第 28 卷 第 6 期
2002 年 11 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 28, No. 6
Nov., 2002

Markov 控制过程基于性能势的平均代价最优策略[1)]

周亚平[1] 奚宏生[2] 殷保群[2] 孙德敏[2]

[1](中国科技大学管理科学系 合肥 230026)

[2](中国科技大学自动化系 合肥 230026)

(E-mail: zhouyp@ustc.edu.cn)

摘 要 研究了一类离散时间 Markov 控制过程平均代价性能最优控制决策问题. 应用 Markov 性能势的基本性质,在很一般性的假设条件下,直接导出了无限时间平均代价模型在紧致行动集上的最优性方程及其解的存在性定理. 提出了求解最优平稳控制策略的迭代算法,并讨论了这种算法的收敛性问题. 最后通过分析一个实例来说明这种算法的应用.

关键词 Markov 控制过程,性能势,平均代价模型,最优平稳策略

中图分类号 TP202

OPTIMALITY STRATEGY OF AVERAGE COST BASED PERFORMANCE POTENTIALS FOR MARKOV CONTROL PROCESS

ZHOU Ya-Ping[1] XI Hong-Sheng[2] YIN Bao-Qun[2] SUN De-Min[2]

[1](*Department of Management Science, University of Science and Technology of China, Hefei* 230026)

[2](*Department of Automation, University of Science and Technology of China, Hefei* 230026)

(E-mail: zhouyp@ustc.edu.cn)

Abstract This paper deals with the average cost optimization problem for a class of discrete time Markov control processes. Under quite general assumptions, the optimality equation is directly established and the existence theorem of optimal solution is proved for infinite time average cost model in a compact action set by using basic properties of the Markov performance potentials. The iterate algorithm for solving optimal stationary control strategy is suggested and the convergence problem of this algorithm is discussed. Finally, a numerical example is analyzed to illustrate the application of the proposed algorithm.

Key words Markov control process, performance potentials, average cost model, optimal stationary strategy

1) 国家自然科学基金(69974037)和国家高性能计算基金(00212)资助

收稿日期 2000-12-07 收修改稿日期 2001-12-18

1　引言

离散时间 Markov 控制过程(DMCP,Discrete-time Markov Control Process)是一类受到一系列控制决策驱动的 Markov 链,过程的状态转移规律与控制决策所采用的方案相互作用决定了过程的演化,由于其在计算机制造和现代通信网络等实际系统中具有广泛的应用而逐渐受到国内外控制界的关注. DMCP 的性能优化问题,通常被归结为 Markov 决策过程(MDP,Markov Decision Process). 文献[1～3]采用 MDP 基本原理研究了 DMCP 的平均代价的最优性方程及其解的存在性问题,并讨论了迭代求解的算法及其收敛性. 这些理论具有较强的约束条件,一些假设条件对于实际系统是难以满足或验证的. 最近文献[4,5]将 Markov 势论引入了 Markov 决策过程. 并揭示了 Markov 势论,摄动分析和 Markov 决策过程三者之间的关系,有关结果也被推广到了一类排队网络系统[6～8]. 本文主要在文献[4,5]的基础上,研究一类具有有限状态空间和紧致行动集的 DMCP 平均代价的优化问题. 在很一般性的假设条件下,应用 Markov 性能势的基本概念和性质直接导出了无限时间平均代价最优性方程及其解的存在性定理,给出了求解最优平稳控制策略的迭代算法,并讨论了这种算法的收敛性.

2　问题的描述

设 $X_n, n\geqslant 0$,是一离散时间 Markov 遍历链,具有有限状态空间 $\Phi=\{1,2,\cdots,M\}$和紧致行动集 A,记 Ω_S 是全体平稳策略集. 对 $v\in\Omega_S, v:\Phi\to A$,即 $i\in\Phi, v(i)=\beta_i\in A(i)\subseteq A$,对给定 $v\in\Omega_S$,令 $P^v=[p_{ij}(v(i))]_{i=1}^{M}{}_{,j=1}^{M}$是 X_n 在策略 v 驱动下的状态转移概率矩阵,并且稳态概率为 $\boldsymbol{\pi}^v=(\pi^v(1),\cdots,\pi^v(M))$. 显然,有

$$\boldsymbol{\pi}^v\boldsymbol{e}=1;\quad P^v\boldsymbol{e}=\boldsymbol{e};\quad \boldsymbol{\pi}^vP^v=\boldsymbol{\pi}^v \tag{1}$$

其中 $\boldsymbol{e}=(1,1,\cdots,1)^{\mathrm{T}}$ 是 M 维列向量. $\boldsymbol{f}^v=(f(1,v(1)),\cdots,f(M,v(M)))^{\mathrm{T}}$ 为性能函数. 我们称 $X=(X_n,\Phi,A,P^v,f^v)$为约束在 Ω_s 上的离散时间 Markov 控制过程.

设 X 的折扣代价性能指标为:$i\in\Phi, v\in\Omega_S$

$$\psi_\alpha^v(i)=E\Big\{\sum_{n=0}^{\infty}\alpha^n f(X_n,v(X_n))\,|\,X_0=i\Big\}=E_i\Big\{\sum_{n=0}^{\infty}\alpha^n f(X_n,v(X_n))\Big\} \tag{2}$$

其中 $0<\alpha<1$ 是折扣因子. 记 $\boldsymbol{\psi}_\alpha^v=(\psi_\alpha^v(1),\cdots,\psi_\alpha^v(M))^{\mathrm{T}}$,并注意到,$\psi_\alpha^v(i)=E_i\Big\{\sum_{n=0}^{\infty}\alpha^n f(X_n,v(X_n))\Big\}=\sum_{n=0}^{\infty}\alpha^n\sum_{j=1}^{M}p_{ij}^{(n)}(v(i))f(j,v(j))$,则 $\boldsymbol{\psi}_\alpha^v=\sum_{n=0}^{\infty}\alpha^n(P^v)^n\boldsymbol{f}^v=\Big[I+\alpha P^v\sum_{n=1}^{\infty}\alpha^{n-1}(P^v)^{n-1}\Big]\boldsymbol{f}^v=\Big[I+\alpha P^v\sum_{n=0}^{\infty}\alpha^n(P^v)^n\Big]\boldsymbol{f}^v=\boldsymbol{f}^v+\alpha P^v\boldsymbol{\psi}_\alpha^v$,即

$$\boldsymbol{\psi}_\alpha^v=(I-\alpha P^v)^{-1}\boldsymbol{f}^v \tag{3}$$

设 X 的平均代价性能指标为 $v=\Omega_S$

$$\eta^v=\boldsymbol{\pi}^v\boldsymbol{f}^v=\lim_{N\to\infty}E\Big\{\frac{1}{N}\sum_{n=0}^{N-1}f(X_n,v(X_n))\Big\} \tag{4}$$

并满足下列假设:

1) 对任意 $i,j\in\Phi$, $p_{i,j}(v(i))$是定义在 $A(i)$上的连续函数;

2) 对任意 $i\in\Phi$, $f(i,v(i))$是 $A(i)$上的连续函数.

对于 DMCP $X=(X_n,\Phi,A,P^v,\boldsymbol{f}^v)$中状态转移规律与控制决策选用的行动方案相互作用决定了过程的发展,问题是如何选择控制决策方案,使过程在性能代价准则下达到最优的运行效果.

3 平均代价最优性方程

在这一节中将引用文献[5]的结果直接导出 DMCP 平均代价模型的最优性方程,并证明最优性方程解的存在性定理.

对任意给定的 $v\in\Omega_S$ 和实数 $0<\alpha\leqslant 1$,广义 Poisson 方程为

$$(I-\alpha P^v+\alpha\boldsymbol{e}\boldsymbol{\pi}^v)\boldsymbol{g}_\alpha^v=\boldsymbol{f}^v \tag{5}$$

其中解向量 $\boldsymbol{g}_\alpha^v=(g_\alpha^v(1),\cdots,g_\alpha^v(M))^{\mathrm{T}}$,当 $\alpha=1$ 时

$$(I-P^v+\boldsymbol{e}\boldsymbol{\pi}^v)\boldsymbol{g}^v=\boldsymbol{f}^v \tag{6}$$

称为标准 Poisson 方程,其解向量 $\boldsymbol{g}^v=(g^v(1),\cdots,g^v(M))^{\mathrm{T}}$.

注意到对任意 $v=\Omega_S$, $(P^v-\boldsymbol{e}\boldsymbol{\pi}^v)$的特征值均在单位圆内[4]. 因此对 $0<\alpha\leqslant 1$, $\alpha(P^v-\boldsymbol{e}\boldsymbol{\pi}^v)$的特征值也均在单位圆内. 由此易知$(I-\alpha P^v+\alpha\boldsymbol{e}\boldsymbol{\pi}^v)$是非奇异的,即

$$\boldsymbol{g}_\alpha^v=(I-\alpha P^v+\alpha\boldsymbol{e}\boldsymbol{\pi}^v)^{-1}\boldsymbol{f}^v;\quad \boldsymbol{g}^v=(I-P^v+\boldsymbol{e}\boldsymbol{\pi}^v)^{-1}\boldsymbol{f}^v \tag{7}$$

并且有

$$\lim_{\alpha\to 1}\boldsymbol{g}_\alpha^v=\boldsymbol{g}^v \tag{8}$$

定义 1. 对任意平稳策略 $v\in\Omega_S$,称 $\boldsymbol{g}^v+\boldsymbol{e}c$ 为 DMCP X 的平均代价性能势向量,它的第 i 个分量 $g^v(i)+c$ 为在状态 i 的性能势,其中 c 是任意常数.

该性能势与文献[5]中定义的性能势具有相同的性质. 易验证,下述结论成立.

引理 1. 对任意 $v\in\Omega_S$, $0<\alpha<1$,有

1) $\boldsymbol{\pi}^v(I-\alpha P^v+\alpha\boldsymbol{e}\boldsymbol{\pi}^v)^{-1}=\boldsymbol{\pi}^v$;

2) $(I-\alpha P^v+\alpha\boldsymbol{e}\boldsymbol{\pi}^v)^{-1}\boldsymbol{e}=\boldsymbol{e}$;

3) $(I-\alpha P^v)^{-1}\boldsymbol{e}=\dfrac{1}{1-\alpha}\boldsymbol{e}$.

由引理 1 的 1)和 3)以及 $\boldsymbol{\pi}^v\boldsymbol{g}^v=\boldsymbol{\pi}^v\boldsymbol{f}^v=\eta^v$,可得

$$(I-\alpha P^v)^{-1}=(I-\alpha P^v+\alpha\boldsymbol{e}\boldsymbol{\pi}^v)^{-1}+\frac{\alpha}{1-\alpha}\boldsymbol{e}\boldsymbol{\pi}^v \tag{9}$$

由式(9)和(3),易得

$$\boldsymbol{\psi}_\alpha^v=\boldsymbol{g}_\alpha^v+\frac{\alpha}{1-\alpha}\boldsymbol{e}\boldsymbol{\pi}^v\boldsymbol{f}^v=\boldsymbol{g}_\alpha^v+(I-\alpha P^v)^{-1}\alpha\boldsymbol{e}\eta^v \tag{10}$$

Blackwell[9]引用 Banach 不动点定理首先给出了折扣代价问题最优性方程及其解的存在性定理,并且最优平稳策略存在的充分必要条件由下面的定理给出.

定理 1. 对任意给定的 $0<\alpha<1$,折扣代价模型(2)存在唯一的最优平稳策略 $v^*\in\Omega_s$ 和 $\boldsymbol{\psi}_\alpha^{v^*}$ 满足最优性方程

$$\boldsymbol{0}=\min_{v\in\Omega_s}\{\boldsymbol{f}^v+(\alpha P^v-I)\boldsymbol{\psi}_\alpha^{v^*}\}.$$

等价地,设$(v(1),v(2),\cdots,v(M))=\beta\in A$,并记最优平稳策略 v^* 对应的最优行动$(v^*(1),v^*(2),\cdots,v^*(M))=\delta^\infty\in A$,并令

$$\delta=\arg\min_{\beta\in A}\{\boldsymbol{f}^\beta+(\alpha P^\beta-I)\boldsymbol{\psi}_\alpha^{\delta^\infty}\}.$$

则有

$$\mathbf{0}=\boldsymbol{f}^\delta+(\alpha P^\delta-I)\boldsymbol{\psi}_\alpha^{\delta^\infty}\leqslant\boldsymbol{f}^v+(\alpha P^v-I)\boldsymbol{\psi}_\alpha^{\delta^\infty},\quad v\in\Omega_s \tag{11}$$

文献[4,5]首次揭示了 Markov 性能势,摄动分析和平均代价 Markov 决策过程之间的关系.其中两个重要结果被引用作为下述引理 2 和定理 2.

引理 2. 对任意 $v,v'\in\Omega_S$,有

$$\eta^v-\eta^{v'}=\boldsymbol{\pi}^v\{(\boldsymbol{f}^v+P^v\boldsymbol{g}^{v'})-(\boldsymbol{f}^{v'}+P^{v'}\boldsymbol{g}^{v'})\}.$$

定理 2. v^* 是平均代价模型(4)的平均代价最优平稳策略的充分必要条件是对任意平稳策略 $v\in\Omega_s$,有

$$\boldsymbol{f}^{v^*}+P^{v^*}\boldsymbol{g}^{v^*}\leqslant\boldsymbol{f}^v+P^v\boldsymbol{g}^{v^*} \tag{12}$$

上述关系符号"≤"表示,对应分量小于或等于关系.

注意到 $e\eta^{v^*}+\boldsymbol{g}^{v^*}=\boldsymbol{f}^{v^*}+P^{v^*}\boldsymbol{g}^{v^*}$,等价地 v^* 是平均代价模型的最优平稳策略的充分必要条件是 v^* 应满足下述最优性方程

$$\mathbf{0}=\min_{v\in\Omega_s}\{\boldsymbol{f}^v-e\eta^{v^*}+(P^v-I)\boldsymbol{g}^{v^*}\} \tag{13}$$

定理 3. 在假设条件 1)和 2)下,存在实数 η 和一个 M 维向量 $\boldsymbol{g}$,满足最优性方程(13).其次,若(η',g')是满足(13)的任意一个解,则 $\eta'=\eta$.

证明. 选择一折扣因子序列 $\alpha_k,\alpha_k\uparrow 1$,由定理 1 知,对每一固定的 α_k,存在唯一最优平稳策略 v_k^*,其对应的行动 $\delta_k^\infty\in A$ 及 $\delta_k=\arg\min_{\beta_k\in A}\{\boldsymbol{f}^{\beta_k}-(\alpha_kP^{\beta_k}-I)\boldsymbol{\psi}_{\alpha_k}^{\delta_k^\infty}\}$.由于 A 是紧致集,故存在一子序列$\{\delta_{k_l}\}$收敛到 $\delta\in A$,且$\{\delta_{k_l}^\infty\}$收敛到 $\delta^\infty\in A$. 为表达简洁起见,将$\{k_l\}$仍记为 k.则由式(11),$\mathbf{0}=\boldsymbol{f}^{\delta_k}+(\alpha_kP^{\delta_k}-I)\boldsymbol{\psi}_{\alpha_k}^{\delta_k^\infty}\leqslant\boldsymbol{f}^{\beta_k}+(\alpha_kP^{\beta_k}-I)\boldsymbol{\psi}_{\alpha_k}^{\delta_k^\infty}$,$\beta_k\in A$. 将(10)式代入上式,并注意到引理 1 中 3)的等式右边不依赖策略 v,有

$$\mathbf{0}=\boldsymbol{f}^{\delta_k}+(\alpha_kP^{\delta_k}-I)[\boldsymbol{g}_{\alpha_k}^{\delta_k^\infty}+(I-\alpha_kP^{\delta_k})^{-1}\alpha_ke\eta^{\delta_k^\infty}]\leqslant\boldsymbol{f}^{\beta_k}+(\alpha_kP^{\beta_k}-I)[\boldsymbol{g}_{\alpha_k}^{\delta_k^\infty}+(I-\alpha_kP^{\beta_k})^{-1}\alpha_ke\eta^{\delta_k^\infty}]$$

由假设条件 1)和 2),易证得$\lim_{k\to\infty}\eta^{\delta_k^\infty}=\lim_{k\to\infty}\boldsymbol{\pi}^{\delta_k^\infty}\boldsymbol{f}^{\delta_k^\infty}=\boldsymbol{\pi}^{\delta^\infty}\boldsymbol{f}^{\delta^\infty}=\eta^{\delta^\infty}$,$\lim_{\substack{k\to\infty\\ \alpha_k\uparrow 1}}\boldsymbol{g}_{\alpha_k}^{\delta_k^\infty}=\lim_{k\to\infty}(I-\alpha_kP^{\delta_k^\infty}+\alpha_ke\boldsymbol{\pi}^{\delta_k^\infty})^{-1}\boldsymbol{f}^{\delta_k^\infty}=(I-P^{\delta^\infty}+e\boldsymbol{\pi}^{\delta^\infty})^{-1}\boldsymbol{f}^{\delta^\infty}=\boldsymbol{g}^{\delta^\infty}$. 令 $k\to\infty$, $\alpha_k\uparrow 1$,有

$$\mathbf{0}=\boldsymbol{f}^\delta-e\eta^{\delta^\infty}+(P^\delta-I)\boldsymbol{g}^{\delta^\infty}\leqslant\boldsymbol{f}^{\beta_k}-e\eta^{\delta^\infty}+(P^{\beta_k}-I)\boldsymbol{g}^{\delta^\infty},\beta\in A.$$

即存在$(\eta,\boldsymbol{g})=(\eta^{\delta^\infty},\boldsymbol{g}^{\delta^\infty})$满足(13).若$(\eta',\boldsymbol{g}')$是式(13)的另一解,则由式(12)和引理 2,$\eta'\leqslant\eta^{\delta^\infty}$,反之亦然,有 $\eta^{\delta^\infty}\leqslant\eta'$,故 $\eta^{\delta^\infty}=\eta'$.

定理 3 同时也表明,存在一个最优平稳策略 v^* 及其对应的最优行动$(v^*(1),v^*(2),\cdots,v^*(M))=\delta^\infty\in A$,满足式(13). 在通常情况下,平均代价模型的最优平稳策略并不唯一,而最优准则函数值是唯一的.

4　迭代算法

这一节给出基于性能势迭代的策略优化算法及收敛性证明.该算法如下.

步 1. 置 $k=0$,选择初始策略 v_k.

步 2. 由式(1)和式(7),计算 π^{v_k}和 $\boldsymbol{g}^{v_k}$,并置 $\boldsymbol{g}^{v_k}=\boldsymbol{g}^k$.

步 3. 选择策略 v_{k+1},对每一状态 $i\in\Phi$,满足

$$v_{k+1}(i)\in\arg\min_{a_i\in A(i)}\left\{f(i,a_i)+\sum_{j=1}^{M}p_{ij}(a_i)g^{v_k}(j)\right\}\tag{14}$$

步 4. 计算 $\boldsymbol{g}^{v_{k+1}}=\boldsymbol{f}^{v_{k+1}}+P^{v_{k+1}}\boldsymbol{g}^{v_k}=L^{k+1}\boldsymbol{g}^{v_k}$.

步 5. 如果 $sp(\boldsymbol{g}^{v_{k+1}}-\boldsymbol{g}^{v_k})<\varepsilon$,则转步 6;否则置 $k+1=k$,转步 3.

步 6. 对每一状态 $i\in\Phi$,选择

$$v_\varepsilon(i)\in\arg\min_{a_i\in A(i)}\{f(i,a_i)+\sum_{j=1}^{M}p_{ij}(a_i)g^{v_k}(j)\}.$$

停止,v_ε 是 ε-最优平稳策略.

其中 $sp(\boldsymbol{g}^{v_{k+1}}-\boldsymbol{g}^{v_k})\triangleq\max_{i\in\Phi}(\boldsymbol{g}^{v_{k+1}}-\boldsymbol{g}^{v_k})(i)-\min_{i\in\Phi}(\boldsymbol{g}^{v_{k+1}}-\boldsymbol{g}^{v_k})(i)$. 另外,注意到 $\arg\min_{v\in\Omega_s}\{\boldsymbol{f}^v+P^v\boldsymbol{g}\}=\arg\min_{v\in\Omega_s}\{\boldsymbol{f}^v+(P^v-I)\boldsymbol{g}-\boldsymbol{e}c\}$. 其中 c 是任意常数.

步骤 3 表明,如果 $v_{k+1}(i)=v_k(i),i=1,2,\cdots,M$,则

$$\boldsymbol{f}^{v_{k+1}}+P^{v_{k+1}}\boldsymbol{g}^{v_k}=\boldsymbol{f}^{v_k}+P^{v_k}\boldsymbol{g}^{v_k}\leqslant\boldsymbol{f}^v+P^v\boldsymbol{g}^{v_k},\quad v\in\Omega_s.$$

由定理 2,v_k 是最优策略. 否则,$\boldsymbol{f}^{v_{k+1}}+P^{v_{k+1}}\boldsymbol{g}^{v_k}<\boldsymbol{f}^{v_k}+P^{v_k}\boldsymbol{g}^{v_k}$,关系符号"<"表示对应分量中至少有一对"<"成立. 于是,由引理 2,$\eta^{v_{k+1}}<\eta^{v_k}$,$v_{k+1}$是可改善性能指标的策略.

下面证明该算法的收敛性. 记 $\boldsymbol{g}^{k+1}=\min_{v\in\Omega_s}\{\boldsymbol{f}^v+P^v\boldsymbol{g}^k\}=L\boldsymbol{g}^k$. 由文献[9],由于 $p_{ij}(a_i)$是定义在紧致集 $A(i)$上的连续函数,Φ 是有限集,故存在 $0<r<1$,使得 $sp(\boldsymbol{g}^{k+1}-\boldsymbol{g}^k)\leqslant rsp(\boldsymbol{g}^k-\boldsymbol{g}^{k-1})$,其中,$r=\max_{i\in\Phi,a_i\in A(i),i'\in\Phi,a_{i'}\in A(i')}\left\{1-\sum_{j=1}^{M}\min[p_{ij}(a_i),p_{ij}(a_{i'})]\right\}$,故 $sp(\boldsymbol{g}^{k+1}-\boldsymbol{g}^k)\leqslant r^ksp(\boldsymbol{g}^1-\boldsymbol{g}^0)$,因此,对任意 $\varepsilon>0$,存在 k_0,当 $k\geqslant k_0$ 时,恒有 $sp(\boldsymbol{g}^{k+1}-\boldsymbol{g}^k)<\varepsilon$.

假如任给 $\varepsilon>0$,对固定的 k,$sp(\boldsymbol{g}^{k+1}-\boldsymbol{g}^k)<\varepsilon$ 成立. 选择 $v_\varepsilon(i)\in\arg\min_{a_i\in A(i)}\left\{f(i,a_i)+\sum_{j=1}^{M}p_{ij}(a_i)g^k(j)\right\}$, $i=1,2,\cdots,M$. 取 $\boldsymbol{g}^k=\boldsymbol{g}^{v_k}$,$v_\varepsilon(i)=v_{k+1}(i),i=1,2,\cdots,M$,则有 $\eta^{v_\varepsilon}=\eta^{v_{k+1}}=\boldsymbol{\pi}^{v_{k+1}}\boldsymbol{f}^{v_{k+1}}=\boldsymbol{\pi}^{v_{k+1}}[\boldsymbol{f}^{v_{k+1}}+P^{v_{k+1}}\boldsymbol{g}^{v_k}-\boldsymbol{g}^{v_k}]=\boldsymbol{\pi}^{v_{k+1}}[L^{k+1}\boldsymbol{g}^{v_k}-\boldsymbol{g}^{v_k}]=\boldsymbol{\pi}^{v_{k+1}}[L\boldsymbol{g}^k-\boldsymbol{g}^k]\leqslant\max_{i\in\Phi}[L\boldsymbol{g}^k-\boldsymbol{g}^k](i)$.

设 η^* 是最优值,并注意到(14),对于任意的 $v\in\Omega_S$,$\boldsymbol{f}^v+P^v\boldsymbol{g}^{v_k}\geqslant\boldsymbol{f}^{v_{k+1}}+P^{v_{k+1}}\boldsymbol{g}^{v_k}$,

$$\eta^*=\boldsymbol{\pi}^*\boldsymbol{f}^*=\boldsymbol{\pi}^*[\boldsymbol{f}^*+P^*\boldsymbol{g}^{v_k}-\boldsymbol{g}^{v_k}]\geqslant\boldsymbol{\pi}^*[L^{k+1}\boldsymbol{g}^{v_k}-\boldsymbol{g}^{v_k}]=\boldsymbol{\pi}^*[L\boldsymbol{g}^k-\boldsymbol{g}^k],$$

$\geqslant\min_{i\in\Phi}[L\boldsymbol{g}^k-\boldsymbol{g}^k](i)$,因而,$\eta^{v_\varepsilon}-\eta^*\leqslant sp(L\boldsymbol{g}^k-\boldsymbol{g}^k)<\varepsilon$.

5 实例

考虑一个具有有限状态空间和紧致行动集的 DMCP 系统. 该系统的运行规律可由一个遍历的马尔可夫链表示. 它有 3 个状态:$\Phi=\{1,2,3\}$. 在状态 i 下的行动为维护系统运行的费用,$v(i)\in A(i)=[0,+\infty)$,策略 v 为 $\Phi\to A$ 的一个映射. 状态转移概率与维护费用的关系为

$$P_{11}=1-\mathrm{e}^{\frac{\ln 0.01}{3000}v(1)}\quad P_{12}=\frac{7}{8}(1-p_{11})\quad P_{13}=\frac{1}{8}(1-p_{11})$$

$$p_{21}=\begin{cases}(833.3*v(2)+1.667*v(2)^2)*10^{-7}, & \text{当 } v(2)\leqslant 1500\\ 1-0.5*\mathrm{e}^{\frac{\ln 20}{2500}(1500-v(2))}, & \text{其它}\end{cases}$$

$$P_{22}=1-p_{21}-p_{23},$$

$$p_{23}=\frac{1}{3}\mathrm{e}^{\frac{\ln 0.3}{1500}v(2)},$$

$$P_{31}=\begin{cases}(4167*v(3)+4*v(3)^2)*10^{-8}, & \text{当 } v(3)\leqslant 3700\\ 1-0.298*\mathrm{e}^{\frac{\ln 10}{2000}(3700-v(3))}, & \text{其它}\end{cases}$$

$$p_{32}=1-p_{31}-p_{33},$$

$$p_{33}=\begin{cases}1-(v(3)+0.001*v(3)^2)*10^{-4}, & \text{当 } v(3)\leqslant 2000\\ 0.4*\mathrm{e}^{\frac{\ln 0.025}{3000}(v(3)-2000)}, & \text{其它}\end{cases}$$

其状态转移概率与维护费用的关系如图 1～图 3 所示.

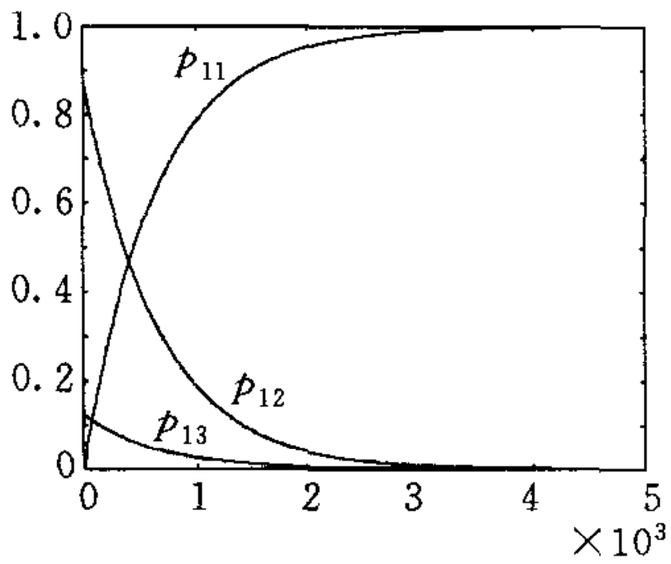

图 1　$P_{1j}(j=1,2,3)$

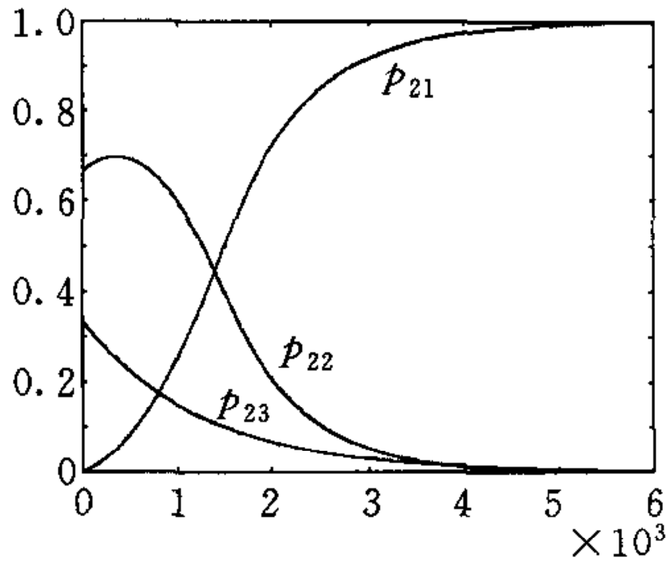

图 2　$P_{2j}(j=1,2,3)$

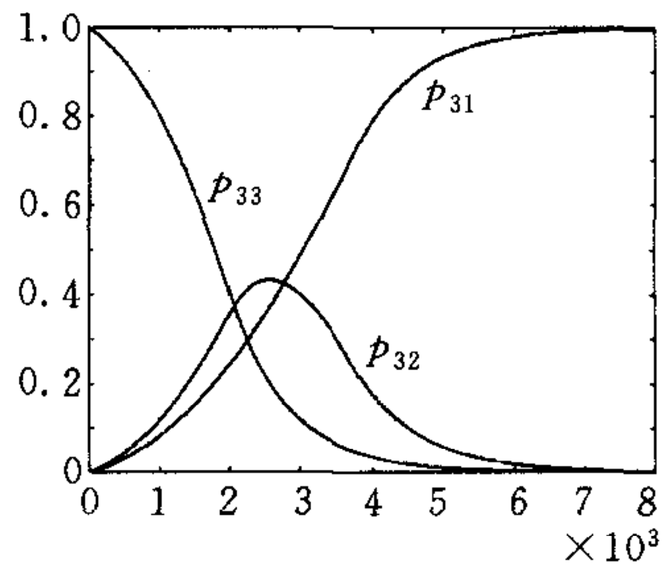

图 3　$P_{3j}(j=1,2,3)$

另外,系统在状态 i 运行的固定代价为 $cost(i)$($cost(1)=0, cost(2)=2000, cost(3)=8000$). 于是,在状态 i,系统的总运行代价为

$$f(i)=cost(i)+v(i)$$

容易看出,性能指标函数 $\boldsymbol{f}$ 与稳态概率 $\boldsymbol{\pi}$ 都取决于策略 v,分别记为 $\boldsymbol{f}^v$ 与 $\boldsymbol{\pi}^v$,平均代价为 $\eta^v=\boldsymbol{\pi}^v\boldsymbol{f}^v$. 我们希望寻找一稳态策略 v,使平均代价 η^v 达到最小.

表 1 给出了基于梯度寻优方法的结果和基于性能势的迭代算法进行寻优的结果与运行时间的比较,从中可以看出本文所用的算法具有明显的速度优势. 如果所研究的问题状态空间更大的话,则这种速度优势将越明显. 因为与基于梯度方法的直接寻优相比,该方法相当于将对一个 M 维向量的整体寻优转化为分别对 M 个分量的寻优.

表 1　两种算法的寻优结果与运行时间的比较

方法	误差	最优策略	最优值	运行时间
基于梯度直接寻优	$\varepsilon=0.00001$	(1246.38, 2402.16, 4127.33)	1897.82	21 分 53 秒
基于性能势迭代	$\varepsilon=0.00001$	(1246.38, 2402.16, 4127.37)	1897.82	4 秒

6 结论

本文是在 Markov 性能势的基础上，以一种简明的方法直接导出了 DMCP 无限时间平均代价模型在紧致行动集上的最优性方程及其最优解的存在性定理，给出了基于策略和性能势向量迭代的收敛算法. 该方法与已有的有关结果相比较具有较大的优越性. 首先，在理论上，减弱了对问题的约束条件，文章中的假设条件 1)和 2)适合大多数实际系统. 其次，本文的迭代算法具有非常快的收敛速度，并且适合于并行处理. 注意到性能势向量能够通过分析实际系统的一条样本轨道被估计，采用文献[8]中的方法，能够建立一种新的高效并行算法. 这项研究的进一步工作是建立一个依赖于性能势并行仿真的简单而快速的迭代算法，为实际 DMCP 系统的性能优化提供一条新的途径.

参 考 文 献

1 Arapostathis A, Borkar V S, Fernandez-Gaucherand E Ghosh M K, Marcus S I. Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM J. Control Optim.*, 1993, **31**(2):282～344

2 Raul Montes-de-Oca. The average cost optimality equation for Markov control processes on Borel spaces. *System & Control Letters*, 1994, **22**(2):351～357

3 Sennott L I. Another set of Conditions for average optimality in Markov control processes. *System & Control Letters*, 1995, **23**(1):147～151

4 Cao Xi-Ren. The relations among potentials, perturbation analysis, and Markov decision processes. *Discrete Event Dynamic Systems: Theory and Applications*, 1998, **8**(1):71～87

5 Cao Xi-Ren. A unified approach to Markov decision problems and performance sensitivity analysis. *Automatica*, 2000, **36**(5):771～774

6 殷保群，周亚平，杨孝先，奚宏生，孙德敏. 状态相关闭排队网络的性能指标灵敏度公式. 控制理论与应用，1999，**16**(2)：255～257

7 Yin Bao-Qun, Zhou Ya-Ping, Xi Hong-Sheng, Sun De-Min. Sensitivity formulas of performance in two-server cyclic queueing networks with phase-type distributed service times. *International Transactions in Operation Research*, 1999, **6**(6):649～663

8 邹长春，周亚平，殷保群，奚宏生，孙德敏. 基于性能势理论对闭排队网络进行梯度估计的并行仿真算法. 中国科技大学学报，1999，**29**(1):21～29

9 Puterman M L. Markov Decision Processes: Discrete Stochastic Dynamic Programming. New York: Wiley, 1994

周亚平 工学博士，现为中国科技大学管理科学系副教授. 研究领域为排队网络性能指标灵敏度仿真估计及优化等.

奚宏生 现为中国科技大学自动化系教授，博士生导师，自动化系副主任. 研究领域为鲁棒控制、离散事件动态系统及其应用等.

殷保群 工学博士，现为中国科技大学自动化系副教授. 主要从事非线性系统展开理论，随机离散事件系统性能分析、优化及在通讯网络中的应用等方面的工作.

孙德敏 中国科技大学自动化系教授，博士生导师，中国自动化学会理事，中国自动化学会控制理论专业委员会委员. 长期从事工业控制过程先进控制和优化及伺服系统综合等方面的研究.