FORMULATION OF MODERN SIGNAL CONTROL OPERATIONS AS A NON-LINEAR MIXED INTEGER PROGRAM

Balaji Ramanathan
Michael G. McNally
R. Jayakrishnan

Department of Civil & Environmental Engineering and
Institute of Transportation Studies
University of California, Irvine, CA 92717

Abstract—This paper presents a mathematical formulation for the problem of optimizing the operations of a modern traffic signal in a possibly off-line evaluation mode. The demand is assumed to be known for the entire optimization period. The signal operations are acyclic, with no fixed staging or stage durations. Solution techniques for the resulting Non-Linear Mixed Integer Program are discussed. It is envisaged that the technique can be used to obtain the optimal signal control over a time horizon as a benchmark to compare the performance of other signal control techniques including adaptive control schemes.

I. Overview

Traffic congestion is an endemic and chronic condition in many urban regions of the world. Congestion is a result of the inability of the transportation infrastructure to accommodate daily demands, particularly demands during commuting peak hours. The recognition that the addition of new infrastructure was not necessarily the best solution to the problem of ever-increasing traffic demand has led to the development of Advanced Transportation Management Systems (ATMS) as one component of Intelligent Vehicle-Highway Systems. As part of ATMS, improved traffic control is an initial step in the better management of urban arterial traffic flow. Over the years, the technology of traffic signal control has progressed from simple two-phase, pretimed control to highly advanced multiple-ring traffic actuated control. Algorithms used in traffic signal controllers have also significantly improved, leading to the use of volume-occupancy control, delayed actuation etc.

This paper begins with a summary of the state of the art in signal control and adaptive control strategies. A formulation of the signal control problem as a mathematical program is presented and issues concerning solution techniques are discussed. The formulation assumes complete knowledge of the arrivals over a given period, and attempts to find the 'best possible' solution in terms of the signal indications. The benefits from such an approach are two-fold. First, it gives an upper bound on benefits from traffic control under a given scenario where there is 'perfect information'. This can be used to compare the performance of existing adaptive control techniques which often include heuristic steps and do not guarantee optimal results. Second, the framework can also be used in an 'imperfect information' case, where information on future traffic is unavailable. In this case future conditions are predicted, but the framework will still produce the 'best expected' performance within the constraints of the accuracy of the prediction. The new technique is intended to produce a significant improvement in the control of traffic signals directed toward minimizing intersection delay. The motivation behind this research is that, despite the volume of past research in traffic engineering, there does not appear to be any complete mathematical models published in the literature that explicitly model the phase sequences and constraints of modern traffic signal controllers (some commercial packages may use unpublished techniques to handle the modelling of dual rings and other features of modern signal controllers).

This should not be surprising when one considers the complexity of formulations such as that presented herein. Indeed, the proposed formulation appears to be convex only under certain conditions such as the arrival function being convex. It is possible that common solution techniques may not produce globally optimal solutions. This should not be a reason for rejecting such a complete formulation as this since special purpose solution techniques could be developed. The initial formulation is directed towards the optimization of traffic signal control at isolated intersections, however, extensions of the approach to network level control are discussed.

II. Motivation

The problem of optimizing signal control problems has been approached along various paths by several researchers. One of the earliest results from such research is the well -known Webster's equation that relates the delay at a pre-timed intersection with the split lengths. There are two interrelated aspects of this problem which most researchers have formulated and solved independently.

The first aspect is that of identifying the optimum phase sequence to be used at a given intersection, taking into

0-7803-2587-7/95 $4.00 © 1995 IEEE

account the peculiarities of the intersection geometry and other factors. This is referred to sometimes as the staging problem. This problem has been investigated by Murchland and Tully using combinatorial techniques [1], [2]. The second aspect is the identification of optimal durations for the signal phases whose order is already fixed by some method. This problem can be formulated such that convex programming techniques can be used to optimize various performance criteria at any given intersection [3], [4].

However, attempts to combine the two parts of the problem have lead to inefficiencies because of the loss of special properties that characterize each of the individual parts. It has been shown that incompatibilities in the solution techniques used for each of these problems prevents their satisfactory combination into a single procedure [5]. Improta and Cantarella have formulated the problem as a Binary Mixed Integer Linear Program that can be solved to obtain the staging implicitly and the splits explicitly [6].

However, Improta and Cantarella assume that demand at the intersection is cyclic (with constant demands on each of the movements) and the signal operates as a pretimed signal. Apparently, there have been no successful attempts to obtain the optimal staging and phase durations when the demand levels change or when the signal is not constrained to operate cyclically. The object of this research is to formulate the problem of optimizing the sequence of phase indications and their durations when the demand levels over a long time horizon are known and the signal is under no constraints except for minimum and maximum green durations.

III. ADAPTIVE CONTROL

Adaptive signal control, in its broadest sense, encompasses any signal control strategy that can adjust parameters of signal operations in response to fluctuating traffic demand. Over the past two decades numerous adaptive control strategies have been proposed for use at isolated intersections. Some of the better known strategies are Traffic Optimization Logic, Modernized Optimized Vehicle-Actuated Strategy, and Step-wise Adjustment of Signal Timing strategy. A detailed review of these strategies is provided by Feng-Bor Lin and others [7], [8], [9].

A characteristic that is common to most of these strategies is their reliance on heuristics rather than formal mathematical techniques to control intersection operations. Some approaches do apply mathematical programming techniques over short time horizons (local optimization) but there is little evidence of achieving global optimality over longer time horizons. This explains in part why the impact of existing adaptive control strategies on delays in the field has sometimes been limited.

OPAC (Optimized Policies for Adaptive Control) is an adaptive control strategy based on the sequencing approach [10]. The technique is formulated as a dynamic program which produces an optimal phasing pattern. The technique imposes restrictions, however, on the number of phase changes allowed during the time horizon. The optimization is applied to independent intersections, however, the formulation leads to synchronization of intersections with surrounding intersections.

The improvements in operational parameters such as delay have often been found modest in most of these techniques. A potential explanation for the sub-optimal performance of most adaptive control techniques is their reliance on heuristics. Most of the techniques rely on arbitrary and ad hoc procedures to determine the phasing and splits, supported primarily by intuitive reasoning. Methods that use optimization techniques and algorithms apply them only in the short run; a series of optimal short run decisions need not necessarily lead to the global optimum in the long run. Such problems have undermined the effectiveness of these techniques.

Performance evaluations of adaptive control techniques have relied on comparisons of various signal operation characteristics (delays and stops) obtained with the adaptive control technique against existing intersection control schemes (actuated control or pretimed control) [9]. However, these comparisons are suspect since the actual degree of improvement depends significantly on various local factors, making it almost impossible to make valid field comparisons between different adaptive control techniques. This suggests the need for a benchmark that can be used to evaluate the performance of the adaptive control technique relative to the theoretical optimum.

IV. FOUNDATIONS OF THE PROPOSED APPROACH

A formulation is proposed which is based on optimization from a purely mathematical point of view. The benefit from such a formulation is that solution properties and performance sensitivity can be comprehensively established. Traffic control has generally been considered too complex to be tackled in a pure mathematical programming environment; here, a traffic control formulation is introduced as a minimization problem to identify the optimum phasing and splits over a suitably long time horizon.

The proposed technique rests on theoretical considerations of the complete operation of an intersection and its controller; the approach can be easily extended to a network of signals. The formulation uses variables defining the state of an intersection. These state variables include, for each phase, the cumulative number of arrivals, the cumulative number of departures, the departure rate, queue lengths, and cumulative vehicular delay due to signal operation.

The operation of an intersection can be simulated at any level of detail if the arrival rates, the phasing and splits, and other design and geometric features of the intersection are known. It is possible from this simulation to find the state of the intersection at any instant in time. However, there apparently have not been any successful attempts to derive the state of the system at any instant in time by using any technique other than simulation. In simulation, the relationship between the inputs and the state of the system is obscured by the numerical process and, frequently, no formal relationship can be identified. This lack of a formal relationship makes it difficult to optimize the system mathematically. Any solution found without mathematical optimization can not be proven to be optimal because the mathematical properties of the functions being optimized are unknown.

The approach proposed here describes the intersection operations in terms of symbolic inputs. A method is proposed to derive the state of the system completely in terms of the inputs (arrival rates, the phasing and the split lengths, the saturation flow rates, etc.). This functional representation of the system can then be used to optimize the operation of the signal with respect to the phasing and splits to produce an optimal signal timing pattern for a specified objective (such as minimizing delays or reducing queues). Such functional representation of intersection operations allows for the critical examination of signal operation and the evaluation of the relative importance of the various components, which could lead to the development of more efficient optimization techniques. Furthermore, since mathematical programming is used in the optimization, the solution would have interesting and useful mathematical properties, which can be used to gain insight into and to improve the operations of other adaptive control techniques.

V. FORMULATION

A formulation is presented for the minimization of the delay resulting from the choice of particular phasing patterns and splits at an isolated intersection. The intersection is assumed to start operations with uniformly zero initial conditions (no initial queues, no initial delay). However, the formulation can be modified easily to incorporate non-zero initial conditions. The problem's variables and parameters are first defined. Please note that the term "split" is used in this formulation to refer to the time interval between successive phase changes.

Exogenous variables

- T: The time horizon under consideration (seconds)
- K: The maximum number of phase changes allowed during the T seconds (the actual number is determined during the optimization)
- k: ($k=1,2,\ldots,K$) phase change index
- i: NEMA phase (1 through 8)
- m_i: The minimum green interval of phase i
- M_i: The maximum green interval of phase i
- $a_i(t)$: The arrival rate of vehicles at time t for phase i
- $A_i(t)$: The cumulative number of arrivals for phase i through time t
- $S_i(t)$: The maximum discharge rate at time t for phase i
- N: Number of sub-intervals in each split
- n: ($n=1,2,\ldots,N$) sub-intervals index
- q_i^0: Initial queue length on phase i

Endogenous variables

- $d_i(t)$: The departure rate of vehicles at time t for phase i
- $D_i(t)$: The cumulative number of departures in phase i though time t
- $q_i(t)$: The queue length at time t in phase i
- δ_i^k: The status of phase i following phase change k (1 if it has green, 0 otherwise)
- X_k: The elapsed time at the end of the kth phase change

The following assumptions are made to operationalize the formulation. First, all the exogenous variables are known during the entire time period under consideration. Second, two non-conflicting movements, as determined by the standard dual ring diagram, are given the green indication during each phase of non-zero length. Finally, the splits are assumed to be effective green times (lost times are not dealt with explicitly).

The following relationships for cumulative arrivals, cumulative departures, and queue lengths follow directly:

$$A_i(t) = \int_0^t a_i(t)dt \tag{1}$$

$$D_i(t) = \int_0^t d_i(t)dt \tag{2}$$

$$q_i(t) = q_i^0 + A_i(t) - D_i(t) \tag{3}$$

Intersection delay is calculated directly as a function of $q_i(t)$ by integrating the area under the queue curves for all NEMA phases, or:

$$Delay\ (t) = \sum_{i=1}^{8} \int_0^t q_i(t)dt \tag{4}$$

The only variable which cannot be determined explicitly as a function of time is the departure rate of vehicles at the time instant under consideration. In general, the departure rate at any instant can be determined only with knowledge of the queue length and the phase indication. If there is a positive queue associated with the phase, the departure rate is equal to the saturation rate during the green; when the queue length is zero, the departure rate during the green is equal to the arrival rate.

Since the queue length itself is a function of the cumulative number of departures (and therefore the departure rates), it is not possible to express the departure rate as a continuously differentiable function of time. However, it is possible to find the departure rate during any arbitrarily small time-step with knowledge of the queue length at the end of the previous time-step. This discretization of time unfortunately makes the mathematical representation of the problem somewhat more complicated and less elegant.

The presence of a queue at the end of the previous time-step does not ensure that the departure rate during the next time-step would remain equal to the saturation rate. It is necessary to independently ensure that the queue length is non-negative on each phase at the end of each time-step. For each time step during which the phase receives a green, vehicles are assumed to depart at the saturation flow rate subject to the constraint that the queue length at the end of the timestep is non-negative. The smaller the time-steps, the greater the number and frequency of this independent check on the queue lengths, leading to lower errors in the calculation of departure rates and delays.

The discretization of time into intervals as long as the splits themselves would introduce unacceptable errors in the calculation of departure rates and, therefore, the queues and delays. Each split was thus divided into a constant number of time-steps, N. By choosing a sufficiently large N, an arbitrary level of accuracy can be ensured in the calculation of the queues and delays. This discretization of time results in the following formulation for the delay experienced at the intersection by all vehicles at the end of the time horizon:

$$Delay = \sum_{i=1}^{8}\sum_{n=1}^{N}[q_i(\frac{nX_1}{N}).\frac{X_1}{N}+ \ldots + q_i(X_{K-1}+\frac{n(X_K-X_{K-1})}{N}).\frac{X_K-X_{K-1}}{N}] \quad (5)$$

The expression $(X_i + n(X_{i+1}-X_i)/N)$ is a time instant at the end of the nth subdivision between the ith and $(i+1)$th phase changes. In the formulation above, the queue length at this instant is multiplied by the length of the subdivision preceding it $(n(X_{i+1}-X_i)/N)$. Notice that the integrals have been replaced by summations and that the representation utilizes the numerical integration technique known as rectangular integration. This method is sufficiently accurate as long as the intervals over which it is applied are small. A similar procedure can be used to calculate the cumulative number of departures from a phase.

The following constraints are introduced to ensure that the queue lengths are non-negative on each phase at the end of any time step:

$$q_i(X_{k-1}+\frac{n(X_k-X_{k-1})}{N}) \geq 0 \quad (6)$$

The number of actual phase splits required during the time horizon is decided during the optimization and all the remaining splits are assigned zero lengths. However, constraints are needed to ensure that the length of each non-zero split satisfies the minimum and maximum green times:

$$X_k-X_{k-1} > 0 \Rightarrow m_i \leq X_k-X_{k-1} \leq M_i \ ; \quad \forall k \quad (7)$$

Constraints are also needed to ensure that conflicting phases are not given the green indication during any split. Of the 28 possible combinations of phases in a dual ring controller, 20 are not permissible. Since δ's are constrained to be binary variables in the formulation, it is easy to ensure that no more than one of any combination of two non-permissible phases has its corresponding δ set to 1. In this formulation, the standard NEMA numbering convention for the different movements is used to identify the different phases. Thus, the inadmissible combinations of NEMA phases are phases numbered 1 and 2 (EW left and opposing through), 1 and 3 (EW left and NS left), 1 and 4 (EW left and NS through), and so forth, through 7 and 8 (NS left and opposing through). The set of twenty constraints is:

$$\delta_1^k+\delta_2^k \leq 1; \quad \delta_1^k+\delta_3^k \leq 1; \ \ldots \ ;\delta_7^k+\delta_8^k \leq 1; \ \forall k \quad (8)$$

The full formulation of the mathematical program that, when solved, provides for the minimization of delay at the intersection can now be expressed as below:

Objective function:
Minimize

$$Delay = \sum_{i=1}^{8}\sum_{n=1}^{N}[q_i(\frac{nX_1}{N}).\frac{X_1}{N}+ \ldots + q_i(X_{K-1}+\frac{n(X_K-X_{K-1})}{N}).\frac{X_K-X_{K-1}}{N}] \quad (9)$$

subject to:

$$d_i(X_{k-1}+\frac{n(X_k-X_{k-1})}{N}) \leq \delta_i^k . S_i(X_{k-1}+\frac{n(X_k-X_{k-1})}{N}) \;;\quad \forall i,\ k,\ n \tag{10}$$

$$D_i(X_{k-1}+\frac{n(X_k-X_{k-1})}{N}) = \int_0^{X_{k-1}+\frac{n(X_k-X_{k-1})}{N}} d_i(t)dt \;;\quad \forall i,\ k,\ n \tag{11}$$

$$A_i(X_{k-1}+\frac{n(X_k-X_{k-1})}{N}) = \int_0^{X_{k-1}+\frac{n(X_k-X_{k-1})}{N}} a_i(t)dt \;;\quad \forall i,\ k,\ n \tag{12}$$

$$q_i(X_{k-1}+\frac{n(X_k-X_{k-1})}{N}) = A_i(X_{k-1}+\frac{n(X_k-X_{k-1})}{N}) - D_i(X_{k-1}+\frac{n(X_k-X_{k-1})}{N}) + q_i^0 \;;\quad \forall i,\ k,\ ,n \tag{13}$$

$$X_k - X_{k-1} \leq M_i \;;\quad \forall k \tag{14}$$

$$\delta_i^k \begin{array}{l} = 1 \text{ if phase } i \text{ is green for } X_{k-1} \leq t \leq X_k \\ = 0 \text{ otherwise} \end{array} \tag{15}$$

$$q_i(X_{k-1}+\frac{n(X_k-X_{k-1})}{N}) \geq 0;\quad \forall i,\ k,\ n \tag{16}$$

$$m_i-(X_k-X_{k-1}) \leq B(1-\delta_i^k);\quad \forall i,\ k \tag{17}$$

where B is some large number.

$$\delta_1^k+\delta_2^k \leq 1;\ \delta_1^k+\delta_3^k \leq 1;\ldots;\delta_7^k+\delta_8^k \leq 1;\ \forall k \tag{18}$$

$$X_k \geq 0;\quad \forall k \tag{19}$$

$$X_k \geq X_{k-1};\quad \forall k \tag{20}$$

$$X_K = T \tag{21}$$

VI. Proposed Solution Approach

The proposed formulation is a mixed integer non-linear program. Several branch and bound techniques have been developed and studied to solve problems of this nature. Most of these techniques are available as computer programs that are sufficiently powerful and that can handle thousands of variables and hundreds of constraints. Such programs, however, often cannot be used if the formulation is non-convex.

An immediate concern in model formulation is the shape of the arrival function which could be non-convex. Conditions could be derived for the problem to remain convex and this is the focus of continuing research. In such convex cases, packages such as GAMS (General Algebraic Modelling System) could be utilized [11]. When the convexity is violated, alternative heuristic solution techniques such as simulated annealing or Tabu search could be applied. It is also possible that other special purpose algorithms could be developed. These identified concerns, however, are not to be considered grounds for rejection of the formulation or it's future refinement since it does provide a mathematical framework which includes explicit consideration of modern signalization and control, details which are not typically found in existing traffic signal optimization methods.

The problem under consideration has K positive continuous variables (the X_k's) and $i \cdot K$ binary variables (the δ's). In addition to these decision variables, the problem also involves the queue lengths ($i \cdot K \cdot N$ variables), the instantaneous departure rates ($i \cdot K \cdot N$ variables), and the cumulative numbers of departures ($i \cdot K \cdot N$ variables), at various time instants. These variables are continuous positive variables.

The number of constraints is of the order of $i \cdot K \cdot N$ because of the need for constraints at every time instant to ensure positive queue lengths. The actual number of constraints would include the number of minimum phase length constraints, maximum phase length constraints, and constraints that ensure that conflicting movements do not receive the green indication during any single split.

It is important to state that the formulation serves a dual purpose. On the one hand, it provides for off-line analysis with known conditions to develop optimal signal strategies, which can be used to evaluate signal operations (even under other adaptive control techniques). On the other hand, the formulation can be incorporated into a real-time adaptive control framework, where it finds optimal signal control strategies under assumed or predicted future conditions. For the former purpose, the framework need not be solvable in real-time, and thus solution approaches can be based on general purpose algorithms. The latter purpose, however, requires that optimal solutions be found in real-time, which may be possible only with special-purpose solution algorithms. The development of specialized solution algorithms focuses on the special structure of the signal phase transitions to develop efficient branch and bound techniques. These focus on the special structure of the signal phase transitions to develop efficient branch and bound techniques.

VII. Contribution to Advancements in Signal Control

The proposed formulation relies on formal mathematical techniques rather than heuristics. For the time horizon over which the technique is applied, and under the few assumptions, it is expected that the optimal solution, would be globally optimal. Since the technique is mathematically based, it may be shown that the solution produced is optimal for the conditions and within the constraints imposed. The solution can thus be used with confidence that there is no better alternative under the given circumstances.

The technique can be applied for an after-the-fact assessment of the efficacy of other techniques. Since the technique produces the theoretical minimum possible delay for the given arrival pattern, it can be used as a benchmark for the evaluation of other techniques. These other techniques have been evaluated in before and after tests against existing systems (typically, fully actuated systems). Such tests cannot however reveal the optimal level at which the system could have performed. This technique would allow for the establishment of a universally acceptable benchmark that is not subject to uncertainties in the test conditions.

Since the technique uses the arrival pattern as a parameter in the optimization, it is possible to see the effects of arrival stochasticity on optimal signal operation. The arrival pattern is not typically known a priori. In such cases, knowledge of the sensitivity of the optimal solution to stochastic changes in arrival patterns (and other variables) in the system is useful in designing robust systems that are near-optimal and stable for a range of arrival patterns that differ to some degree from the predicted arrival patterns.

VIII. Conclusions

As traffic demand on urban streets has increased rapidly, traffic control has advanced from pretimed control to actuated control and adaptive control. Adaptive control techniques are currently among the most versatile tools that can be used to mitigate congestion and oversaturation. However, most current adaptive control techniques are heuristic and do not ensure global optimization. Adaptive control techniques also have not been systematically evaluated due to the lack of practical and acceptable benchmarks.

A new signal control formulation has been proposed that is based entirely on mathematical programming. A functional representation of the intersection system is developed and the formulation for using mathematical programming techniques to optimize system parameters such as delay is presented. It considers intersection operation as well as demand levels to be acyclic and allows for the optimization of both signal sequences and phase durations.

The formulation, once solved, shows promise as a benchmark for use in evaluating other adaptive control techniques. Since the solution obtained is derived as the solution to a mathematical program, it has certain properties that can be demonstrated in mathematical terms. Moreover, the technique can be used to evaluate the sensitivity of traffic signal optimization to errors and stochasticity in input parameters such as arrival rates.

It is expected that the proposed formulation may fill a theoretical gap in traffic control, an applications area which has been dominated by heuristic control techniques. Future research can build on the theoretical foundations of this technique to further advance the state of the art and the state of the practice.

IX Acknowledgments

This research was sponsored by the California Department of Transportation, under its Advanced Transportation Management Systems Testbed Research Program. Their support is gratefully acknowledged. The opinions expressed in this paper are, however, solely those of the authors.

References

[1] I.M.Tully, "Synthesis of sequences for traffic signal controllers using techniques of the theory of graphs," *Report OUEL 1189/77*, Oxford University Engineering Laboratory, 1977.

[2] I.M.Tully, and J.D.Murchland, "Calculation and use of the critical cycle time for a traffic signal controller," *Proceedings of the PTRC Annual Meeting, PTRC-P152*, pp 96-112, 1978.

[3] R.E.Allsop, "Delay minimizing settings for fixed-time traffic signals at a single road junction," *Journal of the Institute of Mathematical Applications*, Vol 8, No. 2, pp 164-185, 1971.

[4] R.E.Allsop, "Estimating the traffic capacity of a signalized road junction," *Transportation Research*, Vol. 6, No. 3, pp 245-255, 1972.

[5] Stephen Gallivan, and Benjamin Heydecker, "Optimizing the control performance of traffic signals at a single junction," *Transportation Research B*, Vol 22B, No. 5, pp 357-370, 1988.

[6] G.Improta, and G.E.Cantarella, "Control system design for an individual signalized junction," *Transportation Research B*, Vol. 18B, No. 2, pp 147-167, 1984.

[7] Feng-Bor Lin, and D.J.Cooke, "Potential performance characteristics of adaptive control at individual intersections," Paper presented at the *65th Annual Meeting of the Transportation Research Board*, Washington, DC, 1986.

[8] Feng-Bor Lin, et al., "Use of predicted vehicle arrival information for adaptive signal control: An assessment," Paper presented at the *66th Annual Meeting of the Transportation Research Board*, Washington, DC., 1987.

[9] Feng-Bor Lin, and Sangaranath Vijayakumar, "Adaptive signal control at isolated intersections," *Journal of Transportation Engineering*, Vol. 114, No. 5, pp. 555-573, 1988.

[10] Nathan H. Gartner, "OPAC: A demand-responsive strategy for traffic signal control," *Transportation Research Record* 906, Transportation Research Board, Washington, DC, 1983.

[11] A.Brooke, D.Kendrick, and A.Meeraus, "GAMS: A User's Guide," World Bank, New York, 1988.