第23卷第1期 控制理论与应用 Vol.23 No.1
2006年2月 Control Theory & Applications Feb. 2006

文章编号：1000－8152(2006)01－0065－04

半Markov决策过程折扣模型与平均模型之间的关系

殷保群[1]，李衍杰[1]，唐 昊[2]，代桂平[1]，奚宏生[1]
（1. 中国科学技术大学 自动化系，安徽 合肥 230026；2. 合肥工业大学 计算机系，安徽 合肥 230009）

摘要：首先分别在折扣代价与平均代价性能准则下，讨论了一类半Markov决策问题. 基于性能势方法，导出了由最优平稳策略所满足的最优性方程. 然后讨论了两种模型之间的关系，表明了平均模型的有关结论，可以通过对折扣模型相应结论取折扣因子趋于零时的极限来得到.

关键词：半Markov决策过程；折扣模型；平均模型；最优性方程；最优平稳策略

中图分类号：TP202 **文献标识码**：A

Relations between discounted models and average models for semi-Markov decision processes

YIN Bao-qun[1], LI Yan-jie[1], TANG Hao[2], DAI Gui-ping[1], XI Hong-sheng[1]
(1. Department of Automation, University of Science and Technology of China, Hefei Anhui 230026, China;
2. Department of Computer, Hefei University of Technology, Hefei Anhui 230009, China)

Abstract: The semi-Markov decision problems are discussed for discounted-cost and average-cost performance criteria, respectively. Based on a potential approach, the optimality equations satisfied by the optimal stationary policies are derived. Then the relation between the discounted model and average model is studied. It shows that the related conclusions for the average model can be obtained by taking the limits of results about the discounted model as the discounted factor tends to zero.

Key words: semi-Markov decision processes; discounted model; average model; optimality equation; optimal stationary policy

1 引言(Introduction)

在Markov决策过程(MDP)的研究中，就准则函数来说，我们考虑最多的是两种性能准则，即折扣准则和平均准则. 也即我们通常所说的折扣模型和平均模型. 目前对这两种模型的研究文献非常之多[1~7]. 在一些文献[5,7]中，也对这两种模型之间的关系进行了研究. 而对半Markov决策过程(SMDP)两种模型之间关系的研究，相应的文献较少，文献[8]讨论了两种模型之间的联系.

本文首先基于Poisson方程，研究一类具有有限状态空间的半Markov决策问题. 我们定义一个矩阵，该矩阵可以作为一个Markov过程的无穷小矩阵，并给出了该矩阵的一些性质. 通过这个矩阵，我们对一个SMDP引入Poisson方程，并根据这个方程，定义α-势与性能势. 然后分别在折扣代价与平均代价性能准则下，导出由最优平稳策略所满足的最优性方程. 最后研究了两种模型之间的关系，并表明平均模型的有关结论，可以通过对折扣模型相应结论取折扣因子趋于零时的极限来得到.

本文实际上发展了文献[7]中所运用的方法. 文献[7]通过一个矩阵M给出了最优性方程和策略迭代算法. 事实上，这个矩阵就是本文中的矩阵Q_α. 在策略迭代中，这个矩阵的物理意义不是太清楚. 而本文是通过矩阵A_α给出最优性方程，这个矩阵A_α可以作为一个Markov过程的无穷小矩阵，因此其物理意义非常清楚. 我们通过两个引理给出了矩阵A_α的一些性质. 在引理1的证明中，我们通过运用文献[9]中的一个定理，直接导出了$pA=0$的结果. 这比文献[8]的相应处理要简单得多. 引理2表明了具有无穷小矩阵A_α的Markov过程是不可约和正常返的，因此这个Markov过程具有唯一严格正的稳态分布. 而此点在本文式(42)的推导中起着重要作用.

收稿日期：2004－04－29；收修改稿日期：2005－04－30.
基金项目：国家自然科学基金资助项目(60274012,60574065)；安徽省自然科学基金资助项目(050420301).

由于本文定义的一些基本量,如折扣性能,α-势等,与文献[8]中的定义有些不同,并且在处理方法上也有不小的差别,故本文对两种模型之间关系的讨论与文献[8]相比也有较大的差异.

2 折扣性能准则与平均性能准则(Discounted performance criteria and average performance criteria)

考虑一个半Markov过程$Y=\{Y_t;\ t\geqslant 0\}$,它有一个有限状态空间$\Phi=\{1,2,\cdots,k\}$.设D为一个行动空间,$D(i)\subset D$是状态i的容许行动集.设对任意的$i\in\Phi$,$D(i)$非空.置$X=\{X_n;\ n\geqslant 0\}$是Y的嵌入Markov链,$0=T_0<T_1<\cdots$是相继的状态转移时刻,则$(X,T)=\{X_n,\ T_n;\ n\geqslant 0\}$是一个具有状态空间$\Phi$的Markov更新过程.记$v=(v(1),v(2),\cdots,v(k))$为一个平稳策略,并令$\Omega_s$是全体平稳策略集.在策略$v$下,$Y$的半Markov核是$Q^v(t)=[Q(i,j,v(i),t)]$.这里$Q(i,j,v(i),t)=P\{X_{n+1}=j,\ T_{n+1}-T_n\leqslant t\,|\,X_n=i,v(i)\}$不依赖于$n$.设在任意策略$v\in\Omega_s$下,相应的$Y$是不可约和非周期的,因而也是正常返的[9].于是,Y存在唯一的稳态分布$p^v=p^v(1),p^v(2),\cdots,p^v(k))>0$,嵌入Markov链$X$也存在唯一的稳态分布$\pi^v=(\pi^v(1),\pi^v(2),\cdots,\pi^v(k))>0$.这里,一个向量$g>0$是指它的每个分量$g(i)>0$,此时我们也称$g$是严格正的.令

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

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

则有

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

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

我们称$(Y,\Phi,D,Q^v(t),f^v)$为一个约束在平稳策略集Ω_s上的SMDP.Y的无限水平折扣性能准则是

$$\eta_\alpha^v(i)=\mathrm{E}\left\{\int_0^{+\infty}\mathrm{e}^{-\alpha t}f(Y_t,v(Y_t))\mathrm{d}t\,|\,Y_0=i\right\},\tag{4}$$

$i\in\Phi,v\in\Omega_s$.

这里$\alpha>0$是一个折扣因子.平均性能准则是

$$\eta^v=\lim_{T\to+\infty}\mathrm{E}\left\{\frac{1}{T}\int_0^T f(Y_t,v(Y_t))\mathrm{d}t\right\},\ v\in\Omega_s.\tag{5}$$

由于Y遍历,故$\eta^v=p^vf^v$[9].

在一个代价模型的SMDP问题中,优化的目标是选择一个策略$v^*\in\Omega_s$,使得$\eta_\alpha^v(i)$对每一个$i\in\Phi$达到最小(折扣模型)或η^v达到最小(平均模型),我们一般称这个策略v^*为最优平稳策略.

3 α-势与性能势(α-potential and performance potential)

为了简化记号,我们暂时省略上标"v".对$\alpha\geqslant 0$,令

$$Q_\alpha=\int_0^{+\infty}\mathrm{e}^{-\alpha t}Q(\mathrm{d}t),\ h_\alpha=\int_0^{+\infty}\mathrm{e}^{-\alpha t}h(t)\mathrm{d}t.\tag{6}$$

注意到$Q(0)=0$,则

$$Q_0=P=[P(i,j)],\quad h_0=(m(1),m(2),\cdots,m(k))^\tau.\tag{7}$$

这里,$P(i,j)=P\{X_{n+1}=j\,|\,X_n=i\}$是嵌入Markov链$X$的转移概率;$m(i)=\int_0^{+\infty}h(i,t)\mathrm{d}t$是过程$Y$在状态$i$的平均逗留时间.从式(3)有

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

对$\alpha\geqslant 0$定义

$$A_\alpha=\alpha I-H_\alpha^{-1}(I-Q_\alpha),\tag{9}$$

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

$$A_\alpha=\Lambda_\alpha(P_\alpha-I).\tag{10}$$

特别当$\alpha=0$时,我们省略下标"0",则

$$A=\Lambda(P-I).\tag{11}$$

因此,$A_\alpha=[A_\alpha(i,j)]$可以作为一个Markov过程的无穷小矩阵.下面我们来对矩阵A_α的性质作一些研究.为此,考虑状态空间Φ上的一个Markov过程$X^\alpha=\{X_t^\alpha;t\geqslant 0\}$,它具有无穷小矩阵$A_\alpha$.首先我们有下列引理.

引理1 Markov过程X^0存在唯一的稳态分布,且这个稳态分布就是半Markov过程Y的稳态分布p.

证 由于嵌入Markov链X不可约,故其转移矩阵P不可约,因此Markov过程X^0也不可约.而状态空间Φ有限,故其也是正常返的.于是,Markov过程X^0存在唯一的严格正的稳态分布,它是方程$xA=0$,$xe=1$的唯一严格正解.若记$\sigma=\sum_{j\in\Phi}\pi(j)m(j)=\pi h_0$,则由文献[9]中定理10.5.22可知

$$p(i)=\frac{\pi(i)m(i)}{\sigma},\ i\in\Phi,\tag{12}$$

或用矩阵形式表为

$$p\Lambda=\frac{1}{\sigma}\pi I.\tag{13}$$

注意到$\pi P=\pi$,则有

$$pA=p\Lambda(P-I)=\frac{1}{\sigma}\pi(P-I)=0.\tag{14}$$

故 p 就是 Markov 过程 X^0 的唯一稳态分布.

引理 2　对任意的 $\alpha \geqslant 0$，方程

$$xA_\alpha = 0,\ xe = 1 \tag{15}$$

存在唯一的严格正解 p_α.

证　当 $\alpha = 0$ 时，由引理 1 即得结论. 当 $\alpha > 0$ 时，显然我们只要证明矩阵 $P_\alpha = [P_\alpha(i,j)]$ 不可约即可. 为此，我们先来证明：当 $i \neq j$ 时，如果 $P_\alpha(i,j) = Q_\alpha(i,j) = 0$，则 $P(i,j) = 0$；反之，如果 $P(i,j) = 0$，则 $P_\alpha(i,j) = Q_\alpha(i,j) = 0$. 事实上，由于 $Q(i,j,t)$ 是 t 的不减函数，故其在 $[0, +\infty)$ 上关于 t 几乎处处可微，且 $Q'(i,j,t) \geqslant 0$ 在 $[0, +\infty)$ 上几乎处处成立. 如果 $\int_0^{+\infty} e^{-\alpha t} Q'(i,j,t)\,dt = Q_\alpha(i,j) = 0$，则 $Q'(i,j,t) = 0$ 在 $[0, +\infty]$ 上几乎处处成立. 于是，$P(i,j) = \int_0^{+\infty} Q'(i,j,t)\,dt = 0$. 反过来，显然亦成立. 因此，矩阵 P 与矩阵 P_α 的不可约性相同. 而 P 不可约，故 P_α 也不可约.

对 $\alpha > 0$，令

$$U_\alpha = \int_0^{+\infty} e^{-\alpha t} P(t)\,dt, \tag{16}$$

这里，$P(t) = [P_t(i,j)]$，$P_t(i,j) = P\{Y_t = j \mid Y_0 = i\}$. 则从文献[9]，我们易得

$$U_\alpha = (\alpha I - A_\alpha)^{-1}. \tag{17}$$

注意到 $A_\alpha e = 0$，则当 $\alpha > 0$ 时，易见

$$(\alpha I - A_\alpha)^{-1} e = \frac{e}{\alpha}. \tag{18}$$

现在再加上上标“v”. 根据式(4)，有

$$\eta_\alpha^v(i) = \int_0^{+\infty} {}_0 e^{-\alpha t} \sum_{j \in \Phi} p_t^v(i,j) f(j, v(j))\,dt. \tag{19}$$

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

$$\eta_\alpha^v = (\alpha I - A_\alpha^v)^{-1} f^v. \tag{20}$$

对任意的 $v \in \Omega_s$，$\alpha \geqslant 0$ 我们定义 SMDP 的折扣 Poisson 方程为

$$(\alpha I - A_\alpha^v) g_\alpha^v = f^v - \frac{e p_\alpha^v f^v}{1 + \alpha}. \tag{21}$$

这里，p_α^v 是方程(15)(在策略 v 下) 的唯一解. 当 $\alpha = 0$ 时，方程(21) 变为平均 Poisson 方程

$$A^v g^v = -f^v + e\eta^v. \tag{22}$$

而其解 g^v 称为性能势，它不是唯一的.

当 $\alpha > 0$，$v \in \Omega_s$ 时，由于矩阵 $(\alpha I - A_\alpha^v)$ 可逆，故方程(21) 存在唯一解 g_α^v，我们称 g_α^v 为 α- 势. 由式(18)(20) 及(21)，可得

$$\eta_\alpha^v = g_\alpha^v + \frac{e p_\alpha^v f^v}{\alpha(1 + \alpha)}. \tag{23}$$

4　最优性方程(Optimality equations)

我们将分别导出 SMDP 在两种性能准则下的最优性方程. 先考虑折扣代价模型，以下设 $\alpha > 0$，则容易得到下列引理.

引理 3　对任意的 v'，$v \in \Omega_s$，有

$$\eta_\alpha^{v'} - \eta_\alpha^v = (\alpha I - A_\alpha^v)^{-1}[(f^{v'} + A_\alpha^{v'} g_\alpha^{v'}) - (f^v + A_\alpha^v g_\alpha^{v'})]. \tag{24}$$

注意引理 3 中的式(24)还有另外一种表现形式：

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

由于本文给出引理 3 是为了导出最优性方程(26)，因此采用式(24)的形式比较方便. 如果是为了给出策略迭代算法，则采用上式是正确的.

根据引理 3，我们可以得到下列最优性定理.

定理 1　$v^* \in \Omega_s$ 是 SMDP$(Y, \Phi, D, Q^v(t), f^v)$ 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

$$f^{v^*} + A_\alpha^{v^*} g_\alpha^{v^*} \leqslant f^v + A_\alpha^v g_\alpha^{v^*},\ v \in \Omega_s. \tag{25}$$

从定理 1，我们可以直接获得最优性方程.

定理 2　$v^* \in \Omega_s$ 是 SMDP$(Y, \Phi, D, Q^v(t), f^v)$ 在折扣代价准则下的一个最优平稳策略的充分必要条件是其满足方程

$$0 = \inf_{v \in \Omega_s}\{f^v + A_\alpha^v g_\alpha^{v^*} - \alpha\eta_\alpha^{v^*}\}. \tag{26}$$

方程(26) 称为 SMDP 基于 α-势的折扣代价最优性方程. 下面我们考虑平均代价模型. 根据引理 1 可知，SMDP$(Y, \Phi, D, Q^v(t), f^v)$ 与 MDP(X^0, Φ, D, A^v, f^v)，在平均性能准则下是等价的. 故我们不难得到下面的几个结论.

引理 4　对任意的 $v', v \in \Omega_s$，我们有

$$\eta^{v'} - \eta^v = p^v[(f^{v'} + A^{v'} g^{v'}) - (f^v + A^v g^{v'})]. \tag{27}$$

定理 3　$v^* \in \Omega_s$ 是 SMDP$(Y, \Phi, D, Q^v(t), f^v)$ 在平均代价准则下的一个最优平稳策略的充分必要条件为

$$f^{v^*} + A^{v^*} g^{v^*} \leqslant f^v + A^v g^{v^*},\ v \in \Omega_s. \tag{28}$$

定理 4　$v^* \in \Omega_s$ 是 SMDP$(Y, \Phi, D, Q^v(t), f^v)$ 在平均代价准则下的一个最优平稳策略的充分必要条件是其满足方程

$$0 = \inf_{v \in \Omega_s}\{f^v + A^v g^{v^*} - e\eta^{v^*}\}. \tag{29}$$

方程(29)称为 SMDP 基于性能势的平均代价最优性方程.

5　两种模型之间的关系(Relations between two models)

为了简化记号，我们仍然暂时省略上标“v”. 首先我们有下列引理.

定理 5 对任意的 $\alpha \geqslant 0$,矩阵 $(\alpha I - A_\alpha + ep_\alpha)$ 可逆. 且当 $\alpha > 0$ 时, 有

$$(\alpha I - A_\alpha + ep_\alpha)^{-1} = (\alpha I - A_\alpha)^{-1} - \frac{ep_\alpha}{\alpha(1+\alpha)}. \tag{30}$$

由 $Q(t)$, Q_α 及 h_α 的定义易知

$$\lim_{\alpha \to +0} Q_\alpha = Q_0 = P, \ \lim_{\alpha \to +0} h_\alpha = h_0. \tag{31}$$

故有

$$\lim_{\alpha \to +0} A_\alpha = A. \tag{32}$$

在式(21) 两边左乘 p_α 并注意到 $p_\alpha A_\alpha = 0$, $p_\alpha e = 1$, 则有

$$p_\alpha g_\alpha = \frac{p_\alpha f}{1+\alpha}. \tag{33}$$

故折扣 Poisson 方程(21)又可写为

$$(\alpha I - A_\alpha + ep_\alpha) g_\alpha = f. \tag{34}$$

由引理5可知,对任意的 $\alpha \geqslant 0$, 方程(34)存在唯一解

$$g_\alpha = (\alpha I - A_\alpha + ep_\alpha)^{-1} f. \tag{35}$$

特别当 $\alpha = 0$ 时, 其解 $g_0 = (-A + ep)^{-1} f$ 是平均 Poisson 方程(22) 的一个解. 而方程(22) 的一切解可表为 $g = g_0 + \beta e$, β 是任意实数.

此外,我们不难证明

$$\lim_{\alpha \to +0} p_\alpha = p. \tag{36}$$

事实上,若记 $\lim\limits_{\alpha \to +0} \sup p_\alpha = \bar{p}$, $\lim\limits_{\alpha \to +0} \inf p_\alpha = \underline{p}$, 则我们有

$$\bar{p}A = 0, \ \bar{p}e = 1; \ \underline{p}A = 0, \ \underline{p}e = 1. \tag{37}$$

由于方程 $pA = 0$, $pe = 1$ 仅有唯一解 p, 故必有 $\bar{p} = \underline{p} = p$.

由式(32)(36), 可知

$$\lim_{\alpha \to +0} (\alpha I - A_\alpha + ep_\alpha)^{-1} = (-A + ep)^{-1}. \tag{38}$$

故有

$$\lim_{\alpha \to +0} g_\alpha = g_0. \tag{39}$$

以下, 简记 g_0 为 g. 在式(23) 两边同乘 α, 并令 $\alpha \to +0$, 则有

$$\lim_{\alpha \to +0} \alpha \eta_\alpha = e\eta. \tag{40}$$

又在式(30) 两边同乘 α, 并令 $\alpha \to +0$, 则得

$$\lim_{\alpha \to +0} \alpha (\alpha I - A_\alpha)^{-1} = ep. \tag{41}$$

现在我们再加上上标"v". 在式(24) 两边同乘 α, 并令 $\alpha \to +0$, 则对任意的 v', $v \in \Omega_s$, 我们有

$$e(\eta^{v'} - \eta^v) = ep^v[(f^{v'} + A^{v'} g^{v'}) - (f^v + A^v g^{v'})]. \tag{42}$$

上式两边左乘 p^v, 即得式(27). 故对平均代价模型, 我们可以直接利用折扣代价模型, 通过对折扣因子取趋于零时的极限, 来得到有关的结果.

6 结论(Conclusion)

我们研究了一类有限半 Markov 决策过程在两种性能准则下的最优性方程. 通过定义一个无穷小矩阵, 对一个 SMDP 引入 Poisson 方程, 并定义 α-势与性能势. 然后分别在折扣代价与平均代价性能准则下, 导出由最优平稳策略所满足的最优性方程. 最后研究了两种模型之间的关系, 并表明了平均模型的有关结论, 可以通过对折扣模型相应结论取折扣因子趋于零时的极限来得到. 由于 MDP 为 SMDP 的一种特殊情况, 故本文的结论可以直接应用于具有有限状态的 MDP. 这些结果可直接用来研究半 Markov 型系统的控制与优化问题.

参考文献(References):

[1] 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, 12: 1258 - 1267.
(SONG J S. Continuous time Markov decision programming with nonuniformly bounded transition rate [J]. *Scientia Sinica(Series A)*, 1987, 12:1258 - 1267.)

[2] GUO X P, LIU K. A note on optimality conditions for continuous-time Markov decision processes with average cost criterion [J]. *IEEE Trans on Automatic Control*, 2001, 46(12): 1984 - 1989.

[3] 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 Appl Prob*, 2002, 39(2): 233 - 250.

[4] 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.

[5] CAO X R. A unified approach to Markov decision problems and performance sensitivity analysis [J]. *Automatica*, 2000, 36(5): 771 - 774.

[6] 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 Qiying, LIU Jianyong. *An Introduction to Markov Decision Processes* [M]. Xi'an: Xidian University Publication, 2000.)

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

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

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

作者简介:

殷保群 (1962—), 男, 副教授, 博士, 主要从事随机离散事件动态系统性能分析、优化及其应用等方面的研究工作, E-mail: bqyin@ustc.edu.cn;

李衍杰 (1978—), 男, 中国科技大学自动化系在读博士研究生, 主要研究方向为离散时间动态系统分析、控制和应用;

唐 昊 (1972—), 男, 合肥工业大学计算机系副教授, 博士, 主要研究方向为离散时间动态系统;

代桂平 (1977—), 女, 中国科技大学自动化系在读博士研究生, 主要研究方向为离散时间动态系统;

奚宏生 (1950—), 男, 中国科学技术大学自动化系常务副主任, 教授, 博士生导师, 主要研究方向为鲁棒控制, 离散事件动态系统及其应用等.