Network-wide Optimum Signal Control by Multi-Agent Reinforcement Learning utilizing Traffic Wave Propagation

Shin Hashimoto[a], Kazuki Fukuda[a], Jun Tanabe[a],
Keisuke Yoshioka[b], Masafumi Kobayashi[c], and Masao Kuwahara[d]

[a]*Regional Futures Research Center, Osaka, Japan*
[b]*Nihon University, Chiba, Japan*
[c]*Sumitomo Electric Co. Ltd., Osaka, Japan*
[d]*Tohoku University, Sendai, Japan*

1 Introduction

This study proposes network-wide optimum signal control using Multi-Agent Reinforcement Learning (MARL) based on the decomposition of the reward into individual intersections considering the property of traffic wave propagation. The proposed MARL-based control is theoretically proved to yield the network-wide optimum for a general network. The proposed method is validated under both undersaturated and oversaturated scenarios with queue blocking back.

In general, a network-wide optimum signal control is a complex problem and the mathematical analysis is not straightforward because the formulation of dynamic traffic flow with queues is difficult due to variety of signal phase configurations and intersection geometries. Among the huge amount of past literatures, Smith (1979, 1984, 2015) provided theoretically interesting discussions on control strategies that maximize network capacity considering route choice but did not fully consider queue blocking back.

Recently, Reinforcement Learning (RL) has been used for signal control because of its flexible applicability to various optimization problems. For a network with multiple intersections, MARL has been employed to obtain the optimum control in most recent studies (e.g., Chow et al. (2020), Li et al. (2021), Haddad et al. (2022)). However, in general, MARL-based approach cannot guarantee to establish the network-wide optimum control, since every individual agent behaves so as to optimize its own return not the return for an entire network. Although previous MARL-based approaches introduce some cooperative arrangements among intersections (agents) to capture their interactions for getting closer to the network-wide optimum, they still do not guarantee to establish the network-wide optimum.

One of the issues on the existing MARL-based approach is that traffic flow characteristics have not been sufficiently addressed. Therefore, this study formulates the signal control problem as Markov Decision Process (MDP) considering propagations of forward and backward traffic waves generated by signal control at intersections. Utilizing the property of wave propagations, we show a gentle condition that makes the MARL-based approach establish the network-wide optimum. Although user choice behavior, especially route choice, is not considered in this study, feasibility of the extension to include choice behavior is discussed.

2 Control Objective and Wave Propagations

2.1 Network and Demand

A node with a traffic light is an intersection node denoted as node *j*, *j* = 1, 2, ……, *J*, and a node generating or absorbing traffic demand is a centroid. A link is described as $(i,\, j)$, a pair of its starting node *i* and terminal nodes *j*. Sets of nodes and links are denoted as *N* and *L* respectively. Traffic demand is assumed given and the diverging ratios at intersection nodes remain fixed due to the assumption of fixed vehicle routes. Nodes are connected by a single directed link for each direction as shown in Figure 1, illustrating a segment of the entire network. For intersection *j*, its adjacent nodes and links are delineated in red, and their sets are denoted as $N_j \in N$, $L_j \in L$ and $L_j \bigcap L_k = \varnothing$ if $j \neq k$.

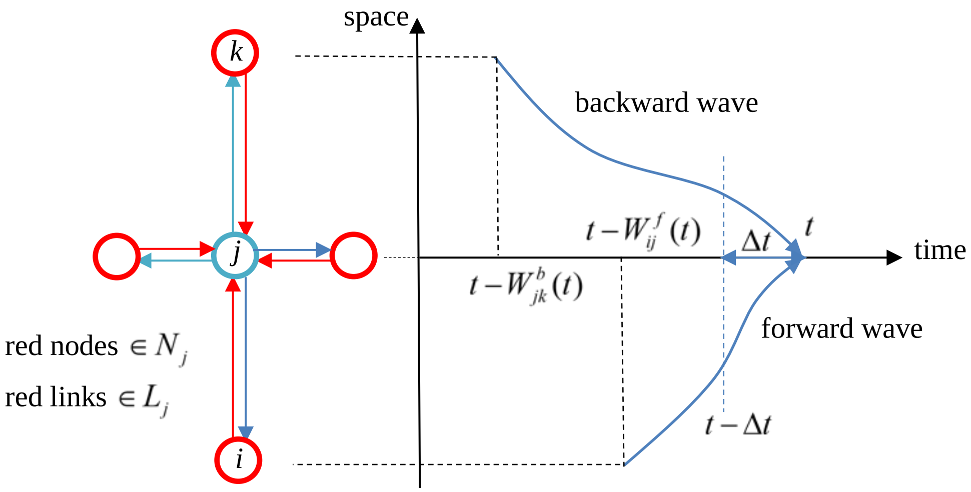


Figure 1: Network configuration and wave propagations on links

2.2 Control Objective

Let us first discuss the signal control objective and its characteristics in relation to the wave propagation of traffic flow on a network based on kinematic wave theory. First of all, the following cumulative vehicle counts are defined for link $(i,\, j)$, $\forall (i,\, j) \in L$.

$A_{ij}(t)$ = the cumulative number of vehicles entering link $(i,\, j)$ by time *t*,

$A'_{ij}(t)$ = the cumulative number of vehicles on link $(i,\, j)$ arriving at node *j* without delay by time *t*

$= A_{ij}(t - W_{ij}^{f})$,

$D_{ij}(t)$ = the cumulative number of vehicles leaving link $(i,\, j)$ by time *t*,

W_{ij}^{f} = free flow travel time on link $(i,\, j)$.

The $A'_{ij}(t)$ represents the hypothetical cumulative vehicles that could have arrived at the downstream end of link (i, j) by time *t* if delays do not occur. It is simply the horizontal translation of $A_{ij}(t)$ by a time displacement of free flow travel time: $A'_{ij}(t) = A_{ij}(t - W_{ij}^f)$. The difference between $A'_{ij}(t)$ and $D_{ij}(t)$ is the appropriate measure of the number of vehicles in a queue on link (i, j) at time *t*:

$$q_{ij}(t) = A'_{ij}(t) - D_{ij}(t) = A_{ij}(t - W_{ij}^f) - D_{ij}(t)$$
$$= \text{the number of vehicles in a queue on link } (i, j) \text{ at time } t, \ \forall (i, j) \in L. \tag{1}$$

We employ the throughput maximization as the control objective. For traffic starting from origins to destinations along various routes on a network, the flow conservation at every node should be satisfied as follows:

$$\sum_{i \in N_j} \frac{dD_{ij}(t)}{dt} + o_j(t) = \sum_{k \in N_j} \frac{dA_{jk}(t)}{dt} + d_j(t), \ \forall j \in N, \tag{2}$$

$o_j(t),\ d_j(t)$ = demand rate originated from and absorbed at node *j* at time *t.*

Because $d_j(t)$ denotes the flow rate arriving at destination *j* at time *t*, the total throughput is the sum of $d_j(t)$ for all nodes throughout the study period: $\int_0^T \left\{ \sum_{j \in N} d_j(t) \right\} dt$. From the flow conservation and Eq.(1), total throughput is expressed as follows:

$$\begin{aligned}\int_0^T \left\{ \sum_{j \in N} d_j(t) \right\} dt &= \int_0^T \sum_{j \in N} \left\{ \sum_{i \in N_j} \frac{dD_{ij}(t)}{dt} - \sum_{k \in N_j} \frac{dA_{jk}(t)}{dt} + o_j(t) \right\} dt \\ &= \int_0^T \sum_{(i,j) \in L} \left\{ \frac{dD_{ij}(t)}{dt} - \frac{dA_{ij}(t - W_{ij}^f)}{dt} \right\} dt + \int_0^T \left\{ \sum_{j \in N} o_j(t) \right\} dt \\ &= -\int_0^T \left\{ \sum_{(i,j) \in L} \frac{dq_{ij}(t)}{dt} \right\} dt + \int_0^T \left\{ \sum_{j \in N} o_j(t) \right\} dt.\end{aligned} \tag{3}$$

Since the demand from origins, $\int_0^T \left\{ \sum_{j \in N} o_j(t) \right\} dt$, is given, throughput maximization is equivalent to the following:

$$\text{Throughput Maximization} \equiv Max \int_0^T \left\{ - \sum_{(i,j) \in L} \frac{dq_{ij}(t)}{dt} \right\} dt. \tag{4}$$

2.3 Forward and Backward Wave Propagations

When an intersection is controlled by the signal, forward and backward waves are generated from the intersection node. Travel times of these waves traversing links are defined as follows:

$W_{ij}^{f}(t)$ = Travel time of forward wave generated from node *i* traveling on link (i, j) and arriving at node *j* at time *t*, $\forall(i, j) \in L$,

$W_{ij}^{b}(t)$ = Travel time of backward wave generated from node *j* traveling backward on link (i, j) and arriving at node *i* at time *t*, $\forall(i, j) \in L$.

As depicted in Figure.1, there must be always some time-lag, which is equal to the wave travel time, between time when an intersection is controlled and time when the generated wave arrives at the adjacent intersection. We will utilize this time-lag later for the MARL-based signal control.

3 Reinforcement Learning

3.1 Action

We propose designing the Reinforcement Learning (RL) process within a discretized time axis, partitioning time into uniform intervals denoted by Δt, and time step *t* is defined as $[t, t+\Delta t)$. Continuous time *t* and discrete time step *t* are used interchangeably according to the context. For each intersection *j*, the phase configuration is assumed given and action a_t^j is defined as the choice of one of the given phases at intersection *j* during time step *t*. If the number of phases at intersection *j* is p_j, action a_t^j is an integer number from 1 to p_j. A set of actions at all intersections at the time step *t* is denoted as $a_t = \left(a_t^1, a_t^2, a_t^3, \ldots\ldots., a_t^J\right)$.

3.2 State

State s_t is the collective traffic conditions covering all intersections at the onset of time step *t*, that corresponds to continuous time *t*. For each intersection *j*, state s_t^j is defined as traffic conditions on links $(i, j) \in L_j$ at time step *t*. The collective array of states for all intersections is denoted as $s_t = \left(s_t^1, s_t^2, s_t^3, \ldots\ldots., s_t^J\right)$.

3.3 Reward

Reward r_{t+1} is compensation at the end of time step *t* by taking action a_t under state s_t. If the traffic environment adheres to the Markov property, reward r_{t+1} depends only on state s_t and action a_t, and it can thus be written as $r_{t+1}(s_t, a_t)$. This reward aligns with the signal control objective as shown in Eq.(4), which is a function of queues. Given $q_{ij}(t)$, $\Delta q_{ij}(t)$ is rewritten on the continuous time as $\Delta q_{ij}(t) = q_{ij}(t+\Delta t) - q_{ij}(t) = \{A'_{ij}(t+\Delta t) - D_{ij}(t+\Delta t)\} - q_{ij}(t)$. The $A'_{ij}(t+\Delta t) - D_{ij}(t+\Delta t)$ is the inflow rate minus outflow rate on link (i, j) during $[t, t+\Delta t)$ and hence depends on state s_t and action a_t during the time step. Given $q_{ij}(t)$, $\Delta q_{ij}(t)$ is therefore written as a function of s_t and a_t: $\Delta q_{ij}(t) = \Delta q_{ij}(s_t, a_t)$. Then, the reward is also written as $r_{t+1}(s_t, a_t) = -\sum_{(i,j)\in L} \Delta q_{ij}(s_t, a_t)$.

As explained using Figure 1, links $(i, j) \in L_j$ associated with individual intersection *j* do not overlap each other, reward $r_{t+1}(s_t, a_t)$ is separable for each associated intersection and written as the sum of rewards at individual intersections:

$$r_{t+1}(s_t, a_t) = \sum_j r_{t+1}^j(s_t, a_t)\,, \qquad (5)$$
$$r_{t+1}^j(s_t, a_t) = -\sum_{(i,j)\in L_j} \Delta q_{ij}(s_t, a_t), \quad j = 1, 2, ..., J\,.$$

3.4 Action-value

The action-value function $Q_\pi(s_t, a_t)$ under policy $\pi(a_t \mid s_t)$ is written below as the expected return starting from state s_t, taking action a_t, and then following policy $\pi(a_t \mid s_t)$:

$$Q_\pi(s_t, a_t) = E_\pi[r_{t+1}(s_t, a_t) + \gamma r_{t+2}(s_{t+1}, a_{t+1}) + \gamma^2 r_{t+3}(s_{t+2}, a_{t+2}) + ... + \gamma^{T-t} r_{T+1}(s_T, a_T) \mid s_t, a_t]\,, \quad (6)$$

where $\gamma \in [0, 1]$ = the discount rate.

Under the Markov property in the study environment, the action-value is written as the following recursive form explicitly using policy $\pi(a_t \mid s_t)$:

$$Q_\pi(s_t, a_t) = r_{t+1}(s_t, a_t) + \gamma E\ [\sum_{a_{t+1}} \pi(a_{t+1} | s_{t+1}) Q_\pi(s_{t+1}, a_{t+1}) \mid s_t, a_t]. \qquad (7)$$

where $\pi(a_t \mid s_t)$ = probability to take action a_t when state is s_t,

3.5 Decomposition of reward into individual intersections

Generally, $r_{t+1}^j(s_t, a_t)$ for individual intersection *j* is not independent each other because they depend on action a_t and state s_t at all intersections. However, as discussed in the previous section, a time-lag always exists between the time of signal control and the time when the wave caused by the control propagates to the adjacent intersections. Let us assume that the time-lag, equal to the wave travel time, always exceeds the discrete time interval Δt:

$$\min_t W_{ij}^f(t) = W_{ij}^f > \Delta t \quad \cap \quad \min_t W_{ij}^b(t) = W_{ij}^b > \Delta t, \quad \forall (i, j) \in L. \qquad (8)$$

This condition seems quite feasible because the order of Δt is a few seconds which would be smaller than wave travel time between intersections normally apart each other at least 100 meters in urban areas; that is, wave travel time would likely be more than a few seconds.

To evaluate $A'_{ij}(t)$, it is sufficient to know $A_{ij}(\cdot)$ until $t-W_{ij}^{f}$ as shown in Figure 1. And, to evaluate $D_{ij}(t)$ which could be affected by queue blocking back from the downstream, it is sufficient to know $D_{jk}(\cdot)$ on the downstream link (j,k) only until $t-W_{jk}^{b}$, $\forall k \in N_j$, $k \neq i$. If Condition (8) is satisfied, $t-W_{ij}^{f} < t-\Delta t$ and $t-W_{jk}^{b} < t-\Delta t$. And, since $A_{ij}(\cdot)$ and $D_{jk}(\cdot)$ are controlled by the upstream and downstream traffic lights, $A'_{ij}(t)$ and $D_{ij}(t)$ can be evaluated without being influenced by signal controls at other intersections after time $t-\Delta t$. Therefore, reward r_{t+1}^{j} is written as a function of only its own action a_t^j:

$$r_{t+1}(s_t, a_t) = \sum_j r_{t+1}^{j}(s_t, a_t^j) . \tag{9}$$

3.6 MARL-based network-wide control

Since action-value $Q_\pi(s_t, a_t)$ is the sum of future reward for an entire network, for network optimum control, action a_t must be taken to maximize $Q_\pi(s_t, a_t)$ for any state s_t considering all intersections simultaneously. This means the following optimum policy should be taken:

$$\pi(a_t \left| s_t \right.) = \begin{cases} 1\ , & a_t \ = \arg\max_a Q_\pi(s_t, a) \\ 0\ , & \text{otherwise} \end{cases} . \tag{10}$$

Let us denote the optimum policy as '*' and plug Eq.(10) into Eq.(7). Then, the following Bellman optimality equation is obtained:

$$Q_*(s_t\ , a_t\ ) = r_{t+1}(s_t, a_t\ ) + \gamma E\ [\max_{a_{t+1}} Q_*(s_{t+1}, a_{t+1}) \big|\ s_t\ , a_t\ ], \qquad 0 \leq t \leq T . \tag{11}$$

On the other hand, let us consider the MARL-based policy, $\pi^j(a_t^j \left| s_t \right.)$, in which individual intersection *j* independently takes action a_t^j to maximize its own action-value for any state s_t:

$$\pi^j(a_t^j \left| s_t \right.) = \begin{cases} 1\ , & a_t^j = \arg\max_a Q_\pi^j(s_t, a) \\ 0\ , & \text{otherwise} \end{cases} . \tag{12}$$

This MARL-based policy is denoted as 'Δ' and it is plugged into Eq.(7) to obtain the following:

$$\begin{aligned} Q_\Delta(s_t\ , a_t\ ) &= r_{t+1}(s_t, a_t\ ) + \gamma E\ [\sum_j \sum_{a_{t+1}^j} \pi^j(a_{t+1}^j \left| s_{t+1} \right.) Q_\Delta^j(s_{t+1}, a_{t+1}) \big|\ s_t\ , a_t\ ] \\ &= \sum_j [r_{t+1}^j(s_t, a_t^j\ ) + \gamma E\ [\max_{a_{t+1}^j} Q_\Delta^j(s_{t+1}, a_{t+1}) \big|\ s_t\ , a_t\ ]] \\ &= \sum_j Q_\Delta^j(s_t\ , a_t\ )\ , \end{aligned} \tag{13}$$

where $\quad Q^j_\Delta(s_t\,, a_t\,) = r^j_{t+1}(s_t, a^j_t\,) + \gamma E\ \left[\max_{a^j_{t+1}} Q^j_\Delta(s_{t+1}, a_{t+1})\right|\, s_t\,, a_t\,]\,.$ (14)

$\qquad\qquad\qquad$ = individual action-value under policy Δ

If $\sum_j \max_{a^j_t} Q^j_\Delta(s_t\,, a_t\,)$ is equal to $\max_{a_t} Q_*\,(s_t\,, a_t\,)$ for any state s_t, we can say that the MARL-based control achieves the network-wide optimum. The following is the proof of this.

First of all, at the final time step *T*, since the next state at time step $T+1$ is out of the study period, $Q^j_\Delta(s_T\,, a_T)$ is equal to $r^j_{T+1}(s_T\,, a^j_T\,)$ which depends on only its own action a^j_T, given s_T. Thus, $\sum_j \max_{a^j_T} Q^j_\Delta(s_T\,, a_T\,)$ becomes equal to $\max_{a_T} Q_*\,(s_T\,, a_T)$ as shown below.

$$\begin{aligned} \sum_j \max_{a^j_T} Q^j_\Delta(s_T\,, a_T\,) &= \sum_j \max_{a^j_T} r^j_{T+1}(s_T\,, a^j_T\,) \\ &= \max_{a_T} \sum_j r^j_{T+1}(s_T\,, a^j_T\,) = \max_{a_T} Q_*\,(s_T\,, a_T) \end{aligned} \tag{15}$$

Next, at one time step earlier $T-1$, based on the above result, $\sum_j \max_{a^j_{T-1}} Q^j_\Delta(s_{T-1}, a_{T-1})$ is shown to be equal to $\max_{a_{T-1}} Q_*\,(s_{T-1}, a_{T-1})$ as follows:

$$\begin{aligned} \sum_j \max_{a^j_{T-1}} Q^j_\Delta(s_{T-1}, a_{T-1}) &= \sum_j \max_{a^j_{T-1}} \left[r^j_T\,(s_{T-1}, a^j_{T-1}\,) + \gamma E\ \left[\max_{a^j_T} Q^j_\Delta(s_T\,, a_T)\right|\, s_{T-1}, a_{T-1}] \right] \\ &= \max_{a_{T-1}} \sum_j \left[r^j_T\,(s_{T-1}, a^j_{T-1}\,) + \gamma E\ \left[\max_{a^j_T} Q^j_\Delta(s_T\,, a_T)\right|\, s_{T-1}, a_{T-1}] \right] \\ &= \max_{a_{T-1}} \left[r_T\,(s_{T-1}, a_{T-1}\,) + \gamma E\ \left[\max_{a_T} Q_*\,(s_T\,, a_T)\right|\, s_{T-1}, a_{T-1}] \right] \\ &= \max_{a_{T-1}} Q_*\,(s_{T-1}, a_{T-1}) \end{aligned} \tag{16}$$

The right-hand side on the first line comes from Eq.(14). Given s_{T-1}, within [ ] for agent *j* on the first line, $r^j_T\,(s_{T-1}, a^j_{T-1})$ depends only on a^j_{T-1}, while $Q^j_\Delta(s_T\,, a_T) = r^j_{T+1}(s_T\,, a^j_T\,)$ depends on state s_T and action a^j_T. However, given s_{T-1}, whatever action each agent independently takes at time step $T-1$, next state $s_T\,(s_{T-1}, a_{T-1})$ is realized by the resulted action $a_{T-1} = \left(a^1_{T-1},\, a^2_{T-1}, \ldots\ldots, a^j_{T-1}, \ldots,\, a^J_{T-1}\right)$. Therefore, given s_{T-1} and a_{T-1}, $Q^j_\Delta(s_T\,, a_T)$ depends only on a^j_T, and hence maximization of $Q^j_\Delta(s_T\,, a_T)$ with respect to a^j_T, $\{\max_{a^j_T} Q^j_\Delta(s_T\,, a_T)\,|\, s_{T-1}, a_{T-1}\}$, is feasible. As a whole, [ ] equal to $Q^j_\Delta(s_{T-1}, a_{T-1})$ for agent *j* depends only on its own action a^j_{T-1} but independent of actions of others. Consequently, given s_{T-1}, if each agent *j* independently chooses action a^j_{T-1} to maximize $Q^j_\Delta(s_{T-1}, a_{T-1})$, next state s_T is realized by chosen a_{T-1} and $Q^j_\Delta(s_{T-1}, a_{T-1})$'s become independent each other under a_{T-1}. For this reason, the second line is derived. Finally, by plugging Eq.(15) into the second line, the third and fourth lines are obtained.

Repeating the same procedure backward along the time step, $\sum_j \max_{a_t^j} Q_\Delta^j(s_t, a_t)$ is proved to be equal to $\max_{a_t} Q_*(s_t, a_t)$ for any time step *t*, $0 \le t \le T$:

$$\sum_j \max_{a_t^j} Q_\Delta^j(s_t, a_t) = \max_{a_t} Q_*(s_t, a_t), \qquad 0 \le t \le T. \tag{17}$$

This is the most important finding in this study; that is, '*action-value* $Q_*(s_t, a_t)$ *representing the expected sum of future reward for an entire network from any state* s_t *can be maximized by the individual maximization of* $Q_\Delta^j(s_t, a_t)$ *under the MARL-based control*'. Normally, under the MARL-based control, the individual maximization of the actin-value does not lead to the maximization of the whole action-value because the reward of an agent is influenced by actions of others. However, if reward $r_{t+1}^j(s_t, a_t^j)$ depends on only its own action a_t^j under Condition (8), the MARL-based control establishes the network-wide optimum. Based on this result, if each intersection learns $Q_\Delta^j(s_t, a_t)$ through MARL and chooses action a_t^j so as to maximize $Q_\Delta^j(s_t, a_t)$ for any state s_t, such decentralized action at every intersection can optimize the network-wide control. Since the proposed method just fit with the standard MARL framework, any of MARL algorithms can be applied to find $Q_\Delta^j(s_t, a_t)$.

Using the above MARL-based approach, considerable action space can be saved. Also, regarding the state space, while the formulation uses state s_t across all intersections for generalization purposes, the concept of a state as an abstract representation of the environment can be tailored to support for an agent to appropriately select its action. Since intersection *j* selects its own action a_t^j in our problem, intersection *j* may not need conditions of intersections apart from *j*, but may suffice to consider state associated with its neighbor or exclusively associated with its own traffic conditions on links $(i, j) \in L_j$. The state design is examined in case studies in next section. Due to the action and state space savings, $Q_\Delta^j(s_t, a_t)$ could be efficiently estimated by less amount of training. Furthermore, actions and states can keep the same designs of their elements even if the number of intersections changes.

4 Case Studies

The case study is performed to confirm if the proposed MARL-based control could yield the network-wide optimum. For this purpose, we intend to design a simple network consisting of one-way links with no turning movements dealing with constant traffic demand to easily compare with control by the conventional signal control theory. It is important to note that the theoretical framework proposed is not limited to these simplified network and demand configurations. Although, only a case study under an oversaturated scenario is presented in this abstract, we have validated the proposed control indeed establishes the network-wide optimum in undersaturated as well as oversaturated conditions with queue blocking back.

4.1 Network and traffic demand

A network consists of 5 intersections with all one-way links as shown in Figure 2. Nodes 1 to 5 are intersections and nodes 6 to 17 are centroids. Major links are running in the *EW* direction with the saturation flow rate of 1800 [veh/h], while all minor links are running in the *NS* direction with the saturation flow rate of 1440 [veh/h]. The lengths of all minor links are 1000 [m], whereas those of the major links are 200 [m] with the exception of four links at both ends. For all links, a triangular fundamental diagram is applied in which the fixed forward and backward wave speeds are assumed $v_{ij}^{f} = 10$ [m/s] (=36 [km/h]) and $v_{ij}^{b} = -2.78$ [m/s] (= -10 [km/h]) respectively. The discrete time step Δt is 5 [s] so that Condition (8) is satisfied by the wave travel times > 5 [s] for all links.

The traffic demand rate is assumed constant and departs as well as arrives uniformly at all intersections (no stochasticity). The constant OD demand rate of 900 [veh/h] is supplied from centroids 6→7 on the major links, whereas the demand rate of 720 [veh/h] is supplied on every minor link in *S*→*N* direction. Therefore, the network is oversaturated because the saturation degrees on both the major and minor links become 0.5 (= 900/1800 = 720/1440) at all intersections.

4.2 Traffic model

To evaluate the cumulative curves $A_{ij}(t)$ and $D_{ij}(t)$ for all $(i, j) \in L$, we use the CTM (Cell Transmission Model; Daganzo, 1995), which is a well-known traffic simulation model based on the kinematic wave theory. For the CTM, a triangular fundamental diagram with the same forward and backward speeds as those mentioned above is used. The length of a cell is 50 [m] and the scan interval is 5 [s] because the free flow travel speed is 10 [m/s].

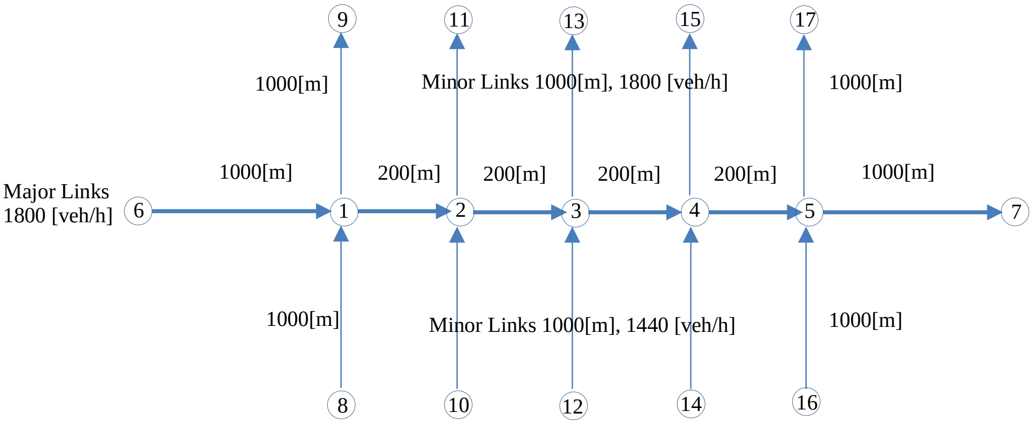


Figure 2: Arterial with 5 signalized intersections

4.3 Signal control

At every intersection, traffic is assumed to go straight only (no turnings) and all intersections are assumed to use a simple two-phase signal control: greens in the *EW* and in *NS* directions. Since oversaturated cases are examined, the green time is bounded by the 75-seconds maximum and the 15-seconds minimum because otherwise the cycle time could be infinity under an oversaturated condition.

4.4 Reinforcement learning

Action a_t^j at intersection *j* takes either 0 or 1: green is given to *EW* (*NS*) direction, if $a_t^j = 0$ (1). State s_t, common for all intersections, is designed to include queues and inflows of all intersections at the end of time step t:

$$s_t = \left(q_{ij}(t),\ \Delta A'_{ij}(t),\ a^j_{t-1} \right),\ \ (i,j) \in L_j,\ \ j = 1, 2, ..., J\ , \tag{18}$$

where $\Delta A'_{ij}(t) = A'_{ij}(t + \Delta t) - A'_{ij}(t)$.

Since every intersection has two approach links, two $q_{ij}(t)$'s and two $\Delta A'_{ij}(t)$'s are included in s_t for one intersection. Furthermore, the action in the previous time step, a^j_{t-1} is included. Therefore, state s_t consists of total 25 (= (2+2+1) * 5 intersections) elements. The reason for the inclusion of last action a^j_{t-1} is to consider the lost time equal to Δt when selected actions (phases) are changed.

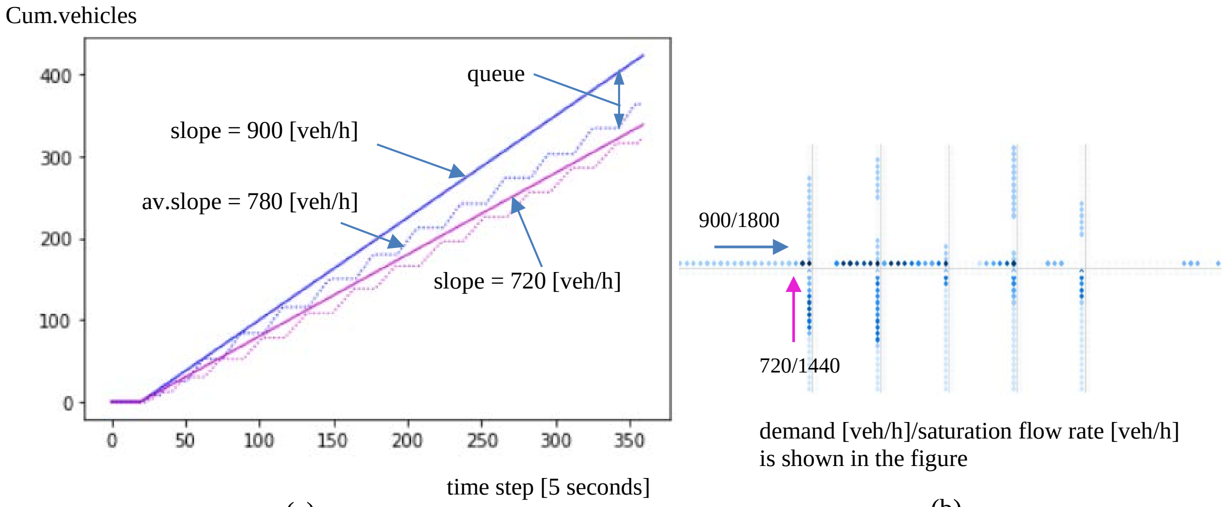
For the MARL training, we use the Deep Q-Network in which three layers with 128 neurons each are designed. The discount rate γ is 0.95 and the ε-greedy parameter is assumed to decrease linearly from an initial value of 0.5 to 0.0 at the final episode. The MARL is trained with 200 [episodes] with 360 [time steps/ episode] for all the cases.

4.5 Case study : Oversaturated Scenario

The total demand from all minor links is 3600 [veh/h] (=720 [veh/h] times 5 origins) which is significantly larger than the demand of 900 [veh/h] on the major links. Also, the total saturation flow rate on 5 minor links is 720*5=3600 [veh/h] larger than 1800 [veh/h] of the major link. Therefore, the priority should be given to the minor links at all intersections to maximize the throughput. To fully discharge demands on the minor links, 50% of the cycle time should be used for them, and the rest of 50% should be assigned to green on the major links and the 10-second lost time. Clearly, to maximize the throughput, the longer cycle time is advantageous by giving the maximum green time to the minor links. The optimal signal parameters are therefore $\hat{G}_{major} = 65$ [s], $\hat{G}_{minor} = 75$ [s], and $\hat{C} = 150$ [s]. Under this control, the average flow rate on the major link is 780 [veh/h] (=1800*65/150), and the queue on the major link grows at the rate of 120 [veh/h] (=900-780).

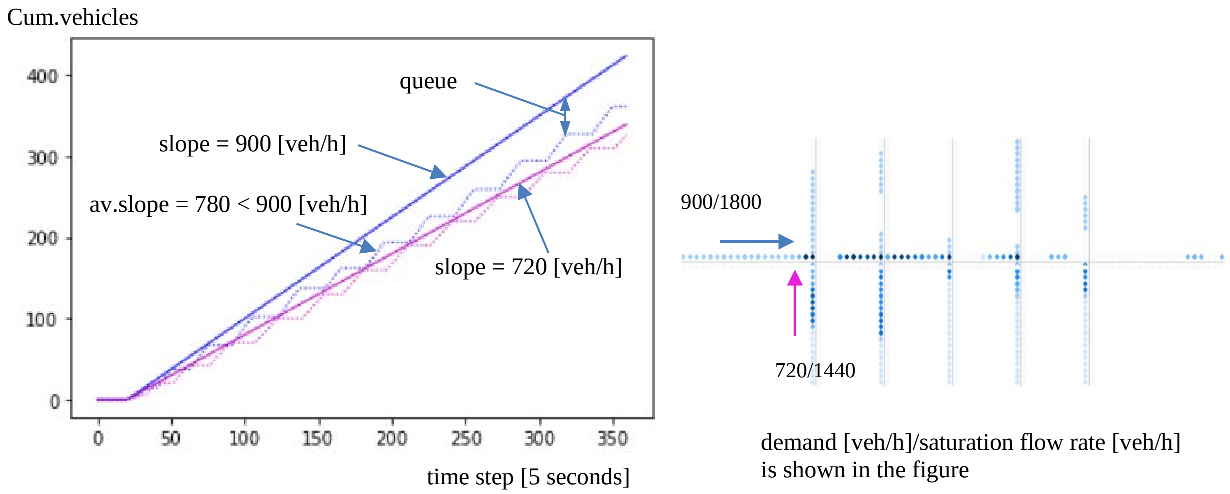
After the training, signal parameters of $G_{major} = 64$ [s] and $G_{minor} = 74$ [s] are obtained as the average values during the last 90 time steps as shown in Figure 3. The yielded green times are almost consistent with the optimum control that prioritizes the minor links. The queues are growing nearly at the expected rate of 120 [veh/h] on the major links but a queue vanishes at the end of green on every minor link due to the given priority. This is the clear evidence that the proposed MARL-based control yields the network-wide optimum with autonomous coordination that maximizes the throughput. Since the network consists of all one-way links, an upstream intersection is not directly influenced by traffic waves from downstream intersections. However, they are still autonomously coordinated each other to give the priority to the minor links. If each intersection were to concern only its own throughput, the priority must have been given to the major link because of the larger demand and saturation flow rate of 1800 [veh/h] than those on the minor link.

Figure 4 shows the result using the state including queues and inflows only associated with an individual intersection; that is, the number of state elements equal to 9 (= (4+4+1)). The yielded signal parameters are $G_{major} = 69$ [s] and $G_{minor} = 75$ [s] also close to the optimum and the queues completely disappear on the minor links. This almost same result as in Figure 3 suggests the possibility that the state element does not have to include the entire traffic condition over the network but only the condition around individual intersections.



(a) (b)

Figure 3: Cumulative curves resulted in oversaturated condition (25 state elements)



(a) (b)

Figure 4: Cumulative curves resulted in oversaturated condition (9 state elements)
(the state includes inflows and queues associated with an individual intersection)

5 Summary

This study theoretically argues that the MARL-based signal control using MARL can achieve the network-wide optimum control even in the absence of collaborative arrangements among intersections. The argument is grounded on the fact that traffic waves caused by signal control at an intersection always take some time to propagate to other intersections. This property, common across diverse traffic flows, suggests the potential extension of this decentralized optimization to other traffic controls, such as ramp control. Moreover, beyond the realm of traffic and transportation, similar MARL-based approaches might find applications in systems where the impact of an agent's action influences others after a certain duration.

References

Chow, A.H.F., Sha, R., and Li, S. (2020). Centralised and decentralised signal timing optimisation approaches for network traffic control. *Transportation Research Part C: Emerging Technologies*, 113:108–123.

Daganzo, C.F. (1995). The cell transmission model, part II: Network traffic. *Transportation Research Part B: Methodological*, 29(2): 79–93.

Haddad, T.A., Hedjazi, D., and Aouag, S. (2022). A deep reinforcement learning-based cooperative approach for multi-intersection traffic signal control. *Engineering Applications of Artificial Intelligence*, 114.

Li, Z., Yu, H., Zhang, G., Dong, S., and Xu, C-Z. (2021). Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning. *Transportation Research Part C: Emerging Technologies*, 125.

Newell, G.F. (1993). A simplified theory of kinematic waves in highway traffic, part II: Queueing at freeway bottlenecks. *Transportation Research Part B: Methodological*, 27(4):289–303.

Smith, M.J. (1979). The existence, uniqueness and stability of traffic equilibria. *Transportation Research Part B: Methodological*, 13(4):295–304.

Smith, M.J. (1984). The Stability of a Dynamic Model of Traffic Assignment – An Application of a Method of Lyapunov. *Transportation Science*, 18(3):245–252.

Smith, M.J., Liu, R., and Mounce, R. (2015). Traffic Control and Route Choice; Capacity Maximization and Stability. *Transportation Research Part B: Methodological*, 81:863–885.