文章编号：1000－8152(2006)02－0292－05

平均和折扣准则MDP基于TD(0)学习的统一NDP方法

唐 昊，周 雷，袁继彬
（合肥工业大学 计算机与信息学院，安徽 合肥 230009）

摘要：为适应实际大规模Markov系统的需要，讨论Markov决策过程(MDP)基于仿真的学习优化问题．根据定义式，建立性能势在平均和折扣性能准则下统一的即时差分公式，并利用一个神经元网络来表示性能势的估计值，导出参数TD(0)学习公式和算法，进行逼近策略评估；然后，根据性能势的逼近值，通过逼近策略迭代来实现两种准则下统一的神经元动态规划(neuro-dynamic programming, NDP)优化方法．研究结果适用于半Markov决策过程，并通过一个数值例子，说明了文中的神经元策略迭代算法对两种准则都适用，验证了平均问题是折扣问题当折扣因子趋近于零时的极限情况．

关键词：Markov决策过程；性能势；TD(0)学习；神经元动态规划
中图分类号：TP202 **文献标识码**：A

Unified NDP method based on TD(0) learning for both average and discounted Markov decision processes

TANG Hao, ZHOU Lei, YUAN Ji-bin
(School of Computer and Information, Hefei University of Technology, Hefei Anhui 230009, China)

Abstract: Motivated by the need of practical large-scale Markov systems, we considered in this paper the learning optimization problems for Markov decision processes (MDPs). Based on the definition of performance potentials, a unified formula of temporal difference is provided for both average and discounted performance criteria. A neural network is then used to represent the estimation of potentials, both the parameterized TD(0) learning formulas and algorithm are also derived for approximating the policy evaluation. By the approximation values of potentials and approximation policy iteration, a unified neuro-dynamic programming (NDP) optimization approach is consequently proposed for both two criteria. The obtained results can be extended to semi-Markov decision processes, and a numerical example is finally used to illustrate the application of the proposed neuro-policy iteration algorithm for both average and discounted criteria. The example also shows that the average problem is the limitation case of the discount ones as discount factor goes to zero.

Key words: Markov decision processes; performance potentials; TD(0) learning; neuro-dynamic programming

1 引言(Introduction)

现实世界中的一类随机序贯决策问题可以用马尔可夫决策过程(MDP)来描述，Markov性能势理论为其性能分析和优化提供了有利途径和理论框架[1,2]．一方面，性能势可以看作是Poisson方程的解，本质上同差分代价、相对代价或bias的概念相同[2~4]，因此在系统模型参数已知时，理论上MDP可以用基于性能势的梯度方法或策略迭代方法来数值求解[1,2,5]；另一方面，性能势又是定义在随机过程样本轨道上的，能够通过仿真或观测实际系统的一条样本轨道来估计，因而可以发展MDP基于性能势仿真的一类优化方法[1,6,7]．无论是数值计算方法还是仿真方法，优化过程中都需要建立状态与性能势或行动的一一对应关系，即计算机保存的是一种数据表格，故有时称为查表法．对大规模系统，一般状态空间很大，利用计算机来求解性能势或用查表法来表示性能势，受计算时间或存储空间的限制，有时不可行．建立在强化学习(reinforcement learning, RL)基础上的神经元动态规划(NDP)是解决大规模离散事件动态系统优化问题的一种有效

收稿日期：2004－09－25；收修改稿日期：2005－07－14.
基金项目：国家自然科学基金资助项目(60404009)；安徽省自然科学基金资助项目(050420303)；合肥工业大学中青年科技创新群体计划资助项目．

方法[4, 8].

性能势理论同 RL 或 NDP 的结合研究是一项具有一定创新意义的工作. 文献[9]研究了 Critic 模式下,半 Markov 决策过程(SMDP)基于 Monte-Carlo 仿真和性能势逼近的优化问题;文献[10]讨论了 Actor 模式下,MDP 基于其一致链的单样本轨道仿真和随机策略逼近的一种 NDP 优化方法. 理论上,平均代价问题是折扣代价问题的一个特例,因此可以预见平均准则下基于性能势的 NDP 算法也是折扣准则下的一个特殊情况. 文中,根据性能势在随机过程样本轨道上的定义,首先针对离散时间 MDP 给出了两种准则下统一的性能势 TD(0)学习公式和参数学习公式,并给出了基于性能势参数 TD(0)学习的一种神经元策略迭代算法,包括 Optimistic TD(0)和 Partially Optimistic TD(0)两种情况. 通过一致链的定义[9],研究结果可以推广运用到连续时间 MDP 或 SMDP 中. 文中,将利用 SMDP 的一个数值例子验证了有关重要结果的正确性和有效性.

2 问题描述(problem formulation)

5-元组 $X=(X_n, \Phi, D, P^v, f^v)$ 表示离散时间 Markov 决策过程,具有有限状态空间 $\Phi=\{1, 2, \cdots, M\}$ 和紧致行动集 D. $X_n(n\geqslant 0)$ 是状态过程,$v=(v(1), \cdots, v(M))$ 表示平稳策略,且行动 $v(i)\in D$,记 Ω_s 是全体平稳策略集. $P^v=[p_{ij}(v(i))]$ 是 X_n 在策略 v 驱动下的状态转移概率矩阵,$f^v=(f(1, v(1)), \cdots, f(M, v(M)))^{\mathrm{T}}$ 为策略 v 下的性能向量. 假设随机过程是不可约的,其稳态概率为 $\pi^v=(\pi^v(1), \cdots, \pi^v(M))$. 无穷时段平均性能准则为

$$\eta^v=\frac{1}{N}\mathrm{E}\Big[\sum_{n=0}^{N-1}f((X_n, v(V_n))\Big]. \tag{1}$$

折扣性能准则为[7]

$$\eta_\beta^v(i)=(1-\beta)\mathrm{E}\Big[\sum_{n=0}^{\infty}\beta^n f(X_n), v(X_n))\mid X_0=i\Big]. \tag{2}$$

这里 $0<\beta<1$ 是折扣因子. 系统优化目标是在 Ω_s 上寻找一个最优策略 v^* 满足 $v^*\in\arg\min\limits_{v\in\Omega_s}\eta^v$ 或 $v^*\in\arg\min\limits_{v\in\Omega_s}\eta_\beta^v$.

对任意 $0<\beta\leqslant 1$,$(I-\beta P^v+\beta e\pi^v)g_\beta^v=f^v$ 称为基于性能势向量的 Poisson 方程,其中 $e=(1, 1, \cdots, 1)^{\mathrm{T}}$ 是 M 维列向量,且 $g_\beta^v=(g_\beta^v(1), \cdots, g_\beta^v(M))^{\mathrm{T}}$. $\beta<1$ 时为折扣情况,g_β^v 为折扣代价性能势;$\beta=1$ 时为平均情况,g_1^v 为平均代价性能势向量,记 $g_1^v=g^v$. 有

$$g_\beta^v=(I-\beta P^v+\beta e\pi^v)^{-1}f^v. \tag{3}$$

对任意 $0<\beta\leqslant 1$,可证一个策略是最优的充要条件是下列基于性能势的 Bellman 最优性方程成立:

$$\min_{v\in\Omega_s}\{f^v+\beta(P^v-I)g_\beta^{v^*}-\eta_\beta^{v^*}\}=0.$$

与连续时间 MDP 类似,这个最优性方程可以通过基于性能势数值计算的策略迭代求解[2, 5]. 但是,求解 Poisson 方程涉及到矩阵求逆,对大状态空间问题,这将耗费较多计算时间,需要大量计算机内存,因而实际中有时需要考虑仿真方法.

3 TD 学习和 NDP(TD learning and NDP)

文献[1]中已提供了多种性能势估计的算法,在此基础上,产生了诸多仿真优化方法,如基于性能势估计的梯度方法和策略迭代方法等. 强化学习是解决 MDP 的一种重要仿真方法,自然地能与性能势理论结合起来,并导致一系列基于性能势学习的有效优化技术. 下面,将讨论基于性能势参数学习的一种 NDP 方法.

3.1 性能势的 TD 学习(TD learning of performance potential)

文献[4]中,折扣准则下的即时差分(temporary difference,TD)定义为

$$d_n=f(X_n)+\beta\eta_\beta(X_{n+1})-\eta_\beta(X_n). \tag{4}$$

其中 $\beta<1$. 理论上,有 $\lim\limits_{\beta\to 1}\eta_\beta^v=e\eta^v$,平均和折扣准则下的优化问题能够统一考虑,且 η_β^v,$\beta<1$ 也可看作是折扣代价性能势的一种定义. 但是,当 $\beta=1$ 时,按照式(4)构造的 TD 学习算法无法保证其收敛性. 故平均和折扣准则下的即时差分无法统一,TD 学习算法亦无法统一. 因为 $\lim\limits_{\beta\to 1}g_\beta^v=g^v$,两种准则下基于性能势数值求解的优化算法具有统一框架,预示基于性能势学习的优化算法也能统一起来.

在策略 v 下,对任意 $0<\beta\leqslant 1$,性能势也可按下式定义[7]:

$$g_\beta^v(i)=\mathrm{E}\Big[\sum_{n=0}^{\infty}\beta^n[f[X_n, v(X_n)]-\beta\eta^v]\mid X_0=i\Big]=$$

$$f(i, v(i))-\beta\eta^v+\beta\sum_{j=1}^{M}p_{ij}^v g_\beta^v(j).$$

其定义同式(3)完全一致. 于是,可定义即时差分 $d_n=f(X_n, v(X_n))-\beta\eta^v+\beta g_\beta^v(X_{n+1})-g_\beta^v(X_n)$,显然 $\mathrm{E}[d_n]=0$. 当 $\beta=1$,上式变为

$$d_n=f(X_n, v(X_n))-\eta^v+g^v(X_{n+1})-g^v(X_n).$$

这同文献[7]中给出的公式一样. 这样,折扣和平均准则下的性能势 TD 学习公式可以统一起来,有

以下的 Robbins-Monro 随机逼近算法

$$g_\beta^v(X_n) := g_\beta^v(X_n) + \gamma d_n .$$

根据随机逼近原理,类似于文献[4],不难证明其收敛性,此处不再赘述.

3.2 性能势的参数 TD(0)学习(parameterized TD (0) learning of performance potential)

为减少优化过程中性能势的查表表示法所需的大量计算机内存,根据 NDP 方法中的 Critic 模式,可利用神经元网络等逼近结构来近似表示性能势,并以性能势的近似即时差分来更新网络结构的参数,即权系数,以改进逼近质量,这就是性能势的参数 TD 学习. 这里,逼近结构可以是神经元网络,也可以是一些特征函数的线性组合或者是多项式. 通常用神经元网络来泛指一般逼近结构,其参数向量 r 就泛称为权系数向量,且参数数目比系统的状态数要少,从而起到节省内存的目的. 参数表示法同查表表示法的区别是,查表表示法在计算机内存里保存的是状态空间中状态 i 同性能势 $g_\beta^v(i)$ 或其估计学习值的一一对应关系,而参数表示法保存的是一个逼近结构,网络结构的输入是状态 i,输出是 $g_\beta^v(i, r)$. 记 $g_\beta^v(r) = (g_\beta^v(1, r), \cdots, g_\beta^v(M, r))^{\mathrm{T}}$. 在策略 v 作用下的一条样本轨道上,对任意 $0 < \beta \leqslant 1$,近似即时差分定义为

$$d_n = d(X_n, X_{n+1}, r_n) = f(X_n, v(X_n)) - \beta\eta_n + \beta g_\beta^v(X_{n+1}, r_n) - g_\beta^v(X_n, r_n) . \tag{5}$$

这里

$$\eta_n = (1 - \delta_n)\eta_{n-1} + \delta_n f(X_n, v(X_n)) \tag{6}$$

为平均代价 η^v 的学习值,δ_n 为学习步长. 于是参数化 TD(0)学习公式为

$$r_{n+1} = r_n + \gamma_n d_n \nabla g_\beta^v(X_n, r_n) . \tag{7}$$

这里 γ_n 也为步长. 步长参数可以为常数,也可以为时间变量.

3.3 基于 TD(0)学习的 NDP 优化(NDP optimization based on TD(0) learning)

根据性能势的参数学习值,可执行逼近策略迭代,对策略进行改进. 逼近策略迭代有两种情况[4],即 Optimistic TD(0)和 Partially Optimistic TD(0). Optimistic TD(0)就是每仿真一步,性能势学习一步,随即进行策略更新;Partially Optimistic TD(0)就是在每次迭代中,仿真多步,学习多步,然后才进行策略更新,记每次迭代的学习步数为 L. 若 $L = 1$,则 Partially Optimistic TD(0)就变为 Optimistic TD(0),即后者为前者的特殊形式. 在给定的策略 v 下,Partially Optimistic TD(0)的性能势参数学习过程如下:

算法 1

Step 1 令 $n = 0$,初始化 L, X_n, r_n, η_{n-1};

Step 2 在状态 X_n 根据 P^v 仿真,随机产生下一状态 X_{n+1};或观测实际系统的运行得到 X_{n+1};

Step 3 选择步长 δ_n 和 γ_n,根据公式(5)~(7),更新神经元网络的权值 r_n,进行性能势的参数学习;

Step 4 令 $L := L - 1$. 若 $L = 0$,退出. 否则令 $n := n + 1$,转 Step 2.

上述算法其实是对策略 v 的性能进行评估,称为逼近策略评估(approximate policy evaluation). 经过 L 步的仿真学习后,神经元网络结构保存的就是性能势的学习值 $g_\beta^v(r)$. 实际中,也可以不固定学习步数,采用随机的学习步数,例如给定一个较小的正常数 ε,判断条件 $\|g_\beta^v(r_{n+1}) - g_\beta^v(r_n)\| < \varepsilon$ 或 $\|\eta_n - \eta_{n-1}\| < \varepsilon$ 是否成立. 有了性能势的参数学习值,就可以进行逼近策略迭代(approximate policy iteration),具体过程如下:

算法 2

Step 1 令 $k = 0$,选择一初始策略 v_k;根据优化目标,确定区间(0 1]内的折扣因子 β;

Step 2 根据 v_k 调用算法 1,得到性能势的参数学习值 $g_\beta^{v_k}(r)$;

Step 3 根据下式,计算更新策略 v_{k+1}:

$$v_{k+1} \in \arg\min_{v \in \Omega_a} \{f^v + \beta P^v g_\beta^{v_k}(r)\} .$$

Step 4 判断停止条件是否满足. 若满足,退出;否则,令 $k := k + 1$,转 Step 2.

上述优化算法称为神经元策略迭代,若取 $\beta = 1$,则为平均情况,若取 $\beta < 1$,则为折扣情况. 其停止条件有多种选择,可以选择固定迭代步数. 也可选择 $\|v_{k+1} - v_k\| < \varepsilon$ 或 $\|\eta_\beta^{v_{k+1}} - \eta_\beta^{v_k}\| < \varepsilon$.

注意的是,对不同的停止准则,需选择合适参数 ε,以保证算法实际运行时能退出循环,正常停止. 一般,可把多种停止条件结合起来运用.

3.4 半 Markov 决策过程(Semi-Markov decision process)

5-元组 $\bar{X} = (X_t, \Phi, D, Q^v(t), f^v)$ 表示的 SMDP 中,$Q^v(t) = [Q(i, j, v(i), t)]$ 是半 Markov 核,$f^v = [f(i, j, v(i))]$ 为性能矩阵. 按文献[11]定义平均准则 $\bar{\eta}^v$ 和折扣准则向量 $\bar{\eta}_\alpha^v$. 根据文献[11]或[12],对任意的 $\alpha \geqslant 0$,有等价无穷小矩阵 A_α^v,等价性能向量 f_α^v,且 $\bar{X}$ 等价于一个连续时间

MDP$X^{\alpha}=(X_t^{\alpha}, \Phi, D, A_{\alpha}^{v}, f_{\alpha}^{v})$. 假设折扣因子有界,即设 $\alpha \in (0, a]$,a 为一固定常数. 按文献[9]定义一个一致化参数 λ,且令随机矩阵 $P_{\alpha}^{v}=A_{\alpha}^{v}/\lambda+I$,则 P_{α}^{v} 对应 SMDP$\bar{X}$ 的一个 α-一致 Markov 链 $X=(X_n, \Phi, D, P_{\alpha}^{v}, f_{\alpha}^{v})$.

令 X 的折扣因子 β 满足 $\beta=\lambda/(\lambda+\alpha)$,且按公式(1)和(2)来定义其性能准则. 则对任意 $\alpha \geqslant 0$,SMDP$\bar{X}$ 的 α-一致 Markov 链 X 的稳态分布等于其等价 Markov 过程 X^{α} 的稳态分布,且有 $\eta_{\beta}^{v}=\bar{\eta}_{\alpha}^{v}$. 另外,当折扣因子 $\alpha=0$ 即 $\beta=1$ 时有 $\eta^{v}=\bar{\eta}^{v}$. 于是,在 SMDP 系统模型参数已知的情况下,可根据其一致链来进行性能势的强化学习,并实现 NDP 优化. 因此,前面的有关结果,可以推广到 SMDP.

4　数值例子(Numerical example)

仍然考虑文献[9]中给出的一个 SMDP,令 $\Phi=\{1, 2, \cdots, 31\}$, $D=[0.5, 3.5]$, 在策略 v 下随机过程嵌入链的转移概率满足

$$p_{ij}(v(i))=\begin{cases}\dfrac{\exp(-v(i)/j)}{M(1+\exp(-v(i)))}, & j\neq i+1,\\ 1-\sum\limits_{j\neq i+1}p_{ij}(v(i)), & j=i+1,\end{cases}$$

且 $i=31$, $j=1$ 时,$p_{ij}(v(i))=1-\sum\limits_{j\neq 1}p_{ij}(v(i))$;性能函数 $f(i, v(i))=\ln[(1+i)v(i)]+\sqrt{i}/(2v(i))$;状态 i 的逗留时间服从三阶 Erlang 分布,即 $F_{ij}(t, v(i))=F_i(t, v(i))=1-e_1\exp(T^{v(i)}t)e$,且 $T^{v(i)}=3v(i)[-1, 1, 0; 0, -1, 1; 0, 0, -1]$, $e=[1, 1, 1]^{\mathrm{T}}$, $e_1=[1, 0, 0]$.

设计一个拓扑结构为 5×3×1 的 BP 神经元网络,逼近性能势的学习值,显然网络结构的参数个数比状态数少. 优化实现时,把 SMDP 转换成其一致链,通过仿真该链的一条样本轨道来进行性能势的参数学习和逼近策略迭代. 利用文中提供的统一的神经元策略迭代算法,可以求解不同折扣因子和不同学习步数下的最优策略. $\beta=1$,即 $\alpha=0$ 时为平均准则情况,对应的优化结果见表 1 所示.

表 1 中,η^{v_s} 表示神经元策略迭代算法的停止策略 v_s 对应的平均代价,分别对应各自一条样本轨道上的一次优化结果;平均代价最优理论值可以通过基于性能势的梯度方法或策略迭代来求解. 可以看到,不同学习步数下的仿真优化结果接近最优理论解,是次优的(suboptimal). 仿真中注意到,基于 TD(0)学习的神经元策略迭优化统计意义上比文献[9]中的基于 Monte-Carlo 仿真的神经元策略迭代的优化结果要好,速度也相对较快;另外,学习步数为 1 时的优化精度在统计意义上要略差于其它学习步数,这是由于性能势学习不充分造成的;并且,对固定的学习步数和折扣因子,不同次仿真的优化速度存在一定差异,其优化时间存在波动. 这是随机过程仿真的固有特性,另一方面也与网络的初始权参数和阈值是随机产生的因素有关.

表 1　不同学习步数下平均代价的优化结果

Table 1　Results of average cost for different learning steps

学习步数 L	1	1000	2000	10000
平均代价 η^{v_s}	3.86937	3.72508	3.72754	3.73175
理论结果		3.71758		

图 1 和图 2 分别是学习步数为 1000 和 10000 时,平均代价的优化曲线,算法 2 分别在迭代次数 k 为 40 多次和 20 多次时停止. 统计意义上优化仿真的迭代次数随着学习步数增加而减少.

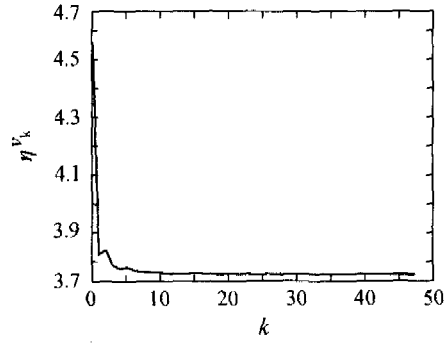

图 1　$N=1000$ 时平均代价 η^{v_k} 的优化曲线

Fig. 1　Plot of average cost η^{v_k} as $N=1000$

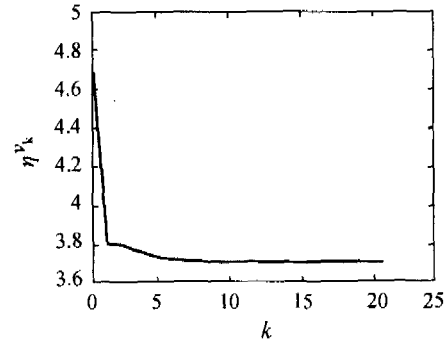

图 2　$N=10000$ 时平均代价 η^{v_k} 的优化曲线

Fig. 2　Plot of average cost η^{v_k} as $N=10000$

图 3 是折扣因子 $\alpha=0.01$ 和学习步数 $N=1000$ 时,5 个折扣代价分量 $\eta_{\alpha}^{v_s}(i)$ $(i=1, 3, 15, 30, 31)$ 的变化曲线,可见各个分量随着算法 2 中迭代次数 k 的增加而逐渐趋向平稳. 图 4 是 $N=2000$ 时,5 个折扣代价分量 $\eta_{\alpha}^{v_s}(i)$ 随 α 变化的曲线. 可以看到,当折扣因子 α 趋向零时,各个分量都趋向于同一个数

值,即平均准则下基于性能势学习的神经元策略迭代是折扣准则情况的一个特例.

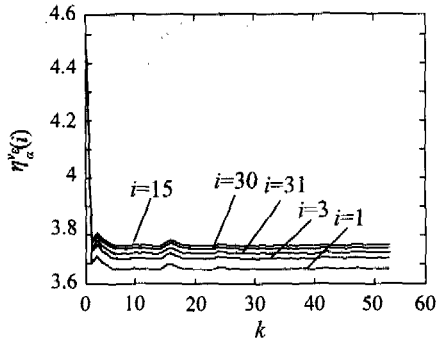

图3 $\alpha = 0.01$ 和 $N = 1000$ 时折扣代价 $\eta_\alpha^{v_k}(i)$ 的优化曲线

Fig. 3 Plot of discounted cost $\eta_\alpha^{v_k}(i)$ as $N = 1000$ and $\alpha = 0.01$

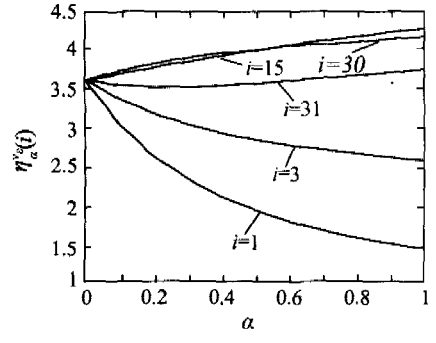

图4 $N = 2000$ 时 $\eta_\alpha^{v_\alpha}(i)$ 随折扣因子 α 的变化曲线

Fig. 4 Plot of discounted cost $\eta_\alpha^{v_\alpha}(i)$ as α changes with N=2000

5 结论(Conclusion)

平均准则和折扣准则下的性能势TD(0)学习和参数TD(0)学习具有统一的表达式,因而建立在强化学习基础上的NDP方法可以统一起来,具有统一的实现算法. 通过选择不同的折扣因子,给出的神经元策略迭代算法能产生对应不同折扣准则或平均准则的优化结果. 文中讨论的主要学习或优化算法可以推广到模型参数已知的SMDP中. 按同样的思路,很容易建立统一的参数TD(λ)学习公式和相应的NDP算法. 但是,针对模型参数未知的情况,需要考虑性能势理论下,基于统一Q因子学习的NDP方法.

参考文献(References):

[1] CAO X R, CHEN H F. Perturbation realization, potentials and sensitivity analysis of Markov processes [J]. *IEEE Trans on Automatic Control*, 1997, 42(10): 1382 - 1393.

[2] CAO X R. The relations among potentials, perturbation analysis, and Markov decision processes [J]. *Discrete Event Dynamic Systems: Theory and Applications*, 1998, 8(1): 71 - 78.

[3] PUTERMAN M L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming* [M]. New York: Wiley, 1994.

[4] BERTSAKEAS D P, TSITSIKLIS J N. *Neuro-Dynamic Programming* [M]. Belmont, MA: Athena Scientific, 1996.

[5] TANG H, Xi H S, YIN B Q. Performance optimization of continuous-time Markov control processes based on performance potentials [J]. *Int J of Systems Science*, 2003, 34(1): 63 - 71.

[6] CAO X R. Single sample path-based optimization of Markov chains [J]. *J of Optimization Theory and Applications*, 1999, 100(3): 527 - 548.

[7] CAO X R. From Perturbation analysis to Markov decision processes and reinforcement learning [J]. *Discrete Event Dynamic Systems: Theory and Applications*, 2003, 13(1/2): 9 - 39.

[8] SUTTON R S, BARTO A G. *Reinforcement Learning: An Introduction* [M]. Cambridge, MA: MIT Press, 1998.

[9] TANG H, YUAN J B, LU Y, et al. Performance potential-based neuro-dynamic programming for SMDPs [J]. *Acta Automatic Sinica*, 2005, 31(4): 642 - 645.

[10] 唐昊, 奚宏生, 殷保群. Markov控制过程基于单个样本轨道的在线优化算法[J]. 控制理论与应用, 2002, 19(6): 863 - 871.
(TANG Hao, XI Hongsheng, YIN Baoqun. On-line optimization algorithm for Markov control processes based on a single sample path [J]. *Control Theory & Application*, 2002, 19(6): 863 - 871.)

[11] CAO X R. Semi-Markov decision problems and performance sensitivity analysis [J]. *IEEE Trans on Automatic Control*, 2003, 48(5):758 - 769.

[12] 殷保群,奚宏生,周亚平. 排队系统性能分析与Markov控制过程 [M]. 合肥:中国科学技术大学出版社,2004.
(YIN Baoqun, XI Hongsheng, ZHOU Yaping. *Queueing System Performance Analysis and Markov Control Processes* [M]. Hefei: Press of University of Science and Technology of China, 2004.)

作者简介:

唐 昊 (1972—),男,副教授,1998年获中国科学院等离子体物理研究所核能科学与工程专业硕士学位,2002年获中国科学技术大学模式识别与智能系统专业博士学位,感兴趣的研究领域包括离散事件动态系统(DEDS)、强化学习(RL)和神经元动态规划(NDP)以及智能优化方法等,现主持1项国家自然科学基金项目以及1项安徽省自然科学基金项目, E-mail: htang@hfut.edu.cn;

周 雷 (1981—),男,硕士研究生,2003年毕业于中南大学,获工学学士学位,感兴趣方向为离散事件动态系统、强化学习以及智能优化方法, E-mail: zhoulei@ialab.hfut.edu.cn;

袁继彬 (1972—),男,硕士研究生,主要研究方向为计算机控制技术、人工智能等, E-mail: yuanjibin@ah163.com.