第 30 卷 第 2 期
2004 年 3 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 30, No. 2
Mar., 2004

A Simulation Optimization Algorithm for CTMDPs Based on Randomized Stationary Policies[1)]

TANG Hao[1,2] XI Hong-Sheng[1] YIN Bao-Qun[1]

[1] (*Department of Automation, University of Science and Technology of China, Hefei* 230026)

[2] (*Department of Computer, Hefei University of Technology, Hefei* 230009)

(E-mail: tangh@ustc.edu; xihs@ustc.edu.cn)

Abstract Based on the theory of Markov performance potentials and neuro-dynamic programming (NDP) methodology, we study simulation optimization algorithm for a class of continuous time Markov decision processes (CTMDPs) under randomized stationary policies. The proposed algorithm will estimate the gradient of average cost performance measure with respect to policy parameters by transforming a continuous time Markov process into a uniform Markov chain and simulating a single sample path of the chain. The goal is to look for a suboptimal randomized stationary policy. The algorithm derived here can meet the needs of performance optimization of many difficult systems with large-scale state space. Finally, a numerical example for a controlled Markov process is provided.

Key words Performance potentials, neuro-dynamic programming, simulation optimization

1 Introduction

Many real systems, such as communication networks and flexible manufacturing systems can be modeled as Markov decision processes (MDPs). Their performance optimization has become an important direction of DEDS research field. Traditional computation based methods involve much computation of matrix inverse thereby consuming much time and rely on the precise model in the form of transition probability matrices or infinitesimal generators. For large-scale systems with curse of dimensionality and curse of modeling, these methods will be inapplicable. The theory of Markov performance potentials, introduced and developed by Cao X R and Chen H F for the sensitivity analysis of MDPs[1], has been used to solve the simulation optimization problems[2]. The proposed optimization methods dispense with matrix inverse, but still need some tables to show the relation between performance potentials or polices and the states one by one. So much storage space is still required. In order to avoid the curse of dimensionality, Bertsekas D P and Tsitsiklis J N developed the neuro-dynamic programming (NDP) based optimization theory for discrete time MDPs (DTMDPs)[3]. Thereafter, further work on NDP has been made[4,5], and paved the way for optimization study of CTMDPs. In this paper, we combine performance potential theory and NDP methodology to study a class of CTMDPs under randomized stationary policies. The policy parameters will be updated according to the gradient estimates by transforming the processes to its uniform Markov chain and simulating a single sample path of the chain. The randomized policies are presented by some approximation architecture with fewer numbers of parameters than the states. Thus the storage space is saved and the curse of dimensionality is avoided.

2 Problem description and fundamental theory

Consider an ergodic continuous-time Markov process $\{X_t, t\in[0,\infty)\}$ with finite state

1) Supported by National Natural Science Foundation of P. R. China(60274012) and the Natural Science Foundation of Anhui Province(01042308)

Received August 8, 2002; in revised form July 22, 2003

收稿日期 2002-08-08; 收修改稿件日期 2003-07-22

space $\Phi=\{1,2,\cdots,M\}$ and finite action space A. Denote $Q(a)=[q_{ij}(a)]$ as the infinitesimal generator of X_t controlled by action a, and $f(i,a)$ as the cost paid for per unit time whenever action a is taken at state i. Assume that $Q(a)$ is conservative. As Φ and A are finite, for any $i\in\Phi, a\in A$, $-q_{ii}(a)$ is uniformly bounded. Let λ be the bound. Suppose any parameter vector $\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots,\theta_K)\in\Theta\subseteq\Re^K$ determines one and only one randomized stationary policy $\mu(\boldsymbol{\theta})$, which assigns a probability distribution function mapping state $i\in\Phi$ to action $a\in A$. Under this policy, action a is taken at state i with probability $\mu_a(i,\boldsymbol{\theta})$. Such policies represented in terms of parameters are called parameterized policies. In NDP approaches, $\mu(\boldsymbol{\theta})$ can be constructed through the outputs of an artificial neural network or other approximation architectures, and θ will be the weights of the network. Define $Q(\boldsymbol{\theta})=[q_{ij}(\boldsymbol{\theta})]$ and $f(\boldsymbol{\theta})=(f(1,\boldsymbol{\theta}),\cdots,f(M,\boldsymbol{\theta}))^{\tau}$, which satisfy

$$q_{ij}(\boldsymbol{\theta})=\sum_{a\in A}\mu_a(i,\boldsymbol{\theta})q_{ij}(a),\quad f(i,\boldsymbol{\theta})=\sum_{a\in A}\mu_a(i,\boldsymbol{\theta})f(i,a),\quad \forall i,j\in\Phi \tag{1}$$

Obviously $Q(\boldsymbol{\theta})$ is well defined, meaning the average transition rate matrix of the process under $\mu(\boldsymbol{\theta})$. We call $X(\boldsymbol{\theta})=\{X_t,\Phi,A,Q(\boldsymbol{\theta}),f(\boldsymbol{\theta})\}$ a parameterized CTMDP constrained on Θ. Let $\pi(\boldsymbol{\theta})=(\pi(1,\boldsymbol{\theta}),\cdots,\pi(M,\boldsymbol{\theta}))$ be the stationary distribution of $X(\boldsymbol{\theta})$, which satisfies

$$\pi(\boldsymbol{\theta})e=1;\quad Q(\boldsymbol{\theta})e=0;\quad \pi(\boldsymbol{\theta})Q(\boldsymbol{\theta})=0 \tag{2}$$

Here, $e=(1,1,\cdots,1)^{\tau}$ is the all-one vector. The average-cost performance measure of $X(\boldsymbol{\theta})$ is

$$\eta(\boldsymbol{\theta})=\lim_{T\to\infty}E\left\{\frac{1}{T}\int_0^T f(X_t,\boldsymbol{\theta})\mathrm{d}t\right\}=\pi(\boldsymbol{\theta})f(\boldsymbol{\theta})$$

For any $\boldsymbol{\theta}\in\Theta$, define the Poisson equation of $X(\boldsymbol{\theta})$ as follows[6]

$$(-Q(\boldsymbol{\theta})+\lambda e\pi(\boldsymbol{\theta}))g(\boldsymbol{\theta})=f(\boldsymbol{\theta})$$

Its unique solution vector is $g(\boldsymbol{\theta})=(-Q(\boldsymbol{\theta})+\lambda e\pi(\boldsymbol{\theta}))^{-1}f(\boldsymbol{\theta})$. We call $g(\boldsymbol{\theta})+ec$ an average-cost performance potential vector for any fixed constant c. The potential of state i can be given by[1]

$$g(i,\boldsymbol{\theta})=\lim_{T\to\infty}\left\{E\left[\int_0^T f(X_t,\boldsymbol{\theta})\mathrm{d}t\,\middle|\,X_0=i\right]-T\eta(\boldsymbol{\theta})\right\} \tag{3}$$

The realization factors of $X(\boldsymbol{\theta})$ are defined as $d_{ij}(\boldsymbol{\theta})=g(j,\boldsymbol{\theta})-g(i,\boldsymbol{\theta})$, $i,j\in\Phi$. For any recurrent state i^* and constant c, let $g(i^*,\boldsymbol{\theta})\equiv c$ and define $T_{ji^*}(\boldsymbol{\theta})=\inf\{t:t>0,X_t=i^*,X_0=j\}$. Then, by equation (3), we have

$$g(j,\boldsymbol{\theta})=d_{i^*j}(\boldsymbol{\theta})+c=E\left\{\int_0^{T_{ji^*}(\boldsymbol{\theta})}[f(X_n,\boldsymbol{\theta})-\eta(\boldsymbol{\theta})]\mathrm{d}t\,\middle|\,X_0=j\right\}+c \tag{4}$$

Assumption 1.

a) For any $i\in\Phi, a\in A$, $\mu_a(i,\boldsymbol{\theta})$ is twice differentiable with respect to parameter vector θ, and has bounded first and second derivatives.

b) For any $i\in\Phi, a\in A, \boldsymbol{\theta}\in\Theta$, there exists a bounded function vector $L_a(i,\boldsymbol{\theta})$ such that $\nabla\mu_a(i,\boldsymbol{\theta})=\mu_a(i,\boldsymbol{\theta})L_a(i,\boldsymbol{\theta})$, where ∇ denotes taking gradient with respect to parameter vector θ.

c) Denote Q as the closure of $\{Q(\boldsymbol{\theta})\mid\boldsymbol{\theta}\in\Theta\}$, and let the Markov process corresponding to every element of Q be ergodic.

3 A simulation optimization algorithm based on a single sample path

3.1 The uniform Markov chain

First, we have the following theorem.

Theorem 1. Under Assumption 1, the gradient of average cost $\eta(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is equal to

$$\nabla\eta(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})(\nabla f(\boldsymbol{\theta}) + \nabla Q(\boldsymbol{\theta})g(\boldsymbol{\theta})) \tag{5}$$

This theorem is easy to verify by applying Poisson equation and balance equation (2), so we omit the details. Define

$$P(a) = I + \lambda^{-1}Q(a); \quad P(\boldsymbol{\theta}) = I + \lambda^{-1}Q(\boldsymbol{\theta}) \tag{6}$$

Obviously, $P(a)$ and $P(\boldsymbol{\theta})$ are all stochastic matrices. From equation (1), we have $p_{ij}(\boldsymbol{\theta}) = \sum_{a\in A}\mu_a(i,\boldsymbol{\theta})p_{ij}(a)$. Define a DTMDP $X'(\boldsymbol{\theta}) = \{X_n, \Phi, A, P(\boldsymbol{\theta}), f(\boldsymbol{\theta})\}$ corresponding to $X(\boldsymbol{\theta})$. $X'(\boldsymbol{\theta})$ is called a uniform Markov chain of $X(\boldsymbol{\theta})$, and λ is the uniformization parameter. On a sample path of Markov process $X'(\boldsymbol{\theta})$, the realization factors are defined as [2]

$$d'_{ji}(\boldsymbol{\theta}) = E\left\{\sum_{n=0}^{N_{ij}(\boldsymbol{\theta})-1}[f(X_n,\boldsymbol{\theta}) - \eta(\boldsymbol{\theta})] \,\middle|\, X_0 = i\right\}$$

where $N_{ij}(\boldsymbol{\theta}) = \min\{n: n>0, X_n = j, X_0 = i\}$. If $g(i^*,\boldsymbol{\theta}) \equiv c$, then the potentials of $X'(\boldsymbol{\theta})$ are given by

$$g'(j,\boldsymbol{\theta}) = d'_{i^*j}(\boldsymbol{\theta}) + c = E\left\{\sum_{n=0}^{N_{ji^*}(\boldsymbol{\theta})-1}[f(X_n,\boldsymbol{\theta}) - \eta(\boldsymbol{\theta})] \,\middle|\, X_0 = j\right\} + c, \quad \forall j \in \Phi \tag{7}$$

Theorem 2. For a fixed parameter vector θ, CTMDP $X(\boldsymbol{\theta})$ and the corresponding DTMDP $X'(\boldsymbol{\theta})$ have the same stationary distribution $\pi(\boldsymbol{\theta})$, and have the same average-cost performance measure $\eta(\boldsymbol{\theta})$. If $c=0$ in (4) and (7), we have

$$\nabla\eta(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})(\nabla f(\boldsymbol{\theta}) + \nabla P(\boldsymbol{\theta})g'(\boldsymbol{\theta})) \tag{8}$$

Proof. Let $\pi(\boldsymbol{\theta})$ be the stationary distribution of $X(\boldsymbol{\theta})$. Since $P(\boldsymbol{\theta})$ is a stochastic matrix, it is easy to verify $\pi(\boldsymbol{\theta})P(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})\boldsymbol{e} = 1$ by using (2) and (6), namely $\pi(\boldsymbol{\theta})$ is also the stationary distribution of $X'(\boldsymbol{\theta})$. In addition, the definition of $X'(\boldsymbol{\theta})$ implies that performance function $f(\boldsymbol{\theta})$ is the same for both $X(\boldsymbol{\theta})$ and $X'(\boldsymbol{\theta})$, thereby the average-cost performance measure of $X'(\boldsymbol{\theta})$ is equal to

$$\lim_{N\to\infty} E\left\{\frac{1}{N}\sum_{n=0}^{N-1} f(X_n,\boldsymbol{\theta})\right\} = \pi(\boldsymbol{\theta})f(\boldsymbol{\theta}) = \eta(\boldsymbol{\theta})$$

By Theorem 2 in [7], we obtain $d_{ij} = d'_{ij}/\lambda$. Therefore, when $c=0$, $g(\boldsymbol{\theta}) = g'(\boldsymbol{\theta})/\lambda$. Furthermore, (6) yields $\nabla Q(\boldsymbol{\theta}) = \lambda \cdot \nabla P(\boldsymbol{\theta})$. Combining these two formulas with (5), we can obtain (8), and the theorem is proved. □

3.2 A simulation-optimization algorithm

Theorem 2 implies that performance optimization problem of CTMDP $X(\boldsymbol{\theta})$ can be solved equivalently through its uniformized Markov chain $X'(\boldsymbol{\theta})$. So the results of [1] may be referenced. By Assumption 1, (8) can be rewritten as[1]

$$\nabla\eta(\boldsymbol{\theta}) = \sum_{i\in\Phi} E[\chi_i(X_n)\nabla f(i,\boldsymbol{\theta})] + \sum_{i,j\in\Phi}\sum_{a\in A} E[\chi_i(X_n)\chi_j(X_{n+1})\chi_a(a_n)L_a(i,\boldsymbol{\theta})g(j,\boldsymbol{\theta})] \tag{9}$$

where a_n is an action selected at state X_n with probability $q_{a_n}(X_n,\boldsymbol{\theta})$, $\chi_i(\cdot)$, and $\chi_a(\cdot)$ are denoting-functions of i and a, respectively. Given a parameter vector θ, we obtain a sample path $(X_{u_0}, X_{u_0+1}, \cdots, X_{u_1}, X_{u_1+1}, \cdots)$ by simulating $X'(\boldsymbol{\theta})$ according to the transition matrix $P(\boldsymbol{\theta})$. Here, $u_0 = 0$, $X_{u_0} = i^*$, and $u_{k+1} = \min\{n: n>u_k, X_n = i^*\}$. Furthermore, let

$$\tilde{g}(X_n,\boldsymbol{\theta}) = \begin{cases} 0, & \text{if } n = u_{k-1} \\ \sum_{t=n}^{u_k-1}\alpha^{(t-n)}(f(X_t,\boldsymbol{\theta}) - \tilde{\eta}), & \text{otherwise} \end{cases}$$

be an estimate of potential $g'(X_n,\boldsymbol{\theta})$, $u_{k-1} \leqslant n < u_k$. Here, $\alpha>0$ is a forgetting factor, and $\tilde{\eta}$ is an estimate of $\eta(\boldsymbol{\theta})$ obtained at time n. Then, according to (9), we define the estimate

of gradient direction as $F^{(k)}(\boldsymbol{\theta},\tilde{\eta})=\sum_{n=u_{k-1}}^{u_k-1}\sum_{a\in A}\mu_a(X_n,\boldsymbol{\theta})(f(X_n,a)-\tilde{\eta})z_n(\boldsymbol{\theta})$, where

$$z_n(\boldsymbol{\theta})=\begin{cases}L_{a_n}(X_n,\boldsymbol{\theta}), & \text{if } X_n=i^*\\ \alpha z_{n-1}(\boldsymbol{\theta})+L_{a_n}(X_n,\boldsymbol{\theta}), & \text{otherwise}\end{cases}$$

Then, for a fixed integer T, our iteration algorithm takes the following form

$$\begin{cases}\theta_{k+1}=\theta_k-\gamma_k\sum_{n=kT}^{(k+1)T-1}(f(X_n,a_n)-\tilde{\eta}_k)z_n(\theta_k)\\ \tilde{\eta}_{k+1}=\tilde{\eta}_k+\beta\gamma_k\sum_{n=kT}^{(k+1)T-1}(f(X_n,a_n)-\tilde{\eta}_k)\end{cases}\tag{10}$$

Here, $\{\gamma_k\}$ is a stepsize sequence, and β is a positive scalar coefficient. Similar to [4], we introduce the following assumptions to ensure convergence property of the above algorithm.

Assumption 2. The stepsize sequence $\{\gamma_k\}$ is nonnegative and nonincreasing. Furthermore, there exist a positive integer p and a positive scalar B such that

$$\sum_{k=0}^{\infty}\gamma_k=\infty,\ \sum_{k=0}^{\infty}\gamma_k^2<\infty;\quad \sum_{k=n}^{n+t}(\gamma_n-\gamma_k)\leqslant Bt^p\gamma_n^2,\quad \forall\, n,t>0$$

Assumption 3. There exists an integer N_0 such that, for every state i and every N_0 matrices $Q_l(\boldsymbol{\theta}),l=1,2,\cdots,N_0$ in Q, the corresponding matrices $P_l(\boldsymbol{\theta}),l=1,2,\cdots,N_0$ defined by (6) satisfy $\sum_{n=1}^{N_0}[\Pi_{l=1}^{n}P_l^{\mathrm{T}}]_{ii^*}>0$.

We then have the following theorem. The proof is similar to [5], so we omit the details.

Theorem 3. For any fixed positive integer T, let Assumptions 1, 2 and 3 hold, denote $\{\theta_k\}$ as the parameter vector sequence generated by the above described algorithm. Then, $\eta(\theta_k)$ converges with probability one, and $\nabla\eta(\theta_k)$ converges to zero with probability one.

4 A numerical example

Consider a CTMDP with three-state space $\Phi=\{1,2,3\}$ and two-action space $A=\{1,2\}$. Under action a, the state transition is illustrated in Fig. 1, the infinitesimal generators and the cost rates are given respectively by $Q(a=1)=[-2,1.6,0.4;1.6,-2,0.4;0,1.6,-1.6]$, $Q(a=2)=[-2,0.4,1.6;0.4,-2,1.6;0,0.4,-0.4]$, and $f(a=1)=f(a=2)=(0;0;1)$. The optimal policy is to choose action 1 at any state, and the associated stationary distribution is $\pi=(16/45,4/9,1/5)$. Thus the optimal average cost is $\eta=\pi\cdot f(1)=0.2$. Let $\lambda=2$. Denote the coding of a state as $x_i,i\in\Phi$, taking the value $(0;1)$, $(1;0)$ and $(1;1)$ respectively. Define the character vector function of a state to be the form

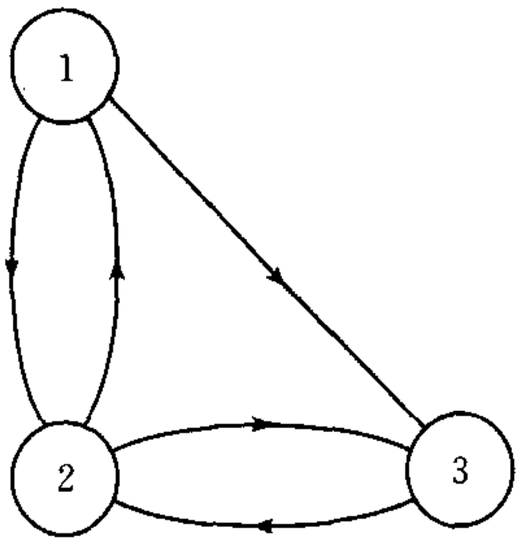

Fig. 1 State transition diagram

$$\phi(x_i)=[\phi_1(x_i),\phi_2(x_i)]=$$
$$[(-7/18,-1/18)\cdot x_i+13/18,(-1/18,-7/18)\cdot x_i+13/18]$$

Let $r_1(x_i)=\phi(x_i)\cdot\theta_1$; $r_2(x_i)=\phi(x_i)\cdot\theta_2$, where θ_1 and θ_2 are parameter vectors. Then the probabilities of taking actions 1 and 2 at state i are of the following forms respectively

$$\mu_{a=1}(i)=\frac{\exp(r_1(x_i))}{\exp(r_1(x_i))+\exp(r_2(x_i))};$$

$$\mu_{a=2}(i)=\frac{\exp(r_2(x_i))}{\exp(r_1(x_i))+\exp(r_2(x_i))}=1-\mu_{a=1}(i)$$

Select the initial parameter vector to be zero, i. e., choose actions 1 and 2 with equal probability at the beginning. Let $T=3$, $\alpha=0.99$, $\beta=0.2$. The simulation results corresponding to 2000 state transitions are given in Fig. 2. The upper plot denotes the value of $\tilde{\eta}$, and the three plots underside represent respectively the probabilities of taking action 1 at states 1, 2, and 3. We see that, after initial updating steps, the resulting policies all take action 1 with large probability at any state, approaching the optimal values.

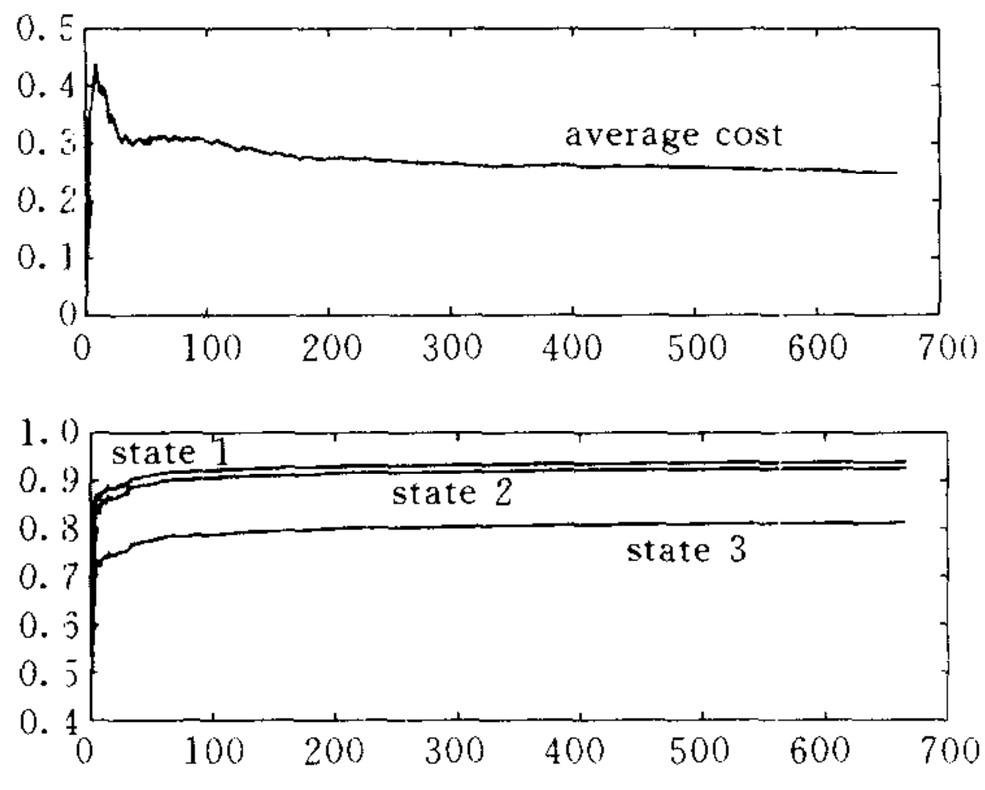

Fig. 2 Curve of simulation results

5 Conclusions

By applying Markov performance potentials and NDP methodology, the optimization of CTMDPs with curse of dimensionality is studied. Although we focus on the processes with finite state space and finite action space, the results can be extended to other cases such as general state space and action space.

References

1 Cao X R, Chen H F. Perturbation realization, potentials and sensitivity analysis of Markov processes. *IEEE Transactions on Automatic Control*, 1997, **42**(10): 1382~1393

2 Cao X R. Single sample path-based optimization of Markov chains. *Journal of Optimization Theory and Applications*, 1999, **100**(3): 527~548

3 Bertsekas D P, Tsitsiklis J N. Neuro-Dynamic Programming. Belmont, MA: Athena Scientific, 1996

4 Marbach P, Tsitsiklis J N. Simulation-based optimization of Markov reward process. *IEEE Transactions on Automatic Control*, 2001, **46**(2): 191~209

5 Tang H, Xi H S, Yin B Q. An on-line optimization algorithm for Markov control processes based on a single sample path. *Control Theory and Applications*, 2002, **19**(6): 865~871(in Chinese)

6 Xi H S, Tang H, Yin B Q. Optimal policies for a continuous time MCP with compact action set. *Acta Automatica Sinica*, 2003, **29**(2): 206~211

7 Liu Z K, Tu F S. Single sample path-based sensitivity analysis of Markov processes. *IEEE Transactions on Automatic Control*, 1999, **44**(4): 872~875

TANG Hao Received his master degree from Institute of Plasma Physics, Chinese Academy of Sciences, in 1998, and Ph. D. degree from University of Science and Technology of China (USTC) in 2002. He is currently an associate professor of Hefei University of Technology. His research interests include DEDS, the methodology of neuro-dynamic programming, and system optimization.

XI Hong-Sheng Received his bachelor and master degrees from USTC in 1977 and 1985, respectively. He is currently a professor of USTC and a vice director of the department of automation. His research interests include DEDS and hybrid dynamic systems.

YIN Bao-Qun Received his bachelor degree from Sichuan University in 1985, the master and Ph. D. degree from University of Science and Technology of P. R. China in 1993 and 1998, respectively. He is currently an associate professor of USTC. His research interests include discrete event dynamic systems, stochastic processes, system optimization, and hybrid dynamic systems.

CTMDP 基于随机平稳策略的仿真优化算法

唐 昊[1,2] 奚宏生[1] 殷保群[1]

[1](中国科学技术大学自动化系 合肥 230026)

[2](合肥工业大学计算机系 合肥 230009)

(E-mail: tangh@ustc.edu; xihs@ustc.edu.cn)

摘 要 基于 Markov 性能势理论和神经元动态规划(NDP)方法，研究一类连续时间 Markov 决策过程(MDP)在随机平稳策略下的仿真优化问题，给出的算法是把一个连续时间过程转换成其一致化 Markov 链，然后通过其单个样本轨道来估计平均代价性能指标关于策略参数的梯度，以寻找次优策略，该方法适合于解决大状态空间系统的性能优化问题。并给出了一个受控 Markov 过程的数值实例.

关键词 性能势，神经元动态规划，仿真优化

中图分类号 TP202