Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-3 (2016) 067–072

Traffic Signal Control based on Markov Decision Process ⋆

Yunwen Xu * **Yugeng Xi** * **Dewei Li** * **Zhao Zhou** *

* *Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China.(e-mail: willing419@sjtu.edu.cn, ygxi@sjtu.edu.cn, dwli@sjtu.edu.cn, zzhou553@gmail.com).*

Abstract:
This paper proposes a Morkov state transition model for an isolated intersection in urban traffic and formulates the traffic signal control problem as a Markov Decision Process(MDP). In order to reduce computational burden, a sensitivity-based policy iteration(PI) algorithm is introduced to solve the MDP. The proposed model is stage-varying according to traffic flow variation around the intersection, and the state transition matrices and cost matrices are updated so that a new optimal policy can be searched by the PI algorithm. The proposed model also can be easily extended from an isolated intersection to a traffic network based on the space-time distribution characteristics of traffic flow, so as the PI algorithm. The numerical experiments of a small traffic network show that this approach is capable of reducing the number of vehicles substantially compared with the fixed-time control particularly for high traffic demand, while being computationally efficient.

Keywords: Markov state transition model; Traffic signal control; Policy iteration algorithm; Markov decision process

1. INTRODUCTION

Traffic congestion is a stringent issue in modern society due to increasing population and economic activity, which motivates the requirement for better utilization of the existing infrastructures and for efficiently control of the traffic flow. Among all the measures, traffic signal control is a major component whose enhancement is the most efficient way to reduce the traffic congestion.

The signal control problem has been studied extensively in the literature. Except for the early developed adaptive signal control systems, such as SCOOT(Hunt et al. (1982)), SCATS(Lowrie (1982)), the existing signal control problem generally contains two aspects: (1) a mathematical model for the complex traffic system, and (2) an appropriate control law such that the behavior of the system meets certain performance indices. Aboudolas et al. (2009), Lin et al. (2011), Zhou et al. (2014), Lo (2001), Beard and Ziliaskopoulos (2006) and Aziz and Ukkusuri (2012) described the signal control as a mathematical programming problem with embedded deterministic traffic flow models. However, for large-scale road networks, most of above strategies are with heavy computational burden.

Abdulhai et al. (2003), Balaji et al. (2010), Prashanth and Bhatnagar (2011), El-Tantawy et al. (2013), Robertson and Bretherton (1974) and Yu and Recker (2006) applied the framework of Markov Decision Process(MDP) to identify signal control in random traffic environment as a sequential decision-making problem, where dynamic programming(DP) and reinforcement learning(RL), as well as their variants, were commonly used. The DP also has a bottleneck of the computational complexity in the recursive calculation of Bellman's equation which is exponential to the size of state space and action space. RL based methods have a major advantage that the optimization could be executed off line, but a stochastically stable environment and constant state-action costs are the premise for getting the eventual optimal stationary control policy, which restrict the random events to be handled, such as a surge of traffic flow and adverse weather.

In this paper, we propose a Morkov state transition model for an intersection to describe the random nature of the traffic system. This model can be easily extended from an isolated intersection to a traffic network according to the space-time distribution characteristics of traffic flow. The entire traffic signal control problem is described as a MDP based on the proposed model. To solve the MDP efficiently, a PI algorithm from the sensitivity viewpoint is introduced. The computational complexity of the PI algorithm is very small when the number of policies is less than 10^{10} (see details in Cao (2007) section 4.1) which is much larger than the optimal problem formulated in this paper. The simulation on microscopic traffic simulation software(CORSIM) shows that the PI algorithm with the proposed model can reduce the number of vehicles significantly compared with the fixed-time control especially under the scenarios of high traffic demand.

⋆ This work is supported in part by the National Science Foundation of China (Grant No. 61374110, 61433002, 61221003), NSFC International Cooperation Project (Grant No. 71361130012).

10.1016/j.ifacol.2016.07.012

The rest of this paper is organized as follows. In Section 2, we give a brief introduction of MDP and the PI algorithm from sensitivity viewpoint. The detailed description of the proposed Markov state transition model is provided in Section 3. Section 4 demonstrates the simulation results and the numerical comparison with the fixed-time control. Finally, Section 5 concludes the paper.

2. MORKOV DECISION PROCESS AND POLICY ITERATION ALGORITHM

2.1 Markov decision process

The MDP has been extensively studied in the literature. Interested readers can find a good introduction to MDP in the book by Puterman (2014). For the sake of completeness, a brief introduction to discrete-time MDP is given below.

In a MDP, at any time k, $k = 0, 1, 2, \ldots$, the system is in a state $X_k \in S$, where $S = \{1, 2, \ldots, n\}$ is a finite state space. According to the Markov property of a state process, its future behavior is determined only by the current state X_k which contains all the information of system history.

In addition to the state space, there is an action space A. If the system is in state s, $s \in S$, we can take (independently from the actions taken in other states) any action $a \in A(s) \subseteq A$ and apply it to the system, where $A(s)$ is the set of actions that are available in state $s \in S$, $A = \cup_{s\in S} A(s)$. A (stationary and deterministic) policy is a mapping from S to A, denoted as $d = (d(1), d(2), \ldots, d(n))$, with $d(s) \in A(s)$, $s = 1, 2, \ldots, n$, that determines the action taken in state s. We use $D = \times_{s\in S} A(s)$ to denote the space of all possible policies, where "$\times$" is called a Cartesian product, which is a direct product of sets.

Therefore, if policy d is adopted, the state transition probability matrix is $P^d = [p^{d(s)}(s_1 \,|s)]_{s,s_1=1}^n$. The reward vector is denoted as $f^d = (f(1, d(1)), f(2, d(2)), \ldots, f(n, d(n)))^T$. The steady-state probability vector of a Markov chain under policy d is denoted as $\pi^d = (\pi^d(1), \pi^d(2), \ldots, \pi^d(n))$ and $\pi^d = \pi^d \cdot P^d$. Considering an ergodic Markov chain under policy d, the long-run average reward is :

$$\eta^d = \lim_{L\to\infty} \left\{ \frac{1}{L} \sum_{l=0}^{L-1} f[X_l, d(X_l)] \right\} = \pi^d f^d \tag{1}$$

in which L is the length of Markov chain.

2.2 Sensitivity-based Policy iteration algorithm

There are two basic approaches to solve the standard MDPs, value iteration and policy iteration. Value iteration is basically a numerical approach, such as Q-learning(Sutton and Barto (1998)). In this subsection, a PI algorithm based on the sensitivity is introduced. Its basic principle can be briefly described as(Cao (2007)): by observing and analyzing the behavior of a system under a policy, find another policy that perform better, if such a policy exists.

Suppose η_h and η_d are the long-run average rewards corresponding to two policies h and d, $h, d \in D$. The comparison of two policies is based on performance difference formula:

$$\eta^h - \eta^d = \pi^h[(f^h + P^h g^d) - (f^d + P^d g^d)] \tag{2}$$

where $g^d = [g^d(1), g^d(1), \ldots, g^d(n)]$ is the performance potential vector of policy d, and $g^d(s)$, $s \in S$, measures the "potential" contribution of state s to the long-run average reward η^d. It can be calculated from the Poisson equation:

$$(I - P^d)g^d + \eta^d e = f^d \tag{3}$$

where I is a unit matrix and $e = (1, 1, \ldots, 1)$. The PI algorithm flowchart is shown in Algorithm 1.

Algorithm 1 Policy Iteration Algorithm (Cao (2007))

1: Guess an initial policy d_0, set $k = 0$.
2: (Policy evaluation) Obtain the potential g^{d_k} by solving the poisson equation $(I - P^{d_k})g^{d_k} + \eta^{d_k} e = f^{d_k}$.
3: (Policy improvement) Choose

$$d_{k+1} \in \arg\{\max_{d\in D}[f^d + P^d g^{d_k}]\} \tag{4}$$

component-wisely(i.e., to determine an action for each state). If in state s, action $d_k(s)$ attains the maximum, then set $d_{k+1}(i) = d_k(i)$.
4: If $d_{k+1} = d_k$, stop; otherwise, set $k := k + 1$ and go to step 2.

Therefor, for a specific control problem, once the state transition probability matrix P and the corresponding reward vector f for each available policy are defined, then by maximizing the long-run average reward η, a policy for choosing an optimal action for each state can be obtained, which represents the optimal strategy that should be followed.

3. MARKOV STATE TRANSITION MODEL FOR AN INTERSECTION

In this section we present a Markov state transition model for an intersection, which consists of two steps: calculating state transition matrices for in-roads of an intersection and then Markov state transition modelling for the intersection.

3.1 Notation and network description

Consider a traffic network composed of a number of intersections and roads. Each intersection $j \in J$ consists of several in-roads, I_j, which are mutually disjoint, and denote $I = \cup_{j\in J} I_j$. Each road $i \in I$ with N_i lanes has a number of upstream roads $I_{i,up}$ and downstream roads $I_{i,down}$. The vehicles entering road come from the $I_{i,up}$, and the vehicles leaving road i drive to the $I_{i,down}$ with different turning options. Service time (green light) for road i is represented by a vector $t_g^i = [t_{g,L}^i, t_{g,T}^i, t_{g,R}^i]^T$, denoting service time for left, through and right turning option respectively.

In the rest of this section we assume that all intersections have a common cycle length T, the time devoted to serving vehicles from different in-roads at the intersection. We also define a uniform cycle length T_m for markov model updating, $T_m = M \cdot T$ with M an integer. Control decisions in our policy are then made at the beginning of each cycle T_m based on the assumption that the traffic flow remains unchanged at the short term (T_m) in the small area of