文章编号: 1001-0920(2006)08-0933-04

可数半Markov 决策过程折扣代价性能优化

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

摘 要: 讨论一类可数半Markov 决策过程(CSMDP)在折扣代价准则下的性能优化问题. 运用等价Markov 过程方法, 定义了折扣 Poisson 方程, 并由该方程定义了 α-势. 基于 α-势, 导出了由最优平稳策略所满足的最优性方程. 较为详细地讨论了最优性方程解的存在性问题, 并给出了其解存在的一些充分条件.

关键词: 可数半Markov 决策过程; 折扣性能准则; 折扣 Poisson 方程; α-势; 最优性方程
中图分类号: O 232 **文献标识码**: A

Performance Optimization for Countable Semi-Markov Decision Processes with Discounted-cost

YIN Bao-qun, LI Yan-jie, ZHOU Ya-ping, XI Hong-sheng
(Department of Automation, University of Science and Technology of China, Hefei 230026, China. Correspondent: YIN Bao-qun, E-mail: bqyin@ustc.edu.cn)

Abstract: The problem of discounted-cost performance optimization is addressed for a class of countable semi-Markov decision processes (CSMDPs). By using an equivalent Markov process, the discounted Poisson equation is proposed for a CSMDP. This equation is used to define the α-potential, based on which the optimality equation satisfied by the optimal stationary policy is derived. The existence of solutions to the optimality equation is discussed and some sufficient conditions for the existence of solutions are given.

Key words: Countable semi-Markov decision processes; Discounted performance criteria; Discounted Poisson equation; α-potential; Optimality equations

1 引 言

对于可数Markov 决策过程(CMDP)性能优化问题的研究已取得了许多成果[1~6], 但对半Markov 决策过程(SMDP), 特别是可数半Markov 决策过程(CSMDP)折扣模型的研究却很少. 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, 讨论了一类可数半Markov 决策过程, 在一定的条件下, 给出了最优性方程. 文献[6]同样研究了一类可数半Markov 决策过程, 通过引入一个矩阵, 给出了最优性方程以及迭代优化算法. 文献[7]研究了一类有限半Makrov 决策过程及性能灵敏度分析问题. 文献[8]讨论了一类有限半Markov 决策过程折扣性能优化问题. 文献[9]讨论了一类有限半Markov 控制过程平均性能优化算法.

本文基于折扣 Poisson 方程, 研究了一类具有可数状态空间的半Markov 决策过程以及在折扣代价准则下的性能优化问题. 实际上, 本文是将文献[8]的结果推广到可数状态空间的情况. 由于在可数状态空间下, 一些可数矩阵的可逆性遇到了问题, 故在具体的处理方法上会与有限状态空间有所不同. 特别是本文在一定的条件下, 证明了最优性方程解的存在性. 最后讨论了最优性方程解存在的一些充分条件, 并举例说明了存在性定理中的条件是可以满足的.

收稿日期: 2005-05-13; **修回日期**: 2005-10-17.
基金项目: 国家自然科学基金项目(60274012, 60574065); 安徽省自然科学基金项目(050420301).
作者简介: 殷保群(1962—), 男, 安徽全椒人, 教授, 博士, 从事随机DEDS、系统优化及其应用等研究;
李衍杰(1978—), 男, 山东青岛人, 博士生, 从事DEDS 等研究.

2 折扣性能准则

考虑一个半Markov过程(SMP), $Y=\{Y_t; t\geq 0\}$, 具有可数状态空间 $\Phi=\{1,2,\cdots\}$ 和行动空间 D. $D(i)\subset D$ 为状态 i 的容许行动集. 本文假设对任意的 $i\in\Phi$, $D(i)$ 非空. 置 $X=\{X_n; n\geq 0\}$ 为 Y 的嵌入Markov链, $0=T_0<T_1<\cdots$ 为相继的状态转移时刻, 则 $(X,T)=\{X_n,T_n; n\geq 0\}$ 是一个具有状态空间 Φ 的Markov更新过程.

本文仅考虑平稳策略. 一个平稳策略是状态空间到行动空间的一个映射 $v:\Phi\to D$, 且对任意的 $i\in\Phi$, $v(i)=d_i\in D(i)$. 记 $v=(v(1),v(2),\cdots)$, 并令 Ω_s 为全体平稳策略集. 在策略 v 下, 嵌入Markov链的转移矩阵为 $P^v=[P(i,j,v(i))]$, 已知当前状态为 i 且下次转移到 j, Y 在状态 i 的逗留时间的分布函数为 $F(i,j,v(i),t)$. 如果设半Markov过程 Y 的半Markov核为 $Q^v(t)=[Q(i,j,v(i),t)]$, 则对 $i,j\in\Phi$, 有 $Q(i,j,v(i),t)=P(i,j,v(i))F(i,j,v(i),t)$. 这里, $Q(i,j,v(i),t)=P\{X_{n+1}=j, T_{n+1}-T_n\leq t\,|X_n=i, v(i)\}$ 不依赖于 n.

本文假设在任意策略 $v\in\Omega_s$ 下, 嵌入Markov链 X 是不可约正常返和非周期的. 令

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

$$h^v(t)=(h(1,v(1),t),h(2,v(2),t),\cdots)^\tau, \tag{2}$$

则有

$$h^v(t)=(I-Q^v(t))e, \tag{3}$$

这里 $e=(1,1,\cdots)^\tau$. 设 f 是一个依赖于策略 v 的性能函数, 且对每一个 $i\in\Phi$, $f(i,\bullet):D(i)\to(-\infty,+\infty)$, 记 $f^v=(f(1,v(1)),f(2,v(2)),\cdots)^\tau$.

称 $(Y,\Phi,D,Q^v(t),f^v)$ 为一个约束在平稳策略集 Ω_s 上的CSMDP, 其无限水平折扣性能准则为

$$\eta_\alpha(i)=E\Big\{\int_0^\infty e^{-\alpha t}f(Y_t,v(Y_t))\,dt\,\Big|Y_0=i\Big\},\quad i\in\Phi, v\in\Omega_s, \tag{4}$$

这里 $\alpha>0$ 是一个折扣因子.

3 α-势

为简化记号, 暂时省略上标"v". 对于 $\alpha\geq 0$, 令

$$Q_\alpha=\int_0^\infty e^{-\alpha t}Q(dt),$$

$$h_\alpha=\int_0^\infty e^{-\alpha t}h(t)\,dt,$$

$$R_\alpha=\int_0^\infty e^{-\alpha t}R(dt). \tag{5}$$

这里 $R(t)=[R(i,j,t)]$ 为Markov更新过程 (X,T) 的Markov更新核[10]. 注意到 $Q(0)=0$, 则有

$$Q_0=\int_0^\infty Q(dt)=Q(+\infty)=P=[P(i,j)], \tag{6}$$

$$h_0=\int_0^\infty h(t)\,dt=(m(1),m(2),\cdots), \tag{7}$$

这里 $m(i)=\int_0^\infty h(i,t)\,dt$ 为SMP Y 在状态 i 的平均逗留时间. 由假设知, 对任意的 $i\in\Phi$, 有 $0<m(i)<+\infty$. 此外, 由式(3) 有

$$\alpha h_\alpha=(I-Q_\alpha)e. \tag{8}$$

根据文献[10], 有

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

特别地, 如果状态空间有限, 则矩阵 $(I-Q_\alpha)$ 可逆, 且 $R_\alpha=(I-Q_\alpha)^{-1}$.

对 $\alpha\geq 0$, 定义

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

其中 $H_\alpha=\mathrm{diag}(h_\alpha(1),h_\alpha(2),\cdots)$. 记 $P_\alpha=\alpha H_\alpha+Q_\alpha$, $\Lambda_\alpha=H_\alpha^{-1}$, 则易知 P_α 是一个Markov矩阵. 从而 A_α 又可表示为

$$A_\alpha=\Lambda_\alpha(P_\alpha-I). \tag{11}$$

因此, $A_\alpha=[A_\alpha(i,j)]$ 可作为一个Markov过程的无穷小矩阵.

对于 $\alpha>0$, 令

$$U_\alpha=\int_0^\infty e^{-\alpha t}P(t)\,dt \tag{12}$$

这里: $P(t)=[P(i,j,t)]$, $P(i,j,t)=P\{Y_t=j\,|Y_0=i\}$. 显然有

$$U_\alpha e=e/\alpha \tag{13}$$

此外, 由文献[10]中的式(10.5.14), 有

$$U_\alpha=R_\alpha H_\alpha \tag{14}$$

故当 $\alpha>0$ 时, 由 A_α 的定义, 易知

$$(\alpha I-A_\alpha)U_\alpha=U_\alpha(\alpha I-A_\alpha)=I. \tag{15}$$

特别地, 如果状态空间有限, 则矩阵 $(\alpha I-A_\alpha)$ 可逆, 且 $U_\alpha=(\alpha I-A_\alpha)^{-1}$.

现在再加上上标"v". 由 η_α 的定义, 易见

$$\eta_\alpha(i)=\int_0^\infty e^{-\alpha t}\sum_{j\in\Phi}P^v(i,j,t)f(j,v(j))\,dt,\quad i\in\Phi, v\in\Omega_s. \tag{16}$$

故若记 $\eta_\alpha=(\eta_\alpha(1),\eta_\alpha(2),\cdots)^\tau$, 则有

$$\eta_\alpha=U_\alpha^v f^v, v\in\Omega_s, \tag{17}$$

于是, 由式(15) 可得

$$(\alpha I-A_\alpha^v)\eta_\alpha=f^v, v\in\Omega_s. \tag{18}$$

对任意的 $v\in\Omega_s$, $\alpha\geq 0$, 定义CSMDP的折扣Poisson方程为

$$(\alpha I-A_\alpha^v)g_\alpha^v=f^v-\frac{ep_\alpha^v f^v}{1+\alpha}, \tag{19}$$

其中 p_α^v 是方程

$$p_\alpha^v A_\alpha^v=0, p_\alpha^v e=1 \tag{20}$$

的一个非负解. 可以证明, 在嵌入Markov链 X 不可

约正常返和非周期的条件下, 此方程存在非负解. 当 $\alpha > 0$ 时, 称 g_α^v 为CSMDP 的 α- 势.

当 $\alpha > 0$ 时, 在式(19) 两边左乘 U_α^v, 并运用式(13), (15) 及(17), 可得

$$\eta_\alpha = g_\alpha^v + \frac{ep_\alpha^v f^v}{\alpha(1+\alpha)}. \tag{21}$$

由式(21) 可知, 折扣 Poisson 方程(19) 存在唯一解 g_α^v.

4　最优性方程及其解的存在性

由式(18) 和(15), 容易得到下列引理:

引理 1　对任意的 $v', v \in \Omega_s$, 有

$$\eta_\alpha' - \eta_\alpha = U_\alpha^v[(f^{v'} + A_\alpha^{v'} g_\alpha^{v'}) - (f^v + A_\alpha^v g_\alpha^{v'})]. \tag{22}$$

由引理 1, 可以得到最优性方程.

定理 1　$v^* \in \Omega_s$ 是 $(Y, \Phi, D, Q^v(t), f^v)$ 在折扣代价准则下的一个最优平稳策略的充分必要条件为 v^* 满足方程

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

方程(23) 称为CSMDP 基于 α- 势的折扣代价最优性方程. 由式(21), 最优性方程(23) 也可表示为

$$0 = \inf_{v \in \Omega_s}\{f^v - (\alpha I - A_\alpha^v)\eta_\alpha^*\}. \tag{24}$$

下面研究最优性方程解的存在性. 先给出下面 2 个假设:

假设 1　f^v 一致有界, 即存在 M, 使对任意的 $d \in D(i)$, $i \in \Phi$, 有 $|f(i,d)| \le M$.

假设 2　存在 $\lambda > 0$, 使对任意的 $\alpha > 0$, $d \in D(i)$, $i \in \Phi$, 有

$$\sum_{j \in \Phi} Q_\alpha(i, j, d) \le \lambda h_\alpha(i, d). \tag{25}$$

假设 2 不太容易验证, 这里研究假设 2 成立的一些充分条件. 由于 $h(i,d,t)$ 是 t 的不增函数, 故其关于 t 在 $[0, +\infty)$ 上几乎处处可微. 从而不难得到下列引理:

引理 2　若存在 $\lambda > 0$, 使对任意的 $d \in D(i)$, $i \in \Phi$, 均有

$$h'(i,d,t) \ge -\lambda h(i,d,t) \tag{26}$$

关于 t 在 $[0, +\infty)$ 上几乎处处成立, 则假设 2 成立.

引理 3　若存在 $\lambda > 0$, 使对任意的 $t \ge 0$, $d \in D(i)$, $i \in \Phi$, 均有

$$h(i,d,t) \ge e^{-\lambda t}, \tag{27}$$

则假设 2 成立.

定理 2　在假设 1 和假设 2 成立的条件下, 对每一个 $\alpha > 0$, 有:

1) 存在唯一有界的 η_α^* 满足最优性方程

$$0 = \inf_{v \in \Omega_s}\{f^v - (\alpha I - A_\alpha^v)\eta_\alpha^*\}; \tag{28}$$

2) 如果存在 $v^* \in \arg\inf_{v \in \Omega_s}\{f^v - (\alpha I - A_\alpha^v)\eta_\alpha^*\}$, 则 v^* 是一个最优平稳策略.

证明　由式(8) 和假设 2, 对任意的 $\alpha > 0$, $d \in D(i)$, $i \in \Phi$, 有

$$|A_\alpha(i,i,d)| = |\alpha - h_\alpha^{-1}(i,d)(1 - Q_\alpha(i,i,d))| = h_\alpha^{-1}(i,d)\sum_{j \ne i} Q_\alpha(i,j,d) \le \lambda \tag{29}$$

故若令 $\tilde{P}_\alpha^v = A_\alpha^v/\lambda + I$, $\beta = \lambda/(\lambda + \alpha)$, 则显然 $\tilde{P}_\alpha^v$ 是一个Markov 矩阵, 且 $0 < \beta < 1$. 而式(28) 变为

$$\eta_\alpha^* = \inf_{v \in \Omega_s}\left\{\frac{f^v}{\lambda + \alpha} + \beta\tilde{P}_\alpha^v\eta_\alpha^*\right\} \tag{30}$$

因此, 根据文献[6] 便可得到该定理. □

例 1　考虑一个CSMDP, 状态空间 $\Phi = \{1, 2, \cdots\}$, 容许行动集 $D(i) = [2, +\infty)$, $i \in \Phi$. 嵌入Markov 链的转移概率如下: $P(1,1,v(1)) = 1 - e^{-(7/8)v(1)}$, $P(1,2,v(1)) = e^{-(7/8)v(1)}$; 对于 $i = 2, 3, \cdots$, $P(i, i-1, v(i)) = e^{-(3/8)v(i)}$, $P(i, i+1, v(i)) = e^{-(1/8)v(i)}$, $P(i,i,v(i)) = 1 - e^{-(3/8)v(i)} - e^{-(1/8)v(i)}$; 其余 $P(i,j,v(i))$ 全为 0. 分布函数为

$$F(i,j,v(i),t) = \begin{cases} \dfrac{t}{v(i)}, & 0 \le t \le v(i); \\ 1, & t > v(i); \end{cases} \quad i, j \in \Phi$$

则可以验证假设 2 成立, 且 $\lambda = 1$.

5　结　　语

本文讨论了一类CSMDP 的折扣代价性能优化问题, 将有限状态空间半Markov 决策过程的相关结果推广到可数状态空间的情况. 考虑到可数矩阵的可逆性问题, 在某些具体的处理过程中, 采取了与有限状态空间有所不同的方式. 特别是在一定的条件下, 证明了最优性方程解的存在性, 并较为详细地讨论了最优性方程解存在的一些充分条件. 这些结果可直接用于研究可数半Markov 系统的控制和优化问题.

参考文献(References)

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

[2] Yushkevich A A. Controlled Markov Models with Countable State Space and Continuous Time [J]. *Theory Probability and Applications*, 1977, 22(2): 215-235.

[3] 宋京生. 转移概率族非一致有界的连续时间马氏决策规划[J]. *中国科学*(A 辑), 1987, 17(12): 1258-1267.

(Song J S. Continuous Time Markov Decision Programming with Nonuniformly Bounded Transition Rate [J]. *Scientia Sinica (Series A)*, 1987, 17 (12): 1258-1267.)

[4] Guo X P, Zhu W P. Denumerable-state Continuous Time Markov Decision Processes with Unbounded Transition and Reward Rates Under the Discounted Criterion [J]. *J of Applied Probability*, 2002, 39 (2): 233-250.

[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.
(Hu Q Y, Liu J Y. *An Introduction to Markov Decision Processes* [M]. Xi'an: Xidian University Publication, 2000.)

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

[7] Cao X R. Semi-Markov Decision Problems and Performance Sensitivity Analysis [J]. *IEEE Trans on Automatical Control*, 2003, 48 (5): 758-769.

[8] 殷保群, 李衍杰, 周亚平, 等. 半Markov控制过程在折扣代价准则下的最优平稳策略 [J]. *控制与决策*, 2004, 19 (6): 691-694.
(Yin B Q, Li Y J, Zhou Y P, et al. Optimal Stationary Policies for Semi-Markov Control Processes with Discounted-cost Criteria [J]. *Control and Decision*, 2004, 19 (6): 691-694.)

[9] Dai G P, Yin B Q, Li Y J, et al. Performance Optimization Algorithms Based on Potential for Semi-Markov Control Processes [J]. *Int J of Control*, 2005, 78 (11): 801-812.

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

(上接第 928 页)

参考文献(References)

[1] Tore H, Astrom K J. Industrial Adaptive Controllers Based on Frequency Response Techniques [J]. *Automatica*, 1991, 27 (4): 599-609.

[2] Galeani Sergio, Grasselli Osvaldo Maria, Menini Laura. Strong Stabilization with Infinite Multivariable Gain Margin through Linear Periodic Control [J]. *Int J of Control*, 2004, 77 (5): 441-460.

[3] Perng J W, Wu B F, Chin H I, et al. Gain-phase Margin Analysis of Dynamic Fuzzy Control Systems [J]. *IEEE Trans on Systems, Man and Cybernetics: Part B*, 2004, 34 (5): 2133-2139.

[4] Wu B F, Perng J W. Gain-phase Margin Analysis of Pilot-induced Oscillations for Limit-cycle Prediction [J]. *J of Guidance, Control and Dynamics*, 2004, 27 (1): 59-65.

[5] Cavicchi Thomas J. Phase Margin Revisited: Phase-robot Locus, Bode Plot and Phase Shifters [J]. *IEEE Trans on Education*, 2003, 46 (1): 168-176.

[6] Krajewski Wieslaw, Lepschy Antonio, Viaro Umberto. Designing PI Controllers for Robust Stability and Performance [J]. *IEEE Trans on Control Systems Technology*, 2004, 12 (6): 973-983.

(上接第 932 页)

[2] Jeffrey T S, Kevin M P. Decentralized Adaptive Control of Nonlinear Systems Using Radial Basis Neural Networks [J]. *IEEE Trans on Automatic Control*, 1999, 44 (11): 2050-2057.

[3] 张颖伟, 王剑, 张嗣瀛. 一类具有不确定性的相似组合系统鲁棒输出反馈镇定 [J]. *控制理论与应用*, 2001, 18 (4): 573-575.
(Zhang Y W, Wang J, Zhang S Y. Decentralized Output Feedback Robust Stabilization for a Class of Nonlinear Interconnected Systems with Similarity [J]. *Control Theory and Application*, 2001, 18 (4): 573-575.)

[4] Yan X G, Zhang S Y. Robust Stability of Nonlinear Large-scale Composite Systems with Uncertain Parameters [A]. *The 13th IFAC World Congress* [C]. Sanfransco, 1996: 467-473.

[5] 刘晓志, 井元伟, 张嗣瀛. 采用还原方法的不确定关联时滞系统的鲁棒分散镇定 [J]. *控制与决策*, 2004, 19 (11): 1218-1222.
(Liu X Z, Jing Y W, Zhang S Y. Robust Decentralized Stabilization for Uncertain Interconnected Delayed Systems Using Reduction Method [J]. *Control and Decision*, 2004, 19 (11): 1218-1222.)