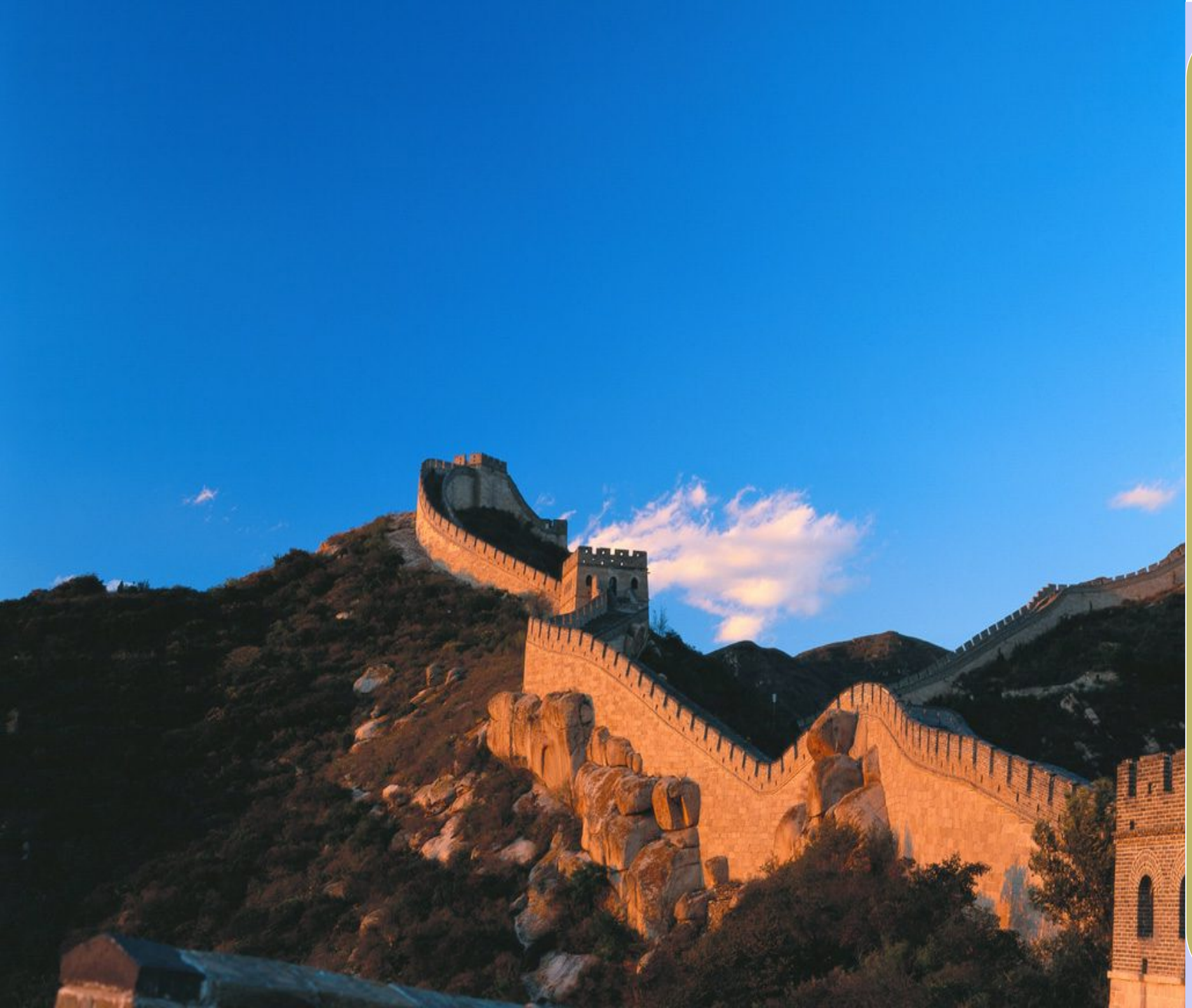

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

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
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

第三部分　马氏决策—折扣模型

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

1 最优准则

2 最优方程

3 最优策略的存在性

4 策略迭代算法

5 值迭代算法

6 改进的策略迭代算法

7 线性规划算法

8 可数状态与行动的模型

9 最优单调策略

10 最优策略的结构

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

1 最优准则

在这一章，我们讨论离散决策时刻、无限阶段的折扣马氏决策问题。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$T = \{0, 1, \cdots, \}$，$r(i,a)$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$0, 1, \cdots,$时依一定的概率获得一串报酬，报酬折现以后累加起来就是该模型的具体效用函数，称为无限阶段折扣模型，或简称为折扣模型，其折现率被称为**折扣因子**。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

对策略$\pi \in \Pi$和固定的折扣因子β，$0 < \beta < 1$，折扣模型的报酬效用函数定义为：

$$V_\beta(i,\pi) \equiv \sum_{t=0}^{\infty} \beta^t E_\pi^i[r(Y_t, \Delta_t)], \quad i \in S, \tag{1}$$

表示使用策略π，在0时刻从状态$i \in S$出发的条件下，系统折扣期望总报酬。用$V_\beta(\pi)$表示第i个分量为$V_\beta(i,\pi)$的列向量，当状态空间S可列时，$V_\beta(\pi)$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折扣期望总报酬，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$V_\beta(\pi)$是有界的，从而可以做出下面的定义。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

定义3.1：令

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

为最优值函数，用向量表示为V_β^*。对$\epsilon \geq 0$，如果策略π^*使得$V_\beta(i, \pi^*) \geq V_\beta^*(i) - \epsilon$对所有状态$i \in S$成立，则称$\pi^*$为折扣模型的$\epsilon$最优策略，简称为$\epsilon$最优策略，当$\epsilon = 0$时简称为最优策略。

同样, 这里还是寻求向量最优的问题。对于一般的向量优化问题，由于不存在良好的序关系，往往不存在最优的值函数。但是，在无限阶段的折扣马氏决策问题中，我们有非常好的结果，既在很一般的条件下，最优策略和最优值函数都是存在的。在这一章里, 我们不做特别的声明, 总认为报酬和转移概率是“平稳的”, 也就是它们不依赖于时间. 因此我们不必要写出它们的下标了. 这里我们还假设报酬是一致有界的, 即对一切$i \in S$和$a \in A(i)$, 有$|r(i, a)| \leq M < \infty$。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

2 最优方程

第一章我们定义一般策略是很广泛的，在最一般的策略类里寻找最优策略是很困难的，而且实用性也不强。但是，由定理1.2，每个初始状态$i \in S$和策略$\pi \in \Pi$, 存在随机马氏策略$\pi' \in \Pi_m$其诱导的过程与策略π所诱导出来的是一致的. 如果初始阶段的状态是根据其分布决定的, 该随机马氏策略π'不依赖于初始状态但可能依赖于那个初始的概率分布. 如此类推, 我们得到π'的期望总折扣报酬与策略π的相同, 所以我们有

$$V_\beta^*(i) = \sup_{\pi \in \Pi} V_\beta(i, \pi) = \sup_{\pi \in \Pi_m} V_\beta(i, \pi). \tag{3}$$

故寻找最优策略的范围就缩小到随机马氏策略类中。

记B为S上的有界实值函数的集合. 并且定义B上的**范数**为$||V|| \equiv \sup_{i \in S} |V(i)|$, 如果$V \in B$. 下面我们给出$B$上**偏序**的定义. 对一切$i \in S$,如果$V_1(i) \geq V_2(i)$, 我们记为: $V_1 \geq V_2$; 如果对一切$i \in S$, 有$V_1(i) = V_2(i)$, 记为$V_1 = V_2$; 如果$V_1 \geq V_2$, 且至少存在一个状态$i \in S$满足$V_1(i) > V_2(i)$, 我们记为$V_1 > V_2$. 记B_M为B的子集, 表示B上Borel可测函数的集合. 如果状态空间S是有限的或者是可数无限(赋予离散的拓扑)时, S上所有有界实值函数都是可测的, 所以这时$B = B_M$; 如果S不可数时, B_M一般是B的真子集.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

当S离散时, B中元素常常被称为向量, 而B上的线性算子被称为矩阵(正象第1节中已经使用过的那样). 用$|S|$表示S中元素的个数. 如果$f \in F$(或者$f^\infty \in \Pi_s^d$), 记向量$r(f)$的分量为$r(i, f(i))$; 转移概率矩阵$P(f)$的(i, j)分量为$p(j|i, f(i))$. 我们这里说的向量，总认为是列向量。在这一章里, 如果不是特别强调, 我们总认为状态空间是有限的或者可数的。

下面定理说明任一个随机马氏策略的总期望折扣报酬可以分为一周期的期望报酬与用第一个决策规则后以$V_\beta(\pi')$为终止报酬的和, 这里π'的定义可以参见定理3.1。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

定理3.1：任取策略$\pi = (\pi_0, \pi_1, \cdots) \in \Pi_m$, 状态$i \in S$,

$$\begin{aligned} V_\beta(i,\pi) &= \sum_{t=0}^{\infty} \beta^t E_\pi^i[r(Y_t, \Delta_t)] \\ &= \sum_{a\in A(i)} \pi_0(a|i) \left\{ r(i,a) + \beta \sum_{j\in S} p(j|i,a) V_\beta(j,\pi') \right\}, \end{aligned} \tag{4}$$

其中$\pi' = (\pi_1, \pi_2, \cdots) \in \Pi_m$. 也可以用向量和矩阵表示为:

$$V_\beta(\pi) = \sum_{t=0}^{\infty} \beta^t P^t(\pi) r(\pi_t) = r(\pi_0) + \beta P(\pi_0) V_\beta(\pi'), \tag{5}$$

其中$r(i,\pi_t) = \sum_{a\in A(i)} \pi_t(a|i) r(i,a)$, $P^t(\pi) = P(\pi_0) \times P(\pi_1) \times \cdots \times P(\pi_{t-1})$这里$P(\pi_{t-1})$的$(i,j)$分量为$p(j|i,\pi_{t-1}) = \sum_{a\in A(i)} \pi_{t-1}(a|i) p(j|i,a)$以及$P^0 = I$为单位矩阵.

证明: 利用状态空间的有限性和报酬函数的有界性，对任意的策略$\pi \in \Pi$，由公式(4)有：

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

$$
\begin{aligned}
V_\beta(i,\pi) &= \sum_{t=0}^{\infty} \beta^t E_\pi^i[r(Y_t,\Delta_t)] \\
&= \sum_{t=0}^{\infty} \beta^t \sum_{a\in A(i)} \pi_0(a|i) E_\pi[r(Y_t,\Delta_t)|Y_0=i,\Delta_0=a] \\
&= \sum_{a\in A(i)} \pi_0(a|i) \left\{ r(i,a) + \beta \sum_{t=1}^{\infty} \beta^{t-1} E_\pi[r(Y_t,\Delta_t)|Y_0=i,\Delta_0=a] \right\} \\
&= \sum_{a\in A(i)} \pi_0(a|i) \left\{ r(i,a) \right. \\
&\quad \left. + \beta \sum_{t=1}^{\infty} \beta^{t-1} \sum_{j\in S} p(j|i,a) E_\pi[r(Y_t,\Delta_t)|Y_0=i,\Delta_0=a,Y_1=j] \right\} \\
&= \sum_{a\in A(i)} \pi_0(a|i) \left\{ r(i,a) \right. \\
&\quad \left. + \beta \sum_{j\in S} p(j|i,a) \sum_{t=1}^{\infty} \beta^{t-1} E_\pi[r(Y_t,\Delta_t)|Y_0=i,\Delta_0=a,Y_1=j] \right\}.
\end{aligned}
$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

考虑到$\pi \in \Pi_m$，就有：

$$\begin{aligned}\sum_{t=1}^{\infty} \beta^{t-1} E_{\pi}[r(Y_t, \Delta_t)|Y_0 = i, \Delta_0 = a, Y_1 = j] &= \sum_{t=0}^{\infty} \beta^{t} E_{\pi'}[r(Y_t, \Delta_t)|Y_0 = j] \\ &= V_{\beta}(j, \pi').\end{aligned}$$

故定理得证。 □

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

引理3.1：. 如果状态空间S是离散的，且对一切$i \in S$和$a \in A(i)$有$|r(i,a)| \leq M$以及$0 \leq \beta \leq 1$. 那么只要$v \in B$和任意的决策规则π_t有:

$$r(\pi_t) + \beta P(\pi_t)v \in B. \tag{6}$$

证明: 由于$||\mathbf{r}(\pi_t)|| \leq M$和$||\mathbf{P}(\pi_t)\mathbf{v}|| \leq ||\mathbf{P}(\pi_t)||||\mathbf{v}|| = ||\mathbf{v}||$，引理得证。 □

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

引理3.2：. 设$\mathbf{P}$是一个$N \times N$的实矩阵，如果当$n \to \infty$时$\mathbf{P}^n \to \mathbf{0}$(元素都是0的$N \times N$的矩阵)。那么，$\mathbf{I} - \mathbf{P}$的逆矩阵存在，而且有：

$$(\mathbf{I} - \mathbf{P})^{-1} = \sum_{n=0}^{\infty} \mathbf{P}^n. \tag{7}$$

证明:对任何$n \geq 1$，总成立：

$$(\mathbf{I} + \mathbf{P} + \mathbf{P}^2 + \cdots + \mathbf{P}^{n-1})(\mathbf{I} - \mathbf{P}) = \mathbf{I} - \mathbf{P}^n.$$

当$n \to \infty$时，矩阵$\mathbf{P}^n \to \mathbf{0}$，所以当$n$充分大时，行列式$det(\mathbf{I} - \mathbf{P}^n) > 0$，故$det(\mathbf{I} - \mathbf{P}) \neq 0$，即$\mathbf{I} - \mathbf{P}$的逆矩阵存在。于是

$$\sum_{k=0}^{n-1} \mathbf{P}^k = (\mathbf{I} - \mathbf{P}^n)(\mathbf{I} - \mathbf{P})^{-1}. \tag{8}$$

令(8)中$n \to \infty$就得到公式(7)。 □

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

在定理3.1中，如果使用的策略是平稳策略，即$f^\infty \in \Pi_s^d$,则公式(5) 可以写成为

$$V_\beta(f^\infty) = \sum_{t=0}^{\infty} \beta^t P^t(f) r(f) = r(f) + \beta P(f) V_\beta(f^\infty). \tag{9}$$

也就是说$V_\beta(f^\infty)$是方程

$$v = r(f) + \beta P(f) v \tag{10}$$

的一个解。事实上，当$0 \leq \beta < 1$，我们将说明解是唯一的。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

当$v \in B$, 对确定性决策规则$f \in F$和随机决策规则π_0, 我们分别定义线性算子T_f和随机线性算子T_{π_0}为：

$$T_f v \equiv r(f) + \beta P(f)v \tag{11}$$

$$T_{\pi_0} v \equiv r(\pi_0) + \beta P(\pi_0)v. \tag{12}$$

我们还定义算子T为

$$Tv \equiv \sup_{f \in F} T_f v. \tag{13}$$

由引理3.1知道上述定义的三个算子都是$B \to B$映射。这里我们特别声明, 公式(13)中取极大是按照分量来取的; 一般来说,公式(13)中的最大行动不一定能够取到, 为了便于下面的叙述, 我们总假设公式(13)中最大的行动是可以取到的(需要某种紧性假设), 就象在第2章中定理2.6 后面的相应假设条件成立那样，具体的条件我们会在下一节详细给出. 我们还有下面的一些性质。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

引理3.3：关于范数$||\cdot||$, T_f, T_{π_0}和T是B中的单调压缩映射, 压缩因子为β. 具体来说, 如果算子Ł是T_f, T_{π_0}或者T之一, 则有

1) 如果$u, v \in B$而且$u \geq v$, 那么Ł$u \geq$ Łv;

2) 如果$u, v \in B$, 那么$||$Ł$u -$ Ł$v|| \leq \beta||u - v||$.

证明: 考虑算子T_f。对$\mathbf{u}, \mathbf{v} \in B$和任意的$i \in S$，考虑分量

$$\begin{aligned}[T_f\mathbf{u}](i) - [T_f\mathbf{v}](i) &= r(i,f) + \beta\sum_{j\in S} p(j|i,a)u(j) - r(i,f) - \beta\sum_{j\in S} p(j|i,a)v(j) \\ &= \beta\sum_{j\in S} p(j|i,a)(u(j) - v(j)).\end{aligned}$$

根据$\mathbf{u} \geq \mathbf{v}$，1) 对于$T_f$成立。不难看出2) 对于$T_f$也成立。类似的证明对于$T_{\pi_0}$的结论都成立。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

对于算子T，我们记

$$a^*(i) \in \arg\max_{a\in A(i)} \left\{ r(i,a) + \beta \sum_{j\in S} p(j|i,a)u(j) \right\}$$

$$b^*(i) \in \arg\max_{a\in A(i)} \left\{ r(i,a) + \beta \sum_{j\in S} p(j|i,a)v(j) \right\}$$

那么就有:

$$\begin{aligned}
[T\mathbf{u}](i) - [T\mathbf{v}](i) &\leq r(i,a^*(i)) + \beta \sum_{j\in S} p(j|i,a^*(i))u(j) \\
&\quad -r(i,a^*(i)) - \beta \sum_{j\in S} p(j|i,a^*(i))v(j) \\
&= \beta \sum_{j\in S} p(j|i,a^*(i))(u(j)-v(j)) \leq \beta||\mathbf{u}-\mathbf{v}||,
\end{aligned}$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

以及

$$\begin{aligned}[T\mathbf{u}](i) - [T\mathbf{v}](i) &\geq r(i, b^*(i)) + \beta \sum_{j \in S} p(j|i, b^*(i))u(j) \\ &\quad -r(i, b^*(i)) - \beta \sum_{j \in S} p(j|i, b^*(i))v(j) \\ &= \beta \sum_{j \in S} p(j|i, b^*(i))(u(j) - v(j)) \geq 0,\end{aligned}$$

其中用到了$\mathbf{u} \geq \mathbf{v}$。将两者结合起来就证明了定理的结论对算子$T$成立。 □

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

定理3.2：(Banach不动点定理) 如果B是Banach空间, $T : B \to B$的压缩映射. 那么:

(i) 存在唯一的$v^* \in B$满足$Tv^* = v^*$;

(ii) 对任意的$v^0 \in B$, 序列$\{v^n\}$定义为:

$$v^{n+1} = Tv^n = T^{n+1}v^0 \tag{14}$$

收敛到v^*。

证明: 考察序列$\{\mathbf{v}^n\}$，对于任意的$m \geq 1$，

$$\begin{aligned} ||\mathbf{v}^{n+m} - \mathbf{v}^n|| &\leq \sum_{k=0}^{m-1} ||\mathbf{v}^{n+k+1} - \mathbf{v}^{n+k}|| = \sum_{k=0}^{m-1} ||T^{n+k}\mathbf{v}^1 - T^{n+k}\mathbf{v}^0|| \\ &\leq \sum_{k=0}^{m-1} \beta^{n+k}||\mathbf{v}^1 - \mathbf{v}^0|| = \frac{\beta^n(1-\beta^m)}{1-\beta}||\mathbf{v}^1 - \mathbf{v}^0||. \end{aligned} \tag{15}$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

因为$\beta \in [0,1)$，所以序列$\{\mathbf{v}^n\}$为Cauchy序列并且根据B的完备性，得到序列$\{\mathbf{v}^n\}$收敛到$\mathbf{v}^* \in B$。

下面证明$\mathbf{v}^*$是不动点。因为：

$$\begin{aligned} 0 \leq & ||T\mathbf{v}^* - \mathbf{v}^*|| \leq ||T\mathbf{v}^* - \mathbf{v}^n|| + ||\mathbf{v}^n - \mathbf{v}^*|| \\ = & ||T\mathbf{v}^* - T\mathbf{v}^{n-1}|| + ||\mathbf{v}^n - \mathbf{v}^*|| \leq \beta||\mathbf{v}^* - \mathbf{v}^{n-1}|| + ||\mathbf{v}^n - \mathbf{v}^*||. \end{aligned}$$

由于$\lim_{n\to\infty} ||\mathbf{v}^n - \mathbf{v}^*|| = 0$，上式右端的两项均趋于0，当$n$充分大的时候右边可以充分的小，而中间的一项与$n$无关，所以必为0。
假设u^*也是一个T的不动点，那么考虑：

$$\begin{aligned} ||\mathbf{u}^* - \mathbf{v}^*|| = & ||T\mathbf{u}^* - T\mathbf{v}^*|| \leq \beta||\mathbf{u}^* - \mathbf{v}^*|| \\ < & ||\mathbf{u}^* - \mathbf{v}^*||. \end{aligned}$$

所以不动点必惟一。 □

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

引理3.4：假设存在$\mathbf{v} \in B$，那么：
1) 如果$\mathbf{v} \geq (>) T\mathbf{v}$，则有$\mathbf{v} \geq (>) \mathbf{V}_\beta^*$；
2) 如果$\mathbf{v} \leq (<) T\mathbf{v}$，则有$\mathbf{v} \leq (<) \mathbf{V}_\beta^*$；
3) 如果$\mathbf{v} = T\mathbf{v}$，那么$\mathbf{v}$就是B中惟一具有性质$\mathbf{v} = \mathbf{V}_\beta^*$的向量。

证明: 首先证明1)。任取策略$\pi = (f_0, f_1, \cdots) \in \Pi_m^d$。根据$T$的定义和上确界是按照分量分别取的，就有：

$$\mathbf{v} \geq \sup_{f \in F} \{\mathbf{r}(f) + \beta \mathbf{P}(f)\mathbf{v}\} = \sup_{\pi_0} \{\mathbf{r}(\pi_0) + \beta \mathbf{P}(\pi_0)\mathbf{v}\},$$

其中π_0的每个分量$\pi_0(\cdot|i)$是行动集合$A(i)$上的随机决策规则，对所有的$i \in S$。所以

$$\begin{aligned} \mathbf{v} &\geq \mathbf{r}(f_0) + \beta \mathbf{P}(f_0)\mathbf{v} \geq \mathbf{r}(f_0) + \beta \mathbf{P}(f_0)(\mathbf{r}(f_1) + \beta \mathbf{P}(f_1)\mathbf{v}) \\ &= \mathbf{r}(f_0) + \beta \mathbf{P}(f_0)\mathbf{r}(f_1) + \beta^2 \mathbf{P}(f_0)\mathbf{P}(f_1)\mathbf{v}. \end{aligned}$$

由归纳可以得到，对任意的$n \geq 1$，

$$\begin{aligned} \mathbf{v} &\geq \mathbf{r}(f_0) + \beta \mathbf{P}(f_0)\mathbf{r}(f_1) + \cdots + \beta^n \mathbf{P}(f_0) \cdots \mathbf{P}(f_{n-1})\mathbf{r}(f_n) + \beta^{n+1} \mathbf{P}^{n+1}(\pi)\mathbf{v} \\ &= \mathbf{r}(f_0) + \beta \mathbf{P}(f_0)\mathbf{r}(f_1) + \cdots + \beta^n \mathbf{P}^n(\pi)\mathbf{r}(f_n) + \beta^{n+1} \mathbf{P}^{n+1}(\pi)\mathbf{v}. \end{aligned}$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

因此，

$$\mathbf{v} - \mathbf{V}_\beta(\pi) \geq \beta^{n+1}\mathbf{P}^{n+1}(\pi)\mathbf{v} - \sum_{t=n+1}^{\infty} \beta^t \mathbf{P}^t(\pi)\mathbf{r}(f_t). \tag{16}$$

取$\epsilon > 0$，由于$||\beta^{n+1}\mathbf{P}^{n+1}(\pi)\mathbf{v}|| \leq \beta^{n+1}||\mathbf{v}||$而且$\beta \in [0,1)$，当$n$充分大的时候有：

$$-\frac{\epsilon}{2}\mathbf{1} \leq \beta^n \mathbf{P}^n(\pi)\mathbf{v} \leq \frac{\epsilon}{2}\mathbf{1},$$

其中$\mathbf{1}$表示分量都是1的列向量。由于报酬函数是有界的且$|r(i,a)| \leq M < \infty$，则有：

$$-\frac{\beta^n M}{1-\beta}\mathbf{1} \leq \sum_{t=n}^{\infty} \beta^t \mathbf{P}^t(\pi)\mathbf{r}(f_t),$$

故当n充分大的时候，(16)右端的第二项也可以被$\pm(\epsilon/2)\mathbf{1}$界住。可是(16)的左端与$n$无关，所以对所有状态$i \in S$和$\epsilon > 0$，总有：

$$v(i) \geq V_\beta(i,\pi) - \epsilon.$$

由ϵ的任意性，可以得到1)。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

2)的证明：因为$\mathbf{v} \leq T\mathbf{v}$，对任意的$\epsilon > 0$存在一个确定性决策规则$f$满足：

$$\mathbf{v} \leq \mathbf{r}(f) + \beta \mathbf{P}(f)\mathbf{v} + \epsilon \mathbf{1}.$$

类似1)中的推理，有：

$$\mathbf{v} \leq \mathbf{r}(f) + \epsilon \mathbf{1} + \beta \mathbf{P}(f)(\mathbf{r}(f) + \epsilon \mathbf{1} + \beta \mathbf{P}(f)\mathbf{v}).$$

经过归纳有：

$$\mathbf{v} \leq \sum_{k=0}^{n-1} \beta^k \mathbf{P}^k(f)(\mathbf{r}(f) + \epsilon \mathbf{1}) + \beta^n \mathbf{P}^n(f)\mathbf{v}.$$

当$n \to \infty$，并利用引理，得到：

$$\mathbf{v} \leq (\mathbf{I} - \beta \mathbf{P}(f))^{-1}(\mathbf{r}(f) + \epsilon \mathbf{1}) = \mathbf{V}_\beta(f^\infty) + \frac{\epsilon}{1-\beta}\mathbf{1}. \qquad (17)$$

因此

$$\mathbf{v} \leq \sup_{\pi \in \Pi} \mathbf{V}_\beta(\pi) + \frac{\epsilon}{1-\beta}\mathbf{1}.$$

再由ϵ的任意性得到2)。关于1)和2)的严格不等情况的证明是类似的，请读者自行证明。结合1)和2)可以得到3)。 □

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

引理3.5：假设存在$\mathbf{v} \in B$和$f \in F$，那么：
1) 如果$\mathbf{v} \geq (>) T_f \mathbf{v}$，则有$\mathbf{v} \geq (>) \mathbf{V}_\beta(f^\infty)$；
2) 如果$\mathbf{v} \leq (<) T_f \mathbf{v}$，则有$\mathbf{v} \leq (<) \mathbf{V}_\beta(f^\infty)$；
3) 如果$\mathbf{v} = T_f \mathbf{v}$，那么$\mathbf{v}$就是$B$中惟一具有性质$\mathbf{v} = \mathbf{V}_\beta(f^\infty)$的向量。
类似的，将上述的平稳策略$f \in F$替换为随机平稳策略$\pi_0^\infty \in \Pi_s$，上面的三条结论依然成立。

证明: 类似引理3.4的证明。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

我们可以证明下面的定理.

定理3.3：(i) 存在$v^* \in B$满足

$$Tv^* = v^* \tag{18}$$

其中T为(13)式所定义. 这个v^*在B里唯一, 而且有$v^* = V_\beta^*$. 等式(18)也称为**折扣模型的最优方程**.
(ii) 对每个f^∞, 存在唯一的v满足$T_f v = v$, 而且有$v = V_\beta(f^\infty)$.

定理3.3是十分有意义的，一方面它保证了最优方程解的存在性，从而可以利用线性方程组实现最优值与最优策略的求解过程。另一方面，就算子T_f而言，不动点恰为平稳策略f^∞的报酬值函数$\mathbf{V}_\beta(f^\infty)$。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

3 最优策略的存在性

我们希望把寻找最优策略的范围限制得越小越好, 比如说最好限制到平稳策略类中. 但是当我们限制寻优的范围小到平稳策略类中时, 能否保证得到的最好策略仍然在最一般的策略类中是最优的呢? 这一节中, 我们将表明这样一个事实: 使公式(13)在$v = v^*$时达到上确界的决策规则的存在性蕴涵了最优平稳策略的存在性。

定理3.4：策略$\pi^* \in \Pi$是最优的充要条件是$V_\beta(\pi^*)$是最优方程(18)的解.

证明: 假设$\pi^* \in \Pi$是最优的，那么有$\mathbf{V}_\beta(\pi^*) = \mathbf{V}_\beta^*$。根据定理3.3的(i)，$\mathbf{V}_\beta(\pi^*)$满足方程：$T\mathbf{v} = \mathbf{v}$。
假设$T\mathbf{V}_\beta(\pi^*) = \mathbf{V}_\beta(\pi^*)$，根据定理3.2知道它是$T\mathbf{v} = \mathbf{v}$的惟一解，即有：$\mathbf{V}_\beta(\pi^*) = \mathbf{V}_\beta^*$，说明$\pi^*$是最优的。 □

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

对于$v \in B$, 我们称决策规则$f_v \in F$为v-**改进规则**, 如果

$$f_v \in \arg\max_{f \in F} \{r(f) + \beta P(f)v\}, \tag{19}$$

而且$Tv \geq v$. 或者等价的说

$$r(f_v) + \beta P(f_v)v = \max_{f \in F} \{r(f) + \beta P(f)v\} \geq v \quad \text{或者} \quad T_{f_v}v = Tv \geq v. \tag{20}$$

这里我们用改进规则的意思是平稳策略f_v的期望折扣报酬至少不比v值小, 即总有$V_\beta(f_v^\infty) \geq v$. 只有当存在某个分量$i \in S$满足$r(i, f_v(i)) + \beta P(f_v)v(i) > v(i)$时, 改进才是严格的, 即$V_\beta(f_v^\infty) > v$.

对于v^*的改进规则有着特殊的意义, 我们称为这样的规则为**保持**的. 即决策规则f^*是保持的, 其充要条件为

$$T_{f^*}v^* = r(f^*) + \beta P(f^*)v^* = v^*. \tag{21}$$

或者等价的说

$$f^* \in \arg\max_{f \in F} \{r(f) + \beta P(f)v^*\}. \tag{22}$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

下面的定理表明由保持的决策规则形成的平稳策略是最优策略, 这也是本章中最为重要的结果. 它为MDP理论提供了一个主要理论结果: 评价平稳策略是否为最优策略的方法.

定理3.5：设状态空间S是离散的且对一切$v \in B$, 公式(13)右端的极大可以取到. 那么

a) 存在保持决策规则$f^* \in F$;

b) 如果f^*是保持的, 则平稳策略$f^{*\infty}$(简记为f^*)是最优的;

c) $v^* = V_\beta^* = \sup_{f \in F} V_\beta(f^\infty)$.

证明: 因为最优值向量$\mathbf{V}_\beta^* \in B$，而且公式(13)右端的极大可以取到，所以a)显然。再根据定理3.3和引理3.4的3)，$\mathbf{V}_\beta^* = \mathbf{v}^*$是$T\mathbf{v}^* = \mathbf{v}^*$的惟一解。因此，由公式(21)知道：

$$\mathbf{V}_\beta^* = T\mathbf{V}_\beta^* = \mathbf{r}(f^*) + \beta \mathbf{P}(f^*)\mathbf{V}_\beta^* = T_{f^*}\mathbf{V}_\beta^*,$$

于是再根据公式(9)知道：

$$\mathbf{V}_\beta(f^{*\infty}) = \mathbf{V}_\beta^*,$$

就得到了b)。而c)可直接由b)导出。 □

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

下面我们给出平稳策略是最优的条件(注意, 下面的定理对一般的状态空间也成立).

定理3.6：假设条件: a) 存在保持决策规则, 或者b) 存在最优策略. 那么存在平稳策略是最优的.

证明: 如果a)成立，由定理3.5的b)部分结论直接可得。下面我们证明情形b)。假设$\pi^* = (\pi_0, \pi_1, \cdots) \in \Pi$是最优策略。则，

$$\mathbf{V}_\beta(\pi^*) = \mathbf{r}(\pi_0) + \beta \mathbf{P}(\pi_0)\mathbf{V}_\beta(\pi^*|h_1), \tag{23}$$

其中符号$(\pi^*|h_1)$表示了一族与策略$\pi^* = (\pi_0, \pi_1, \cdots)$有关的策略，具体是：根据过程在阶段1 的历史$h_1$不同，最优策略$\pi^*$在阶段1对应的决策规则也可能不同。但是在同一个到阶段1的历史下，将来每一步的决策规则是同于给定到阶段1的历史下的策略$\pi^* = (\pi_0, \pi_1, \cdots)$，我们可以用一个策略来表示，它就是这一族策略$(\pi|h_1)$中的一个。由于策略$\pi^*$是最优的，所以它不劣于这一族策略中的每一个。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

故由公式(23)，有，

$$\begin{aligned}\mathbf{V}_\beta(\pi^*) &= \mathbf{r}(\pi_0)+\beta\mathbf{P}(\pi_0)\mathbf{V}_\beta(\pi^*|h_1)\le \mathbf{r}(\pi_0)+\beta\mathbf{P}(\pi_0)\mathbf{V}_\beta(\pi^*)\\ &= T_{\pi_0}\mathbf{V}_\beta(\pi^*)\le T\mathbf{V}_\beta(\pi^*)\\ &= \sup_{f\in F}\{\mathbf{r}(f)+\beta\mathbf{P}(f)\mathbf{V}_\beta(\pi^*)\}=\mathbf{V}_\beta(\pi^*),\end{aligned}$$

最后的等式利用了最优策略的值函数是算子T的不动点。下面需要证明上面倒数第二个等式的上确界可以达到。根据

$$T_{\pi_0}\mathbf{V}_\beta(\pi^*)=\sup_{f\in F}\{\mathbf{r}(f)+\beta\mathbf{P}(f)\mathbf{V}_\beta(\pi^*)\}=\mathbf{V}_\beta(\pi^*)$$

和所有行动集合的有限性，知道上面倒数第二个等式的上确界可以达到，即存在确定的决策规则f是保持的，也即a)成立，由a)知道b)成立。 □

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

为了便于应用定理3.5, 我们给出公式(13)右端取到极大的条件, 以保证平稳最优策略的存在性.

定理3.7：假设状态空间S是离散的,而且下面条件之一成立:

a) $A(i)$对每个状态$i \in S$均有限;

b) 对每个$i \in S$, $A(i)$紧致, $r(i,a)$关于$a \in A(i)$连续而且对$i,j \in S$, $p(j|i,a)$关于$a \in A(i)$连续;

c) 对每个$i \in S$, $A(i)$紧致, $r(i,a)$关于$a \in A(i)$上半连续而且对$i,j \in S$, $p(j|i,a)$关于$a \in A(i)$下半连续.

那么存在最优的平稳策略.

当最优策略不存在时, 我们可以寻找ϵ-最优策略. 而ϵ-最优策略的存在条件则相对较弱.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

定理3.8：假设状态空间S是有限的或者可数的，则对任意的$\epsilon > 0$，存在ϵ-最优的确定性平稳策略.

证明：由定理3.3知道存在$\mathbf{V}_\beta^* \in B$,满足$T\mathbf{V}_\beta^* = V_\beta^*$，其中$\beta$为折扣因子.对$\epsilon > 0$,取$f_\epsilon \in F$满足

$$\begin{aligned} \mathbf{r}(f_\epsilon) + \beta \mathbf{P}(f_\epsilon)\mathbf{V}_\beta^* &\geq \sup_{f \in F}\{\mathbf{r}(f) + \beta \mathbf{P}(f)\mathbf{V}_\beta^*\} - (1-\beta)\epsilon \mathbf{1} \\ &= \mathbf{V}_\beta^* - (1-\beta)\epsilon \mathbf{1}. \end{aligned}$$

两端用算子T_{f_ϵ}作用，并利用保序性质（引理3.3），有：

$$\begin{aligned} T_{f_\epsilon}^2 \mathbf{V}_\beta^* &\geq T_{f_\epsilon}[\mathbf{V}_\beta^* - (1-\beta)\epsilon \mathbf{1}] \\ &\geq \mathbf{V}_\beta^* - (1-\beta)\epsilon \mathbf{1} - \beta(1-\beta)\epsilon \mathbf{1}. \end{aligned}$$

重复上面的做法，可以得到：

$$\mathbf{V}_\beta(f_\epsilon) \geq \mathbf{V}_\beta^* - \epsilon \mathbf{1},$$

即结论得证。 □

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

例3.1：考虑MDP问题. 状态空间$S = \{1, 2\}$. 在状态1的可用行动集$A(1) = \{a_1, a_2\}$，而在状态2的行动集为独点集$A(2) = \{a_3\}$。相应的报酬函数和转移概率分别为：

表 3.1　转移概率和报酬

状态 (i)	可用行动 (a)	转移概率 $p(j\|i,a)$		报酬 $r(i,a)$
		$j=1$	$j=2$	
1	a_1	0.5	0.5	5
1	a_2	0	1	10
2	a_3	0	1	-1

由于状态空间和行动集都是有限的, 我们知道存在最优的平稳策略. 这时只有两个平稳策略, 我们记

$$f = \begin{pmatrix} a_1 \\ a_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad g = \begin{pmatrix} a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

为这两个平稳策略. 下面, 我们通过解最优方程寻求最优策略.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

此时,最优方程是:

$$\begin{aligned} v(1) &= \max\{5+0.5\beta v(1)+0.5\beta v(2), 10+\beta v(2)\}, \\ v(2) &= -1+\beta v(2). \end{aligned}$$

直接解上面的第二个方程, 有$v(2) = -1/(1-\beta)$, 将其带入第一个方程, 有

$$v(1) = \max\left\{5 - 0.5\frac{\beta}{1-\beta} + 0.5\beta v(1), 10 - \frac{\beta}{1-\beta}\right\} \equiv Tv(1). \qquad (24)$$

对$\beta = 0,\ 0.5$和0.9分别解(24), 得到最优值函数分别为:

$$V_0^* = \begin{pmatrix} 10 \\ -1 \end{pmatrix}, \quad V_{0.5}^* = \begin{pmatrix} 9 \\ -2 \end{pmatrix}, \quad V_{0.9}^* = \begin{pmatrix} 1 \\ -10 \end{pmatrix}.$$

将$\beta = 0.9$和0分别带入(24), 得到对应的最优平稳策略分别为g和g. 也可以将最优策略看成是β的函数. 当$V_\beta(f) = V_\beta(g)$时, 必有$\beta^2 - \frac{21}{11}\beta + \frac{10}{11} = 0$. 解得后知道当$\beta \le 10/11$时$g$最优, $\beta \ge 10/11$时f最优.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

在实际建模时, 条件往往不是这样理想, 所以这里有必要提及一般状态空间和行动集的情形. 在这种情况下, 即使公式(13)中的上确界可以取到, 仍然不能保证下面的结论成立: 1)存在可测的ϵ-最优策略; 2)对于$v \in B_M$, Tv不一定还在B_M中; 3)MDP的值V_β^*的可测性. 因此, 我们这里列出相应的结果.

设P为S上的概率测度, $\epsilon > 0$. 我们称策略$\pi \in \Pi$为(P, ϵ)-最优的, 如果

$$P\{i : V_\beta(i, \pi) > V_\beta^*(i) - \epsilon\} = 1. \tag{25}$$

如果取P为在状态i的退化分布, 可以知道(P, ϵ)-最优是ϵ-最优的推广.

定理3.9: 设状态空间S是Polish空间, P为S的Borel子集上的概率测度,$\epsilon > 0$. 则

(a)存在(P, ϵ)-最优(Borel可测)平稳策略;

(b)如果每个$A(i)$可数, 存在ϵ-最优平稳策略;

(c)如果每个$A(i)$有限, 存在最优平稳策略;

(d)如果(i)每个$A(i)$紧致度量空间, (ii)对一切$i \in S$, $r(i, a)$为$A(i)$上的上半连续有界函数, (iii) 对S中每个Borel子集C和状态i, $p(C|i, a)$关于行动$a \in A(i)$连续, 那么存在最优平稳策略.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

4 策略迭代算法

策略迭代(Policy Iteration)算法也称为策略空间逼近法, 它是求解折扣MDP的一个有效方法. 特别是对于状态空间和行动空间有限的MDP问题, 方程(10)中确定$V_\beta(f^\infty)$的值可以通过求解下面的线性方程组得到

$$V_\beta(f^\infty) = (I - \beta P(f))^{-1} r(f). \tag{26}$$

算法3.1 (策略迭代算法)

步骤1: 令$n = 0$且任取$f_0 \in F$.

步骤2: (策略求值过程)解方程(26)或者解

$$(I - \beta P(f_n))v = r(f_n) \tag{27}$$

得到$V_\beta(f_n^\infty)$。

步骤3: (策略改进过程)选取f_{n+1}为一个$V_\beta(f_n^\infty)$-改进规则, 即满足

$$f_{n+1} \in \arg\max_{f \in F} \{r(f) + \beta P(f) V_\beta(f_n^\infty)\}, \tag{28}$$

如有可能, 令$f_{n+1} = f_n$.

步骤4: 如果$f_{n+1} = f_n$, 停止. 这时$f_{n+1} = f_n$为最优策略, $V_\beta(f_{n+1}^\infty) = V_\beta(f_n^\infty)$为最优值函数. 否则令$n = n + 1$,返回到步骤2.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

利用算法3.1可以产生一串平稳策略序列$\{f_n\}$和相应的值函数序列$\{V_\beta(f_n^\infty)\}$. 如果算法中步骤4的停止规则被满足, 这两个序列是有限的序列, 否则就是无限的序列. 下面我们就状态空间和行动空间均为有限的MDP问题和一般的问题给出算法收敛的情况。

定理3.10：设状态空间S是有限的且对每个$i \in S$,$A(i)$也是有限的. 那么策略迭代算法在有限次迭代后必然终止于最优方程的解和最优平稳策略.

证明: 由于f_{n+1}是一个$\mathbf{V}_\beta(f_n^\infty)$-改进规则，利用引理3.5我们有

$$\mathbf{V}_\beta(f_{n+1}^\infty) > \mathbf{V}_\beta(f_n^\infty). \tag{29}$$

因此算法产生的平稳策略不可能循环，再根据F的有限性，定理得证。 □

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

如果平稳策略类F不是有限的, 例如S有限但是$A(i), i \in S$为紧致的或者S是可数集合时, 定理3.10 的证明不再有效. 这是因为无法保证步骤4的停止. 这时算法的收敛问题需要另行考虑. 下面, 一方面我们将步骤3一般化, 即从任一个$v \in B$开始步骤3的计算会有什么结果; 另一方面, 我们将讨论算法的收敛速度如何.

我们用递归方式表示策略迭代过程. 定义算子$L : B \mapsto B$为:

$$\begin{aligned} Lv &\equiv \max_{f \in F}\{r(f) + (\beta P(f) - I)v\} \\ &= Tv - v. \end{aligned} \tag{30}$$

最优方程(18)可以重新表示为

$$Lv = 0. \tag{31}$$

利用这个符号, 求解最优方程, 即求算子T的不动点, 转化为求解算子L的0点. 策略迭代方法等价于用向量空间的牛顿法寻找L的0点.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

对于$v \in B$, 记F_v为所有的v-改进策略全体. 如果

$$f_v \in \arg\max_{f \in F}\{r(f) + (\beta P(f) - I)v\}, \tag{32}$$

那么$f_v \in F_v$. 注意公式(32)中的单位矩阵I并不影响最大行动的选取。

性质3.1：对于$\mathbf{u}, \mathbf{v} \in B$以及$f_{\mathbf{v}} \in F_{\mathbf{v}}$，我们有

$$L\mathbf{u} \geq L\mathbf{v} + (\beta \mathbf{P}(f_{\mathbf{v}}) - \mathbf{I})(\mathbf{u} - \mathbf{v}). \tag{33}$$

证明: 根据L的定义和$f_{\mathbf{v}} \in F_{\mathbf{v}}$，有：

$$\begin{aligned} L\mathbf{u} &\geq T_{f_{\mathbf{v}}}\mathbf{u} - \mathbf{u} = \mathbf{r}(f_{\mathbf{v}}) + \beta \mathbf{P}(f_{\mathbf{v}})\mathbf{u} - \mathbf{u} \\ &= \mathbf{r}(f_{\mathbf{v}}) + \beta \mathbf{P}(f_{\mathbf{v}})\mathbf{v} - \beta \mathbf{P}(f_{\mathbf{v}})\mathbf{v} + \beta \mathbf{P}(f_{\mathbf{v}})\mathbf{u} - \mathbf{u} \\ &= T_{f_{\mathbf{v}}}\mathbf{v} - \mathbf{v} + \mathbf{v} - \beta \mathbf{P}(f_{\mathbf{v}})\mathbf{v} + \beta \mathbf{P}(f_{\mathbf{v}})\mathbf{u} - \mathbf{u} \\ &= T\mathbf{v} - \mathbf{v} + \beta \mathbf{P}(f_{\mathbf{v}})(\mathbf{u} - \mathbf{v}) - \mathbf{I}(\mathbf{u} - \mathbf{v}) \\ &= L\mathbf{v} + (\beta \mathbf{P}(f_{\mathbf{v}}) - \mathbf{I})(\mathbf{u} - \mathbf{v}), \end{aligned}$$

结论得证。 □

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

性质3.2：假定$\{\mathbf{V}^n\}$是通过策略迭代得到的序列,对任意的$f_{\mathbf{V}^n} \in F_{\mathbf{V}^n}$,有:

$$\mathbf{V}^{n+1} = \mathbf{V}^n - (\beta\mathbf{P}(f_{\mathbf{V}^n}) - \mathbf{I})^{-1}L\mathbf{V}^n. \tag{34}$$

证明: 由F_{V^n}和V^{n+1}的定义,

$$\begin{aligned}
\mathbf{V}^{n+1} &= (\mathbf{I} - \beta\mathbf{P}(f_{\mathbf{V}^n}))^{-1}\mathbf{r}(f_{\mathbf{V}^n}) - \mathbf{V}^n + \mathbf{V}^n \\
&= (\mathbf{I} - \beta\mathbf{P}(f_{\mathbf{V}^n}))^{-1}\left[\mathbf{r}(f_{\mathbf{V}^n}) + (\beta\mathbf{P}(f_{\mathbf{V}^n}) - \mathbf{I})\mathbf{V}^n\right] + \mathbf{V}^n \\
&= \mathbf{V}^n - (\beta\mathbf{P}(f_{\mathbf{V}^n}) - \mathbf{I})^{-1}L\mathbf{V}^n.
\end{aligned}$$

□

比较一维的牛顿算法$x_{n+1} = x_n - [f'(x_n)]^{-1}f(x_n)$, 不难看出为什么说策略迭代法等价于向量空间的牛顿法.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

定理3.11：由策略迭代算法得到的序列$\{V^n\}$在$||\bullet||$意义下单调收敛到V_β^*.

推论3.1：设在策略迭代算法的步骤3中, 以任意$V^0 \in B$为初始点, 迭代得到序列$\{V^n\}$, 满足定理3.11 的结论.

定理3.12：设由策略迭代算法得到的序列$\{V^n\}$, 且对所有n,$f_{V^n} \in F_{V^n}$, 以及存在K满足$0 < K < \infty$和

$$||P(f_{V^n}) - P(f_{V_\beta^*})|| \leq K||V^n - V_\beta^*||, \tag{35}$$

对于$n = 1, 2, \cdots$. 则

$$||V^{n+1} - V_\beta^*|| \leq \frac{K\beta}{1-\beta}||V^n - V_\beta^*||^2. \tag{36}$$

策略迭代算法的最大缺点就是每一步都需要求解方程组27. 如果状态空间比较大或者是无限的, 需要很大的计算量或者其他的逼近方法. 这一点也制约了策略迭代算法的使用范围. 下面一节我们介绍一个不需要求解方程组(27)的算法. 在此之前, 我们先给一个数例, 演示策略迭代的过程.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

例3.2：状态空间$S = \{1, 2\}$.在状态1和2的可用行动集$A(1) = A(2) = \{1, 2\}$。折扣因子$\beta = 0.9$. 相应的报酬函数和转移概率分别为：

表 3.2　转移概率和报酬

状态 i	可用行动	转移概率 $p(j\|i,a)$		报酬
	a	$j=1$	$j=2$	$r(i,a)$
1	1	0.5	0.5	6
	2	0.8	0.2	4
2	1	0.4	0.6	-3
	2	0.7	0.3	-5

求最优策略与最优值函数(保留两位小数).

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

解:共有4个平稳策略, 分别记为:

$$f_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad f_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad f_3 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad f_4 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}.$$

下面用策略迭代算法求解.

第一个迭代过程:

1) 取$f_1 \in F$为初始策略作求值运算, 即解线性方程组

$$\begin{cases} 6 + 0.9[0.5v(1) + 0.5v(2)] = v(1), \\ -3 + 0.9[0.4v(1) + 0.6v(2)] = v(2). \end{cases}$$

解得$v(1) = V_{0.9}(1, f_1^\infty) \approx 15.49$,$v(2) = V_{0.9}(2, f_1^\infty) \approx 5.60$.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

2) 策略改进运算.对状态1

$$\max_{a\in A(1)}\{r(1,a)+0.9\sum_{j\in S}p(j|1,a)V_{0.9}(j,f_1^\infty)\}=\max\{15.49,\ 16.16\}=16.16.$$

因此, 在状态1, 行动2达到上式左端的最大值, 故改进的策略在状态1应该采用行动2. 同样对状态2进行策略改进,

$$\max_{a\in A(2)}\{r(2,a)+0.9\sum_{j\in S}p(j|2,a)V_{0.9}(j,f_1^\infty)\}=\max\{5.60,\ 6.27\}=6.27.$$

说明在状态2, 行动2达到上式左端的最大值, 故改进的策略在状态2也应该采用行动2. 这样得到了改进的策略为f_4.

第二个迭代过程: 对f_4重复第一个迭代过程, 解得方程$v(1)=V_{0.9}(1,f_4^\infty)\approx 22.20$, $v(2)=V_{0.9}(2,f_4^\infty)\approx 12.31$. 并且发现在策略改进时满足停止规则. 这样最优策略就是f_4, 最优值函数为向量$(22.20, 12.31)$.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

5 值迭代算法

值迭代(Value Iteration)算法是求解折扣MDP问题的最为广泛的方法. 如果用**逐次逼近**(Successive Approximations), **上松弛**(Over-relaxation), **向后递归**(Backward Induction), **予雅可比迭代**(Pre-Jacobi Iteration)或者动态规划的名字, 读者可能会更熟悉. 总之, 值迭代算法是一个极为简便易行的数值算法.

这一节我们假设对任意的$v \in B$, 公式(13)右端的极大值总能取到, 例如对每个状态$i \in S$, $A(i)$是有限集合. 当然在讨论算法收敛时并不需要这个假设. 为了简便, 我们把式(13)写成分量的形式:

$$v(i) = \max_{a \in A(i)} \left\{ r(i,a) + \beta \sum_{j \in S} p(j|i,a) v(j) \right\}. \tag{37}$$

下面是具体寻求ϵ-最优平稳策略及其逼近值的值迭代算法

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

算法3.2 (值迭代算法)

步骤1: 任取$v^0 \in B$, 给定$\epsilon > 0$并且置$n = 0$.

步骤2: (对每个$i \in S$, 通过计算

$$v^{n+1}(i) = \max_{a \in A(i)} \left\{ r(i,a) + \beta \sum_{j \in S} p(j|i,a) v^n(j) \right\} \tag{38}$$

得到$v^{n+1}(i)$。

步骤3: 如果

$$||v^{n+1} - v^n|| < \frac{\epsilon(1-\beta)}{2\beta}, \tag{39}$$

进入步骤4. 否则将n增加1, 返回步骤2.

步骤4: 对每个$i \in S$, 取

$$f_\epsilon(i) \in \arg\max_{a \in A(i)} \left\{ r(i,a) + \beta \sum_{j \in S} p(j|i,a) v^{n+1}(j) \right\}, \tag{40}$$

算法停止.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

式(38)和(40)我们也记为

$$v^{n+1} = Tv^n, \tag{41}$$

$$f_\epsilon \in \arg\max_{f\in F}\{r(f) + \beta P(f)v^{n+1}\}. \tag{42}$$

下面的定理叙述了值迭代算法收敛的主要结果。

定理3.13：令$v^0 \in B, \epsilon > 0$以及对于$n > 0, \{v^n\}$满足式(41). 则
(1) v^n依模$||\bullet||$收敛到V_β^*,
(2) 存在有限的一个整数N, 使得当$n \geq N$时, 式(39)总成立,
(3) 式(40)定义的平稳策略f_ϵ^∞是ϵ-最优的,
(4) 只要式(39)成立, 则有$||v^{n+1} - V_\beta^*|| < \epsilon/2$.
证明思路: (1)和(2)的证明可以直接从定理3.2得到. 其余部分的证明请参见Puterman中定理6.3.1的证明.

由引理3.2, 不难知道如果$Tv^0 \leq (\geq)v^0$, 则序列$\{v^n\}$收敛到V_β^*是单调降(增)的.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

为了方便的估计值迭代算法的效果，关收敛速度的结果我们总结如下。

定理3.14：设$v^0 \in B$而且$\{v^n\}$是通过值迭代算法得到的序列. 对于值迭代算法的整体收敛情况，我们有

(1) 值迭代算法的一步收敛速度(**压缩率**)是β,

(2) 值迭代算法的渐进平均收敛的压缩率是β,

(3) 值迭代算法整体收敛的阶为$O(\beta^n)$,

(4) 对一切的n,

$$||v^n - V_\beta^*|| \leq \frac{\beta^n}{1-\beta}||v^1 - v^0||, \tag{43}$$

(5) 对一切的$g_n \in \arg\max_{f\in F}\{r(f) + \beta P(f)v^n\}$,

$$||V_\beta(g_n^\infty) - V_\beta^*|| \leq \frac{2\beta^n}{1-\beta}||v^1 - v^0||. \tag{44}$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

从定理3.14可以想象的到, 如果β接近于1时, 收敛的速度会受到很大的影响. 为此, 人们提出了**分解值迭代算法**以增加收敛效果. 手段是对固定的策略$f \in F$,将$I - \beta P(f)$分解为两个部分

$$I - \beta P(f) = Q(f) - R(f). \tag{45}$$

如果一个矩阵A的分量都是非负的, 我们记为$A \geq 0$; 如果矩阵$A-B \geq 0$, 我们记做$A \geq B$. 如果(45)式中的的分解满足$Q^{-1}(f) \geq 0$且$R(f) \geq 0$, 我们称分解$(Q(f), R(f))$为$I - \beta P(f)$ 的**一个正规分解**(Regular Splitting). 最简单的正规分解就是$Q(f) = I$,$R(f) = \beta P(f)$.

下面我们考虑一种改进的值迭代算法—**高斯塞德尔(Gauss-Seidel)值迭代算法**. 尽管这个算法被称为高斯塞德尔值迭代算法, 但是它即不为高斯所知也不是由塞德尔提出来的. 之所以这样叫可能是线性系统的迭代方法最早可以追溯到高斯. Hastings和Kushner等分别独立的提出了这种算法. 为了方便, 我们把状态空间中的所有状态标记为$i_1, i_2, \cdots, i_N$, 并且规定指标集是空集的求和是0.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

算法3.3 (高斯塞德尔值迭代算法)

步骤1: 任取$v^0 \in B$, 给定$\epsilon > 0$并且置$n = 0$.

步骤2: 置$k = 1$, 执行2(a).

2(a). 计算

$$v^{n+1}(i_k) = \max_{a \in A(i_k)} \left\{ r(i_k, a) + \beta \left[\sum_{j<k} p(i_j|i_k, a) v^{n+1}(i_j) + \sum_{j \geq k} p(i_j|i_k, a) v^n(i_j) \right] \right\}$$

2(b). 如果$k = N$, 进入步骤3. 否则令k增加1后回到2(a)。

步骤3: 如果

$$||v^{n+1} - v^n|| < \frac{\epsilon(1-\beta)}{2\beta}, \tag{46}$$

进入步骤4. 否则将n增加1, 返回步骤2.

步骤4: 对每个$i \in S$,取

$$f_\epsilon(i) \in \arg\max_{a \in A(i)} \left\{ r(i, a) + \beta \sum_{j \in S} p(j|i, a) v^{n+1}(j) \right\}, \tag{47}$$

算法停止.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

我们现在介绍用正规分解的方法计算算法3.3的步骤2. 记f为使步骤2式达到最大的决策规则, $P(f) = P^L(f) + P^U(f)$, 其中

$$P^L(f) = \begin{bmatrix} 0 & 0 & \cdot & \cdot & \cdot & 0 \\ p_{21} & 0 & \cdot & & & 0 \\ p_{31} & p_{32} & 0 & & & 0 \\ \cdot & & & \cdot & & \\ \cdot & & & & \cdot & \\ p_{N1} & & & & p_{N,N-1} & 0 \end{bmatrix}; \quad P^U(f) = \begin{bmatrix} p_{11} & p_{12} & \cdot & \cdot & \cdot & p_{1N} \\ 0 & p_{22} & \cdot & \cdot & \cdot & p_{2N} \\ 0 & 0 & p_{33} & & & p_{3N} \\ \cdot & & & \cdot & & \cdot \\ \cdot & & & & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 0 & p_{NN} \end{bmatrix}.$$

按照算法3.3的步骤2, 高斯塞德尔值迭代算法可以写为

$$v^{n+1} = (I - \beta P^L(f))^{-1}(\beta P^U(f))v^n + (I - \beta P^L(f))^{-1}r(f). \qquad (48)$$

令$Q(f) = (I - \beta P^L(f))$及$R(f) = \beta P^U(f)$, 易见$(Q(f), R(f))$是$I - \beta P(f)$的一个正规分解. 式(48)可以重写为

$$v^{n+1} = Q^{-1}(f)R(f)v^n + Q^{-1}(f)r(f). \qquad (49)$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

下面我们给出基于正规分解的算法收敛定理.

定理3.15：设对任意$f \in F$,$(Q(f), R(f))$是$I - \beta P(f)$的一个正规分解并且

$$\alpha \equiv \sup_{f \in F} ||Q^{-1}(f)R(f)|| < 1. \tag{50}$$

则我们有

(1) 对所有$v^0 \in B$,迭代

$$v^{n+1} = \max_{f \in F}\{Q^{-1}(f)r(f) + Q^{-1}(f)R(f)v^n\} \equiv Tv^n \tag{51}$$

收敛到V_β^*.

(2) V_β^*是T的唯一不动点.

(3) 由(51)式定义的序列$\{v^n\}$一步收敛的压缩率小于等于α,其渐进平均收敛的压缩率也小于等于α,进而其整体收敛的阶为$O(\lambda^n)$,其中$\lambda \leq \alpha$.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

为了比较不同的正规分解对算法的影响, 我们需要下面的性质.

性质3.3：假定P是转移概率矩阵并且(Q_1, R_1)和(Q_2, R_2)是$I - \beta P$的两个正规分解,这里$0 \leq \beta < 1$.那么有:

(1) 如果$R_2 \leq R_1 \leq \beta P$,则

$$||Q_2^{-1}R_2|| \leq ||Q_1^{-1}R_1||. \tag{52}$$

(2) 如果还有$R_1 - R_2 \neq 0$,则

$$\sigma(Q_2^{-1}R_2) < \sigma(Q_1^{-1}R_1), \tag{53}$$

其中$\sigma(A)$表示矩阵A的谱半径。

下面就高斯塞德尔值迭代算法, 我们将其收敛情况叙述为一个定理.

定理3.16：对所有$v^0 \in B$, 由高斯塞德尔值迭代算法得到的序列(记为$\{v_{GS}^n\}$)收敛于V_β^*. 而且, 序列$\{v_{GS}^n\}$一步收敛的压缩率小于等于β, 其渐进平均收敛的压缩率也小于等于β, 进而其整体收敛的阶为$O(\lambda^n)$, 其中$\lambda \leq \beta$.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

基于初始点v^0, 固定平稳策略的收敛情况.

如果将$I-\beta P(f)$正规分解成为

$$Q(f) = \begin{bmatrix} 1-\beta p_{11} & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1-\beta p_{22} & 0 & & & 0 \\ 0 & 0 & 1-\beta p_{33} & & & 0 \\ \cdot & & & \cdot & \cdot & \cdot \\ \cdot & & & & \cdot & 0 \\ 0 & & & & 0 & 1-\beta p_{NN} \end{bmatrix},$$

$$R(f) = \beta \begin{bmatrix} 0 & p_{12} & \cdot & \cdot & \cdot & p_{1N} \\ p_{21} & 0 & p_{23} & \cdot & \cdot & p_{2N} \\ p_{31} & p_{32} & 0 & p_{34} & & p_{3N} \\ \cdot & & & \cdot & & \cdot \\ \cdot & & & & \cdot & \cdot \\ p_{N1} & \cdot & \cdot & \cdot & p_{N,N-1} & 0 \end{bmatrix},$$

就是雅可比(Jacobi)值迭代算法. 此时$Q(f)$是对角矩阵,求逆非常方便. 如果对一切$k=1,2,\cdots,N$,$p_{kk}>0$时, 则必有$||Q^{-1}(f)R(f)||<\beta$, 这时雅可比值迭代算法比起值迭代算法收敛的速度要快.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

如果将雅可比值迭代算法中$(I-\beta P(f))v$写成分量形式

$$(1-\beta p_{kk})v(i_k) = \beta(p_{k1}v(i_1)+\cdots+p_{k,k-1}v(i_{k-1}) + p_{k,k+1}v(i_{k+1})+\cdots+p_{kN}v(i_N)). \quad (54)$$

将(54)式两端同时除以$(1-\beta p_{kk})$, 就得到用v的其它项来表示$v(i_k)$的形式.

关于分解算法,我们还可以考虑一些变形:

1) 雅可比与高斯塞德尔值迭代算法的结合;

2) 分块雅可比值迭代算法: 就是类似于式(54)中做法, 将v中元素分为l块, 由此将问题转化为矩阵$l\times l$上的问题;

3) 分块高斯塞德尔值迭代算法;

4) 分块高斯塞德尔与雅可比值迭代算法的结合。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

上面的后三种算法特别适合于并行计算. 算法的进一步发展就是上松弛或者是**下松弛** (under-relaxation)的. 假设$(Q(f), R(f))$是$I - \beta P(f)$的一个正规分解, 对于$0 < \omega < 2$,定义

$$v^{n+1} = v^n + \omega \left[\max_{f \in F} \{Q^{-1}(f)R(f) + Q^{-1}(f)R(f)v^n\} - v^n \right]. \qquad (55)$$

当取$\omega = 1$时,就是标准的分解算法; 当$\omega > 1$时, 也就是比较多的考虑v^n的增加, 即为上松弛算法; 当$\omega < 1$时, 为下松弛算法. 特别是迭代为单调时, 上松弛算法要好于下松弛算法. 在实际的计算中如何选取ω 的值是至关重要的. Kushner等人和Reetz分别提出和分析了高斯塞德尔值迭代算法与上松弛算法的结合算法. 特别是Reetz提出取$\omega = \min_{a \in A(i), i \in S}[1 - \beta p(i|i,a)]^{-1}$并证明了其渐进收敛速率不超过$\beta$.而有人认为$\omega$应该在1.4附近, Porteus则认为$\omega$应该取值在1.2附近收敛更快些.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

例3.3：假设MDP问题是由单个平稳策略组成, 其转移矩阵和报酬函数为:

$$P = \begin{bmatrix} 0.2 & 0.4 & 0.4 \\ 0.3 & 0.3 & 0.4 \\ 0.5 & 0.5 & 0 \end{bmatrix}, \quad r = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

当$\beta = 0.9$时, $V_{0.9}^* = (I - 0.9P)^{-1}r = (18.82, 19.73, 20.35)^T$. 注意到

$$P^L = \begin{bmatrix} 0 & 0 & 0 \\ 0.3 & 0 & 0 \\ 0.5 & 0.5 & 0 \end{bmatrix}, \quad P^U = \begin{bmatrix} 0.2 & 0.4 & 0.4 \\ 0 & 0.3 & 0.4 \\ 0 & 0 & 0 \end{bmatrix},$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

又根据$Q^{-1}R = (I - \beta P^L)^{-1}(\beta P^U)$, 我们有

$$Q^{-1}R = \begin{bmatrix} 0.1800 & 0.3600 & 0.3600 \\ 0.0486 & 0.3672 & 0.4572 \\ 0.1028 & 0.3272 & 0.3677 \end{bmatrix}.$$

通过观察可见$Q^{-1}R$的第一行等于βP的第一行, 这很明显因为普通的值迭代算法与高斯塞得尔对于第一个分量是一致的. 由于此时$||Q^{-1}R|| = ||\beta P|| = 0.9$,所以普通值迭代算法的压缩率是0.9. 而高斯塞得尔值迭代算法的压缩率为0.84. 此时还有$\sigma(Q^{-1}R) = 0.88$, 所以雅可比值迭代算法的压缩率是0.88. 尽管每一步的改进并不大, 但是就平均意义上来说是很明显的。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

6 改进的策略迭代算法

在策略迭代算法中，策略求值时需要解线性方程组(27). 如果状态空间中的元素是N的话, 解方程所需要的N^3次乘除运算. 特别是当N很大时, 计算量是十分可观的. 可是如果采用值迭代类的方法时, 有限步无法得到精确的最优策略和最优值函数. 所以人们考虑了改进的策略迭代算法, 即策略迭代与值迭代的混合型算法。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

令$\{m_n\}$为一非负整数列.

算法3.4 (改进的策略迭代算法)

步骤1: 取$v^0 \in B$且满足$Lv^0 = Tv^0 - v^0 \geq 0$, 给定$\epsilon > 0$并且置$n = 0$.

步骤2: (策略改进)取$f_{n+1} \in F$满足:

$$f_{n+1} \in \arg\max_{f \in F}\{r(f) + \beta P(f)v^n\}, \tag{56}$$

如果可能, 就取$f_{n+1} = f_n (n > 0$时)。

步骤3: (部分策略求值)

3(a). 置$k = 0$且令

$$u_n^0 \equiv \max_{f \in F}\{r(f) + \beta P(f)v^n\}, \tag{57}$$

3(b). 如果$||u_n^0 - v^n|| < \epsilon(1-\beta)/2\beta$, 进入步骤4. 否则进入步骤3的(c).

3(c). 如果$k = m_n$, 进入步骤3的(e). 否则由式(58)计算u_n^{k+1}

$$u_n^{k+1} = r(f_{n+1}) + \beta P(f_{n+1})u_n^k = T_{f_{n+1}}u_n^k. \tag{58}$$

由(56)保证了算子$T_{f_{n+1}}$.

3(d). 令k增加1后回到3(c).

3(e). 置$v^{n+1} = u_n^{m_n}$并将n增加1后, 进入步骤2.

步骤4: 令$f_\epsilon = f_{n+1}$并停止计算.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

该算法是将策略迭代与值迭代结合为一的算法. 象值迭代那样, 算法开始从满足$Lv^0 = Tv^0 - v^0 \geq 0$的v^0迭代, 停止准则在步骤3(b)中体现, 结果是得到ϵ最优策略. 而步骤3(a)中并没有增加额外的计算量, 这是因为在步骤2的式(56)中已经算过了. 算法也包含了策略的(非精确)求值步骤3. 与精确求值相比较, 只是计算了m_n步. 实际上$\{m_n\}$的选取有多种方式, 如:

1) 每次取固定的值($m_n = m$);

2) 逐步精确的选取(m_n随着n增加而增加);

3) 适应性(adaptively)选取, 例如要求$||u^{m_n+1} - u^{m_n}|| < \epsilon_n$, 这里$\epsilon_n$是固定或者变化的.

无论是采用哪种$\{m_n\}$, 下面我们给出了改进策略迭代算法的收敛性和收敛速度. 我们将这些内容总结为一个定理.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

定理3.17：对所有满足$Lv^0 = Tv^0 - v^0 \geq 0$的v^0和任意非负整数列$\{m_n\}$, 我们有:

1). 改进的策略迭代序列$\{v^n\}$以模$||\cdot||$单调收敛到V_β^*;

2). 算法有限次迭代终止于ϵ最优策略; 另外

3). 如果f_n为v^n的改进规则, f^*为V_β^*的改进规则. 那么

$$||v^{n+1} - V_\beta^*|| \leq \left(\frac{\beta(1-\beta^{m_n})}{(1-\beta)} ||P(f_n) - P(f^*)|| + \beta^{m_n+1} \right) ||v^n - V_\beta^*||; \quad (59)$$

4). 如果还有

$$\lim_{n\to\infty} ||P(f_n) - P(f^*)|| = 0, \quad (60)$$

则对任意的$\epsilon > 0$, 存在N, 当$n > N$时有

$$||v^{n+1} - V_\beta^*|| \leq (\beta^{m_n+1} + \epsilon)||v^n - V_\beta^*||. \quad (61)$$

类似于值迭代算法的改进, 也可以考虑改进的高斯塞得尔值迭代算法. 我们这里就不详细给出了.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

7 线性规划算法

在这一节里, 我们关注折扣马氏决策与线性规划之间的关系, 为此在这一节里仅考虑状态空间和行动集为有限的情形, 记状态空间中的元素数为$N < \infty$. 首先我们建立线性规划的模型. 由引理3.3 知道如果$v \in B$, 对所有的$f \in F$满足

$$v \geq r(f) + \beta P(f)v, \tag{62}$$

那么, v是MDP值V_β^*的一个上界. 所以MDP的寻优问题可以转化为如下的线性规划问题:

原规划 $(\mathbf{LP}_\beta)$

$$\min \sum_{i \in S} \frac{1}{N} v(i)$$

满足约束条件:

$$v(i) \geq r(i,a) + \beta \sum_{j \in S} p(j|i,a)v(j), \quad a \in A(i),\ i \in S.$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

实际上,规划里的$1/N$可以被替换为$\gamma(i) > 0, i \in S$而且$\sum_{i\in S}\gamma(i) = 1$.

我们将发现规划(LP_β)的**对偶规划**或者对偶问题会提供更多的信息用于分析折扣MDP问题.

对偶规划 $(\mathbf{DLP}_\beta)$

$$\max \sum_{i\in S}\sum_{a\in A(i)} r(i,a)x(i,a)$$

满足约束条件:

$$\sum_{i\in S}\sum_{a\in A(i)} [\delta(i,j) - \beta p(j|i,a)]x(i,a) = \frac{1}{N}, \tag{63}$$

以及对$i, j \in S$和$a \in A(i)$,$x(i,a) \geq 0$. 其中对所有的$i \in S$, 有$\delta(i,i) = 1$其它的$\delta(i,j) = 0$。

满足对偶规划(DLP_β)约束条件的$x(i,a)$被称为偶规划(DLP_β)的可行解. 下面我们给出对偶规划(DLP_β)的可行解与折扣MDP中的随机平稳策略之间的关系.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

定理3.18：1). 对每个随机马氏策略$\pi \in \Pi_m$, $i \in S$和$a \in A(i)$, 定义

$$x_\pi(i,a) \equiv \sum_{j\in S} \alpha(j) \sum_{n=0}^{\infty} \beta^n P_\pi\{Y_n = i, \Delta_n = a | Y_0 = j\}, \tag{64}$$

其中$\alpha(i) > 0$, $\sum_{i\in S} \alpha(i) = 1$, 为一个初始状态概率分布. 则$x_\pi(i,a)$是对偶规划$(DLP_\beta)$的可行解.

2). 设则$x(i,a)$是对偶规划(DLP_β)的可行解, 那么对于每个$i \in S$, 我们有$\sum_{a\in A(i)} x(i,a) > 0$. 定义随机平稳策略$\pi_0^\infty$为

$$\pi_0(a|i) = \frac{x(i,a)}{\sum_{a'\in A(i)} x(i,a')}. \tag{65}$$

则按照(64)定义的$x_{\pi_0}(i,a)$是对偶规划(DLP_β)的一个可行解, 而且对一切$i \in S$和$a \in A(i)$有$x_{\pi_0}(i,a) = x(i,a)$.

定理3.18的第一部分说明了随机平稳策略对应于对偶问题的一个可行解, 而第二部分则表明从对偶问题的任意一个可行解能够生成一个随机平稳策略。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

定理3.19：对折扣MDP的两个规划问题, 我们有:

1). 规划(LP_β)与其对偶规划(DLP_β)均有有限的最优解且最优值相等.

2). 令$\mathbf{v}^* = (v^*(1), v^*(2), \cdots, v^*(N))^T$为规划$(LP_\beta)$的一个最优解;则它为折扣MDP问题的最优值函数, 即$\mathbf{v}^* = V_\beta^*$.

3). 令$\{x^*(i,a)|a \in A(i),\ i \in S\}$为对偶规划$(DLP_\beta)$的一个最优解并且对每个$i \in S$,定义$x_i^* = \sum_{a\in A(i)} x^*(i,a)$.我们定义一个随机平稳策略$\pi_0^\infty \in \Pi_s$为

$$\pi_0(a|i) = \frac{x^*(i,a)}{x_i^*}, \quad a \in A(i),\ \ i \in S, \tag{66}$$

那么π_0^∞为折扣MDP的最优随机平稳策略. 特别的, 对每个$i \in S$, 由任一个使$x^0(i,a) > 0$的$a \in A(i)$组合起来的平稳策略$f \in F$, 是最优平稳策略.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

在考虑使用线性规划的方法求解折扣MDP问题时,必须要平衡下面的一些因素:

1) 生成线性规划表时会增加额外计算量,
2) 标准的线性规划很容易确定基本初始可行解的便利条件不适合,
3) 最优策略的结构信息在线性规划算法中无法体现,
4) 对于大型问题, (66)中的$x^*(i,a)$可能都很小, 确定最优策略会有困难,
5) 用线性规划做灵敏度分析比较容易, 以及
6) 在线性规划模型中加入约束条件比较方便.

总之，利用线性规划解折扣MDP问题时，要充分考虑到其利弊，才能有效的使用这个方法。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

8 可数状态与行动的模型

尽管前面个节的内容大都适用于可数状态空间，但是在实际的应用中仍然受到了很大的限制. 原因主要是下面的两个方面, 即报酬函数的有界性和针对可数状态空间时算法的可实现性. 下面我们就这两个方面的问题来说明.

为了简便起见, 状态空间设定为$S = \{0, 1, 2, \cdots, \}$. 其实这里的结果适用于任何可数状态的情形, 而且可以被推广到S为$[0, \infty)$或者欧氏空间的任一个无界的Borel子集的情形.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

8.1. 无界报酬的情形

在处理排队控制, 存储管理或者经济计划时, 一方面: 由于很难预先选定或者知道系统状态的上界，所以无法用有限状态的MDP模型对其进行建模和优化; 而另一方面: 在实际当中常常还指明报酬(或者费用)随着状态是非降的(例如,线性增加), 即报酬函数的一致有界性也被破坏了. 这时很自然的就会用到可数状态的MDP模型. 尽管“报酬函数一致有界”的条件不再满足了, 但是我们看到在任何一个状态上采取任一个行动时一周期的期望报酬还是有界的. 因此我们在这样的条件下讨论问题。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

下对于S上的任一个正的实值函数w, 如果它满足$\inf_{i\in S} w(i) > 0$, 我们定义**权重上界范数** $||\cdot||_w$为

$$||v||_w = \sup_{i\in S} w(i)^{-1}|v(i)|, \tag{67}$$

其中$v \in B$, 并且令B_w为S上所有满足w权重有界的实值函数全体. 下面我们需要一些假设条件以保证我们讨论的问题是有意义的。

假设3.1：存在常数$\mu < \infty$满足

$$\sup_{a\in A(i)} |r(i,a)| \leq \mu w(i). \tag{68}$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

假设3.2：1) 存在常数κ满足$0 \leq \kappa < \infty$使得

$$\sum_{j \in S} p(j|i,a)w(j) \leq \kappa w(i). \tag{69}$$

对一切$a \in A(i)$和$i \in S$成立.

2) 对于每个$\beta \in [0,1)$,存在$\alpha \in [0,1)$和一个整数J, 对一切的策略$\pi \in \Pi_m^d$满足

$$\beta^J \sum_{j \in S} p^J(j|i,\pi)w(j) \leq \alpha w(i), \tag{70}$$

其中$p^J(j|i,\pi)$表示从状态i出发J步之后转移到状态j的概率, 也就是乘积矩阵$P(f_0)P(f_1)\cdots P(f_{J-1})$的第$(i,j)$个元素.

上面假设3.2的两个条件并不总是满足的, 下面的例子就能说明问题。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

例3.4：对于本节的S,如果每个$i \in S$,$A(i) \equiv A = \{0, 1, \cdots, \}$,$r(i, a) = i$, 以及当$j = i + a$时$p(j|i, a) = 1$, 其它的$p(j|i, a) = 0$. 如果取$w(i) = \max\{1, i\}$, 很明显式(68)对任意固定的$\mu \geq 1$是成立的. 但是由于

$$\sum_{j \in S} p(j|i, a)w(j) = w(i + a) = i + a,$$

所以式(69)和(70)都不成立. 但是如果$A(i) = \{0, 1, \cdots, M\}$, 那么

$$\sum_{j \in S} p(j|i, \pi)w(j) = i + a \leq i + M \leq (1 + M)w(i)$$

对一切的策略$\pi \in \Pi_m^d$成立, 所以取$\kappa = (1 + M)$,式(69)成立. 又由于对任意的策略$\pi \in \Pi_m^d$有

$$\beta^J \sum_{j \in S} p^J(j|i, \pi)w(j) \leq \beta^J(i + MJ) \leq \beta^J(1 + MJ)w(i).$$

当J足够大, 使得$\beta^J(1 + MJ) < 1$时, 式(70)成立。

如果$r(i, a) = i^2$,$A(i) = \{0, 1, \cdots, M\}$时, 可以取$w(i) = \max\{1, i^2\}$以保证假设的两个条件成立. 由此我们可以知道, 合理的选择$w(i)$及其相关参数, 可以使很多实际问题纳入我们所讨论的范围.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

性质3.4：如果假设3.1和3.2成立, 则对于任意的$\pi \in \Pi_m^d$和$i \in S$, 我们有:

$$|V_\beta(i,\pi)| \leq \frac{\mu}{1-\alpha}[1+\beta\kappa+\cdots+(\beta\kappa)^{J-1}]w(i) \tag{71}$$

并且还有

$$||V_\beta(\pi)||_w \leq \frac{\mu}{1-\alpha}[1+\beta\kappa+\cdots+(\beta\kappa)^{J-1}]. \tag{72}$$

证明思路: 由假设3.1和3.2可以直接验证.

就象有界报酬那样, 性质3.4给出了值函数的界, 而这个界对于最一般的策略类Π也是适用的. 在实际问题中, κ常常会大于1(如在例3.4中那样), 因此求极大或者是上界的算子T(见公式(13)) 在B_w上不是压缩算子. 但是在假设3.1和3.2成立的条件下, 它是J步压缩的. 也就是存在整数J和正数$\lambda \in [0,1)$, 对于任意的$u, v \in B_w$有

$$||T^J u - T^J v|| \leq \lambda ||u-v||.$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

我们将这些总结为下面两个结论。

引理3.6：设B是Banach空间, $T : B \to B$为J步压缩算子($J \geq 1$), 并且存在数M满足$0 \leq M < \infty$, 以及对任意的$u, v \in B$ 有

$$||Tu - Tv|| \leq M||u - v|| \tag{73}$$

那么, 我们有:1) 存在唯一的$v^* \in B$满足$Tv^* = v^*$; 2) 任取$v^0 \in B$, 由$Tv^n = v^{n+1}$生成的序列v^n收敛到v^*.

引理3.7：假设3.1和3.2成立, 那么T为B_w上的J步压缩算子.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

定理3.20：设S为可数集合并且假设3.1和3.2成立, 我们有:

1). 最优方程

$$v = \sup_{f \in F}\{r(f) + \beta P(f)v\}$$

有唯一的解$v^* = V_\beta^* \in B_w$.

2). 令对任意的$\epsilon > 0$, 存在$f_\epsilon \in F$满足$V_\beta(f_\epsilon^\infty) + \epsilon e \geq V_\beta^*$.

3). 如果存在$f^* \in F$满足

$$f^* = \arg\max_{f \in F}\{r(f) + \beta P(f)V_\beta^*\}$$

则$(f^*)^\infty$为最优策略.

4). 对任意的$v^0 \in B_w$, $\lim_{n\to\infty} ||T^n v^0 - V_\beta^*|| = 0$.

5). 对任意的$v^0 \in B_w$, 若有$Tv^0 \geq v^0$, 那么由策略迭代算法或者改进的策略迭代算法生成的序列v^n 收敛到V_β^*.

证明思路: 由假设条件3.1和3.2成立和令$M = \beta\kappa$, 根据引理3.6和3.7可得1)和4)成立. 而部分2)和3)分别是定理3.5和定理3.8 的重新表述. 最后, 部分5)的证明则同于定理3.17.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

假设3.2扮演了重要的角色. 我们下面给出假设3.2成立的一些充分条件.

性质3.5：1). 假设存在一个常数$C > 0$, 对于所有的$a \in A(i)$和$i \in S$, 我们有:

$$\sum_{j \in S} p(j|i,a)w(j) \leq w(i) + C \tag{74}$$

成立. 那么假设3.2成立.

2). 如果存在一个整数M和一个常数$C > 0$, 对于所有的$a \in A(i)$, $i \in S$和$k = 1, 2, \cdots, M$, 我们有:

$$\sum_{j \in S} p(j|i,a)y(j)^k \leq [y(i) + C]^k \tag{75}$$

成立, 其中$y(i) = w(i)^{1/M}$. 那么假设3.2成立.

3). 如果假设3.1成立而且其中的$w(i)$是$[0, \infty)$上的一个可微, 凹且非降函数在集合$\{0, 1, \cdots\}$上的限制时, 以及还对于所有的$a \in A(i)$,$i \in S$满足

$$\sum_{j \in S} jp(j|i,a) \leq K, \tag{76}$$

其中K为一个有界的数. 那么假设3.2成立.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

从理论上讲, 我们可以利用策略迭代, 值迭代或者改进的策略迭代来求解具有无界报酬的可数状态MDP问题了. 但是在实际的问题中这是不可能的, 因为在每一步迭代时都需要计算无限次. 通常补救办法有三个: 其一是利用有限状态空间来逼近可数的情形并且分析界与误差; 其二是分析具有固定结构的v^n,v^n的改进规则以及发展出相关的算法; 再有就是将最优策略分划为有限个(一般是一个) 常返类和一个瞬时状态类,并且解一个有限状态集合的问题, 其有限的状态包括了最优策略的所有常返状态。

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

8.2. 有限状态逼近无限状态的情形

这一小节我们考虑用有限状态的MDP问题来逼近可数状态的MDP问题. 我们认为假设3.1和3.2在这节的所有定理中成立, 不再另行说明了. 我们先通过一个数例观察一下.

例3.5：取状态空间S为$\{0, 1, \cdots\}$.对每个状态$i \in S$,$A(i) = \{a_i\}$是个单点集合,$r(i, a_i) = \mu$,以及当$j = i + 1$时$p(j|i, a_i) = 1$,其它的$p(j|i, a_i) = 0$.很明显, 对于任一个状态$i \in S$,$V_\beta^*(i) = \mu(1 - \beta)^{-1}$. 如果我们用状态空间$S_N = \{0, 1, \cdots, N\}$来替代$S$, 而且将状态$N$的行动集合改为单点集$a'_N$, 转移概率修正为$p(N|N, a'_N) = 1$以及报酬修正为$r(N, a'_N) = 0$, 就得到了一个有限状态的MDP问题.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

我们记$V^*_{\beta,N,0}$为其期望折扣值函数(下面我们会解释这个记号). 那么$V^*_{\beta,N,0}(0) = \mu(1-\beta)^{-1}(1-\beta^N)$. 因此

$$V^*_\beta(0) - V^*_{\beta,N,0}(0) = \mu\beta^N(1-\beta)^{-1},$$

如果N取的很大, 我们可以认为$V^*_\beta(0) - V^*_{\beta,N,0}(0)$可以任意的小. 不难注意到对所有的$N$都有$V^*_\beta(0) > V^*_{\beta,N,0}(0)$,而且有限状态模型的每个状态的值函数都单调增的收敛到V^*_β.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

下面我们一般的讨论利用有限状态空间逼近无限状态空间的问题. 为了便于叙述, 我们把由前$N+1$个状态截得的MDP问题中的各项参数做如下的表示. 状态空间$S_N = \{0, 1, \cdots, N\}$. 对固定的$u \in B_w$(例3.5中的$u = 0$), 我们定义$v \in B_w$,

$$v_{N,u}(i) = \begin{cases} v(i) & i \le N; \\ u(i) & i > N. \end{cases}$$

对于$f \in F$,定义$T_f : B_w \to B_w$为$T_f v = r(f) + \beta P(f)v$. 还定义算子$T_{f,N,u} : B_w \to B_w$为

$$T_{f,N,u}v(i) = \begin{cases} r(i,f) + \beta \sum_{j \le N} p(j|i,f)v(j) + \beta \sum_{j>N} p(j|i,f)u(j) & i \le N; \\ u(i) & i > N. \end{cases}$$

当$u = 0$时是最为常用的, 而且看上去也简洁, 即为:

$$T_{f,N,0}v(i) = \begin{cases} r(i,f(i)) + \beta \sum_{j \le N} p(j|i,f(i))v(j) & i \le N; \\ 0 & i > N. \end{cases}$$

而一般我们选取u时, 总会考虑到尾项$\sum_{j>N} p(j|i,f)u(j)$的易处理性.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

对固定的N, $f \in F$以及$u \in B_w$, 算子$T_{f,N,u}$在B_w上是J步压缩的, 由定理3.20中方法产生的序列在B_w中有唯一的不动点$V_{\beta,N,u}(f)$. 而且当$i > N$时, $V_{\beta,N,u}(i) = u(i)$. 我们把$V_{\beta,N,u}(f)$称为$V_\beta(f)$的一个**N状态逼近**.

我们记$F(N)$为S_N上的决策函数全体(也可以认为是平稳策略类); $V^*_{\beta,N,u} = \sup_{f\in F(N)} V_{\beta,N,u}(f)$; 以及算子$T_{N,u} : B_w \to B_w$为

$$T_{N,u}v \equiv \sup_{f\in F(N)} T_{f,N,u}v. \tag{77}$$

有本小节的假设知道$T_{N,u}$是B_w中J步压缩的, 其不动点为$V^*_{\beta,N,u}$.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

引理3.8：1). 如果对某个$u \in B_w$有$T_f u \geq (\leq) u$,那么对所有的k和N有$(T_{f,N,u})^k u \geq (\leq)(T_{f,N-1,u})^k u$.
2). 如果对某个$u \in B_w$有$T_f u \geq u$或者有$u \geq T_f u$,那么对每个$i \in S$, $V_{\beta,N,u}(i,f)$ 单调收敛于$V_\beta(i,f)$.

当$T_f u \geq u$时, $V_{\beta,N,u}(f)$增加地逼近$V_\beta(f)$, 这时我们称之为$V_{\beta,N,u}(f)$ **从下方逼近**$V_\beta(f)$. 如果$u \geq T_f u$时, $V_{\beta,N,u}(f)$下降地逼近$V_\beta(f)$, 这时我们称之为$V_{\beta,N,u}(f)$ **从上方逼近**$V_\beta(f)$.

下面的定理表明N状态逼近的最优值函数点状单调收敛到原问题的最优值函数.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

定理3.21：我们有:

1). 如果存在$u \in B_w$对一切$f \in F$满足$T_f u \geq u$, 那么对每个状态$i \in S$, $V^*_{\beta,N,u}(i)$ 从下方单调收敛到$V^*_\beta(i)$.

2). 如果存在$u \in B_w$对一切$f \in F$满足$T_f u \leq u$,并且对某些$i \in S$有

$$\lim_{N\to\infty} \sup_{a\in A(i)} \sum_{j>N} p(j|i,a)w(j) = \lim_{N\to\infty} \sup_{f\in F} \sum_{j>N} p(j|i,f(i))w(j) = 0 \qquad (78)$$

那么在这些状态上$V^*_{\beta,N,u}(i)$随着$N \to \infty$而从上方单调收敛到$V^*_\beta(i)$.

证明思路: 部分1)可以利用定理3.20的2)直接估计$V^*_{\beta,N,u}(i)$与$V^*_\beta(i)$ 的差得到.

不难看到定理3.21的两个部分是不对称的, 部分2)的要求要高些, 事实上Fox的例子表明这些要求是必要的. 为了给出逼近的改进的策略迭代算法, 我们要在状态集$S_\nu = \{0, 1, \cdots, \nu\}$上得到可数状态模型的精确逼近值. 首先取好$u \in B_w$使得对一切$f \in F$有$T_f u \geq u$; 或者对一切$f \in F$有$T_f u \leq u$且满足(78)式的要求.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

算法3.5 (逼近的改进策略迭代算法)

步骤1: (初始化) 置$v^0 = u$, $n = 0$, 取$\epsilon > 0$并且选定整数ν以及一个非负整数列$\{m_n\}$.

步骤2: (策略改进) 取f_{n+1}满足:

$$f_{n+1} \in \arg\max_{f \in F(n)} \{r(f) + \beta P(f) v^n\}. \tag{79}$$

步骤3: (部分逼近策略求值) 将n增加1并令

$$v^n = (T_{f_n, n-1, u})^{m_n - 1} v^{n-1}. \tag{80}$$

步骤4: (停止规则) 如果

$$\max_{i \le \nu} \{v^n(i) - v^{n-1}(i)\} - \min_{i \le \nu} \{v^n(i) - v^{n-1}(i)\} < \epsilon,$$

停止并对$i \le \max\{\nu, n\}$取

$$f^*_\nu(i) \in \arg\max_{a \in A(i)} \left\{ r(i,a) + \beta \sum_{j \in S} p(j|i,a) v^n(j) \right\},$$

而对$i > \max\{\nu, n\}$时, 任意取. 否则返回到步骤2.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

这个算法与通常的改进策略迭代算法的区别在于策略求值时的选择和停止的规则. 在每一次的迭代中，利用增加状态进行逼近，而停止规则是仅限制在状态集合S_ν上.

尽管上面的算法给出了停止的规则, 但是在实际当中总希望能够预先的对某个给定集合S^*给出一个界N, 使得这个N状态空间的最优解与原问题的最优解在状态集合S^*上的差别足够小. 为此我们认为满足定理3.21的$u \in B_w$已经找好了.我们将状态空间划分为

$$\emptyset = S_0 \subseteq S_1 \subseteq \cdots \subseteq S_\nu \subseteq S_{\nu+1} = S,$$

并且对某个k_0有$S^* \subseteq S_{k_0}$.

对于两个集合A和B, 如果$A \subseteq B$, 我们用B/A表示A在B中的余集. 对任意的$1 \le l \le \nu$和$1 \le k \le \nu$, 令

$$q(S_l, S/S_k) = \sup_{i \in S_l} \sup_{a \in A(i)} \left\{ \sum_{j \in S/S_k} p(j|i,a) \right\}.$$

这里q的意义在于描述了“一种”从状态集合S_l到状态集合S/S_k的“转移规律”.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

面我们可以定义一个聚合了的转移矩阵Q, 其分量分别为:

$$Q(l, \nu+1) \equiv q(S_l, S/S_\nu), \quad 1 \le l \le \nu,$$
$$Q(l,k) \equiv q(S_l, S/S_{k-1}) - q(S_l, S/S_k), \quad 1 \le l, k \le \nu.$$

实际上$Q(l,k)$提供了从状态集合S_l到状态集合S_k/S_{k-1}转移的概率上的界. 我们用Q表示一个$\nu \times \nu$的矩阵, 其分量为$Q(l,k)$. 注意, 它不包括分量$Q(l, \nu+1)$. 我们还记

$$\begin{aligned}
\Delta(l) &\equiv \sup_{i \in S_l} |V_\beta^*(i) - V_{\beta,N,u}^*(i)| \\
b(l) &\equiv \sup_{i \in S_l} \sup_{a \in A(i)} \left\{ \beta \sum_{j \in S/S_\nu} p(j|i,a) |V_\beta^*(j) - u(j)| \right\} \\
b'(l) &= \sup_{i \in S_l} \{ \beta \kappa w(i) ||V_\beta^* - u||_w \}.
\end{aligned}$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

定理3.22：令$\{S_l\}, l = 1, 2, \cdots, \nu+1, \Delta, Q, b$和$b'$如上定义,又$0 \leq \beta < 1$.我们有:

$$\Delta \leq (I - \beta Q)^{-1} b \leq (I - \beta Q)^{-1} b'. \tag{81}$$

证明思路: 取N使$S_\nu = \{0, 1, \cdots, N\}$. 利用$V_\beta^*$和$V_{\beta,N,u}^*$分别为$T$和$T_{N,u}$的不动点, 直接估计它们的差.

推论3.2：设报酬函数$|r(i, a)| \leq \mu$, $i \in S$和$a \in A(i)$. 那么公式(81)中的b可被替换为:

$$b''(l) = Q(l, \nu + 1) \frac{2\mu\beta}{1 - \beta}. \tag{82}$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

8.3. 设备维修的例子

为了进一步说明状态空间逼近的结果, 我们考虑一个设备维修的例子. 在制造过程中有一台设备在运行, 在每个决策时刻观察到它的运行状态为$S = \{0, 1, \cdots\}$. 状态的数值越大表示其运行条件越差. 而决策者能够选择的行动有两个: 更换新设备(a=1); 或者继续运行(a=0); 即对每个状态$i \in S$而言, 行动集合都是$A(i) = A = \{0, 1\}$. 在决策周期之间设备磨损等级增加j个的概率是$p(j)$(不依赖当前的状态),所以模型的转移概率满足:

$$p(j|i, 0) = \begin{cases} 0 & j < i \\ p(j-i) & j \geq i \end{cases}$$

以及$p(j|i, 1) = p(j), j \geq 0$(这里的意思是: 更换不需要时间, 其后运行了一个决策周期的状态变化规律依然是$p(j)$). 报酬函数考虑到三个因素: 每个周期的固定收入$R > 0$; 与状态相关的运行费用$h(i)$, 关于状态i 非降; 以及更换设备费用K. 所以在状态i的费用函数为:

$$r(i, a) = \begin{cases} R - h(i) & a = 0; \\ R - h(0) - K \equiv R' & a = 1. \end{cases}$$

这里运行费用$h(i)$关于状态i非降表示状态越差, 需要的运行费越高.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

最优方程. 首先我们对$h(i)$的不同选取验证条件3.1和3.2. 记$G = R + K$, 并且用Y记具有概率分布$p(i)$的随机变量. 我们通过性质3.5验证条件3.2.

如果$h(i) = i$. 令$w(i) = i + G$以及$\mu = 1$, 则假设条件3.1成立. 此时

$$\begin{aligned}\sum_{j\in S} p(j|i,0)w(j) &= \sum_{j=i}^{\infty} p(j-i)j + G \\ &= \sum_{k=0}^{\infty} p(k)(k+i) + G = i + E[Y] + G,\end{aligned}$$

而

$$\sum_{j\in S} p(j|i,1)w(j) = \sum_{j=0}^{\infty} p(j)j + G = E[Y] + G.$$

所以, 当$E[Y] < \infty$时, 令$C = E[Y]$, 性质3.5的1)成立.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

如果$h(i) = i^2$,令$w(i) = (i+G)^2$以及$\mu = 1$, 则假设条件3.1依然成立. 此时为了验证条件3.2, 我们利用性质3.5的2), 其中$M = 2$. 由于$y(i) = [(i+G)^2]^{1/2} = i+G$, 只要$E[Y^2] < \infty$,就有

$$\begin{aligned}\sum_{j\in S} p(j|i,0)y(j)^2 &= \sum_{k=0}^{\infty} p(k)(i+k+G)^2 \\ &= (i+G)^2 + E[Y](i+G) + E[Y^2] \\ &\le (i+G+E[Y^2]^{1/2})^2.\end{aligned}$$

再加上

$$\sum_{j\in S} p(j|i,0)y(j) = i+G+E[Y] \le i+G+E[Y^2]^{1/2},$$

令$C = E[Y^2]^{1/2}$, 所以性质3.5的2)成立.

如果h是一个定义在$[0,\infty)$上可微非降凹函数的限制时, 例如$log(i+1)$等, 加上$E[Y] < \infty$, 可以利用性质3.5的3)得到条件3.2.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

由此, 对以上满足条件的h, 最优方程为:

$$v(i) = \max\left\{R - K - h(0) + \beta\sum_{j=0}^{\infty} p(j)v(j), R - h(i) + \beta\sum_{j=0}^{\infty} p(j)v(i+j)\right\},$$

并且在B_w中有唯一解V_β^*, 这个解可用值迭代或者其他的迭代算法求.

N阶段逼近的收敛性. 由于$r(i,a) \le R - h(0)$, 所以$V_\beta^*(i) \le [R - h(0)]/[1-\beta]$. 取

$$u(i) = [R - h(0)]/[1-\beta], \quad i \in S.$$

下面我们要说明对一切的$f \in F$, 有$T_f u \le u$. 对于$i \in S$,当行动取0($f(i) = 0$)时, 由$h(i) \ge h(0)$可推得

$$T_f u(i) = R - h(i) + \beta\frac{R - h(0)}{1-\beta} \le \frac{R - h(0)}{1-\beta} = u(i).$$

由于$K \ge 0$, 对$f(i) = 1$时可类似得到上面的结论.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

对任意的$i \in S$

$$\sum_{j>N} p(j|i,1) \leq \sum_{j>N} p(j|i,0) = \sum_{j>N} p(j-i).$$

所以任给$\epsilon > 0$, 对足够大使

$$\sum_{j>N'} p(j) < \epsilon$$

的N',取$N > i + N'$, 就有

$$\max_{a=0,1} \sum_{j>N} p(j) < \epsilon,$$

这表明了定理3.21中2)的条件满足, 所以对于每个$i \in S$, $V^*_{\beta,N,u}$随着N下降的收敛到V^*_β.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

如果再假设$h(i)$是有上界的, 由$\lambda > 0$界住,比如说$h(i) = \lambda(1 - e^{-i})$, 其中$\lambda > 0$. 那么我们取

$$u'(i) = -(\alpha + K)/(1 - \beta),$$

则对一切$f \in F$有$T_f u' \geq u'$. 由定理3.21中1), 知道$V^*_{\beta,N,u'}$上升的收敛于V^*_β. 这样我们得到了一个自然的关系

$$V^*_{\beta,N,u'} \leq V^*_\beta \leq V^*_{\beta,N,u}.$$

也就是说,可以利用$V^*_{\beta,N,u}(i) - V^*_{\beta,N,u'}(i)$作为$N$状态逼近算法的停止规则.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

预定的界. 基于$V^*_{\beta,N,u}$与V^*_β的差, 我们给出预定的界. 对固定的n, 在

$$\Delta(1) = \sup_{i \le n} |V^*_{\beta,N,u}(i) - V^*_\beta(i)|$$

上寻找一个界. 为了简便起见, 我们仅就满足以下两个条件的情形下讨论界的估计问题: (i).存在$\delta < \infty$, 对某些$i \in S$满足$\delta \ge h(i) > K + h(0)$; 并且(ii). $\{p(j)\}$ 有有限的**支撑**, 即存在一个$M > 0$使得$\sum_{j=0}^{M} p(J) = 1$. 其实条件(i)也包含了下界, 因为如果$R - h(i) > R - K - h(0)$对一切$i \in S$成立, 最优解就会是永远不更换机器. 条件(i)还包含了$|r(i,a)| \le \rho$, 其中

$$\rho = \max\{|R - \delta|, |R - h(0) - K|, |R - h(0)\}.$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

在有限状态逼近时初始化对一切$i \in S$取$u(i)$为$\rho(1-\beta)^{-1}$或者为$-\rho(1-\beta)^{-1}$. 取整数$K' > 0$, 并令$S_0 = \emptyset$, $S_1 = \{0, 1, \cdots, n\}$, $S_2 = \{0, 1, \cdots, n+M\}$以及对$k = 1, 2, \cdots, K'$有$S_k = \{0, 1, \cdots, n+(k-1)M\}$. 不难验证这时候$K' \times K'$的矩阵$Q$为:

$$Q = \begin{bmatrix} p(0) & 1-p(0) & 0 & & 0 \\ 0 & p(0) & 1-p(0) & 0 & 0 \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ 0 & & 0 & p(0) & 1-p(0) \\ 0 & & & & p(0) \end{bmatrix} \tag{83}$$

以及当$j < K'$时, $b(j) = 0$和

$$b(K') = \frac{2\beta\rho}{1-\beta}[1-p(0)].$$

因此, 只要求出$(I - \beta Q)$的逆矩阵就可以利用定理3.22估计$\Delta(1)$的界了.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

为了避免求矩阵$(I-\beta Q)$的逆, 我们介绍一个概率的方法求$\Delta(1)$的界. 以$\{Y_n : n=0,1,\cdots\}$ 记一个状态空间$S=\{1,2,\cdots\}$上的马氏链, 限制在状态$\{1,2,\cdots,K'\}$上的转移矩阵恰为Q. 设随机变量Z表示这个链转移到状态$K'+1$的步数. 那么当$j \geq K'$时

$$P\{Z=j|Y_0=1\} = \begin{pmatrix} K'+j \\ j \end{pmatrix} [1-p(0)]^{K'}[p(0)]^j;$$

而当$j < K'$时, $P\{Z=j|Y_0=1\}=0$. 也就是说, 在$Y_0=1$的条件下Z服从一个非负的二项分布, 所以

$$\Delta(1) \leq \sum_{j=K'}^{\infty} \beta^j P\{Z=j|Y_0=1\} b(K') = \beta^{K'} b(K') \sum_{j=K'}^{\infty} \beta^{j-K'} P\{Z=j|Y_0=1\}.$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

注意到上式的最右边恰是$\beta^{K'}b(K')$与一个非负二项分布的母函数相乘, 于是就有

$$\Delta(1) \leq \left[\frac{1-p(0)}{1-\beta p(0)}\right]^{K'} \frac{2\rho\beta^{k'+1}(1-p(0))}{1-\beta}. \tag{84}$$

我们代一些数来估算这个界. 比如$\beta = 0.9$, $p(0) = 0.2$和$\rho = 1$, 式(72)右端成为$14.4(0.878)^{K'}$. 如果要保证$\Delta(1) < 0.1$, 则要求$K' = 39$. 如果$n = 9$以及$M = 5$, 这意味着我们只需要求解不超过205个状态的问题, 就可以保证$|V^*_{\beta,N,u}(i) - V^*_{\beta}(i)| \leq 0.1$,这里的$i = 0, 1, \cdots, 9$.

特别值得提的是, 找到S的一个好的分解, 我们会得到$\Delta(1)$的很紧的界.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

8.4. 有限状态可数行动的情形

在状态空间有限行动集合为可数或者一般集合的情形下, 对于最优方程存在解和最优策略是否存在的主要影响有以下几个方面: 1)报酬函数的有界性和它关于行动选取的连续性问题; 2)一步转移概率与行动选取的连续性问题. 即便是报酬函数是有界的, 也不能保证在一般的策略类里存在最优策略. 因此, 为了解决这两个问题, 人们通常总是在这两个方面加上一些假设条件, 类似于定理3.7中的条件. 如果放弃寻找最优的策略而求其次, 只需要求ϵ最优, 那么由定理3.8保证了其存在性(只要报酬函数是有界的). 很自然, 有界报酬的一般问题主要是求解ϵ最优的问题.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

由于行动集合的可数性知道, 此时平稳策略类中的策略个数已经不是有限的了. 不失一般性, 我们总认为在每个状态上的可用行动集合为$A = \{1, 2, \cdots\}$, 而状态的个数为N. 由于状态空间是有限性, 我们知道对任意的$f \in F$, 就会存在一个与f相关的整数n_f使得$f \in \times_1^N A_{n_f}$. 为了方便起见, 我们把状态空间有限, 行动集合可数的模型叫做**半无限MDP**. 如果任意从随机平稳策略类Π_s中选一个策略π(随机平稳策略应该记做π_0^∞. 但为了符号简洁, 我们仍然用π表示0时刻的平稳决策规则, 略去了下标0和上标∞), 我们定义一个标量

$$\mu_f(\pi) \equiv \prod_{i=1}^{N} \pi(f(i)|i). \tag{85}$$

$\mu_f(\pi)$的概率解释就是在每个状态时π选取行动恰好与f一样的概率. 很明显, 对任意的$f \in F$, 我们有$\mu_f(\pi) \geq 0$以及$\sum_{f \in F} \mu_f(\pi) = 1$.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

引理3.9：对任意的$\pi \in \Pi_s$,我们有:
1). $\pi = \sum_{f\in F} \mu_f(\pi) f$,即随机平稳策略$\pi$可以被表示为平稳策略的凸组合;
2). $r(\pi) = \sum_{f\in F} \mu_f(\pi) r(f)$以及$P(\pi) = \sum_{f\in F} \mu_f(\pi) P(f)$.

这个结果对于一般的行动集合也成立,只是表述需要利用测度论的一些知识. 尽管定理3.5到定理3.7表明了最优值函数和最优策略存在的条件.我们这里还要明确的表示一下.

定理3.23：对半无限MDP, 只要报酬函数有界, 则我们有:

$$\sup_{\pi\in\Pi} V_\beta(\pi) = \sup_{\pi\in\Pi_s} V_\beta(\pi) = \sup_{f\in F} V_\beta(f). \tag{86}$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

这里我们要用半无限的规划理论来考虑半无限MDP的问题. 首先我们把第3.7节的原规划(LP_β)和其对偶规划(DLP_β)写成数学规划里的标准形式:

$$LP_\beta: \quad \begin{array}{l} \inf b^T x \\ x^T A \geq r^T \\ x \in R^N \end{array} \quad \text{和} \quad DLP_\beta: \quad \begin{array}{l} \sup r^T y \\ Ay = b \\ y \geq 0 \end{array} \tag{87}$$

由于行动集合是可数集合, 所以按照Tijs的方法,我们这里引进一些记号. 定义

$$(\Re^\infty)^C \equiv \{x \in \Re^\infty | \text{存在 } n \in \mathcal{Z}_0, \text{当 } m \geq n,\ x_m = 0\}.$$

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

所以规划(87)可以写成为

$$LP_\beta(\infty): \quad \begin{array}{l} \inf b^T x \\ x^T A \geq r^T \\ x \in R^N \end{array} \quad \text{和} \quad DLP_\beta(\infty): \quad \begin{array}{l} \sup r^T y \\ Ay = b \\ y \geq 0 \\ y \in (\Re^\infty)^C \end{array} \tag{88}$$

在这两个规划中的矩阵A有有限行, 但是有可数列. 我们将规划(88)中的∞写作n时, 则表示它们的一个n截尾问题, 也就是将可数行动集合仅截取前n个行动组成的有限状态有限行动的MDP问题.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

定理3.24：对半无限MDP, 只要报酬函数有界, 则两个规划(88)的最优值相等, 也就是说强对偶定理成立, 特别的, 规划$LP_\beta(\infty)$有有限的最优解.

定理3.25：报酬函数有界的半无限MDP, 只要其对偶规划$DLP_\beta(\infty)$有最优解$\hat{y}$, 则存在有限的整数$\hat{n}$使得当$n \geq \hat{n}$,$DLP_\beta(n)$的最优值与$DLP_\beta(\infty)$的最优值相等. 进而, 当$n \geq \hat{n}$, 如果y_n^0是对偶规划$DLP_\beta(n)$的任一个基本解时, 我们定义

$$f_n^0(i,a) = \begin{cases} 1, & \text{如果 } y_n^0(i,a) > 0, \\ 0, & \text{否则}, \end{cases}$$

那么对于任意$n' \geq n$平稳策略f_n^0也是截取了n'个行动之后的有限MDP问题的最优策略, 也就是这个有限状态可数行动问题的最优平稳策略.

Chen等人、董泽清等人和Fainberg等用不同的方法证明了最优方程解的存在性这一结果.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

9 最优单调策略

下面介绍一下MDP方法中最为常见的一种具有特殊结构的策略–单调策略. 我们知道, 当状态空间可数或者更为一般的时候, 即便是知道存在最优的平稳策略, 但是如何便利的表达和寻找依然是个困难. 就象在第二章中那样, 我们给出具有特殊结构的最优策略, 就能够方便的给出和表达. 我们这里只考虑确定性平稳策略的结构(即在F中考虑).

由于状态空间的复杂性, 某种结构σ在算子T的变换下不一定能保持不变了. 我们特别记$B^\sigma \subseteq B_w$(或B) 和$F^\sigma \subseteq F$分别为具有结构的值函数空间和决策规则空间.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

定理3.26：设$S = \{0, 1, \cdots\}$以及假设条件和成立. 我们还假设对一切的$v \in B_w$, 存在$f \in F$满足$T_f v = Tv$, 另外如果

1). $v \in B^\sigma$则必有$Tv \in B^\sigma$,

2). $v \in B^\sigma$则必然存在$f' \in F^\sigma \cap \arg\max_{f \in F} T_f v$, 以及

3). B^σ为B_w中闭子集.

那么, 存在最优的平稳策略$f^* \in F^\sigma$, 满足$f^* \in \arg\max_{f \in F} \{r(f) + \beta P(f) V_\beta^*\}$.

由定理3.26, 可以建立策略迭代算法的收敛理论, 值迭代算法的收敛理论以及改进的策略迭代算法收敛理论. 我们这里就不一一列出了. 为了方便起见, 我们在这一节的余下部分总认为$S = \{0, 1, \cdots\}$, 对任意的$i \in S$, $A(i) \equiv A$. 回想第二章中的一些概念, 我们叙述下面的两个结果.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

定理3.27：假设条件3.1和3.2成立并且还有

1). $r(i,a)$对所有的$a \in A$关于i非降,

2). $q(k|i,a) \equiv \sum_{j=k}^{\infty} p(j|i,a)$对所有的$k$和$a \in A$关于$i$非降,

3). $r(i,a)$在集合$S \times A$上是上可加(下可加)的函数, 以及

4). $q(k|i,a)$在集合$S \times A$上关于一切k是上可加(下可加)的函数.

那么, 存在最优的平稳策略$f^* \in F^\sigma$,其中$f(i)$关于状态i非降(增).

定理3.28：假设条件3.1和3.2成立并且还有

1). $r(i,a)$对所有的$a \in A$关于i非增,

2). $q(k|i,a)$对所有的k和$a \in A$关于i非降,

3). $r(i,a)$在集合$S \times A$上是上可加的函数, 以及

4). $\sum_{j=0}^{\infty} p(j|i,a)u(j)$在集合$S \times A$上关于非增的$u$是上可加的函数.

那么, 存在最优的平稳策略$f^* \in F^\sigma$, 其中$f(i)$关于状态i非降.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

应用于排队或者存储模型时, 假设$A(i) = A$似乎有点强了, 适当的修正一下定理中的表述, 可以将定理中的条件放松为:

a). 对一切$i \in S$, $A(i) \subset A$,

b). 对一切$i' \geq i$, $A(i) \subset A(i')$, 或者

c). 对每个i, 如果$a \in A(i)$而且$a' \leq a$,那么必有$a' \in A(i)$.

现在我们给出一个保持策略结构寻求单调最优策略的算法. 这里我们假设定理3.27或者定理3.28 的条件成立.而且$S = \{0, 1, \cdots, N\}$以及对一切$i \in S$, $A(i) = A$. 这可以看成可数时的N逼近. 记F^σ为非降的决策规则.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

算法3.6 (单调策略迭代算法)

步骤1: 取$f_0 \in F^\sigma$并置$n = 0$.

步骤2: 通过解方程

$$(I - \beta P(f_n))v = r(f_n),$$

求出v^n。

步骤3: 置$i = 0$, $A(0) = A$.

a). 记

$$A^*(i) = \arg\max_{a \in A(i)} \left\{ r(i,a) + \beta \sum_{j \in S} p(j|i,a) v^n(j) \right\}.$$

b). 如果$i = N$, 进入3(d)步; 否则, 记

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

c). 将i增加1后返回到3(a).

d). 取$f_{n+1} \in F^\sigma \cap \times_{i \in S} A^*(i)$, 如果可能就取$f_{n+1} = f_n$。

步骤4: 如果$f_{n+1} = f_n$, 停止. 我们得到$f^* = f_n$. 否则将n增加1后返回到步骤2.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

10 最优策略的结构

在有限状态可数行动的情形的一节里，我们曾经讨论过随机平稳策略可以分解为确定性平稳策略的凸组合，这一节我们主要叙述最优策略所具有的结构。

如果将Bellman最优化原理纳入MDP的框架下讨论,早在1962年Blackwell就得到了有关最优策略结构的一些结果.

定理3.29：设状态空间S和行动集合A都是有限集合,$\pi = \{f_0, f_1, \cdots, f_n, \cdots\} \in \Pi_m^d$是最优的马氏策略.那么由第一个决策规则构成的平稳策略也是最优的,即策略$f_0^\infty \in \Pi_s^d$或等价的说$f_0 \in F$是最优的.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

随后Chitgopekar在1975年讨论了一些最优策略的组合问题.最后董泽清等人开始讨论最优策略一般所应具有的结构问题.这些结论对于寻找最优策略不无帮助,特别是很多学者继续讨论了ϵ最优策略的结构问题,同时这些结果还可以用到向量值报酬的MDP问题,有限阶段的问题,距最优的问题,递归报酬MDP问题,等等.这里就不详细列出了.我们这里只是给出最为基本的部分, 使读者能够清晰的看到这个问题的本质所在.

首先我们给出**可实现历史**的概念. 如果一个策略$\pi = \{\pi_0, \pi_1, \cdots\} \in \Pi$,对于任一个历史$h_t = (i_0, a_0, i_1, a_1, \cdots, i_t) \in H_t$(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的发生概率不是0,那么这样的历史就是在策略π下的一个可实现历史.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

下面我们还定义最优行动集. 对于每个状态$i \in S$

$$A^*(i) \equiv \arg\max_{a \in A(i)} \left\{ r(i,a) + \beta \sum_{j \in S} p(j|i,a) V_\beta^*(j) \right\} \tag{89}$$

被称为状态i的可用的最优行动集.

定理3.30：策略$\pi = \{\pi_0, \pi_1, \cdots, \pi_t, \cdots\} \in \Pi$是一个最优策略的充分必要条件是: 对任意的$t \geq 0$,如果历史$h_t = (i_0, a_0, i_1, a_1, \cdots, i_t) \in H_t$是策略$\pi$下的一个可实现历史,则当$a \in A(i_t) - A^*(i_t)$时必有$\pi_t(a|h_t) = 0$.

定理3.30的的意义在于: 一个策略是最优策略的充要条件就是每个决策规则在每一个可实现的历史上都要采用最优的行动.

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构

谢谢大家！

最优准则
最优方程
最优策略的存在性
策略迭代算法
值迭代算法
改进的策略迭代算法
线性规划算法
可数状态与行动的模型
最优单调策略
最优策略的结构