第 19 卷 第 6 期 控 制 与 决 策 2004 年 6 月
Vol. 19 No. 6 *Control and Decision* June 2004

文章编号: 1001-0920(2004)06-0691-04

半 Markov 控制过程在折扣代价准则下的最优平稳策略

殷保群, 李衍杰, 周亚平, 奚宏生
(中国科学技术大学 自动化系, 安徽 合肥 230027)

摘 要: 讨论一类半 Markov 控制过程(SMCP)的折扣代价性能优化问题. 通过引入一个矩阵,该矩阵可作为一个 Markov 过程的无穷小矩阵, 对一个 SMCP 定义了折扣 Poisson 方程, 并由这个方程定义了 α-势. 基于 α-势,给出了由最优平稳策略所满足的最优性方程. 最后给出一个求解最优平稳策略的迭代算法,并提供一个数值例子以表明该算法的应用.

关键词: 半 Markov 控制过程; 折扣代价准则; 折扣 Poisson 方程; α-势;最优性方程;最优平稳策略
中图分类号: O232 **文献标识码**: A

Optimal stationary policies for semi-Markov control processes with discounted-cost criteria

YIN Bao-qun, LI Yan-jie, ZHOU Ya-ping, XI Hong-sheng
(Department of Automation, University of Science and Technology of China, Hefei 230027, China. Correspondent: YIN Bao-qun, E-mail: bqyin@ustc.edu.cn)

Abstract: The problems of discounted-cost performance optimization are discussed for a class of semi-Markov control processes (SMCP). A matrix is defined, which can be as the infinitesimal generator of a Markov process. The discounted Poisson equation is proposed for an SMCP by using this matrix, from which the α-potential is defined. Based on the α-potential, the optimality equation satisfied by the optimal stationary policy is given. Finally an iteration algorithm to find an optimal stationary policy is proposed, and an numerical example is provided to illustrate the application of the algorithm.

Key words: semi-Markov control processes; discounted-cost criteria; discounted Poisson equation; α-potential; optimality equation; optimal stationary policy

1 引 言

半 Markov 控制过程(又称半 Markov 决策过程)是一类受一系列控制决策驱动的半 Markov 系统,其状态转移规律和控制决策所采用的行动方案相互作用,决定了系统的演化过程在每个状态的逗留时间是服从一般分布的随机变量. 在平均代价准则下,对半 Markov 控制过程性能优化的研究已有一些结果[1~6];但在折扣代价准则下,相应研究的文献则不多. 文献[5]利用转化成离散时间 Markov 链的方法,在期望折扣总报酬准则下,讨论了一类半 Markov 决策过程,并在一定的条件下给出了最优性方程. 文献[6]研究了一类半 Markov 决策过程,通过引入一个所谓的 M 矩阵,给出了最优性方程以及迭代优化算法.

本文基于折扣 Poisson 方程,研究一类具有有限状态空间的半Markov控制过程,在相对较弱的

收稿日期: 2003-07-17; **修回日期**: 2003-11-19.
基金项目: 国家自然科学基金资助项目(60274012); 安徽省自然科学基金资助项目(01042308).
作者简介: 殷保群(1962—),男,安徽全椒人,副教授,博士,从事随机 DEDS、系统优化等研究; 李衍杰(1978—),男,山东青岛人,博士生,从事 DEDS 等研究.

条件下给出了最优性方程. 首先定义一个矩阵,该矩阵可作为 Markov 过程的无穷小矩阵,通过这个矩阵对 SMCP 引入折扣 Poisson 方程,根据该方程定义 α-势;然后给出最优性定理,并基于 α-势导出由最优平稳策略所满足的最优性方程;最后给出一个求解最优策略的迭代算法,并通过一个数值例子说明该算法的应用.

文献[5]将半 Markov 决策过程(SMDP)转化成离散时间 Markov 决策过程(DTMDP)进行研究,而本文则用势方法直接研究 SMDP 问题,基本上不需要附加的假设条件. 一般而言,由于 DTMDP 是 SMDP 的一种特殊情况,转化成 DTMDP 的方法需要较强的假设条件. 文献[6]通过引入一个所谓的 M 矩阵来描述最优性方程,这个 M 矩阵就是本文中的矩阵 Q_a. 本文通过引入矩阵 A_a 来描述最优性方程,这个矩阵可作为 Markov 过程的无穷小矩阵,因而具有明确的意义. 用该矩阵描述的最优性方程,其形式与 Markov 控制过程完全一致. 如果定义一个等价的 Markov 控制过程,则有关 Markov 控制过程的一些结果(参阅文献[8,9])可直接加以运用,比如求解最优策略的迭代算法等.

2 折扣代价准则

考虑一个半 Markov 过程 $Y=\{Y_t: t\geq 0\}$,具有有限状态空间 $\Phi=\{1,2,\cdots,k\}$ 和行动空间 D, $D(i)\subset D$ 为状态 i 的容许行动集,且 $D(i)$ 非空, $i\in\Phi$. 设 $X=\{X_n, n\geq 0\}$ 是 Y 的嵌入 Markov 链, $0=T_0<T_1<\cdots$ 是相继的状态转移时刻,则 $(X,T)=\{X_n, T_n, n\geq 0\}$ 是具有状态空间 Φ 的 Markov 更新过程. 一个平稳策略是状态空间 Φ 到行动空间 D 的映射 $v:\Phi\to D$,且对 $i\in\Phi$, $v(i)=d_i\in D(i)$. 记 $v=(v(1),\cdots,v(k))$,令 Ω_s 是全体平稳策略集. 在策略 $v\in\Omega_s$ 下, Y 的半 Markov 核为 $Q^v(t)=[Q(i,j,v(i),t)]$. 其中

$$Q(i,j,t)=P\{X_{n+1}=j, T_{n+1}-T_n\leq t\mid X_n=i\} \tag{1}$$

与 n 无关, $i,j\in\Phi$, $t\geq 0$. 假设在任意策略 $v\in\Omega_s$ 下, Y 是不可约的和非周期的. 由于 Φ 有限, Y 也是正常返的. 令

$$h(i,v(i),t)=1-\sum_{j\in\Phi}Q(i,j,v(i),t), \tag{2}$$

$$h^v(t)=[h(1,v(1),t),\cdots,h(k,v(k),t)]^\tau \tag{3}$$

则

$$h^v(t)=(I-Q^v(t))e. \tag{4}$$

其中 $e=(1,1,\cdots,1)^\tau$, τ 表示转置. 令 f 为依赖于 v 的性能函数,记

$$f^v=[f(1,v(1)),f(2,v(2)),\cdots,f(k,v(k))]^\tau.$$

本文称 $(Y,\Phi,D,Q^v(t),f^v)$ 为约束在平稳策略集 Ω_s 上的一个 SMCP. SMCP 关于无限水平折扣性能准则为

$$\eta_\alpha^v(i)=E\left\{\int_0^\infty e^{-\alpha t}f(Y_t,v(Y_t))\,dt\mid Y_0=i\right\},\quad v\in\Omega_s, i\in\Phi. \tag{5}$$

其中 $\alpha>0$ 为折扣因子.

在折扣代价 SMCP 问题中,优化的目标是选择策略 $v^*\in\Omega_s$,使得折扣性能 $\eta_\alpha^v(i)$ 在该策略下对每个 $i\in\Phi$ 达到最小. 一般称此策略为最优平稳策略.

3 α-势

为了简化记号,暂时省略上标 v. 记 $R(t)=[R(i,j,t)]$ 为 Markov 更新过程 (X,T) 的 Markov 更新核[7]. 对于 $\alpha\geq 0$,令

$$Q_\alpha=\int_0^\infty e^{-\alpha t}Q\,dt,\quad h_\alpha=\int_0^\infty e^{-\alpha t}h(t)\,dt, \tag{6}$$

则 $Q_0=P=[P(i,j)]$ 为嵌入 Markov 链 X 的转移矩阵; $h_0=(m(1),m(2),\cdots,m(k))$, $m(i)$ 为过程 Y 在状态 i 的平均逗留时间. 由式(4)有

$$\alpha h_\alpha=(I-Q_\alpha)e. \tag{7}$$

对于 $\alpha>0$,定义

$$R_\alpha=\int_0^\infty e^{-\alpha t}R\,dt, \tag{8}$$

根据文献[7],有

$$R_\alpha=(I-Q_\alpha)^{-1}. \tag{9}$$

定义

$$A_\alpha=\alpha I-H_\alpha^{-1}(I-Q_\alpha), \tag{10}$$

其中 $H_\alpha=\mathrm{diag}\{h_\alpha(1),h_\alpha(2),\cdots,h_\alpha(k)\}$. 记

$$P_\alpha=\alpha H_\alpha+Q_\alpha,\ \Lambda_\alpha=H_\alpha^{-1}. \tag{11}$$

由式(7)可知, P_α 是一个 Markov 矩阵,而 A_α 可表为

$$A_\alpha=\Lambda_\alpha(P_\alpha-I). \tag{12}$$

因此, $A_\alpha=[A_\alpha(i,j)]$ 可作为 Markov 过程的无穷小矩阵.

对于 $\alpha>0$,令

$$U_\alpha=\int_0^\infty e^{-\alpha t}P(t)\,dt. \tag{13}$$

其中: $P_t(i,j)=P\{Y_t=j\mid Y_0=i\}$, $P(t)=[P_t(i,j)]$. 由文献[7]中式(10.5.14)以及本文式(9)和(10),有

$$U_\alpha=(\alpha I-A_\alpha)^{-1}. \tag{14}$$

当 $\alpha>0$ 时,由式(13)和(14)得

$$(\alpha I - A_\alpha^\nu)^{-1} e = e/\alpha. \tag{15}$$

现在加上上标 ν. 易见

$$\eta_\alpha^\nu(i) = \int_0^\infty \mathrm{e}^{-\alpha t} \sum_{j\in\Phi} P_t^\nu(i,j) f(j,\nu(j))\,\mathrm{d}t,$$
$$\nu \in \Omega_s, i \in \Phi. \tag{16}$$

若记 $\eta_\alpha^\nu = (\eta_\alpha^\nu(1), \eta_\alpha^\nu(2), \cdots, \eta_\alpha^\nu(k))^\tau$,则

$$\eta_\alpha^\nu = (\alpha I - A_\alpha^\nu)^{-1} f^\nu, \nu \in \Omega_s. \tag{17}$$

对于任意的 $\nu \in \Omega_s, \alpha \geq 0$,定义折扣 Poisson 方程

$$(\alpha I - A_\alpha^\nu) g_\alpha^\nu = f^\nu - \frac{e p_\alpha^\nu f^\nu}{1+\alpha}. \tag{18}$$

其中 p_α^ν 是方程

$$p_\alpha^\nu A_\alpha^\nu = 0, p_\alpha^\nu e = 1 \tag{19}$$

唯一正解. 当 $\alpha > 0$ 时,由于矩阵 $(\alpha I - A_\alpha^\nu)$ 可逆,此时折扣 Poisson 方程(18) 存在唯一解 g_α^ν. 本文称 g_α^ν 为 α- 势.

当 $\alpha > 0$ 时,由式(15), (17) 和(18) 得

$$\eta_\alpha^\nu = g_\alpha^\nu + \frac{e p_\alpha^\nu f^\nu}{\alpha(1+\alpha)}. \tag{20}$$

4　最优性方程

引理 1　对于任意的 $\nu', \nu \in \Omega_s$,有

$$\eta_\alpha^{\nu'} - \eta_\alpha^\nu = (\alpha I - A_\alpha^\nu)^{-1}[(f^{\nu'} + A_\alpha^{\nu'} g_\alpha^{\nu'}) - (f^\nu + A_\alpha^\nu g_\alpha^\nu)]. \tag{21}$$

证明　由式(17) 有

$$\alpha(\eta_\alpha^{\nu'} - \eta_\alpha^\nu) = f^{\nu'} - f^\nu + A_\alpha^{\nu'}\eta_\alpha^{\nu'} - A_\alpha^\nu\eta_\alpha^\nu = f^{\nu'} - f^\nu + (A_\alpha^{\nu'} - A_\alpha^\nu)\eta_\alpha^{\nu'} + A_\alpha^\nu(\eta_\alpha^{\nu'} - \eta_\alpha^\nu).$$

由此可得

$$\eta_\alpha^{\nu'} - \eta_\alpha^\nu = (\alpha I - A_\alpha^\nu)^{-1}[f^{\nu'} - f^\nu + (A_\alpha^{\nu'} - A_\alpha^\nu)\eta_\alpha^{\nu'}].$$

将式(20) 代入上式右边,并根据 $(A_\alpha^{\nu'} - A_\alpha^\nu)e = 0$,即得式(21). □

定理 1(最优性定理)　$\nu^* \in \Omega_s$ 是半 Markov 控制过程 $(Y, \Phi, D, Q^\nu(t), f^\nu)$ 折扣代价最优平稳策略,其充分必要条件为

$$f^{\nu^*} + A_\alpha^{\nu^*} g_\alpha^{\nu^*} \leq f^\nu + A_\alpha^\nu g_\alpha^{\nu^*}, \nu \in \Omega_s. \tag{22}$$

证明　由式(14) 以及 $P^\nu(t)$ 和 U_α^ν 的定义易知,矩阵 $(\alpha I - A_\alpha^\nu)^{-1}$ 的每一个元素均为正. 故当式(22) 成立时,由引理 1 有 $\eta_\alpha^{\nu^*} \leq \eta_\alpha^\nu, \nu \in \Omega_s$,即 ν^* 是一个最优平稳策略. 反之,如果 ν^* 是一个最优平稳策略,但式(22) 不能成立,则必存在一个 $\nu' \in \Omega_s$,使在某个状态 $i_0 \in \Phi$,有

$$f(i_0, \nu'(i_0)) + \sum_{j\in\Phi} A_\alpha(i_0, j, \nu'(i_0)) g_\alpha^{\nu^*}(j) < f(i_0, \nu^*(i_0)) + \sum_{j\in\Phi} A_\alpha(i_0, j, \nu^*(i_0)) g_\alpha^{\nu^*}(j).$$

现定义一个策略 $\nu'' \in \Omega_s$:除 $\nu''(i_0) = \nu'(i_0)$ 外,其余 $\nu''(i) = \nu^*(i)$. 根据引理 1, 有 $\eta_\alpha^{\nu''}(i_0) < \eta_\alpha^{\nu^*}(i_0)$,从而 ν^* 不是最优平稳策略. 这与假设矛盾,故定理得证. □

定理 2　$\nu^* \in \Omega_s$ 是半 Markov 控制过程 $(Y, \Phi, D, Q^\nu(t), f^\nu)$ 折扣代价最优平稳策略,其充分必要条件为满足方程

$$0 = \inf_{\nu\in\Omega_s}\{f^\nu + A_\alpha^\nu g_\alpha^{\nu^*} - \alpha\eta_\alpha^{\nu^*}\}. \tag{23}$$

证明　根据式(17),有 $\alpha\eta_\alpha^{\nu^*} = f^{\nu^*} + A_\alpha^{\nu^*}\eta_\alpha^{\nu^*}$. 将式(20) 代入上式右边,并注意到 $A_\alpha^{\nu^*} e = 0$,则有

$$\alpha\eta_\alpha^{\nu^*} = f^{\nu^*} + A_\alpha^{\nu^*} g_\alpha^{\nu^*}. \tag{24}$$

根据定理 1,可以直接得到定理 2. □

本文称式(23) 为 SMCP 基于 α- 势的折扣代价最优性方程.

5　算法与算例

根据定理 1 和定理 2,可给出一个求解最优平稳策略的迭代算法. 具体步骤如下:

1) 置 $n = 0$,选择初始策略 ν_0,给定充分小的 $\varepsilon > 0$,折扣因子 $\alpha > 0$;

2) 由式(10) 和(18) 分别计算 $A_\alpha^{\nu_n}$ 和 $g_\alpha^{\nu_n}$;

3) 选择策略 ν_{n+1},使对每个 $i \in \Phi$,满足

$$\nu_{n+1}(i) \in \arg\inf_{\nu(i)\in D(i)}\{f(i,\nu(i)) + \sum_{j\in\Phi} A_\alpha(i,j,\nu(i)) g_\alpha^{\nu_n}(j)\}; \tag{25}$$

4) 如果 $\mathrm{sp}(f^{\nu_{n+1}} + A_\alpha^{\nu_{n+1}} g_\alpha^{\nu_n}) < \varepsilon$(这里 $\mathrm{sp}(h) = \max\{h(i)\} - \min\{h(i)\}$ 为一半范数[6]),则算法终止;否则,置 $n := n+1$,转 2).

为保证式(25) 中的集合非空,一般需要对每个 $i \in \Phi$,假设 $f(i,\nu(i))$ 在紧集 $D(i)$ 上连续;对任意的 $i, j \in \Phi, t \geq 0$,假设 $Q(i,j,\nu(i),t)$ 在紧集 $D(i)$ 上连续. 可以证明在此条件下,上述迭代算法是收敛的,且收敛到的最优策略与初始策略的选择无关. 下面给一个数值例子,以说明该算法的应用.

考虑具有三个状态的半 Markov 过程 $Y, \Phi = \{1,2,3\}, D(i) = [0.5, 10], i = 1,2,3$. 平稳策略为 $\nu = (\nu(1), \nu(2), \nu(3)), \nu(i) \in [0.5, 10], i = 1,2,3$. 嵌入 Markov 链的转移概率分别为

$$P(1,1,\nu(1)) = 1 - \mathrm{e}^{-\nu(1)/2},$$
$$P(1,2,\nu(1)) = \frac{7}{8}\mathrm{e}^{-\nu(1)/2},$$

$$P(1,3,v(1)) =$$
$$1 - P(1,1,v(1)) - P(1,2,v(1)),$$
$$P(2,1,v(2)) = \frac{1 - e^{-v(2)/4}}{1 + e^{-v(2)/2}},$$
$$P(2,3,v(2)) = \frac{1}{3}e^{-v(2)/4},$$
$$P(2,2,v(2)) =$$
$$1 - P(2,1,v(2)) - P(2,3,v(2)),$$
$$P(3,1,v(3)) = \frac{1 - e^{-v(3)/4}}{1 + e^{-v(3)/4}},$$
$$P(3,3,v(3)) = 1 - \frac{1 - e^{-v(3)/3}}{1 + e^{-v(3)/2}},$$
$$P(3,2,v(3)) =$$
$$1 - P(3,1,v(3)) - P(3,3,v(3)).$$

已知过程 Y 处于状态 i 和下一次将转移到状态 j,它在状态 i 的逗留时间服从区间 $[0, jv(i)]$ 上的均匀分布,即分布函数为

$$F(i,j,v(i),t) = \begin{cases} \dfrac{t}{jv(i)}, 0 \leq t \leq jv(i); \\ 1, t > jv(i). \end{cases} \quad (26)$$

对于 $v(i) \in D(i)$, $i, j \in \Phi$, $t \geq 0$,有

$$Q(i,j,v(i),t) = P(i,j,v(i))\,F(i,j,v(i),t).$$

性能函数为

$$f(i,v(i)) = \ln[(1+i)v(i)] + \frac{\sqrt{i}}{2v(i)},$$
$$i = 1,2,3.$$

针对不同的 α 和初始策略 v_0,总取 $\varepsilon = 0.0001$. 本文所作的几种仿真结果如下:

1) 当 $\alpha = 0.1$, $v_0 = (1,1,1)$ 时,迭代 4 次,得到 ε 最优策略为 $(1.296\,8, 0.895\,9, 0.902\,0)$,最优代价为 $(17.511\,4, 18.188\,6, 19.737\,9)$;

2) 当 $\alpha = 0.1$, $v_0 = (8,9,10)$ 时,迭代 5 次,得到的 ε 最优策略和最优代价与 1) 相同;

3) 当 $\alpha = 0.9$, $v_0 = (1,1,1)$ 时,迭代 4 次,得到 ε 最优策略为 $(0.821\,1, 0.702\,0, 0.888\,1)$,最优代价为 $(1.512\,8, 1.820\,0, 2.359\,9)$;

4) 当 $\alpha = 0.9$, $v_0 = (8,9,10)$ 时,迭代 4 次,得到的 ε 最优策略和最优代价与 3) 相同;

5) 当 $\alpha = 10$, $v_0 = (1,1,1)$ 时,迭代 3 次,得到 ε 最优策略为 $(0.539\,3, 0.674\,4, 0.865\,5)$,最优代价为 $(0.105\,6, 0.170\,7, 0.222\,6)$;

6) 当 $\alpha = 10$, $v_0 = (8,9,10)$ 时,迭代 3 次,得到的 ε 最优策略和最优代价与 5) 相同.

由上述结果可以看出,该算法只需很少的迭代次数,便可得到较好的 ε 最优策略,且与初始策略的选择无关.

6 结 语

本文讨论一类半 Markov 控制过程的折扣代价性能优化问题. 对于 SMCP 引入折扣 Poisson 方程,并基于 α-势给出了最优性定理及由最优平稳策略所满足的最优性方程. 据此提出一种求解最优策略的迭代算法. 这些结果可直接用于研究半 Markov 系统的控制和优化问题.

参考文献(References):

[1] Howard R. Semi-Markovian decision processes[J]. *Inst Intcernat Statist*,1963,40:625-652.

[2] Jewell W S. Markov renewal programming: Ⅰ, Ⅱ[J]. *Operat Res*,1963,2:938-971.

[3] Ross S M. *Applied Probability Models with Optimization Applications*[M]. San Franciso: Holden-Day,1971.

[4] Beutler F J, Ross K W. Uniformization for semi-Markov decision processes under stationary policies[J]. *J Appl Prob*,1987,24: 644-656.

[5] 胡奇英,刘建塘. 马尔可夫决策过程引论[M]. 西安: 西安电子科技大学出版社,2000.

[6] Puterman M L. *Markov Decision Processes*[M]. New York: John Wiley,1994.

[7] Cinlar E. *Introduction to Stochastic Processes*[M]. Englewood Cliffs: Prentice-Hall Inc,1975.

[8] Xi Hongsheng, Tang Hao, Yin Baoqun. Optimal policies for a continuous time MCP with compact action set[J]. *Acta Automatica Sinica*,2003,29(2):206-211.

[9] Tang Hao, Xi Hongsheng, Yin Baoqun. Performance optimization of continuous-time Markov control processes based on performance potentials[J]. *Int J of Systems Science*, 2003,34(1):63-71.