Title: Autonomous Multiagent Reinforcement Learning –5 GC Urban Traffic Control?

Authors: Baher Abdulhai
Assistant Professor and Director of ITS Centre
Department of Civil Engineering, University of Toronto
35 St. George Street, Toronto, Ontario, Canada M5S 1A4
Tel. (416) 946-5036
Fax (416) 978-5054/6813
E-mail: baher@ecf.utoronto.ca
URL: http://www.civ.utoronto.ca/its

Rob Pringle
Intelligent Transportation Systems Group
Department of Civil Engineering, University of Toronto
35 St. George Street, Toronto, Ontario, Canada M5S 1A4
and
McCormick Rankin Corporation
2655 North Sheridan Way, Mississauga, Ontario, Canada L5K 2P8
Tel. (905) 823-8500 Ext. 319
Fax (905) 823-8503
E-mail: rpringle@mrc.ca

CALL A5008 – COMMITTEE ON ARTIFICIAL INTELLIGENCE
INTRODUCTION
The last two decades have seen a variety of attempts to develop more effective methodologies for urban traffic control (UTC). Only a few of these have been implemented in practical, real-world situations, and these have produced mixed results.

In several thematic papers, Gartner (1,2) outlined a typology of existing and potential control methods, classifying these into “generations”, each of which offers additional functionality and flexibility over its predecessors. Gartner also offers some insights into the failure of existing methods to generate consistently positive results across the full range of traffic scenarios and suggests, in response, a prescription for future development efforts. Included in his prescription is a possible definition for 4GC and 5GC (4th and 5th Generation) methods, embodying higher levels of “intelligence” than has been achieved to date.

The following discussion suggests that reinforcement learning (R-L), a technique well known in the artificial intelligence (AI) and machine learning (ML) communities, has the potential, at least conceptually, to provide the functionality required of the (perhaps redefined) 5GC level and to address some of the key limitations associated with current systems.

WHERE ARE WE TODAY?
There is no readily discernible answer to this question. The authors of this paper, and we are sure, the authors of similar papers attempting to raise the “bar” in the context of the state of the art, are faced with the difficult task of establishing just where the “bar” is currently positioned. Although a valuable backdrop to any paper of this type, a comprehensive, up-to-date summary of the state of the art in urban traffic signal control is difficult to achieve for a variety of reasons.

The discussion that follows is based on a review of the available literature, supplemented by assessments provided by some prominent researchers in the field. Although these sources may not be entirely up-to-date, available evidence suggests that the issues discussed below have not been completely and successfully addressed.

1. Some of the factors that contribute to the difficulty of evaluating traffic signal control techniques include:
   - Detailed information on the operation and limitations of the more advanced methods in terms of commercial development and real-world application is, understandably, limited for proprietary reasons.
   - It is difficult to draw meaningful comparisons between different techniques. One reason is that the software tends to be sufficiently costly as to be unavailable to most researchers for testing purposes. A second is that reported results tend to be specific to a particular set of testing conditions that, even if known in sufficient detail, are difficult to replicate elsewhere.
   - There are no apparent standard “benchmark” methods or problems that would facilitate comparison. The computer science and machine learning communities appear to have gravitated towards a number of pseudo-standard problems that can serve as the basis for “benchmarking”. Examples include the “pole-balancing”, “block-pushing”, and “robot collecting objects within a grid-world while trying not to drain its battery” problems. In one of his papers, Gartner (?) comments that comparing the performance of a proposed signal control algorithm with that of the well-known TRANSYT package is somehow lacking in validity since he suggests that there are better methods than TRANSYT for the test situation in question. However, this is really missing the point in that the widely-used TRANSYT package might just be a suitable benchmark candidate. The fact that TRANSYT might not be the best method in all situations is less important than its widespread use and familiarity, or more importantly the simple fact that a benchmark has been defined.
   - Finally, the ongoing development and enhancement of many methods, combined with the delays inherent in the publication process, means that the current versions the available techniques, as well as pending enhancements, are difficult to pin down.

In recognition of these difficulties, we apologize beforehand for any statements or suppositions that have been “overtaken by events”.
The apparent state of the art in urban traffic signal control includes a substantial number of sophisticated, complex, and highly developed methods, some of which have been implemented in real-world situations while others are still in the research and development stage. These represent a variety of approaches to the urban traffic signal control problem. However, a number of limitations common to all or most of these methods have so far defied definitive solution, including:

1. Shortcomings in the accuracy and range of application of traffic prediction models.
2. Limited ability to effectively deal with a complete range of traffic conditions, including oversaturated conditions.
3. Limited ability to adapt to a changing environment (beyond perceived variations in traffic demand).
4. Limited ability to deal with traffic conditions spatially or temporally removed from a local control decision yet affecting or affected by it.
5. High computational demands, limiting the ability to deal with all aspects of the control problem or limiting the scope of effective application.

Gartner has suggested the need for new concepts rather than an extension of those already being utilized. The next section introduces R-L by describing it conceptually as a new approach that contrasts with existing approaches in some important and beneficial ways. A more complete description of the basic algorithm and its application can be found in Abdulhai et al (3) as well as in Sutton and Barto (4), Bertsekas and Tsitsiklis (5), Kaelbling et al (6) and many other references. Following this introduction, the potential of the R-L approach to address the limitations listed above is explored, as well as additional benefits of the approach and its possibilities in the context of 5GC control.

HOW IS R-L DIFFERENT FROM OTHER APPROACHES TO ADAPTIVE TRAFFIC SIGNAL CONTROL?

The form of R-L that is the primary focus of the following discussion is Q-Learning as proposed by Watkins (7) and discussed by Watkins and Dayan (8). Q-Learning (Q-L) may most instructively be classified among other R-L methods as a one-step, model-free method. Q-L is also an “off-policy” variant of R-L, meaning that learning can take place even when decisions are being explored that are not currently part of the optimal strategy.

Q-L as Dynamic Programming

Q-L is derived from classic dynamic programming (DP) as developed by Bellman (9) and enhanced by many other researchers. However, it differs markedly and beneficially from the form of dynamic programming used in several contemporary control methods such as RHODES (10), PRODYN (11,12), and the early versions of OPAC (13,14). Like these applications of DP, Q-L addresses the problem of sequential decision making (successive decisions on which traffic movements should receive a green indication next, for how long, and starting when relative to adjacent traffic signals) by recasting the problem as a recursion based on manageable problem subsets. However, whereas other methods typically deal simultaneously with an entire sequence or cycle of phase switching decisions but only one such sequence or cycle at a time, Q-L can both contract and extend this time horizon by addressing each phase-switching decision individually, but in the context of other decisions extending over a potentially much longer time frame than one phase sequence or cycle. This is consistent with the acyclic approach suggested by Gartner (1).

Q-L as a Learning Process with Minimal On-line Computational Burden

Typically, current dynamic programmic approaches, as well as those based on integer programming, exhaustive enumeration, and other optimization methodologies, involve real-time optimization, a computational burden that has necessitated simplification of some of these methods. Boillot et al (15) identifies some examples of such modifications. In contrast, Q-L minimizes the on-line computational burden by “learning” optimal or near-optimal strategies off-line and storing the currently best decisions for each combination of the state variables in a look-up table or by means of a function approximator such as an artificial neural network. Repeated “visits” to the same scenario yields expected transition probabilities and incrementally improved estimates of the “values” of the decisions tried. Best decisions for states not seen before can be obtained through generalization based on previous experiences with similar scenarios. This learning can be accomplished in the background, with the use of an off-line
B. Abdulhai and R. Pringle

simulator, supplemented by on-line fine-tuning of the optimal policy in the context of the real-world environment. The resulting minimization of the on-line computational burden provides the flexibility to re-allocate computational resources to other functions such as frequent decision updates to account for continually changing traffic states. Extensive real-time computation, a difficult problem, is replaced by off-line computation and on-line memory, both significantly easier problems to solve. Recent developments in traffic simulation, by Brockfield et al (16) and others using cellular automata, and by Daganzo (17) using the related cell transmission model, suggest significant improvements in simulation efficiency without debilitating compromises in fidelity.

Performance Feedback in Q-L

Most, if not all, adaptive traffic signal control methods do not include a feedback loop whereby decisions are evaluated as to their optimality in the context of the desired objective. Control decisions are typically based on a theoretical modelling framework tying together signal indications, traffic movement, queues, and delays. Q-L, however, is based on self-evaluation and improvement of the learned policy through trying different actions and observing the outcomes. Basing control decisions on a sample of observed outcomes yields significant benefits in terms of realism and adaptability to the real-world environment. With additional memory, which is inexpensive, Q-L is capable of choosing which among several operational objectives, such as minimizing delay or maximizing throughput, to pursue based on the performance observed previously for the given traffic conditions. This “objective-switching” could be accomplished by training the Q-L agent separately for different objective functions and later, when a control decision is required in the field, comparing the value functions associated with each objective to determine which would yield the best outcome in terms of traffic performance.

Q-L as a Distributed Control Framework

The proposed approach involves completely distributed decision-making and control, in contrast to contemporary adaptive control methods that involve centralization of part or all of the control function. Advantages of the distributed approach include robustness in the event of communications problems, what amounts to an efficient parallel computing environment, and potentially reduced communication requirements. Communication of data and decision information between individual signal control agents can be accommodated directly or through a “blackboard”, depending upon the number of agents involved and the topology of the network.

Other R-L Approaches

Thorpe (18) and Wiering (19) have also reported on R-L-based approaches to traffic signal control. However their formulations of the traffic control problem represent significant departures from the norm and the real-world practicality of these approaches is questionable. Wiering’s approach, in particular, requires an accounting of the value of control decisions to individual drivers as well as real-time knowledge of trip origins and destinations.

Other non-R-L Approaches

There are many recent papers, in addition to those on the more commonly known adaptive control methods such as SCOOT (20,21,22), SCATS (23), OPAC (13,14), RHODES (10), PRODYN (11,12), and UTOPIA (24), that cover a wide range of possible approaches to adaptive signal control. These include optimal control (Diakaki et al (25)), game theory (Porche et al (26), rule-based approaches (Yagar et al (27)), fuzzy logic (Kosonen et al(28)), genetic programming (Park et al (29) and others. The large number of papers on this topic is a reflection of the potential benefit to be gained through true optimization of traffic control and, perhaps, an indication that we are not yet close to achieving that ideal.

HOW DOES R-L/Q-L ADDRESS LIMITATIONS COMMON TO CONTEMPORARY ADAPTIVE CONTROL METHODS?

Shortcomings in the accuracy and range of application of traffic prediction models.

Gartner (2) points out that signal control methods are limited by the accuracy of the models used to predict traffic and the quality of the data input to this process. He goes on to say that the smoothing of
B. Abdulhai and R. Pringle

input data inherent in some traffic prediction models limits the potential for true traffic-responsiveness.

Gartner recognizes the difficulty associated with accurate traffic prediction but also the need for some form of prediction to provide for proactive traffic control. The most common prediction models, the “store-and-forward” model used in several methods and the platoon dispersion and traffic flow profile model used in SCOOT and its off-line companion, TRANSYT, each has limitations. For example, in congested conditions where queues do not clear each cycle, the “store-and-forward” approach can accumulate prediction error. It is not clear that the traffic flow profiles used by SCOOT will be accurate if traffic queues extend beyond the upstream detectors. Neither method deals adequately with significant mid-block traffic sources or sinks. Furthermore, such models are pre-specified and hypothetical constructs, the ongoing validation and calibration of which would be resource-intensive. More recently, a number of researchers have proposed the integration of traffic signal control with dynamic traffic assignment as another method of predicting short-term, future traffic and its response to traffic control decisions. However, this approach adds a further real-time computational burden with respect to the estimation and assignment of trip matrices, as well as a need for ongoing calibration, and is also subject to uncertainty with respect to travel patterns, assignment criteria, and driver responsiveness to control strategies.

The proposed Q-L framework operates without a pre-specified model (although some researchers maintain that model-based R-L can learn an optimal policy more quickly than model-free methods such as Q-L, not a critical issue where the bulk of the learning is accomplished off-line and in the background). Instead, an implicit model is learned based on interaction with the real-world traffic environment. This means that local conditions, such as mid-block traffic additions or subtractions, can also be “learned” without external calibration. This model is implicit in the sense that there is no set of equations describing traffic state transitions. The state transitions are implied in the learned optimal policy or strategy. Learning is of the form: given the state variables such as (i) local queues or occupancies, (ii) queues or occupancies on upstream and downstream approaches, (iii) the time and nature of the next change of indication at the adjacent signals, and a time of day variable, does the action chosen (either the best currently known decision or an exploratory decision) improve on the best performance achieved to date. If so, its “value” is increased or reinforced. The observed performance is related directly to the state variables and the action taken without the intervention of a traffic prediction model, delay or queue estimation model, or on-line optimization model. The stochasticity of traffic means that there is (probably) always going to be some uncertainty in traffic prediction. However, during both training and fine-tuning, the Q-L approach iterates towards an expected performance level, assisted by knowledge of the pending control decisions and traffic states at adjacent traffic signals.

A related issue is the detectorization requirements associated with existing methods. It is possible that the quantity and quality of data collected could be improved and the cost of collection reduced through the application of recent advances in video-based monitoring and machine vision and pattern recognition techniques. Such updated methods could benefit most, if not all, UTC methods and their availability is assumed in conjunction with the proposed Q-L framework.

Limited ability to effectively deal with a complete range of traffic conditions, including oversaturated conditions.

Researchers, including Boillot et al (15), and system suppliers, including Martin (30) in reference to SCOOT and SCATS, have pointed out the inability of existing methods to deal effectively with oversaturated conditions (oversaturation is taken here to be a traffic scenario where queues at signalized intersections fail to clear each cycle and instead grow until they affect the operation of upstream intersections). Part of the problem lies, as pointed out previously, with current traffic prediction models and detectorization setups. However, as discussed in comprehensive reviews penned by Pignataro et al (31), Quinn (32), and the O.E.C.D. (33), there are many different theoretical approaches to dealing with oversaturated conditions. It has also been proposed that the best control policy to prevent the onset of oversaturation is distinct from the best control policy to recover from oversaturated conditions that could not be prevented. Proposed but unproven methods include the “bang-bang” approach of Gazis et al (34), equity offsets, “flared” green phases, and reversed offsets, among others. Brockfield et al (16), through experiments with cellular automata-based simulation, suggest that random offsets, resulting from a local rather than global control strategy, may be an effective means of keeping traffic moving in conditions of high traffic density. Among the existing methods, the “gating” or “metering” capability built into SCOOT
B. Abdulhai and R. Pringle

is perhaps the most prominent example of a current approach to dealing with oversaturation. However, this capability requires pre-planning on a case-by-case basis and is effectively a means of relocating a traffic jam from critical to less-critical links.

The proposed Q-L approach has the advantage of being able, under off-line, simulated conditions, to seek out the best control policy for a variety of oversaturation scenarios, essentially through trial-and-error. Furthermore, the potential ability of the proposed Q-L approach, discussed in more detail below, to address the impacts of control decisions on traffic conditions removed spatially and temporally from the location and time of the actual decision enhances the ability of individual intersection controllers to cooperate towards a successful strategy to avoid, limit the extent of, and recover from severe congestion. This capability is enhanced by the potential for the Q-L approach to choose from among several control objectives and performance measures, recognizing that uncongested and congested conditions may respond best to the pursuit of different objectives. Severely congested central business district streets, or the streets serving a major traffic generator subject to filling and emptying “rushes”, would benefit from effective strategies to address oversaturated conditions. The development of oversaturation remediation strategies for implementation within “conventional” and current adaptive control systems could also be pursued through the application of Q-L as a study tool, much as TRANSYT and other methods are used today.

Limited ability to adapt to a changing environment (beyond perceived variations in traffic demand).

Adaptive traffic signal control techniques are designed to adapt to changes in traffic demand as they occur. However, the parameters defining the traffic prediction model, the integrated assignment model, if used, or ancillary models, such as the gating model in SCOOT, are fixed externally upon implementation and can only be re-calibrated externally as changes in the environment occur.

In contrast, the Q-L approach, by virtue of its ongoing learning capabilities, adapts iteratively and incrementally to new conditions as they occur. For example, a road widening, a change in traffic regulations, or a new parking lot with mid-block access could affect the expected transitions between states for nearby signal controllers. Ongoing self-evaluation on the part of a Q-L control agent could identify the deviation from previously learned transitions and re-start the learning process to adapt to the new conditions. This obviates the need for technical staff to periodically review and update model parameters and avoids the time lag between the change occurring and full adaptation of the controller to the new conditions.

Limited ability to deal with traffic conditions spatially or temporally removed from a local control decision yet affecting or affected by it.

Congestion occurring in one part of a road system can rapidly spread to adjacent areas. A signal control decision at the downstream end of an arterial can, if conditions are close to saturation, produce a debilitating congestion “shock-wave”. At best, existing control methods provide for interaction with adjacent signals, which provides some enlargement of the decision context for individual signal controllers through a “domino” effect.

In addition to this capability, the Q-L approach provides an environment whereby individual signal control agents throughout a signal system can learn how to cooperate towards a common objective. If this objective is global in nature, such as minimizing the time required to empty a downtown area of rush-hour traffic, feeding back to each controller the success or failure of the joint signal control strategy tends to reinforce globally cooperative behaviour. Recent advances in measurement and communication techniques, such as vehicle re-identification as described by Abdulhai (35), provides the capability to utilize more globally oriented state and performance measures, such as travel times through a sub-area, or throughput from the area.

High computational demands, limiting the ability to deal with all aspects of the control problem or limiting the scope of effective application.

The computational demands associated with existing adaptive control algorithms have been alluded to already. The “curse of dimensionality” associated with on-line, real-time dynamic programming has led to limitations on the scope represented by the number of signals controlled and the decision burden, represented by the number of control variables. The limited time available for real-time decision-making,
B. Abdulhai and R. Pringle
particularly where decision updates at short intervals are desired, limits the sophistication of the data processing and optimization functions.

The features of the Q-L approach that reduce the on-line, real-time computational burden have already surfaced in the preceding discussion. The prior learning of an optimal strategy off-line, with real-time fine-tuning operating only in the background, the limited computation required to access the optimal strategy for a current decision or to input some training data, the elimination of a traffic prediction model, and the ability to modularize and re-combine the decision to reduce state space dimensionality and data access times are some of these features. In fact, the minimal requirement for real-time computation is comparable to that of rule-based methods such as SSPORT as presented by Yagar et al (27).

THE Q-L APPROACH IN THE CONTEXT OF 5GC CONTROL

Existing approaches to traffic signal control typically slot into Gartner’s Level 1 (1GC), Level 2 (2GC), or Level 3 (3GC) definitions, based on increasing adaptiveness and emerging intelligence as the Level increases. His definition of Level 4 (4GC) control calls for increased intelligence allowing for integration with dynamic traffic assignment, the implementation of congestion avoidance and relief strategies, and the handling of incidents. However, it is Gartner’s Level 5 (5GC) control that is most intriguing. Despite his earlier call for the development of new approaches rather than the extension of existing ones, Gartner suggests that Level 5 control should be a “superlevel” that uses AI to select from among an array of existing control strategies (taken from Levels 1 through 4) based on traffic conditions. In this way, Gartner feels that best possible use can be made of accumulated experience with the different methods. However, it is difficult to imagine a municipal, provincial, or state agency with the necessary resources to look after the care and feeding of multiple control strategies.

It is our belief that the Q-L approach can achieve the functionality and benefits of Gartner’s Level 5GC “superlevel” without the need to maintain and update a variety of control tools. Some of the features of the Q-L approach that support this belief can be summarized as follows:

1. The ability to learn the best control decision for a given situation through a prior training period that involves an exhaustive trial-and-error process on a simulator followed by fine-tuning and automatic updating through interaction with real-world conditions. While the resulting set of control decisions might not be capable of being labelled as a 3GC or other strategy, they will have evolved to include the successful features of all of these strategies as well as additional strategies to successfully avoid (where possible) or relieve oversaturation, or provide transit priority, emergency vehicle priority, or integration with a route guidance strategy.

2. The ability to choose, in real-time, the best objective to pursue for a given situation. This may be a local objective such as reducing local delay, or a global objective such as maximizing the rate of traffic exiting an area.

3. The ability to incorporate local state information as well as information from adjacent traffic signals without running afoul of the “curse of dimensionality”.

4. The ability to perform self-evaluation and undertake additional training (in the background) when it is recognized that the environment has changed or control decisions are no longer optimal.

5. The ability to make use of emerging detection technology.

6. The ability to simultaneously choose split and offset, although these will be in the form of phase, phase start time, and duration in an acyclic context. The set of possible actions can be constrained in recognition of allowable phase sequences and minimum and maximum phase times based on regulatory or policy requirements.

7. The ability to recognize the effects of a control decision on traffic beyond the intersection where the decision was implemented and for a period of time while that traffic is within the controlled network.

8. The ability to operate within a distributed control framework, achieving efficiency through virtual parallelism and achieving a high degree of robustness.

9. The ability to do the above with a high degree of automation and seamlessness.

These may seem to be optimistic claims but perhaps not when one considers that the RL approach has been used to develop a world-class virtual backgammon player, soccer team strategies, hunting team strategies, packet routing strategies in communication networks, and improved elevator control strategies.
STATUS OF RESEARCH

The Q-L approach has been applied successfully to an isolated intersection (3). Current research efforts are focussed on extending the approach to a linear corridor and a 2-dimensional network with two-way, unevenly-spaced streets and a variety of undersaturated and oversaturated traffic demand profiles but simple two-phase traffic signals without turning movements. Testing of a variety of problem formulation and implementation options is being pursued, including testing with respect to:

1. Alternative state definitions including combinations of local, upstream and downstream queues (difficult to identify and measure when traffic is discharging) or road segment occupancies, planned control decisions at adjacent signals, and time of day variables.
2. Alternative control objectives and related switching mechanisms, including delay, travel time, traffic movement, and throughput-based objectives.
3. Updating and review protocols whereby an individual signal agent can modify its current decision in light of changing traffic conditions and adjacent signal agents can review their decisions in the context of such modifications. The challenge here is to avoid running afoul of the closure aspect of traffic in a connected, 2-dimensional road network.
4. Variations on the Q-L algorithm for both training and actual implementation with respect to modularization to limit the size of the state space if a more comprehensive state definition is used, bounding of rewards (often a compounded discount rate is used but this constrains the recognition of temporally distant outcomes), exploration of currently optimal vs. previously untried control decisions, updating protocols, and training regimens and requirements.
5. The generalization and storage of the learned optimal policies, focussing on the use of CMAC’s (Cerebral Model Articulation Controller(36,37)) stored efficiently in hash tables.
6. Efficient and fast simulation for training purposes based on the concept of cellular automata.

CONCLUSIONS

Existing adaptive traffic signal control approaches are subject to a number of limitations. A new approach, based on reinforcement learning, has been proposed as a possible means of addressing these limitations as well as extending the capability to optimize traffic movement across a signalized road network. The reinforcement learning approach has a number of strengths which potentially could achieve the same objective as Gartner’s proposed multi-strategy 5GC “superlevel” but in a more straightforward and efficient manner.

REFERENCES


