最优准则

有限阶段的策略迭代和...

最优策略的存在性和算法

两个例子

最优策略的结构

单调策略的最优性

中国科学院研究生院

运筹通论II

刘克

中科院数学与系统科学研究院 北京100190

邮箱地址: kliu@amss.ac.cn

最优准则

有限阶段的策略迭代和...

最优策略的存在性和算法

两个例子

最优策略的结构

单调策略的最优性

第三部分　马氏决策—有限阶段模型

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

- 1 最优准则
- 2 有限阶段的策略迭代和最优方程
- 3 最优策略的存在性和算法
- 4 两个例子
- 5 最优策略的结构
- 6 单调策略的最优性

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

1 最优准则

在这一章，我们讨论离散时间决策时刻、有限阶段的马氏决策问题。用第1章的记号，马氏决策过程的五重组为：

$$\{T, S, A(i), p(\cdot|i,a), r(i,a)\},$$

其中$T=\{0,1,\cdots,N-1\}, 0<N<\infty$，$r(i,a)$是有界报酬函数。在选定一个策略并实施以后，决策者在阶段$0,1,\cdots,N$时依一定的概率收到一串报酬，将其累加起来就是该模型的具体效用函数，简称为有限阶段模型或N阶段模型。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

对$N \geq 0$，策略$\pi \in \Pi$下的N阶段期望总报酬效用函数定义为：

$$V_N(i,\pi) \equiv \sum_{t=0}^{N-1} E_\pi^i[r(Y_t, \Delta_t] + E_\pi^i[r(Y_N)], \quad i \in S, \tag{1}$$

表示使用策略π，在0时刻从状态$i \in S$出发的条件下，系统直到N时刻所获得的期望总报酬，这里$r(Y_N)$ 是过程的终止报酬，很多情况下$r(Y_N)$为0。用$V_N(\pi)$表示第i个分量为$V_N(i,\pi)$的列向量，当状态空间S可列时，$V_N(\pi)$为可列维向量。当S和A是一般的Borel集时，我们仍然可以这样定义N时刻所获得的期望总报酬，但是要使(1) 有意义，需要报酬函数$r(i,a)$和策略π的可测性条件，这里就不详细说明了。由$r(i,a)$的有界性知道$V_N(\pi)$是有界的，从而可以做出下面的定义。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

定义2.1：令

$$V_N^*(i) \equiv \sup_{\pi \in \Pi} V_N(i,\pi) \tag{2}$$

为最优值函数，用向量表示为V_N^*。对$\epsilon \geq 0$，如果策略π^*使得$V_N(i,\pi^*) \geq V_N^*(i) - \epsilon$对所有状态$i \in S$成立，则称$\pi^*$为$N$阶段$\epsilon$**最优策略**，简称为$\epsilon$最优策略，当$\epsilon = 0$时简称为最优策略。

我们总认为决策者希望在系统的每个可能的初始状态上都能选取最好的决定。正如定义2.1中所描述的那样，寻求向量最优。在实际的问题中，可能只需要知道一些或一个特定的初始状态是如何选优的就足够了。也就是说只需要极大化$\sum_{i \in S} V_N(i,\pi)\mu\{Y_0 = i\}$，即对$\mu\{Y_0 = i\} > 0$的那些$i$极大化相应的$V_N(i,\pi)$，其中$\mu\{Y_0 = i\}$为初始状态的概率分布。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

2 有限阶段的策略迭代和最优方程

有限阶段模型的理论和计算方法是基于动态规划的期望报酬值向后递归的过程。下面我们介绍动态规划的这一有效的方法。

令$\pi = (\pi_0, \pi_1, \cdots, \pi_{N-1})$为一个一般的策略(由于问题只考虑到阶段$N$，$N$阶段以后的决策规则不再影响策略的值函数，所以为了记号方便，我们依然记为$\pi \in \Pi$)。另外，记函数$u_t^\pi := H_t \to \Re$为用策略π，从时刻t到时刻N的期望报酬总和。如果决策时刻t的历史为$h_t = (i_0, a_0, \cdots, i_t) \in H_t$，则对$t < N$，定义

$$u_t^\pi(h_t) \equiv E_\pi \left\{ \sum_{n=t}^{N-1} R_n(\pi) + R_N(Y_N) \middle| h_t, Y_t = i_t \right\}, \tag{3}$$

很明显，我们希望当$h_0 = i$时$u_0^\pi(i) = V_N(i, \pi)$。这里$V_N(i, \pi)$和$u_t^\pi(i)$的区别在于$V_N(i, \pi)$是表示从决策开始时刻0到决策终止时刻的报酬总和，而$u_t^\pi(i)$则表示从决策时刻t以后到决策时刻终止的报酬和。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

下面我们介绍如何一步一步的计算$u_t^{\pi}(i)$的。这个算法被称为**有限阶段值迭代算法**。为了记号简单，我们只考虑状态空间离散且策略类是Π^d的情形。在实际问题中，随机策略是很不实用的，一般只有理论意义，而实用的是决定性策略、马氏策略和平稳策略等等。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

算法2.1 (对固定的$\pi \in \Pi^d$，有限阶段值迭代)

步骤1: 令$t = N$且对一切$h_N = (h_{N-1}, a_{N-1}, i_N) \in H_N$，$u_N^\pi(h_N) = r_N(i_N)$。

步骤2: 如果$t = 0$，停止. 否则, 令$t - 1 \Rightarrow t$后,进入步骤3。

步骤3: 对每个状态$i_t \in S$和每个历史$h_t = (h_{t-1}, a_{t-1}, i_t) \in H_t$计算$u_t^\pi(h_t)$

$$u_t^\pi(h_t) = r_t(i_t, a_t(h_t)) + \sum_{j \in S} p_t(j|i_t, a_t(h_t)) u_{t+1}^\pi(h_t, a_t(h_t), j), \qquad (4)$$

这里$(h_t, a_t(h_t), j) \in H_{t+1}$，其中用记号$a_t(h_t)$表示这个行动的选取依赖于历史$h_t$。

步骤4: 返回到步骤2.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

公式(4)表示策略π在决策时刻$t, t+1, \cdots, N$时的期望报酬等于决策者采用行动$a_t(h_t)$ 后的一步即得报酬加上直到周期结束的期望报酬。当$t=0$时，有下面的定理。

定理2.1：令$\pi \in \Pi^d$并设$u_t^\pi, t \leq N$，是由算法2.1得到的，则对一切$t \leq N$，u_t^π满足(3).特别的,对一切$i \in S$,我们有$V_N(i,\pi) = u_0^\pi(i)$.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

证明: $t = N$时的结论是明显的。我们假设当$t+1,\ t+2, \cdots, N$结论都成立, 由归纳假设知道:

$$
\begin{aligned}
u_t^{\pi}(h_t) &= r_t(i_t, a_t(h_t)) + \sum_{j \in S} p_t(j|i_t, a_t(h_t)) u_{t+1}^{\pi}(h_t, a_t(h_t), j) \\
&= r_t(i_t, a_t(h_t)) \\
&\quad + \sum_{j \in S} p_t(j|i_t, a_t(h_t)) E_{\pi} \left\{ \sum_{n=t+1}^{N-1} R_n(\pi) + R_N(Y_N) \middle| h_t, a_t(h_t), j \right\} \\
&= r_t(i_t, a_t(h_t)) \\
&\quad + E_{\pi} \left[E_{\pi} \left\{ \sum_{n=t+1}^{N-1} R_n(\pi) + R_N(Y_N) \middle| h_t, a_t(h_t), j \right\} \middle| h_t, i_t \right] \\
&= r_t(i_t, a_t(h_t)) + E_{\pi} \left[\sum_{n=t+1}^{N-1} R_n(\pi) + R_N(Y_N) \middle| h_t, i_t \right] \qquad (5) \\
&= E_{\pi} \left\{ \sum_{n=t}^{N-1} R_n(\pi) + R_N(Y_N) \middle| h_t, Y_t = i_t \right\},
\end{aligned}
$$

由于在决策时刻t时,当前的状态i_t和已经发生的历史h_t都已经知道了, 所以最后一个等式成立. □

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

公式(4)对于一般的策略类Π，我们有类似的定理。

定理2.2：令$\pi \in \Pi$并设$u_t^\pi, t \leq N$，是由递归公式

$$u_t^\pi(h_t) = \sum_{a \in A(i_t)} \pi_t(a|h_t) \left\{ r_t(i_t, a) + \sum_{j \in S} p_t(j|i, a) u_{t+1}^\pi(h_t, a, j) \right\} \qquad (6)$$

得到的，则对一切$t \leq N$，u_t^π满足(3)，而且对一切$i \in S$, $V_N(i, \pi) = u_0^\pi(i)$。

证明:类似于定理2.1用归纳法直接证明。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

令

$$u_t^*(h_t) = \sup_{\pi \in \Pi} u_t^{\pi}(h_t).$$

定义最优方程为：

$$u_t(h_t) = \sup_{a \in A(i_t)} \left\{ r_t(i_t, a) + \sum_{j \in S} p_t(j|i_t, a) u_{t+1}(h_t, a, j) \right\} \tag{7}$$

对一切$t = 0, 1, \cdots, N-1$和一切历史$h_t = (h_{t-1}, a_{t-1}, i_t) \in H_t$。对$t = N$，加上边界条件

$$u_N(h_N) = r_N(i_N) \tag{8}$$

其中$h_N = (h_{N-1}, a_{N-1}, i_N) \in H_N$。当公式(7)中的sup可以达到时，我们使用符号max。例如所有的行动集合$A(i)$都是有限的情形。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

最优方程在马氏决策过程的研究中十分重要，主要是由于

1) 对每个决策时刻t, 最优方程的解就是从时刻t到结束阶段的最优报酬值.

2) 提供了确定策略是否是最优的方法. 也就是说,如果一切决策时刻t, 该策略从t时刻到决策结束的期望总报酬满足$t = 0, 1, \cdots, N$的方程组,那么它就是最优的.

3) 它是计算最优报酬和最优策略的最基本的东西.

4) 它可以被用于确定最优报酬和最优策略的结构性质.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

下面的定理总结了最优方程解的最优性质。

定理2.3：假设$u_t, t \leq N$，是方程(7)的解,而且满足边界条件(8), 那么:
a) $u_t(h_t) = u_t^*(h_t)$对一切$h_t \in H_t, t = 0, 1, \cdots, N$,而且
b) $u_0(i) = V_N^*(i)$, 对一切$i \in S$。

证明：证明分为两部分,首先我们证明$u_t(h_t) \geq u_t^*(h_t)$,对一切$h_t \in H_t$以及$t = 0, 1, \cdots, N$。考察最后阶段N的情形,此时由于不再选择任何决策行动,所以,对任意的$\pi \in \Pi$和一切历史$h_N \in H_N$, 我们都有: $u_N(h_N) = r_N(i_N) = u_N^\pi$. 故$u_N(h_N) = u_N^*(h_N)$.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

我们假设:对所有的$h_t \in H_t$和$t = n+1, \cdots, N$都有$u_t(h_t) \geq u_t^*(h_t)$. 当$t = n$时,由最优方程(7)和归纳假设

$$
\begin{aligned}
u_n(h_n) &= \sup_{a \in A(i_n)} \left\{ r_n(i_n, a) + \sum_{j \in S} p_n(j|i_n, a) u_{n+1}(h_n, a, j) \right\} \\
&\geq \sup_{a \in A(i_n)} \left\{ r_n(i_n, a) + \sum_{j \in S} p_n(j|i_n, a) u_{n+1}^*(h_n, a, j) \right\} \\
&\geq \sup_{a \in A(i_n)} \left\{ r_n(i_n, a) + \sum_{j \in S} p_n(j|i_n, a) u_{n+1}^{\pi'}(h_n, a, j) \right\} \\
&\geq \sum_{a \in A(i_n)} \pi'(a|h_n) \left\{ r_n(i_n, a) + \sum_{j \in S} p_n(j|i_n, a) u_{n+1}^{\pi'}(h_n, a, j) \right\} \\
&= u_n^{\pi'}(h_n), \qquad (9)
\end{aligned}
$$

其中$\pi' \in \Pi$是任意的一个策略. 由π'的任意性知道$u_n(h_n) \geq u_n^*(h_n)$.再由归纳知道结论成立.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

下面我们证明另外一面. 对任意的$\epsilon > 0$,存在策略$\pi' \in \Pi$使得

$$u_t^{\pi'}(h_t) + (N-t)\epsilon \geq u_t(h_t) \tag{10}$$

对一切$h_t \in H_t$以及$t = 0, 1, \cdots, N$成立。事实上,$\pi' = (f_0, f_1, \cdots, f_{N-1}) \in \Pi^d \subset \Pi$是这样构造的:

$$r_n(i_n, f_n(h_n)) + \sum_{j \in S} p_n(j|i_n, f_n(h_n)) u_{n+1}(i_n, f_n(i_n), j) + \epsilon \geq u_n(h_n). \tag{11}$$

利用归纳法很容易证明这样构造的策略π'满足(10)式. 对(10)式两端取sup后再令ϵ趋于0得到结果。 □

定理2.3是在最一般的策略类Π上得到的。根据定理1.2，在Π上的优化问题等价于在策略类Π^d上的优化问题，因此，下面我们局限在Π^d上讨论。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

3 最优策略的存在性和算法

前面一节中，定理2.3指出了最优方程的解就是从决策时刻t直到过程结束决策时刻N的最优值函数，而当$t = 0$时就是我们要求的马氏决策过程的最优值。下面我们介绍如何通过最优方程求得最优策略。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

定理2.4：假设$u_t^*, t \le N$，是方程(7)的解,而且满足边界条件(8)。如果策略$\pi^* = (\pi_0^*, \pi_1^*, \cdots, \pi_{N-1}^*) \in \Pi^d$满足:

$$r_t(i_t, a_t^*(h_t)) + \sum_{j \in S} p_t(j|i_t, a_t^*(h_t)) u_{t+1}^*(h_t, a_t^*(h_t), j)$$

$$= \max_{a \in A(i_t)} \left\{ r_t(i_t, a) + \sum_{j \in S} p_t(j|i_t, a) u_{t+1}^*(h_t, a, j) \right\} \tag{12}$$

其中$\pi_t^*(a_t^*(h_t)|h_t) = 1$为退化分布, 对一切$t = 0, 1, \cdots, N-1$。那么有

a) 对一切$t = 0, 1, \cdots, N-1$,

$$u_t^{\pi^*}(h_t) = u_t^*(h_t), \quad h_t \in H_t. \tag{13}$$

b) π^*是最优策略，而且

$$V_N^{\pi^*}(i) = V_N^*(i), \quad i \in S. \tag{14}$$

证明思路:用归纳法可以直接证明。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

定理2.4表明，在解得最优方程之后，那些满足达到上确界的行动组合成为最优策略。为了方便，我们引入记号

$$\arg\max_{x\in X} g(x) = \{x' \in X | g(x') \geq g(x) \forall x \in X\}.$$

则，公式(12)可以简化为

$$a_t^*(h_t) \in \arg\max_{a\in A(i_t)} \left\{ r_t(i_t, a) + \sum_{j\in S} p_t(j|i_t, a) u_{t+1}^*(h_t, a, j) \right\}. \qquad (15)$$

一般来说(15)右端中的集合可能为空集，也就是说没有行动能够达到上确界，这样的话最优策略就不存在。但是可以证明ϵ最优策略总是存在的。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

定理2.5：令$\epsilon > 0$并假定$u_t^*, t \le N$，是方程(7)的解,而且满足边界条件(8)。如果策略$\pi^\epsilon = (\pi_0^\epsilon, \pi_1^\epsilon, \cdots, \pi_{N-1}^\epsilon) \in \Pi^d$ 满足:

$$r_t(i_t, a_t^\epsilon(h_t)) + \sum_{j \in S} p_t(j|i_t, a_t^\epsilon(h_t)) u_{t+1}^\epsilon(h_t, a_t^*(h_t), j) + \frac{\epsilon}{N}$$

$$\ge \sup_{a \in A(i_t)} \left\{ r_t(i_t, a) + \sum_{j \in S} p_t(j|i_t, a) u_{t+1}^\epsilon(h_t, a, j) \right\} \tag{16}$$

对一切$t = 0, 1, \cdots, N-1$。那么有

a) 对一切$t = 0, 1, \cdots, N-1$,

$$u_t^{\pi^\epsilon}(h_t) + (N-t)\frac{\epsilon}{N} \ge u_t^*(h_t), \quad h_t \in H_t. \tag{17}$$

b) π^ϵ是ϵ最优策略，而且

$$V_N^{\pi^\epsilon}(i) + \epsilon \ge V_N^*(i), \quad i \in S. \tag{18}$$

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

定理2.4和定理2.5中的最优策略和ϵ最优策略是属于策略类Π^d的.但是Π^d中的策略是与历史相关的策略,很不方便使用。下面的定理说明如果最优方程有解,则存在马氏策略是ϵ最优的或者是最优的。

定理2.6：假设$u_t^*, t \leq N$，是方程(7)的解,而且满足边界条件(8)。那么:
a) 对$t = 0, 1, \cdots, N$，$u_t^*(h_t)$对历史$h_t \in H_t$的依赖只是与h_t的最后一个元素$i_t \in S$有关系。
b) 令$\epsilon > 0$，存在ϵ最优策略是马氏策略。
c) 对$t = 0, 1, \cdots, N$和所有的$i_t \in S$，存在$a' \in A(i_t)$满足(15)，那么存在最优策略而且它还是马氏策略。

证明思路:用归纳法可以直接证明。

因此，我们有：

$$V_N^*(i) = \sup_{\pi \in \Pi} V_N(i, \pi) = \sup_{\pi \in \Pi_m^d} V_N(i, \pi), \quad i \in S. \tag{19}$$

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

注2.1：由定理2.6，我们可以比较容易的判断最优马氏策略的存在性.比如当对一切$i \in S$, $A(i)$是有限集合的时候；或者,对一切$t = 0, 1, \cdots, N$和$i \in S$，$A(i)$是紧致的，$r_t(i,a)$关于a(上半)连续[a]且对各个参数一致有界以及$p_t(j|i,a)$关于a（下半）连续[b]时等等。在这一章里我们总认为上面的一组条件成立,以保证(15)右端中的集合非空.

[a] $\Re$上实值函数f被称为**上半连续**的,是指对于$\Re$中每一点x^*和所有收敛于x^*的点列$\{x_n\}$都成立有$\limsup_{n\to\infty} f(x_n) \le f(x^*)$.

[b] $\Re$上实值函数f被称为**下半连续**的,是指$-f$是上半连续的.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

算法2.2 (有限阶段向后递归迭代算法)

步骤1: 令$t = N$且对一切$i_N \in S$,

$$u_N^*(i_N) = r_N(i_N).$$

步骤2: 如果$t = 0$,则$\pi = (f_0^*, f_1^*, \cdots, f_{N-1}^*)$为最优的马氏策略,而$V_N^*(i) = u_0^*(i)$为最优的值函数,算法停止。否则,令$t - 1 \Rightarrow t$后,进入步骤3。

步骤3: 对一切$i_t \in S$计算

$$u_t^*(i_t) = \max_{a \in A(i_t)} \left\{ r_t(i_t, a) + \sum_{j \in S} p_t(j|i_t, a) u_{t+1}^*(j) \right\}, \tag{20}$$

并且记集合

$$A_t^*(i_t) = \arg\max_{a \in A(i_t)} \left\{ r_t(i_t, a) + \sum_{j \in S} p_t(j|i_t, a) u_{t+1}^*(j) \right\}, \tag{21}$$

并任意取定$f_t^*(i_t) \in A_t^*(i_t)$,这样就定义了$t$时刻的决策规则$f_t^*$。

步骤4: 返回到步骤2.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

集合$A_t^*(i_t)$通常被称为**最优行动集**.向后递归的算法正体现了人们常说的“**最优化原理**”.最优化原理这个名词最早出现在Bellman的书里：“最优策略具有如下的性质，无论从那一个初始状态出发和采取了那一个初始行动，对下一个决策时刻来说剩余的决策规则组成的策略是最优策略。”定理2.4 实际上是最优化原理精确表述。Denardo给出了等价的表述：“存在一个策略，它对每一个状态（在每个阶段）都是最优的。”

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

4 两个例子

利用上面的工具，我们讨论两个应用的例子。

4.1. 序贯分配问题

一个决策者拥有M个单位的资源，希望在$N+1$个周期内利用或消费掉。如果记x_t为决策时刻t时所消费的资源数量,$f(x_0, x_1, \cdots, x_N)$则为决策者分配这M个单位的资源的效用函数(f为负值时则表示费用).选择最优消费方式的问题由下面的数学规划给出:

$$\max f(x_0, x_1, \cdots, x_N),$$
$$\begin{cases} \text{s.t.} \\ \quad x_0 + x_1 + \cdots + x_N = M \\ \quad x_t \geq 0, \quad t = 0, 1, \cdots, N \end{cases} \tag{22}$$

为了简化问题，我们这里假设目标函数$f(x_0, x_1, \cdots, x_N)$关于变量是可分的，也就是说它满足：

$$f(x_0, x_1, \cdots, x_N) = \sum_{t=0}^{N} g_t(x_t). \tag{23}$$

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

这个问题可以转化为有限阶段马氏决策过程（也称为动态规划）问题。为了要构造马氏决策问题，先要描述t时刻的状态。我们把从决策时刻t到过程结束N时刻可用于消费的资源总量作为状态，行动是决策时刻t时的消费量。那么，如果t时刻的状态是i（决策者还有i个单位的资源需要分配到时刻$t, t+1, \cdots, N$上去消费），而此时决定这个时刻要消费a个单位的资源，则时刻$t+1, \cdots, N$上可以用于消费的资源还剩$i-a$个单位。在N时刻，决策者是没有什么可以选择的了，因为在前N个周期$0, 1, \cdots, N-1$上的消费量已经都被确定下来了。注意，对于这个问题，状态和行动都不必要是离散的，可以是实直线上的紧致区间。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

具体来说，马氏决策问题可以按照下面的方式定义。

决策时刻：

$$T = \{0, 1, \cdots, N\} \quad N < \infty.$$

可能的状态：

$$S = [0, M] \quad M < \infty.$$

可用的行动集：

$$A(i) = [0, i] \quad 0 \le a \le i.$$

报酬值：

$$\begin{aligned} r_t(i, a) &= g_t(a), \quad i \in S,\ a \in A(i),\ t = 0, \cdots, N-1 \\ r_N(i) &= g_N(i), \quad i \in S. \end{aligned}$$

转移概率：

$$p_t(j|i, a) = \begin{cases} 1 & j = i - a \\ 0 & \text{其他} \end{cases} \qquad i \in S,\ a \in A(i),\ t = 0, \cdots, N-1.$$

这样，我们就定义好了解决这个问题的马氏决策问题。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

但下面我们就一个具体的问题，来解决序贯分配问题。我们希望极大化目标函数$f(x_0, x_1, \cdots, x_N) = -x_0^2 - x_1^2 - \cdots - x_N^2$. 问题可以将规划(22)转化为

$$\min x_0^2 + x_1^2 + \cdots + x_N^2$$

$$\begin{cases} \text{s.t.} \\ \quad x_0 + x_1 + \cdots + x_N = M \\ \quad x_t \geq 0, \ \ t = 0, 1, \cdots, N \end{cases}$$

利用向后递归的算法，我们有下面的步骤：

第一步：取$t = N$，这时没有可以选择的决策，所以$u_N^*(i) = g_N(i) = i^2$, $i \in [0, M]$.

第二步：由于$t \neq 0$，所以令$t = N - 1$而且

$$u_{N-1}^*(i) = \min_{a \in [0,i]} \{a^2 + u_N^*(i-a)\} = \min_{a \in [0,i]} \{a^2 + (i-a)^2\}. \qquad (24)$$

只要求出使公式(24)右端达到极小的a即可。即对其求导数解出极值点和极值分别为：

$$\arg\min_{a \in [0,i]} \{a^2 + (i-a)^2\} = A_{N-1}^*(i) = \frac{i}{2};$$

$$u_{N-1}^*(i) = \frac{i^2}{2}.$$

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

第三步：由于$t \neq 0$，所以令$t = N-2$，则有

$$\begin{aligned} u^*_{N-2} &= \min_{a\in[0,i]}\{a^2 + \frac{1}{2}(i-a)^2\} = \frac{i^2}{3}; \\ A^*_{N-2}(i) &= \frac{i}{3}. \end{aligned}$$

通过观察，我们猜测一般的结论应该是：

$$\begin{aligned} A^*_{N-n+1}(i) &= \frac{i}{n}, \\ u^*_{N-n+1}(i) &= \frac{i^2}{n}. \end{aligned}$$

这一点可以用归纳法证明。所以最优的值函数为$u^*_0 = \frac{M^2}{N+1}$，每个决策时刻的最优行动是$a = \frac{M}{N+1}$.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

4.2. 秘书问题

这是动态规划的一个经典问题，是由Cayley(1875)年为抽彩票寻找最优策略时考虑过的。用秘书问题来说明它，主要是增加一些趣味性。假定需要聘用一名秘书，有N个候选人应聘。按照他们的能力可以排成一队$1, 2, \cdots, N$，排在1位的自然是最好的选择。但是这个排位的顺序并不知道，在面试一位候选人之后，要立即决定是否录用这位候选人。如果不录用他/她，则继续面试下一位候选人；如果录用，则聘用过程停止。问题是既有可能错过了排在1位的候选人，也有可能在聘用过程结束时排在1位的候选人还没有出现。决策者的目标是要保证选到排名第一的候选人的概率最大。决策是在面试之后立即做出的，所以决策的阶段数与候选人数目相同。为了叙述上的方便，我们这里从阶段1开始（与前面的模型从0开始没有本质的区别）到阶段N时结束。状态空间S只有两个元素0和1。其中1表示当前的候选人比前面的所有候选人都好；0则表示当前的候选人不如前面最好的优秀。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

对于每一个状态（面试完当前的候选人并对其与前面已经拒绝的那些候选人进行比较之后）,决策者都有两个行动可以选择：选取行动Q表示接受当前的候选人；选取行动C表示拒绝当前的候选人,并对下一个候选人开始面试.根据要求,除了在过程停止的时候,所有其它时候的报酬都是0,也就是在选取行动C时的报酬总是0.用我们熟悉的记号表示为：状态空间$S' = \{0, 1\}$; 可用的行动集$A(0) = A(1) = \{C, Q\}$; 报酬值$r_t(0, C) = r_t(1, C) = 0, r_t(0, Q) = 0, r_t(1, Q) = \frac{t}{N}, t = 1, 2, \cdots, N.$

上面需要解释的是$r_t(1, Q)$的取值。它表示了选中的候选人是所有候选人中最好的概率，是这样确定的：

$$P\{\text{最好者在前}t\text{个人中}\} = \frac{\text{从1到}N\text{中取}t\text{个且包含1}}{\text{从1到}N\text{中取}t\text{个}} = \frac{C_{N-1}^{t-1}}{C_N^t} = \frac{t}{N}$$

其中C_N^t是二项式展开的系数。另外，为了描述聘用过程的结束情况，我们用Δ表示过程的停止状态。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

转移概率是不依赖于当前状态i的,而且只要采取C,过程就会继续下去.因此$p_t(j|i,C)=p_t(j)$,$i=0,1$。在时刻$t+1$恰好选到前$t+1$个人中的最优秀者的概率是$p_{t+1}(1|i,C)=1/(t+1)$，未选中的概率是$p_{t+1}(0|i,C)=t/(t+1)$, 对于$i=0,1$。

我们给出马氏决策问题的具体细节。

决策时刻：

$$T=\{1,\cdots,N\}\quad N<\infty.$$

可能的状态：

$$S=S'\cup\{\Delta\}=\{0,1,\Delta\}.$$

可用的行动集：

$$A(i)=\begin{cases}\{C,Q\} & i\in S'\\ \{C\} & i=\Delta\end{cases}$$

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

报酬值：

$$r_t(i,a)=\begin{cases}0 & i\in S'\ a=C\\ 0 & i=0,\ a=Q,\ t<N\\ \frac{t}{N} & i=1,\ a=Q,\ t<N\\ 0 & i=\Delta\end{cases}$$

转移概率：

$$p_t(j|i,a)=\begin{cases}\frac{1}{t} & i\in S',\ a=C,\ j=1\\ \frac{t-1}{t} & i\in S',\ a=C,\ j=0\\ 1 & i\in S',\ a=Q,\ j=\Delta\\ 1 & i=j=\Delta,\ a=C\\ 0 & \text{其他}\end{cases}\qquad \text{这里 } t=1,\cdots,N-1.$$

这样，我们就定义好了解决这个问题的马氏决策问题。下面我们给出求解的过程。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

我们用$u_t^*(1)$表示从当前的时段到过程结束决策者能够选到最好候选人的最大概率,而此时刚刚面试过的候选人恰是前面所有面试过的候选人中最好者;用$u_t^*(0)$表示在剩下的时段中(不包括当前的时段,因为此时不会选择接受这个候选人的决策)决策者能够选到最好候选人的最大概率,而此时刚刚面试过的候选人不是前面所有面试过的候选人中最好者.那么，它们满足下面的关系：

$$u_N^*(1) = r_N(1) = 1; \quad u_N^*(0) = r_N(0) = 0; \quad u_N^*(\Delta) = 0.$$

而且对于$t = 1, \cdots, N-1$，有

$$\begin{aligned} u_t^*(1) &= \max\{r_t(1,Q) + u_{t+1}^*(\Delta), p_t(1|1,C)u_{t+1}^*(1) + p_t(0|1,C)u_{t+1}^*(0)\} \\ &= \max\left\{\frac{t}{N}, \frac{1}{t+1}u_{t+1}^*(1) + \frac{t}{t+1}u_{t+1}^*(0)\right\}, \qquad (25) \\ u_t^*(0) &= \max\{r_t(0,Q) + u_{t+1}^*(\Delta), p_t(1|0,C)u_{t+1}^*(1) + p_t(0|0,C)u_{t+1}^*(0)\} \\ &= \max\left\{0, \frac{1}{t+1}u_{t+1}^*(1) + \frac{t}{t+1}u_{t+1}^*(0)\right\}, \qquad (26) \end{aligned}$$

以及

$$u_t^*(\Delta) = u_{t+1}^*(\Delta) = 0.$$

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

我注意到$u_t^* \geq 0$，上面的式子可以化简为

$$u_t^*(0) = \frac{1}{t+1}u_{t+1}^*(1) + \frac{t}{t+1}u_{t+1}^*(0) \tag{27}$$

$$u_t^*(1) = \max\left\{\frac{t}{N}, u_t^*(0)\right\}. \tag{28}$$

求解(27)和(28)可以得到最优策略，而且最优策略具有这样的结构：t时刻如果在状态1，有$\frac{t}{N} > u_t^*(0)$，最优行动是停止；如果$\frac{t}{N} < u_t^*(0)$，最优行动就是继续面试下一个候选人；如果$\frac{t}{N} = u_t^*(0)$，两者都是最优行动。在状态0，继续下去是最优的选择。用语言来说明：当候选人数目N确定以后，最优策略是先观察$\tau(N)$个候选人,然后录用第一个好过前面所有已经面试过的候选人的人.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

我当N很大的时候,$\tau(N) \approx \frac{N}{e}$。关于$\tau(N)$的确定过程,可以参见Puterman的书中4.6.4节或者参见胡奇英,刘建庸书的2.2.2节.这个结果看上去满有点儿经验的味道，1990年9月12日的The Globe and Mail中A22页上有题为“最后的是最好的”的报道：根据加拿大的Runzheimer公司的调查，最后一个面试者得到工作的机会是55.8%,而早期的申请者得到工作的机会只有17.6%.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

5 最优策略的结构

Bellman的最优化原理是作为原理提出来的,后来有很多人讨论这个原理成立的条件. 例如针对确定性动态规划和随机动态规划的一些具体问题, 很多学者探讨了Bellman最优化原理成立的一些条件, 如[156], [178]以及[177]等等. 针对我们这个具体的模型而言, 我们可以一般的进行讨论.

定义2.2: 设$\pi = (\pi_0, \pi_1, \cdots) \in \Pi$。如果历史$h_{t-1} = (i_0, a_0, i_1, a_1, \cdots, i_{t-1}) \in H_{t-1}$, 行动$a_{t-1} \in A(i_{t-1})$和状态$j \in S$满足:

$$\pi_0(a_0|i_0)p(i_1|i_0, a_0)\pi_1(a_1|i_0, a_0, i_1)\cdots p(j|i_{t-1}, a_{t-1}) > 0, \tag{29}$$

则称这个历史h_{t-1}为策略π下直到时刻$t-1$**可实现历史**。特别的, 我们称t时刻的状态j是在π下通过可实现历史h_{t-1}和行动a_{t-1}于时刻t的**可达状态**。此时$h_t = (h_{t-1}, a_{t-1}, i_t)$是$\pi$下到$t$时刻的可实现历史. (29)式有时记做:$P_\pi\{h_t|i_0\} > 0$.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

有时候为了方便，我们把可实现历史做如下描述：对于一个策略$\pi = \{\pi_0, \pi_1, \cdots\} \in \Pi$，有一个历史$h_t = (i_0, a_0, i_1, a_1, \cdots, i_t) \in H_t$(参见1.3.4节的定义)，如果$P_\pi\{h_t|i_0\} > 0$，即在由策略$\pi$生成的概率测度下，事件$h_t$的概率是正的，那么$h_t$被称为策略$\pi$下的一个可实现历史。通俗的解释为：在策略$\pi$下，从状态$i_0$出发，采用行动$a_0$以后状态发生转移，到达状态$i_1$；再采取行动$a_1$，继续下去直到时刻$t$时状态转移到$i_t$，整个事件在策略$\pi$诱导出的概率测度下发生的概率不是0，那么这样的历史就是在策略π下的一个可实现历史。这里描述的可实现历史与定义 5所述的略有不同，但本质是一样的。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

定理2.7：假设$\pi \in \Pi$是最优策略. $1 \leq t \leq N-1, j \leq S$. 如果$j$是在$\pi$下通过可实现历史$h_{t-1}$和行动$a_{t-1}$于时刻$t$可达, 则$u_t^{\pi}(h_{t-1}, a_{t-1}, j) = u_t^*(j)$。

证明: 我们对t进行归纳证明.

由于π是最优策略,所以有$V_N(i, \pi) = u_0^{\pi}(i) = u_0^*(i), i \in S$.

根据定理2.6的a),对任意的历史$h_1 = (i, a, j)$, 我们有$u_1^{\pi}(h_1) \leq u_1^*(j)$.

如果$h_1 = (i_0, a_0, j)$是π的一个可实现历史,那么状态j在时刻$t = 1$可达,且有$\pi_0(a_0|i_0)p(j|i_0, a_0) > 0$.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

进一步,如果$u_1^\pi(i_0,a_0,j)<u_1^*(j)$,对状态$i_0$考虑:

$$\begin{aligned} u_0^\pi(i_0) &= V_N(\pi,i_0)=\sum_{a\in A(i_0)}\pi_0(a|i_0)\left[r_0(i_0,a)+\sum_{i_1\in S}p(i_1|i_0,a)u_1^\pi(i_0,a,i_1)\right] \\ &< \sum_{a\in A(i_0)}\pi_0(a|i_0)\left[r_0(i_0,a)+\sum_{i_1\in S}p(i_1|i_0,a)u_1^*(i_1)\right] \qquad (30) \\ &\le \sum_{a\in A(i_0)}\pi_0(a|i_0)u_0^*(i_0)=u_0^*(i_0). \end{aligned}$$

产生了矛盾。所以，在任意的π下可实现历史$h_1=(i_0,a_0,j)$我们必有：$u_1^\pi(h_1)=u_1^*(j), j\in S$。这就证明了定理对$t=1$时成立。其中(30)式的严格不等号是因为该不等号左边的和式中有一项为：

$$\pi_0(a|i_0)p(j|i_0,a)u_1^\pi(i_0,a,j)<\pi_0(a|i_0)p(j|i_0,a)u_1^*(j),$$

而左边其他各项分别不超过不等式右边的对应项。

仿照上述证明方法,由归纳法易证定理结论成立。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

定理2.7说明,如果$\pi = (\pi_0, \pi_1, \cdots)$是最优策略，则对于时刻$t(1 \leq t \leq N-1)$可达的状态$j$, $(\pi_t, \pi_{t+1}, \cdots)$构成了后面$N-t$个阶段的最优策略，也就是Bellman的最优化原理。
在前面一节我们已经定义了最优行动集合的概念，我们有下面更加深刻的结论。

定理2.8：假设$\pi = (\pi_0, \pi_1, \cdots) \in \Pi$. π是最优策略的充要条件是下面两条成立:
(i) 对任意的时刻$t(1 \leq t \leq N-1)$，如果$j \in S$在π下通过一个可实现历史h_t可达，则当$a \in A(j) - A_t^*(j)$时，有$\pi_t(a|h_t) = 0$;
(ii) 对任意的状态$j \in S$，当$a \in A(j) - A_0^*(j)$时，有$\pi_0(a|j) = 0$。
证明:利用定理2.7的证明方法结合反证法可以证明结论的必要性。由定理2.6可以得到充分性的证明。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

定理2.8使我们清晰的看到了最优策略的结构。粗略的说，一个策略是最优的，当且仅当在每个决策时刻的决策规则，对可达的那些状态都必须选用最优行动。

Evans[76]曾未加证明的指出：如果在随机马氏策略类Π_m中存在最优策略，则必然存在马氏策略$\pi \in \Pi_m^d$在Π_m上是最优的。我们这里有下面的结论。

为了一般的讨论,我们考虑对任意的一个策略$\pi = (\pi_0, \pi_1, \cdots) \in \Pi$,可以定义一个马氏策略与之对应. 具体为:对任意的$j \in S$,显然存在一个$a_0 \in A(j)$使得$\pi_0(a_0|j) > 0$. 令$g_0(j) = a$. 对$1 \leq t \leq N-1$, $j \in S$. (i)如果状态j在π下于时刻t可达, 我们可以取到一个t时可实现历史h_t,使得状态j于t时刻达到.易知,存在一个$a_t \in A(j)$使得$\pi_t(a_t|h_t) > 0$. 令$g_t(j) = a_t$. (ii)如果(i)不成立, 则任取一个$a_t \in A(j)$, 令$g_t(j) = a_t$. 这样我们得到一个策略$\pi' = (g_0, g_1, \cdots)$, 显然有$\pi' \in \Pi_m^d$. 当然,这样的对应并不唯一. 但是我们有下面的结论.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

引理2.1：设$1 \leq t \leq N-1$.如果$j \in S$在策略π'下于时刻t可达, 则j在π下于时刻t也可达.

证明:我们对t进行归纳证明.

如果j在π'下于时刻1可达,即有历史$h_1 = (i_0, a_0, j)$使得$p(j|i_0, a_0) > 0$. 由g_0的定义,$\pi_0(g_0(i_0)|i_0) > 0$知道$h_1 = (i_0, a_0, j)$也是π下的可实现历史,所以状态j在π下于时刻1可达.此即$t = 1$时结论成立.

设$1 \leq t < N-1$时结论成立. 如果状态$j \in S$于时刻$t+1$在π'下可达,那么就存在一个π'的可实现历史$h_{t+1} = (i_0, a_0, \cdots, i_t, a_t, j)$使得$P_{\pi'}\{h_{t+1}|i_0\} > 0$. 很明显有$P_{\pi'}\{h_t|i_0\} > 0$,由归纳假设知道$i_t$在策略$\pi$下于时刻$t$ 可达. 再由g_t的定义和$p(j|i_t, a_t) > 0$,知道h_{t+1}也是策略π的可实现历史,特别的,状态j在π下于时刻$t+1$可达. 因此,结论对$t+1$成立, 引理得证。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

定理2.9：如果在Π上存在最优策略,则必存在马氏策略$\pi \in \Pi_m^d$在Π上是最优的.

证明:设$\pi = (\pi_0, \pi_1, \cdots) \in \Pi$是最优策略. 对任意的$j \in S$,显然存在一个$a_0 \in A(j)$使得$\pi_0(a_0|j) > 0$. 令$g_0(j) = a$. 由定理 知道$a_0 \in A_0^*(j)$.

设$1 \leq t \leq N-1$, $j \in S$. (i)如果状态j在π下于时刻t可达,我们可以取到一个t时可实现历史h_t,使得状态j于t时刻达到.易知,存在一个$a_t \in A(j)$使得$\pi_t(a_t|h_t) > 0$. 令$g_t(j) = a_t$. 再由定理 知道$a_t \in A_t^*(j)$. (ii)如果(i)不成立, 则任取一个$a_t \in A(j)$, 令$g_t(j) = a_t$. 这样我们得到一个策略$\pi' = (g_0, g_1, \cdots)$, 显然有$\pi' \in \Pi_m^d$,而且它的任何一个可实现的历史也是$\pi$的可实现历史.结合引理2.1和定理2.8,知道$\pi'$是一个$\Pi$上的最优策略.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

因此，我们有：

$$V_N^*(i) = \sup_{\pi\in\Pi} V_N(i,\pi) = \sup_{\pi\in\Pi_m^d} V_N(i,\pi), \quad i \in S. \tag{31}$$

这就表明了：如果存在马氏策略π在Π_m^d上是最优的，则π也是Π上的最优策略。否则在Π上不存在最优策略。而寻找最优策略的问题就完全转化到马氏策略类Π_m^d中的相应问题，也可以通过前面的算法找到Π上的最优策略或者ϵ最优策略。事实上，对于S和$A(i)$更一般的情形，我们也有类似的结论。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

6 单调策略的最优性

在第三节中,我们给出了一些条件保证了存在(确定性)马氏策略是最优的.在这一节中我们给出进一步的条件, 以保证最优策略是单调的(关于系统的状态非降或非增).这样会大大简化了最优策略的结构,不仅能满足决策者的要求, 而且具有易操作性和易计算性.为了使这个概念有意义，我们需要状态具有物理解释和自然的序关系。这里**单调策略** 的含义就是单调确定性马氏策略。

定义2.3：令$g(x,y)$为$\Re \times \Re$上的实值二元函数，如果对任意的$x_1 < x_2$和$y_1 < y_2$, 满足：

$$g(x_2,y_2)+g(x_1,y_1) \geq g(x_2,y_1)+g(x_1,y_2). \tag{32}$$

我们称函数$g(x,y)$为**上可加的** (superadditive)或者**上模的** (supermodular). 如果公式(32)的不等号反过来，则称函数$g(x,y)$为**下可加的** (subadditive) 或者**下模的** (submodular)。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

我们用下面的图来说明上可加函数定义。

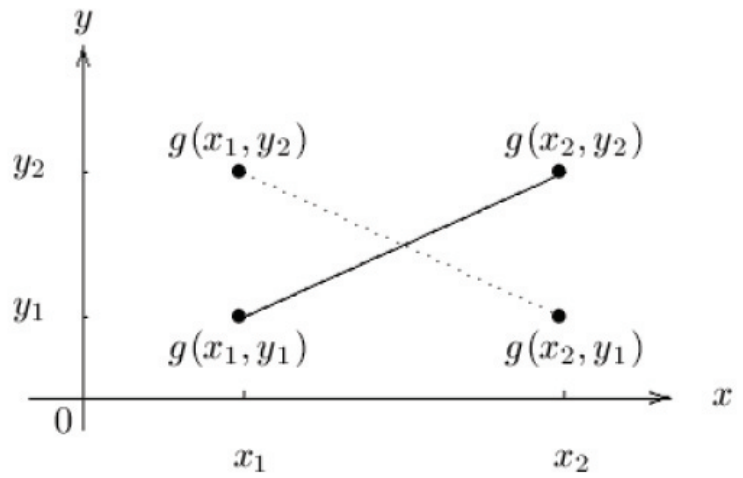

图 2.1　上可加函数，实线相连的两点函数值和不小于虚线相连的两点和

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

下面我们给出一些上可加函数的例子：

1) $g(x,y)=h(x)+e(y)$,其中$h(\cdot)$和$e(\cdot)$为$\Re$上的任意函数.

2) $g(x,y)=h(x+y)$,这里$h(\cdot)$为$\Re$上的凸函数.

3) $g(x,y)=xy$.

特别的,$(x+y)^2$和$-(x-y)^2$都是上可加的,$(x-y)^2$和$-(x+y)^2$是下可加的.如果$g(x,y)$二次可微的话, 若$\frac{\partial^2 g(x,y)}{\partial x\partial y}\geq 0$则$g(x,y)$是上可加的；而$\frac{\partial^2 g(x,y)}{\partial x\partial y}\leq 0$则$g(x,y)$是下可加的。另外，如果定义

$$h(x)=\max\left\{y'\in\arg\max_y g(x,y)\right\}, \tag{33}$$

当$g(x,y)$是上可加的，则$h(x)$关于x是非增的。也就是说随着x的增加，$h(x)$的值可能不变或者减少，但它不会增加。当$g(x,y)$是下可加的，则$h(x)$关于x是非降的。

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

设S是非负整数;一切行动集合都一样,即$A(i) \equiv A \subset \Re$. 记

$$q_t(k|i,a) = \sum_{j=k}^{\infty} p_t(j|i,a) \quad t = 0, 1, \cdots, N-1. \tag{34}$$

公式(34)表示t时,在状态i选取行动a后,$t+1$时超过状态$k-1$的概率.

定理2.10: 假设$t = 0, 1, \cdots, N-1$,而且:

1) 对一切$a \in A$, $r_t(i,a)$关于i非增,

2) 对一切$a \in A$和$k \in S$, $q_t(k|i,a)$关于i非增,

3) $r_t(i,a)$是$S \times A$上的上(下)可加函数,

4) 对一切$k \in S$, $q_t(k|i,a)$是$S \times A$上的上(下)可加函数,

5) $r_N(i)$关于i非降.

那么,存在最优策略$\pi^* = \{f_0^*, f_1^*, \cdots, f_{N-1}^*\}$满足:对一切$t = 0, 1, \cdots, N-1$, 最优决策函数$f_t^*(i)$关于$i$非降(增)[a].

证明思路:用递归的方法证明$u_t^*(i)$关于i非降或者非增,只需要证明

$$w_t(i,a) = r_t(i,a) + \sum_{j=0}^{\infty} p_t(j|i,s) u_t^*(j)$$

为上可加的或者下可加的即可。

[a] 简单的单调决策函数可以描述为:当$i < i^*$时采用a_1,当$i \geq i^*$时采用a_2.实际上可以定义一般的单调决策函数.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

定理2.10只是表明了存在单调最优策略,并没有说最优策略一定是单调的.事实上可能存在非单调的最优策略.

由于存在单调策略是最优的那些条件很不容易验证,所以我们给出下个定理,虽然没有降低验证的难度,但总是从另一个角度给出了一种验证方式.

定理2.11：设$t = 0, 1, \cdots, N-1$,而且:

1) 对一切$a \in A$, $r_t(i, a)$关于i非增,

2) 对一切$a \in A$和$k \in S$, $q_t(k|i, a)$关于i非增,

3) $r_t(i, a)$是$S \times A$上的上可加函数,

4) 对一切非增的u, $\sum_{j=0}^{\infty} p_t(j|i, a)u(j)$是$S \times A$上的上可加函数,

5) $r_N(i)$关于i非降.

那么,存在最优策略$\pi^* = \{f_0^*, f_1^*, \cdots, f_{N-1}^*\}$满足:对一切$t = 0, 1, \cdots, N-1$, 最优决策函数$f_t^*(i)$关于$i$非降.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

在排队或者存储模型的应用中,假设条件$A(i) \equiv A$太强了,通常我们可以将其修正为下面的条件:

1) 对一切$i \in S$, $A(i) \subset A$,

2) 当$i' > i$时, $A(i) \subset A(i')$,

3) 对每个i和$a \in A(i)$, 如果$a' \leq a$, 则必有$a' \in A(i)$.

下面我们给出一个寻求单调决策规则的向后递归算法，在算法中我们假设存在单调决策规则，状态空间有限$S = \{0, 1, 2, \cdots, M\}$,而且对每个$i \in S$有$A(i) = A$.

最优准则

有限阶段的策略迭代和...

最优策略的存在性和算法

两个例子

最优策略的结构

单调策略的最优性

算法2.3 (单调向后递归算法)

步骤1: 令$t = N$且对一切$i_N \in S$,

$$u_N^*(i_N) = r_N(i_N),$$

步骤2: 以$t-1$代t, 置$i = 0$以及$A'(0) = A$。

步骤3: 令

$$u_t^*(i) = \max_{a \in A'(i)} \left\{ r_t(i,a) + \sum_{j \in S} p_t(j|i,a) u_t^*(j) \right\}.$$

步骤4: 令

$$A_t^*(i) = \arg\max_{a \in A'(i)} \left\{ r_t(i,a) + \sum_{j \in S} p_t(j|i,a) u_t^*(j) \right\}.$$

步骤5: 如果$i = M$,进入步骤7,否则令

$$A'(i+1) = \{a \in A | a \geq \max[a' \in A_t^*(i)]\}.$$

步骤6: 以$i+1$代替i,返回到步骤3.

步骤7: 如果$t = 0$,停止;否则回到步骤2.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

在任意t时刻,通过算法2.3得到:在状态i从行动集$A_t^*(i)$中选取行动所组成的规则是单调最优的. 这个算法与3节中的有限阶段向后递归值迭代算法是不同的,主要是选优的范围为不同.一般来说,算法2.3中的$A'(i)$比通常的可用行动集合A要小的多.

例：定价问题

市场的管理人员希望根据当前某产品的销售情况决定其最优的价格.记$i \in S = \{0, 1, \cdots\}$为月销售量. 在每个月的开始,决策者了解上个月的销售情况并依此选择当前的销售价格$a \in A = \{a | a_L \leq a \leq a_U\}$, 这里$a_L$和$a_U$分别为最低和最高销售价格.以$r_t(i, a)$表示在第$t-1$月的销量是$i$而且第$t$月的价格定为$a$时第$t$月的期望收益.在第$t$月的实际销售量为$j$的发生概率为$p_t(j|i, a)$.由于产品的寿命是$N$个月,所以$r_N(i) = 0$, 对一切$i \in S$.决策者希望适当的选取他/她的策略能够极大化在产品淘汰前的总期望利润.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

下面我们讨论一下定理2.7中间的那些条件的意义.请注意$r_t(i,a)$和$p_t(j|i,a)$对固定的i关于a连续, 而且可用的行动集合是个闭区间(紧致性条件满足),因此极大值的点总能达到.

1) $r_t(i,a)$对固定的a关于s非降的意思就是:如果售价a不变,上个月的销售量越大,这个月的期望收益越大.

2) $q_t(k|i,a)$(见公式(34))对固定的a和k关于i非降的意思是:上个月的销售额越大,这个月销售额超过k的概率越大.

3) 由公式(32),$r_t(i,a)$的上可加意味着:

$$r_t(i^+,a^+)-r_t(i^+,a^-)\geq r_t(i^-,a^+)-r_t(i^-,a^-),\quad i^+\geq i^-,\quad a^+\geq a^-.$$

用语言表示这个条件成立的意思就是:只要是上个月的销售额越大,降价会对这个月产生的利润增量就越大.

4) $q_t(k|i,a)$的上可加意味着:

$$q_t(k|i^+,a^+)-q_t(k|i^+,a^-)\geq q_t(k|i^-,a^+)-q_t(k|i^-,a^-).$$

如果这个条件成立,就意味着:如果当前的销售量越大,则由于价格下降使销售量超过某个固定水平的概率会变大.

当上面四个条件都满足时,最优的价格是销售量的非降函数,也就是说当前销售量越大,下个周期的最优价格就越高.

最优准则
有限阶段的策略迭代和...
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性

谢谢大家！

最优准则
有限阶段的策略迭代和 . . .
最优策略的存在性和算法
两个例子
最优策略的结构
单调策略的最优性