Reinforcement Learning for *True* Adaptive Traffic Signal Control

Baher Abdulhai[1]; Rob Pringle[2]; and Grigoris J. Karakoulas[3]

Abstract: The ability to exert *real-time, adaptive* control of transportation processes is the core of many intelligent transportation systems decision support tools. Reinforcement learning, an artificial intelligence approach undergoing development in the machine-learning community, offers key advantages in this regard. The ability of a control agent to *learn* relationships between control actions and their effect on the environment while pursuing a goal is a distinct improvement over prespecified models of the environment. Prespecified models are a prerequisite of conventional control methods and their accuracy limits the performance of control agents. This paper contains an introduction to Q-learning, a simple yet powerful reinforcement learning algorithm, and presents a case study involving application to traffic signal control. Encouraging results of the application to an isolated traffic signal, particularly under variable traffic conditions, are presented. A broader research effort is outlined, including extension to linear and networked signal systems and integration with dynamic route guidance. The research objective involves optimal control of heavily congested traffic across a two-dimensional road network—a challenging task for conventional traffic signal control methodologies.

DOI: 10.1061/(ASCE)0733-947X(2003)129:3(278)

CE Database subject headings: Traffic signal controllers; Intelligent transportation systems; Traffic control; Traffic management; Adaptive systems.

Introduction

The ability to exert real-time, adaptive control over a transportation process is potentially useful for a variety of intelligent transportation systems services, including control of a system of traffic signals, control of the dispatching of paratransit vehicles, and control of the changeable message displays or other cues in a dynamic route guidance system, to name a few. In each case, the controlling actions should respond to actual environmental conditions—vehicular demand in the case of a signal system, the demand for multiple paratransit trip origins and destinations, or the road network topology and traffic conditions in the case of dynamic route guidance. Even more valuable is the ability to control in accordance with an optimal strategy defined in terms of one or more performance objectives. For example, one might wish to have a signal control strategy that minimizes delay, a paratransit dispatching system that minimizes wait time and vehicle kilometers traveled, or a dynamic route guidance system that minimizes travel time.

A key limitation of conventional control systems is a requirement for one or more prespecified models of the environment. The purpose of these might be to convert sensory inputs into a useful picture of current or impending conditions or provide an assessment of the probable impacts of alternative control actions in a given situation. Such models require domain expertise to construct. Furthermore, they must often be sufficiently general to cover a variety of conditions, as it is usually impractical to provide separate models to address each potential situation. For example, some state-of-the-art traffic signal control systems rely on a platoon-dispersion model to predict the arrival pattern of vehicles at a downstream signal based on departures from an upstream signal. A generalized model designed to represent all road links cannot possibly reflect the impacts of the different combinations of side streets and driveways generating and absorbing traffic between the upstream and downstream signals.

What if a controlling agent could directly *learn* the various relationships inherent in its world from its experience with different situations in that world? Not only would the need for model prespecification be obviated or at least minimized, but such an agent could effectively tailor its control actions to specific situations based on its past experience with the same or similar situations. The machine-learning research community, related to the artificial intelligence community, provides us with a variety of methods that might be adapted to transportation control problems. One of these, particularly useful due to its conceptual simplicity, yet impressive in its potential, is reinforcement learning [see Sutton and Barto (1998) or Kaelbling et al. (1996) for comprehensive overviews, or Bertsekas and Tsitsiklis (1996) for a more rigorous treatment].

This paper provides a brief introduction to the concept of reinforcement learning. As a case study, reinforcement learning is applied to the case of an isolated traffic signal with encouraging results. This is the first stage in a research program to develop a signal system control methodology, based on reinforcement learn-

[1] Assistant Professor and Director, Intelligent Transportation Systems Centre, Dept. of Civil Engineering, Univ. of Toronto, Toronto, ON, Canada M5S 1A4. E-mail: baher@ecf.utoronto.ca

[2] PhD Candidate, Intelligent Transportation Systems Centre, Dept. of Civil Engineering, Univ. of Toronto, Toronto, ON, Canada M5S 1A4. E-mail: rob.pringle@utoronto.ca

[3] Dept. of Computer Science, Univ. of Toronto, Pratt Building LP283E, 6 King's College, Toronto, ON, Canada M5S 1A4. E-mail: grigoris@cs.toronto.edu

Note. Discussion open until October 1, 2003. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on October 30, 2001; approved on May 21, 2002. This paper is part of the ***Journal of Transportation Engineering***, Vol. 129, No. 3, May 1, 2003. ©ASCE, ISSN 0733-947X/2003/3-278–285/$18.00.

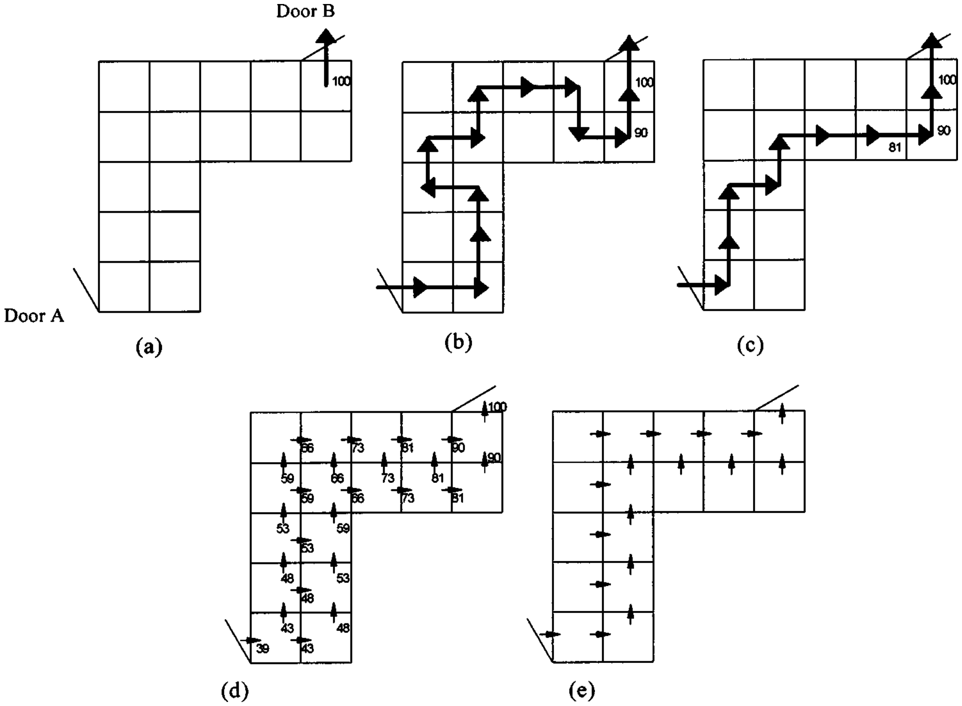

Fig. 1. Illustration of reinforcement learning: (a) Gridworld; (b) first episode; (c) second episode; (d) selected final Q-estimates; (e) one possible optimal policy

ing, which could be integrated with dynamic route guidance to provide effective traffic control in highly congested conditions. Effective traffic control in the face of severe congestion on a two-dimensional road network is a challenging task for existing signal control methodologies.

Reinforcement Learning: Brief Primer

Illustrative Example

In its simplest terms, reinforcement learning involves an agent that wishes to learn how to achieve a goal. It does so by interacting dynamically with its environment, trying different actions in different situations in order to determine the best action or sequence of actions to achieve its goal from any possible given situation. Feedback signals provided by the environment allow the agent to determine to what extent an action actually contributed to the achievement of the desired goal.

To illustrate the concept of reinforcement learning, consider the following simplified example of a mobile robot navigating within the gridworld shown in Fig. 1(a). This is actually an illustration of Q-learning, developed by Watkins (1989; Watkins and Dayan 1992), and is one of a number of possible reinforcement learning algorithms and the one that is used in the case study presented later in this paper. Imagine that the robot starts behind Door A and that its goal is to pass through Door B, for which it gains a reward of 100 units. No other actions are rewarded. Once it passes through Door B, it remains there (perhaps waiting for a further task) and gains no further rewards. At each time step, the robot can move to an adjacent grid square but cannot move diagonally. Let us also define a discount rate that has the effect of reducing the value of future rewards relative to more immediate rewards. In this case, the discount rate also has the effect of encouraging the robot to learn the shortest possible path to Door B. For this example, assume that the discount rate is 0.9.

Initially, all potential moves from any given grid square, except that involving passing through Door B from the square in front of it, have a value of zero, in the sense that no reward appears to be gained by implementing them. On its first journey, therefore, the robot explores in a random fashion, possibly following the path shown in Fig. 1(b), until it eventually passes through Door B. In doing so, it gains a reward of 100 units and remains there, ending the current episode. The value assigned to the move preceding the move through Door B is updated using the reward of 100 units, factored by the discount rate of 0.9, since the reward was gained one time step into the future, to give a net value of 90 units. On its second journey from Door A, the robot explores until it reaches a square adjacent to that in front of Door B. As before, the preceding move is assigned a value of 90 units, factored by the discount rate of 0.9, to give a net value of 81 units, as shown in Fig. 1(c). Each journey or episode thereafter may result in another move being assigned a value. At some point, the robot might find itself confronted with a choice between making a move with zero value and making one with some previously assigned positive value. In this situation, the robot must choose whether or not to explore the move with a current value of zero, on the chance that it might be better than exploiting its current knowledge by making the move that it knows has a positive value.

After a sufficient number of episodes, each move from any given square will have been assigned a value. In most practical problems, particularly in stochastic domains, many episodes are required before these values achieve useful convergence. Fig. 1(d) shows a selection of these values, each of which represents the sum of discounted future rewards if one follows an optimal path from that particular grid square to Door B. The robot has therefore learned an estimate of the value function *Q*. At this point, the

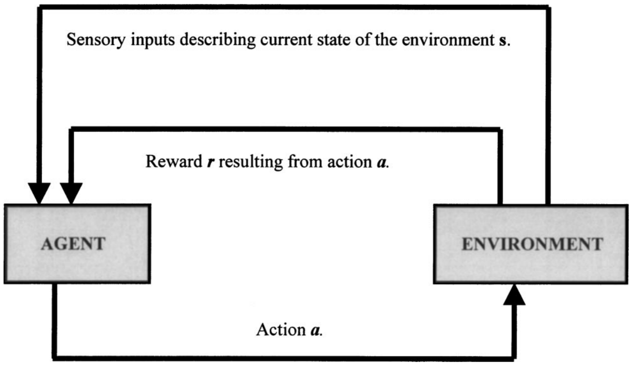
Sensory inputs describing current state of the environment **s**.
Reward *r* resulting from action *a*.
AGENT
ENVIRONMENT
Action *a*.

Fig. 2. Key elements of Q-learning

robot can implement an optimal sequence of actions, or *policy*, by greedily taking the action with the highest value, regardless of where it starts from or finds itself, until it reaches Door B. Fig. 1(e) shows one of several possible optimal policies.

More Precise Definition of Q-learning

Building on the example described in the preceding section, let us now formulate a more precise, although still basic, definition of Q-learning. Consider the system shown in Fig. 2, which shows the key elements of Q-learning.

1. The agent is the entity responsible for interpreting sensory inputs from the environment, choosing actions on the basis of the fused inputs, and learning on the basis of the effects of its actions on the environment. At time **t**, the Q-learning agent receives from the environment a signal describing its current state **s**. The state is a group of key variables that together describe those current characteristics of the environment that are relevant to the problem. Theoretically, the state information must exhibit the Markov property, in that this information, together with a description of the action being taken, is all that is needed to predict the effect on the environment. The agent does not need to know the history of its previous states or actions. In practice, it is assumed that the process is Markovian, although this may not be strictly true.
2. Based on its perception of the state **s**, the agent selects an action **a**, from the set of possible actions. This decision depends on the relative value of the various possible actions, or more precisely on the estimated Q-values $\mathbf{Q}_{s,a}$, which reflect the value to the agent of undertaking action **a** while in state **s**, resulting in a transition to state **s**′, and following a currently optimal policy (sequence of actions) thereafter. At the outset, the agent does not have any values for the Q-estimates and must learn these by randomly exploring alternative actions from each state. A gradual shift is effected from exploration to exploitation of those state/action combinations found to perform well.
3. As a result of taking action **a** in state **s**, the agent receives a reinforcement or reward $\mathbf{r}_{s,a}$, which depends upon the effect of this action on the agent's environment. There may be a delay between the time of the action and the receipt of the reward. The objective of the agent in seeking the optimum policy is to maximize the accumulated reward (or minimize the accumulated penalty) over time. A discount rate may be used to bound the reward, particularly in the case of continuous episodes. The discount rate reflects the higher value of short-term future rewards relative to those in the longer term.
4. The combination of state **s**, action **a**, and reward $\mathbf{r}_{s,a}$ is then used to update the previous estimate of the Q-value $\mathbf{Q}_{t-1(s,a)}$ recursively according to the following training rule:

$$\delta=\alpha_{s,a}\{\mathbf{r}_{s,a}+\gamma_t\cdot\mathbf{MAX}[\mathbf{Q}_{t-1(s',a')}]-\mathbf{Q}_{t-1(s,a)}\} \quad (1)$$

   where δ = increment to be added to the previously estimated Q-value, $\mathbf{Q}_{t-1(s,a)}$, to get $\mathbf{Q}_{t(s,a)}$; $\alpha_{s,a}$ = training rate in the interval [0,1]; $\mathbf{r}_{s,a}$ = reward received for taking action **a**, while in state **s**; γ_t = discount rate in the interval [0,1], applied to future rewards; $\mathbf{MAX}[\mathbf{Q}_{t-1(s',a')}]$ = previously estimated Q-value following the optimum policy starting in state **s**′; and $\mathbf{Q}_{t-1(s,a)}$ = previous estimate of the Q-value of taking action **a** while in state **s**.
   This particular training rule is relevant to stochastic environments such as the traffic environment in the case study outlined in the next section. Decreasing the training rate over time is one of the conditions necessary for convergence of the Q-function in a stochastic environment. The other condition requires that each state-action combination be visited infinitely often, although in most practical problems, the portion of the state-space that is of primary interest will be visited often but not infinitely often. If penalties are received rather than rewards, the **MIN** function is used in place of **MAX**.
5. The updated estimate of the Q-value is then stored for later reuse. The Q-values may be stored in an unaltered form in a look-up table, although this requires a significant amount of memory. They may also be used as inputs to a function approximation process designed to generalize the Q-function so that Q-value estimates may be obtained for state/action combinations not yet visited but similar to combinations that have been visited.

Adaptive Traffic Signal Control—Case Study Using Q-learning

Background

Until relatively recently, capital improvements, such as building new roads or adding traffic lanes, and a variety of operational improvements have been the primary tools used to address increasing congestion due to growth in road traffic volumes. However, increasingly tight constraints on financial resources and physical space, as well as environmental considerations, have required consideration of a wider range of options. Enhancing the intelligence of traffic signal control systems is an approach that has shown potential to improve the efficiency of traffic flow. Offline signal coordination methods, such as the maximization of through-bandwidths using time-space diagrams and optimization with the TRANSYT family of programs, are gradually giving way in larger cities to real-time methods such as the split, cycle, offset optimization technique (SCOOT) (Hunt et al. 1981; Bretherton 1996; Bretherton et al. 1998). Research is continuing into traffic signal control systems that adapt to changing traffic conditions (Gartner and Al-Malik 1996; Yagar and Dion 1996; Spall and Chin 1997; Sadek et al. 1998).

Severe traffic congestion, both recurring and nonrecurring, presents a difficult challenge to existing control methodologies, particularly in the case of two-dimensional road networks. Such congestion is often experienced in conjunction with busy urban cores and, on a more localized basis, in association with major sports and entertainment events, major accidents or other incidents, and road construction and maintenance. There is an apparent need to continue development of traffic signal control techniques to more effectively address these situations and it is to this niche that the research described in the following case study is directed.

Recent research literature includes three related efforts where reinforcement learning or related dynamic programming algorithms were applied to the problem of traffic signal control. Sen and Head (1997) utilized dynamic programming to develop a phasing plan for each cycle based on short-term traffic predictions. However, their approach lacks the ability to learn from experience and requires a traffic prediction model. Thorpe (1997) used reinforcement learning to minimize the time required to discharge a fixed volume of traffic through a road network, but his approach does not appear to be directly applicable to real-time traffic signal control. Bingham (1998) applied reinforcement learning in the context of a neuro-fuzzy approach to traffic signal control, but met with limited success due to the insensitivity of the approach, limited exploration in what is a stochastic environment, and an off-line approach to value updating.

Advantages of Q-learning for Traffic Signal Control

In comparison to other state-of-the-art techniques used for traffic signal control, and many other dynamic programming and machine learning approaches, Q-learning offers some potentially significant advantages, as discussed next.

Q-learning does not require a prespecified model of the environment on which to base action selection. Instead, relationships between states, actions, and rewards are learned through dynamic interaction with the environment. By way of contrast, existing traffic signal control methods usually require prespecified models of traffic flow to generate short-term predictions of traffic conditions or to assess the impacts of possible control decisions. If a single, general model is used, it is possible and even likely that conditions around individual intersections will vary from the conditions upon which the model was based.

Another benefit of Q-learning, and reinforcement learning in general, is that supervision of the learning process is not required. *Supervised* machine-learning algorithms require, for training purposes, a large number of examples, consisting of sets of inputs and associated outcomes, which adequately cover the range of environmental conditions expected on deployment. They involve supervision in the sense that the appropriate outcome is provided for each combination of inputs so that any inherent relationships can be learned. The machine learning methods, such as artificial neural networks, that have been the most widely studied and applied to transportation systems to date, typically involve supervised learning. An example is the work on incident detection by Abdulhai and Ritchie (1999a, b). In the case of Q-learning, which is *unsupervised*, the outcome associated with taking a particular action in any state encountered is learned through dynamic trial-and-error exploration of alternative actions and observation of the relative outcomes. Rather than being presented with a large set of training examples, the generation of which is a challenging task in many cases, even for a domain expert, a Q-learning agent essentially generates its own training experiences from its environment. The learning process can be initiated on a simulator, with refinement and optimization for the intended environment occurring after deployment.

It is important to note that not all *real-time* algorithms are truly *adaptive*. The two terms are often used interchangeably and possibly confused. Real-time algorithms are those able to respond to sensory inputs in real time, although the internal logic and parameters of the controller remain unchanged. On the other hand, an essential feature of adaptive algorithms is their ability to adjust their internal logic and parameters in response to major changes in the environment—changes that may make the knowledge base in a nonadaptive controller obsolete. One of the advantages of reinforcement learning is that such algorithms are truly adaptive, in the sense that they are capable of responding to not only dynamic sensory inputs from the environment, but also a dynamically changing environment, through ongoing learning and adaptation. Since the one-step Q-learning algorithm updates the Q-estimates at short intervals in conjunction with each action, it is also readily adaptable to on-line, real-time learning. Furthermore, Q-learning is an off-policy algorithm, in the sense that it is gaining useful experience even while exploring actions that may later turn out to be nonoptimal.

Key Elements of Case-study Implementation

The initial test application of Q-learning to the problem of traffic signal control involved a single, isolated intersection. This simple example was used to gain experience with this method in the stochastic traffic environment and to establish useful ranges for the various parameters involved. Application of Q-learning in a multiagent context to a linear system of traffic signals is now under way, and this will be followed by extension to a two-dimensional road network and signal system. The isolated signal and linear system implementations involve two-phase operation without turning flows. The network implementation will consider turning movements and more flexible phasing arrangements. The following discussion outlines the essential elements of the isolated signal case study and identifies modifications being tested in the linear, multiagent application as a result of insights gained through the initial application.

Description of Test-beds

The isolated traffic signal test-bed consisted of a simulated two-phase signal controlling the intersection of two two-lane roads. Vehicle arrivals were generated using individual Poisson processes with predefined average arrival rates on each of the four approaches. The average rates could be varied over time to represent different peak-period traffic profiles over the 2 h simulated episodes. In practice, the agent would operate continuously.

In the case of the linear signal system, autonomous Q-learning agents, each controlling a single intersection with two-phase control similar to that used in the isolated signal case, comprise the test-bed. Individual Poisson processes are used to generate vehicle arrivals on each approach to the system. Traffic movement within the system is simulated at a microscopic level. Each road link is divided into blocks; vehicles advance one block per time step, provided the downstream block is not occupied. In cases where there is insufficient space on the downstream link to exit an intersection, vehicles may enter the intersection probabilistically and be trapped there until there is an opportunity to move ahead. This may block following and crossing flows, and allows heavily congested traffic conditions to be simulated more realistically.

State, Action, and Reward Definitions

In the case of the isolated intersection, the state information available to the agent included the queue lengths on the four approaches and the elapsed phase time. The multiagent situation permits additional state information, since communication between agents extends the effective field of view of individual agents. In addition to local queue lengths, various combinations of upstream and downstream queue lengths and the offset of signal changes controlling upstream and downstream movements are being evaluated as state elements. Since the addition of elements to the state definition dramatically increases the size of the state-space, a balance has to be sought between the benefit of this information and its impact on problem tractability. Sensing of queue lengths would be most effectively achieved using video imaging technology in combination with artificial neural network or other pattern recognition techniques.

The isolated signal agent was operated with a fixed cycle length as context. Each second, between a point 10 s into the cycle and a point 10 s from the end of the cycle, arbitrary limits fixed to ensure minimum practical phase lengths, the agent selected an action—either remain with the current signal indication or change it. Considering the potential need to transmit, receive, and process communicated information, 1-s intervals between action-selection decisions were not considered to be sufficiently flexible in the case of the multiagent system. In this case, action selection consists of a decision, made at the time of the previous phase change, as to when to make the next phase change. To provide additional flexibility, cycle lengths are not fixed in this case, but minimum and maximum limits are placed on phase lengths, as before, to ensure practicality. In the case where projected phase-change times at adjacent signals are included as state elements, the agents are provided with an opportunity to respond to this information. At the time of a phase change, and decision on the time of the subsequent change, by any individual agent, the other agents, in order of adjacency, are provided with an opportunity to review and adjust their currently projected change times if the increase in benefit would exceed a minimum threshold. Where these review points are insufficiently close in time, intermediate reviews can be scheduled to allow any significant changes in state to be considered as they occur. This review process is repeated, as required and as time permits, in an attempt to reach equilibrium. In both the single and multiagent settings, it is possible to constrain phase lengths so that they do not vary by more than a prespecified time from the previous phase length. This may be seen as desirable to limit variability in successive cycles, although some degradation of performance is likely.

The definition of reward (actually a penalty in this case) is relatively straightforward in the single-agent case, being the total delay incurred between successive decision points by vehicles in the queues of the four approaches. The delay in each 1-s step, being directly proportional to the queue length, was modified using a power function to encourage approximate balancing of queue lengths. Otherwise, the agent was found to be indifferent between situations involving very long and very short queues and situations involving equal-length queues, both with the same average queue length and therefore delay. In the multiagent case, a key issue is the extent to which *global* rewards (or penalties) are necessary to promote cooperation among the agents. It is hypothesized that interacting agents must respond to a reward structure that incorporates not only *local* rewards, as in the single-agent case, but also global rewards to avoid agents acting solely on the basis of self-interest and compromising overall effectiveness and efficiency. In addition to the local reward used by the isolated agent, alternative global reward formulations are being investigated for the multiagent case, including delay and the incidence of intersection blockage along the main streets and across the network. Weighting of the global rewards relative to local rewards is also being evaluated.

It is possible to define rewards or penalties related to other objectives or priorities. For example, the throughput of the intersection could be used in place of, or in combination with, delay. Delay (or throughput) on main roads could be weighted more heavily than that on lesser streets. Vehicle emissions or fuel consumption could also be incorporated, given suitable methods for their estimation.

Exploration Policies

Convergence of a Q-learning agent on a suitable Q-function, particularly where the process being controlled is stochastic, requires adequate exploration to ensure that all areas of interest across the state space are visited sufficiently often. Limiting attention too soon to promising early results may mean that the optimum policy is not discovered. The single agent test-bed was used to test several exploration policies. An ε-greedy policy was tested, where the *best* action is exploited with probability ε and an exploratory action is chosen randomly with probability $\mathbf{1}-\varepsilon$. A range of values for ε was evaluated, and a value of 0.9 was found to yield good results. A softmax exploration policy was also tested, where the probability of choosing an action was proportional to the Q-estimate or value for that action given the current state. Good results were achieved where the probability of choosing the best action was *annealed*, starting with random exploration and increasing to the point where the best action was chosen with a probability of 0.9, provided that state had been visited at least 35–50 times. Both techniques required that the definition of random exploration be modified to avoid exploratory change actions being implemented consistently within the first few seconds of the phase. With the change in the action-space for the multiagent test-bed, the standard softmax procedure is being used, although further testing is necessary to ensure that the shift from exploration to exploitation is sufficiently gradual so as not to inhibit convergence.

Function Approximation and Generalization

In both the single-agent and multiagent test-beds, the Cerebellar Model Articulation Controller (CMAC), as pioneered by Albus (1975a, b), is used for storage and generalization of the Q-estimates. The CMAC is conceptually similar to an artificial neural network, although the implementation used in this case, as described by Smith (1998), operates more like a sophisticated look-up table. The CMAC fulfills a function approximation and generalization role by allowing Q-estimates for any given state-action pair to influence those of nearby state-action pairs. This effectively smooths the decision hypersurface and enables Q-estimates to be derived for state-action pairs not yet visited, but similar to pairs that have been visited. The actual storage of Q-estimates was accomplished using hash tables. This minimizes memory requirements, since the high-dimensional arrays required, one dimension for each element of the state-space, are typically sparsely populated.

In the case of the isolated intersection, two CMACs were used—one for the *change* action and one for the *don't-change* action. Testing showed that 11 association units or layers in the CMAC, in combination with a resolution of 50% (mapping two adjacent queue lengths—for example, 23 vehicles and 24 vehicles—into the same Q-estimate), yielded good results without requiring excessive memory for the storage of the Q-estimates. Despite the fact that the CMAC implies nonlinear function approximation, possibly problematic in the case of Q-learning in a stochastic environment, lack of convergence did not appear to be an issue. Various values for the training rate $\alpha_{s,a}$ were tested, and the best results were achieved when $\alpha_{s,a}$ was gradually decreased in inverse proportion to the number of visits to that particular state **s** and action **a**. The multiagent structure is designed so that each agent, and therefore each intersection, employs two CMACs—one for changes from a green to a red indication, and one for changes from red to green. Theoretically, all agents could share a single pair of CMACs, implying faster training. However, this limits effective learning of optimal policies in cases where local environments (road section lengths, road configuration, intervening side streets or major generators, etc.) are dissimilar and may result in oscillation of the Q-estimates and hindering of convergence.

Multiagent Architecture and Communication Strategy

One of the key issues being explored with the multiagent test-bed is the role of communication between agents. While it is hypothesized that communication should expand the agents' perceptual horizons and their ability to cooperate toward a globally optimal policy, it is also recognized that excessive information can increase the dimensionality of the problem, increase the computational burden, and reduce robustness should the communication system malfunction. In this research, the benefits of various levels of communication are being compared to each other and to the baseline case without communication, where each agent has only local sensory inputs and acts independently.

There are several key opportunities to incorporate information communicated between agents. The first involves the state of the environment where, as noted previously, tests are being conducted on alternative state definitions that include different forms of communicated information. The communication of intended actions in the form of projected phase-change times provides an opportunity for cooperative, real-time review of proposed actions as the environment changes. Another application of communicated information is in the reward structure where, as discussed earlier, the inclusion of weighted global rewards may assist in convergence toward a policy that is optimal, considering the entire network. In the multiagent test-bed, it is likely that the actions of one agent in the system have an impact on not only adjacent agents, but others as well. If an agent allows a local queue to build up so that it extends through upstream intersections, both following and cross-street traffic may be blocked.

Preliminary Test Results

The following discussion presents selected results obtained from the isolated signal test-bed. In this case, performance was compared with that of a commonly used pretimed signal controller. Comparison with semi- or fully actuated controllers might be considered a more appropriate test of performance, but, in the heavily congested conditions that are the subject of this research, these typically default to what is essentially pretimed control. At the time of writing, testing with the multiagent test-bed had not progressed to the point where useful conclusions could be drawn. These will be reported on at a later time. In the multiagent case, comparisons will be drawn with other commonly used signal system control methodologies such as through-bandwidth maximization and TRANSYT (off-line) and SCOOT (on-line).

Tests were conducted using three different traffic profiles to evaluate the performance of the Q-learning agent under varying conditions. Fig. 3 summarizes the results of these tests. The graphs in Fig. 3 reflect the average vehicular delay across individual sets of 50 test episodes, typically conducted after each of 10, 25, 50, 100, 150, 200, 250, and 500 training episodes. Each training and testing episode was equivalent to a 2-h peak period involving 144 signal cycles. In accordance with typical practice, the pretimed signal-phasing plan used as a baseline for comparison utilized constant phase times based on the critical peak-hour flow rates by direction.

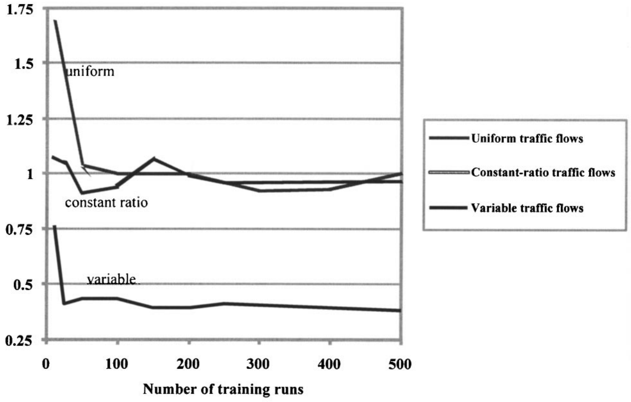

Fig. 3. Isolated traffic signal: Average delay per vehicle ratio (Q-learning/pretimed)

Where traffic flow rates were uniform across the approaches at any given point in time, or where there was a constant ratio between the main and side-street flow rates, the Q-learning agent performed generally on a par with, or slightly better than, the pretimed signal controller. With the uniform flow rates, the differences in mean delay per vehicle between pretimed operation and the results obtained by the Q-learning agent were not statistically significant beyond 50 training episodes. In the case of constant-ratio flow rates, the differences in mean delay per vehicle were statistically significant in favor of the Q-learning agent between 300 and 400 training episodes. The similarity in performance of the pretimed and Q-learning approaches does not argue against the effectiveness of the latter, since these conditions are amenable to pretimed signal control. That the Q-learning agent was able to outperform the pretimed controller at all under these conditions was due to its ability to adapt to minor random fluctuations in flow. The initially higher delays for the Q-learning agent reflect the early stages of training, before the Q-function has stabilized. Starting with zeroed initial Q-estimates, the Q-learning agent was able to achieve effective and reasonably stable performance within 200–400 training episodes. Where the traffic flows were more variable, the Q-learning agent produced delays that were only 38–44% of those obtained using pretimed signal control and outperformed the pretimed controller over virtually 100% of the test episodes.

Tests were also conducted with smoothed signal changes, where the difference between subsequent phase lengths was limited. When the maximum difference was set to 6 s, the delay understandably increased, typically by 5–10%. The ability of the agent to generalize to different cycle lengths was also evaluated. The average delays associated with doubling the cycle length were significantly higher, although this is an extreme case not usually contemplated in practice. This result, in part, motivated the use of a flexible cycle length for the multiagent test-bed.

Implementation Issues

Implementation of reinforcement-learning-based signal control systems is contemplated primarily for the multiagent, signal network situation, as it is in this case that the benefits of learning should be most apparent and useful. Several key deployment issues need to be addressed, as discussed next.

It is envisaged that the agents would be pretrained on a simulator prior to actual deployment. Tests to date with the isolated signal test-bed have shown that pretraining requirements are not onerous. Analogous testing with the multiagent test-bed is required to determine pretraining requirements and to ascertain how close to expected operational conditions the simulated pretraining scenarios would have to be to ensure reasonable generalization to the expected range of operating conditions upon deployment. Once deployed, the agents would be programmed to continue their training to refine the Q-estimates based on actual environmental and operating conditions. Exploration would be necessary in the initial stages of deployment, but this exploration should be incremental and should not produce control decisions that appear to drivers to be obviously inappropriate to the situation. A ‘‘continuing education’’ strategy should also be developed that will enable the agent to assess its performance on an ongoing basis in light of possibly changing conditions and determine when and how much additional on-line training may be required.

A sensory subsystem is needed to provide the required inputs to the agents. This may involve adaptation of existing induction loop technology, although a video imaging system would likely be more effective. A contingency plan would also be required to deal with a potential loss of communications. Evaluation of simulated agent performance using only local state inputs, but in a multiagent context, is planned to provide insights into a possible strategy for this scenario.

Future Research

Following the completion of the current evaluation of reinforcement-learning-based signal control for a linear signal system, extension to a two-dimensional network will be pursued. To fully evaluate the reinforcement-learning approach, it will be necessary to compare its performance with that of state-of-the-art control methodologies, such as SCOOT.

The final stage of this ongoing research effort involves integrating the multiagent traffic control system with dynamic route guidance, also based on reinforcement learning. This is seen as a two-way interaction. Collective perceptions of the Q-learning agents concerning the distribution of congestion across the network could be used as a basis for advising drivers of less-congested routes using variable-message signs, local-area radio broadcasts, or other means. The other side of this interaction would involve the real-time adaptation of the agents across the network to the changes in traffic flows resulting from the reaction of drivers to the guidance information, thus completing the feedback loop. Again, the ability of Q-learning to provide adaptive, real-time control is seen as the key to the effective integration of dynamic route guidance with traffic signal system control.

Conclusions

Reinforcement learning appears to offer significant advantages in the application to transportation processes where real-time, adaptive control is the key to improving effectiveness and efficiency. The ability to learn through dynamic interaction with the environment is seen as a significant benefit relative to control methodologies that rely on prespecified models of these processes.

The current research effort outlined in this paper, one phase of which was presented as a case study, involves the application of reinforcement learning to the problem of traffic signal control, with particular emphasis on heavily congested conditions in a two-dimensional road network. Preliminary results from the application of Q-learning to an isolated, two-phase traffic signal are encouraging. The Q-learning agent performed on a par with pretimed signals under traffic conditions amenable to pretimed control, involving constant or constant-ratio flow rates. Under more variable traffic conditions, the Q-learning agent demonstrated marked superiority due to its ability to adapt to changing circumstances.

Research is currently under way to extend the reinforcement-learning approach to a linear signal system and will be reported on in the near future. Subsequent phases of this research effort will involve extension to control of a two-dimensional system of traffic signals and the integration of traffic signal control based on Q-learning with dynamic route guidance. Comparison of the Q-learning approach to traffic signal control with existing state-of-the-art methods such as SCOOT will also be pursued.

Acknowledgments

The second writer wishes to acknowledge the financial assistance provided by the Natural Science and Engineering Research Council of Canada and the University of Toronto.

References

Abdulhai, B., and Ritchie, S. G. (1999a). ‘‘Enhancing the universality and transferability of freeway incident detection using a Bayesian-based neural network.’’ *Transportation Research—Part C,* 7, 261–280.

Abdulhai, B., and Ritchie, S. G. (1999b). ‘‘Towards adaptive incident detection algorithms.’’ *Proc., 6th World Congress on Intelligent Transport Systems*.

Albus, J. S. (1975a). ‘‘Data storage in the Cerebellar Model Articulation Controller (CMAC).’’ *J. Dyn. Syst., Meas., Control,* 97, 228–233.

Albus, J. S. (1975b). ‘‘A new approach to manipulator control: The cerebellar model articulation controller (CMAC).’’ *J. Dyn. Syst., Meas., Control,* 97, 220–227.

Bertsekas, D. P., and Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*, Athena Scientific, Belmont, Mass.

Bingham, E. (1998). ‘‘Neurofuzzy traffic signal control.’’ Master’s thesis, Dept. of Engineering Physics and Mathematics, Helsinki Univ. of Technology, Helsinki, Finland.

Bretherton, D. (1996). ‘‘Current developments in SCOOT: Version 3.’’ *Transportation Research Record 1554*, Transportation Research Board, Washington, D.C., 48–52.

Bretherton, D., Wood, K., and Raha, N. (1998). ‘‘Traffic monitoring and congestion management in the SCOOT urban traffic control system.’’ *Transportation Research Record 1634*, Transportation Research Board, Washington, D.C., 118–122.

Gartner, N. H., and Al-Malik, M. (1996). ‘‘Combined model for signal control and route choice in urban traffic networks.’’ *Transportation Research Record 1554*, Transportation Research Board, Washington, D.C., 27–35.

Hunt, P. B., Robertson, D. I., Bretherton, D., and Winton, R. I. (1981). ‘‘SCOOT—A traffic responsive method of coordinating signals.’’ *Laboratory Rep. 1014*, Transport and Road Research Laboratory.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). ‘‘Reinforcement learning: A survey.’’ *J. Artif. Intell. Res.,* 4, 237–285.

Sadek, A. W., Smith, B. L., and Demetsky, M. J. (1998). ‘‘Artificial intelligence-based architecture for real-time traffic flow management.’’ *Transportation Research Record 1651*, Transportation Research Board, Washington, D.C., 53–58.

Sen, S., and Head, K. L. (1997). ‘‘Controlled optimization of phases at an intersection.’’ *Transp. Sci.,* 31(1), 5–17.

Smith, R. (1998). ‘‘Intelligent motion control with an artificial cerebellum.’’ PhD thesis, Dept. of Electrical and Electronic Engineering, Univ. of Auckland, Auckland, New Zealand.

Spall, J. C., and Chin, D. C. (1997). ‘‘Traffic-responsive signal timing for system-wide traffic control.’’ *Transp. Res., Part C: Emerg. Technol.*, 5(3/4), 153–163.

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement learning—An introduction*, MIT Press, Cambridge, Mass.

Thorpe, T. L. (1997). ‘‘Vehicle traffic light control using SARSA.’’ *Master’s Project Rep.*, Computer Science Dept., Colorado State Univ., Fort Collins, Colo.

Watkins, C. J. C. H. (1989). ‘‘Learning from delayed rewards.’’ PhD thesis, King’s College, Univ. of Cambridge, Cambridge, U.K.

Watkins, C. J. C. H., and Dayan, P. (1992). ‘‘Q-learning.’’ *Mach. Learn.*, 8, 279–292.

Yagar, S., and Dion, F. (1996). ‘‘Distributed approach to real-time control of complex signalized networks.’’ *Transportation Research Record 1554*, Transportation Research Board, Washington, D.C., 1–8.