STAT0008: Stochastic Processes

Lecture 12 - Queueing Theory

Lecturer: Weichen Zhao　　　　Spring 2025

Key concepts:

- *M/M/1*;
- *M/M/s*;
- *Little* 公式.

排队论是Markov链的一个重要应用，我们先回顾基本的排队模型。

一个排队系统通常记为

$$X/Y/Z$$

其中X表示到达间隔时间的分布，Y表示服务时间的分布，Z表示服务台的数量，这个记号称为Kendall记号。

X/Y的常见类型有：

- M: Memoryless, 指数分布
- G: General, 一般的分布G
- D: Deterministic, 确定型

排队论标准化符号

$$X/Y/Z/A/B/C$$

其中A表示系统容量的限制，B表示客源数量(一般是∞)，C表示服务方式(一般是先到先服务)

排队论的研究内容非常广泛，包含管理科学，运筹决策等方面。本课程只讨论 Markov 链相关的问题，主要关心下面几个量：

(1) 排队系统中顾客的平均数 L；

(2) 排队系统中等待服务的顾客的平均数 L_Q；

(3) 顾客在排队系统中所花费时间的平均值 W；

(4) 顾客在排队系统中用于等候的时间的平均值 W_Q。

我们考虑上面4个“平均值”是因为多数情况下关心的问题不在于排队系统的瞬时变化，而在于系统进入稳态后的情况，所以解决上述问题的关键是排队服务系统的队列长度的极限分布。

这些量对于运筹管理来说是重要的，比如顾客的数量×平均花费就是营业额，顾客在排队等候的时间又会影响顾客的消费意愿。

12.1 M/M/1

M/M/1 模型的到达过程为 Poisson 过程，设顾客流到达强度为λ。仅有一个服务台，服务时间服从参数为μ指数分布。设队列长度为$X(t)$，则$\{X(t)\}$是连续时间 Markov 链，可以用线性齐次生灭过程进行描述，其中$\lambda_n = \lambda$, $\mu_n = \mu$.

12.1.1 队列长度无限制

队列长度无限制的情况下，$X(t)$可以一直增长下去，由 Proposition 11.5, 若

$$1 + \sum_{n=1}^{\infty} \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_1\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_2\mu_1} = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n < \infty$$

即$\lambda < \mu$，上面级数收敛，则过程存在极限分布

$$p_j = \lim_{t\to\infty} p_j(t) = \left(\frac{\lambda}{\mu}\right)^j \left(1 - \frac{\lambda}{\mu}\right)$$

直观来看，如果顾客到达的强度高于服务台的服务速率，$\lambda > \mu$，那么队列将无限制地增长，系统始终无法进入平稳状态。

考虑$\lambda < \mu$，系统进入平稳后，记系统中出现 n 个顾客的概率为 p_n，所以系统中顾客的平均数为

$$L = \sum_{n=0}^{\infty} np_n = \sum_{n=0}^{\infty} n \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) = \frac{\lambda}{\mu - \lambda}$$

这n个顾客中，1人正在接受服务，剩下$n-1$人在排队等候，所以等候的平均人数为

$$L_{\mathrm{Q}} = \sum_{n=1}^{\infty} (n-1)p_n = L - (1 - p_0) = \frac{\lambda}{\mu - \lambda} - \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu - \lambda)}.$$

下面我们计算时间相关的量。假设顾客到达时系统中有n个顾客，由于指数分布的无记忆性，所以从顾客的到达时刻算起，正在接受服务的顾客还要被服务的时间仍然服从指数分布，且参数μ没有变化，所以该顾客逗留在系统的平均时间

$$T_n = \frac{n+1}{\mu}$$

排队等候的平均时间(Gamma分布)

$$T_{n,Q} = \frac{n}{\mu}$$

于是，顾客在系统中的平均逗留时间W以及排队等候的平均时间W_Q 分别为

$$W = \sum_{n=0}^{\infty} T_n p_n = \sum_{n=0}^{\infty} \frac{n+1}{\mu} p_n = \frac{L}{\mu} + \frac{1}{\mu} = \frac{1}{\mu - \lambda}$$

$$W_Q = \sum_{n=0}^{\infty} T_{n,Q} p_n = \sum_{n=0}^{\infty} \frac{n}{\mu} p_n = \frac{L}{\mu} = \frac{\lambda}{\mu(\mu - \lambda)}$$

12.1.2 队列长度受限

现实中，排队的长度往往受到场地空间，顾客排队意愿等限制。

假设队列长度不可以超过N，即M/M/1/N 系统，那么状态空间为$\{0,1,\cdots,N\}$。如果顾客到达时队伍中已经有N个顾客，说明队列已满，新来的顾客将自动离去。此时过程的平稳分布满足

$$\begin{aligned}&\lambda p_0=\mu p_1\\&(\lambda+\mu)p_n=\lambda p_{n-1}+\mu p_{n+1},\quad 1\leqslant n\leqslant N-1\\&\lambda p_{N-1}=\mu p_N\end{aligned}$$

解不变方程，有

$$p_n=\frac{\lambda}{\mu}p_{n-1}=\cdots=\left(\frac{\lambda}{\mu}\right)^n p_0,\quad 1\leqslant n\leqslant N$$

由于有限状态空间，所以无需考虑级数收敛的问题，由归一化条件

$$1=\sum_{n=0}^{N}p_n=p_0\left[1+\frac{\lambda}{\mu}+\cdots+\left(\frac{\lambda}{\mu}\right)^N\right]=p_0\frac{1-\left(\frac{\lambda}{\mu}\right)^{N+1}}{1-\frac{\lambda}{\mu}}$$

可得

$$p_n=\left(\frac{\lambda}{\mu}\right)^n\frac{1-\frac{\lambda}{\mu}}{1-\left(\frac{\lambda}{\mu}\right)^{N+1}},\quad n=0,1,2,\cdots,N$$

那么系统中顾客的平均数为

$$\begin{aligned}L=\sum_{n=0}^{N}np_n&=\frac{1-\frac{\lambda}{\mu}}{1-\left(\frac{\lambda}{\mu}\right)^{N+1}}\sum_{n=1}^{N}n\left(\frac{\lambda}{\mu}\right)^n\\&=\frac{\lambda}{\mu-\lambda}\frac{1-(N+1)\left(\frac{\lambda}{\mu}\right)^N+N\left(\frac{\lambda}{\mu}\right)^{N+1}}{1-\left(\frac{\lambda}{\mu}\right)^{N+1}}\end{aligned}$$

排队等候的平均人数为

$$L_Q=\sum_{n=1}^{N}(n-1)p_n=L-\sum_{n=1}^{N}p_n=L-(1-p_0)$$

下面计算时间相关的量。与队列长度无限制的情况不同，受限情况下有些顾客到达服务台时发现队列已满，会立即自动离去，此时顾客不存在花费时间和等候时间。注意在求和的上限为$N-1$而不是N，这是因为$n=N$时到达的顾客在系统中逗留的时间为 0。

顾客遇到队列满的概率是p_N，我们不考虑计算这部分顾客的花费时间和等候时间，那么需要对队列长度的概率做归一化，即

$$\begin{aligned} W &= \sum_{n=0}^{N-1} \frac{n+1}{\mu} \frac{p_n}{1-p_N} \\ &= \frac{1}{\mu(1-p_N)} \left(\sum_{n=0}^{N-1} n p_n + \sum_{n=0}^{N-1} p_n \right) \\ &= \frac{L-(N+1)p_N+1}{\mu(1-p_N)} \end{aligned}$$

平均的排队等候时间

$$W_Q = \sum_{n=0}^{N-1} \frac{n}{\mu} \frac{p_n}{1-p_N} = \frac{L-Np_N}{\mu(1-p_N)}$$

12.1.3 Little 公式

在计算我们关心的这四个量时，是否有更简单的计算方式？这是由Little公式给出的。

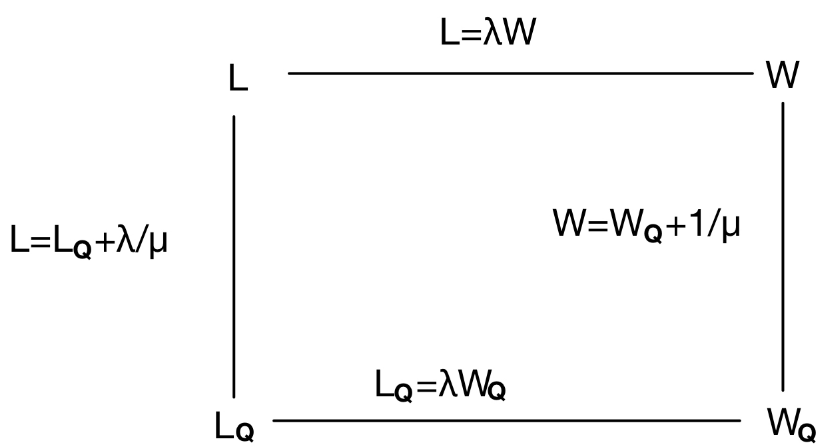

图 12.1: Little 公式

Proposition 12.1 (Little 公式) L，L_Q，W，W_Q具有下面关系：

(1) $L = \lambda W$

(2) $L_Q = \lambda W_Q$

(3) $L = L_Q + \frac{\lambda}{\mu}$

(4) $W = W_Q + \frac{1}{\mu}$

Proof: 只证(1). 当系统进入稳态后，我们考虑取一个充分大的时间间隔T，从两个方面计算T时间内系统内所有顾客花费的总平均时间。

一方面，由于系统已经平稳，系统中的平均顾客数为L，所以在T时间内系统中的所有顾客逗留的总平均时间为 LT。

另一方面，每一个顾客在系统中的平均逗留时间为W， 而T 时间内到达顾客的平均数目为λT，所以所有顾客逗留的总平均时间为λWT，因而有

$$LT = \lambda WT$$

即$L = \lambda W$. ■

思考：对M/M/1/∞和M/M/1/N排队系统验证Little 公式。

注1. 在对M/M/1/N排队系统验证时，注意此时的到达率因为顾客到达服务台后以概率p_N立刻离开，以概率 $1-p_N$进入系统开始排队，那么进入系统的顾客仍然服从 Poisson过程，参数变为 $\lambda(1-p_N)$，即有效到达率为

$$\lambda_{\text{eff}} = \lambda(1-p_N)$$

注2. Little 公式具有广泛的适用性：

(1) 服务规则无关：适用于任意服务规则，如FIFO（先进先出）、LIFO（后进先出）、优先级队列等。

(2) 分布无关性：适用于任意到达间隔分布（如M/M/1、M/G/1、G/G/1等）和服务时间分布。

(3) 系统结构灵活：适用于单服务台（如M/M/1）、多服务台（如M/M/s）、有限容量队列（如M/M/1/N）、批量到达、批量服务系统。

12.2 M/M/s

本节把1个服务台推广为$s > 1$个，仍以系统中的顾客数作为状态空间，假定各服务台的工作统计独立，则当$n < s$时，服务台有空闲，实际工作的服务台个数为n，此时顾客离开系统的速率为$\mu_n = n\mu$。当$n \geqslant s$时，所有s个服务台均在工作，顾客中有$n-s$个在排队等待，顾客离开系统的速率为$\mu_n = s\mu$，即M/M/s 模型仍是生灭过程，参数为

$$\lambda_n = \lambda, \quad \mu_n = \begin{cases} n\mu, & n \leqslant s \\ s\mu, & n > s \end{cases}$$

12.2.1 队列长度无限制 M/M/s/∞

如果队列长度无限制，过程的状态空间为$\{0, 1, 2, \cdots\}$，依然由Proposition 11.5,

$$p_n = \begin{cases} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0, & n \leqslant s, \\ \frac{1}{s!s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n p_0, & n > s \end{cases}$$

若要不变分布的存在，则要求级数

$$\sum_{n=0}^{\infty} p_n = p_0 \left[\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \sum_{n=0}^{\infty}\left(\frac{\lambda}{s\mu}\right)^n\right]$$

收敛，即$\frac{\lambda}{s\mu} < 1$，利用归一化条件：

$$1 = \sum_{n=0}^{\infty} p_n = p_0 \left[\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{(s-1)!}\left(\frac{\lambda}{\mu}\right)^s \frac{\mu}{s\mu - \lambda}\right]$$

即得

$$p_0 = \left[\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{(s-1)!}\left(\frac{\lambda}{\mu}\right)^s \frac{\mu}{s\mu - \lambda}\right]^{-1}$$

利用极限分布我们可以计算关心的量：

$$\begin{aligned}
L = \sum_{n=1}^{\infty} n p_n &= p_0 \left[\sum_{n=1}^{s} \frac{n}{n!} \left(\frac{\lambda}{\mu}\right)^n + \frac{s^s}{s!} \sum_{n=s+1}^{\infty} n \left(\frac{\lambda}{s\mu}\right)^n \right] \\
&= p_0 \left[\frac{\lambda}{\mu} \sum_{n=0}^{s-1} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n + \frac{s^{s+1}}{s!} \frac{\left(\frac{\lambda}{s\mu}\right)^{s+1}}{1-\frac{\lambda}{s\mu}} + \frac{s^s}{s!} \frac{\left(\frac{\lambda}{s\mu}\right)^{s+1}}{\left(1-\frac{\lambda}{s\mu}\right)^2} \right] \\
&= p_0 \left[\frac{\lambda}{\mu} \sum_{n=0}^{s-1} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n + \frac{s^{s+1}}{s!} \left(\frac{\lambda}{s\mu}\right)^{s+1} \frac{s\mu}{s\mu-\lambda} + \frac{s^s}{s!} \left(\frac{\lambda}{s\mu}\right)^{s+1} \left(\frac{s\mu}{s\mu-\lambda}\right)^2 \right] \\
&= \frac{\lambda}{\mu} + p_0 \left(\frac{\lambda}{\mu}\right)^{s+1} \frac{1}{(s-1)!} \left(\frac{\mu}{s\mu-\lambda}\right)^2 \qquad (\text{注意到} p_0 \sum_{k=0}^{s-1} \frac{1}{k!} \left(\frac{\lambda}{\mu}\right)^k = 1 - p_0 \left(\frac{\lambda}{\mu}\right)^s \frac{1}{s!} \frac{s\mu}{s\mu-\lambda})
\end{aligned}$$

由Little 公式可以计算出

$$L_Q = L - \frac{\lambda}{\mu} = p_0 \left(\frac{\lambda}{\mu}\right)^{s+1} \frac{1}{(s-1)!} \left(\frac{\mu}{s\mu-\lambda}\right)^2$$

$$W_Q = \frac{L_Q}{\lambda} = \frac{p_0}{\mu} \left(\frac{\lambda}{\mu}\right)^{s} \frac{1}{(s-1)!} \left(\frac{\mu}{s\mu-\lambda}\right)^2$$

$$W = W_Q + \frac{1}{\mu} = \frac{1}{\mu} + \frac{p_0}{\mu} \left(\frac{\lambda}{\mu}\right)^{s} \frac{1}{(s-1)!} \left(\frac{\mu}{s\mu-\lambda}\right)^2$$

12.2.2 队列长度受限 M/M/s/N

如果队列长度受限，过程的状态空间为$\{0, 1, 2, \cdots, N\}$，则生灭过程的参数为

$$\lambda_n = \lambda, \quad \mu_n = \begin{cases} n\mu, & 0 \leqslant n \leqslant s \\ s\mu, & s < n \leqslant N \end{cases}$$

极限分布为

$$p_n = \begin{cases} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n p_0, & 0 \leqslant n \leqslant s, \\ \frac{1}{s! s^{n-s}} \left(\frac{\lambda}{\mu}\right)^n p_0, & s < n \leq N \end{cases}$$

由于状态有限不需要考虑级数收敛性，由归一化条件：

$$p_0 = \left[\sum_{k=0}^{s} \frac{1}{k!} \left(\frac{\lambda}{\mu}\right)^k + \sum_{k=s+1}^{N} \frac{1}{s!\, s^{k-s}} \left(\frac{\lambda}{\mu}\right)^k \right]^{-1}$$

由Little 公式可以计算出

$$L_Q = p_0 \frac{\rho(s\rho)^s}{s!\,(1-\rho)^2}\left[1-\rho^{N-s}-(N-s)\rho^{N-s}(1-\rho)\right]$$

$$L = L_Q + \frac{\lambda(1-p_N)}{\mu}$$

$$W_Q = \frac{L_Q}{\lambda(1-p_N)}$$

$$W = W_Q + \frac{1}{\mu}$$

其中$\rho = \frac{\lambda}{s\mu}$。