simple method that often finds a feasible solution 83
   3.32 A general method 86
3.4 Existence and uniqueness of delay-minimising settings
   3.41 Existence 88
   3.42 Uniqueness 91
3.5 An algorithm for finding delay-minimising settings 101
   3.51 Choice of direction 102
   3.52 Choice of distance 107
   3.53 Convergence 116
   3.54 Possible effect of the third term in Webster's expression 121
3.6 Computer program 123
3.7 A few illustrative results 125
   3.71 Calculations for three examples 127
   3.72 Comments 135
3.8 Effect of changing a specified cycle time 135
4 Theoretical studies of linked traffic signals 139
4.1 The time-distance diagram 139
   4.11 The basic diagram 139
   4.12 Transformations of the diagram 142
   4.13 An invariance property 143
   4.14 Through bands and bandwidth 143
4.2 Settings for linked signals on one main road 146
   4.21 A one-way road 146
   4.22 A two-way road 151
4.3 Settings for linked signals in a network of roads
4.31 Network geometry
4.32 Some studies relating to special cases
4.33 More general studies
4.34 Calculation of the performance index
4.35 Choice of signal settings
4.36 Comparative results
4.4 Comment

5. An improved method of calculating settings for linked traffic signals
5.1 Choice of cycle time
5.2 Allocations of green time
5.3 Calculation of the performance index
5.4 Choice of offsets
  5.41 Order of choice
  5.42 Dynamic programming procedure
  5.43 Applicability
5.5 Computer programs

6. Synopsis
6.1 Delay at a fixed-time traffic signal
6.2 Settings for a single fixed-time traffic signal
6.3 Settings for linked traffic signals
6.4 Further work
7. Suggestions for further work

References

Appendix: Classified bibliography of related work

A.1 Practical work related mainly to fixed-time signals at one intersection

A.2 Practical work related to linked fixed-time signals

A.3 Studies of vehicle-actuated signals

A.4 Studies of signals at overloaded intersections

A.5 Studies of the behaviour of traffic platoons

A.6 Some relevant simulation studies
1. **INTRODUCTION DEFINITIONS AND NOTATION**

1.1. Introduction

Traffic signals are commonly used to share the carriage-way at busy intersections between conflicting traffic streams. There are two main types of control for traffic signals: fixed-time signals operate always to a preset schedule, whilst the operation of vehicle-actuated signals is controlled partly by vehicles approaching the intersection. Like fixed-time signals, however, vehicle-actuated signals require a preset schedule to which they can operate when there is heavy traffic on all roads approaching the intersection, or when faults develop in the vehicle actuation apparatus. Signals at neighbouring intersections can be linked so that their operation is co-ordinated.

Both for traffic signals at one intersection considered in isolation and for systems of linked signals, the calculation of preset schedules so as to reduce the delay to traffic using the intersections has been the subject of extensive theoretical and practical studies. In this thesis, such theoretical work is reviewed, and then extended in ways that are likely to find early practical application. Detailed surveys of the literature relating to practical studies and to theoretical problems other than the calculation of preset schedules are beyond the scope of this thesis, but a classified bibliography is appended.

1.2. Definitions and notation

Previous authors have used a variety of notations, and some terms have not always been used in the same sense.
Definitions and notation are set out here, in terms of which both earlier work and the author's new work will be described. Only a few of the concepts are new; the sources of the remainder are given in Chapters 2 and 4. In order to give clear definitions, it is sometimes necessary to confine attention to particularly simple cases. When clear definitions have been obtained in this way, however, it is not usually very difficult to apply them to the less straightforward cases that often arise in practice.

1.21. Rule of the road and sequence of signal aspects. It will be assumed throughout that in two-way streets vehicles keep to the left; this means that, at a cross-roads, right-turning vehicles have to cross the path of oncoming traffic and left-turning vehicles do not. It will also be assumed that the sequence of aspects shown by any traffic signal is red, red-and-amber, green, and amber.

1.22. Probability notation. The letter $P$ will be used to denote probability; if $A$ and $B$ are statements, then

\[
P(A) = \text{probability that } A \text{ is true}
\]

\[
P(A|B) = \text{probability that } A \text{ is true, given that } B \text{ is true}.
\]

The letter $E$ will denote expectation, so that if $X$ is any random variable,

\[
E(X) = \text{the expectation, or mean, of } X.
\]

1.23. Traffic signals at a single intersection. A road leading to a signalised intersection may be such that all
vehicles waiting there to enter the intersection from one queue; alternatively, it may be possible for the vehicles to form two or more separate queues according to the direction they intend to take at the intersection. In the former case the road will be said to form one approach to the intersection, whilst in the latter case it comprises as many approaches as there are separate queues.

It should be noted that an approach may consist of several traffic lanes, because where two or more lanes are available for vehicles intending to take the same direction at the intersection, these vehicles tend to behave as a single queue waiting two or more abreast, rather than as several separate queues.

DEFINITIONS: An approach to a signalised intersection is an area of road leading to the intersection and such that any vehicles waiting there to enter the intersection form a single queue.

An approach has right of way when traffic from that approach is allowed to enter the intersection.

In order to define the delay to a vehicle passing through the intersection from a certain approach, suppose that there is a point on the approach so far from the signal that, even if the signal is red, approaching vehicles have not yet begun to slow down when they pass the point. Suppose also that there is another point beyond the intersection such that vehicles that have had to stop at the signal have ceased to accelerate when they pass this second point. Then let the time taken to travel between these two points be called the
transit time, and let the second point, which will be referred to again in Section 1.25, be called the exit point.

**DEFINITION:** The *delay* to a vehicle passing through the intersection from a certain approach is the difference between its transit time and the value that its transit time would take if that approach always had right of way and the intersection were not obstructed by turning traffic.

The purpose of traffic signals is to regulate the times at which the various approaches have right of way. Considerations of safety and convenience usually determine whether any two approaches may have right of way simultaneously.

**DEFINITION:** Two approaches are *compatible* if they may have right of way simultaneously.

At a fixed-time signal, two or more sets of mutually compatible approaches have right of way in turn in cyclic order for predetermined times, and one repetition of this process is called a signal cycle.

**DEFINITIONS:** The *cycle-time* of a fixed-time signal is the duration of one signal cycle, and a *stage* in the cycle is that part of the cycle for which one set of approaches receives right of way.

Suppose that an intersection has $n$ approaches, numbered arbitrarily from 1 to $n$, and that the signal cycle has $m$ stages numbered in the order of their occurrence in the cycle. Then the order in which the approaches receive right of way can be specified by an $m \times n$ matrix $A$. 
DEFINITION: The stage matrix is an m x n matrix $A = (a_{ij})$ such that

$$a_{ij} = \begin{cases} 1 & \text{if approach } j \text{ has right of way in stage } i \\ 0 & \text{otherwise} \end{cases}$$

One restriction is usually imposed in practice on this matrix: if the same approach has right of way in two or more stages, then these stages must be consecutive (stage $m$ and stage $1$ being regarded as consecutive).

Consider now just one approach; the restriction just imposed on the stage matrix means that the approach has right of way for just one period in each cycle. Thus the signal controlling the approach will show red, red-and-amber, green, and amber once each in turn during the cycle. It is found, however, that, when estimating the delay to traffic on one approach, the cycle may be regarded as comprising simply a period of effective red time followed by a period of effective green time. The exact specification of these periods in relation to those for which the signal shows red and green is discussed in Section 2.25; until then, the following qualitative definitions will suffice.

DEFINITIONS: The effective red time for an approach is the duration of the part of the cycle during which, when estimating delay, it is assumed that no traffic passes the signal. The effective green time is the remainder of the cycle time, during which, when estimating delay, traffic is assumed to pass the signal at a constant rate so long as there are vehicles waiting on the approach. The saturation flow for the approach is this constant rate at which traffic is assumed to
pass the signal during the effective green period while there are vehicles waiting on the approach.

The exact relation between the saturation flow and the actual movement of traffic on the approach is discussed in Section 2.25.

When estimating delay the traffic is regarded as consisting of identical vehicles equivalent in size and performance to an average passenger car. Traffic flows are expressed in terms of these passenger car units (p.c.u.), and actual vehicles of various types are regarded as equivalent to different numbers of p.c.u.'s. The saturation flow corresponds to the average flow observed on the approach when the signal has been showing green for several seconds but there are still vehicles waiting. The relation between the effective green time and the time for which the signal shows green depends mainly on two things. The first is the relation between the saturation flow and the flows observed shortly after the signal changes to and from green. The second is the lengths of time for which the signal shows amber and red-and-amber; these are usually determined by considerations of safety in relation to the layout of the intersection. The effective green time usually exceeds slightly the time for which the signal shows green, but by less than the sum of the times for which it shows amber and red-and-amber.

Consider again the whole signal cycle and assume, for the present, that for each stage it is possible to choose one approach that has right of way in that stage only. Then the
sum of the effective green times for the chosen approaches is usually less than the cycle time by an amount that is called the lost time. In Chapter 3 the concept of lost time will be developed to cover intersections that do not satisfy the assumption made at the beginning of this paragraph.

1.24. Notation for studies of a single intersection. The following notation will be used when considering a single approach to a traffic signal; when a whole intersection is being considered, subscripts will be used to distinguish between approaches.

Let $c =$ cycle time in seconds

$g =$ effective green time in seconds

$r =$ effective red time in seconds

$q =$ average arrival rate of traffic on the approach in p.c.u./second

$s =$ saturation flow on the approach in p.c.u./second

$d =$ average delay to p.c.u.'s on the approach, in seconds

$L =$ lost time in seconds

$\lambda = \frac{g}{c},$ i.e. the proportion of the cycle that is effectively green

$y = \frac{q}{s},$ i.e. the ratio of average arrival rate to saturation flow

$x = \frac{qc}{gs},$ i.e. the ratio of the average number of arrivals per cycle to the maximum number of departures per cycle.

Then $r + g = c$ and $\lambda x = y$
DEFINITIONS: \( x \) is the **degree of saturation** of the approach 
\( y \) is the **flow ratio** of the approach

In a number of theoretical studies it has been assumed that events such as the arrival of a vehicle occur only at certain instants equally and quite closely spaced in time. This will be called the **discrete time assumption**.

Let \( \Delta t \) denote the interval in seconds between two successive such instants. Then it is found convenient to make one such instant the time origin \( t = 0 \) so that the others are at times \( t = \pm n\Delta t \) \((n = 1, 2, 3, \ldots)\), to choose \( \Delta t \) so that

\[
\Delta t = \frac{1}{S},
\]

and to assume that the effective red and green times, and therefore the cycle time, are multiples of \( \Delta t \).

Let 
\[
c = C\Delta t \]
\[
r = R\Delta t
\]
\[
g = G\Delta t, \text{ where } C, R \text{ and } G \text{ are integers},
\]
and let 
\[
\alpha = \text{average number of p.c.u.'s arriving in time } \Delta t,
\]
so that 
\[
\alpha = q\Delta t.
\]

1.25. **Systems of linked signals.** When calculating schedules for the co-ordinated operation of linked traffic signals, it is not usually practicable to take into account all the traffic movements in the road network, especially if there are many minor roads not controlled by signals. It is therefore useful to have a simple representation of the network, including just those traffic movements that are to be taken into account in calculating the signal schedules. One such
representation is by means of a graph (Berge 1962). Suppose that signals at \( n \) intersections \( I_1, I_2, \ldots, I_n \) are to be linked; then these intersections are represented by vertices \( A_1, A_2, \ldots, A_n \) respectively in the graph. If traffic flows from intersection \( I_j \) to some approach of intersection \( I_k \) without passing through any of the other intersections \( I_i \), and if this flow of traffic is to be taken into account in calculating the signal schedules, then vertices \( A_j \) and \( A_k \) of the graph are joined by an arc in the direction of the traffic flow. For example, a two-way street joining two adjacent signalised intersections, and forming just one approach to each of them, is represented by two arcs, one in each direction, joining the corresponding vertices of the graph.

Associated with the graph is an \( n \times n \) matrix \( M = (m_{jk}) \) such that \( m_{jk} \) is the number of arcs from \( A_j \) to \( A_k \).

For each arc from \( A_j \) to \( A_k \) in the graph, the length of road used by the corresponding traffic flow from intersection \( I_j \) to intersection \( I_k \) will be called a link in the road network. The signals at \( I_j \) and \( I_k \) respectively will be called the upstream and downstream signals for that link.

It will be convenient to associate delay to traffic in the network with the particular links on which the delay occurs. Consider a vehicle passing along a link and let the time it takes to travel from the exit point (see section 1.23) of the upstream signal to that of the downstream signal be called the transit time for that vehicle on that link.
DEFINITIONS: The delay to a vehicle on a link is the difference between its transit time and the value that the transit time would take if traffic on that link always had right of way at the downstream signal and the downstream intersection were not obstructed by turning traffic. The rate of delay on a link is the average delay incurred per unit time by traffic on the link.

When estimating the delay to traffic on a network of roads it is often impracticable to consider the movement of individual vehicles. The problem is simplified by the fact that, when signals are close together, vehicles on each link tend to travel in bunches as they are released by the upstream signal. Such bunches of vehicles are called platoons, and mathematical representations of the observed behaviour of platoons can be used to estimate the delay to vehicles in them.

The aim of linking traffic signals is usually to reduce both the delay to traffic passing through them and the number of occasions on which vehicles have to stop. The signal schedule should therefore be such that, as far as possible, a platoon released by the upstream signal on a link reaches the downstream signal while the approach on which it arrives has right of way. The calculation of a schedule for a system of linked signals must therefore take into account the time taken to travel along each link. This is usually achieved by specifying for each link a design speed which is representative of the speeds of undelayed vehicles travelling along the link, and variation between the speeds of individual vehicles is allowed for, to some extent, in the representation of platoon behaviour.
So that the relationship between the schedules of individual signals in a linked system may remain the same in successive cycles, it is usual for all the signals in a system to have a common cycle time. Differences in cycle time can be accommodated, however, without making the calculation of schedules more difficult in principle, provided that all the cycle times are submultiples of some common value. A simple case of this kind occurs, for example, if a few of the linked signals have cycle times that are one half of the common cycle time of the remaining signals.

**DEFINITION:** The *master cycle* for a system of linked signals is the lowest common multiple of the cycle times of the individual signals in the system.

It is useful to regard successive repetitions of the master cycle as providing a time scale on which the starting points of the cycles of the individual signals can be marked. This concept has its physical counterpart in the master controller to which the timing devices of individual signals in a linked system are connected. It will usually be convenient to regard the beginning of the effective green time for stage number 1 of an individual signal cycle as the starting point of that cycle, although any specified point in the cycle will serve equally well. The timings of the individual signal cycles relative to the master cycle are determined by the offsets, which are defined as follows:

**DEFINITION:** The *offset at intersection* $I_j$ in a system of linked signals is the time from the start of a master cycle to the next starting point of the signal cycle at intersection
I_j \ (j = 1, 2, ..., n).

The following closely related definition will be useful in considering the behaviour of traffic on individual links in the network.

DEFINITION: The offset on a link of the network is the time from the starting point of a cycle at the upstream signal to the next starting point of a cycle at the downstream signal.

Several methods of calculating offsets for linked signals assume that, for given master cycle, traffic flows, and allocations of green time at individual signals, the estimated rate of delay and average number of p.c.u.'s stopping per unit time on each link of the network can be expressed as a function of the offset on that link. Such functions are called the delay-offset relation and the stops-offset relation for the link.

The schedule for a system of linked signals is completely determined by the master cycle and the offsets and allocations of green time at the individual intersections; these quantities are called the signal settings.

1.26. Notation for linked signals. The following notation will be used when considering linked traffic signals.

Let \[ T = \text{duration of master cycle in seconds}, \]
\[ n = \text{number of intersections where signals are linked}, \]
\[ x_j = \text{offset at intersection } I_j \ (j = 1, 2, ..., n), \]
\[ z_{jk} = \text{offset on link from } I_j \ \text{to} \ I_k, \]
so that \[ z_{jk} = x_k - x_j \pmod{T} \]
Let $d_{jk}(z_{jk})$ be the delay-offset relation for a link from $I_j$ to $I_k$, 
$s_{jk}(z_{jk})$ be the stops-offset relation for a link from $I_j$ to $I_k$, 
$v_{jk}$ = the design speed for a link from $I_j$ to $I_k$, 
t_{jk} = travel time from $I_j$ to $I_k$ at the design speed.

The offsets and travel times will be measured in seconds except where, in some methods of calculating settings for linked signals, the master cycle is divided into a fixed number (usually 50) equal intervals. Where this is the case and

$N = \text{the number of intervals into which the master cycle is divided,}$

the offsets and effective green times at the individual signals, and travel times on the links, are measured in units of $T/N$ and are required to be integer multiples of this quantity. This will be called the network discrete time assumption.
2. THEORETICAL STUDIES OF FIXED-TIME SIGNALS
AT A SINGLE INTERSECTION

2.1. Some aims of the studies

The engineer who is calculating a schedule for traffic signals at one intersection considered in isolation must decide in what order and for what proportions of the signal cycle the approaches are to have right of way, and how long the cycle is to last. In terms of the quantities defined in Chapter 1, he must determine the stage matrix, the cycle time, and, for each approach, the effective green time; these quantities are not independent of each other.

The stage matrix is usually determined by practical considerations, though it has been the subject of one theoretical study (Stoffers 1968), which is discussed in section 2.7. Various criteria have been used in the choice of cycle time and effective green times, but the commonest has been to minimise the estimated average delay to traffic passing through the intersection. For this reason, and because of the theoretical interest of the problem, the aim of many theoretical studies has been to express the average delay to traffic on one approach to a fixed-time signal in terms of the cycle time and the saturation flow, effective green time, and average arrival rate on that approach; such studies are discussed in sections 2.2 - 2.6. The application of the resulting expressions to the calculation of signal schedules has been the aim of further work, which is discussed in section 2.7, together with studies in which other criteria have been used in calculating schedules.
2.2. The arrival and departure of traffic

In order to analyse mathematically the behaviour of traffic passing through a signalised intersection, it is necessary to construct mathematical models describing how the arrival and departure of traffic on each approach vary with time. If the results of the analysis are to be useful in practice, the models must be reasonably consistent with observed traffic behaviour.

To produce a simple departure model it is convenient to regard the traffic as departing one passenger car unit at a time, and this can be made realistic by the appropriate determination of the numbers of p.c.u.'s to which actual vehicles are equivalent (see section 2.25). Most theoretical studies assume, however, that the traffic arrives in the same units as those in which it departs, so that the arrival model must describe the arrival of p.c.u.'s, rather than vehicles; this has to be borne in mind when assessing the realism of the various arrival models that have been proposed. In all models, the average arrival rate can be given any value observed in practice. Another quantity that can be measured in practice on any approach and compared with the theoretical values implied by the various models is the ratio

\[
I = \frac{\text{variance of number of p.c.u.'s arriving in one signal cycle}}{\text{mean }}
\]

first considered by Miller (1963).

The meaning of arrival time and departure time for a p.c.u. on an approach can be made precise with the aid of Figure 1, in which the distance travelled is plotted against
Fig. 1. DIAGRAM TO ILLUSTRATE THE DEFINITIONS OF SUPPOSED ARRIVAL AND DEPARTURE TIMES
time for each of four cars (each equivalent to one p.c.u.) passing through the intersection from the approach being considered. The straight line AB represents the passage of the front of an undelayed car; the distance scale is assumed to be adjusted, if necessary, to make this trajectory a straight line. The horizontal line PQ marks the position of the stop line, where the first vehicle waits when there is a queue at the signal. The trajectory CDEF represents the passage of the front of a car that has to wait at the stop line; the straight portions CD and EF are parallel to AB, and when they are produced to meet PQ at X and Y it follows from the definition in section 1.23 that the length XY on the time scale is the delay to the car. In most of the studies discussed here, X is regarded as the arrival time and Y the departure time of this car, so that the delay is the difference between these supposed arrival and departure times. The other two trajectories represent the passages of a car that stops further back in the queue and a car that is delayed but not stopped. The supposed arrival and departure times of these cars are represented by $X', X'', Y'$ and $Y''$; those of the undelayed car coincide at Z. All delay is thus regarded as time spent at the stop line, and the length of the queue at any time is the number of p.c.u's that are then regarded as waiting at the stop line. This simplification, while satisfactory for estimating delay, has to be dispensed with in estimating, for example, queue lengths or numbers of vehicles stopped.

It is often useful to have a name for the intervals of time between successive supposed arrivals ($ZX, XX', XX''$ in
Figure 1).

**DEFINITION:** An arrival-headway is a difference between the supposed arrival times of successive p.c.u's on an approach.

### 2.21 Regular arrivals.

The regular arrival model may be defined as follows:

**DEFINITION:** An approach has regular arrivals if p.c.u's arrive on the approach with equal headway of 1/q seconds, where q p.c.u./second is the average arrival rate.

This model, first analysed by Clayton (1940), does not claim to be realistic, except in very special circumstances, but the resulting delay expression (see section 2.3) forms part of several other expressions based on more realistic arrival models.

### 2.22 Binomial arrivals.

The binomial arrival model, first analysed by Winsten (Beckmann, McGuire and Winsten 1956), makes the discrete time assumption.

**DEFINITION:** An approach has binomial arrivals if, for some fixed Δt,

\[
P(1 \text{ p.c.u. arrives at time } nΔt) = α
\]

\[
P(\text{no p.c.u. arrives at time } nΔt) = 1-α
\]

for all n, and no arrivals can occur at any other times; α/Δt is the average arrival rate, so that α is independent of n.

It follows (Feller 1962) that if a is the number of p.c.u's arriving in a period containing N instants nΔt, then

\[
P(a = m) = \frac{N!}{(N-m)!m!} α^m(1-α)^{N-m}
\]
The random variable $a$ has probability generating function $(az + 1-a)^N$, mean $aN$ and variance $a(1-a)N$; hence, putting $N = C$, the ratio $I$ for binomial arrivals is $(1-a)$. The arrival headways, measured in units of $\Delta t$, have integer values with a negative binomial distribution.

Although the model remains mathematically valid whatever the value of $\Delta t$, its realism depends strongly on this value because $\Delta t$ represents the minimum arrival headway. $\Delta t$ is usually taken to be $1/s$, where $s$ is the saturation flow on the approach, because this simplifies the departure model and it is usually realistic to assume that the arrival rate will not exceed, even for short periods, the saturation flow on the approach. The minimum headway $\Delta t$ is more realistic where the approach consists of a single lane than where it has several lanes side by side, but the error in the latter case is unlikely to be large because $s$ is then likely to be large and $\Delta t$ small. An unsatisfactory feature of the model is that $I$ is necessarily less than 1, whereas the bunching of traffic on urban roads means that $I$ can be expected often to exceed 1, as observed by Miller (1964).

2.23 Poisson arrivals. Adams (1936) made observations showing that, on London streets with fairly light flows of between 200 and 700 vehicles/hour in each direction, the arrival of vehicles that had not just left a controlled intersection could be regarded in some respects as an example of the classical Poisson process. Adams himself drew some simple conclusions relevant to traffic signals, but the first full analysis of a fixed cycle signal with Poisson arrivals was that of Webster (1958).
DEFINITION: An approach has Poisson arrivals if, for some constant $q$, any small interval $\delta t$, and any time $t$, regardless of the number and distribution of arrivals in $(-\infty, t)$,

$$P(1 \text{ p.c.u. arrives in } (t, t+\delta t)) = q\delta t + o(\delta t)$$
and

$$P(\text{more than } 1 \text{ p.c.u. arrives in } (t, t+\delta t)) = o(\delta t).$$

It follows (Feller 1962) that if $a_T$ is the number of p.c.u's arriving in a period $T$, then

$$P(a_T = n) = \frac{e^{-qT}(qT)^n}{n!},$$

so that $a_T$ has a Poisson distribution with probability generating function $e^{qT(1-x)}$, mean $qT$ and variance $qT$. Hence the ratio $I$ for Poisson arrivals is $1$, and, if time is measured in seconds, $q$ is the average arrival rate as in section 1.24. Arrival headways have a negative exponential distribution.

The Poisson model can be obtained (Feller 1962) as the limiting case of the binomial model as $\alpha$ and $\Delta t$ tend to zero, $\alpha/\Delta t$ remaining equal to $q$. A comparison of the two models with $\Delta t = \frac{1}{2}$ second and $q = \frac{1}{6}$ p.c.u./second is given in section 2.5.

The main objection to the Poisson model on grounds of realism is that it requires a proportion of the headways to be very small, and this is physically impossible in a single lane. The model is, however, found to be quite realistic for an approach that is lightly loaded in relation to the capacity of the road leading to it.
2.24 Some other arrival models. A number of other arrival models have been used, some being generalisations of the above models, others quite distinct from them.

Newell (1956) used a model in which the probability density function of the arrival headway distribution was assumed to be

$$f(t) = \begin{cases} ke^{-(t - \frac{1}{k})} & (t \geq \frac{1}{k}) \\ 0 & (t < \frac{1}{k}) \end{cases}$$

i.e. the arrival headways were assumed to have a displaced exponential distribution. This imposes a minimum headway of 1/s. It corresponds to a Poisson process with rate $k$, interrupted immediately after each arrival by a period 1/s during which all arrivals that would result from the Poisson process are completely ignored; such arrivals are not assumed to occur at the end of the period 1/s or subsequently.

Meissl (1963), making the discrete time assumption, allowed the numbers of p.c.u's arriving at any one of the instants $n\Delta t$ to be integer-valued random variables each having the same generating function, say,

$$A(z) = \sum_{m=0}^{\infty} a_m z^m$$

This represents a generalisation of the binomial model, which is obtained by putting $a_0 = 1 - \alpha$, $a_1 = \alpha$, and $a_m = 0$ for $m > 1$. Meissl's model is particularly suited to the study of approaches having several lanes. It can also be used to approximate to any continuous time model such that the generating function of the number of p.c.u's arriving in the half-closed interval $((n-1)\Delta t, n\Delta t]$ is well-defined and independent
of n; this function is taken as A(z), so that arrivals in the interval are regarded as occurring at time nδt. Almost identical models were analysed by Darroch (1964) and Kleinecke (1964).

Another theoretical generalisation, analysed by McNeil (1968a), is the compound Poisson model, in which batches of p.c.u's are assumed to arrive according to the Poisson model with rate λ, whilst the number of p.c.u's in a batch is an integer-valued random variable with generating function \( φ(z) \). It reduces to the Poisson model when \( φ(z) = z \). The generating function for the number of p.c.u's arriving in a period T is

\[
A_T(z) = \exp \left\{ \lambda T \left[ φ(z) - 1 \right] \right\}
\]

The mean arrival rate \( q = \lambda φ'(1) \) and the ratio I is

\[
1 + \frac{φ''(1)}{φ'(1)}.
\]

I cannot be less than 1 (since \( φ \) is a generating function) but can be given any other positive value by suitable choice of \( φ \); \( q \) can then be given any required value by choice of \( λ \). This model is therefore well suited to situations where I is known to exceed 1, and could also be applied to a case where the arrival of vehicles was Poisson, and the number of p.c.u's equivalent to each vehicle was an integer-valued random variable with generating function \( φ \).

A more practical approach is that of Miller (1963), who assumes no more about the arrival distribution than the arrival rate, the value of the ratio I, and the fact that the numbers of arrivals in disjoint intervals of time are independent. Miller's (1964) observations showed values of I greater than 1 when vehicles were counted. It will now be shown that, under conditions close to those found in practice, if I exceeds 1
for each class of vehicles, then it will also exceed 1 for the corresponding total number of p.c.u's.

**LEMMA 1**. If the arriving traffic consists of $m$ classes of vehicles, and for vehicles in the $i$th class the arrival rate is $q_i$, the I-ratio $I_i$, and the p.c.u. equivalent $p_i$, with $I_i > 1$ and $p_i \geq 1$ for each $i$, and if the numbers of vehicles of different classes arriving in the same signal cycle are not negatively correlated, then the I-ratio for arrivals of p.c.u's exceeds 1.

**PROOF**: Let $q$ be the arrival rate for p.c.u's and $I^*$ the corresponding I-ratio. Let random variables $n_i$ and $n$ be the numbers of vehicles in class $i$ and p.c.u's, respectively, arriving in a cycle. Then, if $\sum$ denotes summation from 1 to $m$,

$$n = \sum_{i=1}^{m} p_i n_i \quad \text{and} \quad q = \sum_{i=1}^{m} p_i q_i$$

Hence

$$I^* = \frac{\text{var}(n)}{qc} = \frac{1}{qc} \text{var}(\sum_{i=1}^{m} p_i n_i)$$

$$= \frac{1}{qc} \left\{ \sum_{i=1}^{m} \frac{\text{var}(n_i)}{p_i} + \sum_{i \neq j} p_i p_j \text{cov}(n_i, n_j) \right\}$$

$$= \frac{\sum_{i=1}^{m} p_i^2 \text{var}(n_i)}{\sum_{i=1}^{m} p_i q_i c} + \frac{1}{qc} \sum_{i \neq j} p_i p_j \text{cov}(n_i, n_j) \quad \ldots \quad (*)$$

Now, for each $i$, $p_i \geq 1$ and $I = \frac{\text{var}(n_i)}{q_i c} > 1$, and therefore $p_i^2 \text{var}(n_i) > p_i q_i c$ for each $i$. Hence the first term in the right hand side of equation $(*)$ exceeds 1. The second term is non-negative because the $n_i$ and $n_j$ are not negatively correlated, which completes the proof that $I^* > 1$.

In practice the number of vehicles with p.c.u. values
less than 1 is unlikely to be large enough to reverse the result, and the correlation between \( n_i \) and \( n_j \) is likely to be negative only if the road leading to the approach is heavily loaded in relation to its capacity.

2.25 Departure models. Greenshields, Schapiro and Erickson (1947) observed that when a traffic signal controlling a single-lane approach turned green, allowing a queue of vehicles to start from rest, the intervals between departures of successive vehicles after the sixth were, on average, equal, so long as there was a queue (allowance being made for the types of vehicles in the queue). The intervals between departures of the first seven vehicles were larger, because these vehicles were still accelerating when they crossed the stop-line. These observations form the basis of the departure model generally used in theoretical studies.

The numbers of p.c.u's equivalent to vehicles of different types are determined so that the flow in p.c.u./second, during the period when vehicles crossing the stop-line are no longer accelerating, but there is still a queue, can be regarded as independent of the composition of the traffic. The saturation flow for an approach is then taken to be the flow in p.c.u./second during this period.

In the theoretical departure model (Clayton 1940; Webster 1958), p.c.u's are assumed to depart at equal time-intervals \( 1/s \) so long as there is a queue, and the departure time of the first vehicle is assumed to coincide with the start of the effective green time. To satisfy the latter assumption, the start of the effective green time is determined as follows. The theoretical departure times of p.c.u's other than the first
few in the queue are assumed to coincide with the average departure times of actual p.c.u's occupying the same positions in the queue. The theoretical departure times of the first few p.c.u's are then determined by assuming equal intervals $1/s$, and the theoretical departure time of the first p.c.u. is taken as the start of the effective green time. This will be later than the time when the signal turns green, by an amount equal to the sum of the amounts by which the first few departure intervals exceed $1/s$, possibly diminished by a small amount to allow for the first vehicle moving off while the signal shows red-and-amber. P.c.u's arriving during the effective green but after the queue has cleared are assumed to be undelayed. The end of the effective green time is determined by making $g_s$ (the maximum number of p.c.u's that can theoretically depart in the effective green time if the queue lasts throughout that time) equal to the average number of p.c.u's departing in an actual green period throughout which there is a queue.

Under the discrete time assumption, $\Delta t$ is taken as $1/s$, so that the first p.c.u. departs at the instant $n\Delta t$, at which the effective green time begins, and one p.c.u. departs at each succeeding instant $n\Delta t$ until either the effective green period ends or the queue clears (Beckmann et al 1956).

Passau (1963) modifies the above departure model slightly; he assumes that the effective green time is immediately preceded by an acceleration time whose duration is $p/s$, where $p$ is an integer. This is part of the effective red time, and no p.c.u. that is waiting at the start of the acceleration time may depart until the start of the effective green time. If, however, there is no queue at the start of the acceleration time,
Passau assumes that traffic arriving during the acceleration time passes undelayed through the intersection.

It sometimes happens that the departure rate tends to decrease as the green time elapses, even while there is a queue. This situation is analysed by Miller (1964) and Webster and Cobbe (1965).

More theoretical generalisations of the departure model have been made by Darroch (1964) and McNeil (1968a & b). Darroch, making the discrete time assumption, assumes binomial departures in the effective green time while there is a queue, and allows the number of departures at each other instant \( n \Delta t \) in the effective green time to have the same general probability generating function, subject to the condition that it may not exceed the number of arrivals at the same instant. McNeil allows the intervals between departures to be independently and identically distributed random variables, subject to an upper bound on the number of departures per cycle.

Other variations relate to the case where right-turning traffic on a two-way street shares one approach with traffic going straight ahead or turning left: Newell (1959), making the discrete time assumption, considered a signalised cross-roads where a two-way road with one lane in each direction crossed a side street, and there were always queues in both directions on the main road. For each of these queues, and for each instant \( n \Delta t \) in the effective green time, the first vehicle in the queue was assumed to depart, provided that either

(a) it was not turning right,

or  (b) it was turning right and so was the first vehicle in the opposing queue
or (c) \( n \Delta t \) was the last such instant in the effective green time.

There does not yet appear to be a satisfactory theoretical treatment of the corresponding situation where either queue clears before the end of the green time.

For practical purposes Webster (1967) suggests that the extra delay caused by right-turners on shared approaches can be allowed for to some extent by regarding each right-turning p.c.u. as a larger number (usually 1.75) of p.c.u's departing at rate \( s \). He also considers how to estimate how many right-turners can depart through gaps in the opposing flow, and how many can depart at the end of the green period, when the lights change. His estimate for the first of these is confirmed by the work of Gordon and Miller (1966).

2.26 The overflow. With all except the regular arrival model, it is possible, even on an approach with a degree of saturation less than 1, for more p.c.u's to arrive in an effective red period and the following effective green period than can depart in that green period. There is, therefore, sometimes a queue at the end of the effective green period, and the length of this queue is referred to so often in the analysis that it is useful to give it a name.

DEFINITION: The overflow on an approach to a traffic signal is the number of p.c.u's in the queue immediately before the beginning of the effective red period. The overflow is an integer-valued random variable and will be denoted by \( Q_0 \).
2.3. Expressions for delay: regular arrivals

Expressions, similar to the one to be derived in Theorem 1, for delay on an approach with regular arrivals were obtained by Clayton (1940), Wardrop (1952) and, working numerically using Greenshields' (et al 1947) flow data, by Matson and McGrath (1954). The present derivation will be rather more general than theirs.

In this section, for any expression \( X \), let \( [X] \) denote the integer part of \( X \).

Consider an approach with regular arrivals. With the notation of section 1.24, p.c.u.'s arrive at intervals \( 1/q \) so that the number of arrivals in one cycle takes the values \( [cq] \) and \( [cq] + 1 \) with frequencies whose ratio, in the long run, is \( (1 + [cq] - cq):(cq - [cq]) \), so that the average number of arrivals per cycle is \( cq \). Since the first p.c.u. always departs at the start of the effective green time, the number of vehicles that could depart, if there were enough waiting, is \( [gs] + 1 \) in every cycle. The approach will not be overloaded, in the long run, provided that

\[
q < [gs] + 1,
\]

from which it follows that

\[
[cq] + 1 < [gs] + 1,
\]

so that the number of p.c.u.'s arriving in a cycle never exceeds the number that can depart in that cycle.

**THEOREM 1**: On an approach with regular arrivals at rate \( q \) p.c.u./sec, saturation flow \( s \) p.c.u./sec, and flow ratio \( y \), if the cycle time is \( c \) seconds, the effective red time \( r \) seconds, and the approach is not overloaded, then the average
delay per p.c.u. is
\[ d = \frac{1}{2c(1-y)} \left\{ \left( r - \frac{1}{2s} \right)^2 + \frac{y(2-y) + \theta(1-y)^2}{12q^2} \right\} \text{ seconds,} \]
where \( \frac{1}{3s} < \theta < \frac{2}{3s} \)

PROOF: Take the time-origin \( t = 0 \) at the beginning of a cycle, so that \((0, r)\) and \((r, r+g)\) are respectively the effective red and green times. Suppose that the first p.c.u. in this cycle arrives at \( t = \frac{u}{q} \), where \( 0 < u < 1 \); it will depart at \( t = r' \), being delayed by an amount \( r' - \frac{u}{q} \). The \( m \)th p.c.u. in the cycle will arrive \( \frac{m-1}{q} \) later, and depart \( \frac{m-1}{s} \) later, being delayed by

\[ r' - \frac{u}{q} - \frac{(m-1)}{q} (1-y) \quad , \quad \text{where} \quad y = \frac{a}{s} \]

This will hold for \( m = 2, 3, \ldots, M+1 \), where \( M = \left\lfloor \frac{a \cdot \frac{r-u}{1-y} - k(u)}{1-y} \right\rfloor \). Thus \( M+1 \) p.c.u.'s are delayed during the cycle, and the total delay is

\[ \sum_{m=1}^{M+1} \left\{ r - \frac{u}{q} - \frac{(m-1)}{q} (1-y) \right\} = (M+1) \left\{ r - \frac{u}{q} - \frac{M}{2q} (1-y) \right\} \]

Let \( M = \frac{a \cdot \frac{r-u}{1-y} - k(u)}{1-y} \), where \( 0 \leq k(u) < 1 \)

Then the total delay is

\[ \frac{M+1}{2} \left\{ r - \frac{u}{q} + \frac{k(u)}{q} (1-y) \right\} \]

\[ = \frac{q}{2(1-y)} \left\{ r \left( r + \frac{1-y}{q} \right) - \frac{u}{q} \left( 2r + \frac{1-y}{q} \right) + \frac{u^2}{q^2} + k(u) \left( 1 - k(u) \right) \left( \frac{1-y}{q} \right)^2 \right\} \quad (*) \]

Now as \( u \) varies from \( 0 \) to \( 1 \), \( \frac{a \cdot \frac{r-u}{1-y}}{1-y} \) varies by an amount \( \frac{1}{1-y} \), so that \( k(u) \) describes the interval \((0,1) \frac{1}{1-y} \) times as a piecewise linear function of \( u \).

Let \( \left\lfloor \frac{1}{1-y} \right\rfloor = N \) and \( \frac{1}{1-y} = N + l \), where \( 0 \leq l < 1 \), and let \( Q \) be the part of the interval \((0, 1)\) that is described \( N+1 \) times by \( k(u) \) as \( u \) varies from \( 0 \) to \( 1 \). Then \( Q \)
is either an interval or the complement in \((0, 1)\) of an interval, and depends on \(q\), \(r\), and \(y\) only.

In the long run, unless \(c\) is, very exceptionally, a rational multiple of \(\frac{1}{q}\) with (in its lowest terms) a small denominator, \(u\) can be regarded as uniformly distributed between 0 and 1. Hence \(E(u) = \frac{1}{2}\) and \(E(u^2) = \frac{1}{3}\), and

\[
E \left\{ k(u)(1 - k(u)) \right\} = \frac{N \int_0^1 z(1-z)dz + \int_q z(1-z)dz}{N + t}
\]

where \(\int_q\) denotes integration over \(Q\).

It follows from the form of the function \(z(1-z)\) in \(0 < z < 1\) that, for given \(N\), \(E \left\{ k(u)(1 - k(u)) \right\}\) is greatest if \(Q\) is an interval of the form \((\frac{1}{2} - v, \frac{1}{2} + v)\), where \(0 < v < \frac{1}{2}\), and least if \(Q\) is the complement of such an interval. In the former case

\[
E \left\{ k(u)(1 - k(u)) \right\} = \frac{1}{6} + \frac{1}{6} \frac{4v^2}{N + 2} = \frac{1}{6} + f(v),
\]

where \(f(v)\) has a unique maximum in the interval \((0, \frac{1}{2})\) and, if this is attained at \(v = v_1\), differentiation shows that

\[
16v_1^3 + 12Nv_1^2 - N = 0.
\]

Ignoring the first term in this equation, a first approximation to \(v_1\) is seen to be \(\frac{1}{2\sqrt[3]{3}}\), and the sign of the first term shows that \(v_1 < \frac{1}{2\sqrt[3]{3}}\).

From the fact that \(f'(v_1) = 0\) it follows that

\[
f(v_1) = \frac{4v_1^3}{3N}
\]

Hence \(f(v_1) < \frac{1}{18N\sqrt[3]{3}} < \frac{1}{18\sqrt[3]{3}}\), since \(N > 1\).
and \( E \{ k(u)(1 - k(u)) \} < \frac{1}{6} + \frac{1}{18\sqrt{3}} \).

In the latter case

\[
E \{ k(u)(1 - k(u)) \} = \frac{1}{6} - \frac{1 - 4v^2}{6(N+1)} = \frac{1}{6} + g(v), \text{ say,}
\]

where \( g(v) \) has a unique minimum in \((0, \frac{1}{2})\). It follows, as in the former case, that, if this is attained at \( v = v_2 \),

then \( \frac{1}{2\sqrt{3}} \) is a first approximation to \( v_2 \), \( v_2 < \frac{1}{2\sqrt{3}} \) and

\[
g(v_2) = -\frac{4v_2^2}{3(N+1)}.\]

Hence \( g(v_2) > -\frac{1}{18(N+1)\sqrt{3}} > -\frac{1}{36\sqrt{3}} \), since \( N \geq 1 \),

and

\[
E \{ k(u)(1 - k(u)) \} > \frac{1}{6} - \frac{1}{36\sqrt{3}}.
\]

Substituting for \( E(u), E(u^2) \) and \( E \{ k(u)(1 - k(u)) \} \)

in the expectation of the expression (*) to obtain the average delay per cycle and dividing by \( q_c \), the average number of p.c.u's passing through the approach in one cycle, the required result follows.

The first term of the expression in Theorem 1 is just Wardrop's formula; i.e., in the present notation,

\[
d = \frac{(\gamma - \frac{1}{2s})^2}{2c(1 - \gamma)}.\]

His formula therefore very slightly underestimates the delay. The difference arises only from the simplifying assumptions in his derivation and is, as he pointed out (1952, p. 358) generally trivial in practice. It seems worthwhile, however, to have a more exact theoretical result, since the difference
could be appreciable in the case of a short red time with a low arrival rate.

Clayton's formula was slightly different again, owing to his assumption that the delay to the last delayed p.c.u. was related to the interval between the beginnings of the actual and effective green times - an assumption that seems hard to justify.

If, in Theorem 1, \( q \) is allowed to increase indefinitely, the flow ratio remaining constant, then the average delay per p.c.u. tends to

\[
d = \frac{\tau^2}{2c(1 - y)}.
\]

This limiting process is equivalent to regarding the traffic on the approach as a fluid. The resulting expression will appear again in sections 2.5.1 and 2.6.

Although the regular arrival model does not pretend to realism, it has been shown by simulation (Wardrop 1952, Webster 1958) that the resulting expression for average delay can be expected to give quite realistic estimates on approaches where the degree of saturation is less than about \( \frac{1}{3} \). This is confirmed in a theoretical study by Newell (1956), which is discussed in section 2.6.

2.4. Expressions for delay: binomial arrivals

The delay on a traffic signal approach with binomial arrivals was first analysed by Winsten (Beckmann et al 1956). He first showed that the average delay incurred during the effective red time was, in the notation of section 1.24,
\[ R \{ E(Q_o) + \frac{1}{2} R (R+1) \} \Delta t \quad \text{p.c.u. seconds,} \]

where \( E \) denotes expectation and \( Q_o \) is the overflow. He next showed that, if \( Q_g \) is the number of p.c.u.'s in the queue just before the beginning of the effective green time, and if the green time were long enough for the queue to clear, then the average delay incurred during the green time would be

\[ \frac{1}{2(1-\alpha)} E \{ Q_g^2 + (2\alpha - 1) Q_g \} \Delta t \quad \text{p.c.u. seconds. \ldots (\star)} \]

Because the queue does not always clear before the green time ends, the expression (\( \star \)) exceeds the average delay incurred during the green time by an amount that Winsten shows to be

\[ \frac{1}{2(1-\alpha)} E \{ Q_o^2 + (2\alpha - 1) Q_o \} \Delta t \quad \text{p.c.u. seconds} \]

Using the fact that \( Q_o \) and \( Q_g \) are not independent, but differ, when the traffic flow on the approach is in statistical equilibrium, by the number of arrivals in the effective red time, Winsten showed that the average delay to a p.c.u. passing through the approach is

\[ d = \frac{R}{(1-\alpha)C} \left( \frac{E(Q_o)}{\alpha} + \frac{R+1}{2} \right) \Delta t \quad \text{seconds} \]

More recently it has been shown by Dunne (1967), and afterwards more simply by Potts (1967), that when there is no overflow, and the green time is long enough for the queue to clear, the total delay (measured in units of \( \Delta t \) p.c.u. seconds) in a cycle in which the effective red time is \( R \Delta t \) has probability generating function
\[ D_R(z) = \prod_{j=1}^{R} \frac{(1-\alpha)}{(1-\alpha z^j)} . \]

Potts obtained this result by observing that the numbers of vehicles delayed by \( R \Delta t, (R-1) \Delta t, \ldots, 2 \Delta t, \) and \( \Delta t \) seconds all have the negative binomial distribution with probability generating function \((1-\alpha)/(1-\alpha z)\).

Winston had remarked that the probability distribution of the overflow could be obtained by suitable analysis, and this was done by Newell (1960), using methods closely analogous to those developed in another context by Bailey (1954). By considering what happens between the end of one effective green time and the end of the next, he showed that if \( H(z) \) is the probability generating function of \( Q_0 \), then

\[ H(z) = \frac{N(z)}{Q(z)} , \]

where \( N(z) \) is a polynomial of degree \( G \) and

\[ Q(z) = z^G - (1 - \alpha + \alpha z)^C . \]

Applying Rouché's Theorem (Whittaker and Watson 1958) to the functions \( z^G \) and \((1 - \alpha + \alpha z)^C\) on the contour \(|z| = 1 + \delta\), when \( \delta \) is positive and sufficiently small, he showed that, if the degree of saturation of the approach is less than 1, \( Q(z) \) has just \( G \) zeros inside or on the unit circle. \( H(z) \) must be regular inside and on the unit circle, so these zeros must be just those of \( N(z) \). It follows that

\[ H(z) = \prod_{j=1}^{R} \frac{(1-z_j)}{(z-z_j)} , \]

where the \( z_j \) are the \( R \) zeros of \( Q(z) \) outside the unit circle.

Hence

\[ E(Q_s) = H'(1) = \sum_{j=1}^{R} \frac{1}{z_j(z_j-1)} . \]
Newell examined the behaviour of the $z_j$ as $\alpha \rightarrow G/C$, i.e. as the degree of saturation approaches $1$, and showed in particular that, in these circumstances,

$$E(Q_s) \sim \frac{RG}{2C(G-\alpha C)}.$$  

Meissl (1962) made an analysis along lines similar to Newell's, but in much greater detail, introducing the probability generating functions of the numbers of p.c.u's on the approach* at each of the instants $n\Delta t$ in the cycle. He obtained the average delay per p.c.u. in the form

$$\hat{a} = \frac{R \Delta t}{(1-\alpha)C} \left\{ \frac{R+1}{2} + \frac{1}{\alpha} \frac{e^{-\frac{1}{\alpha} \Delta t}}{\sum_{j=1}^{R} \frac{1}{(z_j-1)}} \right\} \text{ seconds},$$

which is seen to agree with the results of Wisten and Newell.

2.5. Expressions for delay; Poisson arrivals

This section discusses expressions for delay that apply specifically to an approach with Poisson arrivals; some of the more general work discussed in section 2.6 also applies in particular to Poisson arrivals.

* Meissl includes any p.c.u. that departs at time $n\Delta t$ in the number of p.c.u's on the approach just before instant $(n+1)\Delta t$; this leads, for example, to a slight difference between his expression for the overflow and Newell's.
2.51. **Webster's expression**. The first, and probably most extensively used expression for delay on a traffic signal approach with Poisson arrivals is that of Webster (1958), which is, in the notation of section 1.24,

\[
d = \frac{c(1-\lambda)^2}{2(1-\lambda x)} + \frac{\lambda x}{2q(1-x)} - 0.65 \left( \frac{c}{q^2} \right)^{\frac{1}{3}} x^{(2+5\lambda)}
\]

Having derived this expression partly by means of computer simulation, Webster showed that the estimates of delay given by it agreed closely with actual delays observed at a number of sites. The theory underlying the expression will now be discussed in greater detail than in previous publications, because the author's work described in Chapter 3 is based upon it. Since

\[
c(1 - \lambda) = r \quad \text{and} \quad \lambda x = y
\]

Webster's first term is seen to be the expression obtained in section 2.3 for the average delay to a p.c.u. when arrivals are regular and the traffic is regarded as a fluid.

As Wardrop remarked (1952) the effect of randomness in arrivals is to cause overflow, and therefore, except at low degrees of saturation, substantial extra delay that is not taken account of by the expression obtained in section 2.3. One way of allowing for this effect of randomness is to suppose, for the purposes of analysis, that a queue with constant service time \(1/\lambda s\) is interposed between the arriving traffic and the signal, so that the maximum number of p.c.u.'s that can reach the signal in one cycle is just the maximum number that can depart in one green period. It follows from the Pollaczek-Khintchine formula in the theory of queues, a simple derivation of which is given by Kendall (1951), that the mean waiting time in the interposed queue is \(x^2/2q(1-x)\).
Consider now what would happen at the signal itself if such a queue were interposed. Traffic would arrive in bunches, the number of p.c.u.'s in a bunch would be a random variable with the Exponential distribution (Tanner, 1961), the arrival headways within each bunch would be \( \frac{1}{\lambda} \) s, and the headway between the last p.c.u. of one bunch and the first of the next would exceed \( \frac{1}{\lambda} \) s by an amount having the exponential distribution with mean \( \frac{1}{q} \). No exact analysis of delay at a fixed-cycle signal with such arrivals will be attempted here. The work of Miller (1963, p. 376), however, applied to the present case where the overflow is known to be zero, shows that the average delay incurred during the effective red time is, as in the case of regular arrivals, \( q_c(1 - \lambda)^2/2 \). Moreover, the work of Newell (1965), discussed further in section 2.6, shows that, in the absence of overflow, the delay expression for regular arrivals underestimates the delay incurred in the effective green time only by an amount whose order of magnitude is a fraction \( \frac{1}{sg} \), (or in the region of five per cent) of the total delay per cycle given by the uniform arrival expression. A good approximation to the average delay to a p.c.u. at the traffic signal, after passing through the interposed queue, is therefore given by the expression in Theorem 1, and the last term in that expression can be neglected in view of the approximation already made. To sum up, the average delay to a p.c.u. passing through the interposed queue and the traffic signal would be

\[
\frac{x^2}{2q(1-x)} + \frac{1}{\lambda s} + \frac{1}{2c(1-y)} \left( \gamma - \frac{1}{2s} \right)^2,
\]

or, on expanding the last term, neglecting the term in
\[ \frac{1}{(\tau s)^2}, \text{ and rearranging,} \]

\[
\frac{c(1-\lambda)^2}{2(1-\lambda x)} + \frac{x^2}{2q(1-x)} + \frac{1}{s} \left( \frac{1}{\lambda} - \frac{(1-\lambda)}{2(1-y)} \right)
\]

The first two terms are seen to be those of the Webster expression.

In the above derivation, three small positive terms were neglected; on the other hand, the delay at the signal with interposed queue would exceed the delay at the actual signal because there would be times when traffic was held up in the interposed queue when it could proceed undelayed through the signal. To allow for these effects, Webster made a digital simulation of traffic passing through approaches with various values of \( c \), \( q \), \( \lambda \) and \( x \), and, by regression analysis of the differences between the resulting values of \( d \) and estimates given by the first two terms of his expression, obtained the third term. He found that the third term represented between 5 and 15 per cent of the total, so that for most practical purposes the average delay could be taken as

\[
d = \frac{9}{10} \left\{ \frac{c(1-\lambda)^2}{2(1-\lambda x)} + \frac{x^2}{2q(1-x)} \right\}
\]

Although the delay expression is intended to apply to Poisson arrivals, Webster's simulations assumed binomial arrivals with \( \Delta t = \frac{1}{2} \) second. Figure 2 shows the arrival headway distributions, exponential and geometric respectively, of the two models when the arrival rate is \( \frac{1}{6} \) p.c.u./second, a fairly typical rate for a single-lane approach. Figure 3 shows the probability distributions of the number of arrivals
Fig. 2. ARRIVAL HEADWAY DISTRIBUTIONS IN THE BINOMIAL AND POISSON MODELS WITH ARRIVAL RATE 1/6 p.c.u./second AND, IN THE BINOMIAL MODEL, A TIME INTERVAL OF 1/2 second
Fig. 3. PROBABILITY DISTRIBUTIONS OF THE NUMBERS OF ARRIVALS IN ONE MINUTE IN THE BINOMIAL AND POISSON MODELS WITH ARRIVAL RATE $\frac{1}{6}$ p.c.u./second AND, IN THE BINOMIAL MODEL, A TIME INTERVAL OF $\frac{1}{2}$ second.
in one minute, a typical cycle-time, with the same arrival rate. The smallness of the discrepancies confirms that the third term in Webster's expression can be taken to apply to Poisson arrivals.

2.52. Other work related to Poisson arrivals. The overflow in the case of Poisson arrivals was analysed by Haight (1959), who obtained, in terms of the Poisson and Borel-Tanner (Tanner 1961) distributions, the probability distribution of the overflow, conditional upon the number of p.c.u.'s in the queue at the end of the preceding red time. He deduced, for the case where the flow of traffic on the approach is in statistical equilibrium, simple relations between the probability distributions of the overflow and the number of p.c.u.'s in the queue at the end of the red time.

The usual departure model assumes that p.c.u.'s arriving during the effective green time are undelayed if there is no queue when they arrive. With Poisson arrivals this implies that an arbitrarily large number of departures is possible in one cycle. This possibility prevents the application of the methods of Bailey (1954) to the analysis of the overflow. These methods can be applied, however, (as they have been in the case of binomial arrivals; see section 2.4), if an upper bound (usually \([ge] + 1\)) is placed upon the number of departures per cycle. Newell (1960) showed in this way that as \(x \to 1\),

\[ E(Q_0) \sim \frac{1}{2(1-x)} \]

or \(c/\sim\) times the corresponding expression for binomial
arrivals. The effect of restricting the number of departures must be to increase the overflow, so that the overflow with restricted departures provides an upper bound for the overflow with the usual departure model. Newell also showed that as \( x \to 1 \)

\[
d \sim \frac{1}{2q(1-x)} ,
\]
as does Webster's expression. The details of the corresponding analysis are omitted from Newell's paper, but are given by Haight (1963), who derives equations sufficient to determine the probability generating function of \( Q_0 \). A parallel analysis is given by Meissl (1963) who makes the discrete time assumption and approximates to Poisson arrivals by assuming that all such arrivals in the half-closed interval \( [(n-1)\Delta t, n\Delta t] \) occur at \( n\Delta t \) (an approximation which does not affect the overflow). Meissl obtains the probability generating function \( H(z) \) of the overflow for this case explicitly in the form

\[
H(z) = \frac{(sG - qC)(z - 1)e^{s(z - 1)/z}}{s(z^G - e^{qC(z - 1)/z})} \prod_{j=1}^{G-1} \left\{ \frac{z - z_j e^{qC(z - 1)/z}}{1 - z_j e^{qC(z - 1)/z}} \right\} ,
\]

where the \( z_j \) are the \( G-1 \) zeros inside the unit circle of \( z^G - e^{qC(z-1)/s} \). The resulting expression for the mean overflow confirms Newell's statement of the limiting behaviour.

* In Meissl's analysis the overflow includes the p.c.u. (if any) that departed at the last instant \( n\Delta t \) of the effective green period.
of this quantity as $x \to 1$.

Passau (1963) gives, without derivation, expressions relating to an approach with Poisson arrivals and his modified departure model (see section 2.25). With the notation of section 1.24, and letting the acceleration time be $a$ seconds, his first result is that if $n_r$ is the number of p.c.u.'s delayed when the Poisson flow is interrupted by just one effective red time $\gamma$, then

$$E(n_r) = \frac{q}{1-\gamma} (\gamma - ae^{-\gamma}) .$$

He follows this by an expression for the total delay to these p.c.u.'s, and an adjustment to allow for some of them being further delayed by a second equal red time. It is not clear whether his expressions are claimed to be exact, but the following Lemma shows that the expression for $E(n_r)$ is only an approximation, albeit a good one.

**Lemma 2.** On an approach with Poisson arrivals at rate $q$ p.c.u./second and Passau's departure model with saturation flow $s$ p.c.u./second, the number of p.c.u.'s delayed by just one red period has limit $q(\gamma - ae^{-\gamma}(\gamma - a))$ as $s \to \infty$.

**Proof.** Let $a_r$ and $a_a$ be the numbers of arrivals in the effective red time and the acceleration time respectively. As $s \to \infty$ the time taken by these p.c.u.'s to depart tends to zero, and so does the number of subsequent arrivals that are delayed. Arrivals in the acceleration time are delayed only if there is at least one arrival in the first $\gamma - a$ seconds of the effective red time. Hence
\[
\lim_{s \to \infty} \{ E(n_r) \} = E(a_r) - E(a_d) P(\text{no arrivals in } r-a \text{ seconds}) \\
= q(r - ae^{-q(r-a)}) \text{ as required.}
\]

COROLLARY. Passau's expression for \( E(n_r) \) is not exact.

PROOF: As \( s \to \infty, y \to 0 \) and Passau's expression for \( E(n_r) \) tends to \( q(r - ae^{-q(r-a)}) \), contrary to Lemma 2.

The discrepancy is, however, in the second term, which is usually small compared with the first.

A full analysis of the delay caused when a Poisson flow is interrupted by just one effective red time was made by Buckley and Wheeler (1964). They assumed a departure model like Passau's with acceleration time \( 1/s \), and found an exact expression for the probability distribution of the number of vehicles delayed: they also showed how to determine the moment generating function of the total delay. Their first result can be expressed in the present notation as

\[
P(n_r=j) = \frac{q^i}{i!} \left( r - \frac{1}{s} \right)^i \left( r + \frac{j-1}{s} \right)^{j-i} \exp \left\{ -q(r + \frac{j-1}{s}) \right\} \quad (j = 0, 1, 2, \ldots)
\]

This result can be used to obtain an exact expression for \( E(n_r) \) in Passau's case, as follows.

THEOREM 2. On an approach with Poisson arrivals at rate \( q \) p.c.u./second and Passau's departure model with saturation flow \( s \) p.c.u./second, if \( n_r \) is the number of p.c.u.'s delayed by just one effective red time of \( r \) seconds, including an acceleration time \( a \) seconds, then
\[ E(n_r) = e^{-a} \sum_{j=0}^{\infty} \frac{q^{j+1}}{j!} e^{-\lambda s} \left\{ \left( \frac{r}{s} + \frac{j}{s} \right)^j \left( \frac{r}{s} - \frac{1}{s} \right)^j + \left( a + \frac{j}{s} \right)^j \left( a - \frac{1}{s} \right)^j \right\} \]

**Proof:** For \( j \geq 0 \),

\[ P(n_r = j+1 \text{ and first delayed p.c.u. arrives between } t \text{ and } t+\delta t \text{ seconds after beginning of effective red time}) = e^{-q t} q \delta t P(n_{r-t+\delta t} = j) , \]

and no traffic is delayed if none arrives in the first \( r-a \) seconds of the effective red time.

Hence

\[ P(n_r = j+1) = \int_0^{r-a} q e^{-\lambda t} P(n_{r-t} = j) dt \]

and, using Buckley and Wheeler's expression for the probability under the integral sign,

\[ P(n_r = j+1) = \int_0^{r-a} \frac{q^{j+1}(r-t)}{j!} e^{-\lambda t} \left( \frac{r-t}{s} + \frac{j}{s} \right)^j \exp \left\{ -q \left( \frac{r-t}{s} + \frac{j}{s} \right) \right\} dt \]

\[ = q^{j+1} \exp \left\{ -q \left( r + \frac{j}{s} \right) \right\} \int_0^{r-a} \left\{ \frac{1}{j!} \left( \frac{r-t}{s} + \frac{j}{s} \right)^j - \frac{j}{s(j-1)!} \left( \frac{r-t}{s} + \frac{j}{s} \right)^{j-1} \right\} dt . \]

Hence

\[ E(n_r) = \sum_{j=0}^{\infty} (j+1) P(n_r = j+1) \]

\[ = e^{-a} \sum_{j=0}^{\infty} \frac{q^{j+1} e^{-\lambda s}}{j!} \left[ \left( \frac{r}{s} + \frac{j}{s} \right)^j + \frac{j+1}{s(j+1)!} \left( \frac{r}{s} - \frac{1}{s} \right)^j \right]^{r-a} \]

\[ = e^{-a} \sum_{j=0}^{\infty} \frac{q^{j+1}}{j!} e^{-\lambda s} \left\{ \left( \frac{r}{s} + \frac{j}{s} \right)^j \left( \frac{r}{s} - \frac{1}{s} \right)^j - \left( a + \frac{j}{s} \right)^j \left( a - \frac{1}{s} \right)^j \right\} . \]

**Corollary.** Buckley and Wheeler's expression for \( P(n_r = j) \) is consistent with Lemma 2.

**Proof:** Letting \( s \to \infty \) in the result of Theorem 2, observing
that the series is uniformly convergent for $s \geq \delta$ for any strictly positive $\delta$,

$$
\lim_{n \to \infty} \left\{ E(n) \right\} = e^{-\nu} \sum_{j=0}^{\infty} \frac{q^j}{j!} (r^j - a^{j+1}) \\
= qe^{-\nu}(re\nu - ae\nu) \\
= q(r - ae^{-\nu(r-x)}) \text{ as required.}
$$

Buckley and Wheeler's results, though of considerable theoretical interest and elegance, have, like those of Passau, rather limited practical value because they apply only to a single interruption of the flow.

2.6. Expressions for delay: more general arrival models.

Newell (1956) considered an approach with the usual departure model, and arrival headways distributed identically and independently, but otherwise arbitrarily, subject to a minimum of $1/s$. He observed that with this model, given the arrival time of a p.c.u., the delay to it has a discrete distribution that depends on the arrival time. He obtained integral equations for these distributions in the case when the queue on the approach is in statistical equilibrium, and showed how to solve them in principle by successive approximations, the first approximation being equivalent to assuming zero overflow. In the case of regular arrivals the first approximation gave the exact solution and confirmed that Wardrop's expression (see section 2.3) estimates the delay very closely. As a second example, Newell considered the displaced exponential arrival headway distribution (see section.
the first approximation gave the expression obtained in section 2.3 for regular arrivals with the traffic considered as a fluid, and the second approximation differed appreciably from the first only when the degree of saturation exceeded about 0.6.

Meissl (1963), in an intricate extension of his earlier work, applied the methods of Bailey (1954), with the discrete time assumption, to obtain the probability generating function of the overflow on an approach with arrivals according to a more general model (see section 2.24), and departures according to the usual model but from p lanes side by side, between which the arriving p.c.u.'s are distributed so that the lengths of queues in different lanes never differ by more than one p.c.u. For the case in which the number of arrivals at any instant $n\Delta t$ is binomially distributed between 0 and p, he also derives the probability generating function of the delay to a p.c.u.

Miller (1963) made an analysis assuming only the usual departure model and that the queue on the approach was in statistical equilibrium and that the numbers of arrivals in successive red and green times were independently distributed. He first showed that the mean total delay incurred during the effective red time is

$$\tau \left\{ \mathbb{E} (Q_0) + \frac{1}{2} q \tau \right\}.$$ 

He then analysed the delay incurred during the effective green time, using a method analogous to Winsten's, i.e. obtaining the mean total delay incurred if the green time always continues until the queue has cleared, and then subtracting the average
amount of this delay that would be incurred after the green
time actually ends. Suppose that there are $Q_g$ p.c.u.'s in
the queue just before the beginning of the effective green'
time, that the queue would clear after $k$ departures if the
green time lasted long enough, and that the $j$th departing
p.c.u. leaves behind a queue of $n_j$ p.c.u.'s. Then it can
be shown that with binomial arrivals

$$ E(k) = \frac{Q_g - 1}{1 - y} + 1 $$

and

$$ E(n_j) = Q_g - j + \frac{j - 1}{k - 2} (k - Q_g) \quad (1 \leq j \leq k - 1). $$

By making the reasonable assumption that these two relations
would hold approximately for other arrival patterns, and by
assuming that the parameter $I$ defined in section 2.2 had the
same value for arrivals in the effective red time as for
arrivals in the whole cycle, Miller obtained, in terms of the
overflow, an expression for the total delay incurred in the
effective green time. By considering the relation between the
overflows in successive cycles he showed that, when the queue
on the approach is in statistical equilibrium,

$$ E(Q_0) = \frac{(2x - 1)I}{2(1 - x)}, $$

and that, with this approximation, the average delay per p.c.u.
is given by

$$ d = \frac{(1 - \lambda)}{2(1 - \lambda x)} \left\{ c(1 - \lambda) + \frac{(2x - 1)I}{q(1 - x)} + \frac{1 + \lambda x - 1}{s} \right\} $$

The first term of this expression is the same as that
of Webster's, and the second term $\sim I/2q(1-x)$ as $x \to 1$ so
that, since $I = 1$ for Poisson arrivals, the limiting
behaviour of Miller's expression as $x \to 1$ is the same as
that of Webster's. Miller found (1963 & 1964) that his expression and Webster's gave similar estimates of delay when I was near to 1, but his own expression gave rather higher estimates, and better agreement with observations, when I was appreciably greater than 1.

Darroch (1964) extends to his own discrete time arrival and departure models (see sections 2.24 & 2.25) the analysis of Newell (1960). He obtains a lengthy expression for the average delay per p.c.u., \( d \), in terms of the average overflow \( E(Q_0) \). This expression can be shown to reduce to Winsten's (see section 2.4) when the appropriate arrival and departure parameters are inserted. Darroch then finds upper and lower bounds for \( E(Q_0) \) and hence for \( d \), and evaluates these for several realistic numerical examples. The results confirm those of Newell's (1960) analysis for binomial arrivals except that \( RG/2C(G - \alpha C) \), Newell's asymptotic estimate of \( E(Q_0) \) as \( x \to 1 \), slightly exceeds Darroch's upper bound even when \( x = 0.98 \).

A special case of Darroch's and Meissl's (1963) analyses was investigated independently by Kleinecke (1964), who obtained the probability generating function \( H(z) \) of the overflow for a single-lane approach with Meissl's arrival model (see section 2.24) in the form

\[
H_1(z) \propto \sum_{j=1}^{g} \frac{(A(z) - z z_j^{-R(x)})}{(A(z))^c - z^c},
\]

where the \( z_j \) are the zeros of the denominator inside and on the unit circle. The constant of proportionality can be obtained in terms of the \( z_j \) by putting \( H(1) = 1 \).
Newell (1965) regarded the traffic as a fluid flowing into the approach at a rate that is a random variable with mean $q$, and flowing out at a fixed rate $s$ during the effective green time so long as any accumulated fluid remains. He first confirmed that if the arrival rate is constant, the total delay incurred per cycle is $\frac{r^2 q}{2(1-y)}$ (cf. section 2.3). He then showed that if the arrival rate varies, but not enough to cause overflow, the above expression should be increased by a fraction whose order of magnitude is $1/sg$. He also shows, by approximate analysis of the distribution function of the overflow, that the extra delay due to overflow can be expressed, to first order in $1/sg$, as a multiple of a function $H(\mu)$, say, of the quantity

$$\mu = \frac{sg - q}{1sg} ,$$

where $I$ is as defined in section 2.2; $\mu$ is a dimensionless measure of spare capacity on the approach. $H$ is a decreasing function with $H(0) = 1$ and $H(1) \approx 0.25$, and is obtained by numerical integration. Newell's work leads to the expression

$$d = \frac{c(1 - \lambda)^2}{2(1 - \lambda x)} + \frac{I H(\mu) x}{2q(1 - x)} + \frac{I(1 - \lambda)}{2s(1 - \lambda x)^2} ,$$

which agrees with Webster's in the first term, and in its asymptotic behaviour as $x \to 1$ (since then $\mu \to 0$, and since $I = 1$ for Poisson arrivals). Newell quotes numerical examples in which his expression agrees with Webster's to within 3 to 4 per cent for a wide range of degrees of saturation.

McNeil (1968a) considered an approach with compound
Poisson arrivals (see section 2.24); he obtained \( d \) in terms of \( E(Q_0) \) by considering the delay incurred in the red and green times separately, as did Winsten and Miller. McNeil's method of finding the average delay in a green time that lasts until the queue has cleared is, however, different from theirs. He considers the sequence \( \{a_j \; ; \; j = 0 \; , \; 1 \; , \; 2 \; , \; \ldots \} \) of random variables defined as follows. Let \( a_0 \) be the number of p.c.u.'s in the queue at the end of the effective red time. Let these p.c.u.'s be numbered from 1 to \( a_0 \) in order of their departure, and let succeeding p.c.u.'s be numbered from \( a_0 + 1 \) onwards. Then let \( a_1 \) be the number of arrivals during the departure of p.c.u.'s numbered 1 to \( a_0 \), and for \( j > 1 \) let \( a_j \) be the number of arrivals during the departure of p.c.u.'s numbered \( 1 + \sum_{i=0}^{j-2} a_i \) to \( \sum_{i=0}^{j-1} a_i \), the sequence continuing until one of the \( a_j \) is zero, when the queue clears. By analysing the delay incurred during the departure of these successive groups of \( a_0 \), \( a_1 \), \( a_2 \) \ldots p.c.u.'s, and using the fact that \( a_{j+1} \) depends only on \( a_j \), McNeil showed that the average delay incurred during an effective green time that lasted until the queue cleared would be

\[
\frac{1}{2s(1 - y)^2} E \left\{ (1 + yI - y)a_0 + (1 - y)a_0^2 \right\}
\]

Then, proceeding like Winsten (see section 2.4) he deduced that

\[
d = \frac{(1 - \lambda)}{2(1 - \lambda x)} \left\{ c(1 - \lambda) + \frac{2E(Q_0)}{q} + s \left( 1 + \frac{I}{1 - \lambda x} \right) \right\}
\]

In order to use the methods of Bailey (1954) to investigate the overflow, McNeil had to assume, as did Newell in
the case of Poisson arrivals, an upper bound for the number of
departures per cycle, and thus obtain an upper bound for $E(Q_o)$. He took $gs$ as the maximum number of departures per cycle, and
assumed that $gs$ is an integer. Like Newell (1960) and Meiss (1962) he expressed the upper bound for $E(Q_o)$ in terms of
$$\sum_{j=1}^{gs-1} \frac{1}{1-z_j},$$
where the $z_j$ are the zeros inside the unit circle of
$z^{gs} - A_0(z)$, $A_t(z)$ being the probability generating function
of the number of arrivals in an interval of duration $t$.

By further application of the theory of functions of a
complex variable, however, McNeil obtained the upper bound for
$E(Q_o)$ in the form
$$\sum_{n=1}^{\infty} \frac{1}{n} \sum_{j=1}^{\infty} \frac{j}{(j+gsn)!} \left[ \left( \frac{1}{z} \right)^{j+gsn} A_{cn}(z) \right]_{z=0}.$$ 

In the case of Poisson arrivals, McNeil calculates the average
delay per p.c.u. given by substituting this upper bound for
$E(Q_o)$ in his expression for $d$. For a wide range of values
of $x$ and $gs$, the results are between 6 per cent above and
2 per cent below those given by Webster's expression.

McNeil (1968b) extends his analysis to the case where
the intervals between successive departures in the same cycle
are independently and identically distributed random variables
with mean $1/s$ and coefficient of variation $C_s$. The effect
is to replace $I$ by $I + \lambda x C_s^2$ in the expression for $d$ in
terms of $E(Q_o)$, and to replace $A_{cn}(z)$ in the expression
for the upper bound of $E(Q_c)$ by \( \left\{ z^{A_c}(z)/B(z) \right\}^n \), where $B(z)$ is the probability generating function of the number of departures in a cycle in which the queue does not clear.

Only very limited numerical comparisons of the various expressions for delay described in this and the preceding three sections have so far been made. More extensive comparisons of the expressions with each other and with observed data would be valuable, but are beyond the scope of this thesis.

2.7. Calculation of signal settings.

This section discusses previous methods for the calculation of settings for signals at a single intersection. All these methods depend on the choice, for each stage in the signal cycle, of one representative approach defined as follows.

**DEFINITION:** of the approaches that have right of way in a given stage, the approach that has the highest flow ratio is the representative approach for that stage.

The methods discussed in this section all assume that the representative approach for any stage has right of way in that stage only. This assumption will no longer be needed in Chapter 3.

2.7.1. The stage matrix. The stage matrix is usually determined by the engineer, having regard to the layout of the intersection, the traffic flows, and considerations of safety. In complicated cases, however, it may not be easy for the engineer to identify all the possible sets of mutually compatible approaches, and a technique proposed by Stoffers
(1968) may be helpful. Stoffers constructs a graph (Berge 1962), called the compatibility graph, in which there is one node for each approach, and two nodes are joined by an edge if and only if the corresponding approaches are compatible. Sets of mutually compatible approaches then correspond to the nodes of complete subgraphs of the compatibility graph. Since it is usually desirable to give right of way to as many approaches as possible in each stage of the signal cycle, the engineer will usually confine his attention to maximal complete subgraphs. Having identified these, he must select those that are to correspond to stages in the signal cycle so that the sequence of stages can be chosen to make the stage matrix satisfy the restriction imposed in section 1.23.

2.72. The cycle time. Clayton (1940) noted that with regular arrivals the proportion $\lambda_i$ of the cycle that is effectively green for the $i$th stage must be at least $y_i$, the flow ratio on the corresponding representative approach. If there are $m$ stages, then

$$c \sum_{i=1}^{m} \lambda_i = c - L$$

where $L$ is the lost time. Clayton deduced that if arrivals are regular, then the least cycle-time that will allow the traffic to pass is $L/(1-Y)$, where $Y = \sum_{i=1}^{m} y_i$. He pointed out that in practice a slightly larger cycle-time should be used to allow for variability of arrivals.

The choice of cycle-time for an intersection with two stages and with regular arrivals on each representative
approach was investigated by Wardrop (1952). Numbering the stages so that

\[ \frac{1}{q_1} - \frac{1}{s_1} \geq \frac{1}{q_2} - \frac{1}{s_2} \],

where \( q_1 \) and \( s_1 \) refer to the representative approach of stage \( i \), he showed that the average delay to p.c.u.'s on the representative approaches was minimised by setting

\[ c = L/(1 - y_1 - y_2) \]

(i.e. Clayton's minimum cycle-time) provided that

\[ q_1(1 - y_1) - q_2(1 - y_2) < 2q_1y_2 \]

and otherwise by setting

\[ c = L \left\{ \frac{q_1}{y_2q_1 + (1-y_1)(1-y_2)q_2} \right\}^{\frac{1}{2}} \]

Webster's simulations (1958) showed that when arrivals are Poisson the cycle-time giving minimum average delay to p.c.u.'s on the representative approaches is in the neighbourhood of twice Clayton's minimum, and can, in a wide range of cases, be approximated by

\[ c = \frac{1.5L + 5}{1 - Y} \]

He also showed algebraically that if the average delay per p.c.u. is estimated by the first two terms of his expression (see section 2.51) and \( \lambda_1 \propto y_1 \) (see next section), this delay-minimising cycle-time is given more accurately by

\[ c = \frac{2L}{1 - Y} \left\{ 1 + \frac{(Y^2 - Y + Z)^{\frac{1}{2}}}{2} - Y \right\} \]
where
\[
\frac{1}{\mathcal{L}} = \frac{1}{16mY^2} \sum_{i=1}^{m} \frac{y_i s_i}{(1-y_i)} \left\{ 4(Y_y - y_i)^2 - y_i^2(1-Y)^2 \right\},
\]

This formula results from a simplification that is justified only if \( (Y^2 - Y + Z) \frac{1}{2} - \frac{Y}{2} \) is small compared with 1. If this is not the case, a more complicated formula, also derived by Webster, must be used.

Miller (1963) considered a two-stage intersection with two approaches having right of way in each stage; and sought to minimise the average delay to p.c.u.'s on all approaches, as estimated by applying the first two terms of his expression (see section 2.6) to each approach. He defined first the rate of delay, \( D \), for the whole intersection as the sum of the products \( q_i d_i \) for all four approaches, and then the quantities \( \pi_1 = \lambda_1 c/(c-L) \) and \( \pi_2 = \lambda_2 c/(c-L) \). By obtaining approximate expressions for \( \partial D/\partial c \), \( \partial D/\partial \pi_1 \) and \( \partial D/\partial \pi_2 \), and setting these equal to zero he showed that the required cycle-time is given approximately by
\[
c = \left( 1 - \frac{y_i}{\pi_i} \right)^{-1} \left\{ L + 2 \left( \frac{I_i L}{s_i} \right)^{\frac{1}{2}} \right\},
\]

where stage \( i \) is the stage whose representative approach has the higher flow ratio, and \( y_i \), \( I_i \), and \( s_i \) refer to the representative approach of stage \( i \). The determination of \( \pi_i \) is discussed in the next section.

The choice of cycle-time may also be influenced by queue-length, especially when intersections are close together, and Miller states (1964) that the cycle-time that minimises the queue-length exceeded on average one in 20 cycles on the
approach with highest flow ratio is obtained approximately by replacing \( 2 \) by \( \sqrt{2} \) in the numerator of his cycle-time formula. Webster (1958) had given tables showing the queue-lengths exceeded on average once in 20 and once in 100 cycles on approaches with various values of \( \lambda \), \( q_c \), and degree of saturation, \( x \).

A third criterion considered by some authors is the proportion of cycles in which overflow occurs. Matson and McGrath (1954) suggested that this proportion should not exceed 5 per cent for any approach, and found that in one example with Poisson arrivals a cycle-time about 2\( \frac{1}{2} \) times as long as Clayton's minimum was required to satisfy this condition. Miller (1964) gave a cycle-time formula corresponding to the same condition, but found that it gave impractically long cycle-times except when traffic was very light.

All three criteria were taken into account by Bone, Martin and Harvey (1964). Working largely from Webster's results, they produced a computer program to tabulate, for a number of different cycle-times near to that given by Webster's formula, the average delay per p.c.u., the queue-lengths exceeded on average once in 20 cycles, and the proportion of cycles in which overflow would be expected to occur.

2.73. Allocation of green time. With Clayton's (1940) minimum cycle-time, the total effective green time must be shared between the stages in proportion to the flow ratios on the representative approaches; i.e. \( \lambda_i \alpha y_i \). If not, at least one approach will be overloaded. The results of Webster's
simulations (1958) showed that, when arrivals are Poisson, such an allocation of green time was, in a wide range of cases, very close to that which gave the least average delay per p.c.u.

Miller's theoretical treatment of the two-stage four-approach intersection (see last section) showed that the least average delay per p.c.u. was obtained when, approximately,

$$\pi_1 = \frac{(y_1 I_1)^{\frac{1}{2}} + 1.2(y_1 I_2)^{\frac{1}{2}} \{ (y_1 I_2)^{\frac{1}{2}} - (y_2 I_1)^{\frac{1}{2}} \} }{(y_1 I_1)^{\frac{1}{2}} + (y_2 I_2)^{\frac{1}{2}}}$$

and \( \pi_2 \) is obtained by interchanging the subscripts 1 and 2. In this expression, the constant \( 1.2 \) is an approximation for \( c/(c-L) \). One of the resulting values \( \pi_1 \) is used in calculating \( c \) (see last section), and if the result is such that \( c/(c-L) \) differs greatly from \( 1.2 \), the new value can be inserted in the above expression and the calculation repeated. Miller showed that if \( \pi_1 \) and \( \pi_2 \) are obtained from the above expression \( \pi_1/\pi_2 \rightarrow y_1/y_2 \) as \( y_1 + y_2 \rightarrow 1 \) (i.e. as the intersection becomes fully loaded). Since, for given cycle-time, \( \pi_1/\pi_2 = \lambda_1/\lambda_2 \), this lends support to Webster's results.

The relation between queue-lengths and the allocation of green-time has also been considered. In a theoretical study with the discrete time assumption, Uematu (1958) considered the crossing of two one-way streets controlled by a signal with specified cycle time. He assumed that for each street a certain queue-length that was to be exceeded as rarely as possible was specified. He obtained equations determining how the green time should be allocated so as to maximise the expectation of the time elapsing between the start of a cycle
in which there was no overflow on either street and the next occasion on which either of the specified queue-lengths was exceeded.

2.74. Comment. Whilst some of the work described in the last two sections, and particularly that of Webster, has justly found widespread application, the problem of calculating settings for traffic signals at a single intersection has so far been treated by approximate methods applicable to a limited range of intersections. The next chapter is devoted to a more general treatment of this problem.
3. **CALCULATION OF DELAY-MINIMISING SETTINGS FOR A SINGLE FIXED-TIME TRAFFIC SIGNAL**

This chapter describes a method for calculating the cycle time and allocation of green time at an intersection controlled by a fixed-time traffic signal so as to minimise the estimated average delay to all traffic passing through the intersection. The data required are the average arrival rate and the saturation flow on each approach, and the amounts of lost time occurring after each stage in the signal cycle. The intersection layout and the sequence in which the approaches have right of way are subject only to the restriction that each approach can have right of way for only one period in each cycle; this period may comprise one or more stages. A minimum can be imposed on the effective green time for each stage and a maximum can be imposed on the cycle time. The method can also be used to calculate settings in cases where the green times for some of the stages, or the cycle time, or both, are specified.

3.1. Preliminaries

Two preliminaries are necessary. Firstly, one of the expressions discussed in sections 2.3 - 2.6 is chosen to estimate the delay to traffic passing through the intersection. Secondly, the lost time is defined more precisely than was possible in Chapter 1.

3.11. Choice of expression for delay. The average delay per p.c.u. passing through the intersection will be estimated for each approach by Webster's expression (see section 2.51) in its
simplified form:

\[ d = \frac{9}{10} \left\{ \frac{c(1 - \lambda)^2}{2(1 - \lambda x)} + \frac{x^2}{2q(1 - x)} \right\}. \]

This expression is chosen for the following reasons:

(a) It has been tested by comparison with observed and simulated data (Webster 1958).

(b) It requires measurement of the saturation flow and the average arrival rate only.

(c) It is reasonably simple in form.

Of the other expressions considered in the previous chapter, only Miller's has been similarly tested by comparison with observed data, and his expression is less simple and requires the measurement of the extra quantity \( I \). The various expressions based on complex variable theory, though elegant in their derivation, are difficult to use for the present purpose because their dependence on the parameters to be determined is so complicated.

Whilst it has not so far been possible to include the third term of Webster's expression, an indication whether its inclusion would be likely to affect the results appreciably can be obtained in most cases.

3.12. Lost time and effective green times. The effective green time for an approach was defined in detail in section 2.25; the effective green time for a stage will now be defined in terms of the green times of the approaches.

DEFINITION: The effective green time for a stage in the signal
cycle is the time that is effectively green for every approach that has right of way in that stage.

The amount of lost time between one stage and the next can now be defined.

**DEFINITION:** The *lost time following a stage* in the signal cycle is the time from the end of the effective green time for that stage to the beginning of the effective green time for the next stage.

Whilst the estimated delay is expressed in terms of the effective green times for the approaches, it is those for the stages that have to be determined. To obtain the relation between the two sets of green-times it is necessary to define one further quantity. If an approach does not have right of way in every stage, the effective green time for that approach may overlap either the lost time before the first stage in which the approach has right of way, or the lost time following the last such stage, or both. This overlap is measured by the extra effective green time defined below. The somewhat artificial extension of the definition to an approach that does not have right of way in every stage is designed to enable the same algebraic expression to apply to both kinds of approach.

**DEFINITION:** The *extra effective green time* for an approach that does not have right of way in every stage is the sum of the amounts by which the effective green time for that approach overlaps the lost times before the first and after the last stage in which the approach has right of way. For an approach that has right of way in every stage it is minus the amount by which the cycle-time exceeds the effective green time for that
Let \( m \) = the number of stages in the signal cycle
\( n \) = " " approaches
\( A \) = \((a_{ij})\) be the \( mxn \) stage matrix
\( c \) = the cycle time
\( \lambda_{ic} \) = the effective green time for stage \( i \)
\( L_i \) = the lost time following stage \( i \)
\( L = \sum_{i=1}^{m} L_i \) = the total lost time
\( \lambda_0 = \frac{L}{c} \)
\( \lambda = (\lambda_0, \lambda_1, \lambda_2, \ldots, \lambda_m) \)

Then the signal settings are completely determined by the vector \( \lambda \). By definition of the lost times \( L_i \) it follows from that of \( \lambda_0 \) that
\( \sum_{i=0}^{m} \lambda_i = 1 \)

It will be convenient to write
\( S(\lambda) = \sum_{i=0}^{m} \lambda_i \)

Consider now the approaches; let
\( \Lambda_jc \) = the effective green time for approach \( j \)
\( X_j \) = the extra effective green time for approach \( j \)

Consider first an approach \( j \) that does not have right of way in every stage. Its effective green time comprises the green times for the stages in which it has right of way, the lost time between any two successive such stages, and its extra effective green time, i.e.
\( \Lambda_jc = X_j + a_{mj}a_{ij}L_m + \sum_{i=1}^{m-1} a_{ij}a_{i-1,j}L_i + \sum_{i=1}^{n} a_{ij}\lambda_i c \),
since $a_{ij} = 1$ or $0$ according as approach $j$ does or does not have right of way in stage $i$. If approach $j$ does have right of way in every stage then $a_{ij} = 1$ for all $i$, and the right hand side of the above equation becomes

$$X_j + \sum_{i=1}^{m} (L_i + \lambda_i c), \text{ i.e. } X_j + c,$$

which, by definition of $X_j$ for such an approach, is again the effective green time $\Lambda_j c$. Now let

$$a_{ij} = \frac{1}{L_i} \left\{ X_j + a_{ij} \sum_{i=1}^{m} L_i + \sum_{i=0}^{m} a_{ij} \lambda_i a_{ij} L_i \right\} \quad (j=1,2,\ldots,n)$$

so that

$$\Lambda_j c = a_{0j} L + \sum_{i=1}^{m} a_{ij} \lambda_i c,$$

or, since $\lambda_i = \frac{1}{c}$,

$$\Lambda_j = \sum_{i=0}^{m} a_{ij} \lambda_i \quad (j=1,2,\ldots,n).$$

One property of the $a_{0j}$ will be required; the numerator of $a_{0j}$ is either zero or represents part but not all of the lost time $L$. Hence

$$0 \leq a_{0j} < 1 \quad (j=1,2,\ldots,n).$$

3.2. Mathematical statement of the problem

For approach $j$ ($j=1,2,\ldots,n$) let

- $d_j$ = the average delay per p.c.u. on this approach
- $q_j$ = the average arrival rate
- $x_j$ = the degree of saturation
- $y_j$ = the flow ratio.
Then \( d_j \) is estimated by

\[
d_j = \frac{9}{10} \left\{ \frac{c(1 - \Lambda_j)^2}{2(1 - \Lambda_j x_j)} + \frac{x_j^2}{2q_j(1 - x_j)} \right\}.
\]

It will be convenient to rewrite this in the form

\[
d_j = \frac{9}{10} \left\{ \frac{L(1 - \Lambda_j)^2}{2\lambda_0(1 - y_j)} + \frac{y_j^2}{2q_j\Lambda_j(\Lambda_j - y_j)} \right\},
\]

and to define, for \( j = 1, 2, \ldots, n \),

\[
f_j(\lambda) = \frac{Lq_j(1 - \lambda)^2}{2(1 - y_j)}
\]

and

\[
g_j(\lambda) = \frac{y_j^2}{2\lambda(\lambda - y_j)}.
\]

3.21. Rate of delay. A quantity used by Miller (1963) will now be defined.

DEFINITION: The rate of delay for an intersection is the average total rate at which delay is incurred on all the approaches to the intersection.

Since, on approach \( j \), an average of \( q_j \) p.c.u.'s per second are delayed by an average of \( d_j \) seconds, the rate of delay for the intersection is \( \sum_{j=1}^{n} q_j d_j \).

In the above expression for \( d_j \), the coefficient \( \frac{9}{10} \) is the same for all approaches, so that the rate of delay for the intersection is proportional to

\[
D(\lambda) = \sum_{j=1}^{n} \left\{ \frac{1}{\lambda} f_j(\Lambda_j) + g_j(\Lambda_j) \right\}.
\]

\( \lambda \) will be chosen to minimise \( D \) subject to the constraint
S(λ) = 1 (see section 3.12) and to certain other constraints discussed in the next three sections.

3.22. Capacity constraints. So that, in the long run, every approach has right of way for long enough to allow all arriving traffic to pass through the intersection, the degree of saturation on each approach must be less than 1, i.e.

\[ x_j < 1 \quad (j = 1, 2, \ldots, n) \]

or, since \( y_j / x_j \),

\[ \sum_{i=0}^{\infty} a_j \lambda_i > y_j \quad (j = 1, 2, \ldots, n) \]

These constraints express in terms of \( \lambda \) the condition that, with the resulting signal settings the capacity of each approach must exceed its average arrival rate. They will be called the capacity constraints.

3.23. Cycle time constraint. If a maximum is imposed upon the cycle time, let it be \( c_M \), and if a cycle time is specified, let it be \( c_s \). Then the constraint \( c < c_M \) or \( c = c_s \) can be written

\[ \lambda_o \geq \frac{L}{c_M} \quad \text{or} \quad \lambda_o = \frac{L}{c_s} \]

In any case, the cycle time must be positive, so that, if neither of the above constraints applies, then \( \lambda_o > 0 \) (strictly, \( \lambda_o > 0 \)), but it will be seen in section 3.41 that as \( \lambda_o \rightarrow 0 \), \( D \rightarrow \infty \), so that the \( \lambda \) that minimises \( D \) cannot be affected by permitting, for the present, \( \lambda_o = 0 \).
Let
\[
k_0 = \begin{cases} 
\frac{L/c_M}{c} & \text{if it is required that } c \leq c_M \\
\frac{L/c_S}{c} & \text{"} \quad \text{"} \quad \text{"} \quad \text{"} \quad c = c_S \\
0 & \text{otherwise}
\end{cases}
\]

Then the cycle time constraint can be written
\[
\lambda_o = \begin{cases} 
k_0 & \text{if the cycle time is specified} \\
\geq k_0 & \text{otherwise.}
\end{cases}
\]

From the fact that the lost time can be only part of the cycle time it follows that
\[
0 \leq k_0 < 1.
\]

3.24. Green time constraints. If a minimum is imposed on the effective green time for stage \(i\), let it be \(g_{iM}\), and if no such minimum is imposed let \(g_{iM} = 0\). Then it is required that \(\lambda_i c \geq g_{iM}\), \(i = 1, 2, \ldots, m\). Let
\[
k_i = \frac{g_{iM}}{L} \quad (i = 1, 2, \ldots, m)
\]

Then, since \(\lambda_o = L/c\), the minimum green time constraints can be written
\[
\lambda_i \geq k_i \lambda_o \quad (i = 1, 2, \ldots, m)
\]

If the cycle time is specified, or is subject to a maximum, then this cannot be less than the sum of the lost time and all the minimum green times. Hence the \(k_i\) \((i = 1, 2, \ldots, m)\) must satisfy the condition
\[
\sum_{i=1}^{m} k_i \leq \frac{1}{k_0} - 1 \quad \text{if } k_0 \neq 0.
\]

All the constraints that appear in the mathematical statement of the problem have now been defined. The following definitions distinguish the values of \(\lambda\) that satisfy all these constraints, and those that minimise \(D(\lambda)\).
DEFINITIONS: A feasible solution to the problem is a value of \( \lambda \) that satisfies all the constraints. An optimal solution is a feasible solution that minimises \( D(\lambda) \).

It will later be required that \( k_i < 1 \) \( (i = 1, 2, \ldots, m) \). The following Lemma shows that this restriction implies no loss of generality.

**Lemma 3.** If one or more of the minimum green constraints is \( \lambda_i \geq k_i \lambda_0 \) with \( k_i > 1 \), the problem of choosing an optimal \( \lambda \) can be restated without change of form as that of choosing a vector \( \lambda^* \), where \( \lambda^* \) is related to \( \lambda \) so that the constraints \( \lambda_i \geq k_i \lambda_0 \) become \( \lambda_i^* \geq k_i \lambda_0^* \), where

\[
0 \leq k_i^* \leq 1 \quad (i = 1, 2, \ldots, m).
\]

**Proof.** Suppose, w.l.o.g., that \( g_{1M} = \max_{1 \leq i \leq m} (g_{1M}) \). Let the problem be restated so that an amount \( g_{1M} \) of the effective green time for stage 1 is added to the lost time following stage 1 and to the extra effective green time for any approach that has right of way in stage 1 but not in stage 2. Then, for any given cycle time, the effective green time for each approach is the same as before, but the effective green time for stage 1 is no longer subject to any minimum except zero. Each feasible solution to the new problem corresponds to a feasible solution to the original problem, and vice versa. Denoting parameters of the new problem by the superscript \( * \), the new values are the same as the original ones except that

\[
L_1^* = L_1 + g_{1M},
\]

\[
X_j^* = X_j + a_{1j}(1 - a_{2j})g_{1M},
\]

\[
g_{1M}^* = 0.
\]
and other parameters depending on these three are changed by substituting the new values for the old. In particular

\[ L^* = L + \xi_{iM} \]
\[ \lambda_o^* = L^*/c = \lambda_o L^*/L \]
\[ \lambda_i^* = \lambda_i - \frac{\xi_{iM}}{c} \]
\[ \lambda_i^* = \lambda_i \quad (i = 2, 3, \ldots, m) \]

and \( D^*(\lambda^*) = D(\bar{\lambda}) \) for all corresponding feasible \( \lambda \) and \( \lambda^* \). Hence a \( \lambda^* \) that minimises \( D^* \) corresponds to an optimal \( \bar{\lambda} \).

Consider now the minimum green constraints in the new problem. These are \( \lambda_i^* \geq k_i^* \lambda_o^* \), where \( k_i^* = \xi_{iM}/L^* \).

Thus \[ k_i^* = 0 \]

and \[ 0 \leq k_i^* = \frac{\xi_{iM}}{L + \xi_{iM}} < \frac{\xi_{iM}}{\xi_{iM}} \leq 1 \quad (i=2,3,\ldots,m), \]
i.e. \[ 0 \leq k_i^* \leq 1 \quad (i=1,2,\ldots,m) \] as required.

The other type of green time constraint that can be provided for is the specification of effective green times for one or more stages. This is done not by further mathematical constraints, but by restating the problem so that the determination of the unspecified settings is equivalent to the determination of settings for an intersection whose signal cycle has fewer stages. Suppose, w.l.o.g., that the green time for stage \( m \) is specified as \( \xi_m \); then the procedure is as follows:

(a) For each approach \( j \), replace \( X_j \) by
\[ X_j + a_{m-lj} a_{mj}(1-a_{lj})(L_{m-l}+g_m) + a_{mj} a_{lj}(1-a_{m-lj})(L_m+g_m) + a_{mj}(1-a_{lj})(1-a_{m-lj})g_m - a_{lj} a_{m-lj}(1-a_{mj})(L_{m-l}+L_m+g_m) \]

(b) Replace \( L_{m-l} \) by \( L_{m-l} + L_m + g_m \)

(c) Disregard stage \( m \) and delete the corresponding row of \( A \)

(d) Proceed as for an intersection with \( m-1 \) stages.

The effect of this is that the specified green time \( g_m \) is treated as part of the lost time when determining the remaining settings. The adjustments to the extra effective green times \( X_j \) are such as to keep the effective green time for each approach unchanged. If green times are specified for several stages, these are disregarded one by one in the above manner (with appropriate renumbering of stages if those to be disregarded are not consecutive). There can be at most \( m-1 \) such stages (\( m-2 \) if the cycle time is specified), otherwise the settings would be completely determined by these specifications.

NOTE: If the cycle time is specified, then approaches that have right of way only in stages whose green times are specified can be ignored, because the average delays on these approaches are determined by the specifications. The specified green times must form a sufficiently large proportion of the specified cycle time for the degrees of saturation on these approaches to be less than 1.

3.25. Example. Consider a cross-roads controlled by a signal with a two-stage cycle, traffic from north and south having
right of way in stage 1, and that from east and west in stage 2. Suppose that the traffic forms just one queue on each of the four roads, so that there are four approaches, and let these be numbered in the order N, S, E, and W.

Then

\[ A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \]

and

\[ \lambda = (\lambda_0, \lambda_1, \lambda_2), \]

where

\[ S(\lambda) = \lambda_0 + \lambda_1 + \lambda_2 = 1 \]

Suppose also that the extra effective green time is zero for each approach except for approach number 4, for which it is positive. Then the capacity constraints take the form

\[ \lambda_1 > y_1 \]
\[ \lambda_1 > y_2 \]
\[ \lambda_2 > y_3 \]
\[ a_{04}\lambda_0 + \lambda_2 > y_4 \]

Suppose that a maximum is imposed on the cycle time and minima on the green times; these constraints are

\[ \lambda_0 \geq k_0 \]
\[ \lambda_1 \geq k_1\lambda_0 \]
\[ \lambda_2 \geq k_2\lambda_0 \]

Figure 4 shows the intersection of the plane \( \lambda_0 + \lambda_1 + \lambda_2 = 1 \) with the positive octant, and the lines in this plane corresponding to the various constraints. The part of the plane corresponding to the set of all feasible solutions is indicated.

3.26. Summary. The problem of calculating delay-minimising settings for a single fixed-time signal has been expressed
Fig. 4. CONSTRAINTS AND FEASIBLE SOLUTIONS FOR A CROSS-ROADS CONTROLLED BY A SIGNAL WITH A TWO-STAGE CYCLE SUBJECT TO A MAXIMUM CYCLE TIME AND MINIMUM GREEN TIMES
mathematically in the following form:

To minimise

\[ D(\lambda) = \sum_{j=1}^{n} \left\{ \frac{1}{\lambda_j} f_j(\Lambda_j) + g_j(\Lambda_j) \right\}, \]

where \( \Lambda = (\lambda_0, \lambda_1, \ldots, \lambda_m) \)

\[ f_j(\lambda) = \frac{1+e_j(1-\lambda)^2}{2(1-y_j)} \]

\[ g_j(\lambda) = \frac{y_j^2}{2\lambda(\lambda-y_j)} \] \( \quad (j = 1, 2, \ldots, n), \)

\[ \sum_{i=0}^{m} a_{ij} \lambda_i \]

and \( 0 \leq a_{ij} < 1 \)

subject to the constraints

\[ S(\lambda) = \sum_{i=0}^{m} \lambda_i = 1 \]

\[ \sum_{i=0}^{m} a_{ij} \lambda_i > y_j \] \( \quad (j = 1, 2, \ldots, n) \)

\[ \lambda_0 = k_0 \quad \text{or} \quad \lambda_0 > k_0 \]

and \( \lambda_i > k_i \lambda_0 \) \( \quad (i = 1, 2, \ldots, m), \)

where \( 0 \leq k_i \leq 1 \) \( \quad (i = 0, 1, 2, \ldots, m) \)

and, if \( k_0 \neq 0 \), \( \sum_{i=1}^{m} k_i \leq \frac{1}{k_0} - 1 \).

A value of \( \lambda \) satisfying all these constraints is called a feasible solution, and a feasible solution that minimises \( D(\lambda) \) is called an optimal solution.

This mathematical problem is discussed in sections 3.3 - 3.5; throughout the discussion it will be useful to denote real space of \( p \) dimensions by \( \mathbb{R}^p \) and the vector in \( \mathbb{R}^{m+1} \), whose \( i \)th component is \( 1 \), and all other components are
zero by $e_i (i = 0, 1, 2, \ldots, m)$
e.g. $e_0 = (1, 0, 0, \ldots, 0)$
$e_m = (0, 0, 0, \ldots, 1)$.

3.3. Existence and location of a feasible solution

The first step towards finding a value of $\lambda$ that minimises $D(\lambda)$, subject to all the constraints, is to find some feasible solution; any such solution will serve as a starting point. The conditions imposed on the $k_i$
(i = 0, 1, 2, \ldots, m) in sections 3.23 and 3.24 ensure that there exists at least one $\lambda$ satisfying simultaneously the cycle time constraint, the minimum green constraints, and
the constraint $S(\lambda) = l$. Such a $\lambda$ will not necessarily also satisfy the capacity constraints; whether it does so
(i.e. whether it is a feasible solution) will depend on the stage matrix $A$, the extra effective green times $X_j$, and
the flow ratios $y_j$. If the flow ratios are too large, no feasible solution exists; this corresponds to the intersection being overloaded with traffic. A necessary and sufficient condition for the existence of a feasible solution, and an algorithm for finding one whenever one exists, are given in
section 3.32. In most cases, however, a feasible solution can be found much more easily, and one simple method is described in the next section. The need for the more general method is illustrated by an example.

3.31. A simple method that often finds a feasible solution.
Let $n_j$ be the number of stages in which approach $j$ has right of way; then
\[ n_j = \sum_{i=1}^{m} a_{ij} \quad \left\{ \begin{array}{l}
0 \leq n_j \leq m 
\end{array} \right\} (j = 1, 2, \ldots, n) \]

and

(the value \( n_j = 0 \) will not usually occur unless the problem has been restated so as to specify the green times for one or more stages - see section 3.24).

The method begins by setting \( \lambda_0 = k_0 \). If the cycle time is specified, \( \lambda_0 \) remains equal to \( k_0 \) throughout, and any approach \( j \) for which \( n_j = 0 \) will have been ignored. If the cycle time is not specified, then for any approach \( j \) that has \( n_j = 0 \), replace \( \lambda_0 \) by \( \max (\lambda_0, y_j/a_{0j}) \). Repeat this for each such approach in turn. (\( a_{0j} \) cannot be zero for any such approach, for if it were, the approach would never have right of way). Any further increase in \( \lambda_0 \), however small, will now be sufficient to satisfy the capacity constraints for all approaches \( j \) such that \( n_j = 0 \).

Now let \( \lambda_i = k_i \lambda_0 \) \((i = 1, 2, \ldots, m)\), thus satisfying the minimum green constraints. The capacity constraints for approaches \( j \) such that \( n_j \geq 1 \) are now examined and the \( \lambda_i \) \((i = 1, 2, \ldots, m)\) increased where necessary so that any further increase, however small, will be sufficient to satisfy the constraints. For each such approach in turn, in order of increasing \( n_j \), replace \( \lambda \) by

\[ \lambda + \frac{1}{n_j} \left\{ \max (0, y_j - \sum_{i=0}^{\infty} a_{yi} \lambda_i) \right\} \sum_{i=0}^{\infty} a_{yi} e_i. \]

Since this process leaves \( \lambda_0 \) unchanged and does not decrease the \( \lambda_i \) \((i = 1, 2, \ldots, m)\), the minimum green and cycle
time constraints are still satisfied.

With the resulting \( \lambda \), let \( \sigma = 1 - S(\lambda) \). If \( \sigma > 0 \), replace \( \lambda \) by

\[
\lambda + \frac{\sigma}{m} \sum_{i=1}^{m} e_i
\]

if the cycle time is specified

and

\[
\lambda + \frac{\sigma}{(m+1) \sum_{i=0}^{m} e_i}
\]

if not.

In each case the minimum green constraints are still satisfied (this is true in the latter case despite the increase in \( \lambda_0 \) because \( k_i \leq 1 \) for \( i = 1, 2, \ldots, m \)), and so is the cycle time constraint. Moreover, since \( \sigma > 0 \) this change in \( \lambda \) provides the further increase in its components that is sufficient to satisfy all the capacity constraints. Lastly, with the resulting \( \lambda \), by the choice of \( \sigma \), \( S(\lambda) = 1 \), and thus \( \lambda \) is a feasible solution. If \( \sigma \leq 0 \), the method fails to find a feasible solution; an example will now be given in which this occurs, but nevertheless a feasible solution exists.

**EXAMPLE.** Consider an intersection with 4 approaches controlled by a signal with a two-stage cycle. Let the stage matrix be

\[
A = \begin{pmatrix}
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1
\end{pmatrix}
\]

and let the extra effective green times be \( 2L/3 \) for approaches 1 and 3, and zero for approaches 2 and 4, so that

\[a_{01} = a_{03} = \frac{2}{3}\]

and

\[a_{02} = a_{04} = 0\]

Let the cycle time constraint by \( \lambda_0 \geq \frac{1}{10} \) and let there be no minimum green constraints. Let the flow ratios be
\[ y_1 = \frac{8}{15}, \quad y_2 = \frac{7}{4}, \quad y_3 = \frac{8}{15}, \quad \text{and} \quad y_4 = \frac{1}{4}. \]

Then, when \( \sigma \) is calculated in the above process, \( \lambda \) has the value \( \left( \frac{1}{10}, \frac{7}{15}, \frac{7}{15} \right) \), so that \( \sigma = -\frac{1}{30} \) and the method fails. A feasible solution does, however, exist in this case, viz. \( \lambda = \left( \frac{3}{10}, \frac{7}{20}, \frac{7}{20} \right) \).

3.32. A general method. Consider the set of values of \( \lambda \) satisfying the following amended constraints; in which the capacity constraints have been relaxed so that they can be used as constraints in an application of the simplex method of linear programming (see, e.g. Vajda 1961).

\[
\sum_{i=0}^{m} a_{ij} \lambda_i \geq y_j \quad (j = 1, 2, \ldots, n),
\]

\[ \lambda_o = k_o \quad \text{or} \quad \lambda_o = k_o, \]

and \[ \lambda_i \geq k_i \lambda_o \quad (i = 1, 2, \ldots, m). \]

**Lemma 4.** If any \( \lambda \) satisfying the amended constraints exists, then \( S(\lambda) \) attains a minimum, subject to the amended constraints.

**Proof.** If any \( \lambda \) satisfying the amended constraints exists, let one such \( \lambda \) be \( \lambda_1 \). Then, since \( k_i \geq 0 \) \( (i = 0, 1, 2, \ldots, m) \), the set of \( \lambda \) satisfying the amended constraints and such that \( S(\lambda) \leq S(\lambda_1) \) is a closed and bounded set in \( \mathbb{R}^{m+1} \). Being a continuous function of \( \lambda \), \( S(\lambda) \) therefore attains a minimum on this set; let it do so at \( \lambda = \lambda_0 \).

Then \( S(\lambda_0) \) must also be its minimum subject to the amended constraints, since the construction of the closed set excluded only \( \lambda \) such that \( S(\lambda) > S(\lambda_1) \geq S(\lambda_0) \).
Let $\sigma_0 = 1 - S(\lambda_0)$, with $\lambda_0$ defined as in the Lemma; then $\sigma_0$ is defined if and only if there exists a $\lambda$ satisfying the amended constraints. A necessary and sufficient condition for the existence of a feasible solution will now be obtained.

**THEOREM 3.** A feasible solution exists if and only if $\sigma_0$ is defined and either (a) $\sigma_0 = 0$ and $\lambda_0$ does not lie in any of the hyperplanes $\lambda_0 = 0$ and $\sum_{i=1}^{m} a_{ij} \lambda_i = y_j \ (j = 1, 2, \ldots, n)$, or (b) $\sigma_0 > 0$.

**PROOF:** If $\sigma_0$ is undefined, no $\lambda$ satisfying the amended constraints exists, and any feasible solution would satisfy them; hence no feasible solution exists.

If $\sigma_0 < 0$, then $S(\lambda_0) > l$; but any feasible solution would satisfy the amended constraints with $S(\lambda) = 1$. Hence, by definition of $\lambda_0$, no feasible solution exists.

If $\sigma_0 = 0$ and $\lambda_0$ lies in one or more of the hyperplanes $\lambda_0 = 0$ and $\sum_{i=1}^{m} a_{ij} \lambda_i = y_j$, then, since $a_{ij} \geq 0 \ (i = 0, 1, 2, \ldots, m; j = 1, 2, \ldots, n)$, one or more of the $\lambda_i$ must be increased to satisfy the capacity constraints or to make the cycle time finite. This would make $S(\lambda) > 1$; hence no feasible solution exists.

If $\sigma_0 = 0$ and $\lambda_0$ does not lie in any of these hyperplanes, then $\lambda_0$ is a feasible solution.

If $\sigma_0 > 0$ then $\lambda_0 + \frac{S_0}{m} \sum_{i=1}^{m} e_i$ or $\lambda_0 + \frac{S_0}{(m+1)} \sum_{i=0}^{m} e_i$ is a feasible solution according as the cycle time constraint is $\lambda_0 = k_0$ or $\lambda_0 \geq k_0$, since each of these values of $\lambda$
makes \( S(\lambda) = 1 \), the equal increases in the components of \( \lambda \)
satisfy any capacity constraints not already satisfied by \( \lambda_0 \),
and these equal increases do not violate the minimum green
constraints because \( 0 \leq k_i \leq 1 \) (\( i = 1, 2, \ldots, m \)).

The evaluation of \( \lambda_0 \), and hence the location of a
feasible solution if one exists and the recognition of cases
where none exists is a standard problem in linear programming,
soluble by the simplex method (Vajda 1961).

3.4. Existence and uniqueness of delay-minimising settings.

If no feasible solution exists, then there are no signal
settings satisfying all the constraints discussed in section
3.2, and in order to calculate settings, the possibility of
relaxing some of the constraints must be examined. On the
assumption that a feasible solution does exist, it will be
shown in this section that an optimal solution, and hence
delay-minimising settings, exists and that its uniqueness, or
otherwise, depends only on the stage matrix.

It is now necessary, as foreseen in section 3.23, to
restrict the set of feasible solutions slightly by requiring
that \( \lambda_0 > k_0 \), instead of \( \lambda_0 \geq k_0 \), if \( k_0 = 0 \), i.e. if
no maximum is imposed on the cycle time.

3.41. Existence. The constraint \( S(\lambda) = 1 \) restricts the
feasible solutions to a subspace, isomorphic to \( \mathbb{R}^m \), of
\( \mathbb{R}^{m+1} \). If the cycle time is specified, the feasible solutions
are further restricted to a subspace isomorphic to \( \mathbb{R}^{m-1} \).
Within these subspaces, since the capacity constraints are
strict inequalities, the feasible solutions do not form a
closed set. The existence of an optimal solution is established in Theorem 4 by showing that values of $\Lambda$ that minimise $D(\Lambda)$ are confined to a closed subset of the set of feasible solutions. A corollary confirms that even if the cycle time is not subject to any maximum the optimal solution corresponds to a finite cycle time.

**THEOREM 4.** If a feasible solution exists, then so does an optimal solution.

**PROOF:** For all feasible $\Lambda$, $\lambda_0$ is positive and, for $j = 1, 2, \ldots, n$,

$$f_j(\Lambda_j) > 0 \quad \text{and} \quad g_j(\Lambda_j) > 0 \quad .$$

Hence, for any particular $j$,

$$D(\lambda) > g_j(\Lambda_j)$$

$$= \frac{y_j^2}{2\Lambda_j(\Lambda_j - y_j)}$$

$$> \frac{y_j^2}{2(\Lambda_j - y_j)}$$

$$\to \infty \quad \text{as} \quad \Lambda_j \to y_j \quad .$$

Now let $\Lambda_1$ be any one particular feasible solution. For each $j$ $(j = 1, 2, \ldots, n)$, since $D(\lambda) \to \infty$ as $\Lambda_j \to y_j$, there exists a $Y_j$ such that if $y_j < \Lambda_j < Y_j$ then $D(\lambda) > D(\Lambda_1)$. It follows from the last inequality that when $\lambda = \Lambda_1$, $\Lambda_j > Y_j$ $(j = 1, 2, \ldots, n)$.

In the cases where the cycle time constraint is $\lambda_0 = k_0$ or $\lambda_0 \geq k_0 > 0$, consider the closed sets defined, in the subspaces $\{S(\lambda) = 1\} \cap \{\lambda_0 = k_0\}$ and $S(\lambda) = 1$ respectively, by the minimum green constraints together with the inequalities $\Lambda_j \geq Y_j$. $D(\lambda)$ is continuous on these
closed sets and is therefore bounded below and attains its lower bound there; let it do so at $\tilde{\lambda}_0$. Then for any $\tilde{\lambda}$ in either of these closed sets,
\[ D(\tilde{\lambda}) \geq D(\tilde{\lambda}_0). \]
In particular
\[ D(\tilde{\lambda}_1) \geq D(\tilde{\lambda}_0). \]
But, for any feasible solution $\tilde{\lambda}$ not in the appropriate closed set,
\[ D(\tilde{\lambda}) > D(\tilde{\lambda}_1) \]
by definition of the $Y_j$.
Hence $D(\tilde{\lambda}) \geq D(\tilde{\lambda}_0)$ for all feasible $\tilde{\lambda}$, and $\tilde{\lambda}_0$ is an optimal solution.

It remains to consider the case where the cycle time constraint takes the form $\lambda_0 > 0$. Any feasible $\tilde{\lambda}$ can be expressed in the form
\[ \tilde{\lambda} = \lambda_0 e_0 + (1 - \lambda_0) \mu, \]
where
\[ \mu_0 = 0, \]
\[ S(\mu) = 1, \]
and
\[ 0 < \lambda_0 < 1. \]
Now, for feasible $\tilde{\lambda}$,
\[ D(\tilde{\lambda}) \geq \frac{1}{\lambda_0} \sum_{j=1}^{n} f_j (\Lambda_j) \]
\[ = \frac{1}{2\lambda_0} \sum_{j=1}^{n} \frac{q_j}{(1-y_j)} \left( a_{0j} \lambda_0 + (1-\lambda_0) \sum_{i=1}^{m} a_{ij} \mu_i \right) \]
\[ = \frac{1}{2\lambda_0} \sum_{j=1}^{n} \left( \frac{q_j}{1-y_j} \sum_{i=1}^{m} a_{ij} \mu_i \right) + \frac{1}{2\lambda_0} \sum_{j=1}^{n} \frac{q_j}{(1-y_j)} (a_{0j} \sum_{i=1}^{m} a_{ij} \mu_i) \]
\[ \geq \frac{1}{2\lambda_0} \left( \min_{ij \in \Omega} \left( \frac{q_j}{1-y_j} \sum_{i=1}^{m} \mu_i \sum_{j=1}^{n} a_{ij} \right) \right) + \frac{1}{2\lambda_0} \sum_{j=1}^{n} \frac{q_j a_{0j}}{(1-y_j)} - \frac{1}{2\lambda_0} \left( \max_{ij \in \Omega} \left( \frac{q_j}{1-y_j} \sum_{i=1}^{m} \mu_i \sum_{j=1}^{n} a_{ij} \right) \right). \]
All the $a_{ij}$ are either 0 or 1 and, for each $i$, at least one of them has the value 1, since in every stage at least one approach has right of way. Also $\sum_{i=1}^{m} \mu_i = 1$.

Hence

$$i \leq \sum_{i=1}^{m} \mu_i \sum_{j=1}^{n} a_{ij} < n,$$

and, for feasible $\Delta$,

$$D(\Delta) > \frac{L}{2\lambda_0} \{ \min \left( \frac{a_{ii}}{1-y_i} \right) \} + \frac{L}{2} \sum_{j=1}^{n} y_{ij} \Delta_j - \frac{nL}{2} \{ \max \left( \frac{a_{ii}}{1-y_i} \right) \} \rightarrow \infty \text{ as } \lambda \rightarrow 0 \text{ uniformly for all } \lambda, \lambda_1, \cdots, \lambda_m.$$

Hence there exists a $Y_0 > 0$ such that, for all feasible $\Delta$ with $\lambda_0 < Y_0$, $D(\Delta) > D(\Delta_1)$.

The proof for the case where the cycle time constraint took the form $\lambda_0 \geq k_0 > 0$ shows, on putting $k_0 = Y_0$, that an optimal solution $\Delta_0$ with $D(\Delta_0) < D(\Delta_1)$ exists subject to the constraint $\lambda_0 \geq Y_0$. It has just been shown that, for all feasible $\Delta$ with $\lambda_0 < Y_0$, $D(\Delta) > D(\Delta_1)$. Hence $\Delta_0$ remains optimal when the constraint $\lambda_0 \geq Y_0$ is removed; i.e. an optimal solution exists when the cycle time constraint takes the form $\lambda_0 > 0$.

**COROLLARY.** Even if no maximum is imposed on the cycle time, the delay-minimising cycle time will not exceed $L/Y_0$, where $Y_0$ is as defined in the proof of the theorem.

3.42. **Uniqueness.** The uniqueness or otherwise of optimal solutions will now be investigated. The constraint $S(\Delta) = 1$, and possibly the cycle time constraint, are linear equations that restrict the feasible solutions to a subspace of $\mathbb{R}^{m+1}$. 
All the other constraints are linear inequalities in the $\lambda_i$ $(i = 0, 1, 2, \ldots, m)$. The set of feasible solutions is therefore convex; i.e. any point lying between two feasible solutions on the line joining them is itself a feasible solution. This will be used to show that if more than one optimal solution exists, then the line joining them has the direction of a vector $\mu = (\mu_0, \mu_1, \mu_2, \ldots, \mu_m)$ such that $\mu_0 = 0$ and $(\mu_1, \mu_2, \ldots, \mu_m)$ belongs to the null-space in $\mathbb{R}^m$ of the stage matrix $A$. The optimal solution is therefore unique if $A$ has rank $m$, and if not, any point between two optimal solutions on the line joining them will be shown to be an optimal solution.

The interpretation of these results is that all optimal solutions correspond to the same cycle time and to the same effective green time on each approach; they are therefore indistinguishable so far as the traffic is concerned. There can be more than one optimal solution only if the stages are such that different allocations of green time between the stages can result in the same effective green time on each approach. If there is more than one optimal solution, it is sufficient to find one of them in order to determine delay-minimising settings.

It is convenient first to make a further definition and to extend the notation.

**DEFINITION:** If $\lambda$ is a feasible solution, then the *permissible directions* at $\lambda$ are those of all vectors $\mu$ such that $\lambda + h\mu$ is a feasible solution for all sufficiently small
positive h.

In matrix operations, all vectors will be regarded as row-vectors. Let

\[ a = (a_{01}, a_{02}, \ldots, a_{on}) \]

and

\[ A^* = \begin{pmatrix} 1 & a \\ 0 & A \end{pmatrix} \]

The transpose of a matrix or vector will be denoted by the superscript T, and the gradient operator in \( \mathbb{R}^{m+1} \) by

\[ \nabla = \left( \frac{\partial}{\partial \lambda_0}, \frac{\partial}{\partial \lambda_1}, \frac{\partial}{\partial \lambda_2}, \ldots, \frac{\partial}{\partial \lambda_m} \right) \]

Let \( \Delta = (\Lambda_1, \Lambda_2, \ldots, \Lambda_n) \)

Then \( \lambda A^* = (\lambda, \Delta) \)

Let \( B(\Delta) = (b_{ij}(\Delta)) \)

be the \((m+1) \times (m+1)\) matrix of second derivatives of \( D(\Delta) \), and let the \((n+1) \times (n+1)\) matrix \( B^*(\Delta) = (b_{ij}^*(\Delta)) \) be defined by

\[
\begin{align*}
    b_{00}^*(\Delta) &= \frac{1}{\lambda_0} \sum_{j=1}^{n} f_j'(\Lambda_j) \\
    b_{0j}^*(\Delta) &= b_{j0}^*(\Delta) = -\frac{1}{\lambda_0} f_j'(\Lambda_j) \\
    b_{ij}^*(\Delta) &= \frac{1}{\lambda_0} f_j''(\Lambda_j) + g_{ij}^*(\Lambda_j) \\
    b_{ij}^* (\Delta) &= b_{ji}^* (\Delta) = 0
\end{align*}
\]

\( (1 \leq i < j \leq n) \)

Then \( B^* \) is the matrix of second derivatives of \( D \) when \( D \) is regarded as a function of \( n+1 \) variables \( \lambda_0, \Lambda_1, \Lambda_2, \ldots, \Lambda_n \).

The results proved in the following four lemmas are required in the proof of Theorem 5, the main theorem of this section.
LEMMA 5. If $\lambda$ is a feasible solution, then the function $D$ satisfies the conditions of Taylor's Theorem (Hardy 1955) in the neighbourhood of $\lambda$; in particular, if $\lambda + \mu$ is also a feasible solution, then

$$D(\lambda + \mu) = D(\lambda) + \mu \cdot VD(\lambda) + \frac{1}{2} \mu \cdot B(\lambda + h\mu)\mu^T,$$

where $0 < h < 1$.

PROOF: All the derivatives of the functions $f_j$ and $g_j$ exist in the neighbourhood of $\lambda$ unless $\lambda_j = y_j$ for some $j$, and those of $1/\lambda_0$ exist in the neighbourhood of $\lambda$ unless $\lambda_0 = 0$. Also, $D$ is continuous in the neighbourhood of $\lambda$ unless $\lambda_0 = 0$ or $\lambda_j = y_j$ for some $j$. But for feasible $\lambda$, $\lambda_j \neq y_j$ and $\lambda_0 \neq 0$. Hence, for feasible $\lambda$, $D$ satisfies the conditions of Taylor's Theorem in the neighbourhood of $\lambda$.

If $\lambda$ and $\lambda + \mu$ are each feasible, then, since the set of feasible solutions is convex, $\lambda + h\mu$ is feasible for $0 < h < 1$ and, by Taylor's Theorem,

$$D(\lambda + \mu) = D(\lambda) + \mu \cdot VD(\lambda) + \frac{1}{2} \mu \cdot B(\lambda + h\mu)\mu^T$$

for some $h$ such that $0 < h < 1$.


PROOF: The relation $\frac{\partial g_j}{\partial \lambda_i} = a_{ij}$ ($i = 0, 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$) will be used throughout this proof; it follows immediately from the definition of $\lambda_j$.

Let $G(\lambda) = (c_{ij}(\lambda)) = A^* B^*(\lambda) A^T$ ($0 \leq i, j \leq m$).
Then 
\[ b_{0}(\lambda) = \frac{\partial^{2} P}{\partial \lambda_{0}^{2}} \]

\[ = \frac{2}{\lambda_{0}} \sum_{j=1}^{n} \left\{ -\frac{1}{\lambda_{0}} f_{j}'(\Lambda_{j}) + \left[ \frac{1}{\lambda_{0}} f_{j}'(\Lambda_{j}) + g_{j}'(\Lambda_{j}) \right] a_{ij} \right\} \]

\[ = \sum_{j=1}^{n} \left\{ \frac{a_{ij}}{\lambda_{0}} f_{j}'(\Lambda_{j}) - \frac{2}{\lambda_{0}^{2}} f_{j}'(\Lambda_{j}) a_{ij} + \left[ \frac{1}{\lambda_{0}} f_{j}''(\Lambda_{j}) + g_{j}''(\Lambda_{j}) \right] a_{ij} \right\} \]

\[ = b_{0}^{*}(\lambda) + \sum_{j=1}^{n} a_{ij} b_{0}^{*}(\lambda) + \sum_{j=1}^{n} \left\{ b_{ij}^{*}(\lambda) + \sum_{k=1}^{n} a_{ik} b_{kj}^{*}(\lambda) \right\} a_{ij} \]

\[ = c_{00}(\lambda) \]

Also, for \( i = 1, 2, \ldots, m \)

\[ b_{i0}(\lambda) = \frac{\partial^{2} P}{\partial \lambda_{i} \partial \lambda_{0}} \]

\[ = \frac{2}{\lambda_{0}} \sum_{j=1}^{n} \left\{ -\frac{1}{\lambda_{0}} f_{j}'(\Lambda_{j}) + \left[ \frac{1}{\lambda_{0}} f_{j}'(\Lambda_{j}) + g_{j}'(\Lambda_{j}) \right] a_{ij} \right\} \]

\[ = \sum_{j=1}^{n} \left\{ -\frac{a_{ij}}{\lambda_{0}} f_{j}'(\Lambda_{j}) + a_{ij} \left[ \frac{1}{\lambda_{0}} f_{j}''(\Lambda_{j}) + g_{j}''(\Lambda_{j}) \right] a_{ij} \right\} \]

\[ = \sum_{j=1}^{n} a_{ij} b_{i0}^{*}(\lambda) + \sum_{j=1}^{n} \left\{ \sum_{k=1}^{n} a_{ik} b_{kj}^{*}(\lambda) \right\} a_{ij} \]

\[ = c_{i0}(\lambda) \]

and, by symmetry, \( c_{0i}(\lambda) = b_{0i}(\lambda) \)

Lastly, for \( 1 \leq i, k \leq m \),

\[ b_{ik}(\lambda) = \frac{\partial^{2} P}{\partial \lambda_{i} \partial \lambda_{k}} \]

\[ = \frac{2}{\lambda_{0}} \sum_{j=1}^{n} \left\{ \frac{1}{\lambda_{0}} f_{j}'(\Lambda_{j}) + g_{j}'(\Lambda_{j}) \right\} a_{kj} \]

\[ = \sum_{j=1}^{n} a_{ij} \left[ \frac{1}{\lambda_{0}} f_{j}''(\Lambda_{j}) + g_{j}''(\Lambda_{j}) \right] a_{kj} \]

\[ = \sum_{j=1}^{n} a_{ij} \sum_{k=1}^{n} b_{jk}^{*}(\lambda) a_{kj} \]

\[ = c_{ik}(\lambda) \]

i.e. \( B(\lambda) = C(\lambda) \) as required.
LEMMA 7. If $\lambda$ satisfies the capacity constraints and $\lambda_0 > 0$, the matrix $\mathbf{B}^*(\lambda)$ is positive definite.

PROOF: It is sufficient (Mirsky 1955, p. 400) to show that, with some permutation of the rows and columns of $\mathbf{B}^*(\lambda)$, all the leading minors are positive. It will be shown that this is true when the order of rows and columns is reversed. The first $n$ leading minors of the resulting matrix are

$$
\prod_{j=k}^{n} b_{ij}^*(\lambda) \quad (k=n, n-1, \ldots, 1)
$$

$$
= \prod_{j=k}^{n} \left\{ f_j''(\lambda_j) + q_j''(\lambda_j) \right\}
$$

$$
= \prod_{j=k}^{n} \left\{ \frac{b_j^*}{\lambda_0(\lambda_0-y_j)} + \frac{d}{d\lambda_j} \left( \frac{y_j}{2(\lambda_j-y_j)} - \frac{y_j}{\lambda_j} \right) \right\}
$$

$$
= \prod_{j=k}^{n} \left\{ \frac{b_j^*}{\lambda_0(\lambda_0-y_j)} + y_j \left( \frac{1}{(\lambda_j-y_j)^2} - \frac{1}{\lambda_j} \right) \right\}
$$

$$
> 0
$$

if $\lambda_j > y_j > 0$ (i.e. $\lambda$ satisfies the capacity constraints) and $\lambda_0 > 0$.

It remains only to show that $\det \{ \mathbf{B}^*(\lambda) \} > 0$.

Abbreviating $b_{ij}^*(\lambda)$ to $b_{ij}^*$, let $b_{ij}^*$ be the cofactor of $b_{ij}^*$ in $\det \{ \mathbf{B}^*(\lambda) \}$. Then, expanding by the first row,
\[ \det \{ B^*(\lambda) \} = \sum_{j=0}^{n} b_{0j}^* B_{0j}^* \]

\[
\begin{bmatrix}
 b_{00}^* & b_{01}^* & \cdots & 0 & 0 & \cdots & 0 \\
 0 & b_{11}^* & \cdots & 0 & 0 & \cdots & 0 \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 b_{n-1,0}^* & 0 & \cdots & b_{n-1,n}^* & 0 & \cdots & 0 \\
 b_{n,0}^* & 0 & \cdots & 0 & 0 & \cdots & 0 \\
 0 & \cdots & 0 & b_{n+1,0}^* & b_{n+1,1}^* & \cdots & 0 \\
 \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 0 & \cdots & 0 & b_{m,0}^* & 0 & \cdots & 0 \\
 0 & \cdots & 0 & \cdots & \cdots & \cdots & b_{mm}^*
\end{bmatrix}
\]

\[
= b_{00}^* \prod_{j=1}^{n} b_{jj}^* + \sum_{j=0}^{n} b_{0j}^* (-1)^{j+1} b_{jj}^*
\]

\[
\begin{bmatrix}
 b_{11}^* & \cdots & 0 & b_{00}^* & 0 & \cdots & 0 \\
 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 0 & \cdots & 0 & b_{n-1,0}^* & b_{n-1,n}^* & 0 & \cdots & 0 \\
 0 & \cdots & 0 & b_{n,0}^* & 0 & \cdots & 0 \\
 0 & \cdots & 0 & b_{n+1,0}^* & b_{n+1,1}^* & \cdots & 0 \\
 \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 0 & \cdots & 0 & b_{m,0}^* & 0 & \cdots & 0 \\
 0 & \cdots & 0 & \cdots & \cdots & \cdots & b_{mm}^*
\end{bmatrix}
\]

\[
= \left\{ b_{00}^* - \sum_{j=1}^{n} \frac{(b_{0j}^*)^2}{b_{jj}^*} \right\} \prod_{j=1}^{n} b_{jj}^*
\]

It has already been shown that \( \prod_{j=1}^{n} b_{jj}^* > 0 \) if \( \Lambda_j > y_j \)
and \( \lambda_0 > 0 \). Now
\[ b^{*}_{o} - \sum_{j=1}^{n} \left( \frac{b^*_{jj}}{b_{jj}} \right)^2 = \frac{1}{\lambda_0} \sum_{j=1}^{n} f_j(\lambda_j) - \frac{1}{\lambda_0} \sum_{j=1}^{n} \left\{ \frac{f_j(\lambda_j)}{b_{jj}} \right\}^2 \]

\[ = \frac{1}{\lambda_0} \sum_{j=1}^{n} \frac{1}{b_{jj}} \left\{ 2\lambda_0 g_j''(\lambda_j)f_j(\lambda_j) - 2f_j(\lambda_j)f_j''(\lambda_j) - [f_j'(\lambda_j)]^2 \right\} \]

\[ = \frac{2}{\lambda_0} \sum_{j=1}^{n} \frac{g_j''(\lambda_j)}{b_{jj}} f_j(\lambda_j) \]

\[ > 0 \]

since, as shown in the first part of the proof, \( g_j''(\lambda_j) > 0 \) and \( b^*_{jj} > 0 \) (\( j = 1, 2, \ldots, n \)) if \( \lambda \) satisfies the capacity constraints and \( \lambda_0 > 0 \), and since \( f_j''(\lambda_j) > 0 \).

Hence \( \det \{ B^*(\lambda) \} > 0 \) and \( B^*(\lambda) \) is positive definite.

**Lemma 8.** A feasible solution \( \lambda_0 \) is optimal if and only if, for all permissible directions \( \mu_0 \) at \( \lambda_0 \),

\[ \mu_0 \cdot \nabla D(\lambda_0) \geq 0 \]

**Proof:** Since \( \lambda_0 \) is feasible and \( \mu_0 \) is permissible, \( \lambda_0 + h \mu_0 \) is feasible for all sufficiently small positive \( h \). Hence, by Taylor's Theorem,

\[ D(\lambda_0 + h \mu_0) = D(\lambda_0) + h \mu_0 \cdot \nabla D(\lambda_0) + o(h) \]

Suppose that \( \mu_0 \cdot \nabla D(\lambda_0) < 0 \); then, for all sufficiently small positive \( h \),

\[ D(\lambda_0 + h \mu_0) < D(\lambda_0) \]

and \( \lambda_0 \) is not optimal.

Hence, if \( \lambda_0 \) is optimal, \( \mu_0 \cdot \nabla D(\lambda_0) > 0 \).

Conversely, if \( \lambda_0 \) is not optimal, let \( \lambda_0 + \mu_0 \) be
optimal. Then \( D(\lambda_0 + \mu_0) - D(\lambda_0) < 0 \) and, since \( \lambda_0 \) and \( \lambda_0 + \mu_0 \) are both feasible, \( \mu_0 \) is a permissible direction at \( \lambda_0 \). Hence, by Taylor's Theorem and Lemmas 6 and 7

\[
\mu_0 \cdot \nabla D(\lambda_0) = D(\lambda_0 + \mu_0) - D(\lambda_0) - \frac{1}{2} (\mu_0 \Delta^*) \Delta^* (\lambda_0 + h \mu_0)(\mu_0 \Delta^*)^T \quad \text{where } 0 < h < 1
\]

\[
< D(\lambda_0 + \mu_0) - D(\lambda_0) < 0.
\]

Hence, if, for all permissible \( \mu_0 \) at \( \lambda_0 \), \( \mu_0 \cdot \nabla D(\lambda_0) \geq 0 \), then \( \lambda_0 \) is optimal.

THEOREM 5. If \( \lambda_0 \) and \( \lambda_0 + \mu \) are two optimal solutions, then \( \mu_0 = 0 \) and \((\mu_1, \mu_2, \ldots, \mu_m)\) belongs to the nullspace in \( \mathbb{R}^m \) of \( A \).

PROOF: By Lemma 5, Taylor's Theorem can be applied to expand \( D(\lambda_0 + \mu) \) about \( \lambda_0 \), giving

\[
D(\lambda_0 + \mu) = D(\lambda_0) + \mu \cdot \nabla D(\lambda_0) + \frac{1}{2} \mu \Delta^* (\lambda_0 + h \mu)(\mu \Delta^*)^T,
\]

where \( 0 < h < 1 \).

But \( \lambda_0 \) and \( \lambda_0 + \mu \) are both optimal, so that

\[
D(\lambda_0 + \mu) = D(\lambda_0),
\]

and, using Lemma 6,

\[
\mu \cdot \nabla D(\lambda_0) + \frac{1}{2} (\mu \Delta^*) \Delta^* (\lambda_0 + h \mu)(\mu \Delta^*)^T = 0.
\]

Moreover, by Lemmas 7 and 8, neither term in this expression can be negative. Hence each is zero, and, again by Lemma 7,
\[ \mu^* A^* = 0, \]
\[ \mu_0 = 0, \]

and
\[ \sum_{i=0}^{\infty} \mu_i a_{ij} = 0 \quad (j = 1, 2, \ldots, n) \]

Hence
\[ \sum_{i=1}^{\infty} \mu_i a_{ij} = 0 \quad (j = 1, 2, \ldots, n) \]

i.e.
\[ (\mu_1, \mu_2, \ldots, \mu_m) D = 0 \]
as required.

**COROLLARY 1.** If \( A \) has rank \( m \), there is only one optimal solution.

**PROOF:** If \( A \) has rank \( m \), its nullspace in \( \mathbb{R}^m \) consists of the zero vector only; hence, by the theorem, any two optimal solutions are equal.

**COROLLARY 2.** If \( \lambda_0 \) is an optimal solution such that \( \mu \cdot \nabla D(\lambda_0) > 0 \) for all permissible \( \mu \) at \( \lambda_0 \), then there is no other optimal solution.

**PROOF:** If there were another optimal solution, it could be written as \( \lambda_0 + \mu \), where \( \mu \) is permissible at \( \lambda_0 \). Then, as shown in the proof of the theorem, \( \mu \cdot \nabla D(\lambda_0) = 0 \), which contradicts the given property of \( \lambda_0 \).

**COROLLARY 3.** If there is more than one optimal solution, all such solutions correspond to the same cycle time and the same effective green times on all approaches.

**PROOF:** Let any two optimal solutions be \( \lambda_0 \) and \( \lambda_1 \); then \( \lambda_1 = \lambda_0 + \mu \), where \( \mu \) has the properties proved in the theorem. Hence
\[ \lambda_1 A^* = \lambda_0 A^*, \]
but the components of these two vectors are the values of \( \mathbf{L} / e \) and the \( \Lambda_j \) corresponding to the two solutions, and the corollary follows.

3.5. An algorithm for finding delay-minimising settings

In this section it will be shown how, starting from any feasible solution \( \Lambda_1 \) (found, for example, by the methods of section 3.3) a sequence \( \{ \Lambda_p ; \ p = 1, 2, 3, \ldots \} \) of feasible solutions can be constructed, which converges to an optimal solution. It would in principle be possible to choose \( \Lambda_{p+1} \) so that the direction of \( \Lambda_{p+1} - \Lambda_p \) is that permissible direction at \( \Lambda_p \) in which the gradient of \( D(\Lambda) \) is least (i.e. to use a method of steepest descent). To simplify calculation, however, \( \Lambda_{p+1} \) will (except in special circumstances considered in Lemma 9) be defined in terms of \( \Lambda_p \) by an equation

\[
\Lambda_{p+1} = \Lambda_p + h_p (e_k - e_\ell)
\]

where \( 0 \leq k, \ell \leq m \) if the cycle time is unspecified, \( 1 \leq k, \ell \leq m \) if the cycle time is specified, \( k \neq \ell \), and \( h_p > 0 \). The subscripts \( k \) and \( \ell \), which determine the direction of \( \Lambda_{p+1} - \Lambda_p \), and \( h_p \), which determines the distance of \( \Lambda_{p+1} \) from \( \Lambda_p \), will be determined by \( \Lambda_p \), \( D(\Lambda) \) and the minimum green and cycle time constraints. The form \( h_p (e_k - e_\ell) \) is chosen because it is the simplest form of \( \Lambda_{p+1} - \Lambda_p \) that keeps the constraint \( S(\Lambda) = 1 \) satisfied throughout the sequence \( \{ \Lambda_p \} \).

The choice of \( k, \ell \) and \( h_p \) will be discussed, and then the convergence of \( \{ \Lambda_p \} \) to an optimal solution will be
established. This optimal solution corresponds to delay-minimising signal settings.

The components of $\lambda_p$ will be denoted by $\lambda_{pi}$ ($i = 0, 1, 2, \ldots, m$); i.e.

$$\lambda_p = (\lambda_{po}, \lambda_{p1}, \lambda_{p2}, \ldots, \lambda_{pm})$$

Also, $|_p$ will denote evaluation at $\lambda_p$ of a function of $\lambda$, e.g.

$$\frac{\partial D}{\partial \lambda_p}$$

denotes the value of $\frac{\partial D}{\partial \lambda_p}$ when $\lambda = \lambda_p$.

3.51. Choice of direction. So that $\lambda_{p+1}$ shall be feasible, $k$ and $\ell$ must be chosen so that $e_k - e_\ell$ is permissible at $\lambda_p$. This places no restriction upon $k$ and $\ell$ unless $\lambda_{po} = k_0$, or $\lambda_{pi} = k_i \lambda_{po}$ for one or more $i$ ($i = 1, 2, \ldots, m$), or both. If $\lambda_{po} = k_0$, then $\ell \neq 0$, and if, for any $i$, $\lambda_{pi} = k_i \lambda_{po}$, then $k \neq 0$ and $\ell + i$ for any such $i$. If the cycle time is specified, then $\lambda_{po} = k_0$ for all $p$, and neither $k$ nor $\ell$ can be zero. All these restrictions are embodied in the following rules for the choice of $k$ and $\ell$. Let $k$ be chosen so that

$$\frac{\partial D}{\partial \lambda_k} = \begin{cases} 
\min \left( \frac{\partial D}{\partial \lambda_{k_i}} \right) & \text{if } \lambda_{pi} > k_i \lambda_{po} \ (i = 1, 2, \ldots, m) \text{ and the cycle time is unspecified,} \\
\min \left( \frac{\partial D}{\partial \lambda_{k_i}} \right) & \text{if } \lambda_{pi} = k_i \lambda_{po} \text{ for some } i \ (i = 1, 2, \ldots, m) \text{ or the cycle time is specified.}
\end{cases}$$

Since $m \geq 1$, this rule always yields a value of $k$. There may be more than one value satisfying the rule, and, if so, one of them is chosen arbitrarily. Let $\ell$ be chosen so that
\[ \frac{\partial D}{\partial \lambda_f} |_{p} = \begin{cases} \max \left\{ \frac{\partial D}{\partial \lambda_0} |_{p}, \max_{i \in \mathcal{E}_m} \left( \frac{\partial D}{\partial \lambda_i} |_{p} \right) \right\} \text{ if } \lambda_{po} \neq k_0 \\
\max_{i \in \mathcal{E}_m} \left( \frac{\partial D}{\partial \lambda_i} |_{p} \right) \text{ if } \lambda_{po} = k_0. \end{cases} \]

This rule can fail to yield a value of \( \ell \) only if \( \lambda_{po} = k_0 \) and \( \lambda_{pi} = k_i \lambda_{po} \) (\( i = 1, 2, \ldots, m \)). This is the trivial case in which the maximum or specified cycle time is just the sum of the minimum green times and the lost time, so that \( \lambda_p \) is the only feasible solution, and is therefore optimal, and the problem is solved. In other cases, there may be more than one value of \( \ell \) satisfying the rule, and, if so, one of them is chosen arbitrarily.

It will be assumed in what follows that the rules yield values of both \( k \) and \( \ell \). The following lemma shows that if \( k \) and \( \ell \) are so chosen and \( \lambda_p \) is not itself an optimal solution, then either \( h_p \) can be chosen to make \( D(\lambda_{p+1}) < D(\lambda_p) \), where \( \lambda_{p+1} = \lambda_p + h_p(\varepsilon_k - \varepsilon_{\ell}) \) or this inequality can be made to hold by making a suitable alternative definition of \( \lambda_{p+1} \).

LEMMA 9. If \( k \) and \( \ell \) are chosen by the above rules and \( \lambda_p \) is not optimal, then either \( D(\lambda_{p+1}) < D(\lambda_p) \) for all sufficiently small positive \( h_p \) or \( \lambda_{pi} = k_i \lambda_{po} \) for at least one \( i \) (\( i = 1, 2, \ldots, m \)) and there is a permissible \( \mu \) at \( \lambda_p \), such that \( \mu \cdot \nabla D(\lambda_p) < 0 \), of the form

\[ \mu = \pm \left\{ \varepsilon_0 + \sum_{i \in \mathcal{E}_c} k_i \varepsilon_i - (1 + \sum_{i \in \mathcal{E}_c} k_i) \varepsilon_{\ell} \right\}, \]
where $\lambda_{pi} > k_i \lambda_{po}$ and $C$ is the set of $i$ ($i = 1, 2, \ldots, m$) such that $\lambda_{pi} = k_i \lambda_{po}$.

PROOF: If $\lambda_p$ is not optimal then, by Lemma 8 there exists a permissible $\mu$ at $\lambda_p$ such that $\mu \cdot \nabla D(\lambda_p) < 0$.

Suppose first that $\lambda_{pi} > k_i \lambda_{po}$ for $i = 1, 2, \ldots, m$. Then it follows from the rules of choice of $k$ and $\ell$ that

$$\frac{\partial D}{\partial \lambda_i}(\lambda_p) \geq \frac{\partial D}{\partial \lambda_k}(\lambda_p) \quad \text{for all } i \ (i = 0, 1, 2, \ldots, m) \text{ such that } \mu_i > 0 \quad \text{and} \quad \frac{\partial D}{\partial \lambda_i} \leq \frac{\partial D}{\partial \lambda_k} \quad \text{for all } i \text{ such that } \mu_i < 0.$$ 

Hence

$$\mu \cdot \nabla D(\lambda_p) = \sum_{\mu_i > 0} \mu_i \frac{\partial D}{\partial \lambda_k} + \sum_{\mu_i < 0} \mu_i \frac{\partial D}{\partial \lambda_i} \geq \frac{\partial D}{\partial \lambda_k} \sum_{\mu_i > 0} \mu_i + \frac{\partial D}{\partial \lambda_i} \sum_{\mu_i < 0} \mu_i.$$ 

But $\mu$ is permissible, so that $\sum_{\mu_i > 0} \mu_i = -\sum_{\mu_i < 0} \mu_i$; moreover $\mu \cdot \nabla D(\lambda_p) < 0$.

Hence

$$\left(\frac{\partial D}{\partial \lambda_k} - \frac{\partial D}{\partial \lambda_i}\right) < 0.$$

Also, by the rules of choice $\alpha_k - \alpha_i$ is permissible at $\lambda_p$, and therefore by Lemma 5

$$D(\lambda_{pi}) - D(\lambda) = h_p \left(\frac{\partial D}{\partial \lambda_k} - \frac{\partial D}{\partial \lambda_i}\right) + o(h_p).$$

Hence $D(\lambda_{p+1}) < D(\lambda_p)$ for all sufficiently small positive $h_p$.

Now suppose that $D(\lambda_{p+1}) \geq D(\lambda_p)$ for arbitrarily small positive $h_p$; then it follows from Lemma 5 that

$$\left(\frac{\partial D}{\partial \lambda_k} - \frac{\partial D}{\partial \lambda_i}\right) \geq 0.$$ 

Hence, by the first part of this proof, $\lambda_{pi} = k_i \lambda_{po}$ for at least one $i$ ($i = 1, 2, \ldots, m$). Let the sets $C$ and $F$
be defined by

\[ C = \{ i ; 1 \leq i \leq m \text{ and } \lambda_{pi} = k_i \lambda_{po} \} \]

\[ F = \{ i ; 1 \leq i \leq m \text{ and } \lambda_{pi} > k_i \lambda_{po} \} \]

Then

\[ \frac{\partial D}{\partial \lambda_k} \bigg|_p = \min_{1 \leq i \leq m} \left( \frac{\partial D}{\partial \lambda_i} \bigg|_p \right) \quad (\text{since } \lambda_{pi} = k_i \lambda_{po} \text{ for some } i) \]

\[ \leq \min_{i \in F} \left( \frac{\partial D}{\partial \lambda_i} \bigg|_p \right) \]

\[ \leq \max_{i \in F} \left( \frac{\partial D}{\partial \lambda_i} \bigg|_p \right) \]

\[ \leq \max \left\{ \frac{\partial D}{\partial \lambda_0} \bigg|_p, \max_{i \in F} \left( \frac{\partial D}{\partial \lambda_i} \bigg|_p \right) \right\} \]

\[ = \frac{\partial D}{\partial \lambda_i} \bigg|_p. \]

But \( \frac{\partial D}{\partial \lambda_k} \bigg|_p \geq \frac{\partial D}{\partial \lambda_i} \bigg|_p \), so that all the expressions in the preceding chain of inequalities are equal; let their common value be \( G \).

Then it follows that \( \frac{\partial D}{\partial \lambda_0} \bigg|_p \leq G \), \( \frac{\partial D}{\partial \lambda_i} \bigg|_p \geq G \) for all \( i \) in \( C \), and \( \frac{\partial D}{\partial \lambda_i} \bigg|_p = G \) for all \( i \) in \( F \).

Since \( \lambda_p \) is not optimal, there is a \( \mu \), permissible at \( \lambda_p \), such that \( \mu \cdot \nabla D(\lambda_p) < 0 \); because \( \mu \) is permissible, \( \mu_i \geq k_i \mu_0 \) for all \( i \) in \( C \). It will first be shown that \( \mu_0 \neq 0 \): suppose \( \mu_0 \) were zero. Then \( \mu_i > 0 \) for all \( i \) in \( C \), and, by the above properties of the components of \( \nabla D(\lambda_p) \),

\[ \mu \cdot \nabla D(\lambda_p) = \sum_{i \in C} \mu_i \frac{\partial D}{\partial \lambda_i} \bigg|_p + \sum_{i \in F} \mu_i \frac{\partial D}{\partial \lambda_i} \bigg|_p \]

\[ > G \sum_{i \in C} \mu_i + G \sum_{i \in F} \mu_i \]

\[ = G \sum_{i=1}^m \mu_i \]

\[ = 0, \]
which contradicts the fact that \( \mathbf{\mu} \cdot \nabla D(\lambda_p) < 0 \); hence \( \mathbf{\mu}_0 \neq 0 \). This shows, in particular, that if the cycle time is specified and \( \lambda_p \) is not optimal, then \( D(\lambda_{p+1}) < D(\lambda_p) \) for all sufficiently small \( h_p \).

Now, for \( i \) in \( C \), let \( \mathbf{\mu}^*_i = k_i \mathbf{\mu}_0 \), so that \( \mathbf{\mu}^*_i < \mathbf{\mu}_i \) for all such \( i \). Let \( \mathbf{\mu}^*_0 = \mathbf{\mu}_0 \), let \( f \) be the least \( i \) in \( F \), let \( \mathbf{\mu}^*_f = -\mathbf{\mu}^*_0 - \sum_{i \in C} \mathbf{\mu}^*_i \), and let \( \mathbf{\mu}^*_i = 0 \) for all other \( i \) in \( F \). Then \( \mathbf{\mu}^* = (\mathbf{\mu}^*_0, \mathbf{\mu}^*_1, \ldots, \mathbf{\mu}^*_m) \) is permissible at \( \lambda_p \) and, using the fact that \( \sum_{i=0}^{m} \mathbf{\mu}_i = 0 \),

\[
\mathbf{\mu} \cdot \nabla D(\lambda_p) - \mathbf{\mu}^* \cdot \nabla D(\lambda_p) = G \left( \mathbf{\mu}_0 - \sum_{i \in C} \mathbf{\mu}_i + \mathbf{\mu}^*_0 + \sum_{i \in C} \mathbf{\mu}^*_i \right) + \sum_{i \in C} (\mathbf{\mu}_i - \mathbf{\mu}^*_i) \left. \frac{\partial P}{\partial \lambda_i} \right|_{\lambda_p} \geq G \left\{ \sum_{i \in C} (\mathbf{\mu}^*_i - \mathbf{\mu}_i) + \sum_{i \in C} (\mathbf{\mu}_i - \mathbf{\mu}^*_i) \right\} = 0.
\]

But \( \mathbf{\mu} \cdot \nabla D(\lambda_p) < 0 \); hence \( \mathbf{\mu}^* \cdot \nabla D(\lambda_p) < 0 \), and, on putting \( \mathbf{\mu}_0 = \pm 1 \), \( \mathbf{\mu}^* \) takes the required form

\[
\pm \left\{ \mathbf{e}_0 + \sum_{i \in C} k_i \mathbf{e}_i - (1 + \sum_{i \in C} k_i) \mathbf{e}_f \right\}
\]

To take account of the second alternative in Lemma 9, the definition of \( \lambda_{p+1} \) will be extended as follows: If \( \left| (\mathbf{e}_0 - \mathbf{e}_f) \cdot \nabla D(\lambda_p) (1 + \sum_{i \in C} k_i) \right| < \left| \left\{ \mathbf{e}_0 + \sum_{i \in C} k_i \mathbf{e}_i - (1 + \sum_{i \in C} k_i) \mathbf{e}_f \right\} \cdot \nabla D(\lambda_p) \right| \)
then

\( \lambda_{p+1} = \lambda_p \pm h_p \left\{ \mathbf{e}_0 + \sum_{i \in C} k_i \mathbf{e}_i - (1 + \sum_{i \in C} k_i) \mathbf{e}_f \right\} \),

where \( h_p \) is to be determined, \( C \) is the set of \( i \) (\( i = 1, 2, \ldots, m \)) such that \( \lambda_{pi} = k_i \lambda_{po} \), \( f \) is the least \( i \) such that \( \lambda_{pi} > k_i \lambda_{po} \), and the sign is chosen to
make \((\lambda_{p+1} - \lambda_p) \cdot \nabla D(\lambda_p)\) negative.

It will be assumed in what follows that \(\lambda_p\) is not optimal.

3.52. Choice of distance. The choice of \(h_p\) is influenced by two considerations: given the choice of \(k\) and \(\ell\), \(D(\lambda_{p+1})\) should be as much less than \(D(\lambda_p)\) as possible, but \(h_p\) must not be so large that \(\lambda_{p+1}\) is outside the set of feasible solutions.

It will first be shown that if the direction of \(\lambda_{p+1} - \lambda_p\) is \(e_k - e\ell\) the capacity constraints and the weakest possible form of cycle time constraint (i.e. \(\lambda_0 > 0\)) together define (except in one special case) an upper bound, \(H_p\), say, for \(h_p\). It will then be shown that there is just one value of \(h\) in the interval \(0 < h < H_p\) that minimises \(D(\lambda_p + h(e_k - e\ell))\), and \(h_p\) will be given this value provided that the resulting \(\lambda_{p+1}\) is feasible (i.e. provided that it satisfies the minimum green and cycle time constraints; it must, by definition of \(H_p\), satisfy the capacity constraints). If not, and in the special case where \(H_p\) is undefined, \(h_p\) will be given the largest value that makes \(\lambda_{p+1}\) feasible (i.e. the largest value such that \(\lambda_{p+1}\) satisfies the minimum green and cycle time constraints).

Appropriate extensions of the results will be made to cover the special case where the alternative definition of \(\lambda_{p+1}\) is used.

Consider first the implications of the constraints
where $\lambda_0 > 0$ and $\Lambda_j > y_j$ for $j = 1, 2, \ldots, n$. $\Lambda_p$ satisfies these constraints and $\Lambda_{p+1}$ is to be obtained from $\Lambda_p$ by increasing $\lambda_{pk}$ by $h_p$ and decreasing $\lambda_{pl}$ by $h_p$. Since $a_{ij} \geq 0$ for $i = 0, 1, 2, \ldots, m; j = 1, 2, \ldots, n$, none of the above constraints can be violated by the increase in $\lambda_{pk}$, but, except in one special case, at least some of them will be violated by the decrease in $\lambda_{pl}$ if $h_p$ is too large. To find the upper bound $H_p$ that is thus imposed upon $h_p$, it is necessary to consider three cases:

(a) $k = 0$

(b) $\ell = 0$

(c) $k \neq 0$ and $\ell \neq 0$

and the case, (d), where the alternative definition of $\Lambda_{p+1}$ applies.

In case (a), $\Lambda_{p+1,0}$ is certainly positive, and

$$\Lambda_j|_{p,0} = \Lambda_j|_p + (a_{oj} - a_{oj})h_p, \quad (j = 1, 2, \ldots, n).$$

To make $\Lambda_j|_{p,0} > y_j$ it is therefore necessary that

$$(a_{oj} - a_{oj})h_p < \Lambda_j|_p - y_j, \quad (j = 1, 2, \ldots, n).$$

Since $0 < a_{oj} < 1$, this will certainly be true for those $j$ for which $a_{oj} = 0$.

It is therefore sufficient that

$$h_p < \frac{\Lambda_j|_p - y_j}{a_{oj} - a_{oj}}$$

for those $j$ for which $a_{oj} = 1$.

Hence, if $k = 0$,

$$H_p = \min_{1 \leq j \leq n, a_{oj} = 1} \left\{ \frac{\Lambda_j|_p - y_j}{a_{oj} - a_{oj}} \right\}.$$
In case (b), for $\lambda_{p+1}$ to be positive it is necessary that $h_p < \lambda_{po}$; also
\[ \Lambda_j|_{p+1} = \Lambda_j|_p + (a_{kj} - a_{kj})h_p \quad (j = 1, 2, \ldots, n) \]

To make $\Lambda_j|_{p+1} > y_j$ it is therefore necessary that
\[ (a_{o} - a_{kj})h_p < \Lambda_j|_p - y_j \quad (j = 1, 2, \ldots, n) \]

This will certainly be true for those $j$ for which $a_{kj} > 0$ or $a_{oj} = 0$. It is therefore sufficient that
\[ h_p < (\Lambda_j|_p - y_j)/a_{kj} \]

for those $j$ for which $a_{oj} > 0$ and $a_{kj} = 0$.

Hence, if $l = 0$,
\[ H_p = \min \left\{ \lambda_{po}, \min_{i \in \mathbb{S}, j \in \mathbb{N}, a_{o} > 0} \left( \frac{\Lambda_j|_p - y_j}{a_{oj}} \right) \right\} \]

In case (c), $\lambda_{p+1} = \lambda_{po} > 0$, and
\[ \Lambda_j|_{p+1} = \Lambda_j|_p + (a_{kj} - a_{kj})h_p \quad (j = 1, 2, \ldots, n) \]

To make $\Lambda_j|_{p+1} > y_j$ it is therefore necessary that
\[ (a_{o} - a_{kj})h_p < \Lambda_j|_p - y_j \]

This will certainly be true for all $j$ except those for which $a_{kj} = 1$ and $a_{kj} = 0$, and for these $j$ it is necessary that $h_p < \Lambda_j|_p - y_j$.

Hence, if $k \neq 0$ and $l \neq 0$,
\[ H_p = \min_{\begin{array}{c} i \in \mathbb{S}, j \in \mathbb{N}, a_{o} > 0, a_{kj} = 1, a_{o} = 0 \end{array}} \left\{ \Lambda_j|_p - y_j \right\} \]

If, in case (c), there is no $j$ for which $a_{kj} = 1$
and \( a_{kj} = 0 \), then \( H_p \) is undefined. In such a case, every approach having right of way in stage \( \ell \) also has right of way in stage \( k \). There must therefore be at least one approach that has right of way in stage \( k \) and not in stage \( \ell \); any such approach will benefit, and none suffer, if green time is transferred from stage \( \ell \) to stage \( k \). The green time for stage \( \ell \) should therefore be made as small as possible in such a case, and it will be seen later in this section that the algorithm achieves this.

In case (d), \( \lambda_{p+1} = \lambda_p \pm h_r \{ \varepsilon_0 + \sum_{i \in C} K_i \varepsilon_i - (1 + \sum_{i \in C} K_i) \varepsilon_f \} \)

Suppose first that the positive sign applies; then \( \lambda_{p+1} \) is certainly positive, and

\[
\lambda_j |_{p+1} = \lambda_j |_p + \{ a_{ij} + \sum_{i \in C} K_i a_{ij} - (1 + \sum_{i \in C} K_i) a_{ij} \} h_r
\]

and to make \( \lambda_j |_{p+1} > \gamma_j \) it is therefore necessary that

\[
\{ (1 + \sum_{i \in C} K_i) a_{ij} - a_{ij} - \sum_{i \in C} K_i a_{ij} \} h_r < \lambda_j |_p - \gamma_j
\]

This will certainly be true for those \( j \) for which \( a_{fj} = 0 \). Hence, as in case (a),

\[
H_p = \min_{i \in j} \left\{ \frac{\lambda_j |_p - \gamma_j}{(1 + \sum_{i \in C} K_i) a_{ij} - a_{ij} - \sum_{i \in C} K_i a_{ij}} \right\}
\]

Now suppose that the negative sign applies; then \( \lambda_{p+1} \) will be positive only if \( h_r < \lambda_{po} \). Also

\[
\lambda_j |_{p+1} = \lambda_j |_p - \{ a_{ij} + \sum_{i \in C} K_i a_{ij} - (1 + \sum_{i \in C} K_i) a_{ij} \} h_r
\]

and to make \( \lambda_j |_{p+1} > \gamma_j \) it is therefore necessary that
\[
\left\{ a_{ij} + \sum_{i \in C} k_i a_{ij} - (1 + \sum_{i \in C} k_i) a_{ij} \right\} h_p < \lambda_{ij} p - y_j
\]

Since \( a_{ij} < 1 \) and \( a_{ij} \leq 1 \), this will certainly be true if \( a_{ij} = 1 \); it will also be true if \( a_{ij} + \sum_{i \in C} k_i a_{ij} = 0 \).

Hence, as in case (b)

\[
H_p = \min \left\{ \lambda_p, \min_{a_{ij}, k_i > 0} \left( \frac{\lambda_{ij} - y_j}{a_{ij} + \sum_{i \in C} k_i a_{ij}} \right) \right\}
\]

If \( \lambda_{p+1} = \lambda_p + h_p (e_k - e_{\ell}) \), then, once \( k \) and \( \ell \) have been chosen, \( D(\lambda_{p+1}) \) depends only on the choice of \( h_p \).

Let \( d(h) = D(\lambda_p + h(e_k - e_{\ell})) \).

The following lemma establishes certain properties of this function of \( h \).

**Lemma 10:** If \( H_p \) is defined, then \( d(h) \) has just one turning value in the interval \( 0 < h < H_p \), and this is a minimum. If \( H_p \) is undefined, then \( d(h) \) is decreasing throughout the interval \( 0 < h < \lambda_{p+\ell} \).

**Proof:**

\[
\begin{align*}
d'(h) &= \frac{\partial D}{\partial \lambda_k} - \frac{\partial D}{\partial \lambda_\ell} \\
d''(h) &= \frac{\partial^2 D}{\partial \lambda_k^2} - 2 \frac{\partial^2 D}{\partial \lambda_k \partial \lambda_\ell} + \frac{\partial^2 D}{\partial \lambda_\ell^2}
\end{align*}
\]

Evaluated at \( \lambda = \lambda_p + h(e_k - e_{\ell}) \).

Suppose first that \( H_p \) is defined.

Now,

\[
d''(h) = (e_k - e_{\ell}) A^* B^* \left( \lambda_p + h(e_k - e_{\ell}) \right) (e_k - e_{\ell})^T
\]

\[
= (e_k - e_{\ell}) A^* B^* \left( \lambda_p + h(e_k - e_{\ell}) \right) [A^* (e_k - e_{\ell})] \quad > 0
\]

for \( 0 < h < H_p \) by Lemma 7, since \((e_k - e_{\ell}) A^* \neq 0 \) (for if it were, just the same approaches would have right of way in stage \( k \) as in stage \( \ell \)).
Hence there can be at most one turning value of \( \Phi(h) \) in \( 0 < h < H_p \), and if there is one it is a minimum.

Also \( d''(h) = \left( \frac{2D}{\delta \lambda}\lambda_x - \frac{2D}{\delta \lambda_t}\right)_{\Phi} < 0 \).

It will now be shown that in each of the cases (a), (b) and (c), \( d'(h) \to \infty \) as \( h \to H_p \) from below.

In case (a),
\[
d'(h) = \frac{\partial D}{\partial \lambda_x} - \frac{\partial D}{\partial \lambda_t} \quad \text{evaluated at} \quad \lambda_x + h(e_\alpha - e_t)
\]
\[
= -\frac{1}{\lambda_x} \sum_j f_j(\lambda_x) + \frac{1}{\lambda_x} \sum_j (a_{yj} - a_{ej}) f_j(\lambda_x) + \sum_j (a_{yj} - a_{ej}) \frac{\partial f_j}{\partial \lambda_x}(\lambda_x)
\]
evaluated at \( \lambda_x + h(e_\alpha - e_t) \),

and
\[
g'_j(\lambda_x) = \frac{\partial}{\partial \lambda_x} \left\{ \frac{y_j}{2(\lambda_x - y_j)} - \frac{y_j}{2\lambda_x} \right\}
\]
\[
= -\frac{y_j}{2(\lambda_x - y_j)^2} + \frac{y_j}{2\lambda_x^3}
\]
\[
\to -\infty \quad \text{as} \quad \lambda_x \to y_j.
\]

Now by definition of \( H_p \) in case (a), as \( h \to H_p \)
\( \lambda_x \to y_j \) for one or more \( j \) such that \( a_{xj} = 1 \).

Hence, for these \( j \), since \( a_{ej} < 1 \),
\[
(a_{yj} - a_{ej}) g'_j(\lambda_x) \to \infty \quad \text{as} \quad h \to H_p.
\]

All the other terms in the expression for \( d'(h) \) remain finite. Hence \( d'(h) \to \infty \) as \( h \to H_p \).

In case (b)
\[
d'(h) = \frac{\partial D}{\partial \lambda_x} - \frac{\partial D}{\partial \lambda_t} \quad \text{evaluated at} \quad \lambda_x + h(e_k - e_e)
\]
\[
\frac{1}{\lambda_0} \sum_{j=1}^{n} f_j(\Lambda_j) + \frac{1}{\lambda_0} \sum_{j=1}^{n} (a_{kj} - a_{o_j}) f'_j(\Lambda_j) + \frac{1}{\lambda_0} (a_{kj} - a_{o_j}) g'_j(\Lambda_j)
\]
evaluated at \(\lambda_p + h(e_k - e_o)\). Now if \(H_p = (\Lambda_j | p - y_j)/a_{o_j}\) for some \(j\) for which \(a_{o_j} > 0\) and \(a_{kj} = 0\), it follows from the behaviour of \(g_j(\Lambda_j)\) for this \(j\), as in case (a), that \(d'(h) \to \infty\) as \(h \to H_p\). If, however, \(H_p = \lambda_p o\), then as \(h \to H_p\), \(\lambda_o \to 0\) in the above expression for \(d'(h)\).

The terms in \(\lambda_o\) can be written
\[
\frac{1}{\lambda_0} \sum_{j=1}^{n} \left\{ \frac{1}{\lambda_0} f_j(\Lambda_j) + \lambda_o (a_{kj} - a_{o_j}) f'_j(\Lambda_j) \right\}
\]
and \(f_j(\Lambda_j) > 0\) \((j = 1, 2, \ldots, n)\).

Hence, again \(d'(h) \to \infty\) as \(h \to H_p\).

In case (c)
\[
d'(h) = \frac{\partial D}{\partial \lambda_k} - \frac{\partial D}{\partial \lambda_e} \text{ evaluated at } \lambda_p + h(e_k - e_e)
\]
\[
= \sum_{j=1}^{n} (a_{kj} - a_{o_j}) \left\{ \frac{1}{\lambda_0} f'_j(\Lambda_j) + \frac{1}{\lambda_0} g'_j(\Lambda_j) \right\}
\]
evaluated at \(\lambda_p + h(e_k - e_e)\), since \(k \neq 0\) and \(\ell \neq 0\).

If \(H_p\) is defined, then as \(h \to H_p\), \(\Lambda_j \to y_j\) for some \(j\) such that \(a_{e_j} = 1\) and \(a_{kj} = 0\). It therefore follows from the behaviour of \(g_j(\Lambda_j)\) for this \(j\), as in case (a), that \(d'(h) \to \infty\) as \(h \to H_p\).

Thus, whenever \(H_p\) is defined, \(d'(h) \to \infty\) as \(h \to H_p\). But \(d'(0) < 0\) and \(d'(h)\) is continuous for \(0 < h < H_p\). Hence \(d'(h) = 0\) for some \(h\) in this interval; i.e. \(d(h)\) has at least one turning value in the interval.

It has already been shown that any such value is unique and is a minimum.

Suppose now that \(H_p\) is undefined. The expression for \(d'(h)\) is as in case (c), but there is no \(j\) for which
\[ a_{\ell j} = 1 \text{ and } a_{kj} = 0, \text{ so that } a_{kj} - a_{\ell j} \geq 0 \quad (j = 1, 2, \ldots, n). \] Moreover \( y_j < \Lambda_j < 1 \) for each \( j \), so that

\[
\frac{1}{\lambda_o} f'_j(\Lambda_j) + g'_j(\Lambda_j) = -\frac{L_j(1-\Lambda_j)}{\lambda_0(1-y_j)} - \frac{y_j}{2(\Lambda_j - y_j)^2} + \frac{y_j}{2\Lambda_j^2} < 0
\]

and none of the contributions to \( \sum_{j=1}^{n}(a_{kj} - a_{\ell j}) \left\{ \frac{1}{\lambda_o} f'_j(\Lambda_j) + g'_j(\Lambda_j) \right\} \) can exceed zero. Also \( a_{kj} - a_{\ell j} = 1 \) for at least one \( j \) (for otherwise just the same approaches would have right of way in stage \( k \) as in stage \( \ell \)). Let \( J \) be such a value of \( j \); then throughout the interval \( 0 < h < \lambda_p t \)

\[ d'(h) \leq \frac{1}{\lambda_o} f'_J(\Lambda_J) + g'_J(\Lambda_J) < 0 \]

as required.

If \( \lambda_{p+1} = \lambda_p \pm h_p \left\{ e_o + \sum_{i \in C} k_i e_i - (1 + \sum_{i \in C} k_i)e_f \right\} \) with the sign chosen so that \( (\lambda_{p+1} - \lambda_p)^{-1} D(\lambda_p) < 0 \), let \( d(h) \) be defined by

\[ d(h) = D(\lambda_p \pm h(e_o + \sum_{i \in C} k_i e_i - (1 + \sum_{i \in C} k_i)e_f)) \]

Then \( H_p \) is always defined (see case (d) above) and it can be shown that Lemma 10 remains true with the new definition of \( d(h) \). The proof, which is exactly analogous to that of Lemma 10, but algebraically more tedious, will be omitted.

Let \( h_0 = \left\{ \begin{array}{ll}
\text{the } h \text{ in } 0 < h < H_p \text{ such that } d'(h) = 0, & \text{if } H_p \text{ is defined} \\
\lambda_p t & \text{if } H_p \text{ is undefined.}
\end{array} \right. \)

Then, by this definition, \( \lambda_p \pm h(e_k - e_f) \) or \( \lambda_p \pm h(e_o + \sum_{i \in C} k_i e_i - (1 + \sum_{i \in C} k_i)e_f) \)

satisfies the capacity constraints when \( 0 < h < h_o \). It
remains to examine the minimum green and cycle time constraints. If all these are satisfied with \( h = h_0 \), \( h_p \) will be given the value \( h_0 \); if not, \( h_p \) will be obtained by subtracting from \( h_0 \) an amount just sufficient to make \( \lambda_{p+1} \) satisfy these constraints. Once again, the cases (a), (b), (c) and (d) require separate examination.

In case (a), \( k = 0 \) and any of the minimum green constraints may be violated. The \( i \)th constraint is just satisfied by \( h = h_0 - h^* \), where

\[
\lambda_{pi} - \delta_{i\ell} (h_0 - h^*) = k_i (\lambda_{po} + h_0 - h^*) \quad (i=1,2,\ldots,m),
\]

\( \delta_{i\ell} \) being a Kronecker delta;

i.e.

\[
h^* = h_0 - \frac{\lambda_{pi} - k_i \lambda_{po}}{k_i + \delta_{i\ell}}.
\]

Hence, in case (a),

\[
h_p = h_0 - \max \{ 0, \max_{i \in \ell, \ell = m} \left( h_0 - \frac{\lambda_{pi} - k_i \lambda_{po}}{k_i + \delta_{i\ell}} \right) \}.
\]

In case (b), \( \ell = 0 \), only the cycle time constraint can be violated, and then only if \( k_o > 0 \). This constraint will just be satisfied by \( h = h_0 - h^* \) if

\[
\lambda_{po} - h_0 + h^* = k_o.
\]

Hence, in case (b),

\[
h_p = h_o - \max(0, h_o - \lambda_{po} + k_o).
\]

In case (c), \( k \neq 0 \) and \( \ell \neq 0 \), and the only constraint that can be violated is the \( \ell \)th minimum green constraint. This will just be satisfied by \( h = h_0 - h^* \) if

\[
\lambda_{p\ell} - h_0 + h^* = k_i \lambda_{po}.
\]
Hence, in case (c),

\[ h_p = h_o - \max(0, h_o - \lambda_p \ell + k_\ell \lambda_{po}) \]

In case (d) the minimum green constraints \( \lambda_i \geq k_i \lambda_o \) are automatically satisfied for \( i \) in \( C \).

If the positive sign applies, the constraints that can be violated are the minimum green constraints \( \lambda_i \geq k_i \lambda_o \) for \( i \) in \( F \). The constraint \( \lambda_{\ell} \geq k_\ell \lambda_o \) will just be satisfied with \( h = h_o - h^* \) if

\[ \lambda_{p\ell} - \delta_{ef} (1 + \sum_{i \in C} k_i)(h_o - h^*) = k_\ell (\lambda_{po} + h_o - h^*) \]

and, as in case (a),

\[ h_p = h_o - \max\left\{ 0, \max_{\ell \in F} (h_o - \frac{\lambda_{p\ell} - k_\ell \lambda_{po}}{k_\ell + \delta_{ef} (1 + \sum_{i \in C} k_i)}) \right\} \]

If the negative sign applies, only the cycle time constraint can be violated, and then only if \( k_o \geq 0 \). As in case (b)

\[ h_p = h_o - \max(0, h_o - \lambda_{po} + k_o) \]

This completes the determination of \( h_p \).

3.53. Convergence. It was shown in Theorem 4 that feasible solutions \( \Lambda \) such that \( D(\Lambda) \leq D(\Lambda_1) \) are confined to a closed subset, \( F_1 \), say, of the set of feasible solutions. \( F_1 \) is the set of feasible \( \Lambda \) such that \( \lambda_o \geq Y_o \) and \( \Lambda_j \geq Y_j \) \( (j = 1, 2, \ldots, n) \), where \( Y_o \) and the \( Y_j \) are as defined in Theorem 4. \( F_1 \) is then a closed and bounded
set in $\mathbb{R}^{m+1}$ and the sequence $\{\lambda_p\}$ just constructed, being such that $D(\lambda_p) \leq D(\lambda_1)$ for all $p$, lies in $F_1$, and therefore has at least one limit point there. It will be shown in Theorem 6 that any such limit point is an optimal solution.

Let $\lambda^*$ be a feasible solution that is not optimal. It is first necessary to define, in terms of $\lambda^*$, certain new functions of $\lambda$ for $\lambda$ near to $\lambda^*$, and to show that they are continuous. Let $k^*$ denote the evaluation at $\lambda^*$ of any function of $\lambda$, let $k^*$ and $l^*$ be chosen at $\lambda^*$ as $k$ and $l$ were chosen at $\lambda_p$ in section 3.51 and let $Z$ be the set of all pairs $(r,s)$ such that $\frac{\partial D}{\partial \lambda_r} \leq \frac{\partial D}{\partial \lambda_s}$ and $\frac{\partial D}{\partial \lambda_r} \geq \frac{\partial D}{\partial \lambda_s}$. Four cases will be considered.

Case (a): $\lambda^*_i > k^*_i \lambda^*_i$ ($i = 1, 2, \ldots, m$) and either $\lambda^*_i > k^*_i$ or the cycle time is specified. In this case $\left(\frac{\partial D}{\partial \lambda_k} - \frac{\partial D}{\partial \lambda_s}\right)_r < 0$ (see Lemma 9). Hence by the continuity of the components of $\nabla D$ there is a neighbourhood $N_1$ of $\lambda^*$ in $\mathbb{R}^{m+1}$ or the subspace $\{\lambda^*_i = k^*_i\}$ (according as the cycle time is unspecified or not) such that, for all $\lambda$ in $N_1$, 

(i) $\lambda^*_i > k^*_i$ if the cycle time is unspecified,

(ii) $\lambda^*_i > k^*_i \lambda^*_i$ ($i = 1, 2, \ldots, m$),

(iii) $\frac{\partial D}{\partial \lambda_r} - \frac{\partial D}{\partial \lambda_s} < 0$ for all $(r,s)$ in $Z$, and

(iv) if $k^*$ and $l^*$ are chosen at $\lambda$ as in section 3.51 then $(k^*, l^*) \in Z$. It follows from the sign of $\frac{\partial D}{\partial \lambda_r} - \frac{\partial D}{\partial \lambda_s}$ that, for all $\lambda$ in $N_1$ and $(r,s)$ in $Z$, numbers $h_{rs}(\lambda)$ and $h_{rs}(\lambda)$ can be defined for $\lambda$, $r$ and $s$ just as $h_o$ and $h_o$ respectively were defined for $\lambda_p$, $k$ and $l$ in section 3.52. Then for $\lambda$ in $N_1$ let

$$D_{rs}(\lambda) = D(\lambda) - D\left\{\lambda + h_{rs}(\lambda)(e_r - e_s)\right\}$$

and

$$D^*(\lambda) = \min_{(r,s) \in Z} \left\{D_{rs}(\lambda)\right\}.$$ 

Case (b): $\lambda^*_i > k^*_i \lambda^*_i$ ($i=1, 2, \ldots, m$) and the cycle time is unspecified, but $\lambda^*_i = k^*_i$. In this case $l^* \neq 0$; let $Z_o = \{(r,s); (r,s) \in Z \text{ and } s \neq 0\}$. If $Z_o = Z$ let $N_1$ be a neighbourhood of $\lambda^*$ in $\mathbb{R}^{m+1}$ such that any $\lambda$ in $N_1 \cap \{\lambda^*_i \geq k^*_i\}$
has the properties (ii), (iii) and (iv) of case (a), and let \( D^*(\lambda) \) be defined as in case (a) for all \( \lambda \) in \( N_1 \bigcap \{ \lambda_0 \geq k_0 \} \).

If \( Z_0 \neq \emptyset \) let \( N_1 \) be a neighbourhood of \( \lambda^* \) in \( \mathbb{R}^{m+1} \) such that, for all \( \lambda \) in \( N_1 \), \( \lambda_0 > 0 \) and properties (ii), (iii), and (iv) of case (a) hold. Then for \((r,s)\) in \(Z\), let \( D_{rs}(\lambda) \) and \( D^*(\lambda) \) be defined as in case (a) except that if \( s = 0 \), \( h_{rs}(\lambda) \) is determined as though the cycle time constraint were \( \lambda_0 > 0 \).

Case (c): \( \lambda_j^* = k_j \lambda_0^*, \lambda_i^* > k_i \lambda_0^* \) (\( 1 \leq i \leq m, i \neq j \)) and, at \( \lambda^* \),

\[
(1 + k_j) \left( \frac{\partial D}{\partial \lambda_j} + \frac{\partial D}{\partial \lambda_k^*} \right) < \pm \left( \frac{\partial D}{\partial \lambda_0^*} + k_j \frac{\partial D}{\partial \lambda_j} - (1 + k_j) \frac{\partial D}{\partial \lambda_k^*} \right) \quad \cdots \quad \cdots \quad \cdots \quad (*)
\]

with the sign chosen so that the right hand expression is negative. In this case \( k^* \neq 0 \) and \( f^* \neq j \). Let \( Z_j = \{(r,s); (r,s) \in Z, r \neq 0 \text{ and } s \neq j\} \), and let \( f^* \) and the set \( C^* \) be defined at \( \lambda^* \) just as \( f \) and \( C \) were defined at \( \lambda_p \) in section 3.51, so that \( C^* = \{j\} \). Then if \( Z_j = \emptyset \) let \( N_1 \) be a neighbourhood of \( \lambda^* \) such that for any \( \lambda \) in \( N_1 \bigcap \{ \lambda_j > k_j \lambda_0 \} \) properties (i), (iii), and (iv) of case (a) hold, inequality (*) is true with the sign chosen at \( \lambda^* \), and \( \lambda_i > k_i \lambda_0 \) (\( 1 \leq i \leq m, i \neq j \)). Then let \( D^*(\lambda) \) be defined as in case (a) for all \( \lambda \) in \( N_1 \bigcap \{ \lambda_j \geq k_j \lambda_0 \} \). If \( Z_j \neq \emptyset \) let \( N_1 \) be a neighbourhood of \( \lambda^* \) such that for any \( \lambda \) in \( N_1 \) properties (i), (iii) and (iv) of case (a) hold, inequality (*) is true with the sign chosen at \( \lambda^* \), and \( \lambda_i > k_i \lambda_0 \) (\( 1 \leq i \leq m, i \neq j \)). Then for \((r,s)\) in \(Z\), let \( D_{rs}(\lambda) \) and \( D^*(\lambda) \) be defined as in case (a) except that, if \( r = 0 \) or \( s = j \) or both, the constraint \( \lambda_j \geq k_j \lambda_0 \) is ignored in determining \( h_{rs}(\lambda) \).

Case (d): as case (c) except that, at \( \lambda^* \),

\[
\pm \left( \frac{\partial D}{\partial \lambda_0^*} + k_j \frac{\partial D}{\partial \lambda_j} - (1 + k_j) \frac{\partial D}{\partial \lambda_k^*} \right) < \pm \left( 1 + k_j \right) \frac{\partial D}{\partial \lambda_k^*} \quad \cdots \quad \cdots \quad \cdots \quad (***)
\]

with the sign chosen so that the left hand expression is negative. By definition of \( k^* \), \( f^* \) and \( f^* \), \( \frac{\partial D}{\partial \lambda_j^*} \geq \frac{\partial D}{\partial \lambda_k^*} \) and \( \frac{\partial D}{\partial \lambda_k^*} \leq \frac{\partial D}{\partial \lambda_j^*} \). If both \( \frac{\partial D}{\partial \lambda_j^*} \geq \frac{\partial D}{\partial \lambda_k^*} \) and \( \frac{\partial D}{\partial \lambda_k^*} \leq \frac{\partial D}{\partial \lambda_j^*} \), then it would follow that
\[
\left|(\frac{\partial P}{\partial \lambda_0} + \kappa_j \frac{\partial P}{\partial \lambda_j} - (1 + \kappa_j) \frac{\partial P}{\partial \lambda_k})\right|_* \leq (1 + \kappa_j)\left(\frac{\partial P}{\partial \lambda_0} - \frac{\partial P}{\partial \lambda_k}\right)|_*
\]
which contradicts (**). Hence even if \(\left(\frac{\partial P}{\partial \lambda_k} - \frac{\partial P}{\partial \lambda_k}\right)|_* = 0\) there is at least one pair \((r,s)\) in \(Z\) such that \(\left(\frac{\partial P}{\partial \lambda_r} - \frac{\partial P}{\partial \lambda_s}\right)|_* < 0\).

Let \(Z_{C^*}\) be the set of all such pairs, and let \(N_1\) be a neighbourhood of \(\Lambda^*\) such that for all \(\Lambda\) in \(N_1\), \(\lambda_i \geq k_i \lambda_0\) \((1 < i < m, i \neq j)\), \(\lambda_i \geq k_o\) if the cycle time is not specified, inequality (**) holds with the sign chosen at \(\Lambda^*\) and its left hand side is negative, \(\frac{\partial P}{\partial \lambda_r} - \frac{\partial P}{\partial \lambda_s} < 0\) for all \((r,s)\) in \(Z_{C^*}\), and if \(k\) and \(\ell\) are chosen at \(\Lambda\) as in section 3.51, either \(k = 0\) or \(\ell = j\) or both. Now for \(\Lambda\) in \(N_1\) and \((r,s)\) in \(Z_{C^*}\) let \(D_{rs}(\Lambda)\) be defined as in case (a) except that the constraint \(\lambda_j \geq k_j \lambda_0\) is ignored in determining \(h_{rs}(\Lambda)\) and let \(D_{C^*}(\Lambda) = D(\Lambda) - D(\Lambda + h_{C^*}(\Lambda)\{e_{\phi} + k_j e_{j} - (1 + k_j) e_{r}\})\), where \(h_{C^*}(\Lambda)\) is determined at \(\Lambda\) as \(h_p\) was at \(\Lambda_p\) in case (d) of section 3.52.

Let \(D_{C^*}(\Lambda) = \min \left\{D_{C^*}(\Lambda), \min_{(r,s) \in Z_{C^*}} \left\{D_{rs}(\Lambda)\right\}\right\}\)

**Lemma 11.** \(D^*(\Lambda)\) and \(D_{C^*}(\Lambda)\) are continuous functions of \(\Lambda\) in the respective neighbourhoods \(N_1\) in which they are defined.

**Proof.** Either \(h_{rso}(\Lambda) = \lambda_s\) or \(h_{rso}(\Lambda)\) is a solution of the equation \(\frac{\partial}{\partial h} D(\Lambda + h(e_r - e_s)) = 0\), where the left hand side is a continuous function of \(\Lambda\) for values of \(h\) in the neighbourhood of \(h_{rso}(\Lambda)\) and \(\frac{\partial}{\partial h} D(\Lambda + h(e_r - e_s)) > 0\) for all \(\Lambda\) in \(N_1\). Hence, either trivially or by the implicit function theorem (Hardy 1955), \(h_{rso}(\Lambda)\) is a continuous function of \(\Lambda\) in \(N_1\).

Again \(h_{rs}(\Lambda)\) differs from \(h_{rso}(\Lambda)\) by a continuous function of \(\Lambda\), and \(D(\Lambda)\) is continuous, so that \(D_{rs}(\Lambda)\) is continuous in \(N_1\) for each pair \((r,s)\) in \(Z\). Similarly \(D_{C^*}(\Lambda)\) is continuous in \(N_1\) and so therefore are \(D^*(\Lambda)\) and \(D_{C^*}(\Lambda)\).

**Theorem 6.** If \(\Lambda^*\) is a limit point of the sequence \(\{\Lambda_p\}\) then \(\Lambda^*\) is an optimal solution.
PROOF. Suppose that $\lambda^*$ is not optimal. For any $\delta > 0$ there is, by the continuity of $D(\lambda)$, a neighbourhood $N_2$ of $\lambda^*$ such that for all $\lambda$ in $N_2$ $D(\lambda^*) - \frac{1}{4}\delta < D(\lambda) < D(\lambda^*) + \frac{1}{4}\delta$. It will be shown that in each of the cases considered above there is a $q$ such that if $p > q$, $D(\lambda_p) < D(\lambda^*) - \frac{1}{4}\delta$, so that $\lambda_p$ lies outside $N_2$. This contradicts the fact that $\lambda^*$ is a limit point of $\{\lambda_p\}$ so that $\lambda^*$ must be optimal. The result can be proved similarly but at greater length if $\lambda^*$ lies in two or more of the hyperplanes bounding the set of feasible solutions.

In case (a), in case (b) if $Z_0 = Z$, and in case (c) if $Z_j = Z$, $D_{rs}(\lambda^*) > 0$ for all $(r,s)$ in $Z$. Hence, since $Z$ is finite, $D^*(\lambda^*) > 0$. Let $D^*(\lambda^*) = \delta$. Then by the continuity of $D^*(\lambda)$ there is a neighbourhood $N_3$ of $\lambda^*$ such that $N_3 \subseteq N_1$ and $D^*(\lambda) > \frac{1}{2}\delta$ for all $\lambda$ in $N_3$. Since $\lambda^*$ is a limit point of $\{\lambda_p\}$ there is a $q$ such that $\lambda_q$ is in $N_2 \cap N_3$ and therefore in $N_1$. Hence, for some $(r,s)$ in $Z$, $D(\lambda_{q+1}) = D(\lambda_q) - D_{rs}(\lambda_q) < D(\lambda_q) - D^*(\lambda_q) < D(\lambda^*) + \frac{1}{4}\delta - \frac{1}{2}\delta$. And $\{D(\lambda_p)\}$ is a decreasing sequence, so that $D(\lambda_p) < D(\lambda^*) - \frac{1}{4}\delta$ for all $p > q$ as required.

In case (b) if $Z_0 \neq Z$, let $\delta$ and $N_3$ be as in case (a) and let $N_4$ be the neighbourhood $\{\lambda; |\lambda - \lambda^*| < \varepsilon\}$ of $\lambda^*$, where $\varepsilon > 0$ and is such that $\{\lambda; |\lambda - \lambda^*| < (1 + \sqrt{2})\varepsilon\} \subseteq N_2 \cap N_3$. Let $q$ be such that $\lambda_q$ is in $N_4$ and therefore in $N_1$ and $N_3$. If the $k$ and $\ell$ chosen at $\lambda_q$ as in section 3.51 are such that $(k,\ell)$ is in $Z_0$ or both $\ell = 0$ and $\lambda_q - h_k\ell_0(\lambda_q) > k_0$, then $D(\lambda_{q+1}) = D(\lambda_q) - D_k\ell(\lambda_q)$, and as in case (a) $D(\lambda_p) < D(\lambda^*) - \frac{1}{4}\delta$ for all $p > q$ as required. If, however, $\ell = 0$ and $\lambda_{q_0} - h_k\ell_0(\lambda_q) < k_0$, it follows from section 3.52 that $h_q = \lambda_{q_0} - k_0$ and $\lambda_{q+1} = k_0$. Moreover, since $\lambda_q^* = k_0$, $|\lambda_{q+1} - \lambda^*| < |\lambda_{q+1} - \lambda_q| + |\lambda_q - \lambda^*| = \sqrt{2}(\lambda_{q_0} - k_0) + |\lambda_q - \lambda^*| < (1 + \sqrt{2})|\lambda_q - \lambda^*| < (1 + \sqrt{2})\varepsilon$. Hence, by definition of $\varepsilon$,
\( \lambda_{q+1} \) is in \( N_2 \cap N_3 \). Since \( \lambda_{q+1} \circ = k_0 \), the \( k \) and \( \ell \) chosen at \( \lambda_{q+1} \) as in section 3.51 are such that \( (k, \ell) \) is in \( Z_0 \), so that \( D(\lambda_{q+2}) = D(\lambda_{q+1}) - D_{k\ell}(\lambda_{q+1}) \), and, as in case (a) \( D(\lambda_p) < D(\lambda^*) - \frac{1}{4} \delta \) for all \( p > q+1 \) as required.

In case (c) if \( Z_j \neq Z \) the proof is as for case (b) \( (Z_0 \neq Z) \) with \( Z_0 \) replaced by \( Z_j \) and, in the definition of \( \varepsilon \), \( \sqrt{2} \) replaced by \( (1 + k_j)\sqrt{2}/k_j \).

In case (d) let \( D_{0*}(\lambda^*) = \delta \). Then, as in case (a), \( \delta > 0 \) and there is a neighbourhood \( N_3 \) of \( \lambda^* \) such that \( N_3 \subseteq N_1 \) and \( D_{0*}(\lambda) > \frac{1}{2} \delta \) for all \( \lambda \) in \( N_3 \). Then, as in case (b) \( (Z_0 \neq Z) \) with \( \sqrt{2} \) replaced by \( (1 + k_j)\sqrt{2}/k_j \) in the definition of \( \varepsilon \), if \( \lambda_q \) is in \( N_4 \) either \( D(\lambda_{q+1}) = D(\lambda_q) - D_{rs}(\lambda_q) \) for some \( (r, s) \) in \( Z_{0*} \), or \( D(\lambda_{q+1}) = D(\lambda_q) - D_{0*}(\lambda_q) \), or \( \lambda_{q+1} \) is in \( N_2 \cap N_3 \), \( \lambda_{q+1} \circ = k_j \lambda_{q+1} \) and \( D(\lambda_{q+2}) = D(\lambda_{q+1}) - D_{0*}(\lambda_{q+1}) \). Hence \( D(\lambda_p) < D(\lambda^*) - \frac{1}{4} \delta \) for all \( p > q \) in the first two cases and for all \( p > q+1 \) in the third case, as required.

The sequence generated by the algorithm described in the preceding two sections must thus converge to a value of \( \lambda \) that corresponds to delay-minimising settings.

3.54. Possible effect of the third term in Webster's expression. The use of the function \( D(\lambda) \) as a measure of the rate of delay does not take full account of the third term in Webster's expression for the average delay per p.c.u. on an approach. It has, in effect, been assumed that, for feasible \( \lambda \), the total contribution made by the third term to the rate of delay for the intersection is proportional to the contribution made by the first two terms. Let the
contribution of the third term be $O(\lambda)$.

Then

$$C(\lambda) = -\sum_{j=1}^{n} 0.65 \left( \frac{Lq_i}{\lambda_0} \right)^{1.5} \left( \frac{y_i}{\lambda_j} \right)^{2 + 5\lambda_j} \right)$$

$$= -\sum_{j=1}^{n} c_j \left( \frac{y_i}{\lambda_j} \right)^{2 + 5\lambda_j} ,$$

where

$$c_j = 0.65 \left( Lq_i \right)^{1.5}$$

Differentiation gives:

$$\frac{\partial C}{\partial \lambda_0} = \sum_{j=1}^{n} c_j \left( \frac{y_i}{\lambda_j} \right)^{2 + 5\lambda_j} \left\{ \frac{1}{3\lambda_0} + 5a_{ij} \left( 1 + \frac{2}{5\lambda_j} - \log \frac{y_i}{\lambda_j} \right) \right\}$$

and

$$\frac{\partial C}{\partial \lambda_i} = 5 \sum_{j=1}^{n} \frac{a_{ij} c_j}{\lambda_0^3} \left( \frac{y_i}{\lambda_j} \right)^{2 + 5\lambda_j} \left\{ 1 + \frac{2}{5\lambda_j} - \log \frac{y_i}{\lambda_j} \right\} \quad (i = 1, 2, \ldots, m)$$

The values of these components of $\nabla C(\lambda)$ can be used in most cases to obtain some indication whether the inclusion of the third term would be likely to affect the delay-minimising solutions appreciably. Consider the components of $\nabla D(\lambda)$ when $\lambda$ is optimal; if $k$ and $\ell$ are chosen at $\lambda$ by the rules of section 3.51 then, as in the proof of Lemma 9, it follows that $\frac{\partial D}{\partial \lambda_k} - \frac{\partial D}{\partial \lambda_\ell} = 0$ for all pairs $(k, \ell)$ such that both $e_k - e_\ell$ and $e_\ell - e_k$ are permissible directions at $\lambda$. In most cases there is at least one such pair. Suppose that there is; then $\partial D/\partial \lambda_i$ has the same value for all values of $i$ that appear in such pairs. The values of $\partial C/\partial \lambda_i$ for these $i$ will not in general be equal, and the range of these values, compared with the common value of the $\partial D/\partial \lambda_i$, is an indication of the extent to which the inclusion of $C(\lambda)$ could affect the optimal solution, and hence the delay-minimising settings. This range has so far
been found to be small. If a case arose in which it was large, then it would in principle be possible to start from the optimal solution obtained by the algorithm of sections 3.51 - 3.52 and then apply the algorithm again with $D(\lambda)$ replaced by $D(\lambda) + \frac{9}{10} C(\lambda)$. The convergence properties of the algorithm applied to the latter function have, however, not yet been investigated.

If there is no pair $(k, \ell)$ such that $+ (c_k - c_\ell)$ are both permissible directions at $\lambda$, then the position of the optimal solution is determined entirely by minimum green and cycle time constraints, and is therefore unlikely to be affected by the third term in the delay expression.

3.6. Computer program

A Fortran program has been written for calculating delay-minimising settings using the algorithm described in section 3.5. The following data are required: -

- number of stages and approaches
- maximum or specified cycle time, if any
- minimum green time for each stage
- lost time following each stage
- average arrival rate
- saturation flow
- extra effective green time

The numbers of the stages in which each approach has right of way

The stages are numbered in the order in which they occur in the cycle; the approaches may be numbered in any order. If
required, the saturation flows can be estimated by the program from the dimensions and characteristics of the approaches, using the methods of Webster and Cobbe (1966). Extensive data checking is carried out, and if no feasible solution is found to exist, a warning is printed, together with an estimate of the extent to which the intersection is overloaded.

The results provide the cycle time and effective green times corresponding to the optimal solution found by the algorithm, together with a table of average delays and rates of delay on the approaches, and the total rate of delay for the intersection.

When the calculation for one set of data is complete the program can be required either to amend this set of data and make further calculations, or to proceed to a completely new set of data.

The criteria for convergence of \( \{ \lambda_p \} \) to an optimal solution is the value of \( \left( \frac{\partial D}{\partial \lambda_k} - \frac{\partial D}{\partial \lambda_f} \right) \) or
\[
\frac{\partial D}{\partial \lambda_0} + \sum_{i \in C} k_i \frac{\partial D}{\partial \lambda_i} - (1 + \sum_{i \in C} k_i) \frac{\partial D}{\partial \lambda_f}
\] (see section 3.51), unless the rules of choice fail to yield values of \( k \) and \( \ell \), in which case, as was shown in section 3.51, an optimal solution has been found. If the rules continue to yield values of \( k \) and \( \ell \) for successive \( p \), \( \lambda_p \) is regarded as a sufficiently close approximation to an optimal solution when
\[
\left( \frac{\partial D}{\partial \lambda_k} - \frac{\partial D}{\partial \lambda_f} \right) \bigg|_p < 0.1 \] (which is usually of the order of 0.1 per cent of \( \frac{\partial D}{\partial \lambda_k} \)) and, if one or more of the minimum green constraints is then satisfied,
\[
\left| \frac{\partial D}{\partial \lambda_0} + \sum_{i \in C} k_i \frac{\partial D}{\partial \lambda_i} - (1 + \sum_{i \in C} k_i) \frac{\partial D}{\partial \lambda_f} \right| < 0.1 \]
This criterion has so far been satisfied for values of \( p \) between 3 and 100. It may be possible in the light of experience to relax the criterion somewhat, and to provide guidance about this the values of \( \Lambda_p \) and \( \overline{D(\Lambda_p)} \) are printed for each \( p \).

In determining \( \Lambda_{p+1} \) from \( \Lambda_p \), the value of \( h_o \) (see section 3.52) is determined by successive bisection of the interval \( 0 < h < H_p \), stopping when a value of \( h \) has been found which makes \( \left| d'(h) \right| < \left| d'(0) \right|/100 \).

The running time on the University College IBM 360/65 is a few seconds for each optimal solution found. About 33 000 bytes (8250 words) of core store are required on this machine.

3.7. A few illustrative results

Three examples will now be given to provide brief illustrations of the application of the mathematical technique just described to realistic situations. The first example is a crossing of two similar roads and the second is a crossing of a wide road by a narrow one; at each of these two intersections the signal cycle has two stages and there are four approaches. Webster's method can be used to calculate settings and the results are compared with the settings given by the present method. The third example is one given by Webster and Cobbe (1966); in this case the intersection is more complicated and Webster's method is less easy to apply. The signal cycle in the third example has four stages and there are seven approaches. The layouts of the intersections, with
Example 1

Example 2

Example 3

Fig. 5. LAYOUTS OF THE EXAMPLE INTERSECTIONS, WITH APPROACH NUMBERS CIRCLED AND SATURATION FLOWS IN p.c.u./hour
approach numbers and saturation flows are shown in Figure 5. For simplicity, no minimum green or cycle time constraints are imposed.

3.7. Calculations for three examples. EXAMPLE 1. Suppose that approaches 1 and 2 have right of way in stage 1 and approaches 3 and 4 in stage 2, and that the lost time after each stage is 5 seconds. Settings have been calculated for two ranges of average arrival rates.
In the first range of arrival rates, for which results are given in Table 1A, the flow ratio on the representative approach of stage 1 varies from $1/3$ to 6/10, and the corresponding figure for stage 2 is always half as great. Thus Webster's method allocates twice as much green time to stage 1 as to stage 2. The sum of these two flow ratios varies across the Table from 0.5 to 0.9 and, in Webster's method, determines the cycle time as described in section 2.72; the results are shown in part (a) of the Table. The ratio of the arrival rate on the other approach to that on the representative approach is always taken to be the same for both stages, and varies from 0.2 to 1. The value of this ratio does not affect the signal setting given by Webster's method; it does affect the cycle time given by the present method, though not, in this case, the allocation of green time, as parts (b) and (c) of the Table show. The estimated rates of delay corresponding to the Webster settings and to those given by the present method are compared in part (d) of the Table; the difference is usually negligible and never exceeds about 2 per cent.

In the second range of arrival rates, for which results are given in Table 1B, the flow ratio on each of the representative approaches is kept fixed at 5/12, whilst the ratio of the arrival rate on the other approach to that on the representative approach is allowed to vary from 0.2 to 1 for each stage separately. Webster's method gives the same setting in every case: a cycle time of 120 seconds and equal green times for the two stages. The cycle times and alloc-
Table 1A

Comparison of signal settings given by the present method and Webster's method for the first range of arrival rates in Example 1

| Ratio of arrival rate on other approach to that on representative approach (same for both stages) | Sum of flow ratios on representative approaches |
|---|---|---|---|---|---|
| | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |

(a) **Cycle time (in seconds) given by Webster's method**

- 40.0
- 50.0
- 66.7
- 100.0
- 200.0

(b) **Cycle time (in seconds) given by the present method**

<table>
<thead>
<tr>
<th>Ratio</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>41.4</td>
<td>51.0</td>
<td>67.0</td>
<td>98.6</td>
<td>193.2</td>
</tr>
<tr>
<td>0.4</td>
<td>40.3</td>
<td>49.7</td>
<td>65.1</td>
<td>95.8</td>
<td>187.7</td>
</tr>
<tr>
<td>0.6</td>
<td>39.9</td>
<td>48.7</td>
<td>63.6</td>
<td>93.2</td>
<td>182.3</td>
</tr>
<tr>
<td>0.8</td>
<td>40.2</td>
<td>49.0</td>
<td>63.2</td>
<td>91.7</td>
<td>177.3</td>
</tr>
<tr>
<td>1.0</td>
<td>42.7</td>
<td>52.8</td>
<td>69.2</td>
<td>101.6</td>
<td>198.6</td>
</tr>
</tbody>
</table>

(c) **Ratio of green time for stage 1 to green time for stage 2 given by the present method**

<table>
<thead>
<tr>
<th>Ratio</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>1.89</td>
<td>1.90</td>
<td>1.91</td>
<td>1.94</td>
<td>1.97</td>
</tr>
<tr>
<td>0.4</td>
<td>1.90</td>
<td>1.90</td>
<td>1.92</td>
<td>1.94</td>
<td>1.97</td>
</tr>
<tr>
<td>0.6</td>
<td>1.90</td>
<td>1.90</td>
<td>1.92</td>
<td>1.94</td>
<td>1.97</td>
</tr>
<tr>
<td>0.8</td>
<td>1.89</td>
<td>1.89</td>
<td>1.91</td>
<td>1.94</td>
<td>1.97</td>
</tr>
<tr>
<td>1.0</td>
<td>1.89</td>
<td>1.90</td>
<td>1.91</td>
<td>1.94</td>
<td>1.97</td>
</tr>
</tbody>
</table>

(d) **Estimated rate of delay (in p.c.u.'s) for settings given by the present method, with that for the Webster settings in brackets.**

<table>
<thead>
<tr>
<th>Ratio</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>3.7(3.6)</td>
<td>5.5(5.5)</td>
<td>8.6(8.7)</td>
<td>15.1(15.3)</td>
<td>35.6(36.0)</td>
</tr>
<tr>
<td>0.4</td>
<td>4.1(4.1)</td>
<td>6.2(6.2)</td>
<td>9.6(9.6)</td>
<td>16.6(16.8)</td>
<td>38.5(39.0)</td>
</tr>
<tr>
<td>0.6</td>
<td>4.7(4.7)</td>
<td>7.0(7.0)</td>
<td>10.7(10.8)</td>
<td>18.4(18.7)</td>
<td>42.1(42.9)</td>
</tr>
<tr>
<td>0.8</td>
<td>5.5(5.5)</td>
<td>8.1(8.1)</td>
<td>12.4(12.5)</td>
<td>21.0(21.3)</td>
<td>47.2(48.3)</td>
</tr>
<tr>
<td>1.0</td>
<td>6.5(6.5)</td>
<td>9.9(9.9)</td>
<td>15.7(15.8)</td>
<td>27.8(28.1)</td>
<td>66.3(67.1)</td>
</tr>
<tr>
<td>Rate of arrival rate on other approach to that on representative approach; stage 2</td>
<td>Ratio of arrival rate on other approach to that on representative approach; stage 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.6</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>114.6</td>
<td>112.9</td>
<td>111.2</td>
<td>109.8</td>
<td>116.5</td>
</tr>
<tr>
<td>0.4</td>
<td>111.4</td>
<td>109.8</td>
<td>108.5</td>
<td>107.1</td>
<td>113.5</td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td>108.3</td>
<td>107.1</td>
<td>106.0</td>
<td>112.2</td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td></td>
<td>107.1</td>
<td>112.2</td>
<td>117.9</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Cycle time (in seconds) given by the present method

<table>
<thead>
<tr>
<th>Rate of arrival rate on other approach to that on representative approach; stage 2</th>
<th>Ratio of arrival rate on other approach to that on representative approach; stage 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>0.2</td>
<td>1.00</td>
</tr>
<tr>
<td>0.4</td>
<td>1.00</td>
</tr>
<tr>
<td>0.6</td>
<td>1.00</td>
</tr>
<tr>
<td>0.8</td>
<td>1.00</td>
</tr>
<tr>
<td>1.0</td>
<td>1.00</td>
</tr>
</tbody>
</table>

(b) Ratio green time for stage 1 green time for stage 2 given by the present method

(c) Estimated rate of delay (in p.c.u.'s) for settings given by the present method, with that for the Webster setting in brackets

<table>
<thead>
<tr>
<th>Rate of arrival rate on other approach to that on representative approach; stage 2</th>
<th>Ratio of arrival rate on other approach to that on representative approach; stage 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>0.2</td>
<td>20.7(20.8)</td>
</tr>
<tr>
<td>0.4</td>
<td>22.7(22.8)</td>
</tr>
<tr>
<td>0.6</td>
<td>25.1(25.4)</td>
</tr>
<tr>
<td>0.8</td>
<td>28.5(29.0)</td>
</tr>
<tr>
<td>1.0</td>
<td>38.3(38.3)</td>
</tr>
</tbody>
</table>
ations of green time given by the present method are shown in parts (a) and (b) of the Table; most of the cycle times are appreciably less than 120 seconds. The estimated rates of delay corresponding to the Webster setting and to the settings given by the present method are compared in part (c) of the Table; once again, the difference never exceeds about 2 per cent.

EXAMPLE 2: The lost time and stage matrix are as in Example 1, the difference being that approaches 1 and 2 now have saturation flows twice as great as those of approaches 3 and 4.

Calculations have been made for arrival rates such that the wider road is heavily loaded in at least one direction, whilst the narrower road is less heavily loaded. The results are given in Table 2. The flow ratios on the representative approaches are kept fixed at 5/9 in stage 1 and 2/9 in stage 2. The setting given by Webster's method is therefore always the same: a cycle time of 90 seconds and green time allocated in the ratio 5:2. The ratio of the arrival rate on the other approach to that on the representative approach is allowed to vary from 0.2 to 1 in stage 1 (i.e. on the wider road) and is given the two extreme values 0 and 1 in stage 2 (i.e. on the narrower road). The cycle times and the allocations of green time given by the present method are shown in parts (a) and (b) of the Table; once again the main differences from the Webster setting are in the cycle times. The estimated rates of delay for the Webster setting and the settings given by
Table 2

Comparison of signal settings given by the present method and Webster's method in Example 2

Webster setting: cycle time = 90 seconds; \( \frac{\text{green time for stage 1}}{\text{green time for stage 2}} = 2.5 \)

| Ratio of arrival rate on other approach to that on representative approach; stage 2 | Ratio of arrival rate on other approach to that on representative approach; stage 1 |
|---|---|---|---|---|---|
| 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |

(a) Cycle time (in seconds) given by the present method

<table>
<thead>
<tr>
<th>0</th>
<th>85.6</th>
<th>84.4</th>
<th>83.3</th>
<th>82.5</th>
<th>88.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>83.8</td>
<td>82.8</td>
<td>81.8</td>
<td>81.1</td>
<td>86.6</td>
</tr>
</tbody>
</table>

(b) Ratio \( \frac{\text{green time for stage 1}}{\text{green time for stage 2}} \) given by the present method

<table>
<thead>
<tr>
<th>0</th>
<th>2.44</th>
<th>2.44</th>
<th>2.45</th>
<th>2.47</th>
<th>2.35</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.30</td>
<td>2.31</td>
<td>2.32</td>
<td>2.34</td>
<td>2.43</td>
</tr>
</tbody>
</table>

(c) Estimated rate of delay (in p.c.u.'s) for settings given by the present method, with that for the Webster setting in brackets

<table>
<thead>
<tr>
<th>0</th>
<th>15.7(15.9)</th>
<th>16.8(16.9)</th>
<th>18.2(18.3)</th>
<th>20.2(20.4)</th>
<th>24.4(24.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.6(21.6)</td>
<td>21.7(22.6)</td>
<td>23.1(24.1)</td>
<td>25.2(26.1)</td>
<td>30.0(30.2)</td>
</tr>
</tbody>
</table>
the present method are compared in part (c) of the Table. Once again the difference is usually small, but in one case it is almost 5 per cent.

EXAMPLE 3: (cf. Webster and Cobbe, 1966, pp. 64-65) There are 7 approaches and the signal cycle has 4 stages, in the sense of section 1.23. (Webster and Cobbe use the word 'stage' in a rather different sense; what they call stage 2 is, as they point out, really lost time, and what they call stages 5 and 6 are one stage in the sense of section 1.23)

The stage matrix is

\[
\begin{pmatrix}
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

and the lost times following the four stages are 8, 6, 3 and 3 seconds respectively.

In applying Webster's method to this example the usual definition of representative approach breaks down and Webster and Cobbe make two alternative choices, leading to cycle times of 79 and 87 seconds; they prefer the larger cycle time.

Two calculations have been made by the present method for the same average arrival rates as given by Webster and Cobbe. In the first calculation, no cycle time was specified and the resulting calculated cycle time was 81.4 seconds, giving an estimated rate of delay of 25.7 p.c.u. 's. In the second, Webster and Cobbe's cycle time of 87 seconds was specified, and the resulting estimated rate of delay was
26.0 seconds. In each case the allocation of green time was very close to that given by Webster's method.

Two points considered in sections 3.6 and 3.54 can be illustrated by this example. In section 3.6 the possibility of relaxing the convergence criterion in the computer program was mentioned. In the two calculations in this example the criterion was satisfied after 22 and 13 iterations; the components of $\ldots$ had not changed in the third significant figure after the 16th and 10th iterations respectively. There may therefore be scope for relaxing the criterion slightly, but not greatly.

In section 3.54 the possible effect of the third term in Webster's expression for delay was considered. In the first calculation in this example, since there are no minimum green constraints and no cycle time constraint except $\lambda_0 > 0$, the components of $VD$ must be equal when $\ldots$ is optimal. When the calculation stopped after 22 iterations the components of $VD$ all lay between $-102.39$ and $-102.28$, and the components of $VC$ lay between $7.78$ and $19.39$, a range of $11.61$. The last iteration at which the greatest component of $VD$ exceeded the least by more than 11 was the 12th, and at this iteration the value of every component of $\ldots$ was within 2 per cent of the value accepted as optimal at the 22nd iteration. This suggests that the value of $\ldots$ at which all the components of $VD$ are equal is unlikely to differ much from that at which all the components of $VD$ are equal; i.e. the inclusion of the third term in Webster's expression for delay would probably have little
effect on the resulting signal settings in this case.

3.72. Comments. These few results suggest that Webster's method will give a usually good and often very good approximation to the settings that minimise the estimated delay. The settings given by the present method do sometimes differ appreciably from Webster's approximations, particularly in that the cycle time is often shorter. Although the difference in estimated delay is usually small, the incidence of exceptional delay occurring, for example, when the intersection is partially blocked by an unusual accumulation of turning vehicles, is likely to be reduced by the shortening of the cycle, because such delay is usually confined to one cycle. Webster's expression for delay, like the other expressions discussed in Chapter 2, takes no account of such exceptional delay. The reduction in cycle time may well also be most useful when signals are linked (see section 5.1). The circumstances in which the \( \text{Webster} \) methods give appreciably different settings are not yet clear; to clarify them, more extensive comparisons and practical trials are required. Only one limited practical trial has so far been reported (Taylor and Allsop, 1969). The present method is just as directly applicable to complicated intersection layouts as to simple ones.

3.8. Effect of changing a specified cycle time

It will be seen in the next two chapters that when signals at neighbouring intersections are linked, they usually have a common cycle time. The methods of this chapter can be
applied to find the best allocation of green time at each intersection when the common cycle time has been chosen. The choice of the common cycle time is discussed in section 5.1, and it will then be useful to know how sensitive is the rate of delay at each individual intersection to the choice of the common cycle time. This sensitivity is not difficult to calculate from the preceding results, and, for this purpose, it is convenient to take a slightly different mathematical view of the problem, restricting attention to the case in which the cycle time is specified.

By the methods of this chapter, a value, $\lambda^*$, say, of $\lambda$ has been found that minimises $D(\lambda)$ subject to the constraint: $S(\lambda) = 1$, the capacity constraints, the minimum green constraints, and the cycle time constraint, which now takes the form $\lambda_0 = k_0$. Given the solution $\lambda^*$ it is possible to identify the stages, if any, for which the green times take the specified minimum values. Suppose that there are $p$ such stages; then, since not all the green times can be minimal, $0 < p < m$. If there are such stages, let them have numbers $\pi_i$ ($i = 1, 2, \ldots, p$), where $\pi$ is a suitable permutation of the integers $1, 2, 3, \ldots, m$. Then $\lambda^*$ can also be found by minimising $D(\lambda)$ in the region defined by the capacity constraints and the minimum green constraints for the other stages $\pi_i$ ($i = p+1, p+2, \ldots, m$), subject to the following equalities:

$$S(\lambda) = 1$$
$$\lambda_0 = k_0$$
$$\lambda_{\pi_i} = k_{\pi_i} \lambda_0 \quad (i = 1, 2, \ldots, p).$$
This is a problem that can be solved by means of Lagrange's multipliers. Let \( \mu_S, \mu_0 \) and \( \mu_i \) \((i = 1, 2, \ldots, p)\) respectively be the Lagrange multipliers associated with the above equalities. There are no multipliers associated with the other constraints because \( \lambda^* \) is known not to lie on the boundaries defined by them.

Then by considering the partial derivatives at \( \lambda^* \) of

\[
D(\lambda) + \mu_S(S(\lambda) - 1) + \mu_i(\lambda_i - k_i) + \xi_i \mu_i(\lambda_{m_i} - k_{m_i} \lambda_0)
\]

it follows that

\[
\frac{\partial D}{\partial \lambda_0} + \mu_S + \mu_0 - \xi \sum_{i=1}^{p} k_{m_i} \mu_i = 0,
\]

\[
\frac{\partial D}{\partial \lambda_{m_i}} + \mu_i + \mu_i = 0 \quad (i = 1, 2, \ldots, p)
\]

and \( \frac{\partial D}{\partial \lambda_{m_i}} + \mu_S = 0 \quad (i = p + 1, p + 2, \ldots, m) \).

The last \( m - p \) equations are consistent because, if \( p + 1 \leq k, \ell \leq m \), then, since \( \lambda_{\pi k} > k_{\pi k} \lambda_0 \), \( \frac{\partial}{\partial \lambda_{\pi k}} - \frac{\partial}{\partial \lambda_{\pi \ell}} \) are both permissible directions at \( \lambda^* \). Hence if \( \frac{\partial D}{\partial \lambda_{\pi k}} \) were not equal to \( \frac{\partial D}{\partial \lambda_{\pi \ell}} \), \( \lambda^* \) would not be optimal (cf. Lemma 9). Thus any one of these \( m - p \) equations determines \( \mu_S \); the remaining \( p + 1 \) multipliers are then determined by the first \( p + 1 \) equations (which are independent). In particular, \( \mu_0 \) is so determined. If \( p = 0 \), \( \mu_S \) is determined by any of the \( m \) equations

\[
\frac{\partial D}{\partial \lambda_i} + \mu_S = 0 \quad (i = 1, 2, \ldots, m)
\]

and then \( \mu_0 \) by the equation

\[
\frac{\partial D}{\partial \lambda_0} + \mu_S + \mu_0 = 0
\]
Now by the theory of Lagrange's multipliers (see, e.g. Hadley 1964),

$$\mu_0 = -\frac{\partial D(\tilde{\lambda}^*)}{\partial k_0}$$

where $\partial/\partial k_0$ indicates rate of change with respect to $k_0$, keeping $k_1, k_2, \ldots, k_m$ fixed, when the optimal solution $\tilde{\lambda}^*$ is regarded as a function of $k_0, k_1, k_2, \ldots, k_m$. Also $k_0 = L/\tau$, where $\tau$ is the specified cycle time and $L$ is the total lost time at the intersection being considered.

Hence

$$\frac{\partial D(\tilde{\lambda}^*)}{\partial \tau} = \frac{L}{c^2} \frac{\partial D(\tilde{\lambda}^*)}{\partial k_0} = \frac{L\mu_0}{c^2}$$

Thus $L\mu_0/c^2$ is the required sensitivity, to the choice of specified cycle time, of the rate of delay associated with the corresponding allocation of green time. This sensitivity can be calculated, for each individual intersection in a linked system, for any value of the common cycle time. The application of this result to the choice of the common cycle time for linked signals will be discussed in section 5.1.
4. **THEORETICAL STUDIES OF LINKED TRAFFIC SIGNALS**

The first systems of linked signals to be studied were those in which signals at a number of intersections along one main road were linked, with the principal aim of reducing delay to traffic on the main road; the methods used vary in the extent to which they take account of the needs of side-road traffic. Signal settings for such systems and the movement of traffic through them can be represented in a single diagram called the time-distance diagram. This diagram has long formed the basis of graphical methods of devising good signal settings, and is helpful in theoretical studies; it will therefore be described first in this chapter. A survey will then be made of the theory of various methods of calculating settings for linked signals, beginning with systems in which all the signals are at intersections on one main road, and going on to systems covering networks of intersecting roads.

4.1. **The time-distance diagram.**

4.11. **The basic diagram.** In the time-distance diagram for a system of traffic signals at intersections along a main road, time is the abscissa and distance along the main road is the ordinate. Suppose, for example, that traffic travels northwards along the main road, and that distance is measured northwards from the stopline for northbound traffic at the southernmost intersection. Figure 6 shows a time-distance diagram for a road having three signalised intersections $I_1$, $I_2$ and $I_3$. The positions of the stoplines for northbound traffic at these intersections are represented by horizontal lines, the first of which is the time axis. If traffic also
travels southwards along the main road, these same horizontal lines can represent the positions of the stoplines for southbound traffic, provided that all cross-streets are assumed to have the same width, and the distance scale for southbound traffic is displaced by this width in the negative direction relative to that for northbound traffic.

Provided that the effective green times for northbound and southbound traffic at each intersection coincide, the allocations of green time can be represented by dividing the corresponding horizontal lines into alternate heavily and lightly ruled segments representing the effective red and green times respectively. If the time origin is taken at the start of a master cycle and the starting point of the signal cycle at each intersection is taken as the beginning of the effective green time for the main road, then the offset at an intersection (as defined in section 1.25) is represented by the length of the corresponding horizontal line from the distance axis to the first starting point of a lightly ruled segment (i.e. in Figure 6 the three offsets are represented by $A_1B_1$, $A_2B_2$, and $A_3B_3$). This completes the representation of the signal settings.

The movement of vehicles on the main road is represented in the diagram by trajectories whose gradients represent the speeds of the vehicles. In Figure 6, for example, $C_1D_1$ represents the movement of an undelayed northbound vehicle that travels faster between $I_1$ and $I_2$ than between $I_2$ and $I_3$; $E_1F_1$ represents the movement of a southbound vehicle that is delayed at all three signals and travels at
Fig. 6. AN EXAMPLE OF A TIME-DISTANCE DIAGRAM
the same speed between $I_3$ and $I_2$ as between $I_2$ and $I_1$.

It is often useful to simplify the trajectories so that they consist of straight line segments. This is done by supposing that each vehicle travels at a constant speed on each link - not necessarily the same speed on different links. Acceleration and deceleration are assumed to occur instantaneously at the stoplines, and the standard departure model described in section 2.25 is assumed to apply at each signal. Delay at an intersection is then represented by a horizontal segment in the trajectory. $C_2D_2$ and $E_2F_2$ in Figure 6 are simplified trajectories corresponding to $C_1D_1$ and $E_1F_1$ but displaced by one cycle.

4.12. Transformations of the diagram. Three transformations have been used further to simplify the time-distance diagram. Firstly, the distance scale between each pair of adjacent intersections can be made inversely proportional to the design speed for vehicles travelling in the positive direction between these intersections. Then the trajectory for an undelayed vehicle that travels in this direction at the design speed on each link becomes a single straight line. The trajectory for an undelayed vehicle travelling in the opposite direction at the design speed on each link also becomes a single straight line under this transformation if and only if between each pair of adjacent intersections the design speeds in the two directions are equal. If $t$ denotes time and $y$ denotes the ordinate in this transformed diagram, then the trajectory of an undelayed vehicle travelling in the positive direction at the design speed on each link has an equation of the form $y = a + vt$. 
In a second transformation, particularly useful in the study of a one-way street (Bavarez and Newell 1967), the abscissa \( t \) in the diagram just obtained is replaced by \( t - (y/v) \), so that the trajectories of vehicles moving in the positive direction at the design speed become vertical lines. After this transformation, however, offsets must be measured from the line through the origin with gradient \(-v\), instead of from the distance axis.

A third transformation can be used to equalise, between each pair of adjacent intersections, the gradients of the trajectories of vehicles travelling at the design speeds in the two directions. Newell (1969) showed that this can be achieved by shifting the time scale at each intersection backwards or forwards by an appropriate amount. After this transformation, offsets must be measured from the shifted time-origins instead of from the distance axis.

4.13. An invariance property. Yardeni (1965) and Brooks (1965) each observed that the time-distance diagram remains unchanged, except in scale, if the cycle time is multiplied by some constant and all speeds are divided by the same constant, provided that the offsets (measured in fractions of a cycle) and allocations of green time are not altered. This means that a diagram devised for one cycle time holds good for other cycle times with appropriately adjusted vehicle speeds.

4.14. Through bands and bandwidth. When a number of intersections on a main road are controlled by linked signals, the
signal schedule may allow a vehicle crossing the first stopline at a suitable time and travelling at the design speed on each link to pass through all the intersections without having to stop at any signal. If so, the area of the time-distance diagram covered by the trajectories of vehicles travelling along the main road in this way in either direction is called the through band for that direction. For example, Figure 7 shows that, for the signal system considered in Figure 6, if the design speeds are represented by the gradients of the simplified trajectories in Figure 6, then a through band exists in each direction. These through bands are the shaded areas in Figure 7; they are repeated once in each cycle. The width of a through band, measured parallel to the time axis, is called the bandwidth; in Figure 7 the bandwidths for the positive and negative directions are represented by $X_1Y_1$ and $X_2Y_2$ respectively. It is theoretically possible for the through band in a given direction to be split into more than one portion appearing at different times in the cycle. This possibility is, however, remote, and would result in very narrow bands; it is therefore usually neglected.

Although no widely applicable precise relationship between bandwidth and delay has been derived, it is reasonable to suppose that a signal schedule having wide through bands will result in a freer flow of traffic than one having either no through bands or only narrow ones. Graphical methods of devising signal schedules have therefore usually aimed to produce wide through bands, and some quite systematic procedures have been reported (e.g. Davidson 1960, Kaus 1963). The bandwidth criterion is also the basis of an important
Fig. 7. A TIME-DISTANCE DIAGRAM SHOWING THROUGH BANDS
method of calculating signal settings (Morgan and Little 1964), which is discussed in section 4.22.

4.2. Settings for linked signals on one main road.

In calculating settings for signals in a linked system, the engineer has at his disposal the master cycle and the offsets and allocations of green time at the individual signals. Sometimes also the design speed is at his disposal to the extent that, for each link, a range of speeds is specified, within which any design speed is acceptable. Theoretical studies of the calculation of settings vary as to the parameters they seek to determine, and those that they regard as specified in advance. Studies concerned specifically with the calculation of settings for a system of signals on a single one-way or two-way road will be discussed in the next two sections. In addition, the more general methods, discussed in section 4.3, of calculating settings for signals on a network of roads are also applicable to a system of signals on a single main road.

4.21. A one-way road. In an early theoretical study, Newell (1960) considered a very long sequence of equally spaced signals, at distance $D$ apart, each having the same cycle time $T$ and the same proportion $\lambda$, say, of the cycle effectively green for the main road. Assuming that the offset on each link was $\delta$, and that traffic was so light that any vehicle stopped at a signal would depart immediately the effective red time ended, he considered the delay to a vehicle that travelled at speed $v$ except when stopped at a signal.
He expressed the average delay per signal for such a vehicle as a complicated function of \( v \) and \( \delta \). Regarding \( v \) as a random variable with distribution function \( F(v) \), mean \( \bar{v} \), say, and a small coefficient of variation, he investigated the choice of \( \delta \) that would minimise the average delay to all vehicles. Taking \( D/\bar{v} \) as a first approximation to the required value of \( \delta \), and assuming that the signals are close enough together for \( (D/v) - \delta \) to be a small fraction of \( T \), he examined the average delay, as a function of \( \delta \), near to \( \delta = D/\bar{v} \), and showed that a second approximation to the required value of \( \delta \) is given by

\[
F\left(\frac{D}{\delta}\right) = \lambda
\]

Thus, to minimise delay when traffic is very light the offset \( \delta \) should be chosen so that a proportion \( \lambda \) of vehicles have undelayed speeds less than \( D/\delta \).

The same situation was analysed differently by Haight (1963). He obtained, in terms of the distribution of undelayed speeds \( v \), the proportion \( \pi(j,k) \) of vehicles which, when released at the start of effective green at the first signal, would next be stopped at the \( (j+1) \)th signal in the cycle beginning there \( j\delta + (k-1)T \) later. He then observed that the average speed of such a vehicle along the whole road is its average speed between being released by the first and the \( (j+1) \)th signals - i.e. \( j\bar{v}/(j\delta + kT) \). The average speed for all vehicles is therefore

\[
\sum_{j=1}^{\infty} \sum_{k=0}^{\infty} jD\pi(j,k)/(j\delta + kT)
\]
and Haight suggested choosing $\delta$ and $T$ to maximise the average speed, given $D$ and $F(v)$, but did not pursue the analysis.

A similar sequence of signals was considered again by Newell (1964), this time with a greater flow of traffic. He assumed a perfectly cyclic flow pattern in which the same number of vehicles passed each signal in each cycle, the platoon released by one signal was still accelerating when it reached the next and had not caught up with the preceding platoon, and the flow was high enough for the arrivals at each signal to be spread over a period longer than the effective green time. He showed that, under these assumptions, the delay incurred on any link depends only on the offset on that link. By examining the resulting delay-offset relations and observing that the arrival headways in the accelerating platoon form a decreasing sequence whose sum is less than the cycle time, he showed that delay is minimised by choosing the offset so that the last vehicle in the platoon arriving at any signal is able to cross the stopline just before the effective green time ends.

A more practical method of choosing the offset on a link of a one-way road is proposed by Gazis (Horton 1965). He suggests that, if the offset cannot be chosen so as to prevent a queue forming at the downstream signal, it should be chosen so that the leading vehicle of each platoon passing along the link does not catch up with the last queueing vehicle of the preceding platoon before this vehicle has begun to move forward. Such a choice of offset avoids a platoon being stopped partway along the link.
The analyses discussed so far in this section have considered only the traffic on the main road. A much fuller analysis of a system of signals on a one-way road crossed by n side streets, taking into account the side street traffic, was made by Bavarez and Newell (1967). They regarded the traffic as a continuous fluid arriving uniformly at a rate $q$ at the first intersection on the main road and at a rate $q_j$ on the jth side street ($j = 1, 2, \ldots, n$). They noted that a lower bound for the common cycle time $T$ is provided by the largest of the Clayton minimum cycle times (see section 2.72) for the individual intersections. Considering first the determination of offsets for given cycle time and allocations of green time, they observed that the choice of offsets does not affect the delay on the side streets. They called the intersection having the shortest main road effective green time the critical intersection and showed that, with the given cycle time and allocations of green time, the total delay and number of stops for all traffic are minimised by choosing the offsets so that main road traffic is undelayed after leaving the critical intersection and leaves that intersection at the same time as it would have done if the preceding signals had always been green. They showed how to make such a choice of offsets, and then showed that by varying the allocations of green time, and suitably readjusting the offsets, the delay and number of stops on the side streets can be further reduced without changing the delay and number of stops on the main road. This is done by decreasing the main road effective green times at other intersections to the value at the critical intersection. It remains to determine the resulting
allocation of green time, and the common cycle time, and this they showed to be equivalent to the same problem for a single intersection with \( n + 1 \) approaches controlled by a signal with a two-stage cycle: one approach with regular arrivals at rate \( q \) has right of way in stage 1, and the other \( n \) approaches with regular arrivals at rates \( q_1, q_2, \ldots, q_n \) in stage 2. The number of stops is minimised by choosing the shortest cycle time consistent with capacity requirements, but, as shown by Wardrop (1952) (see section 2.72) such a choice will not in general minimise total delay.

Bavarez and Newell went on to show that in certain circumstances the side street delay at a particular intersection can be further reduced, without increasing main road delay, by allowing the signal at that intersection to operate with a cycle time \( T/m \), where \( m \) is an integer. If this is done, the extra lost time must be subtracted entirely from the side street green time, so that an upper bound on \( m \) is usually imposed by the capacity requirement of the side street. They showed that, provided that this upper bound is not exceeded, side street delay decreases as \( m \) increases so long as \( m(m + 1) < (g/L)^2 \), where \( g \) and \( L \) are respectively the original main street effective green time and the lost time per cycle at the intersection concerned. The possibility of using non-integer values of \( m \) was investigated and found unlikely to reduce delay. The study concluded with a brief discussion of the possibility of dividing the intersections into two groups, one working with the common cycle time previously determined, and the other
with a shorter cycle time; there would be extra main road
delay at the changeover point, but in suitable circumstances
this would be outweighed by the saving in delay on side
streets where the cycle time was shortened.

4.22. A two-way road. Most studies of the calculation of
settings for linked signals on a two-way road have been
directed to the calculation of settings that maximise band-
width (see section 4.14). Before considering such studies,
however, two analyses by Newell (1964 and 1969) of the
relation between signal settings and delay will be discussed.

The assumptions made about traffic behaviour and the
signal system in the first of these analyses (Newell 1964)
were set out in section 4.21. Under these assumptions, Newell
obtained the delay-offset relation for traffic travelling in
one direction between two adjacent signals. He went on to
show how, if the road is two-way, with the traffic in the two
directions behaving independently according to the same
assumptions, the total rate of delay to traffic on the links
in the two directions between adjacent signals can be expres-
sed as a function of the offset on one of the links. This
is done by taking the delay-offset relation for that link
and adding to it the relation for the other link with its
time-scale reversed. The results showed that, for given
flows in the two directions, the offset that gives the least
total delay can be either of the offsets giving least delay
on the separate links; which it is depends on the difference
between the times taken to travel between the two signals by
the last vehicles of the platoons travelling in the two
directions. This conclusion was, however, very sensitive to
the rather strict assumptions made about traffic behaviour.

In the second analysis (1969), Newell considered high flows of traffic through a sequence of signals such that the cycle time and the proportion of the cycle effectively green for the main road was the same at each signal. He assumed that the saturation flow on the main road was $s$ in one direction and $s'$ in the other direction at each signal, that traffic was so heavy that departures in the two directions continued at rates $s$ and $s'$ throughout the green time in every cycle, that all vehicles on any link travelled at the same speed, and that there was no turning traffic. He noted that, under these assumptions, the delay and number of stops on each link depend only on the offset on that link, so that the analysis could be confined to traffic travelling between two adjacent intersections. He then obtained the total delay and number of stops per cycle for such traffic as piecewise linear functions of the offset on one of the two links between the intersections. The form of these functions shows that if $s \neq s'$, the delay and number of stops are both minimised by making the offset on the link with the higher saturation flow equal to the travel time on that link, but that if $s = s'$ there can be a range of offsets (depending on the travel times) that minimise both the delay and the number of stops; in the latter case, other criteria could be taken into account in choosing an offset in this range.

It has been known for some time (see, e.g., Matson, Smith and Hurd, 1954) that if a sequence of signals on a main road has a common cycle time and main road effective green
time, and if, between each pair of adjacent signals, the design speed is the same in both directions, then the bandwidth can be equal to the effective green time in both directions only if, at each intersection, the offset is an integer multiple of half the cycle time. This property follows from the geometry of the time-distance diagram, and is the basis of two frequently used simple choices of offsets for such systems of signals: the so-called simultaneous system and alternate system. In the former, all the offsets are zero, and in the latter they are alternately zero and half the cycle time.

A deeper analysis of the geometry of the time-distance diagram with a view to choosing offsets to maximise bandwidth was made by Morgan and Little (1964). They considered a sequence of \( n \) signals at intersections \( I_1, I_2, \ldots, I_n \) on a main road; the signals had a given common cycle time \( T \), and given allocations of green time (not necessarily the same at each signal). The design speed for each link was also regarded as given, and need not be the same on each link; in particular it need not be the same in the two directions between two adjacent signals. They showed that the offsets can be chosen so that both the sum of the bandwidths in the two directions is maximised and the two bandwidths are equal. This is achieved by choosing the offset \( z_{jj+1} \) on the link from \( I_j \) to \( I_{j+1} \) so that

\[
z_{jj+1} = \frac{1}{2} \left( t_{jj+1} - t_{j+1,j} + m_j T \right) \quad \pmod{T},
\]

where \( t_{jk} \) is the travel time from \( I_j \) to \( I_k \) at the design speed, \( m_j = 0 \) or \( 1 \) (\( j = 1, 2, \ldots, n - 1 \)) and the starting point of each signal cycle is taken at the midpoint of the
main road effective red time. To maximise the equal bandwidths, therefore, it is necessary only to examine the values resulting from the $2^{n-1}$ different combinations of values of $m_j$. The terms $\frac{1}{2}(t_{jj+1} - t_{j+1j})$ correspond to the shifting of the time-scales in Newell's (1969) transformation of the time-distance diagram (see section 4.12). Morgan and Little also show how to adjust the offsets so that the same maximum total bandwidth is shared in any desired proportion between the two directions, provided only that the larger bandwidth may not exceed the least of the main road effective green times. A computer program for all these calculations is given by Little, Martin and Morgan (1966).

Essentially the same method of choosing offsets was used by Yardeni (1964), Brooks (1965) and Koshi (1966), but these authors also considered how to choose the common cycle time. Brooks and Yardeni assumed that the design speed in either direction was the same all along the main road (though the speeds in the two directions might not be the same). They noted that, by the invariance property discussed in section 4.13, the ratio of the bandwidth to the cycle time depends only on the product of the cycle time and design speed. Brooks suggested trying a range of acceptable values of this product and choosing the one that gave the largest ratio of maximum bandwidth to cycle time. Yardeni sought to choose the product by a least squares process so that the travel times between intersections at the design speeds were as nearly as possible integer multiples of the cycle time. Koshi allowed any set of design speeds to be specified, one for each link, and then set out to choose the cycle time on
the assumption that, at each intersection, the main road effective green time was a fixed proportion of the cycle time, whatever the latter. With this assumption he showed that the ratio of the maximum bandwidth (found by Morgan and Little's method) to the cycle time is a piecewise linear function of the reciprocal of the cycle time. Thus, if a range of acceptable cycle times is given, it is not difficult to find for which of these cycle times the ratio of maximum bandwidth to cycle time is greatest.

Little (1966) re-examined the problem solved by himself and Morgan (1964), and restated it as a linear programming problem with $3n$ variables. The first $2n+1$ of these were the bandwidth and the lengths of effective green time to the left of the positive through band and to the right of the negative through band at each intersection. The other $n-1$ variables were constrained to be integers, one associated with each pair of adjacent intersections. The restrictions imposed by the geometry of the time-distance diagram were expressed as linear inequality constraints on the $3n$ variables, so that for any particular values of the integer variables, the bandwidth could be maximised by the simplex method of linear programming (see, e.g. Vajda 1961). Little showed how to find a range within which each integer variable must lie, and devised a branch and bound algorithm to find the set of values of the integer variables that gives the largest maximal bandwidth. Whilst this approach offered no advantages over the earlier one for the problem previously solved by Morgan and Little, Little was able to extend the new method to include the reciprocals of the cycle time and
design speeds as further variables, each with a specified acceptable range. By maximising the bandwidth with respect to all these variables, he was able to obtain, in a realistic example, a ratio of bandwidth to cycle time 20 per cent greater than that obtained by the earlier method.

4.3. Settings for linked signals in a network of roads.
4.3.1. Network geometry. The difficulty of calculating settings for linked signals in a network of roads depends greatly on the layout of the network. The representation of the layout by means of a graph was described in section 1.25, and some examples are shown in Figure 8. Before discussing these examples, it will be useful to define two simple ways in which arcs in the graph (and corresponding links in the network) may be related.

DEFINITIONS. If $A_i$, $A_j$, and $A_k$ are three distinct vertices of the graph and $A_j$ is the endpoint of just two arcs, joining it to $A_i$ and $A_k$ respectively, then these two arcs are said to be in series. The two corresponding links in the network are also said to be in series. If two or more arcs in the graph have the same endpoints, then these arcs are said to be in parallel; the corresponding links in the network are also said to be in parallel.

This terminology was used extensively by Whiting (Hillier 1965), and the idea of links in parallel had already been used by Newell (1964) in a study of a two-way road.

Examples (a) and (b) in Figure 8 show graphs representing
Fig. 8. EXAMPLES OF GRAPHS REPRESENTING ROAD NETWORKS

a) A one-way road
b) A two-way road
c) A tree
d) A ladder
e) A tree of ladders
f) A more complicated network
a one-way and a two-way road, each with three signalised intersections. For many purposes, arcs in parallel between two vertices in a graph can be replaced by a single arc between the same two vertices, and this has been done, for simplicity, in the remaining examples (c)-(f) in Figure 8. In these graphs, therefore, each arc may, for the present, either represent a one-way road or be the result of replacing two or more arcs representing a two-way road; for this reason, directional arrows have been omitted from these graphs. It should be remembered that the graphs represent only the existence of intersections and certain traffic flows between them in the network; the fact that the graphs are drawn as regular rectangular patterns does not imply any similar regularity in the road layout.

Example (c) represents a network in which there are several intersecting roads, but the roads do not form any closed loops; the resulting graph is a tree (Berge 1962). Example (d) represents two main roads, both crossed by a number of other roads; the resulting graph will be described as a ladder, the arcs representing the two main roads will be called side arcs, and those representing the crossing roads at the two ends of the network will be called end arcs. Example (e) represents a network of roads whose graph is a tree of ladders; this is a graph obtained from one ladder by adding further ladders one at a time. At each stage a side arc of one of the ladders already in the graph is used as an end arc of the newly added ladder, or vice versa, provided that no arc may be a side arc of more than one ladder, no two side arcs of different ladders may be in parallel, and the
ladders must not form any closed loops. Example (f) represents a network in which three main roads are crossed by a number of other roads; the resulting graph is more complicated than a tree of ladders because several of the arcs representing the central main road form the side arcs of two ladders.

Various methods of calculating signal settings will be discussed in the next four sections, and the examples in Figure 8 will be useful in distinguishing the types of network to which these methods apply.

4.32. Some studies relating to special cases. Methods applicable to one-way and two-way roads have already been discussed in sections 4.21 and 4.22, and Gazis points out (Horton 1965) that such methods can often be applied to networks whose graphs become trees when arcs in parallel are replaced by single arcs (see section 4.31). It is sufficient to add here that, in particular, the methods of Newell (1964 and 1969) and Bavarez and Newell (1967) can be applied directly to networks whose graphs are trees, provided that the traffic on all the roads satisfies the assumptions made by the respective methods, and that turning traffic can be neglected.

Little (1966) showed how his linear programming method of maximising bandwidth on a two-way road could be extended to apply in a limited way to networks. A number of main roads through the network could be chosen, and any linear combination of the bandwidths on these roads could be maximised by varying the cycle time, the design speeds, and the allocations of green time at intersections of the chosen roads (each within
a given acceptable range), and also the offsets. In addition to the constraints present in the case of a two-way road, there would be, for each closed loop formed by the chosen roads, an equality constraint on the offsets involving a new integer variable. Although Little showed the method to be feasible for a small network, it has since been superseded by methods that seek to minimise delay (see sections 4.33 - 4.35) instead of using the bandwidth criterion. An earlier preliminary exploration of the applicability of linear programming methods was made by Lavallee (1956), but he concluded that the necessary computation would not be feasible. Stoffers (1968) developed a linear programming method for calculating delay-minimising settings for signals at one intersection in a network when the settings at all adjacent intersections are already specified, but this problem seems unlikely to occur often in practice.

Another rather specialised study is that of Buckley et al (1966), who considered a large rectangular grid network consisting of equally spaced north-south roads, alternately one-way northbound and southbound, and equally spaced east-west roads, alternately one-way eastbound and westbound. Travel times were assumed to be \( \tau \), say, on each east-west link and \( k\tau \) (\( k \geq 1 \)) on each north-south link. It was first noted that all circumnavigable rectangles have an odd number of links in each side. Buckley and his co-authors then showed that if the lost time was the same at each intersection, then a sufficient condition for the travel times round all circumnavigable rectangles to be a multiple of the master cycle \( T \) was that the travel times round a 1 link by 1 link
rectangle and a 1 link by 3 link rectangle should each be multiples of $T$. They showed further that if $k = a/b$ in its lowest terms ($a$ and $b$ being integers), then the largest value of $T$ satisfying the above sufficient condition is given by

$$T = \begin{cases} 
4\pi/b & \text{if } a \text{ and } b \text{ are both odd} \\
2\pi/b & \text{if either } a \text{ or } b \text{ is even.}
\end{cases}$$

Though this study is of theoretical interest, its practical value seems limited because it takes no account of traffic passing straight through the network in the direction of either of the sets of roads, and offsets could not be chosen to give free movement both to such traffic and to traffic travelling round rectangles.

4.33. More general studies. More general methods of calculating settings for linked signals in networks of roads (Hillier 1965, Traffic Research Corporation 1966, Chang 1967, Yumoto et al 1967, Robertson 1969a and b, Longley 1969, Martin-Löf 1969) each proceed in two stages. The first stage is to make a mathematical estimate of the rate of delay and the average number of p.c.u.'s stopping per unit time in all or part of the network, as a function of the signal settings. In most of the methods, any desired linear combination of these two quantities can be estimated and used as a criterion for choosing signal settings. The following definition (cf Robertson 1969a) will be useful.

**DEFINITION:** A **performance index** is a linear combination of estimated rate of delay and estimated average number of stops
per unit time, used as a criterion for choosing signal settings.

The estimated rate of delay by itself is a particular example
of a performance index, and has to be used in methods that
make no estimate of the number of stops per unit time.

The second stage is to choose signal settings to minimise
the performance index. All the methods make simplifying
assumptions at the first stage, and most of them use approx-
imate methods at the second. The two stages will be discussed
separately in the next two sections.

4.34. Calculation of the performance index. In each method,
estimates of the rate of delay and (except in the methods of
Chang, Yumoto and Longley) the number of stops per unit time,
and hence the performance index, are obtained from a mathema-
tical model of the movement of traffic in the network. The
mathematical models used have a number of features in common.
All assume a completely cyclic pattern of traffic flow through-
out the network, with period equal to the duration, T, of
the master cycle, which is assumed to be determined in advance;
only Robertson (1969a and b) allows, in his estimate of the
rate of delay, for random fluctuations about the cyclic
pattern. All except Chang and Longley assume regular arrivals
(see section 2.21) on those approaches where traffic enters
the network. Chang (1967) allows arrivals on each such approach
to be spread uniformly over any specified fraction of the
cycle, whilst Longley (1969) allows an arbitrary cyclic arrival
pattern to be specified on each such approach. Flows within
the network depend on the signal settings, but in every case
the amount of traffic per cycle entering each link from each
of the approaches at the upstream signal has to be specified, together with the saturation flow on each approach. All except Chang, Robertson and Longley assume that the rate of delay and number of stops per unit time on each link of the network depend only on the offset on that link, and that the allocations of green time have been determined in advance. Robertson (1969a and b) and Longley (1969), however, allow for the effect throughout the network of changes in individual offsets and allocations of green time. Chang (1967) makes some allowance for such effects, but his ability to do so is limited by the fact that, like the Traffic Research Corporation (1966), he neglects the effect of signal settings on traffic entering each link from all but one of the approaches at the upstream signal; only the effect on traffic from the approach contributing the largest flow is taken into account. This usually means that turning traffic is neglected; all the other methods take such traffic into account. All the methods except Chang's allow for the spreading of vehicle platoons as they travel along the links of the network, although Hillier's did not do so until it was modified by Huddart and Turner (1969).

Some of the methods have not been reported in full detail, and others are as yet untried, but the main principles of each can be set out. Following Robertson (1969a) three so-called traffic patterns will be used to describe the flow of traffic on a link. Each pattern expresses a flow of traffic at the downstream stopline as a function of time; since the flow is assumed to be cyclic, the functions need only be defined in the interval \((0, T)\), with the time-origin at the start of the master cycle.
DEFINITIONS: The **in-pattern** for a link is the rate at which it is assumed that traffic released by the upstream signal would arrive at the downstream stopline if the downstream signal were always green.

The **go-pattern** is the rate at which it is assumed that traffic would leave the downstream stopline if there were always a queue there.

The **out-pattern** is the rate at which it is assumed that traffic actually leaves the downstream stopline.

With the departure model of section 2.25, which is assumed in all the methods, the go-pattern is zero during the effective red time and equal to the saturation flow on the downstream approach during the effective green time. All the methods assume that, when there is no queue during the effective green time, the traffic leaves at the rate at which it arrives. With this assumption, the in-pattern and the go-pattern together determine the out-pattern and also the queue-length. It is assumed (except by Chang - see below) that the queue accumulates at the stopline and that its length is the difference between the integrals of the in-pattern and the out-pattern; since the flow is assumed to be cyclic, the degree of saturation (see section 1.24) on each approach must be less than 1, and the queue-length must be zero at the end of the effective green time. In all the methods, the delay per cycle on the link is taken as the integral of the queue-length over the cycle, and the number of stops is taken as the number of p.c.u.'s that are assumed to reach the stopline while there is a queue there.
An example of the three patterns and the corresponding graph of queue-length as a function of time is given in Figure 9, in which areas representing the delay and number of stops per cycle are shaded. Any such diagram holds good only for one signal schedule, since the go-pattern depends on the offset and allocation of green time at the downstream signal, and other signal settings influence the in-pattern. The methods differ in the extent to which the in-pattern takes account of the effect of the settings at other signals.

Hillier's method (1965), whose mathematical aspects are due largely to Whiting, will be described in some detail because it forms the starting point of the present author's work (see Chapter 5). It requires that the estimated rate of delay and number of stops per unit time on a link should depend only on the offset on that link, i.e. on the difference between the offsets at the upstream and downstream signals. This requirement is satisfied if the shape of the in-pattern is independent of the offsets and its position on the time axis is fixed by the upstream offset, since the position of the go-pattern on the time axis is fixed by the downstream offset and, for given shapes of in-pattern and go-pattern, the estimated rate of delay and number of stops per unit time depend only on the relative position of the two patterns on the time axis. To obtain an in-pattern satisfying these conditions and approximating reasonably well to actual traffic behaviour, at least at high flows, Whiting considered the approaches at the upstream signal from which traffic entered the link. He assumed that if the average flow on such an approach was \( q \) p.c.u./second, the saturation flow
S = Saturation flow on the downstream approach in p.c.u./second

These two shaded areas are equal and represent the estimated number of stops per cycle.

Effective green time on downstream approach.

This shaded area represents the estimated delay per cycle.

Fig. 9. EXAMPLE OF THE DESCRIPTION OF THE TRAFFIC FLOW ON A LINK OF THE NETWORK
s p.c.u./second, and the proportion of traffic from this approach that entered the link in question was p, then this traffic would enter the link in a platoon at rate ps for the first qT/s seconds of each effective green time. An alternative assumption (Huddart and Turner 1969) is that this traffic enters the link at rate ps for the first part, and at rate pq for the remainder, of the effective green time on the approach from which it comes (as would result from regular arrivals on this approach); the rate changes at such a time that the amount of such traffic entering per cycle has the required value pqT. Whiting assumed further that the platoons so formed would travel along the link at the design speed without spreading, thus determining the required in-pattern at the downstream signal.

With these assumptions, and with allocations of green time determined in advance, the rate of delay on each link depends, as required, only on the offset on that link. A delay-offset relation, as defined in section 1.25, is thus obtained for each link. Huddart and Turner (1969) extended the method to obtain also a stops-offset relation for each link (see also Huddart 1969). Using the notation of section 1.26, these relations for a link from intersection I_j to intersection I_k can be denoted by d_{jk}(z_{jk}) and s_{jk}(z_{jk}) respectively, where z_{jk} is the offset on the link. If, as may occasionally happen, there is more than one link (in the same direction) from I_j to I_k it will be convenient to let d_{jk}(z_{jk}) and s_{jk}(z_{jk}) respectively denote the sums of the delay-offset and stops-offset relations for such links; further, let d_{jk}(z_{jk}) and s_{jk}(z_{jk}) be identically zero if
there is no link from $I_j$ to $I_k$. Then, since $z_{jk}$ is determined by the offsets $x_j$ and $x_k$ at $I_j$ and $I_k$ (see section 1.26) any desired performance index (as defined in section 4.33) takes the form

$$p(z) = \sum_{j=1}^{n} \sum_{k=1}^{\infty} \{ a \delta_{jk}(Z_{jk}) + b \delta_{jk}(Z_{jk}) \},$$

where $z = (x_1, x_2, \ldots, x_n)$, and $a$ and $b$ are chosen constants.

A further improvement to this method by Huddart and Turner (1969) was to allow for the spreading of platoons as they travel along the links, using the following exponential smoothing process devised by Robertson (1969a). With the network discrete time assumption (see section 1.26), let $\tau$ be defined so that the average travel time along the link in question for undelayed vehicles is $5\tau T/4N$, where $N$ is the number of intervals into which the master cycle is divided. Suppose further that a platoon begins to leave the upstream signal at time $t = 0$, that the number of p.c.u.'s leaving the upstream signal in the time interval $[(i-1)T/N, iT/N]$ is $q_i$, and that the number arriving at the downstream stopline in the same interval would be $q_i'$ if the downstream signal were always green. Then Robertson showed that the spreading of platoons observed in West London was well reproduced by taking

$$q_i' = \begin{cases} 0 & \text{if } i \leq \tau \\ 2q_{i-\tau} + \tau q_{i-1} & \text{if } i > \tau \end{cases}$$

This smoothing process can be applied to each of the platoons
that contribute to the in-pattern.

The method used by the Traffic Research Corporation (1966) also requires that a delay-offset and a stops-offset relation be obtained for each link, the allocations of green time having been determined in advance. In their method, traffic from just one of the approaches at the upstream signal on each link is selected as the "primary flow" on that link. It is assumed that the primary flow is the only one whose contribution to the in-pattern depends on the offset at the upstream signal. The arrival at the downstream signal of traffic from other upstream approaches is assumed to be spread uniformly over the cycle. For each link, for example that from \( I_j \) to \( I_k \), the desired linear combination of the delay-offset and stops-offset relation is then estimated by a process that allows for spreading of the primary flow platoon. The offset \( z_{jk}^* \), say, that minimises the resulting function is identified, and a parabola of the form \( a_{jk}(z_{jk} - z_{jk}^*)^2 \) is then fitted to the estimated function in the interval \( (z_{jk}^* - \frac{1}{4}T, z_{jk}^* + \frac{1}{4}T) \) and extrapolated by \( \frac{1}{4}T \) in each direction to cover the whole cycle with \( z_{jk}^* \) as its midpoint. This will often overestimate the rate of delay and number of stops per unit time associated with offsets outside the interval \( (z_{jk}^* - \frac{1}{4}T, z_{jk}^* + \frac{1}{4}T) \). Letting \( a_{jk} = 0 \) if there is no link from \( I_j \) to \( I_k \), the performance index takes the form

\[
\varphi(\mathbf{x}) = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} (z_{jk} - z_{jk}^*)^2,
\]

where \( \mathbf{x} \) is, as before, the vector whose components are the offsets at the intersections.
Chang (1967) regards both the offsets and the allocations of green time as variables. He makes the very restrictive assumption, however, that traffic enters each link from only one approach at the upstream signal. With this assumption, he obtains equations determining, for any given signal settings, the queue-length $Q(t)$, say, on each approach in the network at time $t$ after the start of the master cycle. He takes as the performance index the rate of delay for the whole network, i.e.

$$p(x, \lambda_1, \lambda_2, \ldots, \lambda_n) = \frac{1}{T} \sum_{j=0}^{T} Q(t) dt$$

the summation being over all approaches in the network; $x$ is the vector of offsets at the intersections, and $\lambda_j$ is the vector specifying the allocation of green time at intersection $I_j$ ($j = 1, 2, \ldots, n$) (see section 3.12). In estimating the amount of traffic in each queue, Chang takes into account the length of road that the queue occupies; he thus allows for the fact that vehicles spend more time in the queue than they would do if the queue were able to accumulate exactly at the stopline. This leads, however, to an overestimate of delay, because it counts as delay the time that the delayed vehicles would have taken to travel at the design speed along the length of road occupied by the queue. He makes no allowance for the spreading of platoons.

Yumoto, Kitamura and Nakahara (1967) regard the allocations of green time as determined in advance, and seek only to calculate offsets. They obtain delay-offset and stop-offset relations for each link by a method which, though developed independently by Inose, Fujisaki and Hamada (1967),
is essentially the same as Hillier's, and the resulting performance index has the same form.

Robertson (1969a and b) regards both the offsets and the allocations of green time as variables. For any given signal settings he estimates the rate of delay and number of stops per unit time on each link of the network in turn from their in-, go- and out-patterns, taking the links in such an order that, as far as possible, when the calculation is being made for the link $I_j$ to $I_k$, say, it has already been made for all the links having $I_j$ as their downstream intersection. The out-patterns for these links indicate just how the rate at which traffic enters the link from $I_j$ to $I_k$ varies during the cycle. This traffic pattern, adjusted by the smoothing process already described, to allow for platoon spreading, determines the in-pattern for the link from $I_j$ to $I_k$. Where there are closed loops in the network (other than loops formed by links in parallel), no such order exists, and there are certain links for which it is necessary to make the calculation before all the required out-patterns at the upstream intersection are known. In such cases the unknown out-patterns are assumed to be such as would result from regular arrivals on the approaches concerned. By careful choice of the order in which the links are taken, the likely error resulting from this assumption can be made small, and the result is a very realistic model of traffic movement in the network. To allow for random fluctuations about the cyclic flow pattern, Robertson adds to the rate of delay estimated from the in- and out-patterns on each link a
quantity \( x^2/4(1-x) \), where \( x \) is the degree of saturation at the downstream signal (see section 1.24). The form of this expression was suggested by the second term in Webster's expression for delay on a signal approach (see section 2.51). The 2 in the denominator of Webster's term was replaced by 4 on the basis of observations of platoon behaviour in West London; the need for some modification is not surprising, because Webster's expression applies to traffic arriving at random, rather than in platoons. The resulting performance index can be written

\[
p(x, \lambda_1, \lambda_2, \ldots, \lambda_n) = \sum_{j=1}^{n} \sum_{k=1}^{n} \left\{ a\left( d_{k\ell} + \frac{x_{k\ell}}{4(1-x_{k\ell})} \right) + b\|s_{k\ell}\| \right\}
\]

where \( x \) and the \( \lambda_j \) are as previously defined, \( a \) and \( b \) are chosen constants and

\[
\begin{align*}
d_{k\ell} & = \text{rate of delay} \\
s_{k\ell} & = \text{number of stops per unit time} \\
x_{k\ell} & = \text{degree of saturation at downstream signal}
\end{align*}
\]

on the link from \( I_j \) to \( I_k \) if such a link exists, and \( d_{k\ell} = s_{k\ell} = x_{k\ell} = 0 \) otherwise.

Longley (1969), again regarding both the offsets and the allocations of green time as variables, applies a technique from control theory: that of the z-transform. Suppose that a quantity \( q \), say, takes the values \( q_k \) at time \( k\tau \) \((k = 0, 1, 2, \ldots)\), where \( \tau \) is some constant interval. Then the z-transform \( Q(z) \) of \( q \) is defined by

\[
Q(z) = \sum_{k=0}^{\infty} q_k z^k ;
\]

where \( z \) is the complex variable. With the network discrete
time assumption, Longley takes $\gamma = T/N$ (see section 1.26). He then specifies the cyclic arrival pattern on each approach where traffic enters the network by the corresponding numbers $a_k$ of, say, p.c.u.'s arriving in the intervals $[(k-1)T/N, kT/N]$ ($k = 1, 2, \ldots, N$). He then show how, in principle, the $z$-transform of the rate of delay in the network can be obtained, for any given signal settings, in terms of the $z$-transforms of the arrival rates at the approaches where traffic enters the network. The $z$-transform for such an arrival rate takes the form

$$A(z) = \sum_{j=0}^{\infty} z^{nj} \left( \sum_{k=1}^{N} a_k z^{k-1} \right)$$

Martin-Löf (1969) regards the allocations of green time as determined in advance, and seeks only to calculate offsets. He takes as his starting point delay-offset relations such as those obtained by Hillier's method, and considers cases in which these relations can be approximated, within any one cycle, by piecewise linear convex functions. In particular, he considers the following case: if there is a link from intersection $I_j$ to intersection $I_k$, let the rate of delay on this link be least when the offset on the link is $z^*_jk$, and greatest when the offset is $z^{**}_{jk}$, where

$$z^{**}_{jk} - T < z^*_jk < z^{**}_{jk}.$$

Then suppose that the delay-offset relation on the link can be approximated by

$$d_{jk}(z_{jk}) = \begin{cases} a_{jk} + b_{jk} \frac{z^*_jk - z_{jk}}{z^*_jk - z^{**}_{jk} + T} & \text{for } z^{**}_{jk} - T < z_{jk} < z^*_jk \\ a_{jk} + b_{jk} \frac{z_{jk} - z^*_jk}{z^*_jk - z^{**}_{jk}} & \text{for } z^*_jk < z_{jk} < z^{**}_{jk} \end{cases}$$
If there is no link from \( I_j \) to \( I_k \), let \( d_{jk}(z_{jk}) \) be identically zero. Then, since the offsets \( z_{jk} \) on the links are piecewise linear functions of the offsets \( x_i \) \((i = 1, 2, \ldots, n)\) at the intersections (see section 1.26), the performance index

\[
p(x) = \sum_{j=1}^{m} \sum_{k=1}^{n} d_{jk}(z_{jk})
\]

is a piecewise linear function of \( x_1, x_2, \ldots, x_n \).

4.35. Choice of signal settings. In each method the objective is to choose the signal settings so as to minimise the performance index. Before considering the various methods in detail, one observation is relevant to all the methods that seek to determine only the offsets: since the performance index for such a method is a function only of differences between offsets at intersections, the offset at any one intersection can be chosen arbitrarily. This means that if part of the network is joined to the remainder at only one intersection it can be treated as a separate network; it will be assumed that this process has been carried out to the extent that any network either has no such parts or comprises just one road with several signalised intersections.

The method used by Whiting (Hillier 1965) is called the combination method because it is based on successive simplifications of the graph of the network by combining pairs of arcs in parallel and in series (see section 4.31). The procedures for combining links will now be described, and in doing so it will be convenient to denote by a single function
\[ p_{jk}(z_{jk}) \] the contribution, previously denoted by
\[ a \cdot d_{jk}(z_{jk}) + b \cdot s_{jk}(z_{jk}), \] to the performance index from a
link from intersection \( I_j \) to intersection \( I_k \).

Two arcs in parallel in the same direction from vertex \( A_j \) to vertex \( A_k \) are combined by replacing them by a single arc in the same direction between the same two vertices. The contribution to the performance index from the new arc is obtained simply by adding the contributions from the two original arcs. Such pairs of arcs were assumed already to have been combined when the form of the performance index for Hillier's method was derived in section 4.34.

Two arcs in parallel in opposite directions between vertices \( A_j \) and \( A_k \) are combined by replacing them by a single arc in one direction between the same two vertices. Suppose that the direction of this arc is from \( A_j \) to \( A_k \); then in calculating the performance index \( p(x) \), \( p_{jk}(z_{jk}) \) is replaced by \( p_{jk}(z_{jk}) + p_{kj}(z_{kj}) \) and \( p_{kj}(z_{kj}) \) is replaced by zero. The value of \( p(x) \) is thus unaltered, but it now contains one less non-zero term.

Two arcs in series between \( A_i \) and \( A_j \) and \( A_j \) and \( A_k \) are combined by replacing them by a single arc between \( A_i \) and \( A_k \), directed, say, from \( A_i \) to \( A_k \). Vertex \( A_j \) is deleted from the graph, and the performance index \( p(x) \) is replaced by a new index \( p^*(x^*) \), where \( x^* = (x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n) \), i.e. the vector of offsets at the intersections corresponding to the remaining vertices. The index \( p^*(x^*) \) is defined by
\[
p^*(x^*) = p_{ik}^*(z_{uk}) + \sum_{rsj} \sum_{t} p_{rs}(z_{rs})
\]
where

$$p_{ik}^*(z_{ik}) = \min_{x_j} \{ p_{ij}(z_{ij}) + p_{jk}(z_{jk}) + p_{lk}(z_{lk}) \}$$

is the contribution to the performance index from the new arc. This contribution is thus calculated on the assumption that, for any given \( x_i \) and \( x_k \), \( x_j \) is chosen so that the contribution is minimised. Let \( x_j^* \) be a value of \( x_j \) at which the minimum is attained. Then \( x_j^* \) is a function of \( z_{ik} \), and the calculation and recording of this function completes the process of combining the two links in series. In calculating and recording such functions, Whiting makes the network discrete time assumption (see section 1.26), under which \( z_{ik} \) takes just the \( N \) values \( mT/N \), \( m = 0, 1, 2, ..., N-1 \).

A value of \( x \) that minimises \( p(x) \) can be obtained by finding a value of \( x^* \) that minimises \( p^*(x^*) \), thus determining all the components of \( x \) except \( x_j \), and giving \( x_j \) the value \( x_j^* \) determined by \( z_{ik} \) (i.e. by \( x_i \) and \( x_k \)) from the recorded function. The choice of offsets at the \( n \) intersections of the actual network has then been reduced to a choice of offsets at the \( n-1 \) intersections of a hypothetical network.

If the graph of the original network is not more complicated than a tree of ladders (see section 4.31) it can be reduced by successive combinations of arcs in parallel and in series to a single arc between two vertices. The resulting performance index is a function only of the difference
between the offsets at the intersections corresponding to these vertices and, under the network discrete time assumption, offsets that minimise the index can be found by inspection. Offsets at the remaining intersections can then be found in turn from the functions recorded during the series combinations at which the corresponding vertices were deleted from the graph.

Offsets calculated in this way will certainly be such as to minimise the performance index. The procedure is an example of dynamic programming and will be described more rigorously as such in Chapter 5, where it will be extended to apply, in principle, to any network, regardless of whether the corresponding graph can be reduced to a single arc by combining arcs in parallel and in series.

The order in which the series and parallel combinations are made had originally to be specified by the engineer, but Robertson (1967) showed how, using the matrix associated with the graph (see section 1.25), the determination of such an order could be made part of a computer program. Another technique for the same purpose was devised by Huddart and Turner (1969).

In the Traffic Research Corporation method (1966), the performance index \( p(x) \) is a piecewise quadratic function of \( x_1, x_2, \ldots, x_n \), defined in the region \( 0 \leq x_i \leq T \) \((i = 1, 2, \ldots, n)\) of \( \mathbb{R}^n \). With the notation used in section 4.34, there are discontinuities in the gradient of \( p(x) \) at the hyperplanes

\[
x_k - x_j = z_{jk}^* \pm \frac{1}{2} T \quad (\text{mod } T)
\]
for each pair \((j, k)\) such that there is a link from intersection \(I_j\) to intersection \(I_k\). The region in which \(p(x)\) is defined is therefore divided into \(5^{J2^K}\) subregions, where \(J\) is the number of pairs of adjacent intersections between which there are links in both directions, and \(K\) is the number of pairs of adjacent intersections between which there is a link in just one direction. Since \(p(x)\) is quadratic, the local minimum within any one of the subregions can be found by solving a set of linear equations, and the authors show that just one matrix inversion enables these equations to be solved for all subregions. Even so, however, the number of subregions is too large for all the local minima to be evaluated and compared in order to find the global minimum. Instead, the authors show how, having found the local minimum in one subregion, the local minima in all adjacent subregions can be quickly found. They then search for the global minimum as follows: start at the local minimum of a randomly chosen subregion, and move to the least of the local minima of the adjacent subregions (assuming that this is less than the starting point). Continue by such moves until a subregion is found where the local minimum is less than that of any adjacent subregion, and take this local minimum as an estimate of the global minimum. This process is repeated with new randomly chosen starting points until an estimate is obtained which is less than any of the next 10 estimates; it is then assumed that the global minimum has been found. It seems plausible, though it cannot be certain, that this procedure will usually come near to finding the true global minimum.
Chang (1967) seeks to minimise the performance index by a gradient method which he does not specify in detail. As a starting point he takes settings obtained by applying the bandwidth criterion (Morgan and Little 1964) to selected roads in the network.

Yumoto et al (1967) point out that offsets on a set of links in the network can be freely chosen provided that the links in the set do not form any closed loops. They therefore choose a maximal tree in the graph of the network and choose the offsets so that the contribution to the performance index from each arc in the tree is minimised. A maximal tree is one that contains all the vertices of the graph and is such that if any arc not included in the tree is added to it, a closed loop is formed; there are in general many such trees, and Yumoto constructs one in the following way. He defines the offset effectiveness of a link from intersection $I_j$ to intersection $I_k$ as

$$
\frac{1}{T} \int_{0}^{T} p_{jk}(z_{jk}) \, dz_{jk} - \min_{z_{jk}} \{ p_{jk}(z_{jk}) \},
$$

where $p_{jk}(z_{jk})$ is the contribution to the performance index from this link. The maximal tree is formed by taking the arcs of the graph in decreasing order of the offset effectiveness of the corresponding links, and including in the tree any arc that does not form a closed loop with those already included. This construction of the tree leads to a choice of offsets that minimises the contributions to the performance index from those links whose contributions depend most strongly on the offsets. The resulting performance index is likely to be low,
but will not in general be minimal.

Robertson (1969a) makes the network discrete time assumption (see section 1.26) with \( N = 50 \), and searches for the minimum performance index by a "hill-climbing" method in which the offsets are varied in turn by steps of size \( 7T/50 \), \( 20T/50 \) and \( T/50 \), and the allocations of green time by steps of size \( T/50 \). At each stage, the offset or allocation of green time at just one intersection is varied. The use of different step sizes is intended to reduce the chance of finding a local minimum that appreciably exceeds the global minimum, and Robertson gives an example which suggests that this objective is quite well achieved. An alternative selection of step sizes can be specified if required.

Longley (1969) also proposes a hill-climbing method in which the network discrete time assumption is made and the offsets and allocations of green time are each varied in turn by steps of size \( T/N \), but he has not yet reported any results.

Martin-Löf (1969) shows how the restrictions imposed on the offsets by the existence of closed loops in the network can be expressed as linear equality constraints, each containing an integer variable lying within a certain range. For any particular values of these integer variables, and with a performance index that is a suitable piecewise linear function of the offsets (see section 4.34), the choice of offsets to minimise the performance index is shown to be a linear programming problem that is soluble by the simplex method (see e.g. Vajda 1961). Martin-Löf goes on to propose a branch and bound algorithm for finding the set of values of the
integer variables that gives the smallest minimum performance index; he has, however, not yet reported any results.

4.36. Comparative results. Comparative results are available for only two of the methods discussed in the last three sections: those of Hillier and Robertson. Robertson (1969a) gives the values of the rate of delay, as estimated by his traffic model, for a network in the Victoria area of London with signal settings obtained by his own method, by Hillier's combination method, and by a graphical method. The estimated rate of delay with combination method settings is about 10 per cent less than with the settings obtained graphically, whilst the estimated rate of delay with Robertson's settings is about 10 per cent less again than with the combination method settings. These are, however, only estimates; actual observations made by driving cars through a network in Glasgow (Hillier and Holroyd 1968) give a rather different picture. Whilst they show a reduction of about 12 per cent in the rate of delay with combination method settings compared with settings obtained graphically, the rate of delay observed with Robertson's settings was only about 4 per cent less than with the combination method settings. These observations suggest that a large proportion of the benefit to be obtained by calculating settings for linked signals can be obtained by means of the combination method.

4.4. Comment

Of the methods discussed in sections 4.33-4.35 for calculating settings for linked signals in a network of roads,
the best tried are those of Hillier (1965), the Traffic Research Corporation (1966) and Robertson (1969a). The performance index used in the Traffic Research Corporation method seems less satisfactory than the other two, because it takes no account of the effect of signal settings on turning traffic and because its piecewise quadratic form is likely often to be unrealistic. Of the other two methods, Robertson's has the more realistic traffic model and has the advantage of determining the allocations of green time as well as the offsets. It has however the disadvantages that its technique for minimising the performance index is subject to slight uncertainty, and that the user must make a careful choice of the order in which the links of the network are to be dealt with in the calculation. Hillier's method uses a rigorous technique for minimising the performance index and also has the advantage that the user is free to specify how the contribution from any link to the performance index depends on the offset on that link. It has the disadvantage that it applies only to a somewhat restricted range of networks; the relaxation of this restriction is the main subject of Chapter 5. Another problem dealt with in Chapter 5 is the choice of the common cycle time for linked signals - a choice that the methods so far described have left largely to the judgement of the engineer.
5. **AN IMPROVED METHOD OF CALCULATING SETTINGS FOR LINKED TRAFFIC SIGNALS.**

This Chapter sets out a new method of calculating settings for linked traffic signals in a network of roads; it makes extensive use of parts of the earlier work described in sections 4.33-4.35, and complements these where necessary with new techniques. The duration of the master cycle is first determined by a new method, and the allocation of green time at each intersection is obtained by the methods of Chapter 3. Whiting's method, as adapted by Huddart and Turner, is used to determine the contribution to the performance index from each link in the network as a function of the offset on the link. The offsets at the intersections are then chosen in the following way. Whiting's combination method is used to combine arcs in the graph of the network in parallel and series until either the graph is reduced to a single arc or no more such combinations are possible. If the graph is not reduced to a single arc, then a new method is used to choose the offsets at the intersections corresponding to the vertices that remain.

The data required for each intersection in the network include those required in Chapter 3 for the calculation of signal settings at that intersection considered in isolation; it is further necessary to specify, for each approach, what proportion of the traffic arriving on that approach departs along each of the links leading away from the intersection. The layout of the network must be specified by means of its graph, and, for each link, the average time taken by an
undelayed vehicle to travel along the link must be specified.

5.1. Choice of cycle time.

Suppose that the network has \( n \) intersections \( I_j \) 
\((j = 1, 2, \ldots, n)\). If settings are determined by the method
of Chapter 3 for the signals at each of these intersections
considered in isolation, the resulting cycle times will not in
general all be equal, whereas, if the signals are to be linked,
a common value for the cycle time must be chosen. Suppose,
therefore, that settings are calculated by the method of
Chapter 3, but with some common specified cycle time \( c \). At
intersection \( I_j \), let

\[
L_j = \text{the total lost time per cycle, and}
\]

\[
D_j(\lambda_j^*) = \text{the rate of delay corresponding to the}
\]

\[
\text{delay-minimising settings determined by}
\]

\[
\text{the vector } \lambda_j^*.
\]

Then it was shown in section 3.8 that

\[
\frac{\partial D_j(\lambda_j^*)}{\partial c} = \frac{1}{c} \mu_{ij} \lambda_j^*.
\]

where \( \mu_{ij} \) is a quantity that is determined by the components
of \( \nabla D_j(\lambda_j^*) \). Now let

\[
\mathcal{D} = \sum_{j=1}^{n} D_j(\lambda_j^*)
\]

Then

\[
\frac{\partial \mathcal{D}}{\partial c} = \frac{1}{c} \sum_{j=1}^{n} L_j \mu_{ij}
\]

The rate of change of the sum of the minimum rates of delay
at all the intersections (each considered in isolation) with
respect to the specified cycle time is thus known. By varying c in the direction of decreasing D a local minimum of D can be found. Although such a minimum will not in general be unique, the range of values of c that are acceptable in practice is usually small, because too short a cycle time will provide insufficient capacity at heavily loaded intersections, and cycle times exceeding about 120 seconds are usually rejected because they are believed to lead to drivers' disobeying the lights. Within the acceptable range, therefore, the overall minimum of D can reliably be assumed to be little, if at all, less than the least of the local minimum found by taking a few starting values of c distributed over the acceptable range. Further research is needed to discover just how many starting values need to be taken in order to obtain a local minimum acceptably close to the global one.

The resulting estimate of the value of c that minimises D will be taken as the common cycle time T for the linked signals.

5.2. Allocation of green time.

Having chosen the common cycle time T, the allocation of green time at each intersection is determined by the method of Chapter 3 with the cycle time specified as T . In the subsequent calculation of delay-offset and stops-offset relations, however, and in the choice of offsets, the network discrete time assumption (see section 1.26) will be made. The effective green time for each stage and the lost time between any two stages at each intersection must therefore be
integer multiples of $T/N$. This can be achieved by ensuring that the lost times between stages and any specified minimum green times, which are part of the data, are multiples of $T/N$, and then rounding off the resulting effective green times for each intersection in such a way that their sum is unchanged. Since $N$ is usually 50 and $T$ does not usually exceed 120 seconds, the rounded effective green times will not differ by much more than 1 second, at most, from the delay-minimising values. One precaution that should be taken in rounding off, however, is to avoid reducing the effective green time for any approach that has both a short effective green time and a high degree of saturation.

5.3. Calculation of the performance index.

The allocations of green time, together with data about flows on each approach and travel times on each link of the network enable the delay-offset and stops-offset relation for each link to be estimated by Whiting's method, as adapted by Huddart and Turner (see section 4.34). The contribution to the performance index from any link may be any chosen linear combination of the delay-offset and stops-offset relations. The linear multipliers need not be the same for each link. Large multipliers can be used on links for which it is particularly desirable to obtain low delays and to avoid vehicles being stopped; small multipliers are appropriate for links where large delays and numerous stopped vehicles are less unacceptable.

Let $x$ denote the vector $(x_1, x_2, \ldots, x_n)$, where
\( x_j \) is the offset at intersection \( I_j \), and let
\[
\begin{align*}
  z_{jk} &= x_k - x_j \pmod{T}
\end{align*}
\]

Then if arcs in the same direction between any two vertices in the graph of the network have been combined in parallel (see section 4.35) the performance index for the whole network of \( n \) intersections takes the form
\[
\begin{align*}
  p(\mathcal{X}) &= \sum_{j=1}^{n} \sum_{k=1}^{n} p_{jk}(z_{jk})
\end{align*}
\]

where \( p_{jk}(z_{jk}) \) is the total contribution from the links from \( I_j \) to \( I_k \) if there are such links, and \( p_{jk}(z_{jk}) \) is identically zero if there is no such link.

5.4. Choice of offsets

The way in which the offsets are chosen depends strongly on the layout of the network, and, as described in section 1.25, the network will be represented by a graph \( G \), say, in which intersection \( I_j \) in the network is represented by vertex \( A_j \) \((j = 1, 2, \ldots, n)\). The \( nxn \) matrix associated with the graph will be denoted by \( M = (m_{jk}) \). The following definitions (cf. Re 1962) will be useful.

DEFINITIONS: The local degree of a vertex in a graph is the number of arcs that have the vertex as an endpoint.

A vertex \( A \) in a graph is a separating vertex if the vertices other than \( A \) can be divided into two sets such that in order to pass from any vertex in one set to any vertex in the other along arcs of the graph it is necessary to pass through the vertex \( A \).
As noted by Robertson (1967) the local degree of vertex \( A_j \) in \( G \) is \( \sum_i m_{ij} + \sum_k m_{jk} \). The local degree of each vertex in a graph is thus easily found by examining the associated matrix.

5.41. Order of choice. The order in which the offsets \( x_1, x_2, \ldots, x_n \) are chosen will be specified by constructing a permutation \( \pi \) of the integers 1, 2, ..., \( n \).

The first step is, following Whiting, to make all possible combinations of arcs in parallel in \( G \) (see section 4.35). As Robertson showed (1967), such a combination of arcs between vertices \( A_j \) and \( A_k \) is possible if and only if \( m_{jk} + m_{kj} > 1 \). Let the resulting graph be \( G_n \).

As discussed at the beginning of section 4.35, it will be assumed, without loss of generality, that the network either comprises just one road with several signalised intersections or has no parts that are joined to the remainder at only one intersection. If the network comprises just one road with \( n \) signalised intersections, then \( G_n \) consists of \( n-1 \) arcs placed end-to-end. Otherwise, \( G_n \) has no separating vertices, since any such vertex would correspond to an intersection such that part of the network was joined to the remainder only at that intersection.

Let \( A_{\pi n} \) be some vertex with local degree 2 in \( G_n \), if any such vertex exists. Suppose that \( A_j \) and \( A_k \) are the two adjacent vertices; then let the arcs between \( A_j \) and \( A_{\pi n} \) and \( A_{\pi n} \) and \( A_k \) be replaced by a single arc between \( A_j \) and \( A_k \), and let vertex \( A_{\pi n} \) be deleted from
the graph. (cf. the series combination of Whiting's method, described in section 4.35). If there is an arc between \( A_j \) and \( A_k \) in \( G_n \), let the new arc be combined in parallel with it, and let the resulting graph with \( n-l \) vertices be \( G_{n-l} \).

**Lemma 12.** If \( G_n \) has no separating vertices, neither does \( G_{n-l} \).

**Proof.** Suppose \( A_s \) were a separating vertex in \( G_{n-l} \). Then the vertices of \( G_{n-l} \) could be divided into two sets \( S_1 \) and \( S_2 \) such that to pass from any vertex in \( S_1 \) to any vertex in \( S_2 \) along arcs of \( G_{n-l} \) it is necessary to pass through \( A_s \).

Let \( A_j \) and \( A_k \) be the vertices adjacent to \( A_{\pi_n} \) in \( G_n \); then there is an arc joining \( A_j \) and \( A_k \) in \( G_{n-l} \), so that \( A_j \) and \( A_k \) would belong to the same set. Suppose they belonged to \( S_1 \); then the vertices of \( G_n \) other than \( A_s \) could be divided into two sets \( S_1 \cup \{ A_{\pi_n} \} \) and \( S_2 \) such that to pass from any vertex in one set to any vertex in the other along arcs of \( G_n \) it is necessary to pass through \( A_s \). Hence \( A_s \) would be a separating vertex in \( G_n \); but \( G_n \) has no such vertices, so that there can be no such vertex in \( G_{n-l} \).

Let the process by which \( G_{n-l} \) was obtained from \( G_n \) be repeated until a graph is obtained in which there is no vertex of local degree 2. Let the resulting sequence of graphs be \( G_n \), \( G_{n-l} \), \( \ldots \), \( G_r \), there being no vertex of local degree 2 in \( G_r \). If there is no such vertex in \( G_n \), let \( r = n \). Let the vertices deleted in this process be \( A_{\pi_n} \), \( A_{\pi_{n-1}} \), \( A_{\pi_{n-2}} \), \( A_{\pi_{n-3}} \), in that order; this procedure defines \( \pi_n \), \( \pi(n-l) \), \( \ldots \), \( \pi(r+1) \), and is illustrated by an example in Figure 10(a).
Fig. 10(a) ILLUSTRATION OF THE ORDER OF CHOICE OF OFFSETS:
DETERMINATION OF $\Pi n, \Pi (n-1), \ldots, \Pi (r+1)$ WHERE $n=14$ & $r=8$
(STEPS CORRESPONDING TO SERIES AND PARALLEL COMBINATIONS)
THEOREM 7. Either \( r = 2 \), or \( r \geq 4 \) and no vertex of \( G_r \) has local degree less than 3.

PROOF. By definition, \( G_r \) is a connected graph without arcs in parallel and without vertices of local degree 2. The only connected graphs with just 3 vertices and without arcs in parallel are a triangle and a pair of arcs end-to-end; each of these has a vertex of local degree 2. Hence \( r \neq 3 \). It remains only to show that if there is a vertex of local degree 1, then \( r = 2 \). Suppose \( A_m \) is a vertex of \( G_r \) with local degree 1. Let the arc having \( A_m \) as one endpoint have \( A_s \) as the other. Then in order to pass from \( A_m \) to any other vertex of \( G_r \) along arcs of \( G_r \) it is necessary to pass through \( A_s \). \( A_s \) is therefore a separating vertex in \( G_r \) and, by Lemma 12, \( G_{r+1}, G_{r+2}, \ldots, G_n \) each have a separating vertex. But it was assumed that unless \( G_n \) consisted of \( n-1 \) arcs end-to-end it had no separating vertices. Hence \( G_n \) does consist of \( n-1 \) arcs end-to-end. But in such a graph all vertices except the two end ones have local degree 2, and can be deleted in forming the sequence \( G_{n-1}, G_{n-2}, \ldots, G_r \), so that \( r = 2 \).

If \( r = 2 \) the definition of the permutation \( \pi \) is completed by letting the vertices of \( G_2 \) be \( A_{\pi 1} \) and \( A_{\pi 2} \).

In order to complete the definition of \( \pi \) when \( r \geq 4 \) it will be useful to define the star subgraph of a vertex in \( G_r \).

DEFINITION: The star subgraph \( St(A_m) \) of a vertex \( A_m \) of \( G_r \) is the graph comprising \( A_m \) together with those vertices
of $G_r$ adjacent to $A_m$ and the arcs between $A_m$ and such vertices.

Let $A_{\pi r}$ be a vertex in $G_r$ such that no other vertex has a smaller local degree in $G_r$. The choice of $A_{\pi r}$ when several vertices satisfy this condition will be discussed in section 5.43; for the present let one of the vertices satisfying the condition be chosen arbitrarily. Let $H_r = St(A_{\pi r})$.

For $m = r - 1, r - 2, \ldots, 2, 1$ in turn, let $A_{\pi m}$ be a vertex of $G_r$ not in $H_{m+1}$ such that, if $H_m = H_{m+1} \cup St(A_{\pi m})$, the number of vertices in $H_m$ other than $A_{\pi m}, A_{\pi (m+1)}, \ldots, A_{\pi r}$ is no larger than would result from any other choice of $A_{\pi m}$ from among the vertices of $G_r$ not in $H_{m+1}$. $A_{\pi m}$ can be identified by examining the local degrees of the vertices of the graph comprising all the vertices of $G_r$ other than $A_{\pi (m+1)}, A_{\pi (m+2)}, \ldots, A_{\pi r}$ and all the arcs of $G_r$ that do not have both endpoints in $H_{m+1}$. Any one of the vertices having the smallest local degree in this graph can be chosen as $A_{\pi m}$. It will be useful to denote by $B_m$ the set of vertices in $H_m$ other than $A_{\pi m}, A_{\pi (m+1)}: \ldots, A_{\pi r}$.

This completes the construction of the permutation $\pi$, and the second part of the procedure is illustrated by an example in Figure 10(b). The reason for defining $\pi$ in this way will appear in Theorem 9 in the next section.

The offsets $x_1, x_2, \ldots, x_n$ will be chosen in the order $x_{\pi 1}, x_{\pi 2}, \ldots, x_{\pi n}$. 
Fig. 10(b) ILLUSTRATION OF THE ORDER OF CHOICE OF OFFSETS: DETERMINATION OF \( \Pi_r, \Pi(r-1), \ldots, \Pi_1 \) WHERE \( r = 8 \)
5.42. Dynamic programming procedure. In choosing offsets, the network discrete time assumption will be made (see section 1.26). The aim is therefore to minimise the performance index

$$p(x) = \sum_{j=1}^{n} \sum_{k=1}^{n} p_{jk}(z_{jk})$$

where

$$z_{jk} = x_k - x_j \pmod{T}$$

by assigning to each of the offsets, in the order $$x_{\pi_1}, x_{\pi_2}, \ldots, x_{\pi_n}$$, a value which is an integer multiple of $$T/N$$.

For $$m = 1, 2, \ldots, n$$, let $$x_m = (x_{\pi_1}, x_{\pi_2}, \ldots, x_{\pi_m})$$.

Since $$p(x)$$ is a function only of differences between components of $$x$$, the value of $$x_{\pi_1}$$ can be chosen arbitrarily. In order to employ the method of dynamic programming (see, e.g., Hadley 1964) in the choice of the other offsets it is necessary to express the best value $$x_{\pi_m}^*$$, say, of each other offset $$x_{\pi_m}$$ ($$m = 2, 3, \ldots, n$$) as a function of the previously chosen offsets, i.e., of the vector $$x_{m-1}$$. However, in order that the evaluation and storage of these functions in the computer should be practicable, each $$x_{\pi_m}^*$$ must be determined by only a few of the components of $$x_{m-1}$$. Let $$x_m$$ be the vector formed by those components of $$x_{m-1}$$ that determine $$x_{\pi_m}^*$$.

The steps in the dynamic programming procedure are of two kinds, according as $$m > r$$ or $$m < r$$.

The first kind of step is just the series combination of Whiting's method (see section 4.35), and is expressed in a slightly different notation in the following theorem.
THEOREM 8: For $m > r$, $\mathcal{V}_m$ consists of just two components
of $x_m$, and the difference between the components of $\mathcal{V}_m$ determines the difference between $x^*_{m-1}$ and either of these
components.

PROOF: Let the performance index for the graph $G_n$ be
written in the form

$$P_n(x_n) = \sum_{j=1}^{n} \sum_{k=1}^{n} p^{(n)}_{ij,jk}(z_{ij,jk})$$

where $p^{(n)}_{ij,jk}(z_{ij,jk})$ is the contribution from the link from
$A_{ij}$ to $A_{jk}$ after parallel combinations have been made if
there is such a link, and is zero otherwise. Thus for any
pair $(j, k)$, at most one of $p^{(n)}_{ij,jk}(z_{ij,jk})$ and $p^{(n)}_{jk,ik}(z_{jk,ik})$
is non-zero. Then $P_n(x_n) = p(x)$ since $G_n$ is obtained from
$G$ by parallel combinations only, and these do not alter the
value of the performance index (see section 4.35).

Following the series combination procedure, if $A_{ip}$ and
$A_{iq}$ are the vertices adjacent to $A_{in}$ in $G_n$, let

$$P^{(n)}_{in,ip,q}(z_{in,ip,q}) = \min_{x_{in,q}} \left\{ P^{(n)}_{in,ip}(z_{in,ip}) + P^{(n)}_{in,q}(z_{in,q}) + \right. \right.$$

$$\left. + P^{(n)}_{in,q}(z_{in,q}) + P^{(n)}_{ip,ip}(z_{ip,ip}) \right\} \cdots (\star)$$

Then

$$\min_{\mathcal{V}_m} \{ P_n(x_n) \} = \sum_{j, k}^{n} p^{(n)}_{ij,jk}(z_{ij,jk}) + \sum_{j=1}^{n} \sum_{k=1}^{n} p^{(n)}_{jk,ik}(z_{jk,ik})$$

and the second term in this expression for $\min_{\mathcal{V}_m} \{ P_n(x_n) \}$ is
independent of $x_m$, and so cannot affect the value of $x_m$ at which the minimum is attained.
Hence, for any given \( x_{n-1} \), the value of \( x_{\pi n} \) that minimises \( p_n(x_n) \) is the value at which the minimum in the above expression (*) for \( p^{(n)}_{\pi \rho \pi \sigma} \) is attained, and this depends only on \( x_{\pi \rho} \) and \( x_{\pi \sigma} \). Thus the vector \( x_n \) that determines \( x^*_n \) is just \( (x_{\pi \rho}, x_{\pi \sigma}) \) and, since the functions in equation (*) depend only on differences of offsets \( \mod T \), \( x^*_{\pi n} - x_{\pi \rho} \) and \( x^*_{\pi n} - x_{\pi \sigma} \) depend only on \( x_{\pi \sigma} - x_{\pi \rho} \mod T \). The theorem is thus proved for \( m = n \).

The proof is completed by observing that for \( m = r \), \( r+1 \), \( \ldots \), \( n-1 \), and for given \( x_m \), the minimum with respect to \( x_{\pi (m+1)} \), \( x_{\pi (m+2)} \), \( \ldots \), \( x_{\pi n} \) of the performance index \( p(x) \) can be expressed in a form exactly analogous to \( p_n(x_n) \), namely

\[
P_m(x_m) = \sum_{j=1}^{r} \sum_{k=1}^{m} p^{(m)}_{\pi j \pi k} (z_{\pi j \pi k}).
\]

This is achieved by defining the functions \( p^{(m)}_{\pi j \pi k} \) \( (m = r \), \( r+1 \), \( \ldots \), \( n-1 \); \( j = 1, 2, \ldots, m \); \( k = 1, 2, \ldots, m \) \) by means of the following recurrence relation.

\[
p^{(m-1)}_{\pi j \pi k} (z_{\pi j \pi k}) = \begin{cases} 
p^{(m)}_{\pi j \pi k} (z_{\pi j \pi k}) + p^{*(m)}_{\pi j \pi k} (z_{\pi j \pi k}) & \text{if } A_{\pi j} \text{ and } A_{\pi k} \text{ are the vertices adjacent to } A_{\pi m} \text{ in } G_m \\
p^{(m)}_{\pi j \pi k} (z_{\pi j \pi k}) & \text{otherwise},
\end{cases}
\]

where, by analogy with equation (*) in the case \( m = n \), if \( A_{\pi j} \) and \( A_{\pi k} \) are the vertices adjacent to \( A_{\pi m} \) in \( G_m \), then
\[ p_{\pi j \pi k}^{\star (m)}(z_{\pi j \pi k}) = \min_{x_{\pi m}} \left\{ p_{\pi j \pi m}^{(m)}(z_{\pi j \pi m}) + p_{\pi m \pi j}^{(m)}(z_{\pi m \pi j}) + p_{\pi m \pi k}^{(m)}(z_{\pi m \pi k}) + p_{\pi k \pi m}^{(m)}(z_{\pi k \pi m}) \right\} \]  \hfill (**)  

**COROLLARY.** For \( m > r \), just \( N \) values of \( x_{\pi m}^* \) have to be stored, and each of these values is found by comparing \( N \) possibilities.

**PROOF:** The quantity that must be added \((\text{mod } T)\) to either component of \( x_{\pi m} \) to obtain \( x_{\pi m}^* \) depends only on the difference \((\text{mod } T)\) between the components of \( x_{\pi m} \). This difference can take \( N \) values; hence just \( N \) values of \( x_{\pi m}^* \) must be stored. For each \( x_{\pi m} \), \( x_{\pi m}^* \) is the value of \( x_{\pi m} \) at which the minimum in equation (**) , above, is attained; and the expression to be minimised must be evaluated for each of the \( N \) possible values of \( x_{\pi m} \) in order to find the minimum by inspection.

If \( r = 2 \), \( p_2(x_2) \) takes the form

\[ p_2(x_2) = p_{\pi 1 \pi 2}^{(2)}(z_{\pi 1 \pi 2}) + p_{\pi 2 \pi 1}^{(2)}(z_{\pi 2 \pi 1}) \]

which depends only on \( x_{\pi 2} - x_{\pi 1} \) \((\text{mod } T)\). Hence, for the arbitrarily chosen value of \( x_{\pi 1} \), the value of \( x_{\pi 2} \) that minimises \( p_2(x_2) \) can be found by inspection of the values of \( p_2(x_2) \) corresponding to the \( N \) possible values of \( x_{\pi 2} - x_{\pi 1} \) \((\text{mod } T)\). The minimum value of \( p_2(x_2) \) is the required minimum of the performance index \( p(x) \) of the original network. The corresponding values of the offsets \( x_{\pi m} \) \((m = 3, 4, \ldots, n)\) can be obtained from the functions \( x_{\pi m}^*(y_m) \) that have previously been evaluated and stored; in
each case \( \mathcal{V}_m \) is determined by the already chosen \( x_{\pi 1}, x_{\pi 2}, \ldots, x_{\pi (m-1)} \). The choice of offsets is thus completed for the case \( r = 2 \).

To complete the choice of offsets in the case \( r \geq 4 \), the second kind of step is required in the dynamic programming procedure. This second kind of step is described in the following theorem.

**THEOREM 9:** For \( m \leq r \), \( \mathcal{V}_m \) consists of the offsets at the intersections corresponding to the vertices in \( B_m \), and the differences \((\text{mod } T)\) between the components of \( \mathcal{V}_m \) determine the difference between \( x^*_{\pi m} \) and any one of them.

**PROOF:** \( B_r \) is the set of vertices adjacent to \( A_{\pi r} \) in \( G_r \), and, for \( m = r-1, r-2, \ldots, 3, 2 \), \( B_m \cup \{ A_{\pi m} \} \supset B_{m+1} \), since \( H_m \supset H_{m+1} \) and \( B_m \) is the set of vertices in \( H_m \) other than \( A_{\pi m}, A_{\pi (m+1)}, \ldots, A_{\pi r} \). Let \( \mathcal{V}_m \) denote the vector of offsets at intersections corresponding to vertices in \( B_m \); then the components of \( \mathcal{V}_m \) are a subset of those of \( x_{m-1} \), and it remains to show that the components of \( \mathcal{V}_m \) determine \( x^*_{\pi m} \).

Consider first the case \( m = r \). Since \( H_r = \text{St}(A_{\pi r}) \),

\[
\mathcal{P}_r(x_r) = \sum_{A_{\pi j} \subseteq \epsilon H_r} \mathcal{P}_{\pi j}(z_{\pi j}) + \sum_{A_{\pi j} \in B_r} \left\{ \mathcal{P}_{\pi j}(z_{\pi j}) + \mathcal{P}_{\pi r-j}(z_{\pi r-j}) \right\}
\]

The first term in this expression for \( \mathcal{P}_r(x_r) \) is independent of \( x_{\pi r} \) and the second term depends only on \( x_{\pi r} \) and \( \mathcal{V}_r \). Hence \( x^*_{\pi r} \) is the value of \( x_{\pi r} \) that minimises the second term and is determined by \( \mathcal{V}_r \). Further, since the expression
to be minimised is a function only of differences \((\text{mod } T)\) between offsets, the differences between the components of \(y_r\) determine the difference between \(x_{\pi r}^\star\) and any one of them.

The above expression for \(p_r(x_r)\) also shows that the contribution to \(p_r(x_r)\) from the arcs of \(H_r\) depends only on \(y_r\) and \(x_{\pi r}\). Let this contribution be denoted by \(S_r(y_r, x_{\pi r})\); then for given \(y_r\), the minimum value of \(S_r(y_r, x_{\pi r})\) is \(S_r(y_r, x_{\pi r}^\star)\).

Similarly, for \(m < r\), the contribution to \(p_r(x_r)\) from the arcs of \(H_m\) depends only on \(y_m\) and \((x_{\pi m}, x_{\pi(m+1)}), \ldots, x_{\pi r}\). Let this contribution be denoted by \(S_m(y_m, x_{\pi m}, x_{\pi(m+1)}, \ldots, x_{\pi r})\); it will now be shown by induction that its minimum value, for given \(y_m\), is \(S_m(y_m, x_{\pi m}^\star, x_{\pi(m+1)}^\star, \ldots, x_{\pi r}^\star)\), where \(x_{\pi m}^\star\) is determined by \(y_m\) and each \(x_{\pi j}^\star\) \((m < j < r)\) is determined by \(y_m\) and \(x_{\pi k}^\star\) \((m < k < j)\). This result has already been proved for \(m = r\). Suppose it is true for \(m = q + 1\). Now

\[
S_q(y_q, x_{\pi q}, x_{\pi(q+1)}, \ldots, x_{\pi r}) = \sum_{A_{\pi j}A_{\pi k} \in \Pi_q} p_{\pi j \pi k}^{(q)}(z_{\pi j \pi k}) + \sum_{A_{\pi j} \in B_q} \left\{ p_{\pi j \pi q}^{(q)}(z_{\pi j \pi q}) + p_{\pi q \pi j}^{(q)}(z_{\pi q \pi j}) \right\} + S_{q+1}(y_{q+1}, x_{\pi(q+1)}, x_{\pi(q+2)}, \ldots, x_{\pi r})
\]

But for given \(y_q\), the first term in the last expression depends only on \(x_{\pi q}\); moreover, since \(B_q \cup \{A_{\pi q}\} \supset B_{q+1}\), \(y_q\) and \(x_{\pi q}\) together determine \(y_{q+1}\). Hence, by the induction hypothesis, the minimum value of the last expression, for given \(y_q\), is
\[
\min \left\{ \sum_{x_{\pi q}} \left[ p_{\pi q}^{r/y} (z_{\pi q} x_{\pi q}) + p_{\pi q}^{r/y} (z_{\pi q} x_{\pi q}) \right] \right\} + \\
+ \min \left\{ s_{x_{\pi q}} (y_{q+1}, x_{\pi q}^*, x_{\pi q}^{(q+2)}, \ldots, x_{\pi q}^*) \right\} 
\]

in which the first term does not depend on \( x_{\pi (q+1)}, x_{\pi (q+2)}, \ldots, x_{\pi r} \). It follows that the minimum value of \( S_q (y_q, x_{\pi q}, x_{\pi (q+1)}, \ldots, x_{\pi r}) \), for given \( y_q \), is

\[
\min_{x_{\pi q}} \left\{ S_q (y_q, x_{\pi q}, x_{\pi (q+1)}, \ldots, x_{\pi r}) \right\}.
\]

It is therefore only necessary to observe that this minimum is attained at \( x_{\pi q} = x_{\pi q}^* \) because \( x_{\pi q}^* \) is the value of \( x_{\pi q} \) at which, for given \( y_q \), \( p_r (x_r) \) is minimised, and

\[
p_r (x_r) = S_q (y_q, x_{\pi q}^*, x_{\pi (q+1)}^*, \ldots, x_{\pi r}^*) + \sum_{A_{\pi k} \in H_K} p_{\pi q}^{r/y} (z_{\pi q}, y_k).
\]

in which expression the second term depends only on

\( x_{\pi 1}, x_{\pi 2}, \ldots, x_{\pi (q-1)} \).

Hence for \( m \leq r \) the minimum value of \( S_m (y_m, x_{\pi m}, x_{\pi (m+1)}, \ldots, x_{\pi r}) \) for given \( y_m \) is \( S_m (y_m, x_{\pi m}^*, x_{\pi (m+1)}^*, \ldots, x_{\pi r}^*) \).

It follows, as required, that the value of \( x_{\pi m}^* \) is determined by \( y_m \). Again, since each \( S_m (y_m, x_{\pi m}^*, x_{\pi (m+1)}, \ldots, x_{\pi r}) \) is a function only of differences (Mod T) between offsets, the difference between \( x_{\pi m}^* \) and any component of \( y_m \) depends only on the differences between the components of \( y_m \).

**COROLLARY:** For \( m \leq r \), if \( b(m) \) is the number of vertices in \( B_m \), then \( N^{b(m)-1} \) values of \( x_{\pi m}^* \) have to be stored and each of these values is found by comparing \( N \) possibilities.
In addition, \( N^{b(m)-1} \) values of \( S_m(y_m, x_{\pi m}^*, x_{\pi (m+1)}^*, \ldots, x_{\pi r}^*) \) have to be stored for use in the evaluation of \( x_{\pi (m-1)}^* \).

**Proof:** The differences between the \( b(m) \) components of \( y_m \) determine the difference between \( x_{\pi m}^* \) and any one of them; each component can take \( N \) values, and \( b(m)-1 \) of the differences are independent. Hence \( N^{b(m)-1} \) values of \( x_{\pi m}^* \) must be stored. Expression (*) in the proof of the theorem shows that the corresponding \( N^{b(m)-1} \) values of \( S_m(y_m, x_{\pi m}^*, x_{\pi (m+1)}^*, \ldots, x_{\pi r}^*) \) are required in the evaluation of \( x_{\pi (m-1)}^* \).

The choice of offsets is completed at the step \( m = 2 \) when \( x_{\pi 2}^* \) is evaluated for the one arbitrarily chosen value of \( x_{\pi 1}^* \).

**5.43. Applicability.** This method of choosing offsets is applicable in principle to any network. A limitation on its use in practice is imposed by its requirements in computer storage and computing time for the dynamic programming steps of the second kind. If any steps of this kind are required, then \( b(m) \geq 3 \) for at least one step. A value of \( b(m) \) equal to 3 presents no difficulty with present computers; this means, for example, that a network whose graph is that shown in Figure 8(f) can be dealt with. If a large computer is available, steps in which \( b(m) = 4 \) are practicable; steps in which \( b(m) = 5 \) are, however, likely to require unacceptable amounts of computing time.
In constructing the permutation \( \pi \) in section 5.41, each \( \pi_m \) \((m = r, r-1, \ldots, 3)\) was chosen so as to minimise \( b(m) \), given the previous choices of \( \pi_r, \pi(r-1), \ldots, \pi(m+1) \). Ideally, it would be desirable to minimise the largest of the \( b(m) \), but no way of doing this, other than examining all possible permutations, has yet been found. The largest value of \( b(m) \) does, however, depend strongly on the choice of \( A_{\pi_r} \). For example, in the graph shown in Figure 8(f), the largest \( b(m) \) is only 3 if \( A_{\pi_r} \) is taken as one of the end vertices, but is 6 if \( A_{\pi_r} \) is taken as a vertex midway along one side of the graph. It is suggested, therefore, that permutations should be constructed taking each vertex of least local degree in \( G_r \) in turn as \( A_{\pi_r} \). Of these permutations, one whose largest \( b(m) \) is least should be chosen as the permutation \( \pi \). These permutations can be quickly computed from the matrix associated with \( G_r \), and any permutation after the first can be rejected as soon as it leads to a value of \( b(m) \) as great as the largest \( b(m) \) arising from the best permutation so far found.

A somewhat different method of choosing the offsets at intersections corresponding to vertices of \( G_r \) was previously proposed by the author (Allsop 1968a and b). The method described here is an improvement on that method in two respects: the determination of the order of choice of offsets is simpler, and the amount of computation required in some of the dynamic programming steps is substantially reduced. This is because in the earlier method two offsets were often chosen in the same step, which required \( N^2 \) possibilities to be compared. The present method chooses such offsets in two
separate steps, each requiring the comparison of only $N$ possibilities.

5.5. Computer programs.

A Fortran program for calculating signal settings at one intersection considered in isolation was described in section 3.6. As discussed in sections 5.1 and 5.2, this program can be used to calculate the common cycle time and allocations of green time for linked signals. The calculation of offsets by the method of section 5.4 requires that the performance index should be expressed as a function of the offsets as described in section 5.3. Part of a Fortran program developed by Huddart and Turner (Turner 1968) has been adapted for this purpose. Further Fortran programs have been written for the calculation of offsets by the methods of section 5.4. These programs have been developed far enough to verify that the method is practicable, but some further development is required if they are to be made easily usable by engineers unfamiliar with the mathematical details of the method.

There are five programs, which will be called programs A - E, in addition to the one described in section 3.6, and an indication of the amount of core store required by each when compiled on the University College IBM 360/65 is given in Table 3. The amount of store required can be divided into two parts; the amount required for program logic is independent of the network to be dealt with, whilst the amount required for Fortran arrays depends on the size of the largest
Table 3

Approximate core store requirements of programs
for calculating offsets for linked signals

<table>
<thead>
<tr>
<th>Program</th>
<th>Purpose</th>
<th>Approximate store requirement (bytes)</th>
<th>Size of principal arrays (bytes)**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>logic arrays* total</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>Calculation of performance index</td>
<td>27000 54000 81000</td>
<td>466a + 10n</td>
</tr>
<tr>
<td>B</td>
<td>Specification of dynamic programming steps of the</td>
<td>13000 13000 26000</td>
<td>4n² + 50n</td>
</tr>
<tr>
<td>C</td>
<td>Calculations for first kind</td>
<td>15000 51000 66000</td>
<td>2n² + 2r² + 2Nn + 6aN + 64n</td>
</tr>
<tr>
<td>D</td>
<td>Specification of dynamic programming steps of the</td>
<td>10000 13000 23000</td>
<td>4r² + 50r</td>
</tr>
<tr>
<td>E</td>
<td>Calculations for second kind</td>
<td>11000 29000 40000</td>
<td>6N² + 2Nₐr + 40r</td>
</tr>
</tbody>
</table>

* for a network with up to 50 intersections and up to 100 links with the signal cycle divided into 50 intervals.

** Notation: N = number of intervals in signal cycle
n = " " intersections in network
a = " " links in network
r = " " intersections remaining after parallel and series combinations
ₐr = " " links remaining after parallel and series combinations

** Program E assumes that the largest value of b(m) does not exceed 4 (see section 5.43)
network for which offsets are to be calculated, and on the number of intervals into which the signal cycle is divided. The approximate amount of store required for each of these purposes are given in the table, and the amount for arrays is sufficient for networks with up to 50 intersections and up to 100 links, with the signal cycle divided into 50 intervals. An indication of the amounts required for arrays for networks of other sizes is given in the last column of the table, where the sizes of the principal arrays are expressed in terms of characteristics of the network. Programs A and E each make use of a disk or tape backing store.

Running times on the University College IBM 360/65 are of the order of a few seconds for programs B and D, a few minutes for programs A and C, and up to an hour for program E.

The computing capacity required by these programs is therefore such that the methods of this chapter can be applied to all but exceptionally large networks using computers that are widely available to traffic engineers.
6. **SYNOPSIS**

The purpose of this chapter is to give an outline of the contents of Chapters 2 - 5, using the definitions set out in Chapter 1. The marginal numbers in this chapter indicate the relevant sections of the earlier chapters.

6.1. **Delay at a fixed-time traffic signal**

2 A survey is made of previous theoretical studies of the delay to traffic on an approach to a signalised intersection. Three simple mathematical models, known as the regular, binomial, and Poisson models, for the arrival of traffic on an approach are fully defined, together with a model for the departure of traffic from the approach. A number of less straightforward arrival models and some variations on the departure model are discussed more briefly.

2.3-2.6 For each arrival model, a discussion is given of previous theoretical expressions for the average delay per passenger car unit on the approach, in terms of the cycle time and the effective green time for the approach. In discussing a frequently used approximate expression (Wardrop 1952) for the average delay on an approach with regular arrivals, bounds for the error resulting from the approximation are obtained. A new and detailed theoretical examination is made of Webster's semi-empirical expression (1958) for the average delay on an approach with Poisson arrivals.

A new expression, extending the work of Buckley and
2.52 Wheeler (1964), is obtained for the average number of passenger car units delayed on an approach with Poisson arrivals when a red time is followed by a red-and-amber time, during which vehicles are delayed only if a queue has formed during the red time, and then by a green time long enough for any queue that has formed to disperse.

6.2. Settings for a single fixed-time traffic signal

Expressions for the delay at a fixed-time traffic signal in terms of the signal settings have previously been used in several methods of calculating signal settings. Examination of these methods shows that, although they give satisfactory results in straightforward cases, they are difficult to apply when there are approaches having right of way in more than one stage of the signal cycle, and when minimum green times are specified for the stages. They also take into account delay to traffic on only one of the approaches having right of way in each stage. A new method, which overcomes these defects, is therefore developed.

The aim of the new method is to minimise the total delay per unit time to traffic passing through the intersection from all approaches, as estimated by the simplified form of Webster's expression. Reasons are given for choosing this expression for delay from among the various expressions discussed in Chapter 2. The stage matrix is assumed to be specified and is subject only to the restriction that each approach has right of way for just one period in each cycle; this period may,
however, comprise either one stage or several successive stages. The lost time following each stage and the saturation flow and average arrival rate on each approach are also assumed to be specified. The settings to be calculated are the cycle time and the allocation of the green time between the stages. The cycle time may be specified, if required, or a maximum may be imposed on it. A minimum may also be imposed on the effective green time for each stage, and the green time for one or more stages may be specified.

The first step in the calculation is to establish whether there are any signal settings that will both satisfy the conditions imposed on the cycle time and green times and allow the traffic arriving on all approaches to pass through the intersection in the long run. If there are no such settings, the intersection is overloaded, but if there are any such settings it is shown that settings that satisfy the imposed conditions and minimise the estimated total rate of delay exist and are unique. These delay-minimising settings are found by proceeding step by step, starting with any settings that satisfy the imposed conditions and allow the traffic on all approaches to pass through the intersection in the long run. It is shown that the procedure used must converge to the required settings.

A computer program for carrying out this procedure is briefly described, and a few illustrative results are given.
The possible effect of including the third term in Webster's expression for delay is examined, and data indicating the size of this effect are printed by the computer program. Other data printed by the program allow the sensitivity of the minimum rate of delay to the choice of cycle time to be estimated in cases where the cycle time has been specified.

6.3. Settings for linked traffic signals

The settings for linked fixed-time traffic signals comprise the common cycle time and the offset and allocation of green time at each intersection. A survey is made of previous methods of calculating such settings. When all the signals are on one main road, the time-distance diagram, in which distance travelled by vehicles on the main road is plotted against time, is useful for representing the signal settings. This diagram also forms the basis for some methods of calculating settings, and its properties have therefore been investigated and described in some detail.

A number of previous theoretical studies of the movement of traffic through a sequence of signals on a one-way main road are discussed. In one of these (Bavarez and Newell 1967), the calculation of delay-minimising settings, taking into account the delay to traffic on both the main road and the roads that cross it, is reduced to that of calculating the cycle time and allocation of green time at a single hypothetical intersection with a large number of approaches. Previous
studies of settings for signals on a two-way main road are found to have been devoted mainly to maximising the widths of the through bands in the time-distance diagram, but one of these studies (Newell 1964) made use of the delay-offset relation.

Previous methods of calculating settings for linked signals in networks of intersecting roads are then discussed. Such methods also apply to the particular case of a sequence of signals on one main road. All the methods begin by estimating the total delay per unit time and number of stops per unit time on all the links of the network as a function of the signal settings. Some linear combination of the delay per unit time and the number of stops per unit time is taken as the criterion for choosing signal settings and is called the performance index. In some methods this index is a function of both the offsets and the allocations of green time; in other methods the allocations of green time are assumed to be determined in advance, and the performance index depends only on the offsets. In all the methods the common cycle time is assumed to be determined in advance. The performance index for each method is calculated by making a mathematical model of the behaviour of traffic in the network, and these mathematical models are compared in some detail.

The aim in choosing signal settings is to minimise the performance index; each method uses its own technique for this purpose, and these techniques are also
compared in some detail.

4.4 From these comparisons it is concluded that Robertson's method (1969a) uses the best traffic model. The combination method (Hillier 1965), however, although its simpler and more adaptable traffic model is rather less realistic when traffic is light, uses the most reliable technique for minimising the performance index. The combination method is, however, applicable only to a limited range of networks. An improved method of calculating settings for linked signals is therefore developed, in which the performance index is calculated using the same traffic model as in the combination method, but a new dynamic programming technique, which applies in principle to any network, is used to minimise the performance index.

Before the performance index can be calculated, the cycle time and allocations of green time must be determined. The new technique, described in section 6.2, for calculating settings for a single fixed-time signal allows the sensitivity of the rate of delay to the choice of cycle time to be estimated. This feature forms the basis of a new method of choosing the common cycle time for the linked signals. The methods described in section 6.2 can also be applied to each intersection in the network to determine the allocation of green time. The performance index is then calculated as a function of the offsets using the same traffic model as in the
combination method, and the offsets are chosen to minimise the performance index by means of the new dynamic programming technique. This completes the calculation of the settings for the linked signals.

The applicability of the new method is limited by the computing capacity required in minimising the performance index, but it is shown that all but exceptionally large networks could be dealt with using computers widely available to traffic engineers. Exceptionally large networks would have to be treated in parts.

6.4. Further work

A number of suggestions for further research and for the development and application of the work described in this thesis are given in Chapter 7.
7. **SUGGESTIONS FOR FURTHER WORK**

No attempt will be made here to give an extensive list of unsolved problems in the theory of fixed-time traffic signals. The survey given in Chapter 2 shows that many details of the theory of delay at an isolated fixed-time signal remain to be studied, whilst Chapter 4 shows that only a limited understanding of the movement of traffic in signalised networks has so far been attained. Further work in the following areas, however, seems particularly likely to yield useful results.

(a) Comparison of the various expressions for delay on an approach to a fixed-time traffic signal with one another and with observed data over a wide range of arrival rates, saturation flows and signal settings.

(b) Investigation of the possibility of calculating settings for fixed-time traffic signals at a single intersection by integer programming methods, having regard to the fact that signal settings can often only be varied in discrete steps.

(c) Investigation of the sensitivity of the delay-minimising settings and the corresponding rate of delay at a single intersection to errors in the arrival rates and saturation flows used in calculating the settings.

(d) Development of the computer programs described in section 5.5 for calculating settings for linked signals, so that they are readily usable by traffic engineers and so as to reduce, if possible, their requirements in computer storage and running time.
(e) Comparison of settings obtained by the method
developed in Chapter 5 with those given by previous methods.

This thesis has been deliberately confined to the theory
of fixed-time traffic signals at intersections that are not
overloaded by the arriving traffic. The appendix contains a
bibliography covering more practical work related to fixed-
time signals, and work related to vehicle-actuated signals,
signals at overloaded intersections, the behaviour of traffic
platoons, and relevant simulation studies.
REFERENCES


HADLEY G. (1964) Non-linear and dynamic programming. Addison-Wesley.


APPENDIX: CLASSIFIED BIBLIOGRAPHY OF RELATED WORK

A.1. Practical work related mainly to fixed-time signals at one intersection

ALMOND J. (1967) Effect of forward visibility, power/weight ratio and type of transmission of cars on the capacity of traffic signals. Road Research Laboratory Report LR60, Crowthorne, Berks.


A.2. Practical work related to linked fixed-time signals


A.3. Studies of vehicle-actuated signals


A.4. Studies of signals at overloaded intersections


A.5. Studies of the behaviour of traffic platoons


A.6. Some relevant simulation studies


KELL J.H. (1963) Results of computer simulation studies as related to traffic signal operation. Proc. 33rd Meeting of the Institute of Traffic Engineers.


