第 34 卷 第 6 期 应 用 数 学 学 报 Vol. 34 No. 6
2011 年 11 月 ACTA MATHEMATICAE APPLICATAE SINICA Nov., 2011

离散时间马氏决策过程的首达目标准则 *

刘秋丽

(华南师范大学数学科学学院 510631)

(E-mail: liuql2007@yahoo.cn)

摘 要 本文考虑可数状态离散时间马氏决策过程的首达目标模型的风险概率准则. 优化的准则是最小化系统首次到达目标状态集的时间不超过某阈值的风险概率. 首先建立最优方程并且证明最优值函数和最优方程的解对应, 然后讨论了最优策略的一些性质, 并进一步给出了最优平稳策略存在的条件, 最后用一个例子说明我们的结果.

关键词 目标集; 首达时; 风险概率

MR(2000) 主题分类 90C40; 93E20

中图分类 0211.62; 0231.3; 0232

1 引言

在马氏决策过程 (Markov Decision Processes, 简记为 MDPs) 领域中, 受到广泛研究的是期望折扣和期望平均准则. 然而, 上述准则对于刻画某些实际问题的变化特征是不够的. 为了弥补这些准则的不足, 首达目标模型已有许多学者研究, 例如 [1–7] 以及扩展的文献. 这样的优化模型在实际中已经深入到设备维修、排队系统、可靠性工程、风险分析等众多领域. 众所周知, 首达目标准则是最受欢迎的准则之一, 它最大化首次到达某目标集时间内的期望总报酬. Derman[1] 考虑目标集是吸收态的有限状态和行动的离散时间 MDPs 的首达目标准则. 刘等 [3,4] 进一步考察可数状态空间有界报酬的离散时间首达目标模型. 此外, 黄等 [2] 讨论可数状态非负费用的折扣半马氏决策过程的首达目标准则. 一些学者考虑形式为 $\inf\limits_{\pi} P_i^{\pi}\ (\tau_B \le \lambda)$ (等价形式为 $\sup\limits_{\pi} P_i^{\pi}(\tau_B > \lambda)$) 的 MDPs 中特殊效应准则 [5,6] (其中 i 是初始状态, π 是策略, λ 是阈值, τ_B 是给定目标集 B 的首达时). 林等 [6] 得到了连续时间 MDPs 最优策略存在的最优条件并且给出了最优策略的算法. 刘等 [5] 考察形式为 $\sup\limits_{\pi} P_i^{\pi}\ (\tau_B \ge \lambda)$ 的离散时间 MDPs 的优化问题. 他们建立了最优方程, 讨论了最优策略的性质, 证明当最优行动集可数 (有限或者

本文 2009 年 7 月 15 日收到. 2011 年 4 月 21 日收到修改稿.

* 广东省高校优秀青年创新人才培养计划资助项目

无限) 交集非空时最优平稳策略存在. 但这样的最优条件不易验证并且在一般的情形下最优平稳策略可能不存在.

本文考察形式为 $\inf\limits_{\pi} P_i^{\pi}\ (\tau_B \le \lambda)$ 的可数状态和行动空间的离散时间 MDPs 的优化问题，它与 [5] 中的准则相似，是 [7] 中准则的特殊形式. 与 [5,7] 相比，本文有以下特色：

(i) 本文的方法和 [5] 不同，但与 [7] 类似，本文中目标集由吸收态组成并且报酬是免费的 (细节见注 2.2);

(ii) 本文中策略依赖状态，行动和阈值 λ, 而 [5] 中的策略独立于阈值 λ; (引入依赖阈值的策略的好处可以从下面的 (iii) 和 (iv) 看出)

(iii) 我们去掉 [5] 中要求的最优行动集可数 (有限或者无限) 交集非空的条件，并在所有的随机马氏策略类中证明了最优平稳策略的存在性;

(iv) [5] 中的例子 2.2 表明 [5] 的结果和本文结果的区别 (细节见注 6.1).

具体来说，我们首先引入依赖状态、行动和阈值 λ 的策略类，然后建立最优方程并且讨论最优策略的性质以及最优平稳策略存在的条件，并给出独立于阈值的最优策略存在的充分必要条件，更进一步，用 [5] 中的例子说明我们结果的应用.

本文第 2 部分介绍受控模型；第 3 部分讨论风险函数的性质并且建立最优方程，第 4 部分讨论最优条件，第 5 部分给出算法，并在第 6 部分用一个数值算例说明如何应用结果.

2　受控模型

本文所考虑的离散时间 MDPs 的模型如下：

$$\{S, (A(i) \subseteq A, i \in S), B, p(j|i,a)\}, \tag{2.1}$$

其中 S 和 A 分别是可数的状态空间和行动空间， $B \subset S$ 是一给定的目标集， $A(i)$ 表示在状态 $i \in S$ 时的允许行动集. 令 $K := \{(i,a) : i \in S,\ a \in A(i)\}$ 是允许状态 - 行动集. (2.1) 中的 $p(j|i,a)$ 是满足方程 $\sum\limits_{j\in S} p(j|i,a) = 1$ 的转移概率.

下面描述风险概率准则下离散时间 MDPs 的演化过程. 决策者在时刻 $t = 0, 1, \cdots$ 观察到受控随机系统的状态. 假设在初始决策时刻系统处于状态 i_0, 且决策者有一个目标 (阈值) λ_0, 即他要尽量避免系统在 λ_0 单位时间内到达目标集的风险. 于是, 决策者根据当前状态 i_0 和他的目标 λ_0 选取一个行动 a_0, 由于行动的选取, 系统以概率 $p(i_1|i_0,a_0)$ 转移到状态 i_1, 接着新的决策时刻发生. 决策过程以这种方式演化, 从而我们可以得到离散时间 MDPs 的第 n 个决策时刻的允许历史 h_n, 即 $h_n = (i_0, \lambda_0, a_0, i_1, \lambda_1, a_1, \cdots, a_{n-1}, i_n, \lambda_n)$. 令 H_n 表示所有历史 h_n 组成的集合.

下面引入策略的定义.

定义 2.1 依赖于历史的随机策略 $\pi = \{\pi_n,\ n = 0, 1, \cdots\}$ 是满足以下条件的随机核序列：每个 π_n 是给定 H_n 时定义在 A 上的随机核，且 $\pi_n(A(i_n)|h_n) = 1,\ \forall\, h_n \in H_n,\ n \ge 0$. 所有策略构成的集合记为 Π.

注记 2.1 由于历史 h_n 依赖阈值 λ_n, 状态 i_n 以及行动 a_n, 此处的策略与 [5] 中的策略不同，但是与 [7] 中的相似.

令 Φ 表示所有给定 $S\times Z^+$ (Z^+ 表示非负整数的集合) 时定义在 A 上的随机核 φ 组成的集合，对任意 $(i,\lambda)\in S\times Z^+$, 其满足 $\varphi(A(i)|i,\lambda)=1$. F 表示决策函数 f 之集，其中 $f: S\times Z^+\to A$ 满足对任意的 $(i,\lambda)\in S\times Z^+$, $f(i,\lambda)\in A(i)$. 策略 $\pi=\{\pi_n\}$ 称为随机马氏的，如果存在随机核序列 $\{\varphi_n\}$ ($\varphi_n\in\Phi$), 使得对每个 $h_n\in H_n$ 和 $n\geq 0$, $\pi_n(\cdot|h_n)=\varphi_n(\cdot|i_n,\lambda_n)$. 我们把随机马氏策略记为 $\pi=\{\varphi_n\}$. 随机马氏策略 $\pi=\{\varphi_n\}$ 称为随机平稳的，如果 φ_n 与 n 无关. 此时，我们将 $\pi=\{\varphi,\varphi,\cdots\}$ 记为 φ. 进一步，随机马氏策略 $\pi=\{\varphi_n\}$ 称为确定性的，若存在序列 $\{f_n\}$ 使得对每个 $(i,\lambda)\in S\times Z^+$, $n\geq 0$, $\varphi_n(\{f_n(i,\lambda)\}|i,\lambda)=1$. 相似地，我们记这样的策略为 $\pi=\{f_n\}$. 特别地，确定性马氏策略 $\pi=\{f_n\}$ 称为平稳的，若 $\{f_n\}$ 与 n 无关. 为方便，我们将 $\pi=\{f,f,\cdots\}$. 我们分别用 Π_{RM},Π_{RS}, Π_{DM}, Π_{DS}, 表示所有随机马氏策略，随机平稳策略，确定性马氏策略及确定性平稳策略之集. 显然， $\Pi_{RS}\subset\Pi_{RM}\subset\Pi$, $\Pi_{DS}\subset\Pi_{DM}\subset\Pi$.

令 Π_0 表示所有不依赖阈值 λ 的策略之集. 若策略 $\pi=\{f_n\}\in\Pi_{DM}\cap\Pi_0$, 那么对所有的 $(i,\lambda)\in S\times Z^+$ 和 $n\geq 0$, $f_n(i,\lambda)\equiv f_n(i)$. 对策略 $\pi=\{\varphi_0,\varphi_1,\cdots\}\in\Pi_{RM}$ 及 $m\geq 1$, ${}^{(m)}\pi:=\{\varphi_m,\varphi_{m+1},\cdots\}$ 表示 π 的 m- 剩余策略.

对每个 $(i,\lambda)\in S\times Z^+$ 和 $\pi\in\Pi$, 由 [8] 知，存在唯一的概率空间 $(\Omega,\mathcal{F},P^\pi_{(i,\lambda)})$ 和随机过程 $\{x_n,\lambda_n,\Delta_n,\ n\geq 0\}$ 使得对每个 $j\in S$, $a\in A(i)$ 和 $n\geq 0$, 有

$$P^\pi_{(i,\lambda)}(x_0=i,\lambda_0=\lambda)=1, \tag{2.2}$$

$$P^\pi_{(i,\lambda)}(\Delta_n=a|h_n)=\pi_n(a|h_n), \tag{2.3}$$

$$P^\pi_{(i,\lambda)}(x_{n+1}=j|h_n,a_n)=p(j|i_n,a_n), \tag{2.4}$$

其中 $\lambda_n:=\lambda_{n-1}-1$. 关于概率测度 $P^\pi_{(i,\lambda)}$ 的期望算子记为 $E^\pi_{(i,\lambda)}$.

给定目标集 $B\subset S$, 定义

$$\tau_B:=\begin{cases}\inf\{t\geq 0\mid x_t\in B\}, & \text{如果 } \{t\geq 0\mid x_t\in B\}\neq\emptyset,\\ +\infty, & \text{否则}\end{cases}$$

为目标集 B 的首达时. 显然，当 $x_0\in B$ 时， $\tau_B=0$.

定义 2.2 对每个 $(i,\lambda)\in S\times Z^+$ 和 $\pi\in\Pi$, 定义风险概率 (风险函数) 为

$$D^\pi(i,\lambda):=P^\pi_{(i,\lambda)}(\tau_B\leq\lambda) \tag{2.5}$$

及相应的最优值函数

$$D^*(i,\lambda):=\inf_{\pi\in\Pi}D^\pi(i,\lambda), \tag{2.6}$$

其中 λ 表示阈值.

另外，策略 $\pi^*\in\Pi$ 称为最优的，若

$$D^{\pi^*}(i,\lambda)=D^*(i,\lambda),\qquad \forall\,(i,\lambda)\in S\times Z^+. \tag{2.7}$$

注记 2.2

(1) 本文中目标状态不一定是吸收态，而 [7] 中的目标集是吸收态并且如果系统状态为目标状态时报酬为 0.

(2) 若 B 代表系统的失效状态集，则 τ_B 表示系统的失效时间，从而 $D^\pi(i,\lambda)$ 表示采取策略 π 时系统在 λ 单位时间里发生失效的风险概率.

注意到，对每个 $(i,\lambda)\in B\times Z^+$ 和 $\pi\in\Pi$，有 $D^\pi(i,\lambda)=1$. 为避免这种平凡情形，下文我们限制关于 $D^\pi(i,\lambda)$ 的讨论仅在当 $(i,\lambda)\in B^c\times Z^+$ 时的情形，其中 $B^c:=S-B$ 是 B 的补集. 由 [8] 中的定理 5.5.1 的结果，我们有 $D^*(i,\lambda)=\inf\limits_{\pi\in\Pi_{RM}} D^\pi(i,\lambda)$. 因此，我们以下的讨论局限于随机马氏策略类中.

3　最优方程

这一部分建立最优方程并且证明最优值函数是最优方程的解. 为记号方便，我们引入一些算子. 对任意的 $(i,\lambda)\in B^c\times Z^+$, 定义算子 T^a, T^φ 和 T 如下：

$$T^aD(i,\lambda):=\sum_{j\in B}p(j|i,a)+\sum_{j\in B^c}p(j|i,a)D(j,\lambda-1),\tag{3.1}$$

$$T^\varphi D(i,\lambda):=\sum_{a\in A(i)}\varphi(a|i,\lambda)T^aD(i,\lambda),\tag{3.2}$$

$$TD(i,\lambda):=\min_{a\in A(i)}T^aD(i,\lambda).\tag{3.3}$$

给出和证明主要结果之前，我们给出下列简单但重要的引理.

引理 3.1　存在决策函数 $f\in F$ 满足下列方程

$$T^fD(i,\lambda)=TD(i,\lambda).\tag{3.4}$$

证　因为 $A(i)$ 是有限的，存在映射 $f:B^c\times Z^+\to A$ 使得对任意的 $(i,\lambda)\in B^c\times Z^+$, 有 $f(i,\lambda)\in A(i)$ 和 $T^{f(i,\lambda)}D(i,\lambda)=TD(i,\lambda)$.

注记 3.1　为方便，记 $Z^+-\{1\}$ 为 U.

引理 3.2　设 $(i,\lambda)\in B^c\times U$，$\pi=(\varphi_0,\varphi_1,\cdots)\in\Pi_{RM}$, 有

$$D^\pi(i,\lambda)=T^{\varphi_0}D^{^{(1)}\pi}(i,\lambda).\tag{3.5}$$

特别地，当 $\pi=\varphi\in\Pi_{RS}$ 时，$D^\varphi(i,\lambda)=T^\varphi D^\varphi(i,\lambda)$.

证　由策略 π 的马氏性以及性质 (2.2)–(2.4), 有

$$\begin{aligned}
&D^\pi(i,\lambda)=P^\pi_{(i,\lambda)}(\tau_B\le\lambda)=1-P^\pi_{(i,\lambda)}(\tau_B>\lambda)\\
=&1-P^\pi_{(i,\lambda)}(x_0\in B^c,x_1\in B^c,\cdots,x_\lambda\in B^c)\\
=&1-E^\pi_{(i,\lambda)}\big[P^\pi_{(i,\lambda)}(x_0\in B^c,x_1\in B^c,\cdots,x_\lambda\in B^c|x_0,\lambda_0,\Delta_0,x_1,\lambda_1)\big]\\
=&1-E^\pi_{(i,\lambda)}\big[1_{\{x_0\in B^c,x_1\in B^c\}}P^\pi_{(i,\lambda)}(x_2\in B^c,\cdots,x_\lambda\in B^c|x_0,\lambda_0,\Delta_0,x_1,\lambda_1)\big]
\end{aligned}$$

$$
\begin{aligned}
&=1-\sum_{a\in A(i)}\varphi_0(a|i,\lambda)\sum_{j\in B^c}p(j|i,a)P^{\pi}_{(i,\lambda)}(x_2\in B^c,\cdots,x_\lambda\in B^c|x_0=i,\\
&\quad\lambda_0=\lambda,\Delta_0=a,x_1=j,\lambda_1=\lambda-1)\\
&=1-\sum_{a\in A(i)}\varphi_0(a|i,\lambda)\sum_{j\in B^c}p(j|i,a)P^{(1)\pi}_{(j,\lambda-1)}(x_0=j,x_1\in B^c,\cdots,x_{\lambda-1}\in B^c)\\
&=\sum_{a\in A(i)}\varphi_0(a|i,\lambda)\Big[\sum_{j\in S}p(j|i,a)-\sum_{j\in B^c}p(j|i,a)P^{(1)\pi}_{(j,\lambda-1)}(x_0\in B^c,x_1\in B^c,\cdots,x_{\lambda-1}\in B^c)\Big]\\
&=\sum_{a\in A(i)}\varphi_0(a|i,\lambda)\Big[\sum_{j\in B}p(j|i,a)+\sum_{j\in B^c}p(j|i,a)\\
&\quad\cdot\big[1-P^{(1)\pi}_{(j,\lambda-1)}(x_0\in B^c,x_1\in B^c,\cdots,x_{\lambda-1}\in B^c)\big]\Big]\\
&=\sum_{a\in A(i)}\varphi_0(a|i,\lambda)\Big[\sum_{j\in B}p(j|i,a)+\sum_{j\in B^c}p(j|i,a)D^{(1)\pi}(i,\lambda-1)\Big],
\end{aligned}
$$

这意味着 (3.5) 是正确的，即 $D^{\pi}(i,\lambda)=T^{\varphi_0}D^{(1)\pi}(i,\lambda)$. 最后一个结论是显然的，即在前面的证明中策略 $\varphi\in\Pi_{RS}$.

下列定理给出最优方程的解的性质.

定理 3.3 (a) 对任意 $(i,\lambda)\in B^c\times U$, $D^*(i,\lambda)$ 满足最优方程:

$$D^*(i,\lambda)=TD^*(i,\lambda). \tag{3.6}$$

(b) 存在一个决策函数 $f\in F$ 使得

$$D^*(i,\lambda)=T^fD^*(i,\lambda). \tag{3.7}$$

证 (a) 由引理 3.2, 对任意 $\pi=\{\varphi_0,\varphi_1,\cdots\}\in\Pi_{RM}$, 有

$$D^{\pi}(i,\lambda)=T^{\varphi_0}D^{(1)\pi}(i,\lambda)\geq T^{\varphi_0}D^*(i,\lambda)\geq TD^*(i,\lambda). \tag{3.8}$$

因为 π 是任意的, 可得 $D^*(i,\lambda)\geq TD^*(i,\lambda)$. 现证反向不等式. 事实上, 由最优值函数的定义, 对任意 $\varepsilon>0$, $(i,\lambda)\in B^c\times U$, 存在 ${}^{(i,\lambda)}\pi\in\Pi_{RM}$ 使得 $D^{(i,\lambda)\pi}(i,\lambda-1)\leq D^*(i,\lambda-1)+\varepsilon$. 构造策略 $\pi\in\Pi_{RM}$ 如下:

$$\pi=\begin{cases}{}^{(i,\lambda)}\pi, & \text{如果 } x_0=i\in B^c,\\ \text{任意}, & \text{如果 } x_0=i\in B.\end{cases}$$

可得 $D^{\pi}(i,\lambda-1)\leq D^*(i,\lambda-1)+\varepsilon$, 其中 $i\in B^c$. 对每个 $(i,\lambda)\in B^c\times U$, 存在 $a_{(i,\lambda)}\in A(i)$ 使得

$$
\begin{aligned}
&\sum_{j\in B}p(j|i,a_{(i,\lambda)})+\sum_{j\in B^c}p(j|i,a_{(i,\lambda)})D^*(j,\lambda-1)\\
&=\min_{a\in A(i)}\Big[\sum_{j\in B}p(j|i,a)+\sum_{j\in B^c}p(j|i,a)D^*(j,\lambda-1)\Big].
\end{aligned}
$$

定义 $f(i,\lambda)=a_{(i,\lambda)}$, 显然, $f\in F$. 令 $\widehat{\pi}=(f,\pi)$, 由引理 3.2 可得

$$\begin{aligned}D^{\widehat{\pi}}(i,\lambda)&=\sum_{j\in B}p(j|i,f(i))+\sum_{j\in B^c}p(j|i,f(i))D^{\pi}(j,\lambda-1)\\&\le\sum_{j\in B}p(j|i,f(i))+\sum_{j\in B^c}p(j|i,f(i))(D^*(j,\lambda-1)+\varepsilon)\\&\le\sum_{j\in B}p(j|i,f(i))+\sum_{j\in B^c}p(j|i,f(i))D^*(j,\lambda-1)+\varepsilon\\&=\min_{a\in A(i)}\Big[\sum_{j\in B}p(j|i,f(i))+\sum_{j\in B^c}p(j|i,a)D^*(j,\lambda-1)\Big]+\varepsilon,\end{aligned}$$

即是

$$D^*(i,\lambda)\le\min_{a\in A(i)}\Big[\sum_{j\in B}p(j|i,a)+\sum_{j\in B^c}p(j|i,a)D^*(j,\lambda-1)\Big]+\varepsilon,$$

令 $\varepsilon\to 0$, 可得

$$D^*(i,\lambda)\le\min_{a\in A(i)}\Big[\sum_{j\in B}p(j|i,a)+\sum_{j\in B^c}p(j|i,a)D^*(j,\lambda-1)\Big],\tag{3.9}$$

与 (3.8) 一起，可得 $D^*(i,\lambda)=TD^*(i,\lambda)$.

(b) 由 (a) 和引理 3.1, 存在决策函数 $f\in F$ 使得 $D^*(i,\lambda)=TD^*(i,\lambda)=T^fD^*(i,\lambda)$.

4 最优策略的性质及其存在性

本部分讨论最优策略的性质及其存在性. 我们表明在适当的条件下存在最优平稳策略，且达到最优方程最小值的平稳策略都是最优的. 另外，我们也证明不依赖于阈值的最优策略的存在性.

为刻画最优策略，对每个 $(i,\lambda)\in B^c\times U$ ，定义最优行动集如下

$$A^*(i,\lambda):=\big\{a\in A(i)|D^*(i,\lambda)=T^aD^*(i,\lambda)\big\},\qquad A^*(i):=\bigcap_{\lambda\in U}A^*(i,\lambda).\tag{4.1}$$

注记 4.1 由 $A(i)$ 的有限性和定理 3.3(a), $A^*(i,\lambda)\ne\emptyset$, 但 $A^*(i)$ 可能是空集.

现给出最优策略的一些性质.

定理 4.1 设 $\pi=(\varphi_0,\varphi_1,\cdots)\in\Pi_{RM}$ 是最优策略.

(a) 对每个 $(i,\lambda)\in B^c\times U$, 有 $A_{\varphi_0}(i,\lambda)\subset A^*(i,\lambda)$, $\varphi_0(A^*(i,\lambda)|i,\lambda)=1$, 其中 $A_{\varphi_0}(i,\lambda)=\{a\in A(i)|\varphi_0(a|i,\lambda)>0\}$.

(b) 对每个 $(i,\lambda)\in B^c\times U$, $f\in F$ 满足 $f(i,\lambda)\in A_{\varphi_0}(i,\lambda)$, 则 $f^{(1)}\pi:=\{f,\pi_0,\pi_1,\cdots\}$ 是最优的.

(c) 若 $\varphi\in\Phi$ 满足 $D^*(i,\lambda)=T^{\varphi}D^*(i,\lambda)$ 的随机核，则 $\{\varphi,\pi\}$ 是最优的.

证 (a) 因为 $\pi=(\varphi_0,\varphi_1,\cdots)\in\Pi_{RM}$ 是最优策略，由引理 3.2 和定理 3.3, 可得

$$D^*(i,\lambda)=D^{\pi}(i,\lambda)=T^{\varphi_0}D^{^{(1)}\pi}(i,\lambda)\ge T^{\varphi_0}D^*(i,\lambda)\ge TD^*(i,\lambda)=D^*(i,\lambda).\tag{4.2}$$

因此 $T^{\varphi_0}D^*(i,\lambda)=D^*(i,\lambda)$, 这意味着

$$\sum_{a\in A(i)}\varphi_0(a|i,\lambda)\big[T^aD^*(i,\lambda)-D^*(i,\lambda)\big]=0, \tag{4.3}$$

此式联合 $T^aD^*(i,\lambda)\geq TD^*(i,\lambda)=D^*(i,\lambda)$ 得到 (a) 中的结果.

(b) 这部分的证明与 (a) 类似. 对每个 $(i,\lambda)\in B^c\times U$, 有

$$D^*(i,\lambda)=D^\pi(i,\lambda)=T^{\varphi_0}D^{^{(1)}\pi}(i,\lambda)\geq TD^{^{(1)}\pi}(i,\lambda)\geq TD^*(i,\lambda)=D^*(i,\lambda). \tag{4.4}$$

因此 $D^\pi(i,\lambda)=TD^{^{(1)}\pi}(i,\lambda)$, 由此可得

$$D^\pi(i,\lambda)\leq T^aD^{^{(1)}\pi}(i,\lambda). \tag{4.5}$$

为了得到结果, 往证 $D^{f^{(1)}\pi}(i,\lambda)=D^\pi(i,\lambda)$. 如果这不成立, 那么存在某些 $(i,\lambda)\in B^c\times U$ 使得 $D^{f^{(1)}\pi}(i,\lambda)>D^\pi(i,\lambda)$. 因为 $f(i,\lambda)\in A_{\varphi_0}(i,\lambda)$, 由引理 3.2 和 (4.5), 得

$$\begin{aligned}
D^\pi(i,\lambda)=&\sum_{a\in A(i)}\varphi_0(a|i,\lambda)T^aD^{^{(1)}\pi}(i,\lambda)\\
=&\varphi_0(f(i,\lambda)|i,\lambda)T^{f(i,\lambda)}D^{^{(1)}\pi}(i,\lambda)+\sum_{a\in A(i)-f(i,\lambda)}\varphi_0(a|i,\lambda)T^aD^{^{(1)}\pi}(i,\lambda)\\
=&\varphi_0(f(i,\lambda)|i,\lambda)D^{f^{(1)}\pi}(i,\lambda)+\sum_{a\in A(i)-f(i,\lambda)}\varphi_0(a|i,\lambda)T^aD^{^{(1)}\pi}(i,\lambda)\\
>&\varphi_0(f(i,\lambda)|i,\lambda)D^\pi(i,\lambda)+\sum_{a\in A(i)-f(i,\lambda)}\varphi_0(a|i,\lambda)D^\pi(i,\lambda)\\
=&\sum_{a\in A(i)}\varphi_0(a|i,\lambda)D^\pi(i,\lambda)=D^\pi(i,\lambda),
\end{aligned}$$

这是矛盾的, 所以 $D^{f^{(1)}\pi}(i,\lambda)=D^\pi(i,\lambda)$ 对所有的 $(i,\lambda)\in B^c\times U$ 成立, 即 $f^{(1)}\pi$ 是最优策略.

(c) 由引理 3.2 和 π 的最优性, 有 $D^{\{\varphi,\pi\}}(i,\lambda)=T^\varphi D^\pi(i,\lambda)=T^\varphi D^*(i,\lambda)=D^*(i,\lambda)$, 从而 $\{\varphi,\pi\}$ 是最优的.

为保证最优策略的存在, 需要下面的假设.

假设 A 对每个 $(i,\lambda)\in B^c\times U$ 和 $f\in\Pi_{DS}$, $P^f_{(i,\lambda)}(\tau_B<\infty)=1$.

注记 4.2 (1) 因为受控系统的工作时间有限, 所以假设 A 在很多实际情况下成立.

(2) 对每个 $(i,\lambda)\in B^c\times U$ 和 $f\in\Pi_{DS}$, 有

$$P^f_{(i,\lambda)}(\tau_B<\infty)=P^f_{(i,\lambda)}\Big(\bigcup_{k=1}^{\infty}\{x_k\in B\}\Big). \tag{4.6}$$

因此假设 A 和下列形式也等价

$$P^f_{(i,\lambda)}\Big(\bigcup_{k=1}^{\infty}\{x_k\in B\}\Big)=1,\quad \text{或者}\ \ P^f_{(i,\lambda)}\Big(\bigcap_{k=1}^{\infty}\{x_k\in B^c\}\Big)=0,\quad \forall\,(i,\lambda)\in B^c\times U. \tag{4.7}$$

命题 4.2　若存在实数 $\alpha > 0$ 满足

$$p(B|i,a) := \sum_{j\in B} p(j|i,a) \geq \alpha, \qquad \forall\, i \in B^c,\ \ a \in A(i), \tag{4.8}$$

则假设 A 成立.

证　用归纳法证明下式

$$P^f_{(i,\lambda)}\Big(\bigcap_{k=1}^{n}\{x_k \in B^c\}\Big) \leq (1-\alpha)^n. \tag{4.9}$$

当 $n = 1$, 由命题 4.2 的条件和性质 (2.2)–(2.4), 有

$$\begin{aligned} &P^f_{(i,\lambda)}(x_1 \in B^c) = E^f_{(i,\lambda)}\big[P^f_{(i,\lambda)}(x_1 \in B^c|x_0,\Delta_0,\lambda_0)\big]\\ =&P^f_{(i,\lambda)}(x_1 \in B^c|x_0 = i, \Delta_0 = f(i,\lambda), \lambda_0 = \lambda)\\ =&P(B^c|i, f(i,\lambda)) = 1 - P(B|i, f(i,\lambda)) \leq 1-\alpha. \end{aligned}$$

现假设 (4.9) 式对某个 $n \geq 1$ 成立. 由 (4.7) 式和递推假设, 可得

$$\begin{aligned} &P^f_{(i,\lambda)}\Big(\bigcap_{k=1}^{n+1}\{x_k \in B^c\}\Big) = E^f_{(i,\lambda)}\Big[P^f_{(i,\lambda)}\Big(\bigcap_{k=1}^{n+1}\{x_k \in B^c\}\Big| x_0,\Delta_0,\lambda_0,\cdots,x_n,\lambda_n\Big)\Big]\\ =&E^f_{(i,\lambda)}\big[1_{\{x_1\in B^c\}}\cdots 1_{\{x_n\in B^c\}}P^f_{(i,\lambda)}(x_{n+1} \in B^c|x_0,\Delta_0,\lambda_0,\cdots,x_n,\lambda_n)\big]\\ =&E^f_{(i,\lambda)}\big[1_{\{x_1\in B^c\}}\cdots 1_{\{x_n\in B^c\}}P^f_{(x_n,\lambda_n)}(x_1 \in B^c)\big]\\ \leq&(1-\alpha)E^f_{(i,\lambda)}\big[1_{\{x_1\in B^c\}}\cdots 1_{\{x_n\in B^c\}}\big]\\ =&(1-\alpha)P^f_{(i,\lambda)}\Big(\bigcap_{k=1}^{n}\{x_k \in B^c\}\Big) \leq (1-\alpha)^{n+1}. \end{aligned}$$

也即 (4.9) 式对 $n+1$ 也成立. 因此 (4.9) 式对所有的 $n \geq 1$ 成立. 令 (4.9) 式中 $n \to \infty$, 可得

$$P^f_{(i,\lambda)}\Big(\bigcap_{k=1}^{\infty}\{x_k \in B^c\}\Big) = \lim_{n\to\infty} P^f_{(i,\lambda)}\Big(\bigcap_{k=1}^{n}\{x_k \in B^c\}\Big) \leq \lim_{n\to\infty}(1-\alpha)^n = 0,$$

此式联合 (4.7), 可知假设 A 成立.

下面给出对最优策略存在性非常关键的一个引理. 首先引入一些记号.

对每个 $(i,\lambda) \in B^c \times U$ 和 $f \in F$,

$$\widehat{T}^f D(i,\lambda) := \sum_{j\in B^c} P(j|i,f(i,\lambda))D(j,\lambda-1). \tag{4.10}$$

引理 4.3　若假设 A 成立, 令 $f \in \Pi_{DS}$,

(a)　若 $(D-G)(i,\lambda) \leq \widehat{T}^f(D-G)(i,\lambda)$, 则 $D(i,\lambda) \leq G(i,\lambda)$;

(b)　D^f 是方程 $D(i,\lambda) = T^f D(i,\lambda)$ 的唯一解.

证 为了证明 (a), 首先用归纳法证明 $(\widehat{T}^f)^n(D-G)(i,\lambda)\le P^f_{(i,\lambda)}\big(\bigcap\limits_{k=1}^{n}\{x_k\in B^c\}\big)$. 当 $n=1$ 时容易验证上式成立. 事实上, 令 $(i,\lambda)\in B^c\times U$, 由于 $D(i,\lambda)-G(i,\lambda)\le 1$, 可得

$$\begin{aligned}\widehat{T}^f(D-G)(i,\lambda)&=\sum_{j\in B^c}P(j|i,f(i,\lambda))(D-G)(j,\lambda-1)\\&\le\sum_{j\in B^c}P(j|i,f(i,\lambda))=P(B^c|i,f(i,\lambda))=P^f_{(i,\lambda)}(x_1\in B^c).\end{aligned}$$

假设对某一个 $n\ge 1$ 时, 有 $(\widehat{T}^f)^n(D-G)(i,\lambda)\le P^f_{(i,\lambda)}\big(\bigcap\limits_{k=1}^{n}\{x_k\in B^c\}\big)$, 则由递推假设, 有

$$\begin{aligned}(\widehat{T}^f)^{n+1}(D-G)(i,\lambda)=&\widehat{T}^f(\widehat{T}^f)^n(D-G)(i,\lambda)\\=&\sum_{j\in B^c}P(j|i,f(i,\lambda))(\widehat{T}^f)^n(D-G)(j,\lambda-1)\\\le&\sum_{j\in B^c}P(j|i,f(i,\lambda))P^f_{(j,\lambda-1)}\Big(\bigcap_{k=1}^{n}\{x_k\in B^c\}\Big)\\=&P^f_{(i,\lambda)}\Big(\bigcap_{k=1}^{n+1}\{x_k\in B^c\}\Big).\end{aligned}$$

最后一个等式由 (2.2)–(2.4) 得到. 因此 $(\widehat{T}^f)^n(D-G)(i,\lambda)\le P^f_{(i,\lambda)}(\bigcap\limits_{k=1}^{n}\{x_k\in B^c\})$ 对 $n+1$ 也成立. 此外, 对每个 $(i,\lambda)\in B^c\times U$, 直接计算可以得到

$$(D-G)(i,\lambda)\le(\widehat{T}^f)^n(D-G)(i,\lambda)\le P^f_{(i,\lambda-1)}\Big(\bigcap_{k=1}^{n}\{x_k\in B^c\}\Big).\tag{4.11}$$

注意到假设 A 意味着 $P^f_{(i,\lambda)}\big(\bigcap\limits_{k=1}^{\infty}\{x_k\in B^c\}\big)=0$. 因此, 令 (4.11) 式中 $n\to\infty$, 对每个 $(i,\lambda)\in B^c\times U$, 可得 $(D-G)(i,\lambda)\le 0$. 因此 (a) 得证.

(b) 假设 $D(i,\lambda)$ 是方程 $D(i,\lambda)=T^fD(i,\lambda)$ 的解. 由引理 3.2, 可知 $D^f(i,\lambda)=T^fD^f(i,\lambda)$, 因此 $D(i,\lambda)-D^f(i,\lambda)=\widehat{T}^f(D-D^f)(i,\lambda)$, 此式联合 (a) 可得 $D(i,\lambda)=D^f(i,\lambda)$.

下面给出本文的另一个重要结果.

定理 4.4 若假设 A 成立, 则

(a) 任意满足 $D^*(i,\lambda)=T^fD^*(i,\lambda)$ 的策略是最优的.

(b) 存在一个最优平稳策略.

(c) 存在策略 $\pi\in\Pi_0$ 使得 $D^\pi(i,\lambda)=D^*(i,\lambda)$ 当且仅当对所有的 $i\in B^c$, $A^*(i)\ne\emptyset$.

证 (a) 假设存在 $f\in\Pi_{DS}$ 使得 $D^*(i,\lambda)=T^fD^*(i,\lambda)$, 由假设 A 和引理 4.3 (b), 可得 $D^f(i,\lambda)=D^*(i,\lambda)$, 因此 f 是最优的.

(b) 当 $A(i)$ 有限时, 对所有的 $(i,\lambda)\in B^c\times U$, 存在 f 满足 $f(i,\lambda)\in A^*(i,\lambda)$. 因此 $D^*(i,\lambda)=T^fD^*(i,\lambda)$. 由 (a) 可知 f 是最优策略.

(c) 设 $\pi=\{\varphi_0,\varphi_1,\cdots\}\in\Pi_0$ 是最优的, 即 $D^\pi(i,\lambda)=D^*(i,\lambda)$. 对每个 $(i,\lambda)\in B^c\times U$, 由定理 4.1 (a) 知, 对所有的 $(i,\lambda)\in B^c\times U$, $A_{\varphi_0}(i)\equiv A_{\varphi_0}(i,\lambda)\subset A^*(i,\lambda)$, 从而有 $A_{\varphi_0}(i)\subset A^*(i,\lambda)=A^*(i)$. 注意到 $A_{\varphi_0}(i)\neq\emptyset$, 故对所有的 $i\in B^cA^*(i)\neq\emptyset$. 现证明反向不等式. 设 $A^*(i)\neq\emptyset$, $f:B^c\times N\to A$ 使得对每个 $(i,\lambda)\in B^c\times U$, $f(i,\lambda)\equiv f(i)\in A^*(i)$. 于是 $f\in\Pi_0$, 且 $D^*(i,\lambda)=T^fD^*(i,\lambda)$. 由 (a) 可知 f 是最优的.

5　算法

本部分我们给出算法如下，用于计算最优策略.

Step I: 对每个 $i\in B^c$ 和 $\lambda=1$, 设 $D^*(i,1)=0$.

Step II: 对任意 $i\in B^c$ 和 $\lambda\geq 2$, 解方程

$$D^*(i,\lambda)=\min_{a\in A(i)}\Big\{\sum_{j\in B}P(j|i,a)+\sum_{j\in B^c}P(j|i,a)D^*(j,\lambda-1)\Big\},\tag{5.1}$$

可得 $D^*(i,\lambda)$.

Step III: 返回 **Step II**, 用 $\lambda+1$ 代替 λ.

6　例子

本部分给出例子说明我们结果的应用. 这里的例子和 [5] 中的例 2.2 相同. 特别地, [5] 中例 2.2 的最优策略不存在，但是在我们的条件下最优策略存在. 设状态空间 $S=\{0,1,2\}$, 目标集 $B=\{0\}$, 允许行动集 $A(0)=A(1)=\{1\}$, $A(2)=\{1,2\}$, 转移概率为 $p(0|0,1)=1$, $p(0|1,1)=0.25$, $p(1|1,1)=0.5$, $p(2|1,1)=0.25$, $p(0|2,1)=0.15$, $p(1|2,1)=0.15$, $p(2|2,1)=0.7$, $p(0|2,2)=0.1$, $p(1|2,2)=0.6$, $p(2|2,2)=0.3$. 容易验证假设 A 成立. 数值结果见图 1. 由于状态 1 只有一个行动，所以此处不讨论它. 从图 1 和计算过程，我们有以下结论：

(1) 图 1b 中, $T^1D^*(2,20)=T^2D^*(2,20)=1$. 显然，函数 $T^aD^*(2,\lambda)$ 关于 λ 是非降的，且对 $\lambda>20$, 有 $T^1D^*(2,\lambda)=T^2D^*(2,\lambda)=1$.

(2) 图 1b 中，当 $\lambda\in\{3,\cdots,20\}$, $T^1D^*(2,\lambda)$ 在 $T^2D^*(2,\lambda)$ 下方，但是 $T^1D^*(2,2)$ 在 $T^2D^*(2,2)$ 上方. 这意味着当 $\lambda\in\{3,\cdots,20\}$, 行动 1 比行动 2 有更低的风险概率，即最优行动依赖阈值 λ.

(3) 解最优方程 $F^*(i,\lambda)=\min\limits_{a\in A(i)}\{T^aD^*(i,\lambda)\}$, 得到

$$D^*(1,\lambda)=T^1D^*(1,\lambda),\qquad \lambda\in\{1,\cdots,20\},\tag{6.1}$$

$$D^*(2,\lambda)=\begin{cases}T^1D^*(2,\lambda)=T^2D^*(2,\lambda), & \lambda=1,\\ T^2D^*(2,\lambda), & \lambda=2,\\ T^1D^*(2,\lambda), & \lambda\in\{3,\cdots,20\}.\end{cases}\tag{6.2}$$

(4) 由方程 (6.1)–(6.2), 定义策略 f^* 为

$$f^*(1,\lambda)=1,\qquad \lambda\in\{1,\cdots,20\},$$
$$f^*(2,\lambda)=\begin{cases}2, & \lambda\in\{1,2\},\\ 1, & \lambda\in\{3,\cdots,20\},\end{cases}$$

由 (6.1)–(6.2), 对于 $i=1,2,\ \lambda\in\{1,\cdots,20\}$, 得到 $D^*(i,\lambda)=T^{f^*}D^*(i,\lambda)$, 由定理 4.4 可知 f^* 是最优平稳策略.

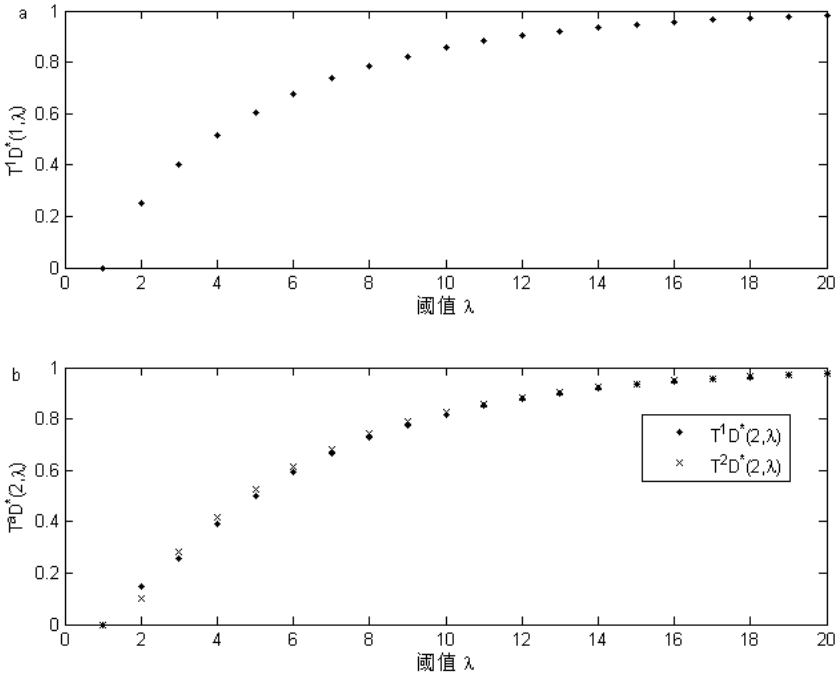

图 1 函数 $T^aD^*(i,\lambda)$.

另外, 由最优行动集 $A^*(i,\lambda)$ 的定义, 有

$$A^*(1,\lambda)=\{1\},\qquad \lambda\in\{1,\cdots,20\},$$
$$A^*(2,\lambda)=\begin{cases}\{1,2\}, & \lambda=1,\\ \{2\}, & \lambda=2,\\ \{1\}, & \lambda\in\{3,\cdots,20\},\end{cases}$$

$A^*(2)=\bigcap\limits_{\lambda\in Z^+}A^*(2,\lambda)=\emptyset$, 即在 Π_0 中对于状态 2 最优策略不存在.

注 6.1 注意到 [5] 中直到 n 的最优策略不存在. 然而, 本文中最优平稳策略存在, 主要是由于这里的策略依赖阈值 λ.

参 考 文 献

[1] Derman C. Finite State Markov Decision Processes. New York: Academic Press, 1970

[2] Huang Y H, Guo X P. First Passage Models for Denumerable Semi-Markov Decision Processes with Nonnegative Discounted Costs. *Acta. Math. Appl. Sin.*, 2011, 27(2): 177-190

[3] Liu, J Y, Liu K. Markov Decision Programming–the Moment Optimal Problem for the Passage Model. *J. Austral. Math. Soc. Ser. B*, 1997, 38: 542–562

[4] Liu J Y, Liu K. Markov Decision Programming—the First Passage Model with Denumerable State Space. *Sys. Sci. Math. Scis.*, 1992, 4: 340–351

[5] Liu J Y, Huang S M. Markov Decision Processes with Distribution Function Criterion of First-passage Time. *Appl. Math. Optim.*, 2001, 43: 187–201

[6] Lin Y L, Tomkins R J, Wang C L. Optimal Models for the First Arrival Time Distribution Function in Continuous Time-with a Special Case. *Acta. Math. Appl. Sin.*, 1994, 10: 194–212

[7] Ohtsubo Y. Optimal Threshold Probability in Undiscounted Markov Decision Processes with a Target Set. *Appl. Math. Comput.*, 2004, 149: 519–532

[8] Puterman M L. Markov Decision Processes. New York: Wiley, 1994

[9] Hernández-Lerma O, Lasserre J B. Discrete-time Markov Control Processes. New York: Springer-Verlag, 1996

Discrete-time Markov Decision Processes with First Passage Models

LIU Qiuli

(*School of Mathematical Sciences, South China Normal University, Guangzhou* 510631)

(*E-mail: liuql2007@yahoo.cn*)

Abstract This paper deals with risk probability for first passage models in discrete-time Markov decision processes with a denumerable state space. The criterion to be minimized is the risk probability (risk function) that a first passage time to a given target set is less than a threshold value. We first establish the optimality equation and show that solutions of the equation correspond to optimal value functions. Then, we discuss some properties of optimal policies and further give suitable conditions under which there exists an optimal stationary policy. Finally, in order to illustrate applications of our results, an example is also displayed.

Key words target set; first passage time; risk probability

MR(2000) Subject Classification 90C40; 93E20

Chinese Library Classification 0211.62; 0231.3; 0232