

离散型冲击折扣半马氏决策过程*

胡奇英

(西安电子科技大学, 710071)

摘要 本文讨论离散型冲击折扣半马氏决策过程, 在建立模型后, 我们将它化成了一个等价的离散时间马氏决策过程.

关键词 半马氏决策过程, 折扣准则, 冲击.

分类号 AMS(1991) 90C40/CC1 O221.5

1 模型的提出

传统马氏决策过程(Markov decision processes, 简记MDP)不考虑环境对系统的影响, 但在很多实际问题中系统可用一MDP表示, 而环境变化将导致系统性能的变化, 于是描述系统的模型参数亦将发生变化. 我们在[1], [2]中分别用状态可数的连续时间冲击MDP、冲击SMDP来描述这种系统. 本文研究Borel状态的离散型冲击SMDP, 其模型为

$$\{f^*(\cdot), S, A^*, \Gamma^*, p^*, h^*, r^*, q^*, R^*, \beta; n \geq 0\}. \quad (1)$$

各元含义如下:

a) 对 $n \geq 0$, ξ_n 为环境对系统的第 n 次冲击时刻, $0 = \xi_0 < \xi_1 < \xi_2 < \dots$, $\Delta \xi_n = \xi_{n+1} - \xi_n$ 是互相独立取整数值随机变量, $f^*(k) = p\{\Delta \xi_n = k\}$ ($k > 0$), $f^*(\infty) \geq 0$

b) 对 $n \geq 0$, 系统在 $[\xi_n, \xi_{n+1})$ 上可用离散型 SMDP(n):

$$\{S, A^*, \Gamma^*, p^*, h^*, r^*\} \quad (2)$$

表示, 其状态集 S , 决策集 A^* 均是非空 Borel 集, 约束集 Γ^* 是 $S \times A^*$ 的解析子集, 对 $s \in S$, $\Gamma_s^* = \{a \in A^* | (s, a) \in \Gamma^*\}$ 非空; 状态转移概率 $p^*(ds' | s, a)$ 是给定 $S \times A^*$ 时 S 上的 Borel 可测随机核; 转移时间为 k 个单位的概率是 $h^*(k | s, a, s')$, $\sum_{k=1}^{\infty} h^*(k | s, a, s') = 1$; $r^*(k | s, a, s')$ 则是转移时间为 k 个单位时于转移开始时所获报酬. 假定 h^*, r^* 均是下半解析函数, 详细含义见[3].

c) 对 $n \geq 0$, $B \in \mathcal{B}(S)$ (S 上的 Borel 集全体), $q^*(B | s) = p\{\xi_{n+1} \text{ 时状态在 } B \text{ 中} | \xi_n = 0 \text{ 时的状态为 } s\}$, 即 q^* 是在第 $n+1$ 次冲击发生时刻的瞬时状态转移概率, 为 S 上的随机核. 假定如 ξ_{n+1} 也刚好是 SMDP(n) 的状态转移时刻, 则只考虑冲击引起的状态转移.

d) 如系统在 $\xi_{n+1} = 0$ 时的状态为 s , 所采取的决策为 a , 则在 $\xi_{n+1} = 0$ 时获得一项报酬 $R^*(s, a)$, 它为 Γ^* 上的下半解析函数. $\beta \in (0, 1)$ 为折扣因子.

* 1992年4月9日收到. 94年3月31日收到修改稿. 国家自然科学基金资助项目.

对 $n, k \geq 0, s \in S$, 记状态 $x = (n, k, s)$ 表示系统在 $\xi_n + k$ 时处于状态 s 且 $\Delta \xi_n > k$, 考虑

$$h_m = (x_0, a_0, t_0, x_1, a_1, t_1, \dots, x_m), \quad (3)$$

其中 a_i, t_i 分别为系统在 $x_i = (n_i, k_i, s_i)$ 时采取的决策和转移时间. h_m 要成为系统在 $\xi_{s_0} + k_0$ 时从 s_0 出发的一个历史当且仅当它满足如下规则

规则 对 h_m 如(3), $l = 0, 1, \dots, m-1$, 或者 $n_{l+1} = n_l$ 且 $k_{l+1} = k_l + t_l$, 或者 $n_{l+1} = n_l + 1$ 且 $k_{l+1} = 0$.

下文总假定历史 h_m 满足如上规则, 策略 $\pi \in \Pi$, 半马氏策略集 Π_{sm} , 随机马氏策略集 Π_m , 随机平稳策略集 Π_s 等同[3]中. 定义决策函数集为 $F = \{f: \text{对 } n, k \geq 0, f(n, k, \cdot) \text{ 是从 } S \text{ 到 } A^*$ 的泛可测映射, 满足 $f(n, k, s) \in F_s^*\}$. 对 $\pi \in \Pi$, 不难构造出 π 下的概率空间 $(\Omega, \mathcal{F}, P_\pi)$.

对 $j \geq n, m \geq 0$, 记 X_m^j, Δ_m^j 分别为 $[\xi_j, \xi_{j+1})$ ($j=n$ 时为 $[\xi_n + k, \xi_{n+1})$) 中第 m 次状态转移后所处的状态和采取的决策, t_m^j 为在 X_m^j 处的停留时间. 令 $T_0^j = k, T_0^j = 0$ ($j > n$), $T_m^j = T_{m-1}^j + t_{m-1}^j$ ($m \geq 0, j \geq n$). 再定义 N_s, N_j ($j > n$) 分别为在 $[\xi_n + k, \xi_{n+1}), [\xi_j, \xi_{j+1})$ 中的状态转移次数 (不包括冲击产生的状态转移). 对 $n \cdot k \geq 0$, 设 $\Delta \xi_n > k$, 记 $V_n(k)$ 为系统在 $[\xi_n + k, \infty)$ 上获得的折扣到 $\xi_n + k$ 时的折扣总报酬, 则

$$V_n(k) = \sum_{j=n}^{\infty} \beta^{T_j^j - \xi_n - k} V_j, \quad (4)$$

$$V_j = \sum_{i=0}^{N_j-1} \beta^{T_i^j} r^j(k_i^j | X_i^j, \Delta_i^j, X_{i+1}^j) + \beta^{T_{N_j}^j} r^j(\Delta \xi_j - T_{N_j}^j | X_{N_j}^j, \Delta_{N_j}^j, X_{N_j+1}^j) + \beta^{T_{N_j}^j} R^j(X_{N_j}^j, \Delta_{N_j}^j). \quad (5)$$

现在定义折扣目标函数为

$$V(\pi, n, k, s) = \bar{P}^n(k) E_\pi \{V_n(k) | X_0 = (n, k, s)\}, \quad (6)$$

其中 $\bar{P}^n(k) = P\{\Delta \xi_n > k\}$, X_m, Δ_m, t_m 分别为系统从 X_0 出发第 m 次转移后所处的状态、采取的决策和转移时间. 令 $T_m = t_0 + t_1 + \dots + t_{m-1}$.

本文假定 r^*, R^* 一致有界, 于是由有界收敛定理知 $V(\pi, n, k, s)$ 存在、一致有界且关于 s 泛可测. 记 $V^*(n, k, s) = \sup_{\pi \in \Pi} V(\pi, n, k, s)$, (π) 最优策略与通常定义.

2 化为离时间 MDP

对 $n, k \geq 0, s \in S, a \in A^*$, 记

$$\begin{aligned} r(n, k, s, a) &= \int_S p^s(ds' | s, a) \sum_{l=1}^{\infty} h^s(l | s, a, s') \{ \bar{P}^s(k+l) r^s(l | s, a, s') \\ &\quad + \sum_{j=1}^l f^s(k+j) [r^s(j | s, a, s') + \beta^j R^s(s, a)] \}, \\ \beta(n, k, s, a) &= \int_S p^s(ds' | s, a) \sum_{l=1}^{\infty} h^s(l | s, a, s') \sum_{j=1}^l \beta^j f^s(k+j). \end{aligned}$$

由于篇幅所限, 以下均省略证明.

定理 1

$$V(\pi, n, k, s) = \int_{A^*} \pi_0(da | n, k, s) \{ r(n, k, s, a) - 120 -$$

$$+ \int_S p^*(ds' | s, a) \sum_{t=1}^{\infty} h^*(t | s, a, s') [\beta^t V(\pi^{(n,k,s),a,t}, n, k+t, s')] \\ + \sum_{j=1}^l f^*(k+j) \beta^j \int_S q^*(ds'' | s) V(\pi^{(n,k,s),a,t}, n+1, 0, s'')], \pi, n, k, s, \quad (7)$$

其中 $\pi^{(n,k,s),a,t} = (\pi'_0, \pi'_1, \dots)$; $\pi'_m(\cdot | h_m) = \pi_{m+1}(\cdot | (n, k, s), a, t, h_m)$ (当 $(n, k, s), a, t, h_m$ 不满足规划时 $\pi'_m(\cdot | h_m)$ 可任意定义).

定理 2 给定 $\pi \in \Pi$ 及 (n, k, s) , 定义 $\pi^* \in \Pi_n$ 如下:

$$\pi_m^*(C | n', k', s') = P_{\pi}\{X_m \in C | X_0 = (n, k, s), X_m = (n', k', s')\}, \\ m \geq 0, (n', k', s'), C \in \mathcal{B}(A^m) \quad (8)$$

则对任意的 $n' \geq n, k' \geq 0, B \in \mathcal{B}(S), C \in \mathcal{B}(A^m), m \geq 0$ 有

$$P_{\pi}\{X_m \in (n', k', B), \Delta_m \in C | X_0 \in (n, k, s)\} \\ = P_{\pi^*}\{X_m \in (n', k', B), \Delta_m \in C | X_0 = (n, k, s)\} \quad (9)$$

定义策略集 $\bar{\Pi} = \{\pi \in \Pi: \pi_m(\cdot | h_m) \text{ 与 } h_m \text{ 中的 } t_0, t_1, \dots, t_{m-1} \text{ 无关}\}$, 则由定理 2 知 $V^*(n, k, s) = \sup_{\pi \in \bar{\Pi}} V(\pi, n, k, s) = \sup_{\pi \in \Pi} V(\pi, n, k, s)$, 于是只须在 $\bar{\Pi}$ 中讨论. 由定理 1 可得:

推论 1

$$V(\pi, n, k, s) = \int_A \pi_0(da | n, k, s) \{r(n, k, s, a) + \beta(n, k, s, a) \int_S q^*(ds' | s) V(\pi^{(n,k,s),a}, n+1, 0, s') \\ + \int_S p^*(ds' | s, a) \sum_{t=1}^{\infty} h^*(t | s, a, s') \beta^t V(\pi^{(n,k,s),a,t}, n, k+t, s')\}, \pi \in \bar{\Pi}. \quad (10)$$

基于上述讨论, 引入离散时间 MDP 模型

$$\{\bar{S}, \bar{A}, \bar{F}, \bar{p}, \tau, \beta\}, \quad (11)$$

其中 $\bar{S} = \{(n, k, s): n, k \geq 0, s \in S\}$, $\bar{A} = \bigcup_{n,k \geq 0} A^*$, $\bar{F}$ 满足 $\bar{F}_{n,k,s} = \{a: (n, k, s, a) \in \bar{F}\} = F_{n,k,s}^*$, 报酬函数为 $\tau(n, k, s, a), \beta$ 同(1), 而

$$\bar{p}(n', k', ds' | n, k, s, a) = \begin{cases} \beta(n, k, s, a) q^*(ds' | s) / \beta, & \text{如 } n' = n+1, k' = 0, \\ p^*(ds' | s, a) h^*(t | s, a, s') \beta^{t-1}, & \text{如 } n' = n, k' = k+t, t \geq 1, \\ 0, & \text{其它.} \end{cases} \quad (12)$$

显然, (11) 的策略集同 $\bar{\Pi}$, 目标函数 $V_{\beta}(\pi, n, k, s)$ 与通常一样定义. 由于 τ 一致有界, $V_{\beta}(\pi)$ 存在, 一致有界且泛可测.

定理 3 对 $\pi \in \bar{\Pi}$, $V_{\beta}(\pi, n, k, s)$ 也满足(10)式.

定理 4 (1)和(11)在如下意义上等价: a) 对 $\pi \in \Pi$, 存在 $\bar{\pi} \in \bar{\Pi}$ 使 $V(\pi) = V(\bar{\pi})$; b) 对 $\pi \in \bar{\Pi}$, $V(\pi) = V_{\beta}(\pi)$.

由定理 4, 我们即可将离散时间 MDP 中的结果直接推广到离散冲击 SMDP 中来. 如由[3]中定理 1, 2, 7 可得

定理 5 i) $V^*(n, k, s)$ 是如下最优方程的唯一的一致有界下半解析解:

$$V(n, k, s) = \sup_{a \in A_s^*} \{r(n, k, s, a) + \beta(n, k, s, a) \int_S q^*(ds' | s) V(n+1, 0, s') \\ + \int_S p^*(ds' | s, a) \sum_{t=1}^{\infty} h^*(t | s, a, s') \beta^t V(n, k+t, s')\}; \quad (13)$$

ii) 对 $\varepsilon \geq 0$, 取到(13)中 ε 上确界的 f 为 $(1-\beta)^{-1}\varepsilon$ 最优策略. $\varepsilon > 0$ 时如此 f 必存在.

下面讨论周期情形, 对 $N > 0$, 如(1)中各元对 n 均有周期 N , 则称(1)有周期 N , 周期为 1 时称之为平稳的. 对周期情形我们有以下结论.

定理 6 若模型(1)有周期 N , 则 $V^*(n, k, s)$ 也有周期 N , 且是(13)的有周期 N 的唯一一致有界的下半解析解.

3 结 论

[2]中讨论的一般冲击 SMDP 不能化为离散时间 MDP, 而模型(1)却可以, 这是它作为一般情形特例所具有的特殊性质. 所化成的(11)也有别于一般的 MDP, 其状态形为 (n, k, s) , 而 n, k 取值于 $\{0, 1, 2, \dots\}$, 其变化服从“规则”和 $f(\cdot)$, 于是可以用[4]中提出的“有限状态逼近可数状态”来讨论近似计算问题. 另一方面, 用(1)来逼近一般的冲击 SMDP 也将是一个很有吸引力的问题.

参 考 文 献

- [1] Hu, Q., *Continuous times shock markov decision processed with discounted criterion*, Optimization, Vol. 25 (1992), 271—285.
- [2] 胡奇英, 随机冲击下的折扣半马氏决策规划, 应用数学学报, Vol. 17, No. 4 (1994)
- [3] S. E. Shreve and D. P. Bertsekas, *Universally measurable policies in dynamic programming*, Math. Ope. Res., Vol. 4 (1979), 15—30.
- [4] D. J. White, *Finite-state approximation for denumerable-state infinite horizon constructed MDP: The Policy Method*, J. Math. Anal. Appl., Vol. 72 (1979), 512—523.

Discrete Type Shook Semi-Markov Decision Processes with Discounted Cirterion

Hu Qiying

(Xidian University, Xi'an 710071)

Abstract

This paper deals with the discrete type shook discounted semi-Markov decision processes. After presenting the model, we transform it into an equivalent discrete time Markov decision processes with discounted cirterion.

Keywords Markov process, Markov decision process, Shook semi-Markov decision process.