第34卷第2期　　中 国 科 学 技 术 大 学 学 报　　Vol. 34, No. 2
2004年4月　JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA　Apr. 2004

文章编号:0253-2778(2004)02-0219-07

连续时间 Markov 控制过程的平均代价最优鲁棒控制策略*

唐　昊,韩江洪,高　隽

(合肥工业大学计算机与信息学院,安徽合肥 230009)

摘要:在 Markov 性能势基础上,研究了一类转移速率不确定但受紧集约束的遍历连续时间 Markov 控制过程(CTMCP)的鲁棒控制问题.根据系统的遍历性,平均代价 Poisson 方程的解可被看作是性能势的一种定义.在平均代价准则下,优化控制的目标是选择一个平稳策略使得系统在参数最坏取值下能获得最小无穷水平平均代价,据此论文给出了求解最优鲁棒控制策略的策略迭代(PI)算法,并详细讨论了算法的收敛性.

关键词:Markov 性能势;连续时间 Markov 控制过程;鲁棒控制策略;策略迭代

中图分类号:TP202　　**文献标识码**:A

0　引言

Markov 控制过程(MCP)或 Markov 决策过程(MDP)描述的一些序贯决策问题,在系统模型参数已知时,可通过一些数值算法来精确求解,如数值迭代算法、策略迭代算法、梯度算法、线性规划和扩展 Kalman 滤波等[1,2],特别是建立在 Markov 性能势理论基础上的一套优化方法已取得了一些理论成果[3~5].但随着科学技术的进步和社会的发展,系统结构越来越复杂,状态空间的规模也越来越大,对实际的 Markov 系统进行数学模型描述,面临两个问题:一是系统模型参数本身确定,但由于受量测等因素限制,决策者对这些参数不全知;其次是由于干扰摄动等原因,参数本身存在不确定性,例如为慢时变的.这两种不确定性因素,最终将导致状态转移概率或无穷小转移速率的不确定性.因此,传统的依赖精确模型参数的优化方法已不再适用.基于实际系统的单个样本轨道,并结合 Markov 性能势理论或神经元动态规划(NDP)的仿真优化方法,为解决系统参数确定但不全知的一类 MCP 的优化问题提供了一个有效途径[6~8].另外,尽管实际系统存在不确定性,但系统参数的取值范围一般是有界的,也是可知的,因而考虑其鲁棒控制问题亦将是一项有意义的研究工作,即考虑在系统

* **收稿日期**:2001-03-08
基金项目:合肥工业大学中青年科技创新群体计划,安徽省优秀青年科技基金(04042044)
作者简介:唐昊,男,1972年9月生,博士,副教授.研究方向:离散事件动态系统、神经元动态规划等优化理论和应用.E-mail:tangh@ustc.edu

参数最坏取值下如何采取最优控制策略问题.

文献[9,10]研究了具有不确定性转移概率的离散时间 Markov 控制过程(DTMCP)在折扣准则下的鲁棒控制问题,文献[11]考虑了有限状态空间和有限行动空间连续时间 Markov 控制过程(CTMCP)在不确定性转移速率下的平均报酬问题. 在此工作的基础上,本文从 Markov 性能势出发,利用 CTMCP 基于性能势的有关优化结果[3~5],重点研究一类具有有限状态空间和一般行动空间的、遍历的 CTMCP,在平均代价准则下的最优鲁棒控制策略求解问题.

1 问题描述和基本理论

一个 CTMCP 可用 5-元组 $X = (X(t), \Phi, D, \boldsymbol{P}^v(t), \boldsymbol{f}^v)$ 来表示,其中 $\{X(t), t \geqslant 0\}$ 是 Markov 状态过程,可在有限状态空间 $\Phi = \{1,2,\cdots,M\}$ 中取值,D 为行动空间. 确定性平稳策略定义为一个映射 $\boldsymbol{v}: \Phi \to D$,即 $v(i) \in D, \forall i \in \Phi$. 记 $\boldsymbol{v} = (v(1), v(2), \cdots, v(M))$,$\Omega_s$ 为全体平稳策略集. 系统的性能函数是把状态空间和策略空间映射到实数空间的一个映射 $\boldsymbol{f}: \Phi \times \Omega_s \to R$,$f(i, v(i))$ 表示系统在状态 i 采用行动 $v(i)$ 时单位时间付出的代价即代价率,称 $\boldsymbol{f}^v = (f(1, v(1)), \cdots, f(M, v(M)))^T$ 为 X 在策略 $\boldsymbol{v}$ 下的性能向量. X 的状态转移矩阵 $\boldsymbol{P}^v(t)$ 和无穷小转移速率矩阵 $\boldsymbol{A}^v = [a_{ij}^v]$ 满足 Kolmogorov 向前向后方程. 若 $\boldsymbol{P}^v$ 为 X 的嵌入 Markov 链的转移矩阵,则 $\boldsymbol{A}^v = \mathrm{diag}(\lambda^v(1), \cdots, \lambda^v(M)) \cdot (\boldsymbol{P}^v - I)$,其中 $\lambda^v(i) > 0$,是 Markov 过程在状态 i 采用行动 $v(i)$ 时的平均转移率,令 $\lambda^v = \max_i \{\lambda^v(i)\}$. 若系统状态过程不可约,则存在稳态分布 $\boldsymbol{\pi}^v = (\pi^v(1), \cdots, \pi^v(M))$, 满足平衡方程

$$\boldsymbol{\pi}^v \boldsymbol{e} = 1,\ \boldsymbol{A}^v \boldsymbol{e} = 0,\ \boldsymbol{\pi}^v \boldsymbol{A}^v = 0. \tag{1}$$

其中 $\boldsymbol{e} = (1,1,\cdots,1)^{\mathrm{T}}$ 是分量均为 1 的 M 维列向量. 在策略 $\boldsymbol{v} \in \Omega_s$ 下,X 的无穷时段平均代价期望值准为 $\eta^v = \lim\limits_{T\to\infty} \dfrac{1}{T} E\left\{\int_0^T f(X(t), v(X(t)))\mathrm{d}t\right\}$, 且

$$\eta^v = \boldsymbol{\pi}^v \boldsymbol{f}^v. \tag{2}$$

一个控制策略 $\boldsymbol{v}^*$ 满足 $\boldsymbol{v}^* \in \arg\min\limits_{\boldsymbol{v}\in\Omega_s} \eta^v$,便是最优的. 对任意 $\boldsymbol{v} \in \Omega_s$,若无穷小转移速率矩阵 $\boldsymbol{A}^v$ 都固定可知,则最优策略可通过平衡方程和 Poisson 方程建立策略迭代或数值迭代算法来求解[5]. 但在一些不确定性实际系统中,对应某些策略 $\boldsymbol{v}$,其无穷小转移速率矩阵 $\boldsymbol{A}^v$ 中的元素 a_{ij}^v, $i,j \in \Phi$ 可能未知或不全知,甚至是变化的. 决策者仅知道其所属范围或可能变化范围 Θ_{ij}^v, 即 $a_{ij}^v \in \Theta_{ij}^v$. 令 Θ_{ij}^v 是紧致集, 且 $\Theta^v = \{\boldsymbol{A}^v = [a_{ij}^v]: a_{ij}^v \in \Theta_{ij}^v,\ i \neq j;\ \sum_i a_{ij}^v = 0\}$. 对这种不确定情况,决策者有时需考虑最坏条件下的最优控制,即考虑在 $\boldsymbol{A}^v$ 取值最不利系统获得最佳性能值的情况下选择一策略 $\boldsymbol{v}^*$,满足

$$\boldsymbol{v}^* \in \arg\min_{\boldsymbol{v}\in\Omega_s} \max_{\boldsymbol{A}^v\in\Theta^v} \eta^v. \tag{3}$$

显然,这种求解极小极大策略的问题是一种最优鲁棒控制问题. 假设对任意给定的策略 $\boldsymbol{v}$,a_{ij}^v 可以独立选取(指 $\boldsymbol{A}^v$ 每行之间不相关),且对任意的 $\boldsymbol{A}^v \in \Theta^v$,由 $\boldsymbol{A}^v$ 决定的状态过程都是不可约的,即遍历的.

CTMCP 的广义平均代价 Poisson 方程[4]

$$(-\boldsymbol{A}^v + \lambda^v \boldsymbol{e}\boldsymbol{\pi}^v)g^v = \boldsymbol{f}^v. \tag{4}$$

根据遍历性,它有唯一解 $\boldsymbol{g}^v = (-\boldsymbol{A}^v + \lambda^v \boldsymbol{e}\boldsymbol{\pi}^v)^{-1}\boldsymbol{f}^v$. 这里,$\boldsymbol{g}^v = (g^v(1),\cdots,g^v(M))^{\mathrm{T}}$ 为 Poisson 方程定义的性能势向量,与曹希仁教授通过样本轨道直接定义的性能势概念本质相同[3,6]. 本文将基于性能势理论来研究 CTMCP 在不确定性参数情况下的鲁棒控制问题.

2　最优鲁棒控制策略

式(3)这样的鲁棒最优问题,可分为两个子问题来求解. 问题一是对给定的策略 $\boldsymbol{v} \in \Omega_s$,确定一个最不利系统获得最佳性能值的转移速率,即寻找一个 $\widehat{\boldsymbol{A}}^v \in \boldsymbol{\Theta}^v$ 满足

$$\widehat{\boldsymbol{A}}^v \in \arg\max_{\boldsymbol{A}^v \in \boldsymbol{\Theta}^v} \eta^v. \tag{6}$$

通过方程(6)建立了策略 $\boldsymbol{v}$ 和 $\widehat{\boldsymbol{A}}^v$ 的一一对应关系. 根据(1)和(2)式,可分别计算对应 $\widehat{\boldsymbol{A}}^v$ 的稳态概率分布 $\widehat{\boldsymbol{\pi}}^v$ 和平均代价 $\widehat{\eta}^v$. 同样,记与 $\widehat{\boldsymbol{A}}^v$ 对应的性能势向量为 $\widehat{\boldsymbol{g}}^v$.

问题二是根据求解问题一得到的性能值,寻求一最优鲁棒控制策略 $\boldsymbol{v}^*$,满足

$$\boldsymbol{v}^* \in \arg\min_{\boldsymbol{v} \in \Omega_s} \widehat{\eta}^v.$$

问题二的求解,其实就变成对一个 CTMCP $\widehat{X} = (\widehat{X}(t), \boldsymbol{\Phi}, D, \widehat{P}^v(t), \boldsymbol{f}^v)$ 求解最优平均代价问题,这里 $\widehat{X}(t)$ 和 $\widehat{P}^v(t)$ 同 $\widehat{\boldsymbol{A}}^v$ 相对应. 下面将给出求解上述两个子问题的迭代算法. 首先有下列引理,其证明较为简单,此处略去,可参考文献[3].

引理 1　对由平衡方程和 Poisson 方程确定的任意两个 4-元组 $(\boldsymbol{f},\boldsymbol{A},\boldsymbol{\pi},\boldsymbol{g})$ 和 $(\boldsymbol{f}',\boldsymbol{A}',\boldsymbol{\pi}',\boldsymbol{g}')$,其对应的稳态性能值 η 和 η' 都满足

$$\eta - \eta' = \boldsymbol{\pi}'[(\boldsymbol{f} + \boldsymbol{A}\boldsymbol{g}) - (\boldsymbol{f}' + \boldsymbol{A}'\boldsymbol{g})]. \tag{7}$$

2.1　求解问题一的迭代算法

首先,根据公式(1)和(4),易证 $\boldsymbol{e}\eta^v = \boldsymbol{f}^v + \boldsymbol{A}^v\boldsymbol{g}^v$,故 $\arg\max_{\boldsymbol{A}^v \in \boldsymbol{\Theta}^v} \boldsymbol{e}\eta^v = \arg\max_{\boldsymbol{A}^v \in \boldsymbol{\Theta}^v} \{\boldsymbol{f}^v + \boldsymbol{A}^v\boldsymbol{g}^v\}$,这里,运算符号"arg max"表示分别对每个分量进行计算. 其次,对一个给定的策略 $\boldsymbol{v} \in \Omega_s$,性能向量 $\boldsymbol{f}^v$ 固定不变,则有 $\arg\max_{\boldsymbol{A}^v \in \boldsymbol{\Theta}^v} \{\boldsymbol{f}^v + \boldsymbol{A}^v\boldsymbol{g}^v\} = \arg\max_{\boldsymbol{A}^v \in \boldsymbol{\Theta}^v} \boldsymbol{A}^v\boldsymbol{g}^v$. 因此,为求解问题一,对给定的策略 $\boldsymbol{v} \in \Omega_s$,我们可构造下列迭代算法来获得最不利于性能值的转移速率矩阵 $\widehat{\boldsymbol{A}}^v$.

算法一

步骤 1　任选一可行的无穷小转移速率矩阵 $\overline{\boldsymbol{A}}^v \in \boldsymbol{\Theta}^v$.

步骤 2　根据 $\overline{\boldsymbol{A}}^v$ 利用平衡方程(1),求解对应的平稳分布 $\overline{\boldsymbol{\pi}}^v$,再根据 *Poisson* 方程(4)来求解对应的性能势 $\overline{\boldsymbol{g}}^v$.

步骤 3　选择一更新的转移速率矩阵 $\widehat{\boldsymbol{A}}^v \in \boldsymbol{\Theta}^v$,满足

$$\widehat{\boldsymbol{A}}^v = \arg\max_{\boldsymbol{A}^v \in \boldsymbol{\Theta}^v} \{\boldsymbol{A}^v\overline{\boldsymbol{g}}^v\}. \tag{8}$$

步骤 4　若 $\widehat{\boldsymbol{A}}^v = \overline{\boldsymbol{A}}^v$,则停止;否则令 $\overline{\boldsymbol{A}}^v = \widehat{\boldsymbol{A}}^v$,转步骤 2.

对于算法一,我们有下列定理.

定理 1　对任意给定的策略 $\boldsymbol{v} \in \Omega_s$,算法一在有限步内停止,且停止时的无穷小转移速

率矩阵 $\overline{\boldsymbol{A}}^{v}$ 或 $\widehat{\boldsymbol{A}}^{v}$ 对应系统在策略 $\boldsymbol{v}$ 下的最大代价,即算法在有限步内就能产生满足方程(6)的移速率矩阵 $\widehat{\boldsymbol{A}}^{v}$.

证明

根据(8)式有

$$\widehat{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}^{v} \geqslant \boldsymbol{A}^{v}\overline{\boldsymbol{g}}^{v}, \quad \forall \boldsymbol{A}^{v} \in \boldsymbol{\Theta}^{v}. \tag{9}$$

其中向量运算中的关系符号"≤"表示对应分量等于或小于关系成立,符号"≥"的定义相似,另外定义符号">"表示向量间的大于等于关系成立,且至少存在一对对应分量大于关系成立,符号"<"的定义相似. 于是

$$\widehat{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}^{v} \geqslant \overline{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}^{v}. \tag{10}$$

根据遍历性,$\widehat{\boldsymbol{\pi}}^{v}(i) > 0, \forall i \in \boldsymbol{\Phi}$. 若(10)式中至少有一对分量的等号关系不成立,即 $\widehat{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}^{v} > \overline{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}$,则由引理1中的方程(7),得 $\overline{\eta}^{v} - \widehat{\eta}^{v} = \widehat{\boldsymbol{\pi}}^{v}[\boldsymbol{f}^{v} + \overline{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}^{v} - (*\boldsymbol{f}^{v} + \widehat{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}^{v})] < 0$,即 $\overline{\eta}^{v} < \widehat{\eta}^{v}$.

若(10)式中的等号成立,即 $\widehat{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}^{v} = \overline{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}^{v}$,则由引理1中的(7)式得 $\overline{\eta}^{v} = \widehat{\eta}^{v}$,且根据(9)式有

$$\overline{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}^{v} \geqslant \boldsymbol{A}^{v}\overline{\boldsymbol{g}}^{v}, \quad \forall \boldsymbol{A}^{v} \in \boldsymbol{\Theta}^{v}.$$

再由引理1得 $\overline{\eta}^{v} - \eta^{v} = \boldsymbol{\pi}^{v}[\boldsymbol{f}^{v} + \overline{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}^{v} - (\boldsymbol{f}^{v} + \boldsymbol{A}^{v}\overline{\boldsymbol{g}}^{v})] \geqslant 0$,因此

$$\overline{\eta}^{v} = \widehat{\eta}^{v} \geqslant \eta^{v} \quad \forall \boldsymbol{A}^{v} \in \boldsymbol{\Theta}^{v}.$$

即对一固定的策略 $\boldsymbol{v} \in \boldsymbol{\Omega}_s$,若 $\widehat{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}^{v} = \overline{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}^{v}$,则 $\widehat{\boldsymbol{A}}^{v}$ 或 $\overline{\boldsymbol{A}}^{v}$ 对应的平均代价是最大的.

因为对任意的 $i \in \boldsymbol{\Phi}$,a_{ij}^{v} 可以独立选取,故在方程(8)的求解中,$\boldsymbol{A}^{v}$ 中第 i 行的元素 a_{ij}^{v} 仅取决于 $\overline{g}^{v}(j) - \overline{g}^{v}(i)$ 的符号. 对任意 $j \neq i$,若 $\overline{g}^{v}(j) - \overline{g}^{v}(i) \geqslant 0$,取 $a_{ij}^{v} = \max\{x, x \in \boldsymbol{\Theta}_{ij}^{v}\}$;若 $\overline{g}^{v}(j) - \overline{g}^{v}(i) < 0$,取 $a_{ij}^{v} = \min\{x, x \in \boldsymbol{\Theta}_{ij}^{v}\}$. 另外,$\boldsymbol{A}^{v}$ 必须满足约束条件 $\boldsymbol{A}^{v}\boldsymbol{e} = 0$,故有 $M^2 - M$ 个参数待选,因此 $\boldsymbol{A}^{v}$ 共有 $2^{M\times(M-1)}$ 个不同的选择. 所以上述算法最多在 $2^{M\times(M-1)} - 1$ 步迭代内就停止. 事实上,在第二步迭代以后若有 $\widehat{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}^{v} = \overline{\boldsymbol{A}}^{v}\overline{\boldsymbol{g}}^{v}$,则必有 $\widehat{\boldsymbol{A}}^{v} = \overline{\boldsymbol{A}}^{v}$;反之亦然. 这意味着,在算法一中,对应更新的转移速率矩阵,其性能值是严格单调上升的,即转移速率矩阵每一步都朝着使性能值变坏的方向变化,直至有限步后达到最大值. 故定理得证.

2.2 求解问题二的策略迭代算法(PIA)

问题二的求解,可以引用文献[3~5]中的有关结果. 首先我们有下面的最优性定理.

引理2 $\boldsymbol{v}^{*} \in \boldsymbol{\Omega}_s$ 是满足(3)式的极小极大最优平稳策略的充分必要条件是

$$\boldsymbol{f}^{v^*} + \widehat{\boldsymbol{A}}^{v^*}\widehat{\boldsymbol{g}}^{v^*} \leqslant \boldsymbol{f}^{v} + \widehat{\boldsymbol{A}}^{v}\widehat{\boldsymbol{g}}^{v^*}, \quad \forall \boldsymbol{v} \in \boldsymbol{\Omega}_s$$

根据这样的定理,同文献[5]相似,可构造下面的策略迭代算法.

算法二

步骤1 任选一初始策略 $\boldsymbol{v}_0 \in \boldsymbol{\Omega}_s$,调用算法一,得到一个产生极大性能值的无穷小转移速率矩阵 $\widehat{\boldsymbol{A}}^{v_0}$,以及对应的性能势 $\widehat{\boldsymbol{g}}^{v_0}$,并计算 $\widehat{\eta}^{v_0}$;令 $k = 0$.

步骤2 选择一个策略 $\boldsymbol{v}_{k+1} \in \boldsymbol{\Omega}_s$,满足

$$\boldsymbol{v}_{k+1} \in \arg\min_{\boldsymbol{v}\in\Omega_s}\max_{\boldsymbol{A}^v\in\boldsymbol{\Theta}^v}\{\boldsymbol{f}^v + \boldsymbol{A}^v\hat{\boldsymbol{g}}^{v_k}\}. \tag{11}$$

步骤 3　针对策略 $\boldsymbol{v}_{k+1}$ 调用算法一,得到对应的转移速率矩阵 $\widehat{\boldsymbol{A}}^{v_{k+1}}$ 和性能势 $\hat{\boldsymbol{g}}^{v_{k+1}}$,并计算 $\hat{\eta}^{v_{k+1}}$.

步骤 4　若 $\hat{\eta}^{v_{k+1}} = \hat{\eta}^{v_k}$,则停止;否则,令 $\boldsymbol{v}_k := \boldsymbol{v}_{k+1}$, $k := k+1$, 转步骤 2.

对于算法二,我们有下列定理.

定理 2　算法二产生的策略序列 $\{\hat{\eta}^{v_k}\}$ 是严格单调下降序列,即每步产生的策略是改进的(improving);若算法停止,则得到一个极小极大最优策略,即最优鲁棒控制策略.

证明

由方程(11)得

$$\max_{\boldsymbol{A}^{v_{k+1}}\in\boldsymbol{\Theta}^{v_{k+1}}}\{\boldsymbol{f}^{v_{k+1}} + \boldsymbol{A}^{v_{k+1}}\hat{\boldsymbol{g}}^{v_k}\} \leqslant \max_{\boldsymbol{A}^v\in\boldsymbol{\Theta}^v}\{\boldsymbol{f}^v + \boldsymbol{A}^v\hat{\boldsymbol{g}}^{v_k}\}, \quad \forall\, \boldsymbol{v}\in\Omega_s.$$

所以

$$\boldsymbol{f}^{v_{k+1}} + \widehat{\boldsymbol{A}}^{v_{k+1}}\hat{\boldsymbol{g}}^{v_k} \leqslant \max_{\boldsymbol{A}^{v_{k+1}}\in\boldsymbol{\Theta}^{v_{k+1}}}\{\boldsymbol{f}^{v_{k+1}} + \boldsymbol{A}^{v_{k+1}}\hat{\boldsymbol{g}}^{v_k}\} \leqslant \max_{\boldsymbol{A}^{v_k}\in\boldsymbol{\Theta}^{v_k}}\{\boldsymbol{f}^{v_k} + \boldsymbol{A}^{v_k}\hat{\boldsymbol{g}}^{v_k}\} = \boldsymbol{f}^{v_k} + \widehat{\boldsymbol{A}}^{v_k}\hat{\boldsymbol{g}}^{v_k}.$$

若 $\boldsymbol{f}^{v_{k+1}} + \widehat{\boldsymbol{A}}^{v_{k+1}}\hat{\boldsymbol{g}}^{v_k} < \boldsymbol{f}^{v_k} + \widehat{\boldsymbol{A}}^{v_k}\hat{\boldsymbol{g}}^{v_k}$, 则根据引理 1 中的方程(7),我们有

$$\hat{\eta}^{v_k} - \hat{\eta}^{v_{k+1}} = \hat{\boldsymbol{\pi}}^{v_{k+1}}[\boldsymbol{f}^{v_k} + \widehat{\boldsymbol{A}}^{v_k}\hat{\boldsymbol{g}}^{v_k} - (\boldsymbol{f}^{v_{k+1}} + \widehat{\boldsymbol{A}}^{v_{k+1}}\hat{\boldsymbol{g}}^{v_k})] > 0,$$

即 $\hat{\eta}^{v_{k+1}} < \hat{\eta}^{v_k}$.

若 $\boldsymbol{f}^{v_{k+1}} + \boldsymbol{A}^{v_{k+1}}\hat{\boldsymbol{g}}^{v_k} = \boldsymbol{f}^{v_k} + \widehat{\boldsymbol{A}}^{v_k}\hat{\boldsymbol{g}}^{v_k}$,则 $\hat{\eta}^{v_k} - \hat{\eta}^{v_{k+1}} = \hat{\boldsymbol{\pi}}^{v_{k+1}}[\boldsymbol{f}^{v_k} + \widehat{\boldsymbol{A}}^{v_k}\hat{\boldsymbol{g}}^{v_k} - (\boldsymbol{f}^{v_{k+1}} + \boldsymbol{A}^{v_{k+1}}\hat{\boldsymbol{g}}^{v_k})] = 0$,即 $\hat{\eta}^{v_k} = \hat{\eta}^{v_{k+1}}$. 再令

$$\hat{\boldsymbol{A}}^v = \arg\max_{\boldsymbol{A}^v\in\boldsymbol{\Theta}^v}\{\boldsymbol{f}^v + \boldsymbol{A}^v\hat{\boldsymbol{g}}^{v_k}\}, \quad \forall\, \boldsymbol{v}\in\Omega_s.$$

由方程(11)得

$$\boldsymbol{f}^{v_k} + \widehat{\boldsymbol{A}}^{v_k}\hat{\boldsymbol{g}}^{v_k} = \boldsymbol{f}^{v_{k+1}} + \widehat{\boldsymbol{A}}^{v_{k+1}}\hat{\boldsymbol{g}}^{v_k} \leqslant \max_{\boldsymbol{A}^v\in\boldsymbol{\Theta}^{v_{k+1}}}\{\boldsymbol{f}^{v_{k+1}} + \boldsymbol{A}^v\hat{\boldsymbol{g}}^{v_k}\} \leqslant \max_{\boldsymbol{A}^v\in\boldsymbol{\Theta}^v}\{\boldsymbol{f}^v + \boldsymbol{A}^v\hat{\boldsymbol{g}}^{v_k}\} = \boldsymbol{f}^v + \hat{\boldsymbol{A}}^v\hat{\boldsymbol{g}}^{v_k}.$$

所以运用引理 1 中的方程(7)得 $\hat{\eta}^{v_k} - \hat{\eta}^v = \hat{\boldsymbol{\pi}}[\boldsymbol{f}^{v_k} + \widehat{\boldsymbol{A}}^{v_k}\hat{\boldsymbol{g}}^{v_k} - (\boldsymbol{f}^v + \hat{\boldsymbol{A}}^v\hat{\boldsymbol{g}}^{v_k})] \leqslant 0$, 即

$$\hat{\eta}^{v_{k+1}} = \hat{\eta}^{v_k} \leqslant \hat{\eta}^v, \quad \forall\, \boldsymbol{v}\in\Omega_s. \tag{12}$$

另外,根据算法一和定理 1,对应 $\hat{\boldsymbol{A}}^v \in \boldsymbol{\Theta}^v$,其性能值 $\hat{\eta}^v$ 满足

$$\hat{\eta}^v \leqslant \hat{\eta}^v, \quad \forall\, v\in\Omega_s. \tag{13}$$

联合公式(12)和(13)得

$$\hat{\eta}^{v_{k+1}} = \hat{\eta}^{v_k} \leqslant \hat{\eta}^v, \quad \forall\, \boldsymbol{v}\in\Omega_s.$$

所以策略 $\boldsymbol{v}_{k+1}$ 和 $\boldsymbol{v}_k$ 是鲁棒最优的.

若停止准则 $\hat{\eta}^{v_{k+1}} = \hat{\eta}^{v_k}$ 成立,即 $\boldsymbol{f}^{v_{k+1}} + \widehat{\boldsymbol{A}}^{v_{k+1}}\hat{\boldsymbol{g}}^{v_{k+1}} = \boldsymbol{f}^{v_k} + \widehat{\boldsymbol{A}}^{v_k}\hat{\boldsymbol{g}}^{v_k}$, 则根据公式(7)得

$$\hat{\eta}^{v_k} - \hat{\eta}^{v_{k+1}} = \hat{\boldsymbol{\pi}}^{v_{k+1}}[\boldsymbol{f}^{v_k} + \widehat{\boldsymbol{A}}^{v_k}\hat{\boldsymbol{g}}^{v_k} - (\boldsymbol{f}^{v_{k+1}} + \widehat{\boldsymbol{A}}^{v_{k+1}}\hat{\boldsymbol{g}}^{v_k})] = 0.$$

因为 $\hat{\boldsymbol{\pi}}^{v_{k+1}}(i) > 0, \forall\, i\in\Phi$,则有 $\boldsymbol{f}^{v_{k+1}} + \widehat{\boldsymbol{A}}^{v_{k+1}}\hat{\boldsymbol{g}}^{v_k} = \boldsymbol{f}^{v_k} + \widehat{\boldsymbol{A}}^{v_k}\hat{\boldsymbol{g}}^{v_k}$ 成立,即条件 $\hat{\eta}^{v_k} = \hat{\eta}^{v_{k+1}}$ 和 $\boldsymbol{f}^{v_{k+1}} + \widehat{\boldsymbol{A}}^{v_{k+1}}\hat{\boldsymbol{g}}^{v_k} = \boldsymbol{f}^{v_k} + \widehat{\boldsymbol{A}}^{v_k}\hat{\boldsymbol{g}}^{v_k}$ 是等价的. 这意味着根据策略迭代算法产生的序列 $\{\boldsymbol{v}_k\}$ 对应的平均

代价序列 $\{\hat{\eta}^{v_k}\}$ 是严格单调的,若停止便得到最优鲁棒控制策略. 于是定理得证.

定理 3

(Ⅰ)若 D 是有限集,则上述策略迭代算法在有限步内停止.

(Ⅱ)若 D 是一般行动集,且 a_{ij}^{v}, $i,j \in \Phi$ 的可能变化范围 Θ_{ij}^{v} 不随策略 $\boldsymbol{v}$ 的变化而变化,即

$$\Theta_{ij}^{v} = \Theta_{ij}, \quad \forall i,j \in \Phi, \ \boldsymbol{v} \in \Omega_s.$$

则上述策略迭代算法在两步内就停止.

证明

(1)若 D 是有限集,则策略集合 Ω_s 是有限集,又因为 $\{\hat{\eta}^{v_k}\}$ 是严格单调下降序列,所以在算法二中,至多有限次,即 $|\Omega_s| = |D|^M$ 次迭代后便有 $\boldsymbol{v}_{k+1} = \boldsymbol{v}_k$,此时条件 $\hat{\eta}^{v_{k+1}} = \hat{\eta}^{v_k}$ 和 $\boldsymbol{f}^{v_{k+1}} + \hat{\boldsymbol{A}}^{v_{k+1}}\hat{\boldsymbol{g}}^{v_k} = \boldsymbol{f}^{v_k} + \hat{\boldsymbol{A}}^{v_k}\hat{\boldsymbol{g}}^{v_k}$ 显然成立,所以算法在有限步内停止.

(2)在 D 是一般行动集的情况下,若对 $\forall \boldsymbol{v} \in \Omega_s$,都有 $\Theta_{ij}^{v} = \Theta_{ij}$, $\forall i,j \in \Phi$,则记 $\Theta = \{\boldsymbol{A} = [a_{ij}] : a_{ij} \in \Theta_{ij}, \ i \neq j; \ \sum_i a_{ij} = 0\}$,即 $\Theta = \Theta^{v}$. 于是有

$$\min_{\boldsymbol{v} \in \Omega_s} \max_{\boldsymbol{A}^{v} \in \Theta^{v}} \{\boldsymbol{f}^{v} + \boldsymbol{A}^{v}\hat{\boldsymbol{g}}^{v_k}\} = \min_{v \in \Omega_s} \max_{A \in \Theta} \{\boldsymbol{f}^{v} + \boldsymbol{A}\hat{\boldsymbol{g}}^{v_k}\} = \min_{\boldsymbol{v} \in \Omega_s} \boldsymbol{f}^{v} + \max_{A \in \Theta} \{\boldsymbol{A}\hat{\boldsymbol{g}}^{v_k}\}.$$

这意味着,公式(11)的求解与 $\hat{\boldsymbol{g}}^{v_k}$ 无关,只与性能函数 $\boldsymbol{f}^{v}$ 有关. 所以

$$v_{k+1} = \arg \min_{\boldsymbol{v} \in \Omega_s} \max_{\boldsymbol{A}^{v} \in \Theta^{v}} \{\boldsymbol{f}^{v} + \boldsymbol{A}^{v}\hat{\boldsymbol{g}}^{v_k}\} = \arg \min_{\boldsymbol{v} \in \Omega_s} \boldsymbol{f}^{v}.$$

因此,算法二最多在第二次迭代后便有 $\boldsymbol{v}_1 = \boldsymbol{v}_2$ 成立,显然 $\hat{\eta}^{v_1} = \hat{\eta}^{v_2}$,算法停止.

3 结论

由于 Markov 模型可用来描述实际生活中的很大一类离散事件动态系统(DEDS),本文关于最优鲁棒控制策略问题的研究结果将为改进这类系统的设计和改善其运行性能提供一定的理论依据,有利于用来提高系统的管理水平和生产效率. 另外,文中有关结果对折扣准则下的鲁棒控制问题也也成立,并能拓展到半 Markov 控制过程(SMCP)描述的实际系统.

参 考 文 献

[1] Puterman M L. Markov decision processes: discrete stochastic dynamic programming [M]. New York: Wiley, 1994.

[2] Bertsekas D P, Tsitsiklis J N. Neuro-Dynamic Programming[M]. Belmont, MA: Athena Scientific, 1996.

[3] 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.

[4] Xi H S, Tang H, Yin B Q. Optimal policies for a continuous time MCP with compact action set[J]. Acta Automatica Sinica, 2003, 29(2): 206-211.

[5] 唐昊,奚宏生,殷保群. Markov 控制过程在紧致行动集上的迭代优化算法[J]. 控制与决策,2003, 18(3): 267-271.

[6] Cao X R, Wan Y W. Algorithms for sensitivity analysis of Markov system through potentials and perturbation realization[J]. IEEE Trans. on Control Systems Technology, 1998, 6(4):

482-494.

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

[8] 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.

[9] Satia J K, Lave R E. Markovian decision processes with uncertain transition probabilities[J]. Operations Research, 1973, 21(3): 728-740.

[10] White C C, Eldeib H K. Markov decision processes with imprecise transition probabilities[J]. Operations Research, 1994, 42(4): 739-749.

[11] Kalyanasundaram S, Chong E K P, Shroff N B. Markov decision processes with uncertain transition rates: sensitivity and robust control [A]. Proceedings of the 41th IEEE Conference on Decision and Control[C]. Nevada: Las Vegas, 2002, 4: 3 799-3 804.

Optimal Robust Control Policy for Continuous-time Markov Control Processes With Average-Cost Criteria

TANG Hao, HAN Jiang-hong, GAO Jun

(*School of Computer and Information, Hefei University of Technology, Heifei* 230009, *China*)

Abstract: Motivated by the needs of optimization and control of practical engineering systems with uncertain parameters, we considered, through the Markov performance potential theory, the robust control problems for a class of continuous-time Markov control processes with uncertain transition rates that are constrained on compact sets. By ergodic property of the processes, the solution of the average-cost Poisson equation can be viewed as a definition for the concept of Markov performance potential. Under average-cost criteria, our goal is to obtain a stationary policy that generates the minimal infinite horizon average cost under the worst choice of the system parameters. Therefore, we developed a policy iteration algorithm for generating an optimal robust control policy, and discussed in detail the convergence of the proposed algorithm.

Key words: Markov performance potentials; continuous time Markov control processes; robust control policy; policy iteration