排队论

11-倪润博　叶昊然　李方圻

目录

1 导言

排队理论是一种眼界顾客以某种随机方式到达一个服务设施的模型，它以概率论和随机过程理论为基本知识。在介绍排队理论之前，我们首先简要介绍随机过程的有关知识。

一个随机过程是一个随机变量的集合$\{X(t), t \in T\}$。即对于每一个$t \in T$，$X(t)$是一个随机变量。对于随机过程的一个朴素的阐释是t表示时间，而$X(t)$表示系统在t时刻的状态。

我们会介绍一些常用的的随机过程实例——马尔可夫过程、泊松过程、生灭过程。其中生灭过程是马尔可夫链性质和泊松性质兼具的一种随机过程，也是排队论模型中分析性质较好的一种原型。

最后我们介绍排队理论的其他一些推广。

这片介绍的大部分内容来自SM.Ross的Introduction to Probability Models，该书作者在前言里提到该书的第5、8章和第4、6章的少许内容足以开设一个排队论的引论课程。本篇文章涵盖了该书这四章的大部分核心知识点，但是在理论框架间的推理思路有所不同。

2 马尔可夫过程

2.1 定义

考虑这样一个随机过程$X(1), X(2), X(3), ...X(N)$，这样一个随机过程可以用所有随机变量的联合分布：

$$p(X(1), X(2), X(3), ...X(N))$$

来描述。

利用贝叶斯公式展开联合分布：

$$p(X(1), X(2), X(3), ...X(N)) =$$

$$p(X(1))p(X(2)|X(1))p(X(3)|X(1), X(2))...p(X(N)|X(1), X(2), ...X(N-1))$$

容易发现对于靠后的项，条件分布因为条件空间的过于庞大而难以描述和计算。马尔可夫假设就是为了简化这种情形：

$$p(X(n)|X(1), X(2), ...X(n-1)) = p(X(n)|X(n-1))$$

此时联合分布为：

$$p(X(1)) \prod_{n=2}^{N} p(X(n)|X(n-1))$$

除了上式，还有其他同等的方式来表述马尔可夫过程：

当前的状态仅与前一个时间的状态有关；

给定某个时间的状态，下一个时间的状态和之前所有时间的状态无关；

此随机过程的概率图模型骨架是一条线（所以被称作马尔可夫链），所有随机变量按照时间顺序排列在这条线上。

在这篇介绍里，我们只考虑在时域和状态空间都离散的马尔可夫过程，即X仅能从有限的状态空间S中取值，而时域是随下标被分割的。

则决定一个马尔可夫链的要素是首个元素的生成概率$p(X(1))$和状态的迁移概率$p(X(n)|X(n-1))$，$p(X(1))$是一个维度为$|S|$的概率向量，而$p(X(n)|X(n-1))$可以由一个$|S| * |S|$的矩阵来唯一描述。我们记这个状态转移矩阵为$\mathbf{P}$，$\mathbf{P}_{ij} = p(S_j|S_i)$，则$\mathbf{P}$的每一行为一个概率向量。

高阶的马尔可夫链是指当前状态和有限个先前状态有关联的情况，譬如考察天气的时候，可能不仅要考虑前一天。此时可以通过扩充状态空间S为S^k（定义状态空间的乘积为笛卡尔积，其中k为此时的马尔科夫链阶数，也即条件概率分布中的条件数量）将高阶马尔可夫链转化为一阶马尔可夫链，但是这会导致状态转移矩阵的非对角部分有很多的零项。

譬如记两种天气为C, S，取链阶数为3，则状态空间为$\{C, S\}^3$，但是很明显$p(CSS|SSS)$等均为零（仅仅$p(SSS|SSS)$和$p(SSC|SSS)$非零且和为一），可以看出矩阵空间有所浪费。

系统的初状态$p(X(1))$记为一个行向量p_1，则经过第一次状态迁移以后，$X(2)$的状态由概率向量$p_2 = p_1\mathbf{P}$描述。可以想象一个仅第一位为1，其他位为0的p_1，则p_2是$\mathbf{P}$的第一行。这一个实例的意义是：给定$X(1) = S_1$时，$X(2)$的分布为$(P_{1,1}, P_{1,2}, ...P_{1,|S|})$。

考虑进行了多次状态转移后的迁移概率：

$$P_{ij}^n = p(X(a+n) = S_j | X(a) = S_i)$$

它可以被递归地计算，首先我们有：

$$P_{ij}^{n+m} = \sum_{k \in S} P_{ik}^n P_{kj}^m$$

它本质上就是贝叶斯全概率公式（我们对于长度$n+m$中的一个步骤的所有可能性进行了求和），被称作C-K方程（查普曼-科尔莫戈罗夫方程）。

它的矩阵形式为：

$$\mathbf{P}^{m+n} = \mathbf{P}^n \mathbf{P}^m$$

所以在迁移了n步以后的概率向量为：

$$p_n = p_0 \mathbf{P}^n$$

2.2 性质

2.2.1 互通和可约性

马尔可夫链中的性质和推论远远比以下罗列的要多，这里我们仅罗列足够的性质以进行下一节例题的推导，更多的性质和理论可以在相关参考书籍文献上找到。

对于状态S_i，如果存在n使得P_{ij}^n不等于零，那么就称状态S_i到状态S_j是可达的。可达的物理意义是：如果初始状态处于S_i的可能性不为零，那么在系统演进的过程中，状态为S_j的可能性不为零。也即“可能从状态S_i迁移到S_j”。两个互相可达的状态是互通的，显然：

1、S_i和自己是互通的，因为$P_{ii}^0 = 1 \neq 0$；

2、如果S_i和S_j互通，则S_j和S_i互通；

3、如果S_i和S_j互通，如果S_j和S_k互通，则S_i和S_k互通；不妨设$P_{ij}^m \neq 0, P_{jk}^n \neq 0$，则$P_{ik}^{m+n} = \sum_{p\in S} P_{ip}^m P_{pk}^n \geq P_{ij}^m P_{jk}^n \neq 0$。

所以互通在状态空间上定义了一个等价关系，继而将状态空间划分成了等价类。整个状态空间是一个等价类的马尔可夫链被称为不可约的。

可约的马尔可夫链进而可能有两种结构，即等价类之间的关系有两种：独立的和单向连接的。独立的等价类可以被直接分为两个马尔可夫链分别处理，单向连接的两个等价类最终会退化为一个，可以直观地进行理解。

2.2.2 常返性

记f_i为从状态S_i有朝一日能够转移回状态S_i的概率，如果$f_i = 1$，那么理论上在马尔可夫链无限的转移过程中，系统将无数次地回到状态S_i，此时称此状态为常返态，否则为暂态。我们设置指示变量$I(t)$，当$X(t) = 1$时$I(t)$为1否则为0，则在马尔可夫链演进的无限过程中进入状态S_i的期望次数为：

$$E[\sum_{n=1}^{\infty} I(t)|X(1) = S_i]$$

而：

$$E[I(t)|X(1) = S_i] = p(X(t) = S_i|X(1) = S_i)$$

所以：

$$E[\sum_{n=1}^{\infty} I(t)|X(1) = S_0] = \sum_{n=1}^{\infty} E[I(t)|X(1) = S_i] = \sum_{n=1}^{\infty} P_{ii}^n$$

根据上述文字的推论，状态S_i常返等价于：

$$\sum_{n=1}^{\infty} P_{ii}^{n} = \infty$$

在可约且状态有限的马尔可夫链中，对于任何非常返状态S_j，可以找到N，使得迁移次数大于N的马尔可夫链再也不返回S_j。因此关于常返性质的讨论可以使得我们在较长的时间下把注意力集中在常返状态而不是所有状态上。

2.2.3 长程性

现在考虑从状态S_j返回自身的期望次数：

$$N_j = min\left\{n \geq 0 : X(n) = S_j\right\}$$

$$m_j = E[N_j | X(1) = S_j]$$

我们考虑在状态转移无穷多步时处于状态S_j的时间，也即此状态的出现次数，我们可以认为状态以如下方式转移，其中每一个箭头表示一次从状态S_j自回归的过程，用时为T次转移： $S_j \to S_j \to S_j \to ... \to S_j$

则对于S_j出现n次，状态S_j占比：

$$\pi_j = \lim_{n\to\infty} \frac{n}{n + \sum_{i=1}^{n-1} T_i}$$

$$= \lim_{n\to\infty} \frac{1}{\frac{1}{n}\sum_{i=1}^{n-1} T_i}$$

$$= \frac{1}{m_j}$$

最后一个等式使用了大数定律。

在渐进意义上，π_i是平均意义上S_i出现的概率，也可以认为是平稳情形下（即各时间的状态的概率向量不变以后）状态处于S_i的概率。而$\pi_i \mathbf{P}_{ij}$则是从状态S_i迁移到S_j的概率，可以在平稳清醒下对前一个状态求和：

$$\pi_j = \sum_{i\in S} \pi_i \mathbf{P}_{ij}$$

以矩阵形式表达，整个演化系统的平稳概率向量π满足：

$$\pi = \pi \mathbf{P}$$

2.2.4 可逆性

从相反的时间序列观察一个马尔可夫链得到的依然是一个马尔可夫链，这可以通过观察它作为一个马尔可夫场的结构得出。对于一个有平稳态的马尔可夫链来说，我们以$\mathbf{Q}$记反向的状态迁移矩阵，则：

$$Q_{ij} = p(X(n-1) = S_j | X(n) = S_i) = \frac{p(X(n-1) = S_j, X(n) = S_i)}{p(X(n) = S_i)}$$

$$= \frac{p(X(n-1) = S_j)p(X(n) = S_i | X(n-1) = S_j)}{p(X(n) = S_i)}$$

$$= \frac{\pi_j \mathbf{P}_{ji}}{\pi_i}$$

如果要求一个对称的逆转$\mathbf{Q}_{ij} = \mathbf{P}_{ij}$，则：

$$\mathbf{P}_{ij}\pi_i = \mathbf{P}_{ji}\pi_j$$

一个满足上式（细致平稳条件）的解的存在是系统平稳的充分条件，譬如我们要寻找解$\mathbf{x}$使得对于任何i, j：

$$x_i \mathbf{P}_{ij} = x_j \mathbf{P}_{ji}$$

$$\sum_i x_i = 1$$

则有：

$$\sum_i x_i \mathbf{P}_{ij} = \sum_i x_j \mathbf{P}_{ji} = x_j \sum_i \mathbf{P}_{ji} = x_j$$

满足为求解唯一的π列出的方程，所以充分性得证。

2.2.5 例1：升级问题

原题表述：
英雄升级，从0级到1级，概率100
从1级升2级，有$\frac{1}{3}$几率成功，$\frac{1}{3}$停留原等级，$\frac{1}{3}$下降到0级；
从2级升3级，有$\frac{1}{9}$几率成功，$\frac{4}{9}$停留原等级，$\frac{4}{9}$下降到1级；
求英雄从0级升到3级平均需要多少次？
解答：

这个问题的状态转移矩阵为：

$$P=\begin{pmatrix}0 & 1 & 0 & 0\\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0\\ 0 & \frac{4}{9} & \frac{4}{9} & \frac{1}{9}\\ 0 & 0 & 0 & 1\end{pmatrix}$$

所求的期望次数为：

$$E(X)=\sum_{n=0}^{\infty}nP(X=n)=\sum_{n=1}^{\infty}P(X>n)$$

第n次升级时，英雄的等级情况由概率向量$(1,0,0,0)P^n$给出，而$P(X<n)$则由这个向量的前三个分量之和给出，所以：

$$E(X)=\sum_{n=0}^{\infty}(1,0,0,0)P^n(1,1,1,0)^T=(1,0,0,0)\left\{\sum_{n=0}^{\infty}P^n\right\}(1,1,1,0)^T$$

因为P的特征值绝对值均小于1，所以：

$$\sum_{n=0}^{\infty}P^n=(1-P)^{-1}$$

最终算出答案为30。

【另：本题和维基百科中stochastic matrix词条最后举出的猫抓老鼠的例子是一致的。】

2.2.6 例2：来自知乎“有哪些类似于七桥问题的有趣数学定理”——醉汉定理

原题表述：

一个喝醉了的醉汉在一条道路上随机走动，以0.5的概率前进一步或者后退一步，但是他最终一定能无数次返回其出发点，这一结论在醉汉在一个二维网格上走动时依旧成立，但是一个喝醉了的鸟随机在三维空间的六个方向游走时，它最终一定无法返回原出发点。

解答：

这个定理的实质是证明对于给定的马尔可夫链常返态是否存在，需要注意限定走动空间无限的前提。（如果走动空间有限，由互通关系划分等价类加上常反性在等价类中一致加上状态空间有限可以直接退出所有状态都是常返的）

在二维的情况下，我们记向一个指定方向走动的概率为p，在给定n（n为偶数）时，醉汉在第n步返回原点的概率：

$$P_{00}^{n} = C_{n}^{\frac{n}{2}} p^{\frac{n}{2}} (1-p)^{\frac{n}{2}}$$

接下来判定此状态的常反性（已代换略去为零的项）：

$$\sum_{n=0}^{\infty} P_{00}^{n} = \frac{(2n)!}{n!n!}(p-p^2)^n$$

斯特林近似后：

$$n! \to n^{\frac{n+1}{2}} e^{-n} \sqrt{2\pi}$$

原合式等于：

$$\sum_{n=0}^{\infty} \frac{(4p(1-p))^n}{\sqrt{\pi n}}$$

易证此和式在p不等于$\frac{1}{2}$时均收敛，即当且仅当$p = \frac{1}{2}$时，原点对于醉汉的随机行走来说是常返态。

在二维的网格情况下，我们类似地计算，为了签单起见我们假设四个方向的游走概率均为$\frac{1}{4}$：

$$P_{00}^{2n} = \sum_{k=0}^{n} C_{2n}^{2k} C_{2k}^{k} C_{2n-2k}^{n-k} (\frac{1}{4})^{2n} = (\frac{1}{4})^{2n} C_{2n}^{n} \sum_{k=0}^{n} C_{n}^{k} C_{n}^{n-k} = (\frac{1}{4})^{2n} C_{2n}^{n} C_{2n}^{n}$$

近似后得到此时和式的通项：

$$P_{00}^{2n} \to \frac{1}{\pi n}$$

在三维空间中，组合数性质给出：

$$P_{00}^{2n} = (\frac{1}{8})^{2n} C_{2n}^{n} \sum_{t=0}^{n} C_{n}^{t} C_{n}^{n-t} C_{2t}^{t}$$

所以：

$$P_{00}^{2n} \leq (\frac{1}{8})^{2n} (C_{2n}^{n}) \to \frac{1}{(\pi n)^{1.5}}$$

对于更高维的情况可以类似地求解，不过级数都是收敛的。

3 生灭过程

3.1 指数分布

3.1.1 无记忆性

如果试图构造一个“没有记忆”的分布，满足：

$$P(X > t+s|X > t) = P(X > s)$$

得到：

$$P(X > t+s|X > t) = \frac{P(X > t+s)}{P(X > t)} = P(X > s)$$

$$P(X > t+s) = P(X > t)P(X > s)$$

若记：

$$g(x) = P(X > x)$$

可以证明$g(x) = e^{\lambda x}$是唯一符合$g(t)g(s) = g(t+s)$的右连续函数，因为：

$$g(\frac{m}{n}) = (g(1))^{\frac{1}{n}}$$

$$g(x) = (g(1))^x = e^{-\lambda x}$$

对于$x \geq 0$，指数分布的累积分布函数为：

$$F(x) = P(X < x) = 1 - e^{-\lambda x}$$

而$P(X > x) = e^{-\lambda x}$
密度函数为：

$$f(x) = \lambda e^{-\lambda x}$$

生成函数为：

$$\phi(t) = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx = \frac{\lambda}{\lambda - t}$$

从中求出其一、二阶矩：

$$E(X) = \phi'(0) = \frac{1}{\lambda}$$

$$E(X^2) = \phi''(0) = \frac{2}{\lambda^2}$$

可得方差：

$$var(X) = \frac{1}{\lambda^2}$$

指数分布也是指数分布族中最基础的成员，也是限定均值情况下熵最大的分布，不过在本文的讨论范围内我们主要关注它的无记忆性质。

指数分布可以看做伽马分布的一个特例($\alpha = 1$)，伽马分布有两个参数λ和α，其分布函数为($x \geq 0$)：

$$f(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}$$

从指数分布族的角度可以看出两者最根本的区别：指数分布的充分统计量为X，故只有一个参数，而伽马分布的充分统计量为X和$\ln X$，所以有两个参数，将其中的一个参数置为0则退化为指数分布。

3.1.2 指数分布的复合效果

我们考虑以两种方法将多个独立同分布的指数分布变量结合为新的随机变量，一个是加法算符，一个是min算符，在这一节中我们考虑n个iid的指数分布随机变量，参数为λ。

我们首先处理服从指数分布的随机变量的和，因为答案可以通过单纯的归纳法检验，所以我们不加证明地给出结论：n个均值均为$\frac{1}{\lambda}$的指数分布随机变量的和是一个伽马分布，其参数为($\alpha = n, \lambda$)。

更一般地，我们给出独立伽马随机变量X(α, λ)和Y(β, λ)的和的分布，记$U = X + Y$，$V = \frac{X}{X+Y}$。利用随机变量的函数联合分布的结论：

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1, x_2)|J|^{-1}$$

首先写出X和Y的联合分布：

$$f_{X_1,X_2}(x, y) = \frac{\lambda^{\alpha+\beta} e^{(-\lambda(x+y))} x^{\alpha-1} y^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}$$

计算雅克比矩阵：

$$J = \left| \begin{pmatrix} \frac{\partial U}{\partial X} & \frac{\partial U}{\partial Y} \\ \frac{\partial V}{\partial X} & \frac{\partial V}{\partial Y} \end{pmatrix} \right| = \left| \begin{pmatrix} 1 & 1 \\ \frac{y}{(x+y)^2} & \frac{-x}{(x+y)^2} \end{pmatrix} \right| = -\frac{1}{x+y}$$

因为$u = x + y, v = \frac{x}{x+y}$解出$x = uv, y = u(1-v)$：

$$f_{U,V}(u,v) = [f_{X,Y}(uv, u(1-v))]u$$

$$= \frac{\lambda e^{-\lambda u}(\lambda u)^{\alpha+\beta-1}}{\Gamma(\alpha+\beta)} \frac{v^{\alpha-1}(1-v)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

所以U和V是相互独立的，且$U = X + Y$服从参数为$(\alpha+\beta, \lambda)$的伽马分布。将指数分布视作伽马分布的特例并应用递归地上述结果可以得到之前给出的关于和的结论。

接下来我们探讨用min算符复合的指数分布变量，设X和Y分别是均值为$\frac{1}{\lambda}$和$\frac{1}{\mu}$的指数分布随机变量，我们试图求：

$$Z = min(X, Y)$$

服从的分布，显然的：

$$P(Z \geq z) = P(X \geq z)P(Y \geq z) = e^{-\lambda z}e^{-\mu z} = e^{-(\lambda+\mu)z}$$

所以$min(X,Y)$服从均值为$\frac{1}{\lambda+\mu}$的指数分布。

递归地调用这个结论，我们有：对于服从均值为$\lambda_1^{-1}, \lambda_2^{-1}, ...\lambda_n^{-1}$的随机变量$X_1, X_2, ...X_n$，$Z = min(X_1, X_2, ...X_n)$服从均值为$(\sum_{i=1}^{n} \lambda_i)^{-1}$的指数分布。

保持X和Y服从上述分布，我们还可以计算出$P(X \leq Y)$的简单封闭形式：

$$P(X \leq Y) = \int_0^\infty P(X = x)P(Y \geq x)dx = \int_0^\infty \lambda e^{-\lambda x}e^{-\mu x}dx$$

$$= \frac{\lambda}{\lambda+\mu}$$

还可以计算给定条件$X \leq Y$时X的条件分布：

$$P(X = x|X \leq Y) = \frac{P(X = x)P(x \leq Y)}{P(X \leq Y)}$$

$$= \frac{\lambda+\mu}{\lambda}\lambda e^{(-\lambda x)} \int_x^\infty P(Y = y)dY = (\lambda+\mu)e^{-(\lambda+\mu)x}$$

所以已知$X \leq Y$时，X的条件分布为参数为$\lambda+\mu$的指数分布，这也和两个变量的较小者的参数为两个变量参数之和的结论一致。

3.1.3 例3：任务分配问题的贪心策略评估

考虑一个规划问题，我们需要将N个任务一对一地分配给N个人，第i个人进行第j个任务的代价由$C_{i,j}$表示，我们试图找到一个任务分配方案使得总和$\sum C$最小。现在考察以下两个贪婪策略：

策略A：对于第一个人，选择使其工作代价最小的任务k分配给他，即对于任何有效的$i, C_{1,i} \geq C_{1,k}$。并从余下所有人的可选任务队列中删除任务k，递归地进行这个过程。

策略B：对于一共n^2个代价，选择一个全局最小值$C_{m,n}$，并将任务n分配给人m，并从任务队列和人物队列中删除m和n，递归地进行这个过程。

现在假设所有的$C_{i,j}$是服从参数为λ的指数分布的独立变量，我们将证明此时两个策略的期望代价是相等的，以C_i表示策略中第i个人的实际工作代价。

对于策略A，C_1是N个iid指数分布变量中的最小者，它服从参数为$N\lambda$的分布；C_2是$N-1$个iid指数分布变量中的最小者，它服从参数为$(-1)N\lambda$的分布，以此类推：

$$E(C) = E(\sum_{n=1}^{N} C_n) = \sum_{n=1}^{N} E(C_n) = \frac{1}{\lambda}\sum_{n=1}^{N}\frac{1}{n}$$

对于策略B，我们不是一般地按照选取顺序重排人的编号使得第n次分配任务给人n，此时C_1服从参数为$N^2\lambda$的指数分布；给定$C_1 = c$时，$C_{i,j} - c$按照无记忆性仍然是参数为λ的指数分布，所以$C_2 - E(C_1)$服从参数为$(n-1)^2\lambda$的指数分布，以此类推：

$$E(C) = \sum_{n=1}^{N} E(C_n) = \frac{1}{N^2\lambda} + (\frac{1}{N^2\lambda} + \frac{1}{(N-1)^2\lambda}) + ... = \frac{1}{\lambda}\sum_{n=1}^{N}\frac{1}{n}$$

3.1.4 例4：银行服务时间

假设你到达银行，两个柜台均处于被使用状态，其完成服务的时间分别服从参数为λ_1, λ_2的指数分布，你进入首先结束服务的柜台接受服务并离开，求你在银行停留的时间期望：

以S_1和S_2分别记当前两个柜台结束服务的时间，t为你在银行停留的总时间，则：

$$E(t) = E(t|S_1 \leq S_2)P(S_1 \leq S_2) + E(t|S_2 \leq S_1)P(S_2 \leq S_1)$$

$$= E(t|S_1 \leq S_2)\frac{\lambda_1}{\lambda_1 + \lambda_2} + E(t|S_2 \leq S_1)\frac{\lambda_2}{\lambda_1 + \lambda_2}$$

而：

$$E(t|S_1 \leq S_2) = E(S_1 + s|S_1 \leq S_2) = E(S_1|S_1 \leq S_2) + E(s|S_1 \leq S_2)$$

$$= \frac{1}{\lambda_1 + \lambda_2} + \frac{1}{\lambda_1}$$

所以：

$$E(t) = \frac{3}{\lambda_1 + \lambda_2}$$

3.2 泊松过程

3.2.1 泊松过程的定义

作为重要随机过程中一员的泊松过程有多种定义，这一节罗列三种定义并说明它们的等价性。利用上一章已有的知识，我们假设有一系列服从参数为λ指数分布的独立随机变量，其取值为$T_1, T_2, ...T_n...$，记$S_n = \sum_{i=1}^{n} T_i, S_0 = 0$，此时形式化地定义泊松分布为一个随机过程：

定义1：

$$\{N(t), t \geq 0\}$$

$$N(t) = max\{n : S_n \leq t\}$$

分布函数可以直截了当地写出：

$$P(N(t) = n_0) = \int_0^{\infty} P(N(t) = n_0, S_{n_0} = T)dT$$

$$= \int_0^{\infty} P(S_{n_0} = T)I[S_{n_0} \leq t]e^{-\lambda(t-S_{n_0})}dT$$

其中积分内第二项是一个示性函数，第三项用来维系最大性，根据上一节的结果，S_n是一个参数为(λ, n)的伽马分布，所以：

$$P(N(t) = n_0) = \int_0^{t} e^{-\lambda(t-T)} \frac{\lambda e^{-\lambda T}(\lambda T)^{n-1}}{(n-1)!}dT$$

$$= \frac{\lambda^n e^{-\lambda t}}{(n-1)!} \int_0^{t} T^{n-1}dT = e^{-\lambda t}\frac{(\lambda t)^n}{n!}$$

根据定义这是一个参数为λt的泊松分布，泊松分布定义为：对于取值为非负整数的随机变量X，对于$\lambda \geq 0$，有：

$$P(X = i) = e^{-\lambda}\frac{\lambda^i}{i!}$$

我们这里不证明泊松分布的数学性质，它们大都可以通过对其生成函数$\phi(t) = e^{\lambda(e^t-1)}$的分析得到。

泊松分布的第二种定义是一个满足如下条件的随机过程：

定义2：

1、$N(t) \geq 0$；

2、$N(t)$取整数值；

3、若$s \leq t$，则$N(s) \leq N(t)$；

（满足以上三个条件的随机过程称为一个计数过程）

4、在两个互斥（不重迭）的区间内所发生的事件的数目是互相独立的随机变量。

5、在区间$[t, t+s]$内发生事件的数目服从参数为λs的泊松分布。

以上的4、5两点分别可以从我们之前的第一个定义中推出，第4点是因为我们假设的到来元素服从无记忆的、独立的指数分布，第5点可以考察独立泊松变量的和的参数，由其生成函数的形式可以推出成立。

泊松分布的第三种定义是一个满足如下条件的计数过程：

定义3：

1、$N(0) = 0$;

2、其拥有平稳增量（即区间$(s, s+t)$间事件数量和s无关）和独立增量（即上一个定义的第4条）；

3、$P(N(t+h) - N(t) = 1) = \lambda h + o(h)$；

4、$P(N(t+h) - N(t) \geq 2) = o(h)$；

我们将从上述定义导出$N(t)$的分布函数来结束本节。

固定$v \geq 0$，构造：

$$g(t) = E(e^{-vN(t)})$$

$$g(t+h) = E(e^{-vN(t+h)}) = E(e^{-vN(t)} e^{-v(N(t+h)-N(t))})$$

$$= E(e^{-vN(t)}) E(e^{-v(N(t+h)-N(t))})$$

$$= g(t) E(e^{-v(N(t+h)-N(t))})$$

对于$N(h)$取条件：

$$E(e^{-v(N(t+h)-N(t))}) = (1 - \lambda h + o(h)) + e^{-v}(\lambda h + o(h)) + o(h)$$

$$= 1 - \lambda h + e^{-v} \lambda h + o(h)$$

继而：

$$\frac{g(t+h) - g(t)}{h} = g(t)\lambda(e^{-v} - 1) + \frac{o(h)}{h}$$

取极限$h \to 0$得到：

$$\frac{g'(t)}{g(t)} = \lambda(e^{-v} - 1) = \frac{d}{dt} \ln g(t)$$

$$\ln g(t) = \lambda(e^{-v} - 1)t + C$$

令$g(0) = 0$得到$C = 0$，我们最终得到的是$N(t)$的拉普拉斯变换：

$$E(e^{-vN(t)}) = e^{\lambda t(e^{-v}-1)}$$

若以v为自变量，上式就是参数为λt的泊松分布的拉普拉斯变换。由于分布的拉普拉斯变换式唯一地确定了一个分布，所以我们得到了想要的结论。

在此处列举的泊松过程三个定义中间，第一个是最能说明其物理意义也最为简单明了的，后两个定义从不同的角度出发，在证明一些其他性质时可以参考使用。

从直观的理解上，想象人们到达一个地点的时间间隔服从一个固定参数的指数分布，则在t时该地点的人数服从一个泊松过程。因为时间间隔的均值为$\frac{1}{\lambda}$，所以从平均意义上可以认为人的到达速率为λ，λ也被称作这个泊松过程的速率。

3.2.2 泊松过程的复合

之前考虑的泊松计数过程是对于同质事件进行的，如果发生的事件以概率p归类于类型A，以$1-p$归类于类型B，我们分别以$\{N_A(t), t \geq 0\}$和$\{N_B(t), t \geq 0\}$来对两种类型的事件计数。很容易验证$N_A(t)$满足上一节第三个定义中的前两个条件，对于$N(h)$取条件可得：

$$\begin{aligned} P(N_A(h) = 1) &= P(N_A(h) = 1|N(h) = 1)P(N(h) = 1) \\ &+ P(N_A(h) = 1|N(h) > 1)P(N(h) > 1) \\ &= \lambda p h + o(h) \end{aligned}$$

$$P(N_A(h) \geq 2)) \leq P(N(h) \geq 2) = o(h)$$

所以N_A单独构成一个以λp为速率的泊松过程，同理N_B是个以$\lambda(1-p)$为速率的泊松过程，再由泊松过程的独立增量性，两个过程是相互独立的。

可以按照这一结论将多类型事件的泊松过程拆分为一系列独立的泊松过程来考察，也可以将类型合并使得多个独立泊松过程合成为一个新的单独过程，其速率为各个单独过程的速率之和。

3.2.3 到达时间的条件分布

我们考虑在给定条件$N(t)=1$时，即在$[0,t]$区间上已经发生了一次事件时，它在该区间上的分布，记其发生时间为T_1：

$$P(T_1<s|N(t)=1)=\frac{P(T_1<s,N(t)=1)}{P(N(t)=1)}=\frac{P(N(s)=1,N(t-s)=0)}{P(N(t)=1)}$$

$$=\frac{P(N(s)=1)P(N(t-s)=0)}{P(N(t)=1)}=\frac{e^{-\lambda s}\lambda s e^{-\lambda(t-s)}}{e^{-\lambda t}\lambda t}=\frac{s}{t}$$

可见此事件在的发生在$[0,t]$上均匀分布。

这一结论可以推广到$N(t)=n$，此时考察事件发生时间($s_n=s_{n-1}+T_n$)的分布：

$$f(s_1,s_2,...s_n|N(t)=n)=\frac{n!}{t^n}$$

因为我们其实附加了条件$s_1<s_2<...<s_n$，所以s在$[0,t]$区间上也是独立均匀分布的。

3.2.4 例5：到达车站的人群

到达车站的人群服从速率为λ的泊松过程，车子在时刻t出发，以X表示在t发车时所有车上的人们等待的时间，$N(t)$为t时间之前到达的人数，试求：

1、$E(X|N(t))$;

2、$var(X)$。

在给定$N(t)=n$时，人到达这一事件在$[0,t]$上独立均匀分布，所以：

$$E(X|N(t))=\frac{tN(t)}{2}$$

首先易知：

$$var(X|N(t))=N(t)var(U(0,t))=\frac{t^2N(t)}{12}$$

从条件方差公式得到：

$$var(X) = E(var(X|N(t))) + var(E(X|N(t)))$$
$$= E(\frac{t^2 N(t)}{12}) + var(\frac{tN(t)}{2}) = \frac{t^2}{12}\lambda t + \frac{t^2}{4}\lambda t = \frac{\lambda t^3}{3}$$

3.3 连续马尔可夫链

3.3.1 定义和科尔莫戈罗夫方程

类比时间离散的马尔可夫链，连续的马尔可夫链定义为一个这样的随机过程$\{X(t), t \geq 0\}$，满足$P(X(t+s) = S_j | X(s) = S_i, X(u) = x(u), 0 \leq u < s) = P(X(t+s) = S_j | X(s) = S_i)$，连续马尔可夫链的平稳性指转移概率$P(X(t+s) = S_j | X(s) = S_i)$与$s$无关，我们假设以下讨论中的马尔可夫链都是平稳的。

以T_i记一个连续马尔可夫链处于状态S_i的时间，试求当其已经处于S_i长达s时间时，之后的t时间仍然处于S_i的概率：

$$P(T_i > s + t | T_i > s)$$

由马尔可夫性与平稳假设：

$$P(T_i > s + t | T_i > s) = P(T_i > t)$$

所以在连续的马尔可夫链中，处于任何一个状态的持续时间服从一个指数分布，记它的速率为v_i。这种指数分布的状态保留性使得连续情况下的一些处理方法和离散状况不同。

记：

$$P_{ij}(t) = P(T_i > s + t | T_i > s)$$

为当前状态为S_i，经过t时间后状态为S_j的概率。对于任意一个中间时间点的可能状态求和给出连续情况下的科尔莫戈罗夫方程：

$$P_{ij}(t+s) = \sum_{k=0}^{\infty} P_{ik}(t) P_{kj}(s)$$

如果目前以速率v_i处在S_i，以P_{ij}最后转移到S_j状态，那么可知：

$$q_{ij} = v_i P_{ij}$$

为从S_i转移到S_j这一事件的速率（应用符合泊松过程的结论）。称q_{ij}为瞬时转移速率，我们有：

$$v_i = \sum_k q_{ik}$$

$$P_{ij} = \frac{q_{ij}}{\sum_k q_{ik}}$$

考虑到连续马尔可夫链的状态迁移可以看做一个泊松过程的初始步骤，我们有（利用定义3的第三个条件）：

$$\lim_{h\to 0}\frac{1-P_{ii}(h)}{h}=v_i$$

$$\lim_{h\to 0}\frac{P_{ij}(h)}{h}=q_{ij}$$

下面我们计算长程的极限概率分布，也即系统处于各个状态的平均时间。

3.3.2 极限概率分布

由C-K方程：

$$P_{ij}(h+t)-P_{ij}(t)=[\sum_{k=0}^{\infty}P_{ik}(t)P_{kj}(h)]-P_{ij}(t)$$

$$=[\sum_{k\neq j}P_{ik}(t)P_{kj}(h)]-[1-P_{jj}(h)]P_{ij}(t)$$

两侧除以h并取极限$h\to 0$得到：

$$P'_{ij}(t)=[\sum_{k\neq j}P_{ik}(t)q_{kj}]-v_jP_{ij}(t)$$

在长程意义上，我们一般假设最终的平衡状态和初始状态无关，所以：

$$P_j=\lim_{t\to\infty}P_{ij}(t)$$

从而：

$$\lim_{t\to\infty}P'_{ij}(t)=[\sum_{k\neq j}q_{kj}P_k]-v_jP_j$$

因为$P_{ij}(t)$是一个有界函数，所以其导数必收敛至零，故：

$$v_jP_j=\sum_{k\neq j}q_{kj}P_k$$

3.3.3 生灭过程

生灭过程是离散情形下的随机游走过程在连续情形下的类比。在离散情况下，我们以当前坐标位置来刻画状态空间；在连续情形下，游走不是一个很好的例子。想象一个服务系统，其中人数的增加途径是用户的到达，减少途径是接受服务完毕以后用户的离开，此时我们可以用当前服务系统中用户的数量来刻画状态空间，但必须假定用户的到达间隔和服务时间服从指数分布（这两个假设至关重要，它们保证了整个系统的马尔可夫性。如果我们赋予生灭途径以非指数的，即有记忆性的分布，那么当前距离上一次状态变化的时间也会影响系统的状态，故系统中人数不能唯一地确定一个状态）。

我们统一记系统中还有n个人时，到达者的速率为λ_n，服务的速率为μ_n，换言之：

$$v_0 = \lambda_0$$

$$v_i = \lambda_i + \mu_i$$

$$q_{i,i+1} = \lambda_i$$

$$q_{i,i-1} = \mu_i$$

$$P_{i,i+1} = \frac{\lambda_i}{\lambda_i + \mu_i}$$

$$P_{i,i-1} = \frac{\mu_i}{\lambda_i + \mu_i}$$

利用上一节得到的极限概率平衡方程：

$$\lambda_0 P_0 = \mu_1 P_1$$

$$(\lambda_i + \mu_i) P_i = \lambda_{i-1} P_{i-1} + \mu_{i+1} P_{i+1}$$

如果你希望套用离散情形的公式会得到：

$$P_i = \sum_k P_k P_{k,i} = P_{i-1} \frac{\lambda_{i-1}}{\lambda_{i-1} + \mu_{i-1}} + P_{i+1} \frac{\mu_{i+1}}{\lambda_{i+1} + \mu_{i+1}}$$

但这是一个错误的结论。如果我们忽略停留在每一个状态的时间，把整个过程离散化来考察，以π_i记离散化情形下的长程比例，确实有：

$$\pi_i = \sum_k \pi_k P_{ki}$$

$$\sum_k \pi_k = 1$$

而平均每一次处于状态S_i的时长为$\frac{1}{\lambda_i}$，所以：

$$P_i = \frac{\pi_i/\lambda_i}{\sum_k \pi_k/\lambda_k}$$

所以生灭过程的个状态平均时间求解也可以先离散化地解出π_i，并加权归一化。

3.3.4 一般生灭过程的解

从生灭过程的概率平衡方程中，依次逐差得到：

$$\lambda_0 P_0 = \mu_1 P_1$$

$$\lambda_1 P_1 = \mu_2 P_2$$

$$\cdots\cdots$$

$$\lambda_n P_n = \mu_{n+1} P_{n+1}$$

所以：

$$P_n = \frac{\lambda_{n-1}}{\mu_n} P_{n-1} = (\prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}) P_0$$

归一化条件等价于：

$$P_0 = ((\sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}) + 1)^{-1}$$

所以：

$$P_n = \frac{\prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}}{\sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} + 1}$$

有解的充分条件是：

$$\sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} < \infty$$

4 排队理论

4.1 定义和肯德尔记号

排队论研究这样一种系统：顾客以某种随机方式进入，并在队列中等候直到其可以享受服务，服务结束后顾客离开系统。我们试图了解这样一个系统的一些宏观性质。

50年代初，美国数学家关于生灭过程的研究、英国数学家D.G.肯德尔提出嵌入马尔可夫链理论，以及对排队队型的分类方法，为排队论奠定了理论基础。其中肯德尔记号(Kendall Notation)是划分排队模型的手段。

一般的排队模型可以表示为模板“X/Y/Z/A/B/C”：

X—顾客相继到达的间隔时间的分布；

Y—服务时间的分布；

（X和Y的取值可以为：M-指数分布，G-一般分布）

Z—服务台个数；

A—系统容量限制（默认为∞）；

B—顾客源数目（默认∞）；

C—服务规则（默认为FCFS”First Come, First Service”）。

后三个特征一般使用默认情况，前三者可以形成组合：

1、M/M/1：顾客到来时间间隔和服务时间服从指数分布，仅有一个队列；

2、M/M/k：顾客到来时间间隔和服务时间服从指数分布，有$k \geq 2$个队列；

（以上两种统称为指数模型）

3、M/G/1：顾客到来时间间隔副从指数分布，服务时间服从一般分布，单条队列；

4、G/M/1：顾客到来时间间隔副从一般分布，服务时间服从指数分布，单条队列；

下文的讨论重点是上面四个模型，另有M/G/K，G/M/K，G/G/1和G/G/k等模型，因为其分析性质一般，可以采取数值模拟的方法定性地研究。而即便是最简单的M/M/1模型，也可以通过设置可变参数和外加条件等等进行扩充和推广。

4.2 目标变量和价格方程

在排队系统上重要的基本量是：

L，系统中顾客的平均数；

L_Q，队列中平均等待顾客数；

W，一个顾客在系统中所耗的平均时间；

W_Q，一个顾客在队列中等待的平均时间。

排队理论就是从排队系统的人员到达分布和服务时间分布计算上述四个基本量的理论，为了说明它们之间的关系，我们赋予这个系统一个支付规则，即每个顾客需要按照此规则向系统支付金钱，则有价格恒等式：

系统赚钱的平均速率=λ_a*进入系统的顾客所支付的平均金额

其中λ_a是顾客抵达系统的速率：

$$\lambda_a = \lim_{t\to\infty} \frac{N(t)}{t}$$

我们假设如下的三种规则：

规则1：对于每个客户，只要其处于系统中一个单位时间，就需要支付一个单位金钱。

此时系统赚钱的速率就是系统中客户的数量，而平均支付金额就是客户在系统中停留的平均时间，所以价格恒等式给出：

$$L = \lambda_a * W$$

规则2：对于每个客户，只要其处于等待队列中一个单位时间，就需要支付一个单位金钱。

此时系统赚钱的速率就是队列中客户的数量，而平均支付金额就是客户在队列中停留的平均时间，所以价格恒等式给出：

$$L_Q = \lambda_a * W_Q$$

规则3：对于每个客户，只要其处于被服务状态中一个单位时间，就需要支付一个单位金钱。

此时系统赚钱的速率就是系统处于服务状态的时间比例，而平均支付金额就是客户被服务的平均时间（记为$E(S)$），所以价格恒等式给出：

$$1 - P_0 = \lambda_a * E(S)$$

把系统中有人在接受服务的时期称为“忙期”，否则称为“闲期”，则系统交替地处于闲期和忙期。一个给定系统的平均忙期时长和平均闲期时长$E(B)$和$E(I)$常常也是度量系统性能的重要因素。

4.3 排队过程作为生灭过程的推广

4.3.1 M/M/1

考虑最简单的排队情况：顾客的到来间隔服从于速率为λ的指数分布，而服务的用时服从于速率μ的指数分布，试求$L, L_Q, W, W_Q, E(I), E(B)$。

此时的系统等同于一个对于所有$n, \lambda_n = \lambda, \mu_n = \mu$的生灭过程，根据3.3.4节中的结论：

$$P_0 = \frac{\mu - \lambda}{\mu}$$

$$P_n = \frac{\mu - \lambda}{\mu}(\frac{\lambda}{\mu})^n$$

所以：

$$L = \sum_{n=0}^{\infty} nP_n = \frac{\mu - \lambda}{\mu} \sum_{n=0}^{\infty} n(\frac{\lambda}{\mu})^n$$

利用：

$$\sum_{n=0}^{\infty} nx^n = \frac{x}{(1-x)^2}$$

得到：

$$L = \frac{\mu - \lambda}{\mu} \frac{\lambda/\mu}{(1-\lambda/\mu)^2} = \frac{\lambda}{\mu - \lambda}$$

$$W = \frac{L}{\lambda} = \frac{1}{\mu - \lambda}$$

$$W_Q = W - E(S) = \frac{\lambda}{\mu(\mu - \lambda)}$$

$$L_Q = \lambda W_Q = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

再利用P_0是系统处于闲期的时间占比：

$$P_0 = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} I_i}{\sum_{i=1}^{n} I_i + B_i} = \lim_{n\to\infty} \frac{\frac{1}{n}\sum_{i=1}^{n} I_i}{\frac{1}{n}\sum_{i=1}^{n} I_i + B_i} = \frac{E(I)}{E(I) + E(B)}$$

最后一步是大数定律的结果，利用$E(I) = \frac{1}{\lambda_0 = \lambda}$：

$$E(B) = \frac{1 - P_0}{\lambda_0 P_0} = \frac{1}{\mu - \lambda}$$

在指数模型里，我们还可以计算一个顾客在系统中停留时间W^*的分布，首先计算一个停留时间$W^* = t$的顾客到达时，系统中到达人数N的分布：

$$P(N = n|W^* = t) = \frac{P(N = n)P(W^* = t|N = n)}{P(W^* = t)}$$

以n为自变量，此时$P(W^* = t|N = n)$是一个伽马密度，因为需要等待的时间是n次服务时间之和：

$$P(N = n|W^* = t) = \frac{1}{P(W^* = t)} P_n \frac{\mu e^{-\mu t}(\mu t)^{n-1}}{n!} = \frac{(1 - \lambda/\mu)\mu e^{-\mu t}}{P(W^* = t)} \frac{(\lambda t)^n}{n!}$$

对n求和：

$$\sum_{n=0}^{\infty} P(N = n|W^* = t) = \frac{(1 - \lambda/\mu)\mu e^{-\mu t}}{P(W^* = t)} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} = \frac{(1 - \lambda/\mu)\mu e^{-\mu t}}{P(W^* = t)} e^{\lambda t} = 1$$

得到：

$$P(W^* = t) = (\mu - \lambda)e^{-(\mu-\lambda)t}$$

即顾客的等待加上服务时间服从一个速率为$\mu - \lambda$的指数分布。

以上将朴素的M/M/1作为恒速率的生灭过程来处理，当引入生灭速率与状态有关的非平凡关系时（譬如提供服务的能力会因为排队人数数量增加而增加），需要使用3.3.4中解的完整形式。

4.3.2 例6：有限容量的M/M/1模型

考虑M/M/1模型的这样一个变型：排队队列的长度上限为N，即如果排队人数已经到达N位，则新到来的顾客将不会进入系统。此时的平衡方程与常规解的不同之处在于一般递推式在$n = N$时变化为：

$$\mu P_N = \lambda P_{N-1}$$

归一时的等比数列求和在有限项上进行，得到：

$$P_0 = \frac{1 - \frac{\lambda}{\mu}}{1 - (\frac{\lambda}{\mu})^{N+1}}$$

此时的到达速率：

$$\lambda_a = \lambda(1 - P_N)$$

因为顾客的到达构成一个泊松过程，按照成功进入/失败进入分为两类，3.2.2节中的结果指出成功进入的计数是一个到达速率为$\lambda(1-P_N)$的新的泊松过程。

利用P_N的递推式以及$L=\sum_n nP_n$和4.2中的关系式可以解出有限容量M/M/1的四个目标变量。

4.3.3 例7：优化排队系统的参数

排队论在了解系统性质以外，也能够利用对于系统性质的定量分析来优化系统参数。

考虑这样一个场景：系统提供速率为μ的服务的价格是$c\mu$，每次服务收费A，来客速率为λ，队列容量为N，试求$\mu|c,A,\lambda,N$使得平均利润最大。

记单位时间的利润为M，利用4.2的价格方程，A即人均支付的金额：

$$M(\mu|c,A,\lambda,N)=\lambda_a A-c\mu=\frac{\lambda A(1-(\frac{\lambda}{\mu})^N)}{1-(\frac{\lambda}{\mu})^{N+1}}-c\mu$$

对μ求导数置零即得最优解。

4.3.4 M/M/k

对应于肯德尔记号M/M/k的是k个可选队列，当k个队列对应的服务台中有一个结束服务时，新的客户就会进入被服务的进程。这一模型可以完全套用生灭过程的一般解，设到达速率为λ，单服务台处理速率为μ：

$$\lambda_n=\lambda$$

$$\mu_n=\begin{cases}n\mu & \text{n}<\text{k}\\ k\mu & \text{n}\geq\text{k}\end{cases}$$

上面第二个等式成立是因为指数分布随机变量由min算符联系时的性质。

此时：

$$\prod_{i=1}^{n}\frac{\lambda_{i-1}}{\mu_i}=\begin{cases}\frac{\lambda^n}{\mu^n n!} & \text{n}<\text{k}\\ \frac{\lambda^n}{\mu^n k^{n-k}k!} & \text{n}\geq\text{k}\end{cases}$$

$$P_0=(1+\sum_{n=1}^{k-1}\frac{\lambda^n}{\mu^n n!}+\sum_{i=k}^{\infty}\frac{\lambda^n}{\mu^n k^{n-k}k!})^{-1}$$

对于$n < k$：

$$P_n = P_0 \frac{\lambda^n}{\mu^n n!}$$

对于$n \geq k$：

$$P_n = P_0 \frac{\lambda^n}{\mu^n k^{n-k} k!}$$

四个关键目标变量的计算方法不变。

4.4 其他排队模型

4.4.1 M/G/1

在服务时间服从一般的任意分布时，设置这样一种付费策略：如果一顾客的剩余服务时间为t，则顾客到达系统到其离开系统时，每单位时间支付金额t。此时系统赚钱的速率为当前所有顾客所剩余的服务时间之和，记平均排队用时W_Q，服务时长S而平均每个顾客支付的金额为：

$$E(cost) = E(W_Q S + \int_{t=0}^{S} (S-t)dt) = W_Q E(S) + \frac{E(S^2)}{2}$$

所以价格方程变为：

$$W_Q = W_Q \lambda_a E(S) + \frac{\lambda_a E(S^2)}{2}$$

如果到达速率服从统一的λ，得到波拉泽克-欣齐内公式（Pollaczek-Khintchine）：

$$W_Q = \frac{\lambda E(S^2)}{2(1-\lambda E(S))}$$

L, L_Q, W在求得W_Q后利用4.3.1中的关系可求。

在求平均闲期长度和忙期长度时，利用4.2中价格方程的最后一种变型：

$$1 - P_0 = \lambda E(S)$$

以及4.3.1中的关系：

$$P_0 = \frac{E(I)}{E(I) + E(B)}$$

和指数到达的性质：

$$E(I) = \frac{1}{\lambda}$$

解得：

$$E(B) = \frac{E(S)}{1 - \lambda E(S)}$$

非指数模型不可以套用连续马尔可夫链的结论，因为状态保持的时长不再服从指数分布。在这种情形下，只能通过设置付费策略来尝试性地提取平均意义上统计量之间的关系。失效的不仅仅是递推关系，就连我们一直以来用系统中剩余人数来划分状态空间的方法也是无效的，因为M/G/1的完全状态空间应该有有序对(n,t)来表示，其中n代表当前系统

人数，t代表当前被服务的客户已经被服务的时间。之所以在上述的推理中使用了P_0是因为在$n=0$时，状态的迁移由一个指数分布随机变量引起，故序对中的t是没有价值的。

4.4.2 G/M/1

当服务时间服从指数分布，而顾客服从任意分布时，按照上一节末尾的讨论，完备的状态空间由有序对(n,t)其中t代表最后一个顾客抵达系统后经过的时间，已服务时间由其无记忆性无需考察。这个模型的处理技巧是：考虑顾客到达时系统中的人数形成的随机过程$\{X_n, n\geq 1\}$，记X_n为第n个到达者看到的系统中的人数（这是一个离散的随机过程!）。

考察此时的状态迁移概率，以j代表新的一位顾客到来用时中被服务而离开的人数，在服从一般分布的到来用时中被服务的人数构成一个泊松过程：

$$P_{i,i+1-j}=\int_{t=0}^{\infty}e^{-\mu t}\frac{\mu^j}{j!}g(t)dt$$

$$P_{i,0}=1-\sum_{j=0}^{i}P_{i,i+1-j}$$

极限概率的平衡方程组为：

$$\sum_{k=0}^{\infty}\pi_k=1$$

$$\pi_k=\sum_{i=k-1}^{\infty}\pi_i\int_{t=0}^{\infty}e^{-\mu t}\frac{\mu^{i+1-k}}{(i+1-k)!}g(t)dt\qquad(k\geq 1)$$

这一方程组的通解形式为：

$$\pi_k=c\beta^k$$

容易验证其中β满足：

$$\beta=\int_0^{\infty}e^{-\mu t(1-\beta)}g(t)dt$$

利用数值方法求得β后可求$\pi_k=(1-\beta)\beta^k$，容易验证在一般分布被取为指数分布时这一解和M/M/1系统解的一致性。

W可以对到达顾客所看见的人数的所有可能性直接求和：

$$W = \sum_{k=0}^{\infty} \pi_k E(waitingtime|k) = \sum_{k=0}^{\infty} (1-\beta)\beta^k \frac{k+1}{\mu}$$

$$= \frac{1-\beta}{\mu} \sum_{k=0}^{\infty} (k+1)\beta^k$$

它正比于$\frac{\beta}{1-\beta}$的微分，所以：

$$W = \frac{1}{\mu(1-\beta)}$$

此时的到达速率：

$$\lambda_a = \frac{1}{\int_{t=0}^{\infty} tg(t)dt}$$

W_Q, L和L_Q由价格方程和基本关系可得。

在G/M/1模型中，我们还可以求得顾客等待时间的分布，它是一个速率为$\mu(1-\beta)$的指数分布，求解的方法和4.3.1保持一致，因为在那里我们并没有使用到顾客到达为泊松过程的前提。

为了计算长程时间占比P_n，首先：系统中人数从n增加到$n+1$的速率和从$n+1$减少到n的速率必须相同，否则系统在长程就不处于平衡态：

$$\lambda\pi_n = \mu P_{n+1}$$

所以：

$$P_n = \frac{\lambda}{\mu}(1-\beta)\beta^{n-1}$$

$$P_0 = 1 - \frac{\lambda}{\mu}$$

分析G/M/1的忙闲期，首先注意到在离散过程中状态0返回自身的平均用时就是一个闲期-忙期周期的平均人数，利用2.2.3中的结论，该人数期望等于：

$$\frac{1}{\pi_0} = \frac{1}{1-\beta}$$

故一个闲期-忙期平均用时：

$$\frac{1}{\lambda(1-\beta)} = E(I) + E(B)$$

再利用4.3.1中P_0与$E(I), E(B)$的关系，得到：

$$E(I) = \frac{1}{\mu(1-\beta)}$$

$$E(B) = \frac{\mu - \lambda}{\lambda\mu(1-\beta)}$$

4.5 排队网络

在以上的情况下，我们都考虑“顾客到达→顾客接受一次服务→顾客离开的模式”，在这一小节我们考虑顾客串联地接受两次服务的情况。为了简单起见，假设顾客到达间隔、第一个服务台服务、第二个服务台服务分别服从速率为λ, μ_1, μ_2的指数分布。

此时状态空间扩展为二维$P_{n,m}$，n和m分别表示第一个服务子系统和第二个服务子系统中人数，在一般情况下：

$$(\lambda+\mu_1+\mu_2)P_{n,m}=\lambda P_{n-1,m}+\mu_1 P_{n+1,m-1}+\mu_2 P_{n,m+1}$$

三个边界情况：

$$\lambda P_{0,0}=\mu_2 P_{0,1}$$

$$(\lambda+\mu_1)P_{n,0}=\lambda P_{n-1,0}+\mu_2 P_{n,1}$$

$$(\lambda+\mu_2)P_{0,m}=\mu_1 P_{1,m-1}+\mu_2 P_{0,m+1}$$

可以联系归一化条件直接求同解，不过更好的方法是利用结论：遍历的生灭过程是时间可逆的（两个直接互连的状态间互相转移的速率必须相等，否则此过程在长程而言不是遍历的）。则第一个服务台输出的顾客也服从速率为λ的泊松过程，从而可以把两个服务子系统看做两个独立的M/M/1模型，所以：

$$P_{n,m}=P_n^1 P_m^2=\frac{\mu_1-\lambda}{\mu_1}(\frac{\lambda}{\mu_1})^n\frac{\mu_2-\lambda}{\mu_2}(\frac{\lambda}{\mu_2})^m$$

$$L=\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(n+m)P_{n,m}=\sum_{n=0}^{\infty}nP_n^1(\sum_{m=0}^{\infty}P_m^2)+\sum_{m=0}^{\infty}mP_m^2(\sum_{n=0}^{\infty}P_n^2)$$

$$=\frac{\lambda}{\mu_1-\lambda}+\frac{\lambda}{\mu_2-\lambda}$$

$$W=\frac{1}{\mu_1-\lambda}+\frac{1}{\mu_2-\lambda}$$

串联系统的“等待”不是很好定义，故不讨论L_Q, W_Q和闲忙期指标。

参考文献

[1] Introduction to Probability Models(14th) SM.Ross 人民邮电出版社 2014

[2] 矩阵理论与应用　张跃辉　科学出版社 2011

[3] Essentials of Stochastic Process(2ed) R.Durrett 机械工业出版社 2014