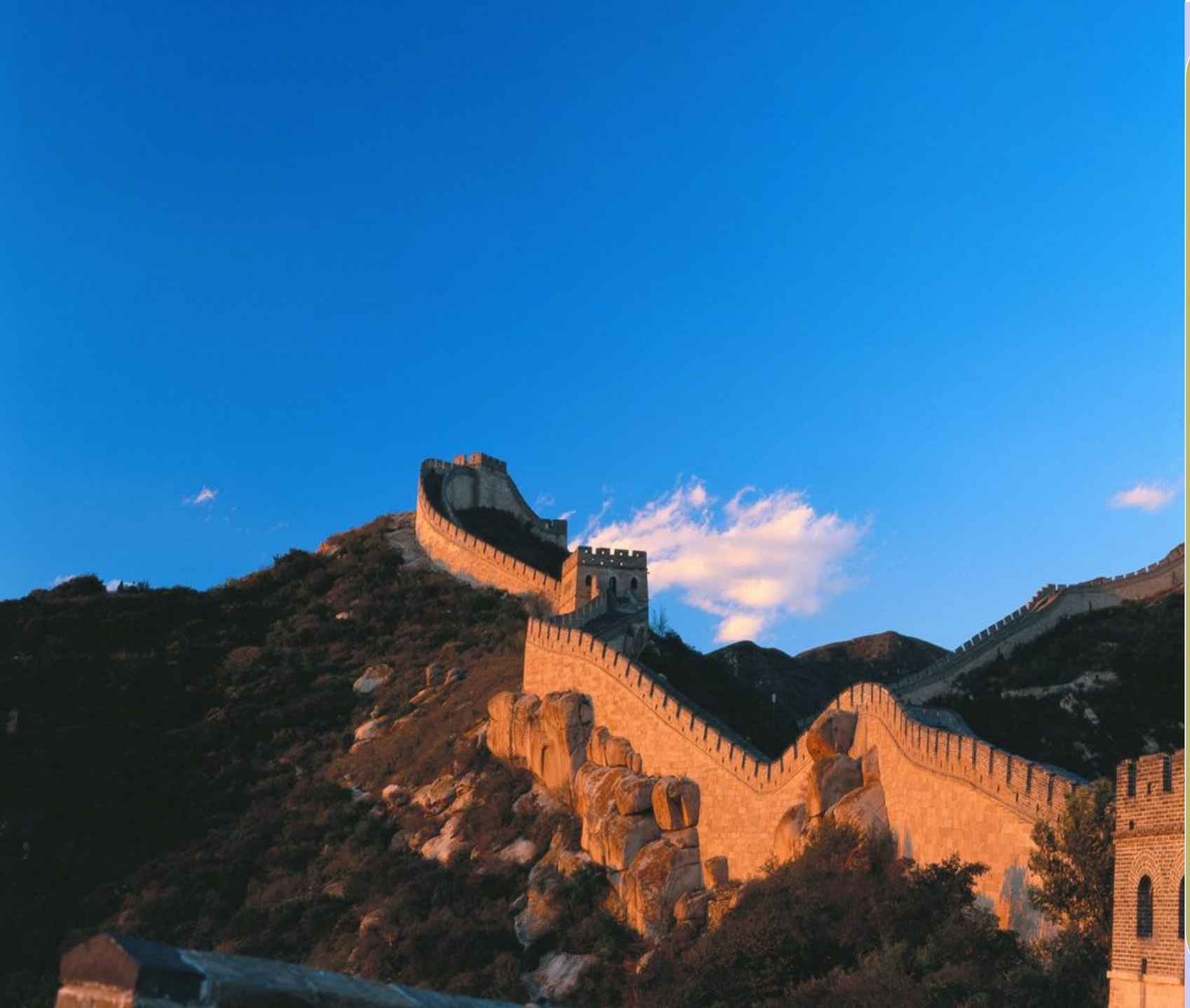

中国科学院研究生院

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

第三部分　马氏决策—连续时间模型

1 连续时间折扣MDP

前面各章介绍的MDP问题有着一个共同的特点就是决策时刻的集合T是等距离的离散点集合.在很多实际的问题中, 往往需要连续的观察系统的状态,随时都有可能发生状态的转移且需要决策者及时的做出决策,例如, 排队控制或者是设备维修等实际问题.因此,这一章里,我们允许决策者在任一个时刻可以采取决策,达到控制系统的目的.这里我们考虑一般的连续时间模型,包括半马氏模型.

1.1. 模型和策略的定义

在离散时间MDP的研究中,我们是用五重组$\{T, S, A(i), p(\cdot|i,a), r(i,a)\}$来刻画的,其中元素$p(\cdot|i,a)$这个元素描述的是一步转移率簇.当给定了初始分布和采用的策略之后,每一个时刻系统所处状态的概率就唯一确定了. 但是我们无法直接将这个方法用于连续时间的MDP问题,因为直接给出每个时刻t的转移概率函数$p_{ij}(t)$是十分困难的.一般只能利用马氏过程中的方法间接的描述$p_{ij}(t)$,即利用转移速率矩阵Q(也称为无穷小生成元)来完成.所以我们仍然可以利用五重组$\{T, S, A(i), q(\cdot|i,a), r(i,a)\}$来刻画,其中$T = [0,\infty)$是非负的半实轴;$S$是状态空间,为一个可数集合;$A(i)$是系统处于状态$i \in S$的可用行动集合, 也是可数集合;$q(\cdot|i,a)$是平稳的转移速率簇,即对任意的状态$i \in S$和行动$i \in A(i)$满足$-\infty < q(i|i,a)$以及当$i \neq j$时$q(j|i,a) \geq 0$且$\sum_{j\in S} q(j|i,a) = 0$; r是有界的报酬率函数,即满足$\sup_{i,a} |r(i,a)| \leq M_1$,这里$M_1 > 0$是一个常数.

我们仍然用f表示决策函数,全体记为F.连续时间MDP可以在任一时刻做决策来控制系统,沿袭原来的记号, 一般策略记为$\pi = (\pi_t, t \geq 0)$,其中π_t为时刻t的决策规则.由于π_t是随着$t \in [0, \infty)$变化的, 而决策过程又是由π诱导出来的,所以需要一些条件保证诱导出来的过程是惟一的.通常的条件是对策略类的限制,即要求策略π是确定性Lebesque可测的马氏策略,也就是说我们限制在确定性Lebesque可测的马氏策略类中讨论问题.一个确定性Lebesque可测的马氏策略定义为$\pi = (f_t, t \geq 0)$,其中$f_t \in F$为确定性决策规则, 而且对任意的行动$a \in A(i)$,集合$\{t|f_t(i) = a\}$是一个Lebesque可测的集合,$i \in S$.全体这样的策略我们仍然称为马氏策略类,记为Π_m^d.如果一个马氏策略$\pi = (f_t, t \geq)$满足$f_t \equiv f \in F$,则称为平稳策略,仍然记为f^∞或者在不引起混淆的时候记为f.全体这样的策略记为Π_s^d或者F.类似的, 我们还可以定义Lebesque可测的随机马氏策略类Π_m和Lebesque可测的随机平稳策略类Π_s.

在受控的排队模型或者维修模型中,决策时刻通常是状态发生变化的时候.如果过程是马氏过程或者是满足正则性条件的半马氏过程时,过程状态的轨迹是片状常数的,在每一条轨迹上状态跳跃的间断点至多是可数点列.所以这些问题的一般控制策略都是属于上面考虑的策略类别的.

1.2. 连续时间MDP的决策过程与折扣准则

当给定一个策略$\pi = (f_t, t \geq 0) \in \Pi_m^d$时,根据连续时间MDP的模型定义,就确定了一个转移速率矩阵簇$(Q(f_t)|t \geq 0)$,其中$Q(f_t) = (q(j|i, f_t(i)))$为$t$时刻的转移速率矩阵.当$\pi$为平稳策略时, 即$\pi = f^\infty$, 它任意时刻的转移速率矩阵都是$Q(f)$.对于一般的策略$\pi = (\pi_t, t \geq 0)$在任意时刻$t$的随机决策规则$\pi_t$来说,转移速率矩阵定义为$Q(\pi_t) = (q(j|i, \pi_t(i)))$,其中对于状态$i, j \in S$, $q(j|i, \pi_t(i)) = \sum_{a \in S} q(j|i, a)\pi(a|i)$.

熟悉马氏过程理论的读者都知道,一般来说,在给定矩阵Q之后,不能惟一的确定一个转移概率矩阵.很自然人们要问满足什么条件的Q矩阵,能够保证由它所确定的决策过程(我们将这个过程称为Q过程)都是惟一的.我们需要一些假设条件以保证我们讨论的Q过程是惟一的.

假设6.1 对任意的状态$i, j \in S$,
1) 如果$i \neq j$且$a \in A(i)$,则必有$q(j|i, a) \geq 0$;
2) 我们还有$\sum_{j \in S} q(j|i, a) = 0$.

假设6.2 存在常数M_2满足$0 < M_2 < \infty$,对任意的$i, j \in S$和$a \in A(i)$有

$$|q(j|i, a)| \leq M_2. \tag{1}$$

假设条件6.1通常被成为保守性条件,满足这个条件的Q矩阵也被称为是保守的. 而条件6.2则被称为Q矩阵的一致有界性条件.在这两个条件的保证下, 根据文献[35],我们知道:对任意平稳策略$\pi = f^{\infty}$, $Q(f)$惟一的确定了一个标准随机的平稳转移概率函数矩阵$P(t, f)$,它是柯尔莫哥洛夫向前微分方程

$$\frac{\partial P(t,f)}{\partial t} = P(t,f)Q(f), \tag{2}$$

和柯尔莫哥洛夫向后微分方程

$$\frac{\partial P(t,f)}{\partial t} = Q(f)P(t,f), \tag{3}$$

的惟一有界解,具有$P(0, f) = I$,其中I为单位矩阵.

对一般的$\pi \in \Pi_m$来说,根据文献[95]知道在假设6.1和6.2的条件下, $(Q(\pi_t), t \geq 0)$也惟一的决定了一个标准随机的转移概率函数矩阵(一般来说是非平稳的)$\{P(s,t,\pi) = (p_{ij}(s,t,\pi))|i,j \in S, t > s \geq 0)\}$,它是柯尔莫哥洛夫向前微分方程

$$\frac{\partial P(s,t,f)}{\partial t} = P(s,t,f)Q(\pi_t), \tag{4}$$

对几乎所有的$t(> s \geq 0)$的惟一有界解,且有$P(s,s,\pi) \equiv I$,其中$p_{ij}(s,t,\pi)$表示系统在s时刻处于状态状态i,在策略π的控制下于时刻t处于状态j的概率.当$s = 0$时,我们记为$p_{ij}(t,\pi)$. 因此,在给定了状态的初始分布$\{p_i, i \in S\}$和策略$\pi \in \Pi_m$之后,就可以概率为1的确定一个马氏过程,其转移概率函数矩阵$P(s,t,\pi)$满足公式(4).这个过程我们记作$\{Y_t(\pi), t \geq 0\}$,其中$Y_t(\pi)$表示在策略π 下系统在时间t所处的状态,被称为由策略π产生的状态过程.而$(Q(\pi_t), t \geq 0)$被称为状态过程的无穷小生成算子.

就象离散时的情形那样(参见定理1.1),我们用$\Delta_t(\pi)$表示在策略π下,系统在时刻t采用的行动,那么当$\pi \in \Pi_m$时,过程$\{(Y_t(\pi), \Delta_t(\pi)), t \geq 0\}$是一个二维的马氏过程.

用策略π时,如果t时刻系统处于状态$i \in S$决策者获得报酬率为$r(i, \pi_t) = r(i, f_t(i))$.因此,在任意区间$[s, t)$ 决策者获得的报酬为

$$\int_s^t r(Y_u, \pi_u(Y_u))du.$$

使用策略$\pi = (\pi_t, t \geq 0)$从任意的状态$i \in S$出发,长期折扣总报酬为:

$$u_\alpha(i, \pi) = \int_0^\infty e^{-\alpha t} \sum_{j \in S} p_{ij}(t, \pi) r(j, \pi_t(j))dt, \tag{5}$$

其中$\alpha > 0$是折扣率因子;而对任意的$j \in S$,由随机决策规则π_t决定的报酬率$r(j, \pi_t(j))$定义为:

$$r(j, \pi_t(j)) = \sum_{a \in A(j)} r(j, a)\pi_t(a|j). \tag{6}$$

关于(5)的可积性问题,由r的有界性和策略π的Lebesque可测性保证, 详细的内容可以参见文献[93].

下令$r(\pi_t)$表示分量为$r(i,\pi_t(i))$的S上列向量,则对任意的策略$\pi = (\pi_t, t \geq 0)$, 公式(5)可以写为向量和矩阵的形式

$$u_\alpha(\pi) = \int_0^\infty e^{-\alpha t} P(t,\pi) r(\pi_t) dt. \tag{7}$$

当策略π分别为马氏策略,公式(5)和(7)中的t时刻决策规则π_t分别被替换为f_t.特别是当$\pi = f^\infty$,公式(5)和(7)分别为

$$u_\alpha(i,\pi) = \int_0^\infty e^{-\alpha t} \sum_{j\in S} p_{ij}(t,\pi) r(j, f(j)) dt, \tag{8}$$

$$u_\alpha(\pi) = \int_0^\infty e^{-\alpha t} P(t,\pi) r(f) dt. \tag{9}$$

定义6.1：如果有策略$\pi^* \in \Pi_m$,对于任意的策略$\pi \in \Pi_m$和状态$i \in S$都满足:

$$u_\alpha(i, \pi^*) \geq u_\alpha(i, \pi), \tag{10}$$

则称策略π^*关于折扣准则是最优的,简称为最优的.式(10)中$u_\alpha(i, \pi^*)$被称为折扣准则的最优值函数, 或简称为最优值函数. 对$\epsilon \geq 0$，如果策略π使得$u_\alpha(i, \pi) \geq u_\alpha(i, \pi^*) - \epsilon$对所有状态$i \in S$成立，则称$\pi$为折扣准则的$\epsilon$最优策略,简称为$\epsilon$最优策略.

1.3. 最优策略的存在性与结构

为了避免过于复杂的条件叙述,我们只是限制在确定性马氏策略类Π_m^d中考虑优化问题.

定义6.2：. 方程

$$\alpha u = \sup_{f \in F}\{r(f) + Q(f)u\}, \tag{11}$$

被称为连续时间折扣MDP的最优方程.

比较离散时刻的折扣MDP问题,方程(11)与第三章中方程(3.13)有着类似的形式,但是连续时间问题的最优方程(11)是基于转移速率和报酬率这些参数建立的. 对任意的$f \in F$,定义算子

$$T_f u = r(f) + Q(f)u. \tag{12}$$

考察平稳策略,我们有下面的定理.

定理6.1：. 任取$\epsilon > 0$, u是一有界向量.那么:

1) 对某$f \in F$,满足$\alpha u \geq T_f u - \epsilon \mathbf{1}$,则有$u \geq u_\alpha(f^\infty) - \alpha^{-1}\epsilon \mathbf{1}$;

2) 对某$f \in F$,满足$\alpha u \leq T_f u + \epsilon \mathbf{1}$,则有$u \leq u_\alpha(f^\infty) + \alpha^{-1}\epsilon \mathbf{1}$;

其中$\mathbf{1}$为每个分量都是1的列向量.进而,如果1)和2)中的条件成立严格不等式,那么结论也成立严格不等式.

证明思路: 参见文献[94],[186](第13章5.3节中定理1的证明)或者文献[187]。

定理6.1说明:如果一个向量u是算子$T_f - \epsilon \mathbf{1}(+\epsilon \mathbf{1})$的上(下)界,那么它也是$u_\alpha(f^\infty) - \alpha^{-1}\epsilon \mathbf{1}(+\alpha^{-1}\epsilon \mathbf{1})$的一个上(下)界.

定理6.2：. 令$\epsilon \geq 0$, u是一有界向量.那么:
1) 若对所有$f \in F$,满足$\alpha u \geq T_f u - \epsilon \mathbf{1}$,则对任何$\pi \in \Pi_m^d$,都有$u \geq u_\alpha(\pi) - \epsilon \mathbf{1}$;
2) 若对所有$f \in F$,满足$\alpha u \leq T_f u + \epsilon \mathbf{1}$,则对任何$\pi \in \Pi_m^d$,都有$u \leq u_\alpha(\pi) + \epsilon \mathbf{1}$.

定理6.2说明:如果一个向量u是所有算子$T_f - \epsilon \mathbf{1}(+\epsilon \mathbf{1})$的上(下)界,那么对于所有策略$\pi \in \Pi_m^d$来说,它也是$u_\alpha(\pi) - \alpha^{-1}\epsilon \mathbf{1}(+\alpha^{-1}\epsilon \mathbf{1})$的一个上(下)界.这样我们就可以得到下面的推论.

推论6.1：. 对任何$f \in F$,如果策略$\pi^* \in \Pi_m^d$的期望连续折扣总报酬向量$u_\alpha(\pi^*)$满足$\alpha u_\alpha(\pi^*) \geq T_f u_\alpha(\pi^*)$,那么策略$\pi^*$是最优的.

结合上面的结论,我们有:

定理6.3:

1) 若$\pi = f^\infty$,那么$u = u_\alpha(f^\infty)$是方程

$$\alpha u = T_f u$$

的惟一有界解.

2) 对于本章的模型,如果假设6.1和6.2 成立,那么最优方程(11)存在有界解u^*,而且满足

$$u^* = \sup_{\pi \in \Pi_m^d} u_\alpha(\pi) = \sup_{f \in F} u_\alpha(f^\infty). \tag{13}$$

因此,对任意的$\epsilon > 0$,总存在平稳ϵ最优策略.进而,如果$A(i)$都是有限集合,那么一定存在最优平稳策略.

证明思路:参见文献[94]或者[186].

在有限阶段MDP问题,折扣MDP问题和平均目标MDP问题的章节里,我们都讨论过最优策略的结构问题.连续时间MDP问题的最优策略结构比较前面各章节来说,反而简单了,这是由于连续时间模型的特点所决定的.

定理6.4：如果$\pi^* = (f_t^*, t \geq 0) \in \Pi_m^d$是最优策略,那么

1) 如果存在t_0满足$0 < t_0 < \infty$,使得在$t \in [0, t_0)$上恒有$f_t^* = f_0$,则f_0^∞对同一折扣率α也是最优策略.

2) 对于任意的固定时刻t_0, π^*也是该系统从t_0时刻开始的最优策略.

3) 对于任意的固定时刻$t_0 > 0$,有

$$u_\alpha({}^{t_0}\pi^*) = u_\alpha(\pi^*), \tag{14}$$

其中${}^{t_0}\pi^*$是由策略π^*截掉$[0, t_0)$这些决策规则构成的策略.

4) 如果存在t_0和T_0满足$0 < t_0 < T_0 < \infty$使得在区间$[t_0, T_0)$上有$f_t^* \equiv f_{t_0}$,则$f_{t_0}^\infty$对同一折扣率α也是最优策略.

证明思路:参见文献[187].

定理6.4之所以比较离散折扣模型中定理3.29要强的多,就是因为定理6.4的3)和4)在离散MDP问题中是不成立的.具体的反例可以参见文献[172]和[173].下面的结论与离散模型的结论类似.

定理6.5：随机平稳策略π_0^∞是最优策略的充分必要条件是对任意的状态$i \in S$,决策规则$\pi_0(\cdot|i)$是最优行动集合$A^*(i)$上的概率分布,其中:

$$A^*(i) \equiv \left\{ a \left| a \in A(i), r(i,a) + \sum_{j \in S} q(j|i,a) u^*(j) = \alpha u^*(i) \right. \right\}, \qquad (15)$$

而$u^* = \sup_{f \in F} u_\alpha(f^\infty)$.

证明思路:参见文献[187].

1.4. 转化为离散时间模型

类似于离散模型,我们也可以构造出策略迭代算法.但是最为方便的还是将连续时间模型转换为离散的模型求解,因为这样转换后离散模型的所有解法对连续时间模型都有效.

我们已经知道,对任意的$f \in F$,它所对应的转移速率矩阵$Q(f)$惟一的确定了一个马氏过程,其转移概率矩阵是柯尔莫哥洛夫向前和向后微分方程组的惟一解.从矩阵$Q(f)$决定的连续时间马氏过程到间断时间的马氏链的转移概率矩阵变换如下:令

$$\tilde{P}(f) = \lambda^{-1}Q(f) + I, \tag{16}$$

其中I为单位矩阵,而λ定义为

$$0 < \lambda = \sup_{i \in S,\ a \in A(i)} -q(i|i,a) \leq M_2. \tag{17}$$

读者很容易检查,矩阵$\tilde{P}(f)$是一个行和为1且每个元素都非负的转移概率矩阵.我们用$\tilde{p}(j|i, f(i))$ 表示矩阵$\tilde{P}(f)$的第(i,j)个元素.

如果我们定义一个离散的MDP模型$\{T, S, A(i), \tilde{p}(\cdot|i,a), r(i,a)\}$,其中$T = (0, 1, \cdots)$,状态空间$S$和行动集合$A(i)$与连续时间的相同,转移概率矩阵为$\tilde{P}(f)$,报酬函数就是连续时间模型的报酬率函数$r$.如果取$\beta \in (0, 1)$,我们就可以建立一个时间离散的无限阶段折扣MDP问题了.如果假设6.1和6.2成立,根据定理6.3的结论2) 知道,对任意的$\epsilon > 0$,存在关于折扣率因子α的ϵ最优平稳策略.所以我们只需要在平稳策略类考虑就足够了.

关于由连续时间MDP问题导出的离散时间MDP问题,我们有下面的结论.

定理6.6：导出的离散时间MDP问题关于折扣因子$\beta \in (0,1)$的ϵ最优平稳策略f^{∞},也是原来连续时间MDP问题关于折扣率因子$\alpha > 0$的ϵ'最优平稳策略,其中这些参数的关系为:

$$\alpha = \frac{1-\beta}{\beta}\lambda, \tag{18}$$

$$\epsilon' = \frac{\epsilon}{\beta}\lambda. \tag{19}$$

反之也成立.

证明思路:参见文献[186].

推论6.2：对所有状态$i \in S$,如果$A(i)$都是有限集合,那么,导出的离散时间MDP问题关于折扣因子$\beta \in (0,1)$的最优平稳策略f^{∞},也是原来连续时间MDP问题关于折扣率因子$\alpha > 0$的最优平稳策略,其中: α和β的关系满足公式(18).反之亦然.

定理6.6和推论6.2保证了一个连续时间MDP问题的求解可以完全通过将其转化为离散时间MDP问题后求解得到,即保证了最优值函数的转换,也保证了最优平稳策略的同一性.充分的利用离散时的结论, 我们能够考虑连续时间MDP问题的各种解法.

1.5. 适用范围的推广

上假设6.2的条件一般来说过于苛刻了.在实际当中,无论是排队模型还是库存模型一般很难估计队长允许的上界和库存量的允许上下界(负的库存量认为是缺货记录),这样一来,一致有界性条件就难以保证了. 如果放弃这个条件,文献[35]的结果不能使用,这样就会造成无法利用转移速率矩阵来惟一的确定过程了,即Q过程的惟一性难以保证.为了保证Q过程的惟一性,我们下面给出定义和需要的条件.

定义6.3：我们称策略类Π_m关于t是连续的,如果对于任意固定的策略$\pi \in \Pi_m$和状态$i, j \in S$,转移速率矩阵的第(i, j)个元素$q_{ij}(t, \pi)$作为t的函数是连续的.

假设6.3：策略类Π_m和Π_m^d都限制在连续的策略类上.

假设6.3是对两个策略类增加一个形式上的限制,即策略类Π_m和策略类Π_m^d都是连续的.我们依然用记号Π_m和Π_m^d表示满足连续性条件的这种策略类.之所以称为形式上的,是因为除了对Π_m和Π_m^d以外, 对于Π_s和Π_s^d中的策略自然满足连续性条件.而对Π_m和Π_m^d的影响也是很小的,因为Π_m和Π_m^d中有实际意义的策略都是连续的,直观上的解释就是:在决策环境和效果什么条件都不变的情况下,突然改变决策规则除了数学上的理论意义以外,一般不是一个实用的策略.

根据文献[59]的结论,在假设6.3下,对每个策略$\pi \in \Pi_m$都可以构造出最小的Q过程, 相应的转移概率矩阵函数我们依然记为$P(s,t,\pi)$.对于这个最小的Q过程,转移概率矩阵函数$P(s,t,\pi)$仅满足柯尔莫哥洛夫向前微分方程(2).但是如果策略$\pi \in \Pi_s$时,相应的转移概率矩阵函数$P(s,t,\pi)$同时满足柯尔莫哥洛夫向前和向后微分方程,即(2)和(3).

尽管如此,确定的转移概率矩阵函数$P(s,t,\pi)$对于任意时刻t来说,只保证是次随机的.为了保证Q过程的惟一性, 只需要对任意的时刻t转移概率矩阵函数$P(s,t,\pi)$是随机矩阵.因此我们还需要一个假设条件.

假设6.4：假设6.3成立,且对每个$\pi \in \Pi_m$,状态$i \in S$和时间$T > s \geq 0$,有

$$\int_s^T \sum_{j \in S} p_{ij}(t, \pi) q(j, \pi_t) dt < \infty. \tag{20}$$

定理6.7：假设6,4成立,任取$\pi \in \Pi_m$,对所有的状态$i \in S$和时间$t > s \geq 0$有$\sum_{j \in S} p_{ij}(s, t, \pi) = 1$,其中$p_{ij}(s, t, \pi)$是矩阵$P(s, t, \pi)$的第$(i, j)$个元素.
证明思路:参见文献[187].

定理6.7说明假设条件6.4也是保证Q过程惟一的一个充分条件.这样,我们前面叙述的所有结论都成立.我们这里就不再一一赘述了.关于更多的Q过程惟一性条件和讨论参见文献[179].

2 连续时间平均MDP

模型和策略的定义和连续时间折扣MDP模型的一样,只是平均准则函数定义有所区别.对于任意的策略$\pi = (\pi_t, t \geq 0) \in \Pi_m$和状态$i \in S$,平均准则函数$U(i, \pi)$定义为:

$$U(i,\pi) = \liminf_{T\to\infty} \frac{1}{T} \int_0^T \sum_{j\in S} p_{ij}(t,\pi) r(j, \pi_t(j)) dt; \tag{21}$$

或者写成向量的形式:

$$U(\pi) = \liminf_{T\to\infty} \frac{1}{T} \int_0^T P(t,\pi) r(\pi_t) dt. \tag{22}$$

定义6.4：如果有策略$\pi^* \in \Pi_m$,对于任意的策略$\pi \in \Pi_m$和所有的状态$i \in S$都满足:

$$U^*(i) \equiv U(i, \pi^*) \geq U(i, \pi), \tag{23}$$

则称策略π^*关于平均准则是最优的,简称为最优的.公式(23)中$U^*(i)$的向量形式为U^*被称为平均准则最优值函数,或简称为最优值函数. 对$\epsilon \geq 0$，如果策略π使得$U(i, \pi) \geq U^*(i) - \epsilon$对所有状态$i \in S$成立，则称$\pi$为平均准则的$\epsilon$最优策略,简称为$\epsilon$最优策略.

这一节里我们仍然认为假设6.1和6.4成立,并且报酬率函数r是有界的.除此以外, 我们还假设所有的行动集合都是有限的.根据连续时间折扣MDP的结果(定理6.3),我们知道对于任意的折扣率因子$\alpha > 0$,总存在α折扣平稳最优策略,记为f_α^*.因此,对于任意的一个序列$\{f_{\alpha_n}^*, n \geq 1\}$,总有一个$F$中的$f^*$为其极限点,不妨设对每个$i \in S$,有

$$\lim_{n \to \infty} f_{\alpha_n}^*(i) = f^*(i). \tag{24}$$

记$i_0 \in S$为一个固定的状态.对任意的$\alpha > 0$和$i \in S$,我们定义α折扣平稳最优策略值u_α^*的相对差为:

$$u_{ii_0}(\alpha) = u_\alpha(i, f_\alpha^*) - u_\alpha(i_0, f_\alpha^*). \tag{25}$$

假设6.5：对某一列单调下降趋于0的序列$\{\alpha_n\}$和一个固定的状态$i_0 \in S$,存在一个常数M满足: 对所有$n \geq 1$和所有状态$i \in S$,有$|u_{ii_0}(\alpha_n)| \leq M$.

假设6.5是说有一串单调下降趋于0的折扣率因子$\{\alpha_n\}$,对应的最优折扣报酬与固定的状态i_0上的最优报酬值的相对差被一个预先给定的常数M界住.下面的定理6.9将说明假设6.5是保证最优平稳策略存在的一个重要条件,文献[73],[95]或者[187]都做过类似的假设并且将它推广到更一般的情形. 尽管如此,假设6.5仍然是比较难以验证的一个条件,下面我们给出一个比较容易验证的充分条件. 首先我们给出一个记号.如果$f \in F$是一个平稳策略,由它诱导出来的状态变化马氏过程$\{Y_t(f), t \geq 0\}$是时齐的马氏链,我们用$\tau_{ij}(f)$表示从状态i出发,首次到达状态j的平均时间.

定理6.8：如果对某个单调下降趋于0的序列$\{\alpha_m\}$和一个固定的状态$i_0 \in S$,存在一个正的常数M,满足对一切$m \geq 1$和$i \in S$均有$\tau_{ii_0}(f_{\alpha_m}) \leq M$.那么,$\{u_{ii_0}(\alpha_m)\}$是一致有界的.

证明思路:参见文献[48],[180].

为面的定理说明:通过从任意状态$i \in S$出发首次到达状态i_0的平均时间的一致有界性这个条件能够得到相对差的一致有界性.

定理6.9：设假设所有状态的行动集合是有限的,并且假设6.4和6.5成立.我们有:

1) 存在一个常数g^*,一个有界函数u以及一个下降趋于0的序列$\{\alpha_m\}$,对于任意状态$i \in S$满足

$$g^* = \lim_{m\to\infty} \alpha_m u_{\alpha_m}(i, f^*_{\alpha_m}), \tag{26}$$

$$u(i) = \lim_{m\to\infty} u_{ii_0}(\alpha_m); \tag{27}$$

$$\begin{aligned} g^* &\le r(i, f^*(i)) + \sum_{j\in S} q(j|i, f^*(i))u(j) \\ &\le \sup_{a\in A(i)} \left\{ r(i,a) + \sum_{j\in S} q(j|i,a)u(j) \right\} \end{aligned} \tag{28}$$

其中f^*由公式(24)所定义.

2) 对任意的平稳策略$f \in F$,状态$i \in S$,时间$t \ge 0$和有界函数u,有

$$\int_0^t \sum_{k\in S} p_{ik}(s,f) \left[\sum_{j\in S} q(j|k, f(k))u(j) \right] ds$$

$$= \sum_{j\in S} \left[\int_0^t \sum_{k\in S} p_{ik}(s,f) q(j|k,f(k)) ds \right] u(j).$$

定理6.9（续）：3) 对给定的平稳策略$f \in F$,如果存在一个常数g和有界的函数u对一切状态$i \in S$满足

$$g \leq r(i, f(i)) + \sum_{j \in S} q(j|i, f(i))u(j),$$

那么对一切状态$i \in S$有$g \leq U(i, f)$.

证明思路：通过选取子列的方法,能够得到g^*的收敛结论.再利用公式(12)并结合Fatou引理(见文献[80]中定理8.3.7)得到公式(28)成立.详细证明可以参见文献[73],但需要注意到我们这里的目标函数是极大化一个下极限函数,文献[73]中是极小化一个上极限函数,所以不等号恰好相反.

公式(28)还被称为平均准则下的最优不等式,如果它成立等式,我们称其为平均准则的最优方程.

定理6.10：假设所有状态的行动集合是有限的,并且假设6.4和假设6.5成立.则存在平均准则的最优平稳策略$\pi^* = f^* \in F$,而且$U(f^*) = g^*$,其中g^*由式(26)所定义.

证明思路:参见文献[73].

上面的结论要求假设6.5成立,实际上这个条件可以被放宽,比如说可以放宽为第4章中的以某个函数为权重的上界范数有界的情况,我们这里就不再详细叙述了,有兴趣的读者可以参见文献[73]及其参考文献.

平均模型MDP问题也可以转换为离散平均MDP问题求解,具体的转换方式依然可以使用公式(16)进行. 问题等价的变成了寻求转换后平均模型MDP的最优策略和最优值函数问题.具体的手法与折扣时的一样,这里我们就不赘述了.另外,也可以利用后面处理半马氏模型的手段来处理.

3 折扣半马氏模型

离散时间MDP问题和连续时间MDP问题有着一个共同的特点:在离散时间MDP问题中,平稳策略诱导出来的马氏决策过程Ł(π)是一个二维时齐的马氏链,相应的状态过程Ł$_S(\pi)$也是一个时齐的马氏链(参见第1.3.5节); 同样,在连续时间MDP问题中, 平稳策略诱导出来的马氏决策过程Ł(π)是一个二维时齐的马氏过程,相应的状态过程Ł$_S(\pi)$也是一个时齐的马氏过程.

熟悉马氏过程的读者知道在马氏过程中,相邻的状态跳跃时间服从指数分布,这在一般的过程中是不满足的.人们研究的马氏更新理论拓广了马氏过程的不足.在这个框架下研究的MDP问题,就是半马氏模型(简记为SMDP)的理论.它适用于一般的排队模型,延迟时间不是指数分布的库存模型和维修更新模型等等.

为了刻画半马氏模型,我们用六重组

$$\{T, S, (A(i), i \in S), t, p, r\}$$

来表示.其中$T = [0, \infty)$是非负的半实轴;S是状态空间,为一个可数集合;$A(i)$是系统处于状态$i \in S$的可用行动集合,也是可数集合; $t_{ij}(\cdot|a)$是系统从状态i出发在行动a的控制下系统发生转移并且转移到状态$j \in S$的时间分布函数;$p(j|i, a)$ 则表示转移时刻发生时,系统转移到状态j的概率;r是报酬函数.这里需要说明的是:

1) 转移发生时间的分布可以依赖于转移到的状态.

2) 报酬函数r可以分为两个部分,一个是做出决策时的瞬时报酬部分$R(i, a)$,一个是停留在状态i尚未转移时的报酬率$r(i, a)$.这里的r是形式上的记号.

这里的转移概率$p(j|i,a)$可以是**次随机**的,也就是说对任意的状态$i \in S$和$a \in A(i)$只需满足条件$p(j|i,a) \geq 0$而且$\sum_{j \in S} p(j|i,a) \leq 1$.至于前面的各种模型,也允许转移概率是次随机的,而且并不影响结果的成立.如果必要的话,通常可以利用增加一个虚拟的状态,使得转移概率成为随机的.

系统的动态过程可以这样的描述:系统在某个时刻处于状态$i \in S$,决策者根据情况选取一个可用的行动$a \in A(i)$, 系统遵从一个概率分布$t_{ij}(\cdot|a)$确定状态的转移时间并且依照分布$p(j|i,a)$转移到状态j,并且获得即时报酬$R(i,a)$ 和累积报酬$t \times r(i,a)$(折扣模型和平均模型会有区别,后面会一一介绍).过程这样继续进行下去.

通过对上面SMDP模型的动态过程的观察,我们注意到SMDP与一般的MDP的差异在于两个相邻的状态跳跃间隔的不同. 为了清晰的描述这一点,我们需要重新定义SMDP的历史.记

$$h_n := (i_0, a_0, t_0, i_1, a_1, t_1, \cdots, i_{n-1}, a_{n-1}, t_{n-1}, i_n) \tag{29}$$

为一个从0时刻到第n次跳跃的历史,其中$i_k \in S$, $a_k \in A(i_k)$和$t_k \in [0, \infty)$分别为第k次跳跃系统所处的状态,采用的行动和下一次跳跃发生的时间,$k = 0, 1, \cdots, n-1$.除了历史的定义有所差别以外,策略的定义也需要特别说明.在连续时间的MDP问题里,尽管决策是任何时候都可以做的,但是为了能够处理,我们限定了只在惟一确定Q过程的类别里.实际上,决策时刻只是一些离散的点,即过程发生跳跃的时刻,这样过程的轨迹是片状连续的. 我们这里也约定,决策时刻只允许在过程发生跳跃的时刻,所以我们这里的策略具有形式$(\pi_0, \pi_1, \cdots)$, 其中第n个决策规则π_n表示了过程第n次跳跃时刻的决策方式,它也是定义在新的历史集合上的分布函数.我们依然保留一般MDP中定义的策略类的相应名称与记号,例如一般的策略类Π,平稳策略类F或者Π_s^d等等.

为了使讨论的问题有意义,必须保证在有限的时段内系统发生跳跃的次数不是无穷的,所以我们需要一个被称为**正则性条件**的假设.

假设6.6：存在$\delta > 0$和$1 > \epsilon > 0$,使得对所有的状态$i \in S$和行动$a \in A(i)$都有

$$\sum_{j \in S} p(j|i,a) t_{ij}(\delta|a) \leq 1 - \epsilon. \tag{30}$$

假设6.6的直观描述就是从任意个状态$i \in S$出发,使用任何行动$a \in A(i)$,下一次发生跳跃的时间不超过δ的概率严格小于1.假设6.6的条件有一个等价的条件,在叙述这个条件之前,我们先给出一些记号.对于$\alpha > 0$,状态$i, j \in S$以及行动$a \in A(i)$,记

$$\gamma_\alpha(i,a,j) = \int_0^\infty e^{-\alpha s} dt_{ij}(s|a), \tag{31}$$

$$\gamma_\alpha(i,a) = \sum_{j \in S} p(j|i,a)\gamma_\alpha(i,a,j), \tag{32}$$

$$\gamma_\alpha = \sup_{i \in S,\ a \in A(i)} \gamma_\alpha(i,a). \tag{33}$$

其中$\gamma_\alpha(i,a)$被称为在状态i采用行动a时一阶段期望折扣因子,这是由于α对应于连续时间MDP的折扣率因子的缘故.可以这样认为,在系统处于状态i而使用行动a时,下一个决策时刻(系统的下一个跳跃点)所获得的单位报酬仅仅折现为当前时刻的$\gamma_\alpha(i,a)$倍.通常来说,有$\gamma_\alpha \leq 1$.

假设6.7：对$\alpha > 0$,有$\gamma_\alpha < 1$.

引理6.1 1) 假设6.6和假设6.7是等价的.进而,如果假设6.6或者假设6.7 成立,则在任意有限时段内系统不会发生无穷次状态转移.
2) 对任意状态$i \in S$和行动$a \in A(i)$,如果我们记

$$\tau(i,a) \equiv \sum_{j \in S} p(j|i,a) \int_0^\infty u t_{ij}(du|a). \tag{34}$$

那么,正则性条件成立必有

$$\inf_{i \in S,\ a \in A(i)} \tau(i,a) \geq \delta\epsilon > 0. \tag{35}$$

证明思路:证明参见文献[182]中第6章引理1.1和引理1.2的证明.

本章的余下部分,总认为这两个正则性条件(假设6.7和假设6.7)之一成立.那么公式(35) 总成立,其中$\tau(i,a)$的实际含义就是系统进入状态i后,采用行动a的**条件周期长度**,也被称为在状态i 用行动a的**平均逗留时间**

半马氏模型的折扣准则 我们仍然用Y_n和Δ_n表示过程第n次跳跃进入的状态和采用的行动.令$\tau_{-1}=0$,并且τ_n表示从第n次跳跃到第$n+1$次跳跃的间隔时间($n\geq 0$,而且依赖于决策者使用的策略.为了记号简便,我们略去了记号π). 对给定的策略$\pi\in\Pi$和折扣率因子$\alpha>0$,定义

$$V_\alpha(i,\pi) \equiv E_\pi^i\left[\sum_{n=0}^{\infty} e^{-\alpha(\tau_0+\tau_1+\cdots+\tau_{n-1})}\left(R(Y_n,\Delta_n) + \int_0^{\tau_n} e^{-\alpha t} r(Y_n,\Delta_n)dt\right)\right], \tag{36}$$

为从初始状态$i\in S$出发使用策略π的期望总折扣报酬.

定义6.5：如果有策略$\pi^* \in \Pi$,对于折扣率因子$\alpha > 0$,对任意的策略$\pi \in \Pi$和所有的状态$i \in S$都满足:

$$V_\alpha^*(i) \equiv V_\alpha(i, \pi^*) \geq V_\alpha(i, \pi), \tag{37}$$

则称策略π^*是α折扣最优的,简称为最优的.公式(27)中$V_\alpha^*(i)$的向量形式为V_α^*被称为α折扣最优值函数,或简称为最优值函数. 对$\epsilon \geq 0$，如果策略π使得$V_\alpha(i, \pi) \geq V_\alpha^*(i) - \epsilon$对所有状态$i \in S$成立，则称$\pi$为$\alpha$折扣准则的$\epsilon$最优策略,简称为$\epsilon$最优策略.

类似于离散模型,寻找最优策略的范围可以从一般的策略类Π中缩减到策略类Π_m.

引理6.2：任给策略$\pi \in \Pi$和状态$i \in S$,存在策略$\tilde{\pi} = (\tilde{\pi}_0, \tilde{\pi}_1, \cdots,) \in \Pi_m$对所有的$n \geq 0, j \in S$和$a \in A(j)$,满足

$$P_\pi\{Y_n = j, \Delta_n = a | Y_0 = i\} = P_{\tilde{\pi}}\{Y_n = j, \Delta_n = a | Y_0 = i\}, \tag{38}$$

以及

$$V_\alpha(i, \pi) = V_\alpha(i, \tilde{\pi}). \tag{39}$$

证明思路:证明参见文献[182].

定理6.11：策略$\pi=(\pi_0,\pi_1,\cdots)\in\Pi_m$.对任意的初始状态$i\in S$,我们有

$$V_\alpha(i,\pi) = \sum_{a\in A(i)}\pi_0(a|i)\left[\bar{R}(i,a) + \sum_{j\in S}p(j|i,a)\int_0^\infty e^{-\alpha u}V_\alpha(j,\pi^1)dt_{ij}(u|a)\right], \quad (40)$$

其中$\pi^1\in\Pi_m$为一个随机马氏策略,由策略π去掉0时刻的决策规则π_0所构成,以及

$$\bar{R}(i,a)=R(i,a)+\sum_{j\in S}p(j|i,a)\int_0^\infty\int_0^s e^{-\alpha u}r(i,a)dudt_{ij}(s|a) \quad (41)$$

是在状态i用行动a的期望一步转移折扣报酬.

证明思路: 利用全概率公式,将$V_\alpha(i,\pi)$按照π_0和首次跳跃时间τ_0展开,可以直接得到.也可以参见文献[186]中13章6.2节定理1的证明.

定理6.11说明:在平稳的马氏策略类中,总期望报酬可以分解为一步期望报酬与首次跳跃之后到达的那些状态上用策略π^1的折扣期望总报酬的折扣期望和.对任意的状态$i, j \in S$和行动$a \in A(i)$,我们令

$$\beta_\alpha(i,a,j) = \int_0^\infty e^{-\alpha u} dt_{ij}(u|a). \tag{42}$$

公式(40) 可以写成为

$$V_\alpha(i,\pi) = \sum_{a\in A(i)} \pi_0(a|i) \left[\bar{R}(i,a) + \sum_{j\in S} p(j|i,a)\beta_\alpha(i,a,j)V_\alpha(j,\pi^1) \right]. \tag{43}$$

定理6.12：1) 对任意的$f \in F$,相应的报酬值函数$\{V_\alpha(i,f), i \in S\}$是如下方程组

$$u(i) = \bar{R}(i, f(i)) + \sum_{j \in S} p(j|i, f(i))\beta_\alpha(i, f(i), j)u(j), \quad i \in S, \tag{44}$$

的惟一有界解.

2) 如果状态和所有行动集合都是有限集合,则$\{u^*(i), i \in S\}$是最优方程

$$u^*(i) = \max_{a \in A(i)} \left[\bar{R}(i,a) + \sum_{j \in S} p(j|i,a)\beta_\alpha(i,a,j)u^*(j) \right], \quad i \in S, \tag{45}$$

的惟一有界解,而且对一切的状态$i \in S$有

$$u^*(i) = V_\alpha^*(i) = \max_{f \in F} V_\alpha(i,f). \tag{46}$$

证明思路: 参见文献[186]中13章6.2节定理2和定理3的证明.

定理6.12不仅说明平稳策略具有优势,而且还说明存在最优平稳策略.当然,定理6,12的2)要求状态和行动集合都是有限的.对于任意的$f \in F$,我们定义算子T_f为:

$$T_f u(i) = \bar{R}(i, f(i)) + \sum_{j \in S} p(j|i, f(i))\beta_\alpha(i, f(i), j)u(j), \tag{47}$$

其中$i \in S$,而u为一有界向量.那么,我们有如下的结论.

引理6.3：设u, v为任意的有界向量,$f \in F$为任意决策函数,则
1) 如果$u \leq v$,则$T_f u \leq T_f v$;
2) 总有$T_f V_\alpha(f^\infty) = V_\alpha(f^\infty)$;
3) 当$n \to \infty$时,有$T_f u \to V_\alpha(f^\infty)$.
证明思路: 证明参见文献[113].

也就是说T_f是保序的压缩算子.类似于离散时间的折扣模型,可以定义最优算子T,而且也具有离散模型的那些性质.实际上,通过公式(43)可以看到我们已经将SMDP问题转化为离散问题的形式了.但是我们注意到这时的折扣因子$\beta_\alpha(i,a,j)$依赖于当前的状态i,采用的行动a和将要转入的状态j. 尽管这与我们前面介绍的折扣模型有所区别,但是只需要对离散的折扣模型稍做修改,就可以得到适应解这个模型的所有对应算法,例如策略迭代算法,值迭代算法,改进的策略迭代算法,线性规划算法等等. 我们这里就不一一例举了,有兴趣的读者可以自行得到,或者参见文献[166].同样,最优策略的结构问题可以参见文献[172]和[176].

总之,SMDP问题在满足正则性条件(假设6.6或者假设6.7之一成立)下,按照状态跳跃造成的决策时刻将策略划分开,很自然的就可以转化为离散时间模型.针对已经转换了的问题,最优策略的存在性,相应的求解算法等等都可以在转化后的离散模型中考虑.但是有一点要注意,就是正则性条件是SMDP所特有的要求,不可以忽略这一点. 至于在状态空间和行动集合为可数的情形下的结论,可以类似的参照离散折扣模型中第3.8节附加相应的条件后得到,这里就不赘述了.

4 平均半马氏模型

我们这里讨论的SMDP模型依然是六重组$\{T, S, (A(i), i \in S), t, p, r\}$,其中所有的参数都和第6.3节的定义相同.下面我们考虑平均准则的具体定义.

平均准则在离散时间模型中的定义比较容易理解.在SMDP模型中,由于看问题的角度不同,定义的形式也不同.如果我们用Y_n和Δ_n表示过程第n次跳跃进入的状态和采用的行动.令$\tau_{-1} = 0$,并且τ_n表示从第n次跳跃到第$n+1$次跳跃的间隔时间($n \geq 0$,而且依赖于决策者使用的策略.为了记号简便,我们略去了记号π). 设$Z(t,\pi)$为时段$[0,t]$内的总报酬,而$Z_n(\pi) = R(Y_n, \Delta_n) + \tau_n r(Y_n, \Delta_n)$为从第$n$次跳跃时刚刚采取决策到第$n+1$次跳跃后尚未采取决策时刻内的总报酬.对任意的策略$\pi$,定义

$$\Phi^1(i,\pi) = \liminf_{t\to\infty} E_\pi^i \left[\frac{Z(t,\pi)}{t}\right], \tag{48}$$

以及

$$\Phi^2(i,\pi) = \liminf_{t\to\infty} \frac{E^i_\pi\left[\sum_{j=0}^n Z_j\right]}{E^i_\pi\left[\sum_{j=0}^n \tau_j\right]}. \tag{49}$$

很明显Φ^1是通常意义下的平均期望报酬.由于正则性条件成立,根据引理6.1的2)知道,公式(49) 右端的分母不会是0,所以Φ^2是有定义的,而且Φ^2是某种意义的平均期望报酬.之所以说是"某种意义"的平均期望报酬,实际上,$Z_n(\pi)$可以理解为第n个更新周期内的总报酬而τ_n为相对应的第n个更新周期的周期长度.所以, 公式(49)右端未取极限时的比值可以理解为直到第$n+1$次跳跃时刻的单位时间报酬.它与Φ^1的区别在于Φ^1是对时间一致的取平均(在同一时间内,系统的跳跃次数对于不同的样本轨迹可能有所不同),而Φ^2则是对系统发生跳跃的次数一致的取平均(对于相同的跳月次数来说,对不同的样本轨道,发生的时间一般不一致). 虽然Φ^1看起来更直接,但是Φ^2用起来更方便.在一定的条件下,在随机平稳策略类Π_s中, Φ^1和Φ^2是相等的.

对于任意的随机平稳策略$\pi \in \Pi_s$,系统诱导的状态变化过程是$\{Y_t, t \geq 0\}$.再由正则性条件知道:系统在任意的有限时间区间内状态跳跃的次数是有限的,而且每一条轨迹都是片状连续的.因此,0时刻开始,系统从某个状态出发, 我们记做Y_0,直到时刻τ_0时跳到一个新的状态Y_1,停留了时间τ_1以后再跳到了Y_2,系统在状态Y_n 停留了时间长度τ_n.这样我们就可以从连续的状态变化过程$\{Y_t, t \geq 0\}$得到了一个离散的二元过程$\{(Y_n, \tau_n)|n \geq 0\}$.对任意的状态$i \in S$,定义

$$T(i) = \inf\{t > 0|Y_t = i, Y_{t-} \neq i\}, \tag{50}$$

$$N(i) = \min\{n > 0, Y_{n+1} = i\}, \tag{51}$$

其中Y_{t-}表示时刻t点的左极限状态(注意,过程的轨迹具有左极限存在,右连续的片状常数性质).这里$T(i)$是一个随机变量,表示过程首次进入状态i的时间(注意我们的定义,就会知道不包括初始时刻的状态).而$N(i)$也是一个随机变量,表示过程首次到达状态i的状态转移次数.

引理6.4：对任意的随机平稳策略$\pi \in \Pi_m$和状态$i \in S$,由策略π诱导的系统状态变化过程$\{Y_t, t \geq 0\}$从状态i出发首次进入(返回)状态i的时间的数学期望有限,即$E_\pi^i[T(i)] < \infty$.那么,我们有$E_\pi^i[N(i)] < \infty$而且$T(i) = \sum_{n=0}^{N(i)} \tau_n$.

证明思路: 证明参见文献[186]中第13章6.3节引理4的证明或者参见文献[182]中第6章引理1.3的证明.

定理6.13：对一个平稳策略$\pi_0 \in \Pi_s$,如果$E^i_{\pi_0}[T(i)] < \infty$,那么我们有

$$\Phi^1(i,\pi_0) = \Phi^2(i,\pi_0) = \frac{E^i_{\pi_0}[Z(T(i))]}{E^i_{\pi_0}[T(i)]}. \tag{52}$$

证明思路: 参见文献[182]中6章1.3节定理1.1的证明,或者参见文献[186]中13章6.3节定理4的证明.

定理6.13说明:设π_0是一个平稳的策略,如果由它诱导出的状态过程从某个状态$i \in S$出发,首达状态i的平均时间$T(i)$是有限的.那么,不仅有$\Phi^1(i,\pi_0)$和$\Phi^2(i,\pi_0)$相等,而且一个周期的平均报酬$E^i_{\pi_0}[Z(T(i))]$除以这个周期的平均长度$E^i_{\pi_0}[T(i)]$就是该平稳策略在状态i的长期平均报酬. 如果出发的状态如果不是i而是别的状态$j \in S$,如果从状态j出发,过程以概率1最终进入状态i的话, 这个过程对于状态i来说是一个延迟的更新过程,依然有定理6.13的结论,而且从状态j出发的报酬等于从状态i出发的报酬值.如果上面的条件对于所有的$j \neq i$都成立,问题就集中在如何求解从状态i 出发的一周期平均报酬和平均返回时间了.尽管如此,一般的求解一个平均准则的SMDP问题也是困难的.

平均准则SMDP问题也可以转化为离散时间的模型,有时候求解问题会容易一些,特别是从Φ^2出发考虑问题的离散化.回忆引理6.1中公式(34)关于$\tau(i,a)$的记号,并且定义

$$\tilde{R}(i,a) = R(i,a) + r(i,a)\tau(i,a). \tag{53}$$

我们可以将$\tau(i,a)$理解为在状态i使用行动a直到下一次系统状态发生转移的期望时间(也称为在状态i的平均逗留时间),而$\tilde{R}(i,a)$则是在状态i的平均逗留报酬.

如果我们考虑一个特殊的问题,在状态$i \in S$采用行动$a \in A(i)$系统发生跳跃的时间间隔是一个确定的数记为$\tau(i,a)$, 获得的报酬是$\tilde{R}(i,a)$,转移到状态$j \in S$的概率是$p(j|i,a)$的话,问题自然变成了离散的了.由于跳跃时间是一个与状态和行动相关的数,这个离散模型的最优方程也有所变化.

假设6.8：

$$\sup_{i\in S,\ a\in A(i)} \tau(i,a) < \infty. \tag{54}$$

定理6.14：如果假设6.8成立,而且存在一个有界函数$\{h(i), i \in S\}$和一个常数g满足

$$h(i) = \sup_{a \in A(i)} \left\{ \tilde{R}(i,a) + \sum_{j \in S} p(j|i,a)h(j) - g\tau(i,a) \right\}, \quad i \in S. \tag{55}$$

则存在平稳策略$f^{*\infty}$满足

$$g = \Phi^2_{f^{*\infty}}(i) = \max_{\pi \in \Pi} \Phi^2_{\pi}, \quad i \in S, \tag{56}$$

而且f^*在每个状态上采取的行动是由使公式(55)右端达到最大的行动构成.

证明思路: 结合文献[145]和文献[182]第6章中定理2.3的证明即得.

请读者注意到此时的最优方程(55)与离散的平均准则模型的最优方程的差异在于常数g有一个系数$\tau(i,a)$.实际上这并不难理解,原因在于系统在状态$i \in S$上采用行动$a \in A(i)$时的平均逗留时间是$\tau(i,a)$而不是1的缘故.

针对离散化后的SMDP问题,有很多学者给出了一些有效的算法,我们建议读者参见文献[135],[166]或者[169].其实读者不难根据第4章的算法直接得到.

5 服务率受控的一个排队模型

考虑一个$M/G/1$排队模型.这是具有一个服务员的系统,系统每次到达一个顾客,而且相邻的到达时间间隔是独立同参数的指数分布.顾客在系统里接受服务的服务时间也与顾客到达时间相互独立,而且是一个一般的分布.决策者可以控制服务的速率.当然,一般来说服务的速度越快,费用会越高.

我们假设顾客到达时间分布是参数γ的指数分布.服务时间的分布函数是$G_b(\cdot)$,而且具有密度函数$g_b(\cdot)$,其中参数b是属于一个有限的集合B.我们认为只有在完成一个服务之后或者在一个顾客到达时系统刚刚不空时,决策者才可以改变参数b,以获得其他的服务速度.决策者考虑顾客等待的费用$r_1(i)$,其中i是系统中的顾客数;连续的服务费用$r_2(b)$以及改变服务速度的固定费用K.

状态用一个数对(i,b)表示,其中i表示系统中的顾客数,b表示决策者使用的服务分布参数.我们用SMDP的语言来描述这个过程.

状态空间为$S=\{0,1,\cdots\}\times B$或者是$S=\{0,1,\cdots,M\}\times B$,其中前者表示系统允许有无穷个顾客等待,而后者则表示系统的等待空间是有限的.每个状态$(i,b)\in S$的可用行动集合$A(i,b)=B$.

在这个问题里,系统在任何一个状态上的停留时间分布不仅依赖于系统的状态,而且依赖于决策者所采用的行动.如果$i \geq 1$时,下一个决策时刻(可以改变服务速度的时刻)是恰好服务结束一个顾客的时刻,所以它服从分布

$$t(u|(i,b),b') = G_{b'}(u).$$

如果$i = 0$,下一个决策时刻就是一个顾客到达的时刻,于是它服从分布

$$t(u|(0,b),b') = 1 - e^{-\gamma u}.$$

为了满足正则性条件,我们需要条件:存在$\epsilon > 0$和$\delta > 0$,对所有的$b \in B$满足$G_b(\delta) < 1 - \epsilon$. 由于我们假设$B$是有限的集合,正则性条件就等价于:对一切$b \in B$,有$G_b(0) < 1$.

下面我们对$M = \infty$的情况分析转移概率.当系统空闲的时候,下一步转移是发生在一个顾客到达的时候,所以

$$p((1, b')|u, (0, b), b') = 1.$$

同样,对于$i > 0$,

$$p((i + k, b')|u, (i, b), b') = \frac{e^{-\gamma u}(\gamma u)^k}{k!}.$$

注意,这里的转移概率是依赖于时间u的,我们前面介绍的模型中为了简便,将转移时间分布和转移概率分开了,而且假设转移概率不依赖于时间.我们这里的例子不满足这个假设,但是这对我们的前面的分析没有本质影响,因为凡是涉及到对时间积分时,只是不能将转移概率拿到积分号外面来而已.至于离散化求解时,变化稍微有些复杂,但原则上也是将转移概率$p(j|t, i, a)$放到积分号里,例如在折扣模型问题中需要重写折扣因子或者在平均模型中重写为平均转移概率就行了. 这样并不改变我们需要的任何条件或者附加什么别的条件.

模型的费用结构为:改变服务水平的固定费用

$$R((i,b),b') = \begin{cases} 0 & b' = b \\ -K & b' \neq b, \end{cases}$$

以及连续费用率

$$r((i,b),b',(j,b')) = -r_1(j) - r_2(b').$$

这样我们就完成了这个模型的基本定义.

谢谢大家！