

一类可数 Markov 控制过程的最优平稳策略

殷保群, 李衍杰, 奚宏生, 周亚平
(中国科学技术大学 自动化系, 安徽 合肥 230026)

摘要: 研究了一类具有可数状态空间的 Markov 控制过程在无限水平平均代价准则下的最优平稳策略问题. 对此类过程, 引入了折扣 Poisson 方程, 运用无穷小矩阵和性能势的基本性质, 导出了平均代价模型在紧致行动集上的最优性方程, 并证明了其解的一个存在性定理.

关键词: 可数 Markov 控制过程; 性能势; 平均代价准则; 紧致行动集; 最优平稳策略
中图分类号: TP202 **文献标识码:** A

Optimal stationary policies for a class of countable Markov control processes

YIN Bao-qun, LI Yan-jie, XI Hong-sheng, ZHOU Ya-ping

(Department Of Automation, University of Science and Technology of China, Hefei Anhui 230026, China)

Abstract: The problems of optimal stationary policies are studied for a class of Markov control processes with countable state spaces and infinite horizon average-cost criteria. The discounted Poisson equation is introduced for these processes. Using the basic properties of infinitesimal generators and performance potentials, the optimality equation is given for the average-cost model on a compact action set, and an existence theorem of the solution to this equation is proved.

Key words: countable Markov control process; performance potential; average-cost criteria; compact action set; optimal stationary policy

1 引言 (Introduction)

Markov 控制过程是一类受到一系列控制决策驱动的 Markov 系统, 它的状态转移规律与控制决策所采用的行动方案相互作用决定了过程的演化. 文献[1~3]研究了连续时间 Markov 控制过程 (CTMCP) 的平均代价最优性方程及其解的存在性问题, 并给出了收敛的迭代求解算法, 但这一般需要较强的约束条件来保证. Cao 将 Markov 性能势理论引入了 Markov 控制过程的性能分析中^[4~6], 放宽了约束条件的限制, 其结果已被推广到排队网络系统的研究中^[7,8]. 文献[9]在此基础上, 研究了一类具有有限状态空间和紧致行动集的 CTMCP 在无限水平平均代价准则下的优化问题. 本文进一步考虑具有可数状态空间的 Markov 控制过程在无限水平平均代价准则下的优化问题. 通过引入折扣 Poisson 方程, 并运用无穷小矩阵和性能势的基本性质, 导出了此类过程在紧致行动集上的最优性方程, 并证明了其解的一个存在性定理.

2 可数 Markov 控制过程 (Countable Markov control process)

考虑连续时间 Markov 过程 $X = \{X_t; t \geq 0\}$, 具有可数状态空间 $\Phi = \{1, 2, \dots\}$ 和行动空间 D . 对每一个 $i \in \Phi$, $D(i) \subset D$ 为状态 i 的容许行动集, 设 $D(i)$ 非空. 记 $\bar{D} = D(1) \times D(2) \times \dots$, 并设 $\bar{D}$ 是紧集. 我们仅考虑平稳策略, 一个平稳策略是状态空间 Φ 到行动空间 D 的一个映射 $v: \Phi \rightarrow D$, 即对 $i \in \Phi$, $v(i) = d_i \in D(i)$. 记 $v = (v(1), v(2), \dots)$, 令 Ω_v 是全体平稳策略集. 假设在任意策略 $v \in \Omega_v$ 的驱动下, X 是不可约的, 正常返的和强遍历的, 无穷小矩阵为 $A^v = [a_{ij}^v]$. 又设在每一个策略 $v \in \Omega_v$ 下, X 具有标准转移概率矩阵 $P^v(t) = [p_{ij}^v(t)]$, 且对任意给定的 $j \in \Phi$, 当 $t \rightarrow 0$ 时, $p_{ij}^v(t)/t$ 关于 $i \in \Phi$ 一致收敛到 a_{ij}^v , $i \neq j$. 此假设保证 Kolmogorov 前瞻方程和后退方程成立^[10].

设 $\{X_n; n \geq 0\}$ 是 X 的嵌入 Markov 链, 其一步转移概率矩阵为 $P^v = [p_{ij}^v]$, 则

$$A^v = \text{diag}\{\lambda^v(1), \lambda^v(2), \dots\}(P^v - I). \quad (1)$$

这里 $\lambda^v(i) > 0$ 为 Markov 过程 X 在策略 v 下状态 i 的转移率. 假设 P^v 和 $\lambda^v(i)$, $i \in \Phi$ 均在 $\bar{D}$ 上连续, 且 $\sup_{i \in \Phi, v \in \Omega_s} \{\lambda^v(i)\} = \lambda < +\infty$. 由文献[10]中定理 8.5.11 及其说明可知, 在任意策略 v 下, X 存在唯一的稳态分布 $p^v = (p^v(1), p^v(2), \dots) > 0$, 且 $P^v(t) \rightarrow ep^v (t \rightarrow +\infty)$. 我们还假设 $P^v(t)$ 关于 $v \in \Omega_s$ 一致收敛到 ep^v . 则不难验证, p^v 也在 $\bar{D}$ 上连续. 事实上, 由于 $P^v(t) = e^{A^v t}$, 故对每一个 $t \geq 0$, $P^v(t)$ 在 $\bar{D}$ 上连续, 从而由上述假设即得. 根据 Markov 过程理论, 应有

$$p^v e = 1, A^v e = 0, p^v A^v = 0. \quad (2)$$

这里, $e = (1, 1, \dots)^T$, “ T ” 表示转置. 令 f 为依赖于策略 v 的性能函数, 且 $f(i, \cdot): D(i) \rightarrow \mathbb{R}$, $i \in \Phi$. 记 $f^v = (f(1, v(1)), f(2, v(2)), \dots)^T$, 假设 f^v 在 $\bar{D}$ 上连续, 且一致有界.

称 (X, Φ, D, A^v, f^v) 为约束在平稳策略集 Ω_s 上的一个可数 Markov 控制过程 (CMCP). X 关于无限时间水平的折扣代价准则为

$$\eta_\alpha^v(i) = E\left\{\int_0^{+\infty} e^{-\alpha t} f(X_t, v(X_t)) dt \mid X_0 = i\right\}, \quad v \in \Omega_s, i \in \Phi. \quad (3)$$

这里 $\alpha > 0$ 为折扣因子; 平均代价准则为

$$\eta^v = \lim_{T \rightarrow \infty} \frac{1}{T} E\left\{\int_0^T f(X_t, v(X_t)) dt\right\}, v \in \Omega_s. \quad (4)$$

因为 X 遍历, 故有

$$\eta^v = \sum_{i \in \Phi} p^v(i) f(i, v(i)) = p^v f^v. \quad (5)$$

在 CMCP 的代价问题中, 优化的目标是选择一个控制策略, 使 $\eta_\alpha^v(i)$ (折扣代价) 对每一个 $i \in \Phi$ 或 η^v (平均代价) 达到最小.

3 性能势 (Performance potential)

记 $\eta_\alpha^v = (\eta_\alpha^v(1), \eta_\alpha^v(2), \dots)^T$, 则由文献[10]中定理 8.4.18, 有

$$(\alpha I - A^v) \eta_\alpha^v = f^v. \quad (6)$$

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

$$(\alpha I - A^v + \lambda ep^v) g = f^v. \quad (7)$$

特别, 当 $\alpha = 0$ 时, 我们得到 CMCP 的平均代价 Poisson 方程

$$(-A^v + \lambda ep^v) g = f^v. \quad (8)$$

引理 1 对任意 $\alpha \geq 0$, 矩阵 $(\alpha I - A^v + \lambda ep^v)$ 可逆.

证 令 $\tilde{P}^v = A^v/\lambda + I$, $\beta = \lambda/(\lambda + \alpha)$, 则 $0 < \beta \leq 1$, $\tilde{P}^v$ 是一随机矩阵, 且易验证

$$\beta(\alpha I - A^v + \lambda ep^v) = \lambda[I - \beta(\tilde{P}^v - ep^v)],$$

故只要证明矩阵 $[I - \beta(\tilde{P}^v - ep^v)]$ 可逆即可.

由式(2), 易验证 $p^v \tilde{P}^v = p^v$, 即当 $n \rightarrow \infty$ 时, $(\tilde{P}^v)^n \rightarrow ep^v$. 又由前面假设可知, $\tilde{P}^v$ 是强遍历的, 故根据文献[11]中定理 33 可知, 矩阵 $(I - \tilde{P}^v + ep^v)$ 可逆, 且

$$(I - \tilde{P}^v + ep^v)^{-1} = \sum_{n=0}^{\infty} (\tilde{P}^v - ep^v)^n. \quad (9)$$

因此, 当 $0 < \beta < 1$ 时, 级数 $\sum_{n=0}^{\infty} \beta^n (\tilde{P}^v - ep^v)^n$ 收敛, 从而矩阵 $[I - \beta(\tilde{P}^v - ep^v)]$ 可逆, 且

$$[I - \beta(\tilde{P}^v - ep^v)]^{-1} = \sum_{n=0}^{\infty} \beta^n (\tilde{P}^v - ep^v)^n.$$

由引理 1 可知, 方程(7)存在唯一解

$$g_\alpha^v = (\alpha I - A^v + \lambda ep^v)^{-1} f^v. \quad (10)$$

特别, 令 $\alpha = 0$, 并省略下标, 有

$$g^v = (-A^v + \lambda ep^v)^{-1} f^v. \quad (11)$$

称 $g^v = (g^v(1), g^v(2), \dots)^T$ 为性能势. 在式(8)两边左乘 p^v , 并由式(2)和(5)得

$$\lambda p^v g = \eta^v. \quad (12)$$

因此, 平均代价 Poisson 方程又可表示为

$$A^v g = -f^v + e\eta^v. \quad (13)$$

故这里的性能势与文献[4]中的定义是一致的.

当 $\alpha > 0$ 时, 由式(2), 直接验证可得

$$\{(\alpha I - A^v + \lambda ep^v)^{-1} + \lambda ep^v / [\alpha(\lambda + \alpha)]\}(\alpha I - A^v) = I. \quad (14)$$

在式(6)两边左乘

$$(\alpha I - A^v + \lambda ep^v)^{-1} + \lambda ep^v / [\alpha(\lambda + \alpha)],$$

并由式(5), (10)和(14)可得

$$\eta_\alpha^v = g_\alpha^v + \lambda e\eta^v / [\alpha(\lambda + \alpha)]. \quad (15)$$

4 最优性方程 (Optimality equation)

本节主要讨论 CMCP 的平均代价最优性方程及其在紧致行动集上解的存在性.

引理 2 对任意的 $v', v \in \Omega_s$, 有

$$\eta^{v'} - \eta^v = p^v[(f^{v'} + A^{v'} g^{v'}) - (f^v + A^v g^v)]. \quad (16)$$

证 由式(13), 有

$$\begin{aligned} e(\eta^{v'} - \eta^v) &= f^{v'} - f^v + A^{v'} g^{v'} - A^v g^v = \\ &= (f^{v'} + A^{v'} g^{v'}) - (f^v + A^v g^v) + A^v(g^{v'} - g^v). \end{aligned} \quad (17)$$

上式两边左乘 p^v , 并注意到 $p^v e = 1, p^v A^v = 0$, 即得式(16).

由引理 2, 并注意到 $p^v(i) > 0, i \in \Phi$, 可得下列最优性定理.

定理 1 $v^* \in \Omega_s$ 是可数 Markov 控制过程 (X, Φ, D, A^v, f^v) 平均代价最优平稳策略的充分必要条件为

$$f^{v^*} + A^{v^*} g^{v^*} \leq f^v + A^v g^{v^*}, v \in \Omega_s. \quad (18)$$

再由式(13), 并根据定理 1, 可直接得到下面定理.

定理 2 $v^* \in \Omega_s$ 是可数 Markov 控制过程 (X, Φ, D, A^v, f^v) 平均代价最优平稳策略的充分必要条件是满足方程

$$0 = \min_{v \in \Omega_s} \{f^v + A^v g^{v^*} - e\eta^{v^*}\}. \quad (19)$$

式(19)称为 CMCP 基于性能势的平均代价最优性方程. 为了证明最优性方程解的存在性, 我们再回到 α -折扣代价问题. 根据文献[12], 可得到下列引理.

引理 3 对任意 $\alpha > 0$, 存在一个最优平稳策略 $v^* \in \Omega_s$ 和相应的有界实向量 η_α^* , 满足折扣代价最优性方程

$$0 = \min_{v \in \Omega_s} \{f^v - (\alpha I - A^v) \eta_\alpha^*\}. \quad (20)$$

证 令 $\tilde{P}^v = A^v / \lambda + I, \beta = \lambda / (\lambda + \alpha)$, 则 $0 < \beta < 1$, 且方程(20)变为

$$\eta_\alpha^* = \min_{v \in \Omega_s} \{f^v / (\alpha + \lambda) + \beta \tilde{P}^v \eta_\alpha^*\}, \quad (21)$$

从而由文献[12]可得引理 3.

若记

$$\begin{aligned} (v(1), v(2), \dots) &= d \in \bar{D}, \\ (v^*(1), v^*(2), \dots) &= \delta^\infty \in \bar{D}, \end{aligned}$$

并取

$$\delta \in \arg \min_{d \in \bar{D}} \{f^d - (\alpha I - A^d) \eta_\alpha^*\}. \quad (22)$$

则对任意的 $v \in \Omega_s$, 有

$$0 = f^\delta - (\alpha I - A^\delta) \eta_\alpha^* \leq f^v - (\alpha I - A^v) \eta_\alpha^*.$$

定理 3 存在实数 η 和向量 g , 满足

$$0 = \min_{v \in \Omega_s} \{f^v + A^v g - e\eta\}. \quad (23)$$

证 选择一折扣因子序列 $\alpha_k \downarrow 0$. 由引理 3 可知, 对每一固定的 α_k , 存在 $\delta_k \in \bar{D}$ 和相应的一个最优策略 v_k^* 的 $\delta_k^\infty \in \bar{D}$, 使

$$\delta_k \in \arg \min_{d \in \bar{D}} \{f^d - (\alpha_k I - A^d) \eta_{\alpha_k}^*\}. \quad (24)$$

由于 $\bar{D}$ 是紧的, 故可选择 $\{\delta_k\}$ 的一个子序列 $\{\delta_{k_j}\}$ 收敛到一个 $\delta \in \bar{D}$, 且 $\{\delta_k^\infty\}$ 的相应子序列 $\{\delta_{k_j}^\infty\}$ 收敛

到 $\delta^\infty \in \bar{D}$. 为表达简洁起见, 仍将 $\{k_j\}$ 记为 $\{k\}$, 则有

$$\begin{aligned} 0 &= f^{\delta_k} - (\alpha_k I - A^{\delta_k}) \eta_{\alpha_k}^* \leq \\ &f^v - (\alpha_k I - A^v) \eta_{\alpha_k}^*, v \in \Omega_s. \end{aligned} \quad (25)$$

由于 f^v 一致有界, 即存在常数 M , 使得对任意的 $v \in \Omega_s, i \in \Phi$, 有 $|f(i, v(i))| \leq M$. 故对任意的 k , 有

$$\sum_{i \in \Phi} |p^{\delta_k}(i) f(i, v_k^*(i))| \leq M \sum_{i \in \Phi} p^{\delta_k}(i) = M.$$

由于 p^v 和 f^v 均在 $\bar{D}$ 上连续, 故有

$$\begin{aligned} \lim_{k \rightarrow \infty} \eta_{\alpha_k}^* &= \lim_{k \rightarrow \infty} \sum_{i \in \Phi} p^{\delta_k}(i) f(i, v_k^*(i)) = \\ \sum_{i \in \Phi} \lim_{k \rightarrow \infty} \eta_{\alpha_k}^*(i) f(i, v_k^*(i)) &= p^{\delta^\infty} f^{\delta^\infty} = \eta^{\delta^\infty}. \end{aligned}$$

为了得到 $\lim_{k \rightarrow \infty} g_{\alpha_k}^* = g^{\delta^\infty}$, 先来证明

$$\lim_{k \rightarrow \infty} (\alpha_k I - A^{\delta_k} + \lambda e p^{\delta_k})^{-1} = (-A^{\delta^\infty} + \lambda e p^{\delta^\infty})^{-1}.$$

为此, 记 $C_k = \alpha_k I - A^{\delta_k} + \lambda e p^{\delta_k}, B_k = (\alpha_k I - A^{\delta_k} + \lambda e p^{\delta_k})^{-1}, C = -A^{\delta^\infty} + \lambda e p^{\delta^\infty}$, 则对任意的 k , 有

$$B_k C_k = C_k B_k = I.$$

显然, $C_k \rightarrow C$, 故若记 $\lim_{k \rightarrow \infty} B_k = \bar{B}, \lim_{k \rightarrow \infty} B_k = \underline{B}$, 则有

$$\bar{B} C = C \bar{B} = I, \underline{B} C = C \underline{B} = I.$$

由于矩阵 C 可逆, 故满足方程 $BC = CB = I$ 的矩阵 B 是唯一的, 从而必有 $\bar{B} = \underline{B} = C^{-1}$. 又易验证, 对任意的 k , 有

$$(\alpha_k I - A^{\delta_k} + \lambda e p^{\delta_k})^{-1} e = \frac{e}{\lambda + \alpha_k}.$$

故若记 $(\alpha_k I - A^{\delta_k} + \lambda e p^{\delta_k})^{-1} = [D_k(i, j)]$, 则对任意的 i 及 k , 有

$$\sum_{j \in \Phi} |D_k(i, j) f(j, v_k^*(j))| \leq$$

$$M \sum_{j \in \Phi} D_k(i, j) = \frac{M}{\lambda + \alpha_k}.$$

从而, 最终可得

$$\begin{aligned} \lim_{k \rightarrow \infty} g_{\alpha_k}^* &= \lim_{k \rightarrow \infty} (\alpha_k I - A^{\delta_k} + \lambda e p^{\delta_k})^{-1} f^{\delta_k} = \\ (-A^{\delta^\infty} + \lambda e p^{\delta^\infty})^{-1} f^{\delta^\infty} &= g^{\delta^\infty}. \end{aligned}$$

将式(15)代入式(25), 并由式(2), 可得

$$0 = f^{\delta_k} - (\alpha_k I - A^{\delta_k}) g_{\alpha_k}^* - \frac{\lambda e \eta_{\alpha_k}^*}{\lambda + \alpha_k} \leq$$

$$f^v - (\alpha_k I - A^v) g_{\alpha_k}^* - \frac{\lambda e \eta_{\alpha_k}^*}{\lambda + \alpha_k}, v \in \Omega_s.$$

令 $k \rightarrow \infty (\alpha_k \rightarrow 0)$, 则对任意的 $v \in \Omega_s$, 有

$$0 = f^\delta + A^\delta g^{\delta^\infty} - e\eta^{\delta^\infty} \leq f^v + A^v g^{\delta^\infty} - e\eta^{\delta^\infty},$$

即 $(\eta^{\delta^\infty}, g^{\delta^\infty})$ 满足式(23).

5 结论(Conclusions)

基于 Markov 性能势, 讨论了一类具有可数状态空间的 Markov 控制过程的性能优化问题. 通过引入折扣 Poisson 方程, 并运用无穷小矩阵和性能势的基本性质, 在无限水平平均代价准则下, 给出了此类控制过程在紧致行动集上的最优性方程, 及其解的存在性定理. 由这些结果可给出寻找最优或 ε -最优策略的一系列迭代算法, 并且此方法还易于建立相应的并行算法, 从而加快寻优速度. 因此它为大规模实际 Markov 控制系统的性能优化提供了一条新途径, 特别可应用于 Markov 型排队网络的性能优化问题.

参考文献(References):

- [1] ARAPOSTATHIS A, BORKAR V S, FERNANDEZ-GAUCHER, et al. Discrete-time controlled Markov processes with average cost criterion: a survey [J]. *SIAM J of Control Optimization*, 1993, 31(2): 282 - 344.
- [2] RAUL Montes-de-Oca. The average cost optimality equation for Markov control processes on Borel spaces [J]. *System and Control Letters*, 1994, 22(5): 351 - 357.
- [3] SENNOT L I. Another set of conditions for average optimality in Markov control processes [J]. *Systems and Control Letters*, 1995, 23(2): 147 - 151.
- [4] 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.
- [5] 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.
- [6] CAO X R. A unified approach to Markov decision problems and per-

formance sensitivity analysis [J]. *Automatica*, 2000, 36(5): 771 - 774.

- [7] 殷保群, 周亚平, 杨孝先等. 状态相关排队网络中的性能指标灵敏度公式[J]. *控制理论与应用*, 1999, 16(2): 255 - 257.
(YIN Baoqun, ZHOU Yaping, YANG Xiaoxian, et al. Sensitivity formulas of performance in closed state-dependent queuing networks [J]. *Control Theory & Applications*, 1999, 16(2): 255 - 257.)
- [8] YIN Baoqun, ZHOU Yaping, XI Hongsheng, et al. Sensitivity formulas of performance in two-server cyclic queuing networks with phase-type distributed service times [J]. *Int Trans on Operation Research*, 1999, 6(6): 649 - 663.
- [9] 奚宏生, 唐昊, 殷保群. 连续时间 MCP 在紧致行动集上的最优策略[J]. *自动化学报*, 2003, 29(2): 206 - 211.
(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.)
- [10] CINLAR E. *Introduction to Stochastic Processes* [M]. Englewood Cliffs, NJ: Prentice-hall, 1975.
- [11] KEMENY J G, SNELL J L. Potentials for denumerable Markov chains [J]. *J of Mathematical Analysis and Applications*, 1960, 3(1): 196 - 260.
- [12] PUTERMAN M L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming* [M]. New York: Wiley, 1994.

作者简介:

殷保群 (1962—), 男, 中国科学技术大学自动化系副教授, 博士, 1985 年毕业于四川大学数学系, 主要从事随机离散事件动态系统、系统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攻读博士研究生, 2001 年 7 月毕业于青岛大学数学系计算数学及应用软件专业, 研究方向为离散时间动态系统分析、控制和应用;

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, 1977 年毕业于中国科学技术大学数学系, 长期从事鲁棒控制、离散事件动态系统及其应用等方面的研究;

周亚平 (1963—), 男, 中国科学技术大学管理科学系副教授, 博士, 1984 年毕业于中国科学技术大学自动化系, 从事经济管理系统、排队网络性能灵敏度仿真估计及优化等方面的研究.