云南大学学报（自然科学版），2003，25（5）：377~380　　CN 53-1045/N　ISSN 0258-7971
Journal of Yunnan University

参数 Markov 决策过程的随机逼近算法*

胡光华
（云南大学 数学系，云南 昆明　650091）

摘要：讨论平均报酬参数马氏决策过程的随机梯度算法，利用与折扣报酬的关系，给出了目标函数的梯度的一个新的表达式．同时得到了基于单一样本路径的随机逼近算法，最后证明了算法以概率 1 收敛到其梯度．

关键词：参数 Markov 决策过程；随机逼近；平均报酬
中图分类号：O 211.5　　文献标识码：A　　文章编号：0258-7971(2003)05-0377-04

用仿真的方法求解马氏决策过程(MDP)近几年来引起了不少学者的注意．求解的方法可分 3 种，一是逼近值函数从而求出最优策略，如 Watkins[1]的 Q-学习算法和 Williams[2]的梯度算法；二是直接通过仿真的方法逼近由一参数向量所确定的最优策略[3]，另外就是将两者结合起来的“行动-评价”方法[4]．

本文考虑用第 2 种方法求解平均报酬参数马氏决策过程，利用随机逼近的方法直接求出最优参数所确定的参数 MDP 的最优策略；给出了一种与文献[3]不同的基于折扣准则的随机逼近算法．并讨论了算法以概率 1 的收敛性．

1　模　型

考虑离散时间、有限状态集 $S=\{1,2,\cdots,N\}$ 的时齐马氏链 $\{X_t, t=0,1,2,\cdots\}$．其状态转移概率矩阵依赖于某 R^K 空间中的参数向量 $\theta\in R^K$，即 $\boldsymbol{P}(\theta)=(p_{ij}(\theta))_{N\times N}$．其中，对 $\forall i,j\in S$，

$$\boldsymbol{p}_{ij}(\theta)=\boldsymbol{P}(X_{t+1}=j\mid X_t=i,\theta)$$

在状态 $i\in S$ 系统获一瞬时报酬也与参数向量有关，记为 $r(i,\theta)$．这样参数 θ 充当了控制策略的作用；取不同的 θ，系统就会有不同的状态转移矩阵和报酬函数．称该模型为参数 MDP．

通常意义下的马氏决策过程 $\langle S,A,P,r\rangle$ 很容易转换为上面的参数 MDP．例如，若设行动集 $A=\{1,2,\cdots,M\}$，因一个随机平稳策略 μ 是满足对任意 $i\in S$，$\sum_{a\in A}\mu_a(i)=1$ 的 S 到 $[0,1]^M$ 上的映射，对 $\theta\in R^K$，若给定关于 θ 的连续可微的基函数族 $\{f_a(i,\theta): a\in A, i\in S\}$，则取策略 μ 在状态 i 取行动 a 的概率为

$$\mu_a(i,\theta)=\frac{f_a(i,\theta)}{\sum_{a'\in A}f_{a'}(i,\theta)},$$

则不同 θ 的便确定不同的策略．此时参数 MDP 中的转移概率和报酬函数为

$$p_{ij}(\theta)=\sum_{a\in A}\mu_a(i,\theta)p_{ij}(a),$$

$$r(i,\theta)=\sum_{a\in A}\mu_a(i,\theta)r(i,a).$$

设 $Q=\{\boldsymbol{P}(\theta):\theta\in R^K\}$，$\overline{Q}$ 表示 Q 的闭包，其元素仍为随机矩阵．本文始终假设：

假设 **1**　$\forall\boldsymbol{P}\in\overline{Q}$ 所确定的马氏链是不可分、非周期的．

假设 **2**　对 $\forall i,j\in S$ 及 $\theta\in R^K$，$p_{ij}(\theta)$ 和 $r(i,\theta)$ 有界，有连续二阶导数；且其一、二阶导数有限．

目标函数有 2 种准则，分别是

(1) 折扣报酬：对任意 $i\in S$，

$$J_\alpha(i,\theta)=\lim_{T\to\infty}E_\theta\Big[\sum_{t=0}^{T-1}\alpha^t r(X_t,\theta)\mid X_0=i\Big],$$

这里 E_θ 表示对转移概率矩阵为 $\boldsymbol{P}(\theta)$ 的马氏链 $\{X_t\}$ 求期望．

(2) 平均报酬：

* 收稿日期：2003-05-05
基金项目：云南省教育厅基金资助项目(K1050401)；云南大学理(工)科校级科研项目资助(K1059040)．
作者简介：胡光华(1962-　)，男，云南人，博士，副教授，主要从事随机控制、人工神经网络技术与计算智能方面的研究．

$$\lambda(\theta) = \lim_{T\to\infty} E_\theta\left[\frac{\sum_{t=0}^{T-1} r(X_t, \theta)}{T}\right].$$

由假设1可知, 对任意给定的 $\theta \in R^K$, $\lambda(\theta)$ 有定义且与初始状态无关; 此时由 $\boldsymbol{P}(\theta)$ 所确定的马氏链为遍历的且存在平稳分布 $\pi(\theta) = (\pi_1(\theta), \cdots, \pi_N(\theta))^{\mathrm{T}}$ 满足

$$\pi^{\mathrm{T}}(\theta)\boldsymbol{P}(\theta) = \pi^{\mathrm{T}}(\theta), \tag{1}$$

$$\pi^{\mathrm{T}}(\theta)e = 1$$

这里, π^{T} 表示列向量 π 的转置, $e = (1, 1, \cdots, 1)^{\mathrm{T}} \in R^N$. 此时, 由文献[5]知, 平均报酬等于

$$\lambda(\theta) = \sum_{i=1}^{N} \pi_i(\theta) r(i, \theta) = \pi^{\mathrm{T}}(\theta) r(\theta), \tag{2}$$

其中 $r(\theta) = (r(1, \theta), \cdots, r(N, \theta))^{\mathrm{T}}$ 为报酬向量函数. 有了平稳分布的概念后, 便可定义期望折扣报酬 $\lambda_\alpha(\theta)$ 如下

$$\lambda_\alpha(\theta) = \sum_{i=1}^{N} \pi_i(\theta) J_\alpha(i, \theta) = \pi^{\mathrm{T}}(\theta) J_\alpha(\theta). \tag{3}$$

引理 **1** 折扣报酬目标函数满足

$$J_\alpha(\theta) = r(\theta) + \alpha\boldsymbol{P}(\theta) r(\theta) + \alpha^2\boldsymbol{P}^2(\theta) r(\theta) + \cdots \tag{4}$$

且 $J_\alpha(\theta)$ 满足 Bellman 方程

$$J_\alpha(\theta) = r(\theta) + \alpha\boldsymbol{P}(\theta) J_\alpha(\theta).$$

引理 **2** $\pi(\theta)$, $\lambda_\alpha(\theta)$ 和 $\lambda(\theta)$ 均二阶可微, 且具有有限的一、二阶导数.

引理 1 的证明见文献[5]; 而引理 2 的证明见文献[3].

本文求解的问题是: 求参数 $\theta \in R^K$ 使得 $\lambda(\theta)$ 或 $\lambda_\alpha(\theta)$ 最大. 下定理说明两目标函数是一致的.

定理 **1** 对任意 $\theta \in R^K$ 和 $\alpha \in (0, 1)$, 有

$$\lambda_\alpha(\theta) = \frac{\lambda(\theta)}{1-\alpha}.$$

证明 由(3)式和(4)式

$$\lambda_\alpha(\theta) = \pi^{\mathrm{T}}(\theta) J_\alpha(\theta),$$

$$J_\alpha(\theta) = r(\theta) + \alpha\boldsymbol{P}(\theta) r(\theta) + \alpha^2\boldsymbol{P}^2(\theta) r(\theta) + \cdots$$

考虑到平稳策略满足 $\pi^{\mathrm{T}}(\theta)\boldsymbol{P}(\theta) = \pi^{\mathrm{T}}(\theta)$, 故有

$$\lambda_\alpha(\theta) = \pi^{\mathrm{T}}(\theta) \lim_{n\to\infty} \sum_{t=0}^{n} \alpha^t \boldsymbol{P}^t(\theta) r(\theta) =$$

$$\lim_{n\to\infty} \sum_{t=0}^{n} \alpha^t \lambda \pi^{\mathrm{T}}(\theta) \boldsymbol{P}^t(\theta) r(\theta) =$$

$$\lim_{n\to\infty} \sum_{t=0}^{n} \alpha^t \pi^{\mathrm{T}}(\theta) r(\theta) = \frac{\lambda(\theta)}{1-\alpha}$$

证毕.

2 随机逼近算法

文[3]中给出了 $\lambda(\theta)$ 的梯度为

$$\nabla\lambda(\theta) = \pi^{\mathrm{T}}(\theta)(\nabla r(\theta) + \nabla\boldsymbol{P}(\theta) v(\theta),$$

其中 v 为相对值函数, 第 i 个分量定义为

$$v(i, \theta) = E_\theta\left[\lim_{T\to\infty} \sum_{t=0}^{T-1} (r(X_t, \theta) - \lambda(\theta)) \mid X_0 = i\right].$$

其计算过程涉及到平稳分布 $\pi(\theta)$, $\lambda(\theta)$ 及值函数 $v(\theta)$, 较为复杂. 利用平均报酬与折扣报酬的关系, 本文给出另外一个 $\lambda(\theta)$ 的梯度的表达式.

定理 **2**

$$\nabla\lambda(\theta) = (1-\alpha)\nabla\pi^{\mathrm{T}}(\theta) J_\alpha(\theta) + \pi^{\mathrm{T}}(\theta)(\nabla r(\theta) + \alpha\nabla\boldsymbol{P}(\theta) J_\alpha(\theta))$$

证明 由 Bellman 方程得

$$r(\theta) = J_\alpha(\theta) - \alpha\boldsymbol{P}(\theta) J_\alpha(\theta).$$

又对(1)式两边求梯度, 得

$$\nabla\pi^{\mathrm{T}}(\theta)(I - \boldsymbol{P}(\theta)) = \pi^{\mathrm{T}}(\theta)\nabla\boldsymbol{P}(\theta), \tag{5}$$

再对(2)两边求梯度并将(5)式代入得

$$\begin{aligned}\nabla\lambda(\theta) &= \nabla\pi^{\mathrm{T}}(\theta) = \\ &\nabla\pi^{\mathrm{T}}(\theta)(J_\alpha(\theta) - \alpha\boldsymbol{P}(\theta) J_\alpha(\theta)) + \\ &\pi^{\mathrm{T}}(\theta)\nabla r(\theta) = \\ &\nabla\pi^{\mathrm{T}}(\theta)((I - \alpha 1) + \alpha(I - \\ &\boldsymbol{P}(\theta)) J_\alpha(\theta) + \pi^{\mathrm{T}}(\theta)\nabla r(\theta) = \\ &(1-\alpha)\nabla\pi^{\mathrm{T}}(\theta) J_\alpha(\theta) + \\ &\pi^{\mathrm{T}}(\theta)(\nabla r(\theta) + \alpha\nabla\boldsymbol{P}(\theta) J_\alpha(\theta))\end{aligned}$$

证毕.

若给定参数向量 θ, 则 $\pi(\theta)$, $J_\alpha(\theta)$ 也就唯一确定, 从而可由定理 2 求出梯度 $\nabla\lambda(\theta)$, 再用梯度下降法

$$\theta_{k+1} = \theta_k + \gamma_k \nabla\lambda(\theta_k)$$

便可求出使目标函数 $\lambda(\theta)$ 极大的参数 θ. 其中 γ_k 为步长. 但求解 $\pi(\theta)$, $J_\alpha(\theta)$ 较困难, 有时甚至是不可能的(如当系统模型全部或部分未知时). 此时若能求出梯度 $\nabla\lambda(\theta)$ 的一个带噪声的观察值 $d(\theta)$, 就可由如下的随机梯度法求解

$$\theta_{k+1} = \theta_k + \gamma_k d(\theta_k).$$

下面讨论如何给出恰当 $d(\theta)$ 的的表达式以估计梯度 $\nabla\lambda(\theta)$.

引理 **3** $\lim_{\alpha\uparrow 1}(1-\alpha) J_\alpha(\theta) = e\lambda(\theta).$ (6)

证明见文献[5].

定理 **3**　设

$$D(\alpha,\theta)=\pi(\theta)^{\mathrm{T}}(\nabla r(\theta)+\alpha\boldsymbol{P}(\theta)J_\alpha(\theta),\tag{7}$$

则有

$$\lim_{\alpha\uparrow 1}D(\alpha,\theta)=\nabla\lambda(\theta).$$

证明　由 $\pi^{\mathrm{T}}(\theta)e=1$ 可知

$$\nabla\pi^{\mathrm{T}}(\theta)e=0.$$

于是, 由定理 2

$$\begin{aligned}\nabla\lambda(\theta)&=(1-\alpha)\nabla\pi^{\mathrm{T}}(\theta)J_\alpha(\theta)+\\&\quad\pi^{\mathrm{T}}(\theta)(\nabla r(\theta)+\alpha\nabla\boldsymbol{P}(\theta)J_\alpha(\theta))=\\&\quad\nabla\pi^{\mathrm{T}}(\theta)(1-\alpha)J_\alpha(\theta)+D(\alpha,\theta),\end{aligned}$$

两边取 α 单增趋于 1 的极限并注意到(6) 式得

$$\begin{aligned}\nabla\lambda(\theta)&=\nabla\pi^{\mathrm{T}}(\theta)e\lambda(\theta)+\lim_{\alpha\uparrow 1}D(\alpha,\theta)=\\&\quad\lim_{\alpha\uparrow 1}D(\alpha,\theta),\qquad\text{证毕.}\end{aligned}$$

由定理 3 可知, 当 α 趋于 1 时, $D(\alpha,\theta)$ 便是 $\nabla\lambda(\theta)$ 的近似. 下面考虑如何通过一样本路径来逼近 $D(\alpha,\theta)$. 首先对转移概率 $p_{ij}(\theta)$ 作如下假设:

假设**3**　对任意 $i,j\in S$ 和 $\theta\in R^K$, 存在有界函数 $L_{ij}(\theta)$, 使得

$$\nabla p_{ij}(\theta)=p_{ij}L_{ij}(\theta).$$

易见, 当 $p_{ij}\neq 0$ 时, $L_{ij}(\theta)=\dfrac{\nabla p_{ij}(\theta)}{p_{ij}(\theta)}=\nabla\ln p_{ij}(\theta)$, 相当于梯度似然比. 另外, 当存在 $\varepsilon>0$ 使对任意 $i,j\in S$ 都有

$$p_{ij}(\theta)=0\text{ 或者 }p_{ij}(\theta)\geqslant\varepsilon$$

的话, 则假设 3 成立.

将(7) 式展开

$$\begin{aligned}&D(\alpha,\theta)=\\&\sum_{i=1}^{N}\pi_i(\theta)\left(\nabla r(i,\theta)+\alpha\sum_{j=1}^{N}p_{ij}(\theta)J_\alpha(j,\theta)\right)=\\&\sum_{j=1}^{N}\pi_i(\theta)\nabla r(i,\theta)+\alpha\sum_{i=1}^{N}\sum_{j=1}^{N}\pi_i(\theta)p_{ij}(\theta)J_\alpha(j,\theta).\end{aligned}$$

设给定参数向量 $\theta\in R^K$ 和折扣因子 $\alpha\in(0,1)$, 状态转移概率矩阵为 $\boldsymbol{P}(\theta)$ 的一个样本状态系列为$\{i_0,i_1,i_2,\cdots,i_n,\cdots\}$. 由渐进平稳性, 若任意的 $i,j\in S$, $r(i,\theta)$ 和 $L_{ij}(\theta)$ 是已知的. 这样, 上式右端的第 1 项便可用 $\nabla r(i_0,\theta)$, $\nabla r(i_1,\theta)$, … 的平均来逼近; 同样, 第 2 项也可视为 $L_{i_0i_1}(\theta)J_\alpha(i_1)$, $L_{i_1i_2}(\theta)J_\alpha(i_2)$, … 的平均的 α 倍.

定理 **4**　设假设 **3** 成立, 由 $\boldsymbol{P}(\theta)$ 产生的一个样本状态序列为$\{i_0,i_1,i_2,\cdots,i_n,\cdots\}$, 令

$$\begin{aligned}&d_T(\alpha,\theta)=\\&\frac{1}{T}\sum_{t=0}^{T-1}\left[\nabla r(i,\theta)+\alpha L_{i_1i_{t+1}}(\theta)\sum_{s=t+1}^{T}\alpha^{s-t-1}r(i_s,\theta)\right],\end{aligned}$$

则以概率 1

$$\lim_{T\to\infty}d_T(\alpha,\theta)=D(\alpha,\theta)$$

成立.

证明　设转移概率矩阵 $\boldsymbol{P}(\theta)$ 产生的马氏链为 $\{X_0,X_1,X_2,\cdots\}$, 而$\{i_0,i_1,i_2,\cdots\}$ 为一样本路径. 由假设 1 知样本序列$\{i_0,i_1,i_2,\cdots,i_n,\cdots\}$ 是渐进平稳的. 由遍历理论知, 等式(8) 和(9) 以概率 1 成立.

$$E_\theta(\nabla r(X_t,\theta))=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\nabla r(i_t,\theta),\tag{8}$$

$$\begin{aligned}&E_\theta(L_{X_tX_{t+1}}(\theta)J_\alpha(X_{t+1},\theta))=\\&\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}L_{i_ti_{t+1}}(\theta)J_\alpha(i_{t+1},\theta),\end{aligned}\tag{9}$$

又

$$\begin{aligned}&E_\theta(\nabla r(X_t,\theta))=\sum_{i=1}^{N}\pi(i,\theta)\nabla r(i,\theta)=\\&\qquad\pi^{\mathrm{T}}(\theta)\nabla r(\theta),\\&E_\theta(L_{X_tX_{t+1}}(\theta)J_\alpha(X_{t+1},\theta))=\\&\sum_{i=1}^{N}\sum_{j=1}^{N}\pi(i,\theta)p_{ij}(\theta)L_{ij}(\theta)J_\alpha(j,\theta)=\\&\pi^{\mathrm{T}}(\theta)\nabla\boldsymbol{P}(\theta)J_\alpha(\theta),\end{aligned}$$

所以, (8), (9) 式变为

$$\pi^{\mathrm{T}}(\theta)\nabla r(\theta)=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\nabla r(i_t,\theta),\quad\text{w.p.1}\tag{10}$$

$$\begin{aligned}&\alpha\pi^{\mathrm{T}}(\theta)\nabla\boldsymbol{P}(\theta)J_\alpha(\theta)=\\&\alpha\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}L_{i_ti_{t+1}}(\theta)J_\alpha(i_{t+1},\theta),\end{aligned}\quad\text{w.p.1}\tag{11}$$

另一方面

$$\begin{aligned}&\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}L_{i_ti_{t+1}}(\theta)J_\alpha(i_{t+1},\theta)=\\&\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}L_{i_ti_{t+1}}(\theta)\cdot\\&\left(\sum_{s=t+1}^{T}\alpha^{s-t-1}r(i_s,\theta)+\sum_{S=T+1}^{\infty}\alpha^{s-t-1}r(i_s,\theta)\right),\end{aligned}$$

由假设 2 和假设 3 可知, 上式的右端展开后第 2 项满足

$$\begin{aligned}&\left|\frac{1}{T}\sum_{t=0}^{T-1}L_{i_ti_{t+1}}(\theta)\sum_{s=T+1}^{\infty}\alpha^{s-t-1}r(i_s,\theta)\right|\leqslant\\&\frac{1}{T}\sum_{t=0}^{T-1}|L_{i_ti_{t+1}}(\theta)|\sum_{s=T+1}^{\infty}\alpha^{s-t-1}|r(i_s,\theta)|\leqslant\end{aligned}$$

$$\frac{C}{T}\sum_{t=0}^{T-1}\sum_{s=T+1}^{\infty}\alpha^{s-t-1}=$$

$$\frac{C}{T}\frac{1-\alpha^{T}}{(1-\alpha)^{2}}\to 0\quad (T\to\infty).$$

于是

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}L_{i_t i_{t+1}}(\theta)J_{\alpha}(i_{t+1},\theta)=$$

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}L_{i_t i_{t+1}}(\theta)\sum_{s=t+1}^{T}\alpha^{s-t-1}r(i_s,\theta),$$

将其代入(11)并加上(10)即得

$$D(\alpha,\theta)=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\Big[\nabla r(i_t,\theta)+\alpha L_{i_t i_{t+1}}(\theta)\sum_{s=t+1}^{T}\alpha^{s-t-1}r(i_s,\theta)\Big]=\lim_{T\to\infty}d_T(\alpha,\theta),$$

证毕.

于是,本文给出的基于随机梯度的参数 MDP 求解方法如下:

给定正整数 T,折扣因子 $\alpha\in(0,1)$ 及 k 时刻的参数向量 $\theta_k\in R^K$,如下产生下一个参数向量

$$\theta_{k+1}=\theta_k+\gamma_k d_T(\alpha,\theta_k),$$

其中,$d_T(\alpha,\theta_k)$ 由(7)式所确定,γ_k 为步长,满足

$$\sum_{k=0}^{\infty}\gamma_k=\infty \text{ 和 } \sum_{k=0}^{\infty}\gamma_k^2<\infty,$$

一个易实现的求 $d_T(\alpha,\theta_k)$ 的递推算法如下:

设由转移概率矩阵 $\boldsymbol{P}(\theta)$ 产生的样本路径为:$\{i_0,i_1,i_2,\cdots,i_T\}$.

(i) 令 $z_0=0$, $d_0=0$, $z_0,d_0\in R^K$;

(ii) $t=0$ 到 $T-1$,循环做

$$\begin{cases} z_{t+1}=\alpha(z_t+L_{i_t i_{t+1}}(\theta)),\\ d_{t+1}=d_t+\dfrac{1}{t+1}(\nabla r(i_{t+1},\theta_k)+r(i_{t+1},\theta_k)z_{k+1}-d_t);\end{cases}$$

(iii) 返回 $d_T(\alpha,\theta_k)=d_T$.

说明:

(a) 理论上 α 应十分接近 1,但实际上 α 并不需要很接近 1(如 0.9 即可),否则还会使算法方差很大.

(b) 每次返回 d_T 后,下一样本路径的初始状态 i_0 可取为上一样本路径的终止状态 i_T. 事实上,由于步长很小,θ_{k+1} 的 θ_k 差异不大,即平稳分布变化不大,这样可保证样本状态较快服从平稳分布.

参考文献:

[1] WATKINS C J C H, DAYAN P. Q-learning[J]. Machine Learning, 1992, 8: 279—292.

[2] WILLIAMS R L. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine Learning, 1992, 8: 229—256.

[3] MARBACH P, TSITSIKLIS J N. Simulation-based optimization of Markov reward processes[J]. IEEE Transactions on Automatic Control, 2001, 46(2): 191—209.

[4] KONDA V R, BORKAR V S. Actor-critic type learning algorithms for Markov decision processes[J]. SIAM J Control Optimization, 1999, 38(1): 94—123.

[5] ARAPOSTATHIS A, BORKAR V S, Fernandez-Gaucherand E, et al. Discrete-time controlled Markov processes with average cost criterion: a survey[J]. SIAM J Control Optimization, 1993, 31(2): 282—344.

A stochastic approximation for parameters Markov decision processes

HU Guang-hua

(Department of Mathematics, Yunnan University, Kunming 650091, China)

Abstract: A stochastic gradient algorithm for average reward Markov decision processes (MDP) that depends on a parameter vector is proposed. A new gradient of the object function is given and a stochastic approximation algorithm that bases on a single sample path is presented. Finally, a convergence of the gradient (with probability 1) is provided.

Key words: parameters Markov decision processes; stochastic approximation; average rewards

MSC(2000): 90C40