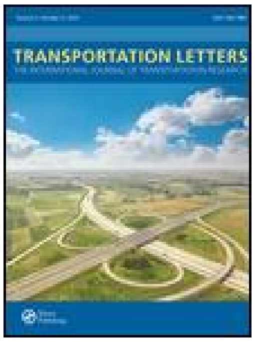

Taylor & Francis
Taylor & Francis Group

Transportation Letters
The International Journal of Transportation Research

ISSN: 1942-7867 (Print) 1942-7875 (Online) Journal homepage: https://www.tandfonline.com/loi/ytrl20

Towards multi-agent reinforcement learning for integrated network of optimal traffic controllers (MARLIN-OTC)

Samah El-Tantawy & Baher Abdulhai

To cite this article: Samah El-Tantawy & Baher Abdulhai (2010) Towards multi-agent reinforcement learning for integrated network of optimal traffic controllers (MARLIN-OTC), Transportation Letters, 2:2, 89-110, DOI: 10.3328/TL.2010.02.02.89-110

To link to this article: https://doi.org/10.3328/TL.2010.02.02.89-110

Published online: 07 Sep 2013.

Samah El-Tantawy[1*] *and Baher Abdulhai*[2]

Towards multi-agent reinforcement learning for integrated network of optimal traffic controllers (MARLIN-OTC)

ABSTRACT: Traffic congestion can be alleviated by infrastructure expansions; however, improving the existing infrastructure using traffic control is more plausible due to the obvious financial resources and physical space constraints. The most promising control tools include ramp metering, variable message signs, and signalized intersections. Synergizing the aforementioned strategies in one platform is an ultimate and challenging goal to alleviate traffic gridlock and optimally utilize the existing system capacity; this is referred to as *Integrated Traffic Control* (*ITC*). *Reinforcement Learning* (*RL*) techniques have the potential to tackle the optimal traffic control problem. *Game Theory* (*GT*) fits well in modelling the distributed control systems as multiplayer games. *Multi-Agent Reinforcement Learning* (*MARL*) achieves the potential synergy of *RL* and *GT* concepts, providing a promising tool for optimal distributed traffic control. The objective of this paper is to clarify the opportunities of game theory concepts and *MARL* approaches in creating an adaptive optimal traffic control system that is decentralized but yet integrated through agents' interactions. In this paper, we comparatively review and evaluate the relevant existing approaches. We then envision and introduce a novel framework that combines *GT* concepts and *MARL* to achieve a *Multi-Agent Reinforcement Learning for Integrated Network of Optimal Traffic Controllers* (*MARLIN-OTC*).

KEYWORDS: Traffic Control, Reinforcement Learning, Game Theory, Multi-Agent Reinforcement Learning

LIST OF ABBREVIATIONS:

ITC: Integrated Traffic Control
RL: Reinforcment Learning
GT: Game Theory
MARL: Multi-Agent Reinforcement Learning
MARLIN-OTC: Multi-Agent Reinforcement Learning for Integrated Network of Optimal Traffic Controllers
ITS: Intelligent Transportation Systems
RM: Ramp Metering
VMS: Variable Message Signs
SI: Signalized Intersections
MAS: Multi-Agent Systems
OOC: Open-Loop Optimal Control
COC: Closed-Loop Optimal Control
DP: Dynamic Programming
TD: Temporal Difference
CBRL: Case-Based Reinforcement Learning
NAC: Natural Actor Critic
CRL: Collaborative Reinforcement Learning
CS: Coalition Structure
SG: Stochastic Game
CoLF: Change or Learn Fast
OAL: Optimal Adaptive Learning
FP: Fictitious Play
ARR: Adaptive Round Robin
ARR-CRL: *CRL*-based *ARR*
ARR-RL: *RL*-based *ARR*
VSC : Variable Speed Control
SFP: Sampled Fictitious Play
SRL: Strategic Reinforcement Learning
TRL: Tactical Reinforcement Learning

*Corresponding Author

[1] PhD Candidate, Toronto ITS Centre and Testbed, Department of Civil Engineering, University of Toronto, 35 St. George St., Toronto, Ontario, Canada M5S 1A4, Email: samah.el.tantawy@utoronto.ca, Tel. 416-978-5049, Fax. 416-978-5054

[2] Ph.D., P.Eng., Canada Research Chair in ITS, Director, Intelligent Transportation Systems Centre and Testbed, Department of Civil Engineering, University of Toronto, 35 St. George St., Toronto, Ontario, Canada M5S 1A4, Email: baher.abdulhai@utoronto.ca, Tel. 416-946-5036, Fax. 416-978-5054

Transportation Letters: The International Journal of Transportation Research (2010) 2: (89-110)
DOI 10.3328/TL.2010.02.02.89-110

1. INTRODUCTION

Population is steadily increasing worldwide. Consequently the demand for mobility is increasing, traffic congestion is deteriorating severely, and undesirable changes in the transportation environment (e.g, air pollution, greenhouse gases, speed reductions, delays, and safety problems) are becoming major concerns for economies and societies. Infrastructure improvements have been used primarily to handle congestion surge until relatively recently. However, tight constraints on financial resources and physical space, as well as environmental considerations, have accentuated the consideration of a wider range of options.

Therefore, the emphasis has been shifted to improving the existing infrastructure by optimizing the utilization of the available capacity. *Advancements in Intelligent Transportation Systems* (*ITS*) have the potential to significantly reduce delay and alleviate traffic congestion through several strategies. *ITS* endeavours to achieve an optimum management plan for the transportation systems using telecommunication and information technology, and advanced control techniques. The essence of *ITS* relies on the need to dynamically control the transportation networks. The most common control tools include *Ramp Metering* (*RM*) to control the entrance flow to the freeway, *Variable Message Signs* (*VMS*) to divert traffic at bifurcation points, and surface streets' *Signalized Intersections* (*SI*). The term agent is used to denote an intelligent control system which acts according to the state of its environment. This state is typically measured using sensors; for example, *RM*, *VMS*, and *SI* are agents in the traffic control environment in which the sensors could be loop detectors, cameras, etc. The independent use of these control strategies might limit their potential benefits. Therefore, *Integrated Traffic Control* (*ITC*) by combining *RM*, *VMS*, and *SI* simultaneously can be synergetic and beneficial. However, such integration certainly adds more complexity to the problem.

One of the promising approaches to optimal control of traffic networks is *Reinforcement Learning* (*RL*) (Abdulhai *et al*. 2003; Liu, 2007). Due to the stochastic nature and the dynamics incorporated in traffic networks, traffic control problems can be conveniently cast as an *RL* problem. The control system would learn from its own experience and adapt itself to the environment (Abdulhai and Kattan, 2003).

In *RL*, a control *agent*, the learner or the decision-maker, learns optimal strategies directly through feedback *reward* from its *environment*. One key property of *RL* that distinguishes it from other control approaches is the fact that the control agent can learn the mapping between control actions and their corresponding effects on the traffic network environment without a pre-specified model of the environment. This property is crucial for traffic control problems since it is hard to accurately model the transport system.

The integration level of the traffic control strategies (*RM*, *VMS*, and *SI*) can be classified as follows (Wang et al., 2007):

Level 0: Completely independent design and operation of control strategies.

Level 1: Extends Level 0 to account for one of the following properties:

- Considering common global reward
- Considering common global measurements
- Exchanging the rewards in real time.

Level 2: Extends Level 1 to account for two or more of the above properties.

Level 3: Extends Level 2 to include a coordination mechanism that coordinates the actions of the individual strategies in real time.

Level 4: Fully integrated design and operation of a central control unit that accounts for all control strategies.

In traffic control applications, coordination has been typically approached in a centralized way (Level 4) (Hunt et al., 1981; Sims and Dobinson, 1979). However, this is doable only if the communication channels are available and efficiently used for the coordination mechanism without consuming too much processing and communication resources. Therefore, this type of coordination is still infeasible in most cities due to the real-time and interoperability constraints. Therefore, decentralization using *Multi-Agent Systems* (*MAS*) is promising to allow for this coordination to emerge even in the absence of a central authority. The following are several potential advantages that the decentralized *MAS* offers over centralized control (Shohman et al., 2003):

- Computationally less demanding, due to the parallel computation of *MAS* in which each agent explores its action space. In contrast, the centralized agent explores the joint action space which is exponentially larger.
- Robustness is plausible in *MAS* because if one or more agents fail, the remaining agents will continue their tasks.
- Scalability, since *MAS* allows easy insertion of new agents into the system.
- Dimensionality is less in instances where not all the state information is relevant to all the learning agents. For instance, in a team of signalized intersections at a given time, the queue lengths for the intersection approaches that are quite far from the intersection of interest might not have a great influ-

ence on it; in such cases, the learning agents can consider only the relevant state components and therefore further decrease the dimensionality of the problem.

MAS is a sub-field of *Artificial Intelligence* (*AI*) that provides principles for constructing complex multi-agent systems and mechanisms for coordinating the behaviour of independent agents. Although achieving cooperation and coherent coordination of *MAS* is not a trivial task, *Game Theory* (*GT*) provides the tools to model the *MAS* as a multi-player game and provide the rational strategy to each player.

Multi-Agent Reinforcement Learning (*MARL*) combines *RL* and *GT* concepts. The decentralized traffic control problem is an excellent testbed for *MARL* because of the inherent dynamics and stochastic nature of the traffic system (Bazzan, 2009). Achieving coordinated decentralized traffic control and at the same time reducing communication requirements seem to be conflicting goals, since the more we attempt to integrate the actions of decentralized agents, the more the communication requirements. Given this challenging problem and conflicting objectives, this paper endeavours to bridge this gap by highlighting the *MARL* approaches that seek coordination with reduced communication and which have not yet been investigated in the traffic control problem, especially the *ITC* problem.

Therefore, the objective of this paper is to identify the opportunities of *GT* concepts and *MARL* approaches for *MARLIN-OTC*.

The main contributions of this paper with respect to the current state of the art are:

- Amalgamating, summarizing and categorizing the current *RL* approaches
- Highlighting the *MARL* approaches that seek coordination with reduced communication
- Developing a taxonomy for the recent traffic control studies that have utilized *GT* concepts and/or *RL* algorithms
- Evaluating recent traffic control studies based on the integration level and the solution optimality
- Proposing a framework for a distributed traffic control system that attempts to optimally integrates all control units and seeks coordination between these units.

In the proposed framework, the decentralization process is represented by a coalition form game. The coalition formation process aims to find the best combination of agents to cooperate in solving the traffic control problem. Within each coalition, coordination is achieved through solving the internal game between coalition members.

Section 2 reviews the theoretical foundation of Automatic Control, Dynamic Programming, and *RL* for the single agent case. Brief descriptions of *GT* and *MARL* approaches are then presented in Section 3. Section 4 provides a review of the traffic control problem in the light of classical *GT* and *RL*. A summary of the research contributions and gaps in the literature review is presented in section 5. A brief description of the proposed framework is presented in Section 6. A hypothetical example that illustrates the mapping between *GT* concepts and the traffic control problem is described in Section 7, and finally a summary, conclusions, and future work are presented in Section 8.

2. SINGLE AGENT: AUTOMATIC CONTROL, DYNAMIC PROGRAMMING, AND REINFORCEMENT LEARNING THEORETICAL FOUNDATION

2.1 Automatic Control Theory

A controller aims to specify, based on available measurements, the control strategy so as to achieve a pre-specified goal despite the influence of various disturbances. In optimal control, the control goal can be expressed as the minimization or maximization of an objective function that depend on the *inputs* and the *state variables* of the process.

The state variables s depend upon the inputs a and disturbances d according to a mathematical model. Most of the dynamic processes can be described by the following state space model:

$$s' = f(s,a,d) \quad [1]$$

If the initial condition s_0, that is, the value of s at time $t = 0$, is known and the time trajectories $a(t)$, $d(t)$, $t \in [0,T]$ are also given, the differential equations can be solved to find the corresponding state trajectory $s(t)$ over the same time period $[0,T]$. The choice set of a is usually limited due to the constraints that define the admissible control region.

The optimal control problem may be formulated in the form of a mathematical optimization problem. Papageorgiou (1998) defined two examples of optimal control problems:

Open-Loop Optimal Control (OOC) Problem: Given the initial condition s_0, and the disturbance trajectory $d(t)$, $t \in [0,T]$ find the input trajectory $a(t)$, $t \in [0,T]$ that minimizes the criterion J subject to the model equations and constraints.

The solution of the *OOC* problem, $a^*(t)$, is applied during the period $[0,T]$ and is optimal only for the specific initial conditions. Unexpected disturbances could lead to totally undesirable results.

Closed-Loop Optimal Control (COC) Problem: Given the disturbance trajectory $d(t)$, $t \in [0,T]$, find a function R, $a(t) = R[s(t), t]$, $t \in [0,T]$ that minimizes the criterion J subject to the model equations and the constraints.

The solution of *COC* problems is the function $R(s, t)$, which is defined as the control law (policy). $R(s, t)$ is a function of feedback state measurements. Therefore, the solution is independent of the initial condition and hence it is applicable anywhere in the space (s, t). Analytical solutions are only feasible for simple problems; therefore, the need exists to utilize numerical solutions such as *Dynamic Programming* (*DP*) techniques to model and solve more complex problems

The stochastic nature of the traffic system limits the application of open-loop strategies where the control strategy should be adjusted according to the most recent traffic conditions. Therefore, the traffic control problem fits well in the *stochastic control problem* in which a *COC* solution is approached using *DP* methods (Bertsekas, Shreve, 1978). The next sections highlight the stochastic control problem and the *DP* approach.

DP is the main tool to solve the infinite horizon stochastic control problem. Value iteration and policy iteration methods are two *DP* techniques that can be used to find the optimal policy (Sutton and Barto, 1998). The policy iteration method converges in a smaller number of iterations compared to the value iteration method. However, the policy iteration method requires solving a system of linear equations in each iteration.

Compared to classical linear and nonlinear optimization algorithms, *DP* can easily handle nonlinearities. In addition, in contrast to traditional methods, with the increase in the number of constraints; *DP* reaches the optimal solution faster. This is primarily due to the reduction in the action

Despite the former advantages, the main challenge of *DP* is the curse of dimensionality in which it requires enormous storage space to allow for large matrices of transition probabilities and rewards. This limits its application to simple structures; otherwise, the computation time becomes practically infeasible. In addition, *DP* requires pre-specified perfect values of the transition probabilities and rewards for every state–action pair which represents the theoretical model of the environment, the existence of which is questionable for traffic networks. Unfortunately, because of these limitations, an optimal solution to the *DP* problem is impractical (Sutton and Barto, 1998).

2.2 Reinforcement Learning

Reinforcement learning combines the two fields of supervised learning and *DP*, yielding a powerful machine learning system. *RL* is distinguished from the supervised learning approaches by its emphasis on learning from direct interaction with the environment. However, in supervised learning the learning system is provided with a set of representative training examples in the form of input–output pairs, which is often impractical to obtain in interactive environments such as traffic networks (Sutton and Barto, 1998). On the other hand, *RL* endeavours not only to achieve as much as *DP* but also to do so with less computation and without always assuming a perfect model of the environment.

The most common single agent *RL* algorithm is Q-learning (Watkins and Dayan, 1992). The Q-Learning agent learns the optimal mapping between the environment's state s and the corresponding optimal control action a based on accumulating rewards $r(s,a)$. Each state-action pair (s,a) has a value called Q-Factor that represents the expected reward for the state-action pair (s,a). In each iteration, k, the agent observes the current state s, chooses and executes an action a that belongs to the available set of actions A, and then the Q-Factors are updated according to the reward $r(s,a)$ and the state transition to state s' as follows (Sutton and Barto, 1998);

$$Q^k(s,a) = (1-\alpha)Q^{k-1}(s,a) + \alpha[r(s,a) + \gamma \max_{a' \in A} Q^{k-1}(s',a')] \qquad [2]$$

where a, $\gamma \in (0,1)$ referred to as the learning rate and discount rate, respectively.

The agent can simply choose the greedy action at each iteration based on the stored Q-Factors, as follows;

$$a \in \arg \max_{b \in A} [Q(s,b)] \qquad [3]$$

However, the sequence Q^k is proven to converge to the optimal value only if the agent visits the state–action pair an infinite number of iterations (Watkins and Dayan, 1992). This means that the agent must sometimes explore (try other actions) rather than exploit the best actions. To balance the exploration and exploitation in Q-Learning, algorithms such as $\in$-greedy and softmax are typically used (Sutton and Barto, 1998).

One important advantage of *RL* is that such algorithms are truly adaptive in the sense that they are capable of responding not only to dynamic inputs from the environment but also to a dynamically changing environment, while continuing the learning and adaptation in the field. Therefore, *RL* appears to offer promising results in application to transportation systems control where optimal real-time adaptive control is a key element in improving the effectiveness and efficiency (Abdulhai et al., 2003; Liu, 2007).

- **Reinforcement Learning Approaches**

Numerous Single Agent RL (*SARL*) algorithms have been investigated in the literature (Sutton and Barto, 1998; Gosavi, 2003); the most relevant algorithms to this study are high-

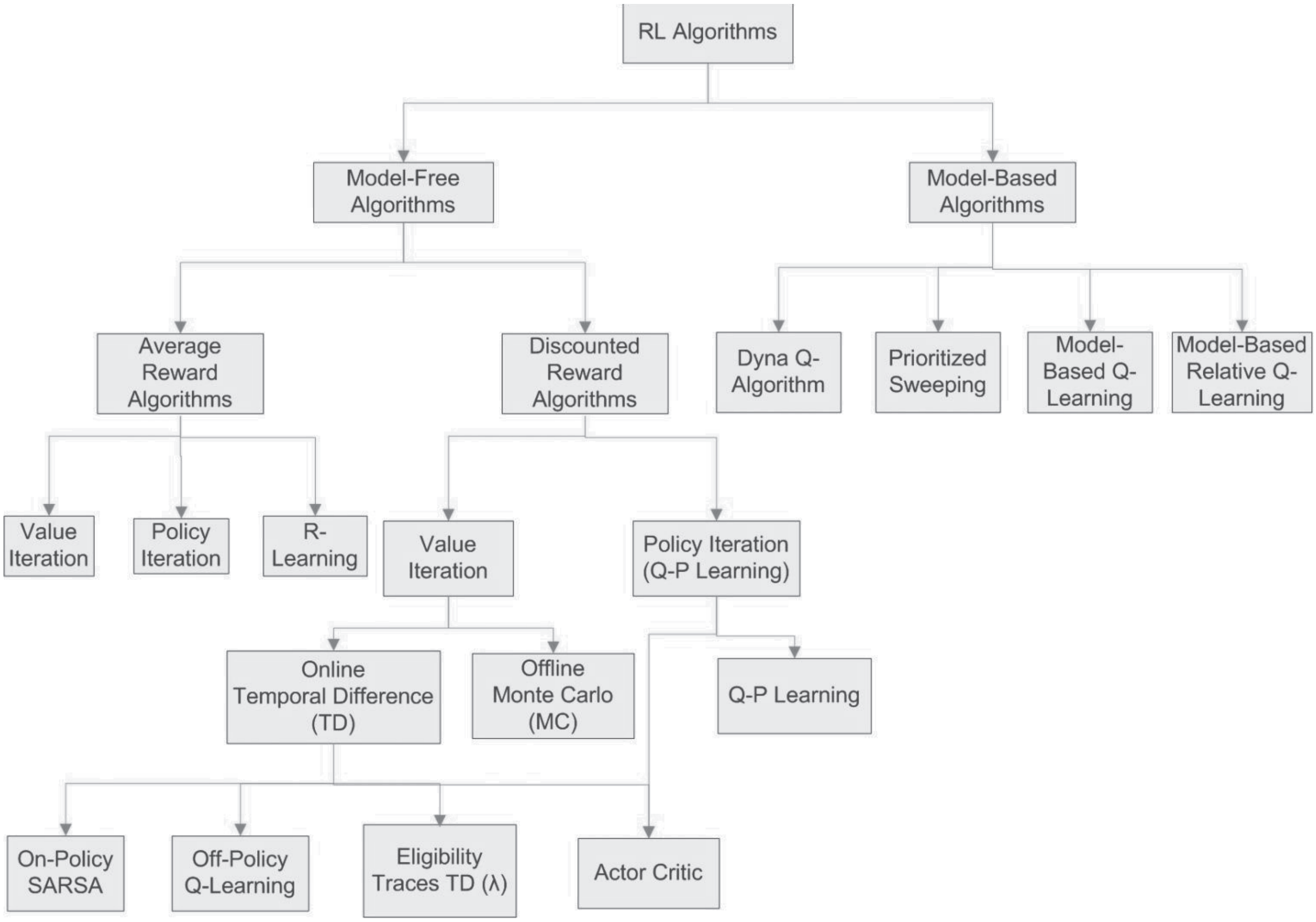

Figure 1. Reinforcement Learning Approaches.

lighted next. *RL* methods can be categorized into model-free and model-based learning algorithms as shown in Figure 1. Model-based algorithms first learn a transition model that estimates the state–action transition probabilities and then compute the rewards associated with these state–action transitions. Second, the algorithm uses a *DP* approach to compute the value function. The Dyna-Q algorithm, Prioritized Sweeping, Model-Based Q-Learning, and Model-Based Relative Q-Learning are examples of model-based algorithms.

On the other hand, model-free learning algorithms can learn without having a transition model in which the algorithm directly updates the value functions according to the observed sequence of state–action pairs during the interaction of the agent with the environment. Model-free approaches can be categorized according to the reward function form into average reward and discounted reward algorithms. For each category, the policy iteration and the value iteration approaches can be used to explore the optimal policy. Value iteration-discounted reward *RL* can run in either offline or online mode according to the internal updating mechanism. In online updating procedures, such as *Temporal Difference* (*TD*) algorithms, the Q-factor of a state–action pair is immediately updated after being visited. On the other hand, in the offline updating procedure, the Q-factors are updated after a finite number of state transitions such as the Monte Carlo algorithm. Four types of *TD* algorithms are presented in the literature, namely SARSA, Q-Learning, Eligibility Traces, and Actor Critic algorithms.

SARSA is an on-policy algorithm in which the agent makes attempts to improve the policy that is used to make decisions. On the other hand, Q-Learning is an off-policy in which the agent attempts to improve a policy while following another one. The Eligibility Traces algorithm is also known as the *TD*(λ) algorithm, where λ refers to a temporary record of the occurrence of an event, defined as the eligibility trace, such as visiting a state or taking an action. The trace marks the memory parameters associated with the event as eligible to undergo the learning changes. If a *TD* error occurs, only the eligible states or actions are assigned credit for that error. The Actor Critic approach combines the characteristics of both the *TD* and the policy iteration algorithms

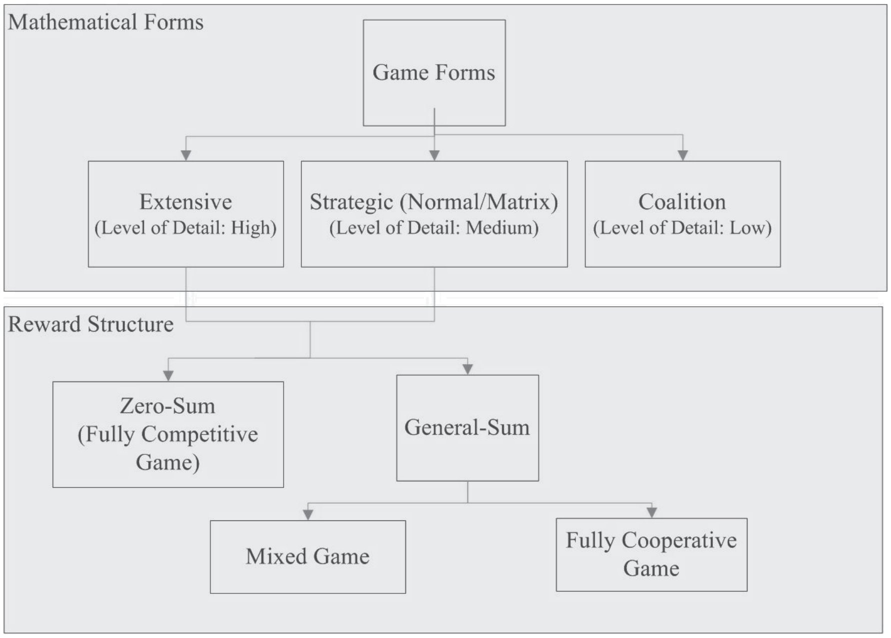

Figure 2. Classification of Games According to Mathematical Forms and Reward Structure.

in which the policy structure is known as the *actor* and the estimated value function is known as the *critic*. As shown in Figure 1, the Q-P Learning algorithm is a policy iteration discounted reward algorithm which is similar to the Q-Learning algorithm in the policy iteration outlook. R-Learning is similar to the Q-Learning algorithm except that it learns a policy by using an average reward function.

3. MULTI-AGENT: GAME THEORY AND REINFORCEMENT LEARNING

As noted in Littman (1994), no agent lives in a vacuum; the agent must interact with other agents in the environment to achieve its target. The traffic system is dynamic and hard to predict not only due to the stochastic nature of the traffic patterns but also because the reward that one agent receives strongly depends on the actions performed by many other agents. Therefore, the traffic control system should be modelled as *MAS*. *MAS* consists of interacting agents that share a common environment in which each player is supposed to behave rationally.

GT provides the tools to formally model and analyse *MAS* as a multi-player game and to further prescribe the rational strategies of the different players (Koller and Pfeffer, 1997) which fits well in the traffic control problem.

Traffic control systems can therefore be framed as generalized Markov Decision Processes in the multi-agent case that is Stochastic Games (Bazzan, 2009). In the next sections, the most relevant *GT* concepts and particularly stochastic games are reviewed (Osborni, 2004; Mendelson, 2004).

3.1 Game Theory

- **Games' Mathematical Forms**

As shown in Figure 2, three main mathematical forms are defined in *GT*: the *extensive form*, the *strategic form*, and the *coalitional form*. What distinguishes each form is the level of detail built into the model. The most detailed form is the extensive form.

❍ **Extensive Form Game**

A game in the extensive form specifies the complete order of actions (along the direction of time) in which players carry out their actions sequentially. If the player knows all the precedent actions by all other players, the game is defined as a *perfect information game*.

❍ **Normal Form Games**

In *strategic form* or *normal form* games, also called *matrix games*, many of the game details are lost. In strategic form games, the players decide on their actions simultaneously or at least without knowing the actions of the others.

As shown in Figure 2, extensive and normal form games can be categorized according to the reward structure of the players. A game is defined as *zero-sum* or *competitive* if the sum of the rewards to the players is zero no matter what actions were chosen by the players. *Non-zero-sum games*, on the other hand, are categorized into two types: *cooperative games* in which all agents have the same reward *and non-cooperative games* or *mixed games*, which are neither fully cooperative nor fully competitive.

❍ Coalition Form Game

If the number of players grows extremely, the strategic form game, although less detailed than the extensive form, becomes too complex for analysis. In the *coalitional form* game, the main features are those of a *coalition* and the *value* of the coalition. The coalitional form game is a part of *cooperative game theory* in which no individual rewards are given.

• Game Theory Solution Concepts

In a case where only one decision maker is involved in the game, the decision maker maximizes the expected reward. However, in *GT* there are many rational players, and therefore it is necessary to explore solution concepts in which all the agents are satisfied, to some extent, with the current action and do not alter this action.

❍ Extensive and Normal Form Games

The solution concept takes the form of equilibrium, which is the best response of an agent to others' actions.

Formally, an action a_i^b is a best response to actions a_{-i} of the other agents if

$$r_i(a_i^b,a_{-i}) \geq r_i(a_i,a_{-i}) \quad \forall a_i \qquad [4]$$

The set of best responses to a_{-i} is denoted $B_i(a_{-i})$

The Nash equilibrium is the best response for all agents (Basar and Olsder, 1999). Formally, a joint action a^*, which regroups the actions for all agents, is a Nash equilibrium if

$$\forall i, a_i^* \in B_i(a_i^*) \qquad [5]$$

where a_i^* is the action of the *i*th agent and a_{-i}^* are the actions of the other agents.

A relatively simple linear program (the minimax algorithm) can be used to achieve such equilibrium (Nash and Sofer, 1996). However, solving equilibria in general-sum games requires a more difficult quadratic programming solution (Filar and Vrieze , 1997).

The Nash equilibrium is not the only equilibrium concept, however; there are many other concepts, such as Stackelberg equilibria. The Stackelberg equilibrium (Basar and Olsder, 1999) is a best response assuming the existence of a hierarchy between agents in which some agents are leaders and some are followers.

❍ Coalition Game

A *Coalition Structure* (*CS*) is a partition of P that contains exhaustive and disjointed coalitions. The *coalition formation* problem endeavours to find the best *CS* to solve a given task.

The formation process consists of the following phases (Chalkiadakis and Boutilier, 2004):

(a) Searching for an optimal coalition structure;
(b) Solving a joint problem in each coalition; and
(c) Distributing the value of the generated solution among the coalition's members; this process is called allocation.

The solution of a coalition game consists of (Milano, 2006):

- A partition of the players, that is, *CS*
- An efficient allocation: The allocation process concerns assigning to each agent at least the gain it may get if the agent is in the singleton coalition.

The group of agents is assumed to maximize its joint reward.

A coalition configuration is called stable if no agent has an incentive to leave its coalition due to its assigned reward. Most of the current coalition formation methods aim at building stable coalitions. Each notion of stability defines particular solution spaces for coalition games which vary depending on the application domain and discipline; among them, for example, are the Shapley-Value, the Core, the Bargaining Set, and the Kernel (Kahan and Rapoport, 1984).

• Stochastic Games

The *Stochastic Game* (*SG*) is played in a sequence of stages. At each stage, the game is in a certain state in which the players select actions and each player receives a reward that depends on the current state and the chosen joint action. This process can be perceived as a *static game* that can take any mathematical game form. The game then moves to a new random state whose distribution depends on the previous state and the joint action chosen by the players. The procedure is *repeated* in the new state and play continues for a finite or infinite number of stages.

Therefore, the *SG* can be perceived as a natural extension of *MDPs* to multiple agents or an extension of repeated static game to multiple states. Thus, stochastic games are the theoretical model for multi-agent learning.

As explained above, *RL* is a well-established field that provides satisfactory convergence and consistent solutions for solving single-agent problems; therefore, *RL* is found to be an attractive approach to tackle multi-agent learning problems.

MARL is an intermediate area between reinforcement learning and game theory that aims to solve the stochastic games. The equilibrium value is the ultimate goal of most *GT* algorithms in which the environment model is required. In contrast, *RL* endeavours to find the agents' optimal policies without assuming a perfect environment model. *MARL* achieves synergy between the two by finding the agents' policies that achieve the game equilibrium without always assuming a perfect model for the environment.

3.2 *MARL* Approaches

In *MARL*, the effect of the agent's action on the environment is non-stationary, i.e. depends on the actions taken by the other agents; therefore, coordination between agents is necessary. The nonstationarity nature of the multiagent learning problem exists due to the fact that all the agents are learning simultaneously. Therefore, each agent is faced with a moving-target learning problem in which the agent's optimal policy changes as the other agents' policies change. Therefore, a coordination mechanism is required so that all agents coherently choose their actions from the optimal joint policy. Achieving this is not trivial, even for the fully cooperative tasks. *MARL* algorithms can be classified according to the level of coordination (Busoniu et al., 2008).

The following section touches upon some of the *MARL* approaches without any pretension of being comprehensive, given the huge number of approaches suggested. Reviews of *MARL* in the literature mainly considered the normal form games. (as in Busoniu et al. (2008)); however, this study considers the other two forms as well (coalition and extensive games), which are very relevant to multiagency and hierarchical traffic control. Also, the present review highlights the *MARL* function approximation algorithms.

- **Extensive Game Approaches**

It is worth noting that most of the work conducted in *MARL* focuses on learning in strategic form games in which the agents, at each step, simultaneously choose their actions. The application of *RL* techniques in extensive form games, especially for games without perfect information, is still largely unexplored.

Huang and Sycara (2003) showed that a group of self-interested agents can learn to play the Stackelberg equilibrium, which is known, in the context of complete-information extensive games, as sub-game perfect equilibrium,by repeatedly reinforcing their previous experiences of success or failure.

Laslier and Walliser (2005) proposed a *MARL* approach that considers repeated finite extensive form games with perfect information. They used the cumulative proportional reinforcement learning rule to stipulate that an agent chooses an action with a probability proportional to the cumulative reward obtained in the past.

Lazaric et al. (2007) investigated the application of *RL* techniques in extensive form games with incomplete information in which a general learning principle (*CoLF*: *Change or Learn Fast*) is introduced. *CoLF* works by opportunely modifying the learning rate to reduce the non-stationary effects induced by explorative actions performed by the other learning agents. Akramizadeh et al. (2009) extended Q-Learning to be used in extensive form games with perfect information, using SPE points. This results in a new version of Markov games called extensive Markov games in which a new concept called associative Q-values is introduced to provide estimation of the SPE actions. Associative Q-values are the probability of reaching a joint action with respect to subsequent agents' preferences.

- **Coalition Game Approaches**

Most of the current coalition formation methods aim at building stable coalitions. The definition of coalition stability varies according to the considered application domain and discipline. Most of the coalition formation algorithms depend on game-theoretic concepts such as the Shapley-Value, the Core, the Bargaining Set, or the Kernel (Kahan and Rapoport, 1984). However, these models assume that the values of coalitions are known with certainty, which is not typically the case in real-life applications such as the traffic control problem.

Chalkiadakis and Boutilier (2004) proposed the Bayesian *RL* model which reflects the uncertain knowledge about the abilities of the coalition members. In this model, the agents maintain explicit beliefs about each other's type, and choose their coalitions according to the value of information (i.e., what can be learned about other agents). They showed that the agents in this framework can reach stable coalition structure given the agents' beliefs while learning the types of their partners and the values of coalitions.

Li and Soh (2004) implemented a *Case-Based Reinforcement Learning* (*CBRL*) approach to a multi-agent coalition formation problem in dynamic, uncertain, real-time, and noisy environments. Their *CBRL* approach integrates case-based reasoning and *RL* to utilize the agent's past coalition formation experience in solving the current problem.

- **Normal Form Game Approaches**
  - **Cooperative Games**

As reviewed in the literature, the *MARL* algorithms are capable of solving cooperative games in many ways (Busoniu et al., 2008). Coordination-free methods such as the Team

Q-Learning algorithm (Littman, 2001), sometimes referred to as distributed Q-Learning (Camponogara and Kraus Jr, 2003), assume that the optimal joint actions are unique, which is rarely the case.

The agents can be indirectly coordinated by learning empirical models of the other agents and adapting to these models. The *Optimal Adaptive Learning* (*OAL*) algorithm (Claus and Boutilier, 1998) applies an indirect-coordination algorithm. *OAL* is the only known algorithm that is proven to converge to optimal joint policies in any cooperative dynamic *SG* (Busoniu et al., 2008).

The agents' action choices can also be directly coordinated using, for example, coordination graphs (Kok et al., 2005) that indicate when coordination between agents is required to prevent the agents from wasting time in unnecessary coordination activities.

❍ **Mixed Game**

Numerous mixed game *MARL* algorithms are designed only for static tasks. In dynamic tasks, most of the algorithms have a common structure based on Q-Learning. However, policies and state values are computed with game-theoretic solvers. Examples of *MARL* algorithms for mixed games include Nash Q-Learning (Hu and Wellman, 2003) and Asymmetric Q-Learning (Könönen, 2003) in which the game theoretic solvers compute Nash, or Stackelberg equilibria, respectively.

If multiple equilibria exist, the equilibrium selection problem arises. For this purpose, some algorithms estimate models of the other agents' strategies or policies and act using some form of best-response to these models, for example the *Fictitious Play* (*FP*) algorithm (Brown, 1951). However, an *FP* algorithm is designed for the static repeated tasks (Wang and Sandholm, 2002). In dynamic games, the Non-Stationary Converging Policies algorithm (Weinberg and Rosenschein, 2004) computes a best response to the models and uses it to estimate state–action values.

❍ **Competitive Games**

The Minimax-Q algorithm (Littman, 1994) employs the minimax optimization problem to compute the policies and values for the static games arising in the states of the *SG* and employs a rule similar to Q-Learning to update the Q-factors of state-action pairs. In case multiple solutions for the minimax-Q exist, the opponent awareness problem arises. An opponent model can be estimated using the M* Algorithm (Carmel and Markovitch, 1996). However, the M* Algorithm is capable of solving only static tasks; on the other hand, Busoniu et al. (2008) introduced an extension to the M* Algorithm for dynamic tasks.

- **Generalization and Function Approximation**

If the Q-factor of each state–action pair is stored individually, it is called a look-up table approach. This approach is suitable for systems with relatively small numbers of states and actions. For complex problems with a very large state–action space, it is impractical to explicitly represent the Q-factors of all the state–action pairs in the memory. Moreover, the agent may perform poorly if it comes across some new states that are never visited during its training or learning processes. A generalization from the previously experienced states to unexplored states has the potential to overcome the drawbacks associated with the look-up approach. The generalization used in conjunction with *RL* is defined as function approximation. Sutton and Barto (1998) offer an overview of value-function approximation in which the approximate function is represented as a linear function of parameters. There is a vector of features corresponding to every state, with the same number of components as the parameter vector. Features may be constructed from the states in different ways; for example, Coarse Coding, Radial Basis Functions, Kanerva Coding, and Tile Coding are among the most common methods. CMAC (Albus, 1975) is a popular non-linear generalization method in which neural networks of adjustable weights can be trained to approximate non-linearities that are not known explicitly.

In *MARL*, the complexity is also exponential in the number of agents, because each agent adds its own variables to the joint state–action space. Although most *MARL* algorithms are applied only to small problems such as static games, a fair number of approximate *MARL* algorithms have been proposed for large state–action spaces (Fujita and Matsuo, 2005; Abul et al., 2000). Although that function approximation-based *MARL* does not guarantee a global optimal as it would be in a fully dynamic programming approach, techniques discussed in (Lin, 1992) can help reduce the likelihood of this issue.

4. TRAFFIC CONTROL APPROACHES USING *GT* CONCEPTS AND *RL* TECHNIQUES

An essential distinction between vehicle-oriented and road-oriented traffic control problems should be highlighted first (Katwijk et al., 2005):

- *Vehicle-oriented traffic control* focuses on controlling traffic through vehicle based (internal) signals such as advanced driver assistance systems.
- *Road-oriented traffic control* focuses on controlling traffic through external road-based measures at fixed locations, such as traffic signals and variable message signs.

This paper focuses on the road-oriented traffic control problems. In this section, traffic control strategies that primarily utilized *GT* and *RL* approaches are reviewed and discussed according to the following taxonomy:

- **Centralized Traffic Control:** The traffic control problem can be managed through a traffic control centre that monitors the traffic network and performs global optimization techniques to better utilize the existing infrastructure. Despite the fact that this approach is the ultimate goal of traffic control systems, its efficient deployment is questionable for large scale systems (e.g. cities with thousands of signalized intersections).
- **Decentralized Traffic Control:** Due to the ease of implementation and the efficient use of communication resources, there has been a trend to examine the merits of replacing the centralized philosophy of traffic control by decentralized traffic control systems. The decentralized control structure includes the following forms: no coordination, hierarchical controller coordination, inter-controller coordination, and intra-controller coordination.
    - **No Coordination:** This type of control is completely decentralized whereas local control units act independently while performing their tasks. Hence, each agent is ignoring the non-stationary effects induced by the actions performed by the other learning agents (see section 3). It is important to note that no coordination means the absence of the coordination mechanism that helps the agents choose the optimal joint action. However, the integration level of the traffic control system can be classified as level 0, 1 or 2 (see section 1).
    - **Hierarchical Controller Coordination:** Higher-level and lower-level agents constitute the hierarchical control system in which the higher-level agents are able to monitor and act according to the lower-level agents' actions. Although this seems to reduce communication and to increase reliability by allowing data to be processed quickly relative to centralized control, it requires more sophisticated software and data structures than the case of no coordination.
    - **Inter-Controller Coordination:** The agent is "aware" of the effect of other agents' actions in the network while all agents act simultaneously. In this case, the coordination mechanism enables the agents to coherently choose their actions from the optimal joint policy
    - **Intra-Controller Coordination:** The control mechanism of a single agent is the result of a negotiation process between multiple internal subagents where each subagent has its own control objective. Although the coordination between the internal subagents' actions is achieved, there is no integration between different control agents (level 0-integration).

4.1 Centralized Traffic Control Systems

Jacob and Abdulhai (2006) applied Q-Learning algorithms to integrated corridor control. The Q-Learning model developed in this study is trained and tested in a microscopic simulation model of a key corridor in Toronto using the Paramics micro-simulation suite. In order to apply the integrated control, the corridor is divided into two segments; each part includes one *VMS* and one on-ramp controlled by a centralized Q-Learning agent in which the optimal joint policy is derived. Because of the huge state space, a CMAC based function approximation with hash coding is used to store the Q-values. The obtained results are encouraging and suggest that this approach is promising in dealing with integrated freeway control applications. Although each freeway segment is centrally controlled, there is a lack of coordination between freeway segments since each segment selects its action independently.

4.2 Decentralized Traffic Control Systems

- **No Coordination**

Mikami and Kakazu (1994) combined the *RL* of a local agent with global optimization by Genetic Algorithms in which they conducted the periodic modification of parameters of *RL* by genetic search. An agent is associated to each signal and learns using a simple *RL* algorithm. The targeted parameter for the intersection controller is the cycle time. The genetic algorithm decision variables are represented by the cycle time for all the intersections in the area. The cooperation of agents is achieved by maximizing a global objective function that is the sum of the performances of all the agents after a certain running interval. The combination of *RL* and global optimization with traffic signal problems yielded good results especially in crowded traffic environments. However, there are a number of system parameters that should be carefully designed. Moreover, this approach results in a sub-optimal solution since *OOC* cycle times' values are obtained for each running interval.

Thorpe (1997) applied the SARSA *RL* algorithm to a simulated traffic light control problem using eligibility traces and greedy action selection methods. The objective of the *RL* agent was to minimize the time required to release a fixed volume of traffic through a road network. The performance of the SARSA was analysed according to different traffic state representations: vehicle counts, relative distance of vehicles from the intersections, and vehicle counts with signal light duration. In addition, four performance measures were calculated: the total number of simulation steps required for all vehicles to reach their destinations, the average vehicle travel time, the total number of stops made by all vehicles, and the average vehicle wait time. The tests conducted with the above four different state representations show that the choice of state representation is critical to the success of the SARSA strategy. The SARSA with relative distance representations are found to be the most effective in reducing average vehicle waiting time.

Wiering (2000) utilized model-based *RL* to control traffic-light agents to minimize the overall waiting time of vehicles in a small grid network. Although agents are the traffic signals, the learning process is formulated such that the state representation is vehicle-based (i.e., based on waiting times for individual vehicles) and aggregated over all vehicles in the intersection. This study introduced an interesting co-learning approach in which the value functions are learned not only by traffic signals but also by the vehicles, which can compute policies to select optimal routes to their destinations. Therefore, the control perspective is a global one, although actions are local to the agents. As for the experiments and results, the author investigated the use of multiple representations of *RL* under various saturation conditions in a grid-like network. Three different *RL* systems were designed: TC-1, in which no communication between traffic lights is required, TC-2, in which communication is required to make the transition function of the first car dependent on the number of cars standing at the next light, and TC-3, which uses this information to compute transition probabilities for all cars. It is shown that *RL* systems outperform the non-adaptive systems and that the TC-3 scenario is the best. This study provides a significant contribution since it is the first control method that applies optimal signal control over multiple signals as opposed to the widely studied single signal approach. However, deployment of this approach is questionable firstly because of high investment in communication resources and secondly because the cooperation between the agents is achieved by exchanging only their state information without coordination between the agents' actions; that is, actions are taken independently.

Abdulhai et al. (2003) applied a simple Q-Learning technique to an isolated traffic signal in a two-phase-signal two-dimensional road network. According to the state information that includes the queue lengths on the four approaches, the agent chooses either to remain in the current signal or to change it with the goal of minimizing the average number of waiting vehicles in all approaches. Three different traffic profiles (uniform traffic flows, constant-ratio traffic flows, and variable traffic flow) are tested to evaluate the performance of the Q-Learning agent under varying conditions. Q-Learning for the isolated traffic-light controller showed that it outperformed the pre-timed control scheme for variable traffic flows. Q-Learning either slightly outperformed or was equal to the pre-timed control when traffic flows were uniform or constant.

Camponogara and Kraus Jr (2003) formulated the traffic signal control problem as a cooperative stochastic game in which agents employ a distributed Q-Learning algorithm. A small network of two intersections with limited capacity roads is modelled. To account for traffic dynamics, traffic conditions were assumed to vary by applying the following policies: uniformly random policy (assigns the same probability to all actions available to an agent), best-effort policy (green indication to the lane with the longest queue), and Q-Learning implemented by agents. The results of Q-Learning show a significant reduction in the average waiting time compared to the other two policies. However, the distributed Q-Learning algorithm assumes that the optimal joint actions are unique, which is rarely the case. Consequently, the performance of the algorithm is found to be questionable in case coordination between agents is required.

Bazzan (2005) utilized techniques of evolutionary game theory to control individually-motivated agents (traffic signals) in a dynamic environment in which the global goal is considered. It is assumed that each agent can obtain its local knowledge by sensing its local environment. This approach enables agents to respond to their local environment states. However, agents receive a global reward according to their joint action. During the learning process, a fitness for each joint action is computed that influences the next generation of actions. Depending on the frequency of the stochastic events, agents are able to coordinate better towards the global goal. Several scenarios were simulated and tested by varying the learning and mutation rates. The results showed that a central coordination performs better in stable scenarios where few or no conflicts occur. However, in scenarios where the directional split is nearly equal, the central progression mechanism does not perform well compared to the agent-based mechanism. This study provides fruitful results; however, the computational time is the bottleneck. Also, there is no joint control policy to follow; instead, an open-loop optimal joint action is computed each time interval.

Jacob and Abdulhai (2006) applied Q-Learning algorithms to single and multiple ramp metering, *VMS*, and integrated corridor control. The Q-Learning model developed in this study is trained and tested in a microscopic simulation model of key corridors in Toronto using the Paramics micro-simulation suite. In case of multiple ramp control, the control of each ramp is carried out independently using the Q-Learning agent. The results showed that Q-Learning performed better than the ALINEA algorithm (Papageorgiou et al. ,1991) which is commonly used as a benchmark. In addition, this study endeavoured to control the flow distribution between a collector and express corridor in Toronto through the use of a *VMS* which implements the Q-Learning algorithm. The Q-Learning agent performed better than all the fixed diversion scenarios.

Oliveira et al. (2006) proposed an *RL* method called Reinforcement Learning with Context Detection (*RL*-CD) to control traffic lights at isolated junctions. The algorithm can handle stochastic traffic patterns which can occur due to the traffic dynamics. Experiments were designed using the ITSUMO tool that implements the Nagel-Schreckenberg micro-simulation model. The authors believe that the use of multiple partial models of the environment is an appealing approach for dealing with the non-stationary nature of the flow patterns. For each model, an optimal policy is assigned in which mapping from traffic conditions to signal plans is achieved by a model-based *RL* method such as Prioritized Sweeping and Dyna. Then, the creation of new models is controlled by a continuous evaluation of the prediction errors generated by each partial model. The empirical results showed that *RL*-CD is more efficient than the classical Q-Learning techniques and that the real traffic states must be discretized in a finer-grained representation. Although this mechanism was tested in a network of nine traffic signals, it remains a single-agent based learning method and an extension is necessary in order for the agents to map the states and joint actions to rewards.

Richter et al. (2007) exploited the *Natural Actor Critic* (*NAC*) algorithm in which four algorithms are used: policy gradient, natural gradient, temporal difference, and least-square temporal difference. In their simplified simulation, five scenarios were tested and each junction (intersection) on the grid constituted four phases. Agents used local rewards but with global observations. *NAC* managed to outperform the SCATS technique (the traffic control system of Sydney Traffic Authority) in a 10 × 10 junction grid simulation while optimizing for average vehicle travel time. Although every junction accounts for global observations, coordination between agents' actions is still missing.

Salkham et al. (2008) utilized *Collaborative Reinforcement Learning* (C*RL*) to provide adaptive and efficient urban traffic control. In their study, each signalized junction runs a *CRL*-based traffic agent that follows an adaptive phase cycle, namely *Adaptive Round Robin* (*ARR*), and observes the local traffic patterns from local vehicle location data. Q-Learning is used and a common advertisement strategy is utilized to allow for a given *ARR-CRL* agent to exchange rewards with its neighbours. The authors applied *CRL*-based *ARR* (*ARR-CRL*) and *RL*-based *ARR* (*ARR-RL*) controllers to a large-scale urban traffic control optimization scheme (in Dublin's inner city centre) which comprises 64 signalized junctions. Despite the uniform nature of the pattern tested, the average reductions in waiting time per vehicle achieved by both *ARR-RL* and *ARR-CRL* scenarios are quite significant. Although cooperation is achieved in the *CRL* algorithm by exchanging the rewards, cooperation between the agents' actions is still largely missing.

Wen et al. (2009) applied a Distributed *RL* approach to control *RM* and dynamic route guidance. In the Distributed *RL* (*eligibility traces Q-Learning*) approach, each on-ramp or *VMS* is defined as an agent and all agents have the same target: minimizing the total time experienced by all drivers in the network. The simulation results show significant improvements over the traditional local control (ALINEA) approach, and this is more vivid in the case of congested traffic networks. However, the distributed *RL* algorithm, as noted before, ignores the need for coordination by assuming that the optimal joint actions are unique, which is not often the case.

It is worth noting that all the above approaches have a common feature in that each agent selects its action independently and disregards the fact that the environment is not stationary as it depends on the policy implemented by other agents. Therefore, most of the presented approaches are limited to exploring only the isolated intersections (e.g. Thorpe, 1997; Abdulhai et al., 2003).

- **Hierarchical Controller Coordination**

Zhenlong (2005) formulated *variable speed control* (*VSC*) and *RM* as an extensive form game, namely, a Stackelberg game with a leader and multi-follower structure in which *VSC* is the leader and on-ramps are the followers. The study presented a Simulated Annealing algorithm to solve the Stackelberg model. This approach is illustrated by a simple network in which the cases of *RM* only and the coordination of *VSC* and *RM* based on Stackelberg game are compared. Simulation results are found to be satisfactory, and confirm the effectiveness of the proposed approach. However, this approach explores the optimal control in an open-loop structure.

- **Inter-Controller Coordination**

Cheng et al. (2004) formulated the traffic signal control problem as a repeated cooperative game in which all players have the same reward. A *Sample Fictitious Play* (*SFP*) algorithm, a modified form of the *FP* algorithm, is applied to explore the local optimal coordinated signal timing plans. A case study in the city of Troy, Michigan, was investigated in which they experienced delay and throughput savings for a network model of 75 signalized intersections using INTEGRATION-UM as a deterministic meso-scopic traffic simulator (Aerde et al.,1989).The significant merit of their algorithm is that it is robustly scalable for networks of realistic sizes. However, a signal timing plan has to be obtained in each planning horizon which represents an *OOC* solution.

Ghods and Rahimi-Kian (2008) applied a combined game theoretical model with a model predictive control to explore the optimal coordination of ramp metering and variable speed limits. A scalable algorithm that is capable of handling a real large-scale network with high accuracy was the ultimate goal in this study. The study formulated the control problem as a repeated cooperative game between players; each player has a similar objective function, which is the minimization of the total travel time in the network. The players in this case are *VSC* and *RM*. The game-theoretic paradigm of *SFP* is utilized to find the Nash equilibrium. The model predictive control is combined with the *SFP* algorithm to constitute a closed-loop structure. However, the approach presented results in an approximate *COC* solution in which the optimal signal plans are obtained as an *OOC* solution in every rolling horizon.

Kuyer et al. (2008) extended the *RL* approach proposed by Wiering (2000) to include an explicit coordination between neighbouring traffic lights using the coordination graphs. An efficient method for finding optimal joint actions is approached using the max-plus algorithm, which estimates the optimal joint action by sending locally optimized messages among connected agents. This approach implicitly casts the signal traffic control problem as a stochastic cooperative game. Three scenarios are investigated: the base case scenario, in which each route has a single intersection and each vehicle's destination is chosen from a uniform distribution; the non-uniform destinations scenario; and the long routes scenario in which destinations are uniformly distributed but each route has two intersections. The results are compared to Wiering's method where the base case scenario implies no significant difference between the two methods' performances; nevertheless, in the other two scenarios, the max-plus method substantially outperforms Wiering's method. However, this approach is fairly complex and assumes that all agents are connected, which makes the field deployment questionable.

- **Intra-Controller Coordination**

Alvares et al. (2008) modelled each intersection as a competitive game between its phases where each player endeavours to minimize its queue. At each iteration, a game is solved to find an $\in$-Nash equilibrium point. The approach is applied to simple isolated intersections in which an improvement over the adaptive control is not guaranteed; nevertheless, in the worst case their control technique should perform as well as the adaptive control. What limits the applicability of this approach is that it considers only isolated intersections and does not account for the non-stationary nature of the traffic network.

5. SUMMARY OF RESEARCH CONTRIBUTIONS AND GAPS IN THE LITERATURE

To summarize the studies reviewed in the previous section, Table 1 is presented to compare and contrast these studies. The rows represent the method or approach introduced to solve the problem, while the columns represent the comparison criteria. The criteria include the game type (if any), solution algorithm (*RL*, *MARL*, or other game theoretic approach), size of the problem (represented by the number of traffic elements involved), control integration level (described below), and solution optimality.

From the above survey of the traffic control applications that utilize *GT* concepts and/or *RL* approaches, it is important to highlight the following remarks.

- As illustrated in Table 1, some of the previous works were unable to achieve any level of integration (Level 0). These studies were primarily conducted to control an isolated traffic element; for example Thorpe (1997), Oliveira et al. (2006), Abdulhai et al. (2003), and Alvares et al. (2008).
- Another stream of studies investigated the integration of the traffic control problem at Level 1 in which no action coordination between the traffic control elements is reported; for example Salkham et al. (2008), Wiering (2000), Mikami and Kakazu (1994), and Richter et al. (2007).
- A third stream of studies accounted for integrated traffic control by considering a global reward and measurements; for example Wen et al. (2009) and Camponogara and Kraus Jr (2003); however, the coordination problem is avoided by implicitly assuming the uniqueness of the optimal joint policy. Therefore, Level 2 is assigned to such approaches.
- Few studies solved the coordination problem using *GT* concepts, such as Zhenlong (2005) and Bazzan (2005), *RL* approaches, such as Kuyer et al. (2008), or both *GT* solution concepts and *RL*, such as Cheng

Table 1: Traffic Control Approaches Using GT and RL: Comparative Literature Survey

Criterion Study			Game Type	Solution Method			Elements	Integration Level	Solution Optimality
				Single RL	MARL	Other			
Centralized Control		Jacob and Abdulhai (2006)	--	Q-Learning	--	--	1 RM 1 VMS	4	COC
Decentralized Control	**Hierarchical Controller Coordination**	Zhenlong (2005)	Repeated Static Extensive Game	--	--	Simulated Annealing	2 RM 1 VMS	3	OOC
	Inter-Controller Coordination	Cheng et al. (2004)	Repeated Static Cooperative Game	--	Modified FP	--	75 SI	3	OOC
		Ghods and Rahimi-Kian (2008)	Repeated Static Cooperative Game	--	Modified FP	--	2 RM 1 VMS	3	OOC
		Kuyer et al. (2008)	Dynamic Cooperative Game	--	Coordination Graphs	--	3 SI	3	COC
	Intra-Controller Coordination	Alvares et al. (2008)	Finite Dynamic Competitive Game	--	--	ϵ - Equilibrium Point Algorithm	1 SI	0	OOC
	No Coordination	Jacob and Abdulhai (2006)	--	Q-Learning	- -	--	1 VMS Or 1 RM	0	COC
		Jacob and Abdulhai (2006)	--	Q-Learning	--	--	2 RM 2 VMS	0	COC
		Abdulhai et al. (2001)	--	Q-Learning	--	--	1 SI	0	COC
		Oliveira et al. (2006)		Prioritized Sweeping and Dyna	--	--	9 SI	0	COC
		Thorpe (1997)	--	SARSA	--	--	1 SI	0	COC

Table 1: Traffic Control Approaches Using GT and RL: Comparative Literature Survey

		Mikami and Kakazu (1994)	--	Similar to Q-Learning	--	--	36 SI	1	COC
		Wiering (2000)	--	Model based Q-Learning	--	--	6 SI	1	COC
		Salkham et al. (2008)	--	Modified Q-Learning	--	--	64 SI	1	COC
		Richter et al. (2007)	--	Modified Actor Critic	--	--	100 SI	1	COC
		Campono gara and Kraus Jr (2003)	Dynamic Cooperative Game	--	Distributed Q-Learning	--	2 SI	2	COC
		Bazzan (2005)	--	--	--	Evolutionary Game Theory Technique	10 SI	2	OOC
		Wen et al. (2009)	--	--	Distributed Eligibility Traces Q-Learning	--	2 RM 1 VMS	2	COC

et al. (2004) and Ghods and Rahimi-Kian (2008). However, these approaches are either impractical or sub-optimal ones in which an *OOC* solution is obtained at each time interval. Therefore, Level 3 is assigned to such approaches.

- The centralized control approaches represent the highest level of integration in which all the traffic elements are controlled using a single agent (Jacob and Abdulhai, 2006). Nevertheless, the computational complexity and the communication requirements of the approach increases exponentially with the number of traffic control elements.
- No study, to the best of the authors' knowledge, has tackled the *ITC* (*RM*, *VMS*, and *SI*) problem to find a *COC* solution using a coordination mechanism that minimizes the communication requirements.

It is clearly shown that significant accomplishments in *GT* and *MARL* have been achieved by several researchers as presented above. However, some gaps still exist. Therefore, our ultimate goal is to propose, test, and implement a prototype for *MARLIN-OTC*.

6. THE PROPOSED MULTI-AGENT REINFORCEMENT LEARNING FOR INTEGRATED NETWORK OF OPTIMAL TRAFFIC CONTROLLERS (*MARLIN-OTC*)

The objective of the proposed framework is to achieve a decentralized optimal closed-loop control law (*COC* solution) for the *ITC* problem using *MARL* approaches and game theory concepts.

Due to the huge number of agents associated with real applications, we partition the agents into coalitions, which results in the coalition game problem. Within each coalition, decentralized agents can play different types of games according to the nature of the control task as will be shortly explained.

As shown in Figure 3, the proposed system is composed of three phases:

- *Coalition Structure Generation*, in which the set of agents partition themselves into exhaustive and disjointed coalitions and form a new coalition structure (*CS*).
- *Coalition Implementation*, in which the agents in each coalition play a game to explore the optimal joint policy.

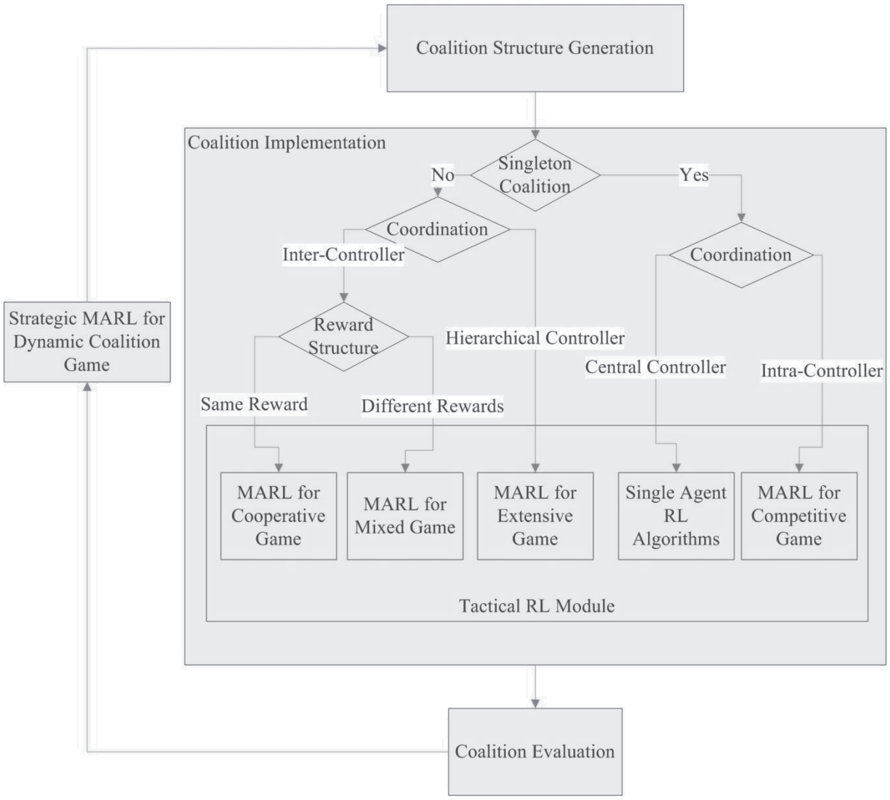

Figure 3. The Proposed MARLIN-OTC Framework.

- *Coalition Evaluation*, in which the coalition's value is calculated.

The three former phases are performed every time interval while each agent is interacting with a *MARL* approach to learn the best coalition structure. The objective of any agent is to join a coalition that guarantees a higher reward than can be obtained by acting alone. Accordingly, each agent will decide from the following: changing its coalition, staying in its coalition, or acting alone and forming a singleton coalition. In Section 4, various *MARL* approaches that solve the dynamic coalition form game are reviewed.

In each coalition, the members play a game every time step in order to find the optimal joint policy. Therefore, the *RL* approach is implemented in each coalition according to the coalition type:

- *Singleton Coalition:* when an agent chooses to act alone, this results in a coalition of single agent or singleton coalition in which the agent can use a SARL to choose its action.
- *Multiple Agents Coalition:* when multiple agents are forming a coalition. According to the coordination mechanism, an appropriate *MARL* approach could be used in that coalition. If the agents are using hierarchical controller coordination, in which actions are taken sequentially, the agents implement one of the *MARL* algorithms for extensive form games. Alternatively, an inter-controller coordination can be obtained in which all agents are acting independently and simultaneously. In this case the solution approach varies according to the reward structure; if the agents have the same reward, a *MARL* for cooperative games is implemented; if the rewards are different, a *MARL* for mixed games is implemented.

The agent structure is represented in Figure 4. In this model, the agent has to interact with two learning modules at each time step: *strategic reinforcement learning* (*SRL*) and *tactical reinforcement learning* (*TRL*). In *SRL* agents learn about the utilities of *CS*, while in *TRL* agents learn about other agents' behaviours. This is because *SRL* provides a strategic learning approach to facilitate the planning of *CS* in the coalition form game, while *TRL* provides a tactical learning approach to learn how to implement the planned *CS* and occurs during the coalition implementation phase. As shown in Figure 3, the *SRL* module can be represented by a *MARL* approach for the dynamic coalition games, while *TRL* can be replaced by either the SARL or the *MARL* approach accord-

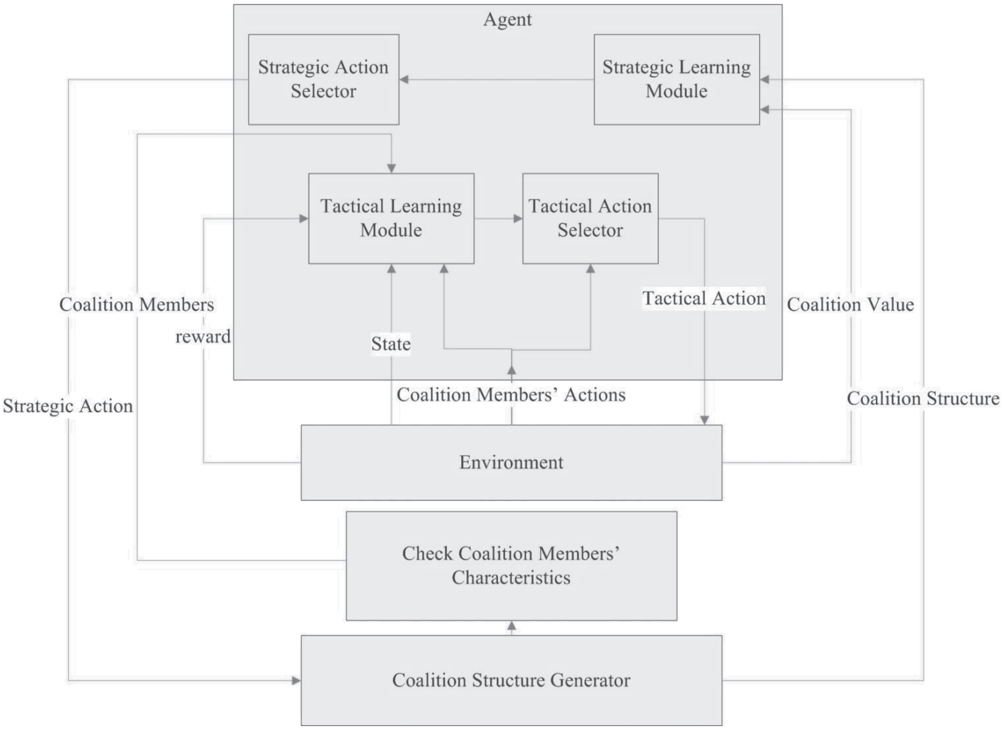

Figure 4. Agent Model.

ing to the coalition characteristics; that is, the coordination strategy between the coalition members and the reward structure. Although there are mathematical proofs of convergence to stable equilibria for *TRL* and *SRL* approaches; there is no mathematical proof of convergence if we loop between the two levels as proposed in *MARLIN-OTC*.

7. MAPPING BETWEEN *MARLIN-OTC* AND TRAFFIC CONTROL APPLICATIONS: ILLUSTRATIVE EXAMPLES

This section aims to provide a vision for implementing the proposed framework. The concept of utilizing *MARLIN-OTC* for various traffic control applications is illustrated for a small network as shown in Figure 5. Both surface streets and freeways are included in the network in which the 401 Freeway in Toronto (the second busiest freeway in NA) intersects with a major collector highway (Bayview Avenue). This network includes two *VMS*s associated with a bifurcation point, a variable speed controller and two *RM* and several minor–major and major–major signalized intersections. Therefore, the following set of players composes the coalition game:

$$P = \{RM_1, RM_2, VMS_1, VMS_2, I_1, I_2, I_3, I_4, I_5\}$$

We envision a one-shot static coalition game in a certain time step as shown in Figure 6. A possible *CS* can be illustrated as follows:

- Coalition C_1: RM_1, VMS_1, I_1, and I_2
- Coalition C_2: RM_2, VMS_2
- Coalition C_3: I_3 and I_4
- Coalition C_4: I_5

The coalition implementation phase is performed according to the coordination strategy and the reward structure of each coalition member.

In coalition C_1, it is assumed that the following hierarchy will be followed while managing the freeway and surface streets systems: freeway controllers (e.g, heavily trafficked provincial freeway) are given priority to take an action first, followed by the surface street controllers (e.g, less critical municipal roads). Also, in the freeway control it is assumed that the VMS_1 would take the action first by controlling the speed, followed by the RM_1 action which depends -in part- on the VMS_1 action. Similarly, we can assume that in the surface street control process, the major intersection I_1 should take the action first, followed by the minor intersec-

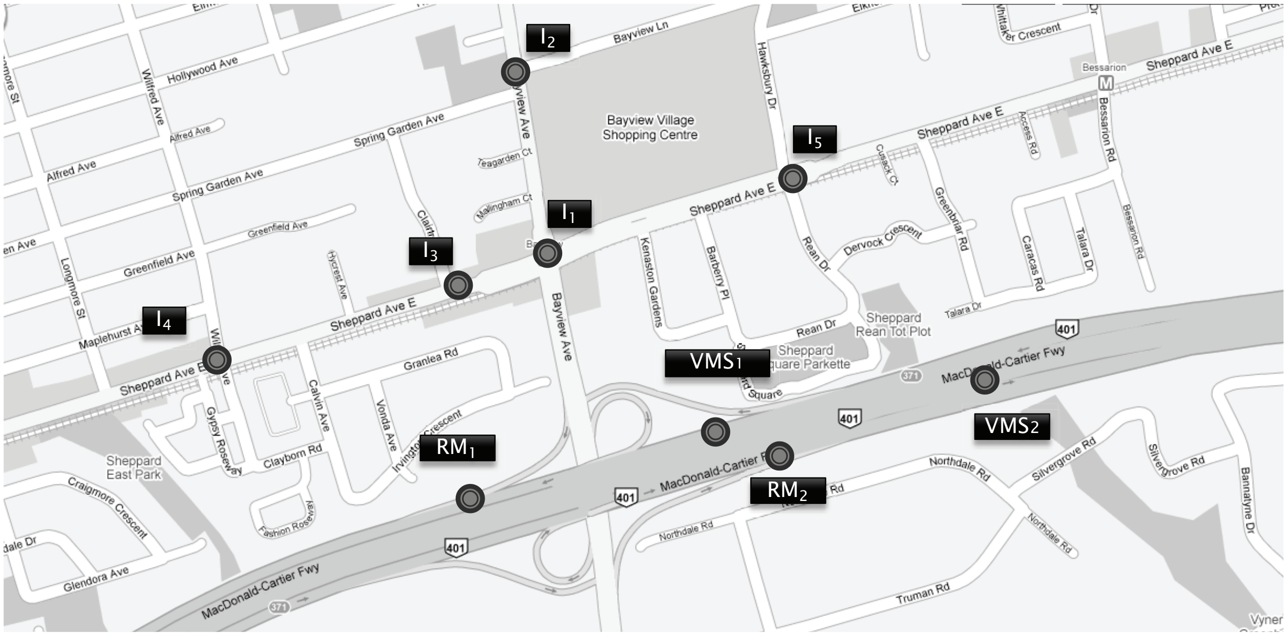

Figure 5. Illustrative Application of MARLIN-OTC.

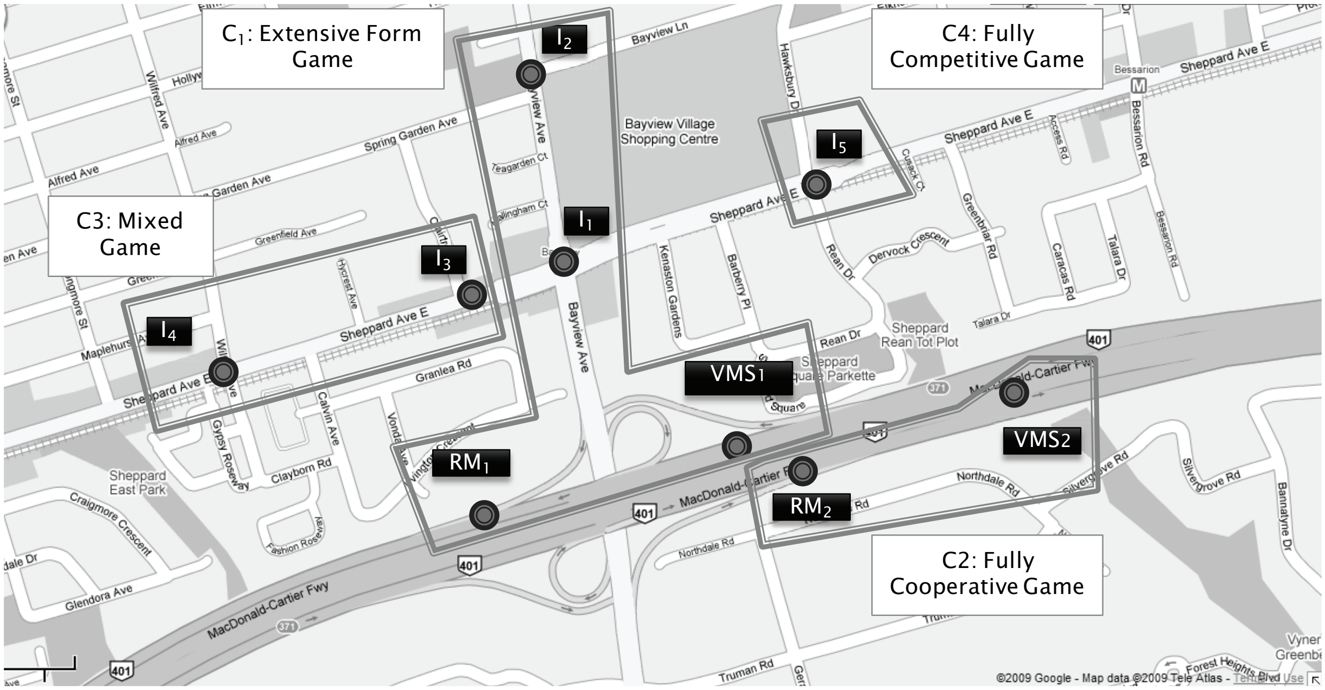
C1: Extensive Form Game
C4: Fully Competitive Game
C3: Mixed Game
C2: Fully Cooperative Game
I2
I5
I1
I3
I4
VMS1
VMS2
RM1
RM2
Bayview Village Shopping Centre
Sheppard Ave E
Bayview Ave
MacDonald-Cartier Fwy
401
371
©2009 Google - Map data ©2009 Tele Atlas - Terms of Use

Figure 6. Example of Coalition Formation Game.

tions I_2. Hence, a hierarchical-controller coordination can be achieved by solving an extensive game between the players in this coalition.

In coalition C_2, it is assumed that RM_2 and the VMS_2 have the same objective of maximizing the exit flow. Thus, the *RM* and *VMS* agents will take part in a cooperative game to implement inter-controller coordination.

In coalition C_3, it is assumed that the two signalized intersections (I_3 and I_4) each have an independent and different reward (a penalty in this case) equal to the squared sum of the queues of the four approaches.

In coalition C_4, the signalized intersection can be controlled as a single agent controller where the objective is to minimize the sum of squared queues of the four approaches. On the other hand, the signalized intersection control can be perceived as a competitive game between the signal phases where agents have competitive rewards. Assume that the signalized intersection, I_5, has two phases and each phase has a reward function (a penalty in this case) equal to the difference between the queue length in both phases.

The above is but one example of how *MARLIN-OTC* can be applied in a real network. This example can be extended to a variety of similar situations in different size networks. The size of the coalition can vary with the size and nature of the road network being controlled and of course with the available computational power available.

8. SUMMARY, CONCLUSIONS, AND FUTURE WORK

The Traffic congestion can be alleviated by infrastructure expansion; however, this is indeed unfeasible due to the obvious financial resources and physical space constraints. The more practical approach is better utilizing the existing infrastructure by optimally integrating traffic control strategies, this includes: *RM*, *VMS*, and *SI*. *RL* is a promising technique to tackle the optimal control problem. Integrated traffic control strategies are broadly categorized to five levels depending on the degree of integration. The higher the level of the integration, the more coordination between the agents is required. Hence, the highest integration level can be achieved by a centralized controller. However, due to the expensive processing and communication resources associated with the central authority, decentralization using *MAS* is more practical. The decentralized traffic control problem is an excellent testbed for *MARL* that combines *RL* and *GT* concepts. While the GT concepts provide the tools to model the *MAS* as a multiplayer stochastic game, *RL* provides the optimal control policies for the players involved in the game. Therefore, a significant progress in the optimal integrated traffic control is approachable by exploring the synergy between the fields of control theory, game theory, and machine learning, which motivated this research.

In section 2, the theoretical foundations of Automatic Control are briefly presented. The distinction between the *OOC* and the *COC* is emphasized. It is obvious that the *COC* is the ultimate goal; however, traditional methods can only solve simple problems analytically. Dynamic Programming, as an efficient approach to achieve *COC* for the stochastic process, is briefly discussed. Due to the limitations of the *DP* that are overcame by *RL*, a comprehensive review and categorization of the single agent-*RL* algorithms is presented.

Section 3 conceptually visits the problem from a multi-agent perspective by highlighting the basic concepts of *GT* and *MARL*. First, three main mathematical game forms are defined: the extensive form, the strategic form, and the coalitional form. Then, the solution concept for each form is presented. The stochastic game, as a generalized stochastic process with multiple players, is defined next. The stochastic game can take any form from the previously mentioned forms. Second, *MARL* algorithms, as efficient approaches to solve the stochastic games while coordinating the agents' actions, are reviewed. The *MARL* algorithms are reviewed according to the following taxonomy: game form, dynamics of the game, and the coordination level. Finally, the function approximation methods that are aiming to reduce the dimensionality issue of the *MARL* approaches are briefly outlined.

Section 4 and 5 comprehensively compile and critique the traffic control strategies that primarily utilized *GT* and *RL* approaches in the literature. The control approaches are categorized into: centralized and decentralized. Decentralized control methods can take one of the following coordination forms: no coordination, hierarchical controller coordination, inter-controller coordination, and intra-controller coordination. Also, the reviewed studies are compared according to the following criteria; the game form (if any), solution algorithm (*RL*, *MARL*, or other game theoretic approach), size of the problem (represented by the number of traffic elements involved), control integration level, and solution optimality.

In an attempt to extend the literature mosaic and to bridge the gaps in the traffic control studies, a new and comprehensive roadmap for *MARLIN-OTC* is presented in section 6. The core concept of *MARLIN-OTC* is to breakdown the network controllers into groups (coalitions) by solving a stochastic coalition form game. Within each coalition, decentralized agents are playing a game at the lower level. The game form in each coalition depends on the objectives of the players and the required coordination level. In the proposed model, each agent interacts with two learning modules: *SRL* and *TRL*. *SRL* is responsible for learning the best set of coalitions while the *TRL* is responsible for learning the best policy for each agent within its coalition. For illustration purposes, the *MARLIN-OTC* is envisioned for a potential application in Toronto's road network.

REFERENCES

Abdulhai, B. and Kattan L. (2003) Reinforcement learning: Introduction to theory and potential for transport applications. *Canadian Journal of Civil Engineering*, 30 (6), 981–991.

Abdulhai, B., Pringle, R., and Karakoulas, G.J. (2003) Reinforcement learning for true adaptive traffic signal control. *ASCE Journal of Transportation Engineering*, 129 (3), 278–285.

Abul, O., Polat, F. and Alhajj, R. (2000) Multiagent reinforcement learning using function approximation, *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Application and Reviews*, 30 (4), 485–497.

Aerde, M. V., Voss, J. and McKinnon, G. (1989) *INTEGRATION Simulation Model User's Guide*, Queen's University.

Akramizadeh, A., Afshar, A., and Menhaj, M.B. (2009) Multiagent reinforcement learning in extensive form games with perfect information. *Journal of Applied Sciences*, ISSN 1812-5654.

Albus, J.S. (1975) A new approach to manipulator control: the cerebellar model articulation controller (CMAC), *Trans. ASME, Series G. Journal of Dynamic Systems, Measurement and Control*, 97, 220–233.

Alvares, I., Poznyak, A., and Malo, A. (2008) Urban traffic control problem: agame theory approach. In: *CDC, 47th IEEE Conference on Decision and Contro*l, pp. 2168–2172.

Basar, T., and Olsder, G.J. (1999) Dynamic Noncooperative Game Theory. *Classics in Applied Mathematics. 2nd edition*. London, U.K.

Bazzan, A.L.C. (2005) A distributed approach for coordination of traffic signal agents. *Autonomous Agents and Multi-Agent Systems*, 10 (2), 131–164.

Bazzan, A.L.C. (2009) Opportunities for multiagent systems and multiagent reinforcement learning in traffic control. *Autonomous Agents and Multi-Agent Systems*, 18 (3), 342–375.

Bingham, E. (2001) Reinforcement learning in neurofuzzy traffic signal control. *European Journal of Operation Research*, 131, 232–241.

Boutilier, C. (1996) Planning, learning and coordination in multiagent decision processes. In: *Proceedings of the 6th Conference on Theoretical Aspects of Rationality and Knowledge*, De Zeeuwse Stromen, The Netherlands, pp. 195–210.

Brown, G.W. (1951) Iterative solutions of games by fictitious play. In: Koopmans, T.C. (ed.) *Activity Analysis of Production and Allocation*. Wiley, New York, pp. 374–376.

Busoniu, L., Babuska, R., and De Schutter, B. (2008) A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man and Cybernetics*, 38 (2), 156–172.

Bertsekas, DP, and Shreve, SE (1978) tochastic optimal control: the discrete time case, *Mathematics in Science and Engineering*, Academic Press, Inc., New York.

Camponogara, E. and Kraus Jr, W. (2003) Distributed learning agents in urban traffic control. In: *Progress in Artificial Intelligence, Proceedings of the 11th Portuguese Conference on Artificial Intelligence*, pp. 324–335.

Carmel, D. and Markovitch, S. (1996) Opponent modeling in multi-agent systems. In: Weiss, G. and Sen, S. (eds.), *Adaptation and Learning in Multi-Agent Systems*, Springer-Verlag, New York, Ch. 3, pp. 40–52.

Chalkiadakis, G. and Boutilier, C. (2004) Bayesian reinforcement learning for coalition formation under uncertainty. In: *3rd International Joint Conference on Autonomous Agents and Multiagent Systems*, vol. 3, pp. 1090–1097.

Cheng, S.F., Epelman, M.A., and Smith, R.L. (2004) CoSIGN: A fictitious play algorithm for coordinated traffic signal control. *INFORMS Annual Meeting*, Denver, CO.

Claus, C. and Boutilier, C. (1998) The dynamics of reinforcement learning in cooperative multiagent systems. In: *Proceedings of the 15th National Conference on Artificial Intelligence and 10th Conference on Innovative Applications of Artificial Intelligence*, Madison, US, pp. 746–752.

Filar J. and Vrieze , K. (1997) *Competitive Markov decision processes*. Springer-Verlag, New York.

Fischer, F., Rovatsos, M., and Weiss, G. (2004) Hierarchical reinforcement learning in communication-mediated multiagent coordination. In: *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems*, New York, US, pp. 1334–1335.

Fujita, K. and Matsuo, H. (2005) Multi-agent reinforcement learning with the partly high-dimensional state space, *Trans. of IEICE*, J88-D-1 (4), 864–872.

Ghods, A.H. and Rahimi-Kian, A. (2008) A game theory approach to optimal coordinated ramp metering and variable speed limits. *IEEE Chinese Control and Decision Conference*, Yantai, China.

Gosavi, A. (2003) *Simulation based optimization: parametric optimization techniques and reinforcement learning.* Kluwer Academic, Norwell, MA.

Hu, J. and Wellman, M.P. (2003) Nash Q-Learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4, 1039–1069.

Huang, P. and Sycara, K. (2003) Multi-agent learning in extensive games with complete information, *Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems*, ACM Press, NY, USA, pp. 701–708.

Hunt, P. B. , Robertson, D. I., Bretherton, R. D. and Winton, R. I. (1981) SCOOT-a traffic responsive method of coordinating signals, *Technical Report, Transport and Road Research Laboratory*, Crowthorne, England.

Jacob, C., and Abdulhai, B. (2006) Automated adaptive traffic corridor control using reinforcement learning: approach and case studies, *Transportation Research Board Annual Meeting*, Washington D.C.

Kahan, J.P. and Rapoport, A. (1984) Theories of coalition formation. Lawrence Erlbaum Associates.

Kok, J.R., Spaan, M.T.J., and Vlassis, N. (2005) Non-communicative multirobot coordination in dynamic environment. *Robotics and Autonomous Systems*, 50 (2/3), 99–114.

Koller, D. and Pfeffer, A. (1997) Representations and solutions for game theoretic problems. *Artificial Intelligence*, 94 (1), 167–215.

Könönen, V. (2003) Asymmetric multiagent reinforcement learning. In: *Proceedings IEEE/WIC International Conference on Intelligent Agent Technology*, Halifax, Canada, pp. 336–342.

Kuyer, L., Whiteson, S., Bakker, B., and Vlassis, N. (2008) Multiagent reinforcement learning for urban traffic control using coordination graphs. In: *Proceedings of the 19th European Conference on Machine Learning, September 2008*, pp. 656–671.

Laslier, J.F. and Walliser, B. (2005) A reinforcement learning process in extensive form games, *International Journal of Game Theory*, 33 (2), 219–227.

Lazaric, A., Munoz de Cote, E., Gatti, N., and Restelli M. (2007) Reinforcement learning in extensive form games with incomplete information: the bargaining case study. In: *Proceedings of the International Joint Conference on Autonomous Agents and Multi Agent Systems,* Honolulu, USA, pp. 216–218.

Li, X. and Soh, L.-K. (2004) Investigating reinforcement learning in multiagent coalition formation. *Technical report no. WS-04-06, American Association of Artificial Intelligence Workshop on Forming and Maintaining Coalitions and Teams in Adaptive Multiagent Systems*, pp. 22–28.

Lin, L.(1992)Self-improving reactive agents based on reinforcement learning, planning and teaching, *Machine Learning*, vol 8, no 3

Littman, M. (1994) Markov games as a framework for multi-agent reinforcement learning. In: *Proceedings of the 11th International Conference on Machine Learning*.

M. L. Littman (2001) Value-function reinforcement learning in Markov games. *The Journal of Cognitive Systems Research*., 2(1), 55–66.

Liu, Z. (2007) A survey of intelligence methods in urban traffic signal control. *International Journal of Computer Science and Network Security*, 7 (7), 105–112.

Mendelson, E. (2004) *Introducing Game Theory and Its Applications*. CRC Press

Mikami, S., and Kakazu, Y. (1994) Genetic reinforcement learning for cooperative traffic signal control. In: *International Conference on Evolutionary Computation*, pp. 223–228.

Milano, P. D. (2006) *Dynamic coalition formation via reinforcement learning: easing optimization problems*. Master's Thesis.

Nash, S. and Sofer, A. (1996) *Linear and nonlinear programming*. McGraw-Hill, NY.

Oliveira, D., Bazzan, A., Silva, B., Basso, E., Nunes, L., Rossetti, R., Oliveira, E., Silva, R., and Lamb, L. (2006) Reinforcement learning-based control of traffic lights in non-stationary environments: a case study in a microscopic simulator. In: *4th European Workshop on Multi-Agent Systems*, Lisbon, pp. 31–42.

Osborne, M. J. (2004) An introduction to game theory. Oxford University Press, New York.

Papageorgiou, M., Hadj-Salem, H., and Blosseville, J. M. (1991) ALINEA: A local feedback control law for on-ramp metering. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1320, TRB, National Research Council, Washington, D.C., pp. 58-64.

Papageorgiou M. (1998) Automatic control methods in traffic and transportation. In: *Operations Research and Decision Aid Methodologies in Traffic and Transportation Management*, P. Toint, M. Labbe, K. Tanczos, G. Laporte, Editors, Springer-Verlag, pp. 46-83.

Richter, S., Aberdeen, D., and Yu, J. (2007) Natural actor-critic for road traffic optimisation. In: *Advances in Neural Information Processing Systems*, vol. 19, MIT Press, Cambridge.

Salkham, A., Cunningham, R., Garg, A., and Cahill, V. (2008) A collaborative reinforcement learning approach to urban traffic control optimization. In: *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 2, pp. 560–566.

Shoham, Y., Powers, R., and Grenager, T. (2003) *Multi-agent reinforcement learning: A critical survey*. Tech. Rep., Computer Science Dept., Stanford University, California, US.

Sims, A. G., and Dobinson, K.W. (1980) The Sydney coordinated adaptive traffic (SCAT) system philosophy and benefits, *Vehicular Technology, IEEE Transactions on*, 1980. 29(2): p. 130-137

Smith, J.M. (1982) *Evolution and the Theory of Games*. Cambridge Univ. Press, UK.

Steingröver, M., Schouten, R., Peelen, S., Nijhuis, E., and Bakker, B. (2005) Reinforcement learning of traffic light controllers adapting to traffic congestion. In: BNAIC, pp. 216–223.

Sutton, R., and Barto, A. (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 Department, Colorado State University, Fort Collins, Colorado.

Katwijk, R.T. van, P. van Koningsbruggen, B. De Schutter and J. Hellendoorn (2005) A test bed for multi-agent control systems in road traffic management, in Applications of Agent Technology in Traffic and Transportation (F. Klügl, A. Bazzan, and S. Ossowski, eds.), Whitestein *Series in Software Agent Technologies*, Basel: Birkhäuser Verlag, ISBN 3-7643-7258-3, pp. 113-131.

Wang, X., and Sandholm,T., (2002), Reinforcement learning to play an optimal Nash equilibrium in team Markov games. In: *Advanced Neural Information Processing Systems*, vol. 15, pp. 1571–1578.

Wang, Y., Diakaki, C., Kotsialos, A., and Papageorgiou, M. (2007) *Towards integrated network traffic control. In: Modeling, Information, and Control of Intelligent Transportation Systems* (in French), Hermès Science Publishing Ltd, UK, pp. 281–313.

Watkins, C. and Dayan, P. (1992) Q-learning, *Machine learning*. vol. 8, pp. 279-292.

Weinberg, M. and Rosenschein, J.S. (2004) Best-response multiagent learning in non-stationary environments. In: *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems*, New York, NY, pp. 506–513.

Wen, K., Qu, S., and Zhang, Y. (2009) A machine learning method for dynamic traffic control and guidance on freeway networks, In: *2009 International Asia Conference on Informatics in Control, Automation and Robotics*, pp. 67–71.

Wiering, M.A. (2000) Learning to control traffic lights with multiagent reinforcement learning. In: *1st World Congress of the Game Theory Society Games*, Utrecht, Netherlands, Basque Country University and Foundation B.B.V., Bilbao, Spain.

Zhenlong, L. (2005) Optimal coordination of variable speed and ramp metering based on Stackelberg game. In: *12th World Congress on Intelligent Transport Systems*, San Francisco, USA.