第 26 卷第 2 期
2009 年 2 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 26 No. 2
Feb. 2009

文章编号: 1000−8152(2009)02−0139−06

受控排队系统的平均最优与约束平均最优

张兰兰, 郭先平

(1. 南方医科大学 公共卫生与热带医学学院, 广东 广州 510515; 2. 中山大学 数学与计算科学学院, 广东 广州 510275)

摘要: 根据连续时间马尔可夫决策过程的平均准则, 给出了一种特殊的马尔可夫决策过程−受控排队系统平均最优以及约束最优的新条件. 这个新条件仅使用模型的初始数据, 但利用了生灭过程的遍历性理论. 可以证明受控排队系统存在平均最优平稳策略与约束平均最优策略.

关键词: 连续时间马尔可夫决策过程; 平均准则; 受控排队系统; 平均最优平稳策略; 约束平均最优策略

中图分类号: O232　　**文献标识码**: A

Average optimality and constrained average optimality for controlled queuing systems

ZHANG Lan-lan, GUO Xian-ping

(1. Department of Biostatistics School of Public Health and Tropical Medicine, Southern Medical University, Guangzhou Guangdong 510515, China;
2. The School of Mathematics and Computational Science, Sun Yat-sen University, Guangzhou Guangdong 510275, China)

Abstract: For a special Markov decision process based on the continuous-time Markov decision processes with the average criterion, a new set of conditions is proposed for both the optimality and constrained optimality for a controlled queuing system. These conditions only employ the initial data of the controlled system, but make use of the ergodicity of a birth and death process. By using the Lagrange multipliers approach, the existence of an average optimal stationary policy and a constrained average-optimal policy can be confirmed.

Key words: continuous-time Markov decision processes; average criterion; controlled queuing systems; average optimal stationary policy; constrained average optimal policy

1　引言(Introduction)

人们对连续时间马尔可夫决策过程(continuous-time Markov decision processes, 简记为CTMDP)的平均最优策略以及约束平均最优策略存在性的研究已取得了许多令人满意的结果. 如文献[1∼3]分别在转移率无界,报酬率有界与转移率一致有界, 报酬率无界的情形下给出了平均最优策略存在的条件, 文献[4,5]则在转移率与报酬率均无界的情形下给出了约束平均最优策略存在的条件. 本文在转移率一致有界, 报酬与费用率有界的情况下, 给出了一种特殊的CTMDP −受控排队系统平均最优以及约束平均最优的新条件.

一般使得平均最优策略存在的假设可由两种方法构造: 一种是利用模型中给出的初始数据来构造, 如文献[1,2,4]中的漂移条件; 另一种则是利用非初始数据来构造, 如文献[3]中直接将假设建立在相对值函数上. 本文则是利用文献[6]中的遍历性理论来建立一致几何遍历的条件(见假设5), 从而可保证相对值函数有界(见引理2). 由于这个新条件仅建立在模型的初始数据之上, 而且有关排队系统收敛率的估计已有了相当完善的结果, 如文献[6]中针对生灭过程利用微分方程的理论得到了其指数收敛率与时间无关的上界; 文献[7]中利用谱间隙对马氏链的指数收敛率进行了估计; 文献[8]则概括了生灭过程的收敛率. 因此相对于文献[3]中的假设来说, 本文假设的验证要简单的多(见注3). 在转移率一致有界, 报酬与费用率有界的情况下, 利用本文给出的新条件, 加上连续紧与不可约的假设, 可以证明受控排队系统平均最优策略的存在性.

另一方面, 约束最优策略的存在性证明通常是

收稿日期: 2007−07−12; 收修改稿日期: 2008−05−16.
基金项目: 国家自然科学基金资助项目(60874004); 教育部博士点基金资助课题(20050558022).

在最优性的基础上进行的. 如在文献[4,9]中, 均分别在已有的平均最优以及折扣最优的假设下, 加上了一个常用的约束最优的假设, 再利用针对单约束MDP的Lagrange算子, 就可以得到约束最优策略的存在性.

由于排队系统是由服务率与到达率来控制的, 所以最后就N个服务设施的排队系统, 针对其服务到达的同时控制给出一个例子, 能验证这个例子能满足本文的所有假设, 但却很难满足文献[1,2,4]中的漂移条件(见注4).

2 受控排队系统(Controlled queueing systems)

一个CTMDP的模型是具有如下意义的五重组:

$$\{S, A, (A(i), i \in S), q(j|i,a), r(i,a)\},$$

其中: $S = \{0, 1, 2, \cdots\}$是可数状态空间, A为行动空间, $\mathcal{B}(A)$为其上的Borel $\sigma-$代数, $A(i)(\subset \mathcal{B}(\mathcal{A}), i \in S)$为系统处于状态$i$时的允许行动集, 令$K := \{(i,a)|i \in S, a \in A(i)\}$, $q(j|i,a)$是转移函数, 满足对所有的$(i,a) \in K$且$i \neq j$, $q(j|i,a) \geqslant 0$, 且$q(j|i,a)$是保守的和稳定的, 即

$$\sum_{j\in S} q(j|i,a) = 0, \quad \sup_{a\in A(i)} [-q(i|i,a)] < \infty, \qquad (1)$$

并且对每个固定的$i, j \in S$, $q(j|i,a)$是$A(i)$上的可测函数. $r(i,a)$是给定的K上的报酬函数, 对每个固定的$i \in S$, $r(i,a)$为$A(i)$上的可测函数.

如果在CTMDP模型的基础上, 加上一个费用函数$c(i,a)$(对任意固定的$i \in S$, $c(i,a)$为$A(i)$上的可测函数), 及约束常数$\beta(> 0)$, 则可得到带约束的CTMDP:

$$\{S, A, (A(i), i \in S), q(j|i,a), r(i,a), c(i,a), \beta\}.$$

定义 1 称一族随机核$\pi := (\pi_t, t \geqslant 0)$为随机马氏策略, 若

1) 对每个$i \in S$及$t \geqslant 0$, $\pi_t(\cdot|i)$是$\mathcal{B}(A(i))$上的概率测度,且

2) 对每个$i \in S$及$B \in \mathcal{B}(A)$, $\pi_t(B|i)$作为t的函数是Lebesgue可测的. 如果对所有的$i \in S$及$t \geqslant 0$, $\pi_t \equiv \pi(\cdot|i)$,则称$\pi = (\pi(\cdot|i), i \in S)$是随机平稳的. 记$F = \{f : f(i) \in A(i), i \in S\}$为决策函数集. 如果一个随机平稳策略$\pi = (\pi(\cdot|i), i \in S)$是退化分布, 即存在$f \in F$, 使得对所有的$i \in S$, $\pi(\{f(i)\}) = 1$, 则称π是一个平稳策略, 此时π由f决定, 所以可视F为平稳策略集. 分别用Π和Π_s表示随机马氏策略集和随机平稳马氏策略集.

对K上的任意一个实值可测函数$u(i,a)$以及策略每个$\pi = (\pi_t, t \geqslant 0) \in \Pi$及$t \geqslant 0$, 令

$$u(i, \pi_t) := \int_{A(i)} u(i,a)\pi_t(\mathrm{d}a|i), \ \forall t \geqslant 0. \qquad (2)$$

特别地, 当$\pi = f(\in F)$时, 记$u(i,\pi)$为$u(i,f)$. 考虑转移率矩阵$Q(\pi_t) := [q(j|i,\pi_t)]$, 由式(1)可知$Q(\pi_t)$也是保守稳定的, 以$Q(\pi_t)$为转移率矩阵的转移函数总是存在的, 称其为$Q$-过程. 这里考虑一种特殊的马尔可夫决策过程—受控的排队系统.

定义 2 受控的排队系统是一个特殊的CTMDP, 其转移率满足: 对任意的$a \in A(i)$,

$$q(0|0,a) = -q(0|0,a) = \lambda_0(a).$$

当$i \geqslant 1$时,

$$q(j|i,a) := \begin{cases} \mu_i(a), & \text{若 } j = i-1, \\ -\mu_i(a) - \lambda_i(a), & \text{若 } j = i, \\ \lambda_i(a), & \text{若 } j = i+1, \\ 0, & \text{其他.} \end{cases}$$

其中: 对任意的$i \geqslant 1$, 服务率$\mu_i(a)$与到达率$\lambda_i(a)$为$S \times A(i)$到$[0, \infty)$上的可测函数.

特别地, 当$\pi = f \in F$时, 令

$$q(j|i,f) := \begin{cases} \mu_i^f, & \text{若 } j = i-1, \\ -\mu_i^f - \lambda_i^f, & \text{若 } j = i, \\ \lambda_i^f, & \text{若 } j = i+1, \\ 0, & \text{其他.} \end{cases}$$

其中: 对任意的$i \geqslant 1$, $\mu_i^f := \mu_i(f(i))$, $\lambda_i^f := \lambda_i(f(i))$.

本文的讨论基于下列条件:

假设 1 a) 转移率是一致有界的, 即

$$\sup_{i\in S, a\in A(i)} (\mu_i(a) + \lambda_i(a)) < \infty.$$

b) 对每一个$i \in S$, $A(i)$是紧集.

c) 对任意的$i \in S$, $r(i,a)$, $c(i,a)$, $\lambda_i(a)$及$\mu_i(a)$关于$a \in A(i)$连续.

d) 存在常数$M_1 \geqslant 0$及$M_2 \geqslant 0$, 使得

$$\sup_{(i,a)\in K} |r(i,a)| \leqslant M_1, \quad \sup_{(i,a)\in K} c(i,a) \leqslant M_2,$$

且对所有的$(i,a) \in K$, $c(i,a) \geqslant 0$.

e) 对任意的$i \geqslant 1$与$a \in A(i)$,

$\mu_i(a) > 0$, $\lambda_i(a) > 0$, $\lambda_0(a) > 0 (a \in A(0))$, 且$\sum_{j\geqslant 1} \dfrac{\lambda_0^f \lambda_1^f \cdots \lambda_{j-1}^f}{\mu_1^f \mu_2^f \cdots \mu_j^f} < \infty$.

注 1 1) 若a)成立, 则矩阵$Q = [q(j|i,a)]$对应的$Q-$过程是正则的(见文献[1]), 用$\{x(t), t \geqslant 0\}$表示对应的状态过程, 用$p(s,i,t,j,\pi)$表示相应正则的转移函数. 特别地, 对任意的$f \in F$, 令$p_t^f(i,j) := p(0,i,t,j,f)$. 对

给定的S上的初始分布ν, 分别用P_ν^π与E_ν^π表示对应的概率测度与期望算子. 如果初始分布ν集中于某个状态$i \in S$(即$\nu(\{i\}) = 1$, 且对所有的$j \neq i$, 有$\nu(\{j\}) = 0$), 则将P_ν^π与E_ν^π分别记为P_i^π与E_i^π.

2) b)和c)就是常用的连续紧性假设, 如在文献[4,5]中均有用到. d)是为了保证平均准则(定义3)的定义有意义. 另一方面, 结合a)与e), 由文献[10]中的命题5.4.1知受控生灭系统$\{x(t), t \geqslant 0\}$有一个唯一不变的概率测度θ_f, 满足对任意的$i, j \in S$, 有$\theta_f(j) = \lim\limits_{t\to\infty} p_f^t(i, j)$.

对任意的$\pi \in \Pi$, 平均报酬与费用准则定义如下:

$$\bar{V}_r(\pi, \nu) := \lim_{T\to\infty} \sup \frac{1}{T} E_\nu^\pi [\int_0^T r(x(t), \pi_t) \mathrm{d}t], \quad (3)$$

$$\bar{V}_c(\pi, \nu) := \lim_{T\to\infty} \sup \frac{1}{T} E_\nu^\pi [\int_0^T c(x(t), \pi_t) \mathrm{d}t]. \quad (4)$$

报酬最优值函数为

$$\bar{V}_r^*(\nu) := \sup_{\pi\in\Pi} \bar{V}_r(\pi, \nu). \quad (5)$$

定义 3　1) 如果策略$\pi^* \in \Pi$满足

$$\bar{V}_r(\pi^*, \nu) = \bar{V}_r^*(\nu),$$

则称π^*为平均最优策略.

2) 对给定约束因子β, 称

$$U := \{\pi \in \Pi : \bar{V}_c(\pi, \nu) < \beta\}$$

为约束策略集. 如果策略$\pi^* \in U$满足

$$\bar{V}_r(\pi^*, \nu) = \sup_{\pi\in U} \bar{V}_r(\pi, \nu),$$

则称π^*为约束平均最优策略.

为了讨论需要, 定义如下:

定义 4　对每一个折扣因子$0 < \alpha < 1$, 固定一个任意状态$i_0 \in S$, 相对值函数$u_\alpha(i)$定义为

$$u_\alpha(i) := V_\alpha^*(i) - V_\alpha^*(i_0), \quad (6)$$

其中$V_\alpha^*(i) := \sup\limits_{\pi\in\Pi} V_\alpha(\pi, i)$是$\alpha-$折扣最优值函数,

$$V_\alpha(\pi, i) := E_i^\pi \int_0^\infty \mathrm{e}^{-\alpha t} r(x(t), \pi_t) \mathrm{d}t. \quad (7)$$

称一个策略$\pi^* \in \Pi$是$\alpha-$折扣最优的, 如果对所有的$i \in S$,有$V_\alpha(\pi^*, i) = V_\alpha^*(i)$.

对任意定义于S上的实值函数u, 定义范数$\|u\| := \sup\limits_{i\in S} |u(i)|$以及Banach空间$B(S) := \{u : \|u\| < \infty\}$.

假设 2　对任意的$f \in F$, 受控生灭系统的状态过程$\{x(t), t \geqslant 0\}$是一致几何遍历的, 即存在不变概率测度θ_f, 使得对所有的$i \in S$, $u \in B(S)$及$t \geqslant 0$, 有

$$\sup_{f\in F} |E_i^f u(x(t)) - \theta_f(u)| \leqslant M\mathrm{e}^{-Rt}\|u\|,$$

其中$R > 0$与$M > 0$是不依赖于f的常数, 且$\theta_f(u) := \sum\limits_{i\in S} \theta_f(i)u(i)$.

注 2　假设2是用以证明平均最优策略存在性的常用假设之一, 在第3节中将用一个新的条件(假设4)来验证它. 在假设1与2的保证下, 可以证明平均最优策略的存在性. 而下面的假设3到是为了进一步得到约束平均最优策略.

假设 3　$U_0 = \{\pi \in \Pi : \bar{V}_c(\pi, \mu) < \beta\}$非空.

定理 1　如果假设1与2成立, 则

a) 存在一个常数g_r^*, 函数$u_r^* \in B(S)$及平稳策略$f^* \in F$, 满足平均报酬最优方程, 即对任意的$i \in S$,

$$g_r^* = \sup_{a\in A(i)} \{r(i, a) + \sum_{j\in S} q(j|i, a)u_r^*(j)\} = r(i, f^*(i)) + \sum_{j\in S} q(j|i, f^*(i))u_r^*(j). \quad (8)$$

b) g_r^*是平均报酬最优值函数, 即

$$g_r^* = \bar{V}_r^*(i),\ \forall i \in S.$$

c) 一个策略$f^* \in F$是平均最优策略, 当且仅当f^*满足(8).

d) 若假设3也成立，则必存在一个约束最优策略,或者是平稳的, 或者是仅于一个状态在两个行动之间随机取值的随机平稳策略. 即: 存在两个平稳策略f^1, f^2, 状态$i^* \in S$, $p \in [0, 1]$, 使得对所有的$i \neq i^*$, 有$f^1(i) = f^2(i)$, 且随机平稳策略$\pi^p(\cdot|i)$:

$$\pi^p(a|i) = \begin{cases} p, & \text{若 } a = f^1(i^*), \\ 1-p, & \text{若 } a = f^2(i^*), \\ 1, & \text{若 } a = f^1(i) = f^2(i),\ i \neq i^* \end{cases}$$

是约束最优的(证明见第3节).

3　平均最优的新条件(New conditions for average optimality)

在这一节中, 首先给出一个相比之下较容易验证的新条件, 结合假设1可以验证假设2成立.

假设 4　存在一个正数序列$\{d_i, i \geqslant 0\}$及常数$\delta > 0$, 满足

a) $\inf\limits_{i\geqslant 0} d_i = d > 0$;

b) 对所有的$i \in S$,

$$\frac{d_{i+1}}{d_i}\lambda_{i+1}(a) + \frac{d_{i-1}}{d_i}\mu_i(a) - \lambda_i(a) - \mu_{i+1}(a) \leqslant -\delta,\ d_{-1} \equiv 0.$$

c) 常数$L := \sum\limits_{j\geqslant 1}\alpha_j \sup\limits_{f\in F}\dfrac{\lambda_0^f\lambda_1^f\cdots\lambda_{j-1}^f}{\mu_1^f\mu_2^f\cdots\mu_j^f} < \infty$, 其中$\alpha_j := \sum\limits_{i=0}^{j-1} d_i$.

注 3 假设4是新的, 且它仅建立在模型所给的初始数据的基础上, 能用以证明受控排队系统是一致几何遍历的(见引理1), 并且还可以进一步证明相对值函数$u_\alpha(i)$有界(见引理2). 显然, 相对于文献[3]中直接假设$u_\alpha(i)$关于折扣因子α一致有界来说, 本文假设4的验证要简单得多. 因为$u_\alpha(i)$是由α–折扣最优值函数定义出的, 但α–折扣最优值函数本来就很难得到, 直接验证$u_\alpha(i)$的有界性自然不太容易.

另外在连续时间马尔可夫决策过程中还常用另一条件:

假设 5 (漂移条件) 存在S上的非降函数$\omega \geqslant 0$, 常数$C > 0$, $b > 0$, 使得对所有的$i \in S$与$a \in A(i)$,

$$\sum_{j\in S} q(j|i,a)\omega(j) \leqslant -C\omega(i) + bI_{\{0\}}(i). \tag{9}$$

假设5也是用以证明一致几何遍历性的条件之一(如在文献[1,2]中), 但它很难满足本文在第4节中所给出的例子.

引理 1 若假设1与4成立, 则受控排队系统是一致几何遍历的, 即假设2成立, 且其中的$M = \dfrac{4L}{d}$, $R = \delta$.

证 由文献[6]中的定理1可知对任意的$f \in F$及$t \geqslant 0$, 存在一个不变概率测度θ_f, 使得

$$\sum_{j\in S}|p_t^f(i,j) - \theta_f(j)| \leqslant \frac{4}{d}\mathrm{e}^{-\delta t}\sum_{j\geqslant 1}\alpha_j\theta_f(j),$$

且其中:

$$\theta_f(0) = (1 + \sum_{j=1}^{\infty}\frac{\lambda_0^f\lambda_1^f\cdots\lambda_{j-1}^f}{\mu_1^f\mu_2^f\cdots\mu_j^f})^{-1},$$

$$\theta_f(j) = \frac{\lambda_0^f\lambda_1^f\cdots\lambda_{j-1}^f}{\mu_1^f\mu_2^f\cdots\mu_j^f}\theta_f(0)(j \geqslant 1),$$

因此有

$$\sum_{j\in S}|p_t^f(i,j) - \theta_f(j)| \leqslant \frac{4}{d}\mathrm{e}^{-\delta t}L,$$

从而

$$\sup_{f\in F}|E_i^f u(x(t)) - \theta_f(u)| \leqslant \frac{4L}{d}\mathrm{e}^{-\delta t}\|u\|.$$

引理 2 若假设1与2成立, 则对任意的$i \in S$及$0 < \alpha < 1$, 有

$$|u_\alpha(i)| \leqslant \frac{2M_1M}{R}. \tag{10}$$

证 由引理1, 文献[11]中的定理3.2以及式(6)知

$$|u_\alpha(i)| \leqslant M_1M\int_0^\infty 2\mathrm{e}^{-(\alpha+R)t}\mathrm{d}t \leqslant \frac{2M_1M}{R}.$$

定理1的证明 由引理2知文献[1]中的引理4.1成立. 由假设1可得$|V_\alpha(\pi,i)| \leqslant \dfrac{M_1}{\alpha}$. 从而仿照文献[1]中定理4.1的证明可知本文的定理1-a)1-b)成立, 且满足式(8)的策略$f^* \in F$是平均最优策略. 现只需证明如果f^*是平均最优策略, 则f^*满足(8). 反设f^*不满足式(8), 则存在$i_0 \in S$及常数$\rho > 0$(可能依赖于i_0与f^*), 使得对所有的$i \in S$,

$$g_r^* \geqslant r(i,f^*) + \rho\delta_{ii_0} + \sum_{j\in S} q(j|i_0,f^*)u_r^*(j). \tag{11}$$

其中$\delta_{ii_0} = \begin{cases} 1, \text{若} i = i_0, \\ 0, \text{若} i \neq i_0. \end{cases}$

另一方面, 由文献[1]中的(4.2)及引理5.1知

$$g_r^* = \bar{V}_r(f^*,i) = g(f^*) := \sum_{j\in S} r(j,f^*(j))\theta_{f^*}(j).$$

由假设1-e)知, 对每个$j \in S$, 有$\theta_{f^*}(j) > 0$, 因此由式(11), 类似于文献[1]中(4.12)的证明可得

$$g_r^* \geqslant \sum_{j\in S}[r(j,f^*(j)) + \rho\delta_{ii_0}]\theta_{f^*}(j) > g(f^*)$$

与f^*为平均最优策略矛盾.

最后证明d): 由假设1-d)知对每个$\pi \in \Pi$及$i \in S$, 有$|\bar{V}_r(\pi,i)| + |\bar{V}_c(\pi,i)| \leqslant M_1 + M_2$, 仿照 [4]中的引理3.2的证明可知$\bar{V}_r(f,\nu)$与$\bar{V}_c(f,\nu)$关于$f \in F$连续.

下面需要引入一个拉格朗日乘子$\lambda \geqslant 0$.

对每个$i \in S$及$a \in A(i)$, 令

$$b^\lambda(i,a) := r(i,a) - \lambda c(i,a). \tag{12}$$

可将$b^\lambda(i,a)$看作新的报酬率, 则类似地可以定义$b^\lambda(\pi,i)$, $\bar{V}_b^\lambda(\pi,i)$, $\bar{V}_b^\lambda(\pi,\nu)$以及$\bar{V}_b^{*\lambda}(i)$, 且定理1中的a)b)c)仍然成立. 因此由a)知, $\bar{V}_b^{*\lambda}(i)$是一个常数, 用$g_b^{*\lambda}$表示, 且存在一个函数$u_b^\lambda \in B(S)$使得

$$g_b^{*\lambda} = \sup_{a\in A(i)}\{b^\lambda(i,a) + \sum_{j\in S} q(j|i,a)u_b^\lambda(j)\},\ \forall i \in S, \tag{13}$$

对每个$i \in S$, 式(13)的上界可以达到, 即集合

$$A_\lambda^*(i) := \{a \in A(i)|g_b^{*\lambda} = b^\lambda(i,a) + \sum_{j\in S} q(j|i,a)u_b^\lambda(j)\} \tag{14}$$

是非空的, 因此集合

$$F_\lambda^* := \{f \in F : f(i) \in A_\lambda^*(i),\ \forall i \in S\}\text{以及}$$

$$\Pi_s^\lambda := \{\pi \in \Pi_s : \pi(A_\lambda^*(i)|i) = 1,\ \forall i \in S\}$$

也是非空的, 且由文献[4]中的引理3.3知Π_s^λ是凸集.

现对每个$\lambda>0$, 取定一个策略$f^\lambda\in F_\lambda^*$, 分别用$\bar V_r(\lambda)$, $\bar V_c(\lambda)$ 及 $\bar V_b(\lambda)$ 表示 $\bar V_r(f^\lambda,\nu)$, $\bar V_c(f^\lambda,\nu)$ 与 $\bar V_b^\lambda(f^\lambda,\nu)$, 则仿照文献[4]中引理3.4的证明知它们关于 $\lambda\in[0,\infty]$ 是非增的. 由定理 1-b) 及式(14)知 $\bar V_b(\lambda)=g_b^{*\lambda}$. 仿照文献[4]中引理3.5的证明知$\bar V_b(\lambda)$关于$\lambda\in[0,\infty)$连续. 若存在序列$\{\lambda_k\}\to\lambda$且$f^{\lambda_k}\in F_{\lambda_k}^*$, 满足$\lim\limits_{k\to\infty} f^{\lambda_k}=f$, 则仿照文献[4]中引理3.6的证明即知$f\in F_\lambda^*$. 若存在$\lambda_0\geqslant 0$及$\pi^*\in\Pi_s$, 满足$\bar V_c(\pi^*,\nu)=\beta$且$\bar V_b^{\lambda_0}(\pi^*,\nu)=g_b^{*\lambda_0}$, 同文献[4]中引理4.1的证明可知策略$\pi^*$是约束平均最优的.

最后仿照文献[4]中定理2.1的证明知d)成立.

4　例子(An example)

考虑一个排队系统, 设该系统中服务台的数目为$N(N\geqslant 1)$, 状态i表示每个时刻$t\geqslant 0$时的顾客数, λ与μ分别为固定的到达率和服务率. $h_1(a)$为由决策者控制的到达率, $h_2(a)$为由决策者控制的服务率, 当系统状态$i\in S:=\{0,1,\cdots\}$时, 决策者从给定的集合$[0,a^*]\subset A$中选择一个行动a, 使得到达参数增加的同时($h_1(a)\geqslant 0$), 相应的服务参数也增加($h_2(a)\geqslant 0$). 这些行动可带来一个单位时间里的费用($c(a)\geqslant 0$). $p\min\{i,N\}$为系统处于状态i时, 决策者获得的报酬(p为单位顾客带来的固定报酬). 笔者知这是一个受控排队系统, 对应的转移率, 报酬与费用函数分别定义如下:

当$i=0$时, 对任意的$a\in A(0)$:

$$q(1|0,a)=-q(0|0,a)=\lambda\geqslant 0; \tag{15}$$

当$i\geqslant 1$时, 对任意的$a\in A(i)$:

$$q(j|i,a):=\begin{cases}\mu\min\{i,N\}+h_2(a), & j=i-1,\\ -\mu\min\{i,N\}-h_2(a)-\lambda-h_1(a), & j=i,\\ \lambda+h_1(a), & j=i+1,\\ 0, & \text{其他},\end{cases}$$

$$r(i,a):=p\min\{i,N\}-c(a),\ c(i,a):=c(a). \tag{16}$$

对于上述排队系统，假设

E1) 当$N=1$时, $\mu_*>\lambda^*>0$; 当$N\geqslant 2$时, $(N-1)\mu>\lambda^*>0$, 且$h_1(a)\geqslant 2h_2(a)$. 其中

$$\lambda^*:=\sup_{a\in[0,a^*]}\{\lambda+h_1(a)\},\ \mu_*:=\inf_{a\in[0,a^*]}\{\mu+h_2(a)\}.$$

E2) $h_1(a)$, $h_2(a)$与$c(a)$关于$a\in[0,a^*]$连续且有界.

E3) $\beta>c(a^*)$ (β是给定的约束常数).

命题 1　当E1)E2)与E3)成立时, 本例所给出的排队系统满足假设1,2与3，因此由定理1和定理2知平均最优平稳策略与约束最优策略均存在.

证　由E2)知

$$\sup_{i\in S,a\in[0,a^*]}(\mu_i(a)+\lambda_i(a))\leqslant N\mu+\lambda+\sup_{a\in[0,a^*]}|h_1(a)|+\sup_{a\in[0,a^*]}|h_2(a)|<\infty,$$

即假设1-a)成立, 因为对任意的$i\in S$, $A(i)\equiv[0,a^*]$, 故假设1-b)成立. 再由E2)与式(16)知假设1-c)与1-d)成立, 由E1)与式(15)知假设1-e)成立.

由引理1知, 可以通过验证假设4来代替假设2的验证:

i) 当$N=1$时, 令$d_i=(\frac{\sqrt{\mu_*}}{\sqrt{\lambda^*}})^i(i\geqslant 0)$, $d_{-1}\equiv 0$. 由$d=\inf\limits_{i\geqslant 0}d_i=1>0$知假设4-a)成立. 对任意的$i\geqslant 0$,

$$\frac{d_{i+1}}{d_i}\lambda_{i+1}(a)+\frac{d_{i-1}}{d_i}\mu_i(a)-\lambda_i(a)-\mu_{i+1}(a)=\frac{\sqrt{\mu_*}}{\sqrt{\lambda^*}}[\lambda+h_1(a)]+\frac{\sqrt{\lambda^*}}{\sqrt{\mu_*}}[\mu+h_2(a)]-\lambda-h_1(a)-\mu-h_2(a)\leqslant\frac{\sqrt{\mu_*}}{\sqrt{\lambda^*}}\lambda^*+\frac{\sqrt{\lambda^*}}{\sqrt{\mu_*}}\mu_*=-(\sqrt{\mu_*}-\sqrt{\lambda^*})^2.$$

可取$\delta=(\sqrt{\mu_*}-\sqrt{\lambda^*})^2>0$使得假设4-b)成立.

另一方面,

$$L=\sum_{j\geqslant 1}\alpha_j\sup_{f\in F}\frac{\lambda_0^f\lambda_1^f\cdots\lambda_{j-1}^f}{\mu_1^f\mu_2^f\cdots\mu_j^f}\leqslant\sum_{j\geqslant 1}\alpha_j(\frac{\lambda^*}{\mu_*})^j,$$

其中

$$\alpha_j=\sum_{i=0}^{j-1}(\frac{\sqrt{\mu_*}}{\sqrt{\lambda^*}})^i=\frac{\sqrt{\lambda^*}}{\sqrt{\mu_*}-\sqrt{\lambda^*}}[(\frac{\sqrt{\mu_*}}{\sqrt{\lambda^*}})^j-1],$$

从而

$$L\leqslant\frac{\sqrt{\lambda^*}}{\sqrt{\mu_*}-\sqrt{\lambda^*}}[\sum_{j\geqslant 1}(\frac{\sqrt{\lambda^*}}{\sqrt{\mu_*}})^j-\sum_{j\geqslant 1}(\frac{\lambda^*}{\mu_*})^j]=\frac{\lambda^*}{\mu_*-\lambda^*}\frac{\sqrt{\mu_*}}{\sqrt{\mu_*}-\sqrt{\lambda^*}},$$

即假设4-c)成立.

ii) 当$N\geqslant 2$时, 令$d_i=(\frac{N}{N-1})^i(i\geqslant 0)$, $d_{-1}\equiv 0$. 由 $d=\inf\limits_{i\geqslant 0}d_i=1>0$ 知假设4-a)成立. 对任意的$i\geqslant 0$,

$$\frac{d_{i+1}}{d_i}\lambda_{i+1}(a)+\frac{d_{i-1}}{d_i}\mu_i(a)-\lambda_i(a)-\mu_{i+1}(a)=\frac{N}{N-1}[\lambda+h_1(a)]+\frac{N-1}{N}[N\mu+h_2(a)]-\lambda-h_1(a)-N\mu-h_2(a)-(\mu-\frac{\lambda}{N-1}).$$

取$\delta = \mu - \dfrac{\lambda}{N-1} > 0$知假设4-b)成立.

另一方面,

$$L \leqslant \sum_{j=1}^{N-1} \alpha_j (\frac{\lambda^*}{\mu})^j + \sum_{j=N}^{\infty} \alpha_j (\frac{\lambda^*}{N\mu})^j,$$

而级数

$$\sum_{j=N}^{\infty} \alpha_j (\frac{\lambda^*}{N\mu})^j =$$

$$(N-1) \sum_{j=N}^{\infty} (\frac{N}{N-1} \frac{\lambda^*}{N\mu})^j - (N-1) \sum_{j=N}^{\infty} (\frac{\lambda^*}{N\mu})^j,$$

从而$L < \infty$, 假设4-c)成立. 结合i)与ii), 由引理1知假设2成立. 最后令$f \equiv a^*$, 由式(16)与式(4), 对应的平均费用$\bar{V}_c(f,\nu) \equiv c(a^*)$, 结合E3)知假设3成立, 因此由定理1知, 上述模型存在平均最优平稳策略与约束最优策略.

注 4 下面考虑在第3节中提到的假设5(漂移条件). 在本例中, 这一条件很难满足. 事实上, 当$i \geqslant N$时,

$$\sum_{j \in S} q(j|i,a)\omega(j) =$$
$$N\mu[\omega(i-1) - \omega(i)] + \lambda[\omega(i+1) -$$
$$\omega(i)] + h_1(a)[\omega(i+1) - \omega(i)].$$

若取$\omega(i) = a_k i^k + a_{k-1} i^{k-1} + \cdots + a_0$, 则上式应为一个最高次等于$k-1$的多项式, 不可能满足式(9). 而在利用假设5的文章中往往取$\omega(i) = i + 1$(如文献[1,2,4]或者$\omega(i) = i$(如文献[11,12]), 这对于本例来说都无法成立.

5 结论(Conclusions)

本文在转移率, 报酬与费用函数均有界的情况下, 针对一种特殊的CTMDP －受控排队系统给出了其平均最优以及约束平均最优的新条件. 相比于文献[3]中利用非初始数据所建立的条件来说，本文的新条件要容易验证; 而文献[1,2,4,11,12]中的漂移条件又不能满足本文例子中的要求. 不过由于本文所给出的新条件是针对于受控排队系统的，因此笔者的下一步工作可致力于找出对一般CTMDP的平均最优以及约束平均最优的新条件.

另一方面, 本文证明了受控排队系统约束最优策略的存在性, 但没有给出策略的具体算法, 这也是值得笔者进一步研究的问题.

参考文献(References):

[1] GUO X , HERNÁNDRZ-LERMA O. Drift and monotonicity conditions for continuous-time controlled Markov chains with an average criterion[J]. *IEEE Transactions on Automatic Control*, 2003, 48(2): 236 – 244.

[2] GUO X P, CAO X R. Optimal control of ergodic continuous-time Markov chains with average sample-path reward[J]. *SIAM Journal on Control and Optimization*, 2005, 44(1): 29 – 30.

[3] 侯振挺, 郭先平. 马尔可夫决策过程[M]. 长沙: 湖南科学技术出版社, 1998.
(HOU Zhenting, GUO Xianping. *Markov Decision Processes*[M]. Changsha, China: Science and Technology Press, 1998.)

[4] ZHANG L L, GUO X P. Constrained continuous-time Markov decision processes with average criteria[J]. *Mathematical Methods of Operations Research*, 2008, 67(2): 323 – 340.

[5] GUO X P. Constrained optimality for average cost continuous-time Markov decision processes in Polish spaces[J]. *IEEE Transactions on Automatic Control*, 2007, 52(6): 1139 – 1143.

[6] ZEIFMAN A I. Some estimates of the rate of convergence for birth and death processes[J]. *Journal of Applied Probability*, 1991, 28(2): 268 – 277.

[7] CHEN M F. Estimate of exponential convergence rate in total variation by spectral gap[J]. *Acta Mathematica Sinica*, 1998, 14(1): 9 – 16.

[8] VAN DOORN E. Conditions for exponential ergodicity and bounds for the decay parameter of a birth and death processes[J]. *Advances in Applied Probability*, 1985, 17(3): 514 – 530.

[9] GUO X P, HERNáNDEZ-LERMA O. Constrained continuous-time Markov controll processes with discounted rewards[J]. *Stochastic Analysis and Applications*, 2003, 21(2): 379 – 399.

[10] ANDERSON W J. *Continuous-Time Markov Chains*[M]. New York: Springer-Verlag, 1991.

[11] GUO X P, HERNáNDEZ-LERMA O. Continuous-time controlled Markov chains with discounted rewards[J]. *Acta Applicandae Mathematicae*, 2003, 79(3): 195 – 216.

[12] GUO X P. Constrained nonhomogeneous Markov decision processes with expected total reward criterion[J]. *Acta Mathematicae Applicatae Sinica*, 2000, 23(1): 230 – 235.

作者简介:

张兰兰 (1980—), 女, 南方医科大学公共卫生与热带医学学院生物统计系教师, 目前研究方向为随机过程的最优控制及其应用, E-mail: katiezll@yahoo.com.cn;

郭先平 (1964—), 男, 中山大学数学与计算科学学院教授, 博士生导师, 目前研究方向为随机过程的最优控制及其应用、随机对策及其应用, E-mail: mcsgxp@mail.sysu.edu.cn