Automatica 54 (2015) 29–35

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

# Parameterized Markov decision process and its application to service rate control☆

Li Xia, Qing-Shan Jia [1]

*CFINS, Department of Automation, TNLIST, Tsinghua University, Beijing 100084, China*

ARTICLE INFO

*Article history:*
Received 3 September 2013
Received in revised form
15 October 2014
Accepted 19 December 2014
Available online 11 February 2015

*Keywords:*
Markov decision process
Discrete event systems
Parameterized policy
Policy iteration
Service rate control

ABSTRACT

In this paper, we discuss the optimization of Markov decision processes (MDPs) with parameterized policy, where the state space is partitioned and a parameter is assigned to each partition. The goal is to find the optimal parameters which maximize the long-run average performance. The traditional policy iteration is usually inapplicable to parameterized policy because the parameter tuning at different states are correlated. With some appropriate assumptions and special conditions, we develop a modified policy iteration type algorithm to find the optimal parameters. Compared with the traditional gradient-based approaches for MDP with parameterized policy, this policy iteration type approach is much more efficient. Finally, as an example, we apply this approach to a service rate control problem in closed Jackson networks. As compared with the gradient-based approach which is trapped into local optimum, our approach is demonstrated to efficiently find the optimal service rates in global scope.



## 1. Introduction

Markov decision process (MDP) is a fundamental mathematical model to study the performance optimization of stochastic dynamic systems and it has been extensively studied in the literature (Feinberg & Shwartz, 2002; Guo & Hernandez-Lerma, 2009; Puterman, 1994). In the theory of MDP, the policy is a mapping from the state space to the action space. However, in many practical problems, the parameterized policy is widely used since its form is much simpler. The state transition probabilities and rewards of Markov systems will change according to the value of these parameters. The parameterized policy does not fit the standard definition of policy in MDP and the traditional approaches, such as the policy iteration, cannot be directly applied to this problem. Our target is to find the optimal parameters which maximize the average performance of Markov systems and we call it *parameterized Markov decision process* (Baxter & Bartlett, 2001; Xia & Jia, 2013).

In the literature, the gradient-based method is the main thread to optimize the parameterized policy of Markov systems. As the Markov system is stochastic, the stochastic approximation is widely used. The key problem is transformed to how to efficiently calculate or estimate the performance gradient of the system performance w.r.t. (with respect to) the parameters. Among all the research efforts for this problem, perturbation analysis (PA) is one of the successful approaches. PA was proposed by Ho and Cao (1983) and it can provide an unbiased and strongly consistent estimate of the gradient only based on single sample path when the sampled function is stochastically Lipschitz continuous (Ho & Cao, 1991). Likelihood-ratio (LR) (Glynn, 1990) and simultaneous perturbation (SP) (Spall, 1992) are other commonly used approaches to efficiently estimate the gradient with much fewer samples in simulation. Thus, these approaches are especially efficient for the problems with high dimensional parameter vectors. Along the direction of PA, Cao and Chen (1997) proposed the direct-comparison theory of MDP. This is a new sensitivity-based framework to optimize the performance of Markov systems and some efficient algorithms are also proposed (Cao, 2007; Cao & Zhang, 2004). In the society of artificial intelligence, a so-called policy gradient method was proposed (Baxter & Bartlett, 2001; Marbach & Tsitsiklis, 2001; Sutton, McAllester, Singh, & Mansour, 2000) and it can also be unified in the framework of sensitivity-based approach. However, the gradient-based methods suffer from the intrinsic deficiencies, such as the slow convergence speed, difficulty of selecting the step-size, dependence on the initial value of parameters, and being trapped into a local optimum. For example, in a service rate control problem discussed in Section 4, we will show that the gradient-based

☆ The material in this paper was partially presented at the Asian Control Conference 2013, June 23–26, 2013, Istanbul, Turkey. This paper was recommended for publication in revised form by Associate Editor Bart De Schutter under the direction of Editor Ian R. Petersen.

*E-mail addresses:* xial@tsinghua.edu.cn (L. Xia), jiaqs@tsinghua.edu.cn (Q.-S. Jia).

[1] Tel.: +86 10 62773006; fax: +86 10 62796115.

http://dx.doi.org/10.1016/j.automatica.2015.01.006
0005-1098/

algorithm is often trapped into a local optimum, as illustrated in Fig. 2. Although we can utilize some global search techniques (Hong & Nelson, 2006) to improve the exploration ability for global optimum, the deficiencies of gradient-based method cannot be thoroughly solved.

Therefore, a question follows naturally: Can we use the policy iteration to solve the parameterized MDP since the policy iteration is much more efficient than the gradient-based method and it can find the global optimum? In this paper, we study a special category of parameterized MDP, where the state space is partitioned and an action (parameter tuning) is assigned to each partition. We use the direct-comparison theory to develop a policy iteration type algorithm for such parameterized MDP. The key idea of direct-comparison theory is the difference equation, which quantifies the performance difference of Markov systems under any two policies or parameter settings (Cao, 2007). Difference equation gives a straightforward perspective to study the relation between the system performance and parameters. The performance difference may provide more sensitivity information than the performance gradient. With the difference equation as a basis, we can clearly analyze the optimization of parameterized MDP and obtain the sufficient conditions to develop the policy iteration algorithm. This gives us a new direction to study the parameterized MDP, besides the traditional gradient-based methods. Finally, as an example, we study a service rate control problem in closed Jackson networks to illustrate our approach. Numerical experiments are conducted to demonstrate the algorithm efficiency.

## 2. Problem formulation

Consider a discrete time Markov chain $\boldsymbol{X} := \{X_t,\ t = 0, 1, 2, \ldots\}$, where $X_t$ is the system state at time epoch $t$. The state space $\mathcal{S}$ is assumed finite. Without loss of generality, we denote $\mathcal{S} := \{1, 2, \ldots, S\}$, where $S$ equals the size of the state space. The Markov chain is controlled by a parameterized policy (Baxter & Bartlett, 2001; Bhatnagar, Sutton, Ghavamzadeh, & Lee, 2009) and the parameters are denoted as a vector $\boldsymbol{\theta} := (\theta_1, \theta_2, \ldots, \theta_k)$ in a $k$-dimensional real number space $\mathbb{R}^k$. The parameters $\boldsymbol{\theta}$ affect the state transition probability and the reward function. We denote the state transition probability as $p^{\boldsymbol{\theta}}(s'|s)$ and the reward function as $f^{\boldsymbol{\theta}}(s)$, $s, s' \in \mathcal{S}$.

In many cases, the effects of parameters $(\theta_1, \theta_2, \ldots, \theta_k)$ on transition probabilities $p(\cdot|s)$ and $f(s)$ are decomposable. That is, if we change the value of one parameter, say $\theta_i$, it affects only part of transition probabilities $p(\cdot|s)$'s and reward functions $f(s)$'s. For example, consider a service rate control problem in a closed Jackson network with 3 servers and 6 customers (the detailed formulation of a closed Jackson network can be referred to Section 4 and Gordon & Newell, 1967). We want to optimize the service rates of server 1. Thus, the parameter $\theta_i$ is the load-dependent service rate $\mu_{1,i}$, $i = 1, 2, \ldots, 6$. The system state $s$ is a vector representing the queue length (include the customer being served) of these 3 servers. Suppose the reward function is $f^{\boldsymbol{\theta}}(s) = s(1) + \mu_{1,s(1)}$, where $s(1)$ is the first element of state vector $s$ (i.e., the queue length of server 1). If we change the value of parameter $\theta_2$, i.e., $\mu_{1,2}$, the transition probabilities $p(\cdot|s)$'s and reward function $f(s)$'s are affected only when $s \in \{(2, 0, 4), (2, 1, 3), (2, 2, 2), (2, 3, 1), (2, 4, 0)\}$, where the queue length of server 1 is 2. For other system states, such as $s = (1, 3, 2)$ or $s = (4, 1, 1)$, the change of parameter $\mu_{1,2}$ will not affect the value of $p(\cdot|s)$ or $f(s)$. Therefore, we can have the following definition to partition the state space $\mathcal{S}$.

**Definition 1.** $\mathcal{S}_i$ is defined as the set of states $s$ whose transition probabilities $p(\cdot|s)$ and reward function $f(s)$ are affected by $\theta_i$, $i = 1, 2, \ldots, k$.

Different parameters $\theta_i$'s have different $\mathcal{S}_i$'s and we have the following assumption

**Assumption 1.** $\mathcal{S}_i$'s are mutually exclusive, i.e., $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$ when $i \neq j$ and $i, j = 1, 2, \ldots, k$.

Assumption 1 means that the state space $\mathcal{S}$ can be partitioned by parameters $\boldsymbol{\theta}$ and every state's transition probability $p(\cdot|s)$ and reward function $f(s)$ are controlled by only one parameter $\theta_i$, where $s \in \mathcal{S}_i$. Therefore, we can further denote $p^{\boldsymbol{\theta}}(\cdot|s)$ and $f^{\boldsymbol{\theta}}(s)$ as $p^{\theta_i}(\cdot|s)$ and $f^{\theta_i}(s)$ respectively, where $s \in \mathcal{S}_i$.

Usually, the partition results of $\mathcal{S}_i$'s are not affected by the value of parameters $\boldsymbol{\theta}$. This is determined by the problem structure. That is, we have the following assumption.

**Assumption 2.** The value change of parameters $\boldsymbol{\theta}$ does not affect the structure of $\mathcal{S}_i$'s, $i = 1, 2, \ldots, k$.

For completeness, we further define $\mathcal{S}_0$ as the set of states whose transition probability and reward function are not affected by the parameters $\boldsymbol{\theta}$. With Assumptions 1 and 2, we see that the state space $\mathcal{S}$ is partitioned as a series of subsets $\mathcal{S}_i$'s. That is, $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_k$ and $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$, $i, j = 0, 1, \ldots, k$. The transition probability $p^{\theta_i}(\cdot|s)$ and the reward function $f^{\theta_i}(s)$ are affected only by the parameter $\theta_i$ and independent of other $\theta_j$'s, where $s \in \mathcal{S}_i$. Still use the aforementioned service rate control problem as an example. The parameter $\theta_i$ is the load-dependent service rate $\mu_{1,i}$, $i = 1, 2, \ldots, 6$. The state space $\mathcal{S}$ is partitioned as a series of subsets according to $\theta_i$'s. That is $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_6$, where $\mathcal{S}_0 = \{(0, 0, 6), (0, 1, 5), (0, 2, 4), (0, 3, 3), (0, 4, 2), (0, 5, 1), (0, 6, 0)\}$, $\mathcal{S}_1 = \{(1, 0, 5), (1, 1, 4), (1, 2, 3), (1, 3, 2), (1, 4, 1), (1, 5, 0)\}$, $\ldots$, $\mathcal{S}_5 = \{(5, 0, 1), (5, 1, 0)\}$, $\mathcal{S}_6 = \{(6, 0, 0)\}$.

The steady state probability of the Markov system staying at state $s$ is denoted as $\pi(s)$ and $\boldsymbol{\pi} := (\pi(1), \pi(2), \ldots, \pi(S))$ is a row vector. The long-run average performance of the Markov system is denoted as $\eta$. To reflect the effect of parameter $\boldsymbol{\theta}$, we rewrite $\boldsymbol{\pi}$ and $\eta$ as $\boldsymbol{\pi}^{\boldsymbol{\theta}}$ and $\eta^{\boldsymbol{\theta}}$, respectively. For ergodic chains, $\eta^{\boldsymbol{\theta}}$ can be written as follows.

$$\eta^{\boldsymbol{\theta}} = \lim_{T\to\infty} \frac{1}{T} E\left\{\sum_{t=0}^{T-1} f^{\boldsymbol{\theta}}(X_t)\right\}, \tag{1}$$

which is independent of the initial state $X_0$. Obviously, we can rewrite the above definition as

$$\eta^{\boldsymbol{\theta}} = \sum_{s\in\mathcal{S}} \pi^{\boldsymbol{\theta}}(s) f^{\boldsymbol{\theta}}(s) = \boldsymbol{\pi}^{\boldsymbol{\theta}} \boldsymbol{f}^{\boldsymbol{\theta}}, \tag{2}$$

where $\boldsymbol{f}^{\boldsymbol{\theta}} := (f^{\boldsymbol{\theta}}(1), f^{\boldsymbol{\theta}}(2), \ldots, f^{\boldsymbol{\theta}}(S))^T$ is a corresponding column vector. We denote $\boldsymbol{P}^{\boldsymbol{\theta}}$ as the corresponding transition probability matrix. We have $\boldsymbol{P}^{\boldsymbol{\theta}}\boldsymbol{e} = \boldsymbol{e}$, $\boldsymbol{\pi}^{\boldsymbol{\theta}}\boldsymbol{P}^{\boldsymbol{\theta}} = \boldsymbol{\pi}^{\boldsymbol{\theta}}$, and $\boldsymbol{\pi}^{\boldsymbol{\theta}}\boldsymbol{e} = 1$, where $\boldsymbol{e}$ is an $S$-dimensional column vector of 1.

The value domain of parameter $\theta_i$ can be a real-number interval denoted as $\mathbb{D}_i$, $i = 1, 2, \ldots, k$. Thus, the value domain of $\boldsymbol{\theta}$ is denoted as $\mathbb{D} := \mathbb{D}_1 \times \mathbb{D}_2 \times \cdots \times \mathbb{D}_k$, where $\times$ is the Cartesian product. Our goal is to find the optimal parameter $\boldsymbol{\theta}^*$ which maximizes the average performance of the parameterized MDP. This optimization problem is mathematically formulated as below.

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}\in\mathbb{D}} \{\eta^{\boldsymbol{\theta}}\}. \tag{3}$$

Assumptions 1 and 2 limit our study to a special category of parameterized MDP, where the state space is partitioned and a parameter to be tuned is assigned to each partition. Please note, our parameterized MDP problem is different from another parameter optimization problem called LSPI (Least Squares Policy Iteration) in MDP. LSPI aims to find the optimal parameters (weights) of the basis functions to approximate the value function (Lagoudakis & Parr,