Journal of Mathematical Modelling and Algorithms **2:** 379–406, 2003.

A Lane-Based Optimization Method for Minimizing Delay at Isolated Signal-Controlled Junctions

C. K. WONG and S. C. WONG
Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong, P.R. China

Abstract. This paper presents a lane-based optimization method for minimizing delay at isolated signal-controlled junctions. The method integrates the design of lane markings and signal settings, and considers both traffic and pedestrian movements in a unified framework. While the capacity maximization and cycle length minimization problems are formulated as Binary-Mix-Integer-Linear-Programs (BMILPs) that are solvable by standard branch-and-bound routines, the problem of delay minimization is formulated as a Binary-Mix-Integer-Non-Linear Program (BMINLP). A cutting plane algorithm and a heuristic line search algorithm are proposed to solve this difficult BMINLP problem. The integer variables include the permitted movements on traffic lanes and successor functions to govern the order of signal displays, whereas the continuous variables include the assigned lane flows, common flow multiplier, cycle length, and starts and durations of green for traffic movements, lanes and pedestrian crossings. A set of constraints is set up to ensure the feasibility and safety of the resultant optimized lane markings and signal settings. A numerical example is given to demonstrate the effectiveness of the proposed methodology. The heuristic line search algorithm is more cost-effective in terms of both optimality of solution and computing time requirement.

Mathematics Subject Classifications (2000): 65K05, 90B20, 90C11.

Key words: lane-based optimization method, delay minimization, capacity maximization, cycle length minimization, isolated signal-controlled junction.

1. Introduction

An extensive literature is devoted to the optimization of signal settings at isolated junctions, in which the arrival pattern of traffic is assumed to be of the Poisson type. One of the first systematic mathematical frameworks for traffic signal design was that of Webster [24], in which the signal cycle was divided into separate stages. The green times were allocated to the stages according to the ratio of arrival flow to saturation flow for the representative arm in each stage. An empirical formula for the optimal cycle length to minimize overall junction delay was also given.

A more general approach to the problem was developed by Allsop [2] in which the calculation of signal settings was formulated as a convex mathematical programming problem to minimize the total delay on all approaches to an isolated junction. An ad hoc procedure was derived to solve the problem. Allsop [4] extended the approach to the determination of capacity-maximizing signal settings, and formulated the problem as a linear program that was solvable by any standard linear programming routine. These mathematical programming approaches were implemented in computer programs [3, 5, 6], and were later incorporated into a commercial package called OSCADY [8]. As the stage sequence has to be specified in the calculation, all such methods are called “stage-based”.

Due to the rapid development of urban road systems and the continuous growth of road traffic, junctions in urban streets have become increasingly complicated. The increasing number of traffic and pedestrian streams at a typical road junction makes the manual design of optimal stage sequences almost impossible. Tully [23] developed a procedure to identify all possible stage sequences at any given junction by using graph theory. The maximal cliques (maximal sets of compatible traffic and pedestrian streams) were first determined and then combined to form a stage sequence while satisfying certain constraints. However, when applied to a complicated junction with a large number of traffic and pedestrian streams, the large number of possibilities limits the application of this procedure in calculating optimal signal settings.

Improta and Cantarella [15] proposed a different approach to the problem of calculating signal settings. They directly considered the time domain for each group of streams with right of way. The cycle-structure was specified by a set of binary variables relating to incompatible signal groups. A branch-and-bound technique was employed to tackle the Binary-Mixed-Integer-Linear Program (BMILP). Gallivan and Heydecker [11] and Heydecker and Dudgeon [14] developed a related approach, in which the cycle-structure was specified by one of the stage sequences that were generated by Tully [23]. This stage sequence need not be maintained during optimization, and hence provides much flexibility for the calculation of the signal variables that are determined by the initial stage sequence. The calculations of delay-minimizing and capacity-maximizing timings were formulated as a convex mathematical program and a linear program respectively. However, as already mentioned, Tully’s method produces a large number of stage sequences.

Heydecker [13] reduced this difficulty by introducing a procedure to group all possibilities into a much smaller number of equivalence classes. The cycle-structure of each equivalence class was represented by a successor function. This makes the method computationally more attractive. Most importantly, these methods can include the structure of interstage periods and some aspects of cycle length and sequence selection into the optimization of timings. As these methods deal directly with groups of traffic/pedestrian streams without the need to maintain the stage-structure during optimization, they are called “group-based” (or “phase-based” in British terminology). Silcock [21] reported a more detailed mathematical framework for the group-based method for isolated junctions. This method was also implemented in a commercial package called SIGSIGN [20], and was extended to linked signals [29–33] and combined signal control and assignment problems [34–36].

All of these methods assume that the layout of a junction, including the lane markings, is given as exogenous input. Based on the lane markings, traffic engineers usually start by grouping the traffic lanes into traffic streams, irrespective of the traffic conditions, and then determining the signal settings with the stage- or group-based methods that are mentioned above. However, in reality, traffic may not distribute in the same pattern as the ideal grouping, especially under an unbalanced flow distribution. Therefore, the conventional approach of pre-grouping traffic lanes into traffic streams may not be adequate for some situations. On the other hand, the design of lane markings is usually considered as a pre-requisite for signal timing calculation. However, for complicated junctions, it is very difficult to come up with an optimal set of lane markings for traffic signal design. The conventional approach is to design the lane markings on a trial and error basis, in which an initial set of lane markings is first assumed, and the signal settings are then determined. After assessing the performance of different approaches to the signal controlled junction with the optimal settings, the lane markings are revised (if necessary) based on the engineer's experience, and the procedure is repeated until the performance of the junction is satisfactory. This, however, may not always produce the truly optimal set of lane markings for the junction concerned.

Lam *et al.* [18] combined the design of lane markings and signal timing calculation. They first enumerated all possible sets of lane markings from each approach, then formulated a mixed integer program to maximize the sum of flow factors from all approaches, and finally solved the program by a heuristic solution procedure that consisted of three stages of green time allocations to traffic and pedestrian movements. They showed that by including the lane marking design in the optimization procedure, substantial improvement in the junction performance could be achieved. However, the maximization of the sum of flow factors may not always lead to the maximization of the junction capacity. Moreover, the allocation of green times to pedestrian crossings was conducted at the later stage of the optimization heuristics, and is usually subject to lower priority. The separate consideration of traffic and pedestrian movements may also produce sub-optimal results. To overcome these difficulties, Wong and Wong [26] proposed a lane-based optimization method for determining the capacity-maximizing and cycle-length-minimizing signal settings for an isolated junction, which was formulated as a BMILP that was solvable by any standard branch-and-bound technique.

However, for operational design, the junction delay is considered an important performance indicator for the operational efficiency of a signal-controlled junction. Therefore, this study extends the lane-based optimization method for determining a combined set of lane markings and signal settings to minimize the traffic delay at an isolated signal-controlled junction. Both traffic and pedestrian movements are considered in a unified framework. The problem is formulated as a Binary-Mix-Integer-Non-Linear-Program (BMINLP). A cutting plane algorithm and a heuristic line search algorithm are proposed to solve the delay-minimization problem. The integer variables include the permitted movements in traffic lanes and successor functions to govern the order of signal displays, whereas the continuous variables include the assigned lane flows, cycle length, and starts and durations of green for traffic movements, lanes and pedestrian crossings. A set of linear constraints is set up to ensure the feasibility and safety of the resultant optimized lane markings and signal settings. A numerical example is given to demonstrate the effectiveness of the proposed methodology. The heuristic line search is more cost-effective in terms of both optimality of solution and computing time requirement.

2. The Lane-Based Model

2.1. GENERAL NOTATION

Consider an isolated junction with N_T traffic arms and N_P pedestrian crossings, with each arm i consisting of L_i approaching lanes and E_i exit lanes, $i = 1, 2, \ldots, N_T$. The traffic lanes are numbered consecutively from the curbside, with the nearest side lane as 1 and the farthest offside lane as L_i. For traffic maneuvers, a movement, u, is defined as (i, j, k), which represents the traffic turning via lane k from arm i to arm j, and for pedestrian maneuvers, each pedestrian crossing defines a unique movement $u = 1, 2, \ldots, N_P$. For each pair of incompatible movements, u and v, a clearance time, $\omega_{u,v}$, is defined as the earliest time that movement v receives right of way after movement u terminates. Note that this clearance time is lane dependent, and its value is determined according to the geometry of the junction. Denote the set of incompatible movements as Ψ. For traffic maneuvers, a signal group is defined as (i, j), which represents the control of all movements from arm i to arm j, and for pedestrian maneuvers, each pedestrian crossing defines a unique signal group. Denote the set of incompatible signal groups as Ψ_s.

The saturation flow of a turning lane is defined by the following expression, which is generalized from that of Kimber *et al.* [17]:

$$s_{i,k} = \frac{\bar{s}_{i,k}}{1 + 1.5 \sum_{j=1}^{N_T - 1} \frac{P_{i,j,k}}{r_{i,j,k}}}, \tag{1}$$

where $s_{i,k}$ is the saturation flow of lane k in arm i, $r_{i,j,k}$ and $P_{i,j,k}$ are the radius of the turning trajectory ($= \infty$ for straight-ahead movement) and the proportion of flow from arm i to arm j via lane k, and $\bar{s}_{i,k}$ is the saturation flow of the lane if it is for straight-ahead movement only. When the lane is shared by a straight-ahead movement and a turning movement, or is used exclusively for one turning movement, the expression reduces to the formulae that were developed by Kimber *et al.* [17]. For straight-ahead movement only, the saturation flow is $\bar{s}_{i,k}$ by default. The formula in Equation (1) is derived according to the weighted average radius of the curvature for different turning movements.

2.2. CONTROL VARIABLES

In the lane-based model, the control variables can be specified as follows. Let $\Lambda = (\Lambda_b, \Lambda_c)$ be the set of control variables, where Λ_b and Λ_c are the subsets of binary and continuous variables respectively.

The subset, Λ_b, consists of the following binary variables.

Permitted movements: $\boldsymbol{\delta} = (\delta_{i,j,k}, j = 1, \ldots, N_T - 1; k = 1, \ldots, L_i; i = 1, \ldots, N_i)$

Successor functions: $\boldsymbol{\Omega} = (\Omega_{i,j,l,m}, ((i, j), (l, m)) \in \Psi_s)$

The subset, Λ_c, consists of the following continuous variables.

Allocated flows: $\mathbf{q} = (q_{i,j,k}, j = 1, \ldots, N_T - 1; k = 1, \ldots, L_i; i = 1, \ldots, N_T)$

Common flow multiplier: μ

Cycle length: ζ

Starts of green for movements: $\boldsymbol{\theta} = (\theta_{i,j}, j = 1, \ldots, N_T - 1; i = 1, \ldots, N_T$ and $j = 1; i = 1, \ldots, N_P)$

Durations of green for movements: $\boldsymbol{\varphi} = (\phi_{i,j}, j = 1, \ldots, N_T - 1; i = 1, \ldots, N_T$ and $j = 1; i = 1, \ldots, N_P)$

Starts of green for traffic lanes: $\boldsymbol{\Theta} = (\Theta_{i,k}, k = 1, \ldots, L_i; i = 1, \ldots, N_T)$

Durations of green for traffic lanes: $\boldsymbol{\Phi} = (\Phi_{i,k}, k = 1, \ldots, L_i; i = 1, \ldots, N_T)$

For the set of binary variables, $\delta_{i,j,k}$ defines the permitted movement (lane marking) on traffic lane k in arm i for $k = 1, \ldots, L_i$; $j = 1, \ldots, N_T - 1$; $i = 1, \ldots, N_T$, where $\delta_{i,j,k} = 1$ if the movement from arm i to arm j is permitted in lane k, and $\delta_{i,j,k} = 0$ otherwise. Shared lanes are allowed, for which there are more than one j with non-zero $\delta_{i,j,k}$ for traffic lane k in arm i. $\Omega_{i,j,l,m}$ is defined as a successor function [13] that governs the order of signal displays between two incompatible signal groups (i, j) and (l, m), where $\Omega_{i,j,l,m} = 0$ if the start of the green of signal group (l, m) follows that of signal group (i, j), and $\Omega_{i,j,l,m} = 1$ if the opposite is true.

For the set of continuous variables, $q_{i,j,k}$ is the allocated flow to lane k from arm i that is turning to destination arm j, for $k = 1, \ldots, L_i$; $j = 1, \ldots, N_T - 1$; $i = 1, \ldots, N_T$. The allocation of these traffic flows is based on the queuing theory, according to which the degrees of saturation on any pair of adjacent lanes with any common permitted movement must be identical [7]. This is because, at equilibrium, users that belong to the permitted movement will arrange themselves in these adjacent lanes, such that they incur the same delay no matter which lane they choose. Consequently, the degrees of saturation in these adjacent lanes are identical. However, if such a distribution of users is not feasible, this constraint will prevent the occurrence of a common permitted movement between adjacent lanes, and hence the lanes will belong to different traffic streams with separate queues. The junction is operated at cycle length $c = 1/\zeta$. The right of way of each traffic movement (irrespective of which lane it takes) or pedestrian crossing is specified by two variables, as in the group-based control method [11, 14, 7]: the start of green, $\theta_{i,j}$, and the duration of green, $\phi_{i,j}$, for all turning movements from arm i to arm j, $j = 1, \ldots, N_T - 1$; $i = 1, \ldots, N_T$ or pedestrian crossing $i = 1, \ldots, N_P$, with both expressed as a fraction of the cycle length (i.e. the actual start and duration of green are $\theta_{i,j}/\zeta$, and $\phi_{i,j}/\zeta$). For pedestrian movements, $j = 1$ only. $\Theta_{i,k}$ and $\Phi_{i,k}$ denote the start and duration of green display that is received in lane k in arm i, expressed as a fraction of the cycle length. For safety reasons, all traffic (irrespective of turning direction) in the same lane should be subject to a single set of signal settings. Therefore, we must ensure that $\Theta_{i,k} = \theta_{i,j}$ and $\Phi_{i,k} = \phi_{i,j}$ for all j and k, such that $\delta_{i,j,k} = 1$ by means of appropriate constraints (to be discussed later). μ is the common flow multiplier of all traffic flows that are entering the junction.

2.3. CONSTRAINTS

2.3.1. *Flow Conservation*

Assume that the traffic demand matrix **Q** is multiplied by a common flow multiplier μ to represent the extent to which the traffic can be increased and the junction can still perform reasonably well. With these increased demands, the flow conservation constraints can be set as follows:

$$\mu Q_{i,j} = \sum_{k=1}^{L_i} q_{i,j,k}, \quad \forall j = 1, \ldots, N_T - 1; i = 1, \ldots, N_T, \tag{2}$$

where $Q_{i,j}$ is the demand from arm i to arm j.

2.3.2. *Minimum Permitted Movement in a Lane*

Each traffic lane should possess at least one permitted turning movement to ensure that every lane which approaches the junction is utilized. This can be specified as

$$\sum_{j=1}^{N_T-1} \delta_{i,j,k} \geqslant 1, \quad \forall k = 1, \ldots, L_i; i = 1, \ldots, N_T. \tag{3}$$

2.3.3. *Maximum Permitted Movements at the Exit*

For each turning movement from arm i to arm j, the number of exit lanes on the destination arm j should always be greater than or equal to the number of approaching lanes that permit such a turning movement on arm i, i.e.

$$E_{\Gamma(i,j)} \geqslant \sum_{k=1}^{L_i} \delta_{i,j,k}, \quad \forall j = 1, \ldots, N_T - 1; i = 1, \ldots, N_T, \tag{4}$$

where $\Gamma(i, j)$ is the global arm number of the jth destination arm counting from arm i.

2.3.4. *Permitted Movements Across Adjacent Lanes*

Taking the numbering convention that is described in the previous section, for any two adjacent traffic lanes, k (left) and $k+1$ (right) in arm i, if the turning movement to arm j is permitted at lane $k+1$, then for safety reasons the turning movements to all arms, $j+1, \ldots, N_T - 1$, should be prohibited in lane k to eliminate potential conflicts within an arm. This can be specified by the following constraints:

$$1 - \delta_{i,j,k+1} \geqslant \delta_{i,m,k}, \quad \forall m = j+1, \ldots, N_T - 1; \; j = 1, \ldots, N_T - 2; \\ k = 1, \ldots, L_i - 1; \; i = 1, \ldots, N_T. \tag{5}$$

Given the binary nature of the variables, if $\delta_{i,j,k+1} = 1$, then $\delta_{i,m,k}$ must vanish for all $m = j+1, \ldots, N_T - 1$, i.e. the movement (i, m, k) is prohibited. However, if $\delta_{i,j,k+1} = 0$, then $\delta_{i,m,k}$ can take on any value of 0 or 1.

2.3.5. *Cycle Length*

Let the minimum and maximum cycle lengths in the junction be $c_{\min}$ and $c_{\max}$. Instead of using the cycle length as the control variable, we use the reciprocal of cycle length $\zeta = 1/c$. The constraints on the cycle length can now be specified as

$$\frac{1}{c_{\min}} \geqslant \zeta \geqslant \frac{1}{c_{\max}}, \tag{6}$$

which ensures that the cycle length will fall in the feasible range of $(c_{\min}, c_{\max})$.

2.3.6. *Lane Signal Settings*

For safety reasons, if a lane permits two or more shared lane movements, then they must be controlled by an identical signal indication. Consider lane k in arm i. If a turning movement j is permitted in this lane, then the following constraints can be set to satisfy the above requirement,

$$M(1 - \delta_{i,j,k}) \geqslant \Theta_{i,k} - \theta_{i,j} \geqslant -M(1 - \delta_{i,j,k}), \tag{7}$$

and

$$M(1 - \delta_{i,j,k}) \geqslant \Phi_{i,k} - \phi_{i,j} \geqslant -M(1 - \delta_{i,j,k}), \tag{8}$$

$\forall j = 1, 2, \ldots, N_T - 1$; $k = 1, \ldots, L_i$; $i = 1, \ldots, N_T$, where M is an arbitrary large positive constant. If a movement (i, j) is permitted in lane k, then we have $\delta_{i,j,k} = 1$ and hence the values on both sides of the inequalities in Equations (7) and (8) become zero. This ensures that $\Theta_{i,k} = \theta_{i,j}$ and $\Phi_{i,k} = \phi_{i,j}$ from the constraints, i.e. there are identical settings for all j that are permitted in this lane. However, if this movement is not permitted in the lane, then the constraints are ineffective because $(1 - \delta_{i,j,k})$ is equal to unity, and hence the signal settings could be different.

2.3.7. *Start of Green*

As the signal settings at the junction are cyclical, the start of green variables can be quite arbitrary as long as they satisfy other constraints in the problem. However, for convenience, all of the starts of green variables are confined within the range of (0, 1), i.e.

$$1 \geqslant \theta_{i,j} \geqslant 0, \quad \forall j = 1, \ldots, N_T - 1;\ i = 1, \ldots, N_T, \tag{9}$$

for traffic movements and

$$1 \geqslant \theta_{i,1} \geqslant 0, \quad \forall i = 1, \ldots, N_P, \tag{10}$$

for pedestrian crossings.

2.3.8. *Duration of Green*

The duration of green for a traffic lane or a pedestrian crossing is subject to a minimum value. These constraints can be set as

$$1 \geqslant \phi_{i,j} \geqslant g_{i,j}\zeta, \quad \forall j = 1, \ldots, N_T - 1;\ i = 1, \ldots, N_T, \tag{11}$$

for traffic movements and

$$1 \geqslant \phi_{i,1} \geqslant g_{i,1}\zeta, \quad \forall i = 1, \ldots, N_P, \tag{12}$$

for pedestrian crossings, where $g_{i,j}$ is the minimum duration of green for a signal group (for traffic or pedestrian crossings) for turning movements from arm i to arm j, $j = 1, \ldots, N_T - 1$; $i = 1, \ldots, N_T$ or pedestrian crossings $i = 1, \ldots, N_P$ $(j = 1$ only).

2.3.9. *Order of Signal Displays*

Any two signal groups, (i, j) and (l, m), are said to be incompatible if a movement that is controlled by (i, j) is incompatible with a movement that is controlled by (l, m). The set of incompatible signal groups Ψ_s can therefore be derived from Ψ, which is the set of incompatible movements. For any two incompatible signal groups (i, j) and (l, m) in Ψ_s, the order of signal displays is governed by a successor function [13], $\Omega_{i,j,l,m}$, where $\Omega_{i,j,l,m} = 0$ if the start of green of signal group (l, m) follows that of signal group (i, j), and $= 1$ if the opposite is true. Therefore, the following constraints can be set for the successor functions,

$$\Omega_{i,j,l,m} + \Omega_{l,m,i,j} = 1, \quad \forall((i, j), (l, m)) \in \Psi_s. \tag{13}$$

2.3.10. *Clearance Time*

For any pair of incompatible movements, the clearance time constraints are needed only when both movements are permitted. If both are traffic movements that are defined as $u = (i, j, k)$ and $v = (l, m, n)$, then the following constraints can be set to ensure satisfaction of clearance time requirements:

$$\begin{aligned}&\theta_{l,m} + \Omega_{i,j,l,m} + M(2 - \delta_{i,j,k} - \delta_{l,m,n})\\ &\quad \geqslant \theta_{i,j} + \phi_{i,j} + \omega_{u,v}\zeta, \quad \forall(u, v) \in \Psi,\end{aligned} \tag{14}$$

where M is an arbitrary large positive constant. These constraints are effective only when the two incompatible movements are permitted, i.e. $\delta_{i,j,k} = \delta_{l,m,n} = 1$. If $u = (i, j, k)$ is a traffic movement and $v = (l, 1)$ is a pedestrian crossing, then the above constraints can be modified to

$$\theta_{l,1} + \Omega_{i,j,l,1} + M(1 - \delta_{i,j,k}) \geqslant \theta_{i,j} + \phi_{i,j} + \omega_{u,v}\zeta, \quad \forall(u, v) \in \Psi. \tag{15}$$

However, if $u = (i, 1)$ is a pedestrian crossing and $v = (l, m, n)$ is a traffic movement, then this formula becomes

$$\theta_{l,m} + \Omega_{i,1,l,m} + M(1 - \delta_{l,m,n}) \geqslant \theta_{i,1} + \phi_{i,1} + \omega_{u,v}\zeta, \quad \forall(u, v) \in \Psi. \tag{16}$$

2.3.11. *Prohibited Movement*

If a traffic movement is prohibited in a particular traffic lane, then the allocated flow of the effective turning movement should vanish in this lane, which can be set as

$$\begin{aligned}&M\delta_{i,j,k} \geqslant q_{i,j,k} \geqslant 0,\\ &\quad \forall k = 1, \ldots, L_i;\ j = 1, \ldots, N_T - 1;\ i = 1, \ldots, N_T,\end{aligned} \tag{17}$$

where M is an arbitrary large positive constant. If $\delta_{i,j,k} = 0$, i.e. the movement is prohibited, then the allocated flow must vanish. However, if $\delta_{i,j,k} = 1$, i.e. the movement is permitted, then the allocated flow can take on any non-negative value, as long as it satisfies the flow conservation in (2).

2.3.12. *Flow Factor*

As the allocation of traffic flow is based on the queuing theory, the degrees of saturation on a pair of adjacent lanes with at least one common permitted movement must be identical. Moreover, from the constraints that were set in Section 2.3.6, the signal settings on this pair of adjacent lanes must be identical. Therefore, to ensure identical degrees of saturation, it suffices to equalize the flow factors (which are defined as the total allocated flow divided by the saturation flow) of these adjacent lanes. Let $y_{i,k}$ be the flow factor on lane k in arm i, which can be expressed as

$$y_{i,k} = \frac{\sum_{j=1,\ldots,N_T-1} q_{i,j,k}}{s_{i,k}}, \quad \forall k = 1, \ldots, L_i;\ i = 1, \ldots, N_T. \tag{18}$$

Because

$$\mathrm{P}_{i,j,k} = \frac{q_{i,j,k}}{\sum_{m=1,\ldots,N_T-1} q_{i,m,k}},$$
$$\forall j = 1, \ldots, N_T - 1;\; k = 1, \ldots, L_i;\; i = 1, \ldots, N_T, \tag{19}$$

we can show that

$$y_{i,k} = \frac{1}{\bar{s}_{i,k}} \sum_{j=1,\ldots,N_T-1} \left(1 + \frac{1.5}{r_{i,j,k}}\right) q_{i,j,k},$$
$$\forall k = 1, \ldots, L_i;\; i = 1, \ldots, N_T. \tag{20}$$

Therefore, the following set of constraints can be used to enforce the equalized flow factors,

$$\begin{aligned} &M(2 - \delta_{i,j,k} - \delta_{i,j,k+1}) \\ &\quad \geqslant \frac{1}{\bar{s}_{i,k}} \sum_{j=1,\ldots,N_T-1} \left(1 + \frac{1.5}{r_{i,j,k}}\right) q_{i,j,k} - \\ &\quad - \frac{1}{\bar{s}_{i,k+1}} \sum_{j=1,\ldots,N_T-1} \left(1 + \frac{1.5}{r_{i,j,k+1}}\right) q_{i,j,k+1} \\ &\quad \geqslant -M(2 - \delta_{i,j,k} - \delta_{i,j,k+1}), \quad \forall k = 1, \ldots, L_i - 1; i = 1, \ldots, N_T, \end{aligned} \tag{21}$$

where M is an arbitrary large positive constant, and k (left) and $k+1$ (right) are adjacent lanes in arm i. Again, these constraints are effective only when $\delta_{i,j,k} = \delta_{i,j,k+1} = 1$.

2.3.13. *Maximum Acceptable Degree of Saturation*

Let $p_{i,k}$ be the maximum degree of saturation at lane k in arm i. For traffic lane k from arm i, the degree of saturation can be expressed as

$$\rho_{i,k} = \frac{y_{i,k}}{\Phi_{i,k} + e\zeta}, \quad \forall k = 1, \ldots, L_i;\; i = 1, \ldots, N_T, \tag{22}$$

where $\rho_{i,k}$ and e are the degree of saturation at lane k in arm i and the difference between actual and effective greens (measured in time units, which are usually taken as 1 second). From Equation (20), the following constraints can be set to ensure that the degree of saturation is below the maximum acceptable limit:

$$\Phi_{i,k} + e\zeta \geqslant \frac{1}{p_{i,k}\bar{s}_{i,k}} \sum_{j=1,\ldots,N_T-1} \left(1 + \frac{1.5}{r_{i,j,k}}\right) q_{i,j,k},$$
$$\forall k = 1, \ldots, L_i;\; i = 1, \ldots, N_T. \tag{23}$$

2.3.14. *Other Signal Group Constraints*

There may be some practicable constraints in setting up the relative timing of starts and ends of greens for different signal groups. These constraints can be set as follows. Let $z_{i,j,l,m}$ be the relative time (for the starts or ends of greens) that is required to appear when measured from signal group (i, j) to group (l, m). For the starts of green,

$$\theta_{i,j} + z_{i,j,l,m} = \theta_{l,m}, \tag{24}$$

and for the ends of green,

$$\theta_{i,j} + \phi_{i,j} + z_{i,j,l,m} = \theta_{l,m} + \phi_{l,m}. \tag{25}$$

These constraints are mainly used to constrain two signal groups to start and/or end simultaneously (i.e. $z_{i,j,l,m} = 0$) for the practicability of signal timing at a junction.

3. Criteria for Optimization in Isolated Junctions

For isolated signal control junctions, there are three common criteria for optimizing the signal settings: capacity maximization, cycle length minimization, and delay minimization. Alternative signal settings can be generated to fulfill specific operating requirements if different optimizing objectives are chosen.

3.1. CAPACITY MAXIMIZATION

The capacity maximization problem can be effectively formulated as a BMILP and solved by standard branch-and-bound routines. Based on the assumption that the traffic flows for the turning movements in the junction will increase in proportion to the demand matrix [4, 11, 31, 34], the problem becomes one of determining the largest common multiplier $\mu_{\max}$ that can be accommodated without violating any of the constraints that are specified in the constraint section. A value of $\mu_{\max} < 1$ then indicates that the junction is overloaded by $100(1 - \mu_{\max})$ percent, and a value of $\mu_{\max} > 1$ indicates a reserve capacity of $100(\mu_{\max} - 1)$ percent. The lane-based method for determining capacity-maximizing settings can be found in the work of Wong *et al.* [27] and Wong and Wong [26].

The problem can be formulated as the mathematical program below.

$$\underset{\Lambda=(\Lambda_b,\Lambda_c)}{\text{Maximize}} \quad \mu \tag{26}$$

subject to the linear constraints in (2)–(17), (21), and (23)–(25).

3.2. CYCLE LENGTH MINIMIZATION

Another important traffic signal setting is the question of how small the cycle length can be maintained to handle the existing traffic flow pattern, i.e. $\mu = 1$, and geometric configuration. The resultant minimum cycle length always refers to critical cycle length. This is particularly useful when the junction is located within an area traffic control system [7], where the cycle length minimization settings provide a useful indication of the feasible cycle length range. The problem becomes one of maximizing the reciprocal of cycle length ζ with $\mu = 1$, subject to the set of constraints that was specified in Section 3.1. This can be formulated as the following BMILP:

$$\underset{\Lambda=(\Lambda_b,\Lambda_c)}{\text{Maximize}} \quad \zeta \tag{27}$$

subject to the linear constraints in (2)–(17), (21), and (23)–(25) and $\mu = 1$.

3.3. DELAY MINIMIZATION

Although the lane-based optimization method has been successfully formulated and implemented for the determination of capacity-maximizing and cycle-length-minimizing settings, a problem remains that such settings may be sub-optimal as far as total junction delay is concerned. Moreover, it can be shown that the optimized cycle length is always pushed to the maximum limit when the junction capacity is maximized. Hence, it is generally difficult to obtain the optimal signal settings for minimizing total delay at the junction.

In general, the delay function D that is used as the objective for optimization is non-linear. Therefore, the problem has to be formulated as the BMINLP that is shown below.

$$\underset{\Lambda=(\Lambda_b,\Lambda_c)}{\text{Minimize}} \quad D \tag{28}$$

subject to the linear constraints in (2)–(17), (21), and (23)–(25) and $\mu = 1$, where D defines the total delay of the junction. Gallivan [10] found that Webster's two-term delay expression is convex in the variables of the reciprocal of cycle length and effective green times. He applied a piecewise linearization of the delay expression to break down the original BMINLP problem into numerous sub-problems in which the standard branch-and-bound technique could attain a global solution. However, the solution quality depends on the number of segments that are considered in the analysis [15]. In the present lane-based formulation, the lane permitted movements and the assigned lane flows are introduced as two new sets of variables, and the convexity of the delay expression depends not only on the reciprocal of cycle length and effective green times, but also the distribution patterns of the assigned lane flows. Unfortunately, however, the convexity of the delay function will no longer hold, even if Webster's two-term delay expression is adopted.

4. Solution Algorithms for Delay Optimization

In this section, two solution methods to solve the lane-based delay minimization problem are proposed. As mentioned in the previous section, a piecewise linearization approach of the objective function of a BMINLP is effective only when the non-linear delay function is convex in the feasible region. Unfortunately, however, it can be proven that, through examining either the eigenvalues or the principal minors of the Hessian matrix, that Webster's delay expression is non-convex if the assigned flows are also considered as explicit control variables together with the reciprocal of cycle length and effective green times. Due to the complicated multi-dimension solution space and the non-convexity of the objective delay function, the piecewise linearization approach is not applicable to the present problem. A classical cutting plane algorithm and a heuristic line search algorithm are proposed and discussed in the following sections to solve the BMINLP problem.

4.1. A CUTTING PLANE ALGORITHM

There are several approaches to solving a general Mixed-Integer-Non-Linear Program (MINLP) problem, such as the generalized benders decomposition (GBD) method [12], the outer approximation (OA) method [9], the linear programming/ non-linear programming (LP/NLP) based branch-and-bound method [1], and Kelley's cutting plane (CP) method [16, 19]. One of the advantages of the cutting plane method is that it obviates the need of solving NLP sub-problems in the solution process, whereas other methods have to solve both NLP and Mixed-Integer-Linear-Program (MILP) sub-problems in an iterative manner. Therefore, the solution quality that is obtained from the cutting plane methods does not depend on the availability of a reliable NLP solver. Moreover, a carefully designed cutting plane algorithm can ensure a global convergence for a class of pseudo-convex problems [22], and offers promising solution characteristics for general non-convex MINLP problems [19]. In the cutting plane approach, the problem size of the MILP increases gradually until a sufficient number of hyper-planes (in the form of linear constraints) are constructed to replicate the solution space as it was represented by the corresponding non-linear constraints. The methodology is quite similar to the piecewise linearization approach discussed earlier, but the cutting plane method creates hyper-planes in a guided manner at the time that they are needed, and thus the resultant computational efforts can be considerably reduced. Each hyper-plane can be regarded as a cutting plane if it appears in the form of an inequality constraint which is also responsible for reductions in the solution region such that the approximate solution can somehow be pushed toward the solution point that is restrained by the original (nonlinear) constraints. The underlying philosophy of this contraction in the search space is similar to the bound tightening updates in the branch-and-bound approaches.

As discussed in Section 3.3, the delay optimization problem is to minimize a non-linear objective function that is subject to a set of linear constraints. For standardization, the problem can be rewritten with a very simple objective (i.e. with a single variable only), as it is for the cases of capacity maximization and cycle length minimization that are subject to a set of constraints. This reformation is made by introducing the auxiliary continuous variable λ as the objective function, together with the following additional non-linear constraint:

$$D - \lambda \leqslant 0. \tag{29}$$

This non-linear constraint is linearized by a first-order Taylor's series expansion to form a linear function L^{κ}, where

$$\begin{aligned} L^{\kappa} = \Bigg[\Bigg(\sum_{i=1}^{N_T} \sum_{k=1}^{L_i} D_{i,k}^{\kappa} - \lambda^{\kappa} \Bigg) + \\ + \alpha^{\kappa} \Bigg(\sum_{i=1}^{N_T} \sum_{k=1}^{L_i} \Bigg(\left(\frac{\partial D_{i,k}}{\partial \gamma_{i,k}} \right)^{\kappa} (\gamma_{i,k} - \gamma_{i,k}^{\kappa}) + \left(\frac{\partial D_{i,k}}{\partial \zeta} \right)^{\kappa} (\zeta - \zeta^{\kappa}) + \\ + \sum_{j=1}^{N_T - 1} \left(\frac{\partial D_{i,k}}{\partial q_{i,j,k}} \right)^{\kappa} (q_{i,j,k} - q_{i,j,k}^{\kappa}) \Bigg) - (\lambda - \lambda^{\kappa}) \Bigg) \Bigg] \leqslant 0, \end{aligned} \tag{30}$$

in which κ specifies the κth linearization point, $\gamma_{i,k} = \Phi_{i,k} + e\zeta$, and α^{κ} is a parameter that controls the degree of a cut in the feasible region. The lane-based delay minimization problem becomes

$$\underset{(\Lambda_b, \Lambda_c, \lambda)}{\text{Minimize}} \quad \lambda \tag{31}$$

subject to constraints in (2)–(17), (21), (23)–(25), $\mu = 1$, and all of the constraints (30) that were generated in previous iterations.

For each additional constraint (30), as the value of α^{κ} increases, the part of the feasible region to be cut away decreases. The solution process then iteratively adds increasingly more cutting planes to the problem, and each current solution point is the location for generating the next cutting plane (linear approximation). To avoid a deep cut in the feasible region, which may eliminate the global solution or even make the approximated BMILP problem infeasible, an update for the parameter α^{κ} can take place before the insertion of a new cutting plane. For the κth linearization, there should be a sufficiently large α^{κ} at each iteration point to ensure that the linearized function always underestimates the original non-linear function. If all of the linearized functions are valid underestimators, then the approximate solution region becomes a valid outer approximation of the original feasible region for the problem. Details of updating α^{κ} in the cutting plane algorithm can be found in the work of Westerlund and Porn [25].

A stringent termination criterion for the cutting plane algorithm can be established as

$$\tilde{D} - \tilde{\lambda} \leqslant 0, \tag{32}$$

where $\tilde{D}$ and $\tilde{\lambda}$ are the numerical values at the solution point. However, it is usually a very time consuming process to achieve the above absolute convergence. For practical considerations, a weaker stopping criterion is specified as

$$(\tilde{D} - \tilde{\lambda})/\tilde{D} \leqslant \varepsilon, \tag{33}$$

where ε defines a pre-specified tolerance (a small positive value).

4.2. A HEURISTIC LINE SEARCH ALGORITHM

While the cutting plane algorithm that is presented in Section 4.1 is a general mathematical algorithm for the BMINLP program, it does not take into account some of the physical properties of the problem concerned that can be used to develop a more efficient solution algorithm. First, for the delay optimization problem of signal settings, the cycle length variable affects the optimization results, because it is generally true that a shorter cycle time will lead to a larger proportion of wasteful intergreen times between conflicting movements, whereas a longer cycle length means a longer waiting time when a vehicle arrives at the beginning of a red period. Second, due to the general property of a delay function, the total junction delay that is incurred with the capacity-maximizing settings at a particular cycle length is reasonably close to that with the delay-minimizing settings at the same cycle length, i.e. though sub-optimal in nature, the solution of capacity maximization serve as a good starting point in the search for delay-maximizing settings [4, 7]. Third, the optimized lane marking patterns for capacity maximization and delay minimization are usually quite similar at the same cycle length.

Based on the above observations, a heuristic line search algorithm that makes use of the lane-based method of capacity maximization [26] and the group-based optimization technique [14] is developed for the determination of optimized lane permitted movements and delay-minimizing settings of a signalized junction. When the lane permitted movements and assigned lane flows are fixed, the traffic can be easily grouped into a set of traffic streams at the junction. Moreover, when the successor functions, which are used to govern the order of signal displays, are also fixed, the problem becomes a standard group-based minimization problem for junction delay [14]. The former Transport Studies Group of the University College of London developed and implemented an efficient gradient-based solution algorithm in a commercial computer package called SIGSIGN (SIGnal deSIGN) [20].

The following modules are defined for the development of optimization heuristics: LB1(c; $\boldsymbol{\delta}$, $\mathbf{q}$, $\boldsymbol{\Omega}$, μ) is the abstract form of the lane-based capacity maximization module, where the input is the fixed cycle length c and the output includes the sets of optimized permitted movement function on lanes $\boldsymbol{\delta}$, assigned lane flows $\mathbf{q}$, successor functions $\boldsymbol{\Omega}$, and maximized flow multiplier μ. LB2(μ; $\boldsymbol{\delta}$, $\mathbf{q}$, $\boldsymbol{\Omega}$, c) is the abstract form of the lane-based cycle length minimization module, where the input is the flow multiplier μ, and the output includes the sets of optimized permitted movement function on lanes $\boldsymbol{\delta}$, assigned lane flows $\mathbf{q}$, successor functions $\boldsymbol{\Omega}$, and minimized cycle length c. GB$(c, \boldsymbol{\delta}, \mathbf{q}, \boldsymbol{\Omega}; \boldsymbol{\psi}, D)$ is the abstract form of the group-based delay minimization module with specified cycle length, where the input is the vector $(\boldsymbol{\delta}, \mathbf{q}, \boldsymbol{\Omega})$ that is obtained from the lane-based module and a specified cycle length c, and the output is the set of optimized starts and durations of signal groups $\boldsymbol{\psi}$ and the total junction delay D. Based on these four modules, the optimization heuristics are described as follows.

The lane-based method is first used to generate the feasible cycle length range for subsequent delay minimization analysis. The optimized lane permitted movements and assigned lane flows, together with the successor functions, at maximum cycle length $c_{\max}$ can be obtained by LB1$(c = c_{\max}; \boldsymbol{\delta}_{\mathrm{U}}, \mathbf{q}_{\mathrm{U}}, \boldsymbol{\Omega}_{\mathrm{U}}, \mu_{\mathrm{U}})$, where $(\boldsymbol{\delta}_{\mathrm{U}}, \mathbf{q}_{\mathrm{U}}, \boldsymbol{\Omega}_{\mathrm{U}}, \mu_{\mathrm{U}})$ are the corresponding optimized results that are associated with maximum cycle length. This forms the upper limit of the feasible cycle length range. If $\mu_{\mathrm{U}} < 1$, then the junction is overloaded and the delay-minimizing settings cannot be determined using Webster's delay expression, unless the sheared delay formula is used for the delay evaluation [39]. In such a case, the feasible cycle length range is reduced to an epoch at $c_{\max}$. However, if $\mu_{\mathrm{U}} \geqslant 1$, then the lower limit of the feasible cycle length range can be obtained by the cycle length minimization problem LB2$(\mu = 1; \boldsymbol{\delta}_{\mathrm{L}}, \mathbf{q}_{\mathrm{L}}, \boldsymbol{\Omega}_{\mathrm{L}}, c_{\min})$, where $(\boldsymbol{\delta}_{\mathrm{L}}, \mathbf{q}_{\mathrm{L}}, \boldsymbol{\Omega}_{\mathrm{L}})$ are the corresponding sets of optimized results associated with minimum cycle length $c_{\min}$. The minimum junction delays at the lower and upper limits of the feasible cycle length can be obtained by GB$(c = c_{\min}, \boldsymbol{\delta}_{\mathrm{L}}, \mathbf{q}_{\mathrm{L}}, \boldsymbol{\Omega}_{\mathrm{L}}; \boldsymbol{\psi}_{\mathrm{L}}, D_{\mathrm{L}})$ and GB$(c = c_{\max}, \boldsymbol{\delta}_{\mathrm{U}}, \mathbf{q}_{\mathrm{U}}, \boldsymbol{\Omega}_{\mathrm{U}}; \boldsymbol{\psi}_{\mathrm{U}}, D_{\mathrm{U}})$, respectively, where D_{L} and D_{U} are the minimum delays at $c_{\min}$ and $c_{\max}$ respectively. It is anticipated that an optimal cycle length with respect to the minimum total delay is located at a point within the range of feasible cycle length. Wong *et al.* [28] proposed a golden-section line search algorithm to locate the optimal cycle length together with the optimal signal settings and lane configurations such that the total junction delay is optimized. As the delay function is non-convex with respect to the cycle length in the lane-based formulation, a true optimal point can possibly be omitted during the golden-section search. Therefore, a more exhaustive line search scheme is adopted to ensure the resultant solution quality.

In the search process, a uniform step size S is used to govern the resolution of choices in the feasible cycle length range. For the tth interior point evaluation, the initial cycle length is taken as $c = c_{\max} - tS \in (c_{\min}, c_{\max})$. The lane-based module, LB1$(c; \tilde{\boldsymbol{\delta}}, \tilde{\mathbf{q}}, \tilde{\boldsymbol{\Omega}}, \tilde{\mu})$, is first used to determine the sets of optimized lane permitted movements, assigned lane flows, and successor functions. All of these lane-based module results are then entered into the group-based model, GB$(\tilde{\boldsymbol{\delta}}, \tilde{\mathbf{q}}, \tilde{\boldsymbol{\Omega}}; \tilde{c}, \tilde{\boldsymbol{\psi}}, \tilde{D})$, as fixed inputs to optimize the overall junction delay $\tilde{D}$. The outputs also include the optimized cycle length $\tilde{c}$ and signal settings $\tilde{\boldsymbol{\psi}}$. The evaluation procedure terminates when $c < c_{\min}$. The minimized delay is then taken as the smallest delay value among all of the evaluation cases with different initial cycle lengths. A flow chart that summarizes the algorithm is shown in Figure 1.

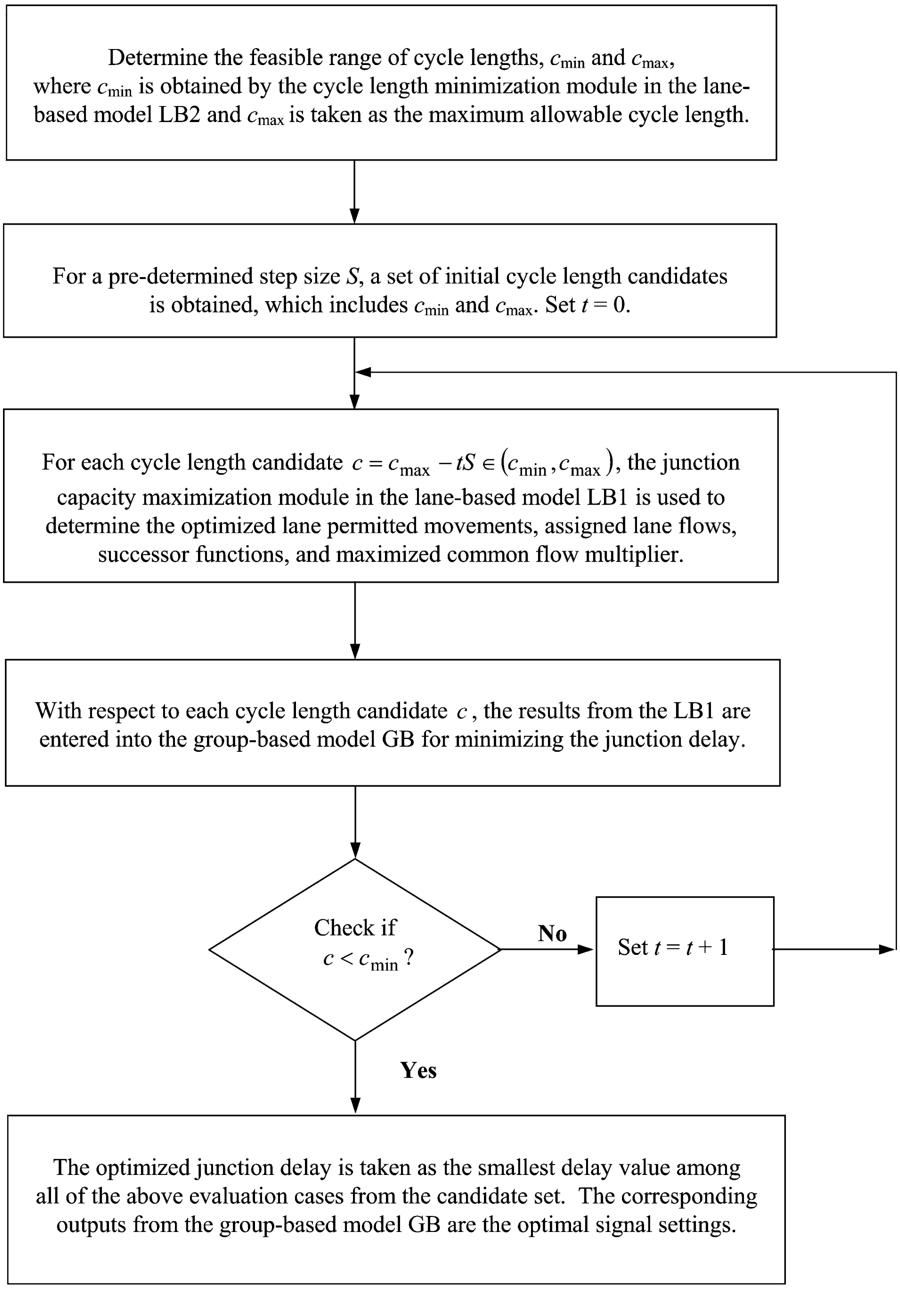

Determine the feasible range of cycle lengths, c_{min} and c_{max}, where c_{min} is obtained by the cycle length minimization module in the lane-based model LB2 and c_{max} is taken as the maximum allowable cycle length.
For a pre-determined step size S, a set of initial cycle length candidates is obtained, which includes c_{min} and c_{max}. Set $t = 0$.
For each cycle length candidate $c = c_{\mathrm{max}} - tS \in (c_{\mathrm{min}}, c_{\mathrm{max}})$, the junction capacity maximization module in the lane-based model LB1 is used to determine the optimized lane permitted movements, assigned lane flows, successor functions, and maximized common flow multiplier.
With respect to each cycle length candidate c, the results from the LB1 are entered into the group-based model GB for minimizing the junction delay.
Check if $c < c_{\mathrm{min}}$?
No
Set $t = t + 1$
Yes
The optimized junction delay is taken as the smallest delay value among all of the above evaluation cases from the candidate set. The corresponding outputs from the group-based model GB are the optimal signal settings.

Figure 1. A flow chart for the heuristic line search algorithm.

5. Numerical Example

The following steady-state delay formula, based on random arrivals and regular departure patterns, is used [24, 3]:

$$D_{i,k} = \frac{9}{10}\left(\frac{\sum_j q_{i,j,k}(1-\gamma_{i,k})^2}{2\zeta(1-y_{i,k})} + \frac{(y_{i,k}/\gamma_{i,k})^2}{2(1-y_{i,k}/\gamma_{i,k})}\right), \tag{34}$$

where $D_{i,k}$ is the rate of delay for the traffic lane k on arm i of a signal junction. The total rate of delay of the junction, D, is the sum of the delays for all traffic lanes on all arms $D = \sum_{i=1}^{N_T}\sum_{k=1}^{L_i} D_{i,k}$. This formula is widely used for the estimation of junction delay. The methodology is also applicable for other delay formulae such as the sheared delay expressions [39].

Consider a four-arm junction with four traffic lanes on each arm, as shown in Figure 2. The number of exit lanes is equal to the number of approaching lanes, except for Arm 4, where there is only one exit lane for all cases. Two pedestrian crossings are located at Arm 3. The maximum cycle length is set at 120 seconds, and the maximum acceptable degree of saturation is 90% in all lanes. The minimum greens are 5 and 20 seconds for traffic and pedestrian movements, respectively. The traffic demands are given in Table I. The required clearance time for any

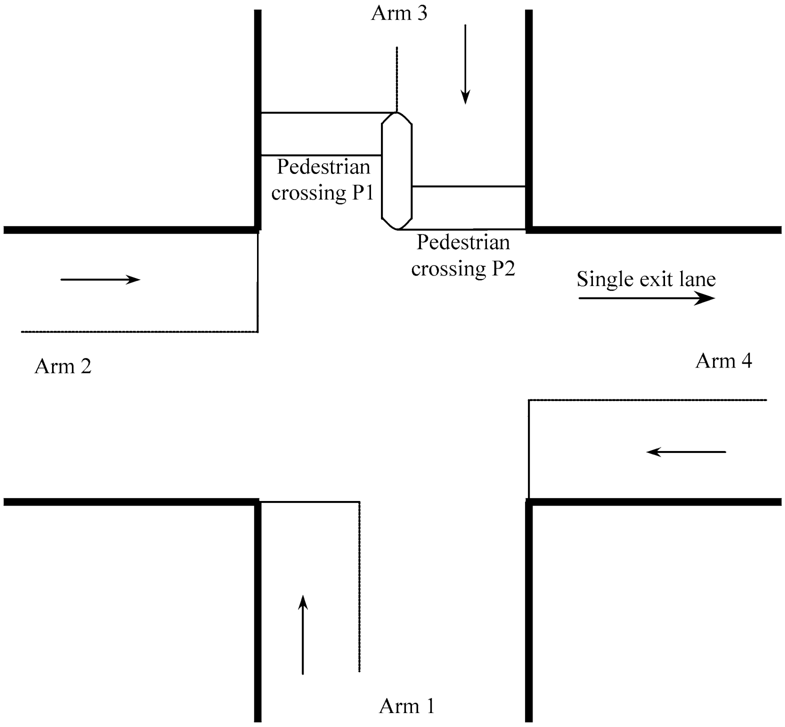
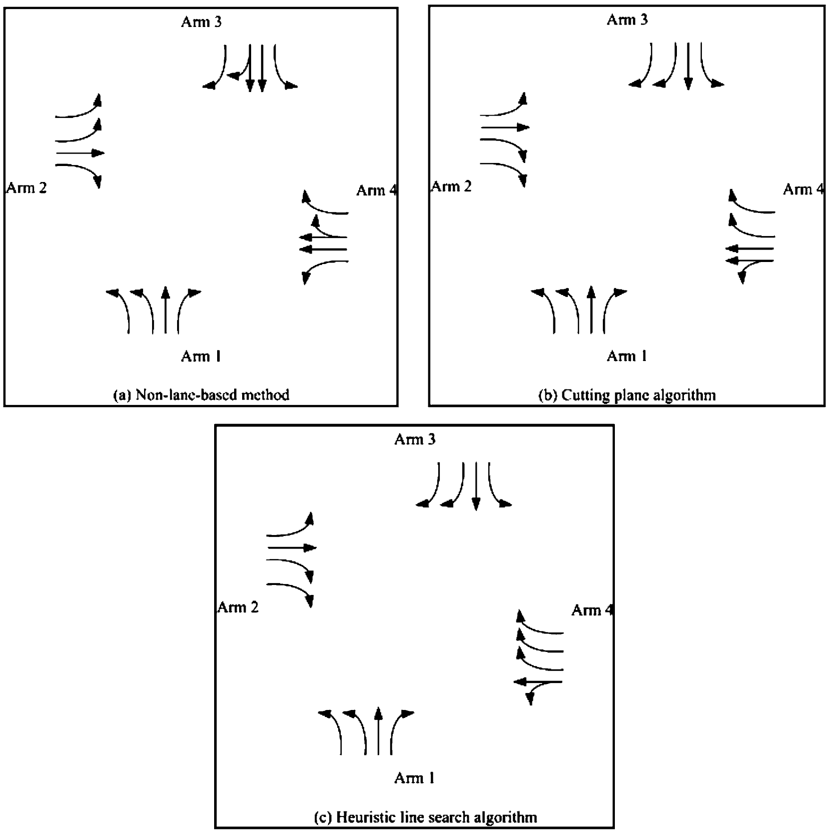


Figure 2. The layout of the four-arm example junction.

two conflicting movements (including both traffic and pedestrian movements) is 6 seconds. A two-second reduction in the clearance time is set for the following conflicting pairs: a traffic movement following a pedestrian movement or the pair of traffic and pedestrian movements that belong to the same approach. All left-

Table I. The traffic demand for the example junction

Traffic demand in veh/h		To arm			
		1	2	3	4
	1	–	500	200	100
From arm	2	100	–	100	500
	3	300	300	–	300
	4	100	400	400	–



(a) Non-lane-based method

(b) Cutting plane algorithm

(c) Heuristic line search algorithm

Figure 3. Lane markings for the example problems.

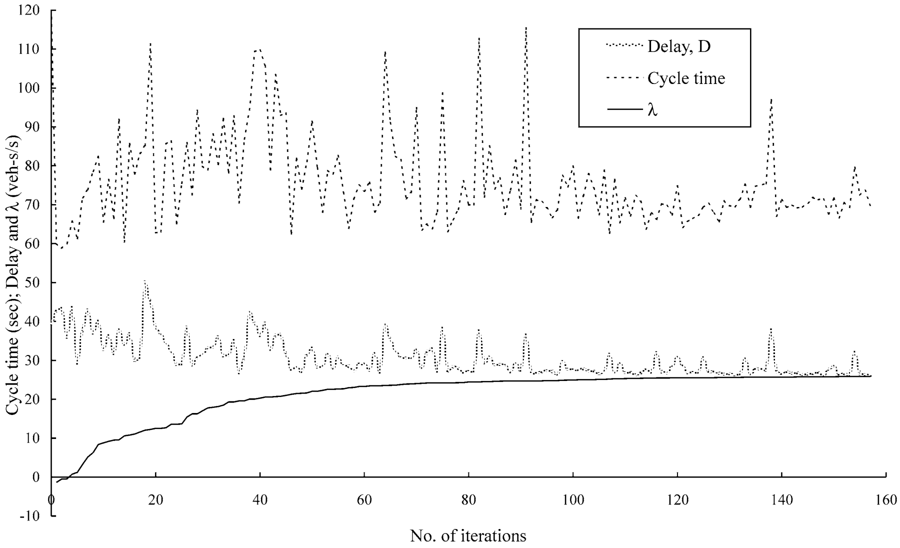


Figure 4. Convergence characteristics of the cutting plane algorithm.

turning and straight-ahead timings are set to end at the same time for all cases (which is a usual practice for safety concerns).

For comparison purpose, we first conduct a non-lane-based analysis for a prespecified set of lane markings that are designed according to the conventional traffic engineering principles. In this example, the lane markings for the non-lane-based analysis are shown in Figure 3(a). There is only one permitted movement from each of the Arms 1, 2, and 3 to Arm 4, and the numbers of permitted movements on each approach were allocated roughly in proportional to the turning flow values on the approach. Note that these lane markings will not change during the optimization process of the signal plan. The software SIGSIGN was used to optimize the junction delay. Up to 20 stage sequences were first generated, and then the group-based signal settings for each of them were optimized individually. The signal plan that corresponds to the smallest optimized delay for all of these cases was considered as the optimized signal plan. In this example, the optimized junction delay is 28.70 veh-s/s and the corresponding cycle length is 75.5 seconds.

For the heuristic line search algorithm, we use 2 seconds as the incremental time step for the initial cycle length. For the cutting plane algorithm, we take the solution of capacity maximization as the starting point for linearization. Set all initial α's = 1.0, and a multiplication factor 10.0 is applied to update the α's each time there is an invalid underestimator [22]. Instead of using a very small positive number to specify the convergence criterion, a more practical stopping requirement $(\tilde{D} - \tilde{\lambda})/\tilde{D} < 1\%$ is adopted for the numerical example. In this example, the minimum cycle time from the cycle time minimization module is 57.8 seconds. The convergence characteristics of the cutting plane algorithm and heuristic line search algorithm are shown in Figures 4 and 5, respectively.

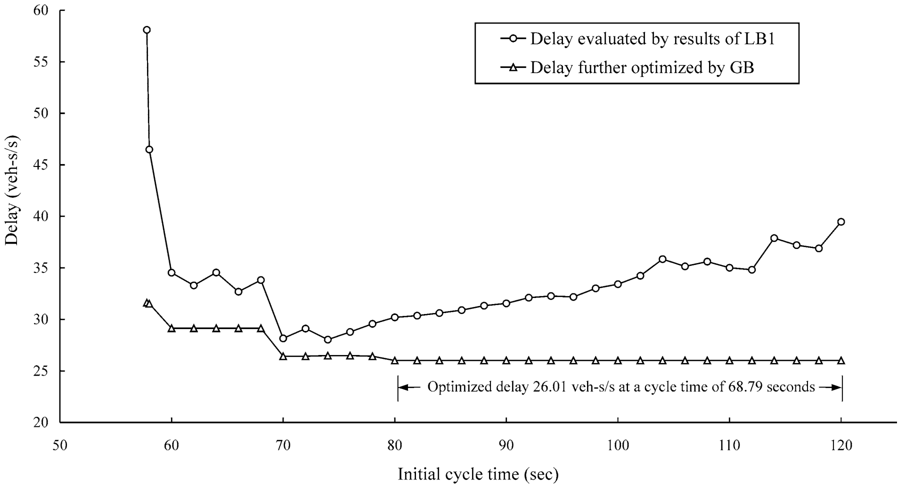


Figure 5. Convergence characteristics of the heuristic line search algorithm.

For the convergence pattern of the cutting plane algorithm, as shown in Figure 4, the top curve describes the changes in cycle length; the middle curve gives the true delays; and the bottom curve represents the auxiliary variable λ. The abscissa-axis corresponds to the iteration number in the algorithm, which is also equal to the number of cutting planes that add up to that iteration. In the example, the oscillation pattern of the cycle length in the solution process occurs as follows. In the first 90 iterations, a wide feasible cycle length range is explored between the upper bound of 120 seconds and the lower bound of around 58 seconds. Although the lower bound is not specified as an exogenous constraint, it is implicitly embedded in the set of constraints of the mathematical program. The search for the cycle length exhibits an obvious reduction and an intensive search within a narrower range of cycle time takes place after the 90th iteration. This may contribute to the fact that a considerable feasible solution region is cut away if a sufficient number of cutting planes are constructed and added to the solution process. For measuring the effectiveness of the cutting plane algorithm, we use the gap between the delay that is associated with the actual set of constraints (including the non-linear constraint) and that with the set of linear constraints. This gap is measured by the difference $\tilde{D} - \tilde{\lambda}$. The gap is also shown graphically as the physical separation between the middle and bottom curves in Figure 4. The gap is steadily reduced with the increase in the number of cutting planes, from a value of over 100% discrepancy to that of less than 1%, which is achieved at the 156th iteration. The optimized delay from the cutting plane algorithm is 26.10 veh-s/s, which is 9.1% lower than that from the non-lane-based method. The corresponding optimized cycle length is 71.4 seconds. The number of BMILP evaluations, which is a time-consuming process in the algorithm, is 156, and the total computing time is about 68 hours.

Figure 5 shows the results of the heuristic line search algorithm with a 2-second step size. It takes 33 evaluations in the feasible cycle length range. The upper curve shows the total delay of the junction, which is directly evaluated at the capacity maximization settings that are obtained from the lane-based model LB1, whereas the lower curve shows the total delay that is obtained by further optimization with the group-based optimization module GB. The group-based module that is used in the example calculation is from SIGSIGN, which takes the optimization results from the lane-based model and further optimizes the signal timings so that the overall junction delay is minimized. Hence, the delays that are obtained from the group-based module are consistently lower than those from the lane-based capacity maximization model. The optimized delay is 26.01 veh-s/s, which is 9.4% lower than that from the non-lane-based method. The optimized cycle length of 68.8 seconds. From the lower curve, there are two sudden drops in the delay at the cycle times of 68.0 and 58.0 seconds due to changes of the lane permitted movement patterns (different lane markings). This reveals that lane permitted movements are important variables in the determination of optimal delay settings for a signalized junction. The number of BMILP evaluations is 33, and the total computing time is about 40 minutes.

The detailed results of the two solution algorithms, including the lane permitted movements, assigned lane flows, and the signal settings are given in Tables II and III. Column (1) states the origin arms and column (2) specifies the traffic lanes. Columns (3) to (6) denote the destination arms. The assigned lane flows can then be distributed accordingly. If the assigned lane flows are summed vertically for each arm, then they return to the input demand flows. If they are summed horizontally for each traffic lane, then they give the total lane flows in column (7). The resultant turning proportions are collected in column (8). Lane saturation flows that take the turning proportions into consideration are then entered in column (9). The flow factor (ratio) that is given in column (10) can easily be evaluated by dividing the sum of lane flows by the saturation flows. With the effective green times in column (11) optimized from the models, the degree of saturation for each traffic lane can be deduced, as shown in column (12). The green start times are also given in column (13). The total delays, based on Webster's delay function, are calculated and put in the last column of the tables. Only the starts and durations of green are reported for the two pedestrian movements because they do not constitute any delay at the junction. The green durations are all longer than or equal to 20 seconds, which is the minimum bound for these example calculations. It can also be verified that the degrees of saturation for all traffic lanes are within the specified upper limit of 90%. Furthermore, the lane flow factors and the signal settings are all identical if these lanes contain the same turning movement. If a traffic lane involves more than one turning flow, then there is a share lane movement assigned to that particular traffic lane.

According to the assigned flow patterns from Tables II and III, the two solution algorithms produce two alternative lane marking designs, although there are only

Table II. Optimization results from the cutting plane algorithm. Optimal cycle length = 71.4 seconds

(1)	(2)	(3–6)				(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
From arm	Lane	To arm				Sum of lane flow	Turning proportion	Saturation flow	y	Effective green	Degree of sat.	Green start	Total delay
		1	2	3	4								
1	1		241.4			241.4	1.00	1746.7	0.1382	26.2	0.3774	0.0	1.1079
1	2		258.6			258.6	1.00	1871.1	0.1382	26.2	0.3774	0.0	1.1795
1	3			200.0		200.0	0.00	2105.0	0.0950	10.2	0.6641	15.9	2.0399
1	4				100.0	100.0	1.00	1871.1	0.0534	6.0	0.6363	34.6	1.2927
2	1			100.0		100.0	1.00	1746.7	0.0573	11.3	0.3626	55.2	0.7645
2	2				500.0	500.0	0.00	2105.0	0.2375	20.9	0.8134	45.6	4.5307
2	3	50.0				50.0	1.00	1871.1	0.0267	10.9	0.1745	71.4	0.3456
2	4	50.0				50.0	1.00	1871.1	0.0267	10.9	0.1745	71.4	0.3456
3	1				300.0	300.0	1.00	1746.7	0.1718	29.6	0.4148	0.0	1.2430
3	2	300.0				300.0	0.00	2105.0	0.1425	13.6	0.7465	15.9	3.0341
3	3		150.0			150.0	1.00	1871.1	0.0802	9.4	0.6081	31.2	1.5222
3	4		150.0			150.0	1.00	1871.1	0.0802	9.4	0.6081	31.2	1.5222
4	1	100.0	134.9			234.9	0.43	1865.7	0.1259	20.9	0.4312	45.6	1.3502
4	2		265.1			265.1	0.00	2105.0	0.1259	20.9	0.4312	45.6	1.5045
4	3			200.0		200.0	1.00	1871.1	0.1069	10.9	0.6981	71.4	2.1607
4	4			200.0		200.0	1.00	1871.1	0.1069	10.9	0.6981	71.4	2.1607
													26.1038
										Actual green			
	P1									20.0		31.2	
	P2									20.0		47.4	

Table III. Optimization results from the heuristic line search algorithm. Optimal cycle length = 68.8 seconds

(1)	(2)	(3–6)				(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
From arm	Lane	To arm				Sum of lane flow	Turning proportion	Saturation flow	y	Effective green	Degree of sat.	Green start	Total delay
		1	2	3	4								
1	1		241.4			241.4	1.00	1746.7	0.1382	21.0	0.4534	0.0	1.3332
1	2		258.6			258.6	1.00	1871.1	0.1382	21.0	0.4534	0.0	1.4161
1	3			200.0		200.0	0.00	2105.0	0.0950	10.0	0.6516	10.9	1.9351
1	4				100.0	100.0	1.00	1871.1	0.0534	6.0	0.6127	29.3	1.1931
2	1			100.0		100.0	1.00	1746.7	0.0573	11.0	0.3574	50.0	0.7327
2	2				500.0	500.0	0.00	2105.0	0.2375	20.7	0.7882	40.3	4.0724
2	3	50.0				50.0	1.00	1871.1	0.0267	5.9	0.3095	0.0	0.4312
2	4	50.0				50.0	1.00	1871.1	0.0267	5.9	0.3095	0.0	0.4312
3	1				300.0	300.0	1.00	1746.7	0.1718	27.1	0.4366	66.0	1.2984
3	2	300.0				300.0	0.00	2105.0	0.1425	13.3	0.7360	10.9	2.8796
3	3		150.0			150.0	1.00	1871.1	0.0802	9.3	0.5936	26.0	1.4392
3	4		150.0			150.0	1.00	1871.1	0.0802	9.3	0.5936	26.0	1.4392
4	1	100.0	400.0			500.0	0.20	1917.1	0.2608	23.5	0.7625	40.3	3.6194
4	2			133.3		133.3	1.00	1871.1	0.0713	8.7	0.5609	66.0	1.2630
4	3			133.3		133.3	1.00	1871.1	0.0713	8.7	0.5609	66.0	1.2630
4	4			133.3		133.3	1.00	1871.1	0.0713	8.7	0.5609	66.0	1.2630
													26.0100
										Actual green			
	P1									20.0		26.0	
	P2									23.7		38.3	

Table IV. Statistics of the example problem

Model variables	Number of variables	Constraints	Number of constraints
Binary variables	**142**	Cutting planes	156
Permitted movements	40	Starts of green	58
Successor functions	102	Durations of green	60
		Order of signal displays	51
Continuous variables	**110**	Clearance times	68
Lane flows	48	Cycle length	1
Starts of green	29	Maximum acceptable degree of saturation	16
Durations of green	30	Flow conservations	12
Cycle length	1	Minimum permitted movements in a lane	8
Common flow multiplier	1	Maximum permitted movements at exits	3
Auxiliary variable	1	Prohibited movements	40
		Permitted movements across adjacent lanes	72
		Flow factors	72
		Lane signal settings	176
		Others	53
Total number	**252**		**846**

marginal differences in the total junction delays and optimized cycle length. For a clearer illustration, the two distinct lane marking designs are plotted in Figure 2(b) and (c). A summary of the model statistics for the example calculations is presented in Table IV. The left-hand column collects the model variables. There are 252 variables defined in the lane-based model, in which 142 are binary and the remaining 110 are continuous. The right-hand column summarizes the model constraints. There are 846 linear constraints established for the solution region in the example calculations. The auxiliary variable and 156 linearized constraints (which increase with the number of iterations) are only needed in the cutting plane algorithm, and are not required in the lane-based models for capacity maximization and cycle length minimization. The computing time for the heuristic line search algorithm (40 minutes) is a hundred times lower than that for the cutting plane algorithm (68 hours). Despite longer computing time for the cutting plane algorithm, the optimized results that are obtained from the cutting plane algorithm is no better than those from the heuristic line search algorithm. The junction delay that is obtained from heuristic line search algorithm is 0.3% lower than that from the cutting plane algorithm. This is because the Webster's delay formula is a non-convex function with respect to the lane-based variables. Therefore, there is no guarantee that the cutting plane algorithm can obtain the global optimal solution. In this example, the solution of the cutting plane algorithm is trapped into a slightly poorer local minimum, when compared with the line search method.

6. Conclusions

Lane-based optimization models have been presented in this paper. While the capacity maximization and cycle length minimization for isolated signal-control junctions are formulated as BMILP problems that can be solved by the standard branch-and-bound technique [26], the delay-minimization problem for isolated junctions is a BMINLP because of the non-linear delay function. Two solution algorithms have been proposed to solve the problem: the cutting plane algorithm and the heuristic line search algorithm. For the cutting plane algorithm, the non-linear delay function is approximated by successive linearizations. The BMINLP problem is transformed into a series of BMILP problems that are solved by the standard branch-and-bound technique. The heuristic line search algorithm takes advantage of the physical properties of the problem, in which the initial cycle length is considered as the line search control variable. The lane-based module for cycle length minimization is used to determine the feasible cycle length range, within which the delay minimization results are sought. At each initial cycle length, the lane-based module for capacity maximization is first adopted to evaluate the set of integer variables that represent the lane markings and successor functions, based on which the group-based module is used to further optimize the junction delay by means of the commercial package SIGSIGN. Both cutting plane and heuristic line search algorithms are capable of obtaining much lower junction delays than that from the non-lane-based method. The improvements are around 9% for both cases. Not only does the heuristic line search algorithm obtain a marginally lower delay than that from the cutting plane algorithm, the computing time that is required for the line search algorithm is considerably less than that for the cutting plane algorithm. Hence, the heuristic line search algorithm is a more cost-effective method for the lane-based optimization of delay at isolated signalized junctions.

Acknowledgements

We should like to gratefully acknowledge the support given to this project by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No.: HKU 7018/01E). This research was also partially supported by an Outstanding Young Researcher Award 2000 from the University of Hong Kong.

References

1. Adjiman, C. S., Androulakis, I. P. and Floudas, C. A.: Global optimization of MINLP problems in process synthesis and design, *Comput. Chem. Eng. S* **21** (1997), S445–S450.
2. Allsop, R. E.: Delay-minimising settings for fixed-time traffic signals at a single road junction, *J. Inst. Math. Appl.* **8** (1971), 164–185.

3. Allsop, R. E.: SIGSET: A computer program for calculating traffic signal settings, *Traffic Engg Control* **13** (1971), 58–60.
4. Allsop, R. E.: Estimating the traffic capacity of a signalized road junction, *Trans. Res.* **6** (1972), 245–255.
5. Allsop, R. E.: Computer program SIGCAP for assessing the traffic capacity of signal-controlled road junctions – description and manual for users, Transportation Operations Research Group Working Paper 11, University of Newcastle upon Tyne, 1975.
6. Allsop, R. E.: Computer program SIGSET for calculating delay-minimising traffic signal timings – description and manual for users, Transport Studies Group Research Report, University College London, 1981.
7. Allsop, R. E.: Evolving application of mathematical optimisation in design and operation of individual signal-controlled road junctions, In: J. D. Griffiths (ed.), *Mathematics in Transport and Planning and Control*, Clarendon Press, Oxford, 1992, pp. 1–24.
8. Burrow, I. J.: OSCADY: A computer program to model capacities, queues and delays at isolated traffic signal junctions, TRRL Report RR 105, Transport and Road Research Laboratory, Crowthorne, 1987.
9. Duran, M. A. and Grossmann, I. E.: An outer approximation algorithm for a class of mixed-integer-non-linear programs, *Math. Programming* **36** (1986), 307–339.
10. Gallivan, S.: A geometric proof that Webster's two term formula for a delay at a traffic signal is convex, Research Report, Transport Studies Group, University College of London, 1982.
11. Gallivan, S. and Heydecker, B. G.: Optimising the control performance of traffic signals at a single junction, *Trans. Res. B* **22** (1988), 357–370.
12. Geoffrion, A. M.: Generalized benders decomposition, *J. Optim. Theory Appl.* **10** (1972), 237–260.
13. Heydecker, B. G.: Sequencing of traffic signals, In: J. D. Griffiths (ed.), *Mathematics in Transport and Planning and Control*, Clarendon Press, Oxford, 1992, pp. 57–67.
14. Heydecker, B. G. and Dudgeon, I. W.: Calculation of signal settings to minimise delay at a junction, In: *Proceedings of 10th International Symposium on Transportation and Traffic Theory*, Elsevier, New York, 1987, pp. 159–178.
15. Improta, G. and Cantarella, G. E.: Control system design for an individual signalized junction, *Trans. Res. B* **18** (1984), 147–167.
16. Kelley, J. E.: The cutting plane method for solving convex programs, *J. Soc. Industr. Appl. Math.* **8** (1960), 703–712.
17. Kimber, R. M., McDonald, M. and Hounsell, N.: The prediction of saturation flows for road junctions controlled by traffic signals, TRRL Report RR 67, Transport and Road Research Laboratory, Crowthorne, 1986.
18. Lam, W. H. K., Poon, A. C. K. and Mung, G. K. S.: Integrated model for lane-use and signal-phase designs, *ASCE J. Trans. Eng.* **123** (1997), 114–122.
19. Porn, R. and Westerlund, T.: A cutting plane method for minimizing pseudo-convex functions in the mixed integer case, *Comput. Chem. Eng.* **24** (2000), 2655–2665.
20. Sang, A. P. and Silcock, J. P.: *SIGSIGN User Manual*, Steer Davies and Gleave Ltd and Transport Studies Group, University College London, 1989.
21. Silcock, J. P.: Designing signal-controlled junctions for group-based operation, *Trans. Res. A* **31** (1997), 157–173.
22. Still, C. and Westerlund, T.: The extended cutting plane algorithm, In: *Encyclopedia of Optimization*, Vol. 2, Kluwer Academic Publishers, 2000, pp. 53–61.
23. Tully, I. M. S. N. Z.: Synthesis of sequences for traffic signal controllers using techniques of the theory of graphs, PhD Thesis, OUEL Report 1189/77, University of Oxford, 1976.
24. Webster, F. V.: Traffic signal settings, Road Research Technical Paper No. 39, HMSO, London, 1958.

25. Westerlund, T. and Porn, R.: Solving pseudo-convex mixed integer optimization problems by cutting plane techniques, *Opt. Engg* **3** (2002), 253–280.
26. Wong, C. K. and Wong, S. C.: Lane-based optimization of signal timings for isolated junctions, *Trans. Res. B* **37** (2003), 291–312.
27. Wong, C. K., Wong, S. C. and Tong, C. O.: Lane-based optimization method for maximizing reserve capacity of isolated signal-controlled junctions, In: *Proceeding of the Fifth Conference of Hong Kong Society for Transportation Studies*, 2 December, Hong Kong, 2000, pp. 176–184.
28. Wong, C. K., Wong, S. C., Tong, C. O. and Lam, W. H. K.: Lane-based optimization method for minimizing delay of isolated signal-controlled junctions, In: *Proceedings of the 7th International Conference on Applications of Advanced Technology in Transportation*, 5–7 August, Cambridge, MA, USA, 2002, pp. 199–206.
29. Wong, S. C.: Derivatives of performance index for the traffic model from TRANSYT, *Trans. Res. B* **29** (1995), 303–327.
30. Wong, S. C.: Group-based optimisation of signal timings using the TRANSYT traffic model, *Trans. Res. B* **30** (1996), 217–244.
31. Wong, S. C.: On the reserve capacities of priority junctions and roundabouts, *Trans. Res. B* **30** (1996), 441–453.
32. Wong, S. C.: Group-based optimisation of signal timings using parallel computing, *Trans. Res. C* **5** (1997), 123–139.
33. Wong, S. C., Wong, W. T., Leung, C. M. and Tong, C. O.: Group-based optimization of a time-dependent TRANSYT traffic model for area traffic control, *Trans. Res. B* **36** (2002), 291–312.
34. Wong, S. C. and Yang, C.: An iterative group-based signal optimization scheme for traffic equilibrium networks, *J. Adv. Trans.* **33** (1999), 201–217.
35. Wong, S. C., Yang, C. and Lo, H. K.: A path-based traffic assignment algorithm using the TRANSYT traffic model, *Trans. Res. B* **35** (2001), 163–181.
36. Wong, S. C., Yang, C., Tong, C. O. and Wong, C. K.: Group-based optimization of signal timings for traffic equilibrium network, *J. Eastern Asia Soc. for Transportation Studies* **4**(4) (2001), 133–148.
37. Wong, S. C. and Yang, H.: The estimation of reserve capacity in traffic control, *Hong Kong Inst. Eng. Transact.* **4** (1997), 21–30.
38. Wong, S. C. and Yang, H.: Reserve capacity of a signal-controlled road network, *Trans. Res. B* **31** (1997), 397–402.
39. Wong W. T., Wong S. C. and Tong C. O. Sheared delay formulae for TRANSYT traffic model: Review and calibration, *Transport Reviews* **23** (2003), 1–20.