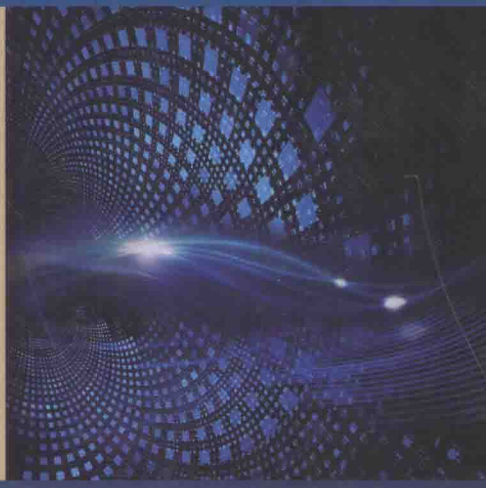
“十二五”江苏省高等学校重点教材

排队论及其应用



唐加山　编著

科学出版社

“十二五” 江苏省高等学校重点教材
重点教材编号：2015-2-049

排队论及其应用

唐加山 编著

科学出版社
北京

内 容 简 介

本书介绍排队论的基本概念、基本理论、基本方法和应用举例，主要内容包括：基本概念及术语介绍、基本单节点排队模型、研究方法简介、广义单节点排队模型、排队网络模型、应用举例等. 全书从相对较低的起点出发详细介绍排队论的基本内容，让读者掌握较为扎实的基础知识，对于理论前沿和应用方面的内容，做相对简明的介绍，同时列出重要的参考文献，让有兴趣的读者可以继续进行深入的探索.

本书适合数理基础较好的高年级本科生、相关专业的研究生或教师使用，也可以作为对排队论有兴趣的科技工作人员的初级读物.

图书在版编目(CIP)数据

排队论及其应用/唐加山编著. —北京：科学出版社, 2016.6
“十二五” 江苏省高等学校重点教材
ISBN 978-7-03-049389-7

Ⅰ. ①排… Ⅱ. ①唐… Ⅲ. ①排队论-高等学校-教材 Ⅳ. ①O226

中国版本图书馆 CIP 数据核字(2016) 第 162858 号

责任编辑：张中兴 / 责任校对：钟 洋
责任印制：徐晓晨 / 封面设计：迷底书装

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京中石油彩色印刷有限责任公司 印刷
科学出版社发行 各地新华书店经销
*
2016 年 6 月第 一 版 开本：720×1000 1/16
2016 年 6 月第一次印刷 印张：15 1/2
字数：313 000

定价：45.00 元

(如有印装质量问题，我社负责调换)

前　言

排队论，又称随机服务系统理论，它是应用概率统计与运筹学的交叉学科之一，具有非常广泛的应用范围，笔者在教授这门课程时，同学们希望有一本起点相对较低，又能引领他们进入排队论前沿的教材，这促使笔者有了编写这样一本教材的想法.

排队论的理论非常丰富，应用也极其广泛，要在一本书中涵盖所有的内容是不现实的，这就要对相关的内容进行取舍，笔者编写的原则是坚持较详细地介绍排队论的基本内容，让学生掌握较为扎实的基础知识，对于理论前沿和应用方面的内容，做相对简明的介绍，同时列出重要的参考文献，让有兴趣的读者可以继续进行深入的探索，为了节省篇幅，假设读者已经具备工科概率论和随机过程的基础知识，基于这样的思想，全书内容共分六个章节.

第 1 章介绍排队论的基本概念，特别是一些标准记号以及普适的结论，为排队论课程的学习做一些技术性的准备.

第 2 章介绍基本的单节点排队模型，这些模型是排队论研究的核心要素，实际上，任何复杂的排队网络模型都是由数目不等的单节点排队模型组成的，本章内容是进一步学习和研究复杂排队模型的基础.

第 3 章通过例子介绍研究排队模型的一些方法，以便读者对这些方法有一定的了解，在排队论的学习以及以后的研究中掌握较多的工具，解决相关问题选择方法时更有针对性.

第 4 章是第 2 章的扩展，旨在介绍一些广义的单节点排队模型，它们既是经典的单节点排队模型的变形或推广，本身具有相当重要的研究意义，而且又是很多现代复杂排队网络模型重要的组成部分，读者可以在大致了解有关模型的基础上，沿着自己感兴趣的方向在科学研究的道路上继续探索前进.

第 5 章介绍排队网络模型，它们是现代排队论应用的基础，读者不但可以学习复杂模型的知识，更可以了解处理这些大型模型的方法.

第 6 章介绍排队论在不同学科领域中的一些应用，这些内容进一步表明了排队论知识的强应用性，可以激发读者对于排队论课程的学习兴趣和研究热情，笔者特别期望读者能结合自身的知识背景，在自己的研究领域中找到排队论应用的结合点.

笔者在北京师范大学求学期间，师从王梓坤院士和李占柄教授学习随机过程，在加拿大卡尔顿大学访学期间，师从 Donald A. Dawson 院士和 Yiqiang Q. Zhao

教授开始进行排队论的学习和研究，是他们引领笔者进入随机过程理论及其应用这一广阔的研究领域，笔者时刻谨记恩师的教诲，感谢他们多年来的培养，在笔者的成长过程中，还得到过多位前辈和同行们的关心、支持和帮助，在此致以诚挚的感谢！本书讲义版本的部分内容在南京邮电大学 2010~2014 级博士生班上讲授过，他们提出了很多有价值的意见和建议，借此机会，对他们致以衷心的感谢. 在本书的出版过程中，得到了科学出版社高教数理分社社长昌盛和首席策划编辑张中兴的热情关心和大力帮助，另外，本书获选江苏省和南京邮电大学“十二五”高等学校重点建设教材和立项资助以及得到国家自然科学基金项目 (11371191) 的部分资助，笔者在此均致以衷心的感谢.

本书适合数理基础较好的高年级本科生、相关专业的研究生或教师使用，也可以作为对排队论有兴趣的科技工作人员的初级读物. 由于笔者的水平有限，书中在材料的取舍、章节的安排以及内容的表述上定有很多不当之处，敬请读者批评指正.

唐加山

2016 年 2 月于南京

目　录

第 1 章　基本概念及术语介绍

在日常生活中, 人们经常会遇到排队现象, 例如, 去超市买东西, 如果超市只有一个收银员, 那么当买东西的顾客比较多时, 则在收银处就可能出现排队现象. 在某些大型超市中, 收银员的人数可能比较多, 即便如此, 也避免不了买东西的顾客出现排队的现象.

除了在超市以外, 排队也经常出现在其他的地方, 例如到银行办理业务会排队、在食堂用餐会排队、在加油站为汽车加油会排队、在机场以及车站买票或者候车也会排队等, 这些排队是人们能够看得见的, 而另有一些排队现象却并不那么明显, 例如在计算机通信中, 当用分组的方式传输数据时, 在数据包等待 CPU 处理的过程中就会出现排队现象, 诸如此类的例子举不胜举. 排队现象在自然界中是非常普遍的, 实际上, 在任何一个能够提供服务的系统中, 只要提供服务的资源有限, 且服务要求具有某种随机特性, 一般就会出现排队的现象.

对于排队系统 (queueing system) 或者随机服务系统 (stochastic service system) 进行研究具有非常重要的意义, 例如, 在上面关于超市的排队问题中, 一方面, 当服务员的人数比较少, 并且买东西的顾客比较多时, 顾客将花很多的时间在排队付钱上, 从而导致顾客购物的满意度下降, 长此以往顾客可能会到别的超市去购物, 由此造成顾客的流失对超市将造成巨大的经济损失. 另一方面, 如果超市聘用较多的服务员来收银, 顾客在收银这个环节上花费的时间较少, 满意度上升了, 然而超市将为支付收银员的薪水而增加不小的开销, 那么对于超市的管理层来说, 聘用多少名收银员才能使得超市具有最大的盈利是个非常值得研究的问题.

排队论 (queueing theory), 又叫随机服务系统理论, 就是对排队问题进行研究的一种理论, 它是应用概率统计与运筹学的交叉学科之一, 有着非常广泛的应用领域, 为了对排队理论进行研究, 本章介绍排队论的基本概念和知识, 以及有关的符号术语.

1.1　排队论术语及记号

1.1.1　单节点排队模型

在单节点排队系统中, 有两个方面的要素, 一个是能够提供服务的系统, 另一个就是需要服务的顾客. 在这样的系统中, 顾客如何到达, 服务系统的空间有多大,

服务员的数量有多少, 服务员的服务速度如何等都与随机服务系统的性能有着密切的联系, 通常用下面的六个特性来描述一个排队系统.

顾客输入 (或到达) 过程　用以刻画或描述顾客来到随机服务系统的规律;

服务时间　即服务系统为每一位顾客提供服务所需要的时间;

服务员数量　为顾客提供服务的服务员的数量;

等待空间　服务系统为排队等待的顾客准备的等待空间的大小;

客源数量　对服务有需求的潜在顾客的数量;

排队规则 有时也被称为 "服务规则", 即确定顾客按照什么样的顺序被接受服务的约定.

在排队论中, 我们经常用下面的信息来表示一个排队服务系统:

输入过程/ 服务时间/ 服务员数量/ 等待空间/ 客源数量/ 排队规则

这种既实用又简短的排队系统的描述是在 1953 年由 Kendall, D. G. 提出的 ([98]). 实际上, Kendall 当时只提出了前面三个描述变量, 即

输入过程/ 服务时间/ 服务员数量

这种描述排队系统的方法被后人推广到了今天的情形, 值得指出的是, Kendall 当初提出的记号在现今的论文中也是非常流行的.

对于一个排队系统来说, 顾客的到达通常是随机的, 也即顾客的到达构成了一个随机点过程, 一种特殊的随机点过程就是更新过程, 即相继到达的顾客之间的时间间隔是独立同分布的非负随机变量, 因此在很多情况下可以用这个共同的分布来描述一个随机服务系统中顾客的到达规律, 不妨假设这个分布函数用 A 来表示, 则我们可以用这个分布 A 来表征服务系统中的输入过程.

在一个简单的服务系统中, 一个服务员每次仅为一位顾客提供服务, 服务时间是一个非负随机变量, 如果不同顾客的服务时间是独立同分布的, 我们记这个共同的分布为 B, 则 B 将被用来描述服务系统中的服务时间规律. 为了更简洁的描述排队系统, 下面介绍一些在排队论中常用的特别符号.

顾客到达时间间隔 (或服务时间) 的缩写符号

D: 定长分布, 即退化为单点 $d \in (0, \infty)$ 的概率分布, 通常假设 $d = 1$ (D 是 deterministic 的首字母);

M: 指数分布 (M 是 Markovian 或 memoryless 的首字母);

E_k: k 阶埃尔朗 (Erlang) 分布;

H_k: k 阶超指数分布 (hyper-exponential distribution);

PH: 位相型 (phase type) 分布;

GI 或 G: 不对变量加特别限制的一般非负随机变量分布 (GI 是 general independent 或 general input 的首字母, G 是 general 的首字母; 沿用历史文献中的规

则, 通常用 GI 表示到达时间间隔的分布, 而用 G 表示服务时间的分布).

在排队论中, 等待空间一般有两种理解, 举例来说, 一个银行有三名办理金融业务的服务员, 顾客坐在服务员前面的旋转座椅上接受服务, 另外在大厅中还有 20 把空椅子, 方便顾客休息等待, 在这个例子中, 等待空间的一种理解就是 20, 即那 20 把空椅子, 等待空间的另外一种理解是 23, 即除了大厅中的 20 把空椅子外, 还包括三名服务员前面的旋转座椅. 如无特别说明, 本书中的等待空间是指后者.

人们通常使用缩写符号来表示不同的**排队规则**, 常见的排队规则有以下几种.

FIFO 表示先入先出 (first in, first out), 也称为 FCFS, 它表示先到先服务 (first come, first served). 在此规则下, 顾客是按照到达的先后顺序接受服务的, 这种与人们的直观相一致的假设在历史文献中得到了广泛的研究, 例如在银行、超市等服务系统中一般都采用这样的服务规则.

LIFO 后入先出 (last in, first out), 也叫 LCFS, 它表示后来先服务 (last come, first served). 在此规则下, 当服务员为一位顾客服务完毕后, 他/她将为最新到达的顾客进行服务, 例如, 在仓库系统中, 后进入仓库的物品需要先出仓.

SJF 服务时间短的先服务 (shortest job first), 或者叫 SJN(shortest job next), 或者叫 SPN(shortest process next), 即当服务员完成前一个顾客的服务后, 将在等待队列中选择要求服务时间最短的那个顾客进行服务, 这个排队规则适用于等待队列中顾客的服务时间是已知的, 或者可以提前预估的情形.

SIRO 随机序服务 (service in random order) 规则. 当服务员为一位顾客服务完毕后, 将从等待服务的顾客中随机挑选一位顾客进行服务, 例如, 某工人在对一批半成品进行加工时, 通常在完成一个加工后, 将会在剩下未加工的半成品中随机挑选一个进行加工.

PS 处理器共享 (processor sharing) 规则. 所有的顾客共享服务员的服务, 也即, 当系统中有 n 位顾客时, 每一位顾客都得到服务员服务能力的 $1/n$ 的服务, 计算机操作系统中 CPU 同时处理多个任务可以被认为是处理器共享典型的例子.

RR 轮循 (round robin) 规则. 在此情况中, 服务员每次只为一位顾客服务, 而且服务时间都为 δ (如果服务完成, 则服务时间可以更少), 服务没有结束的顾客将被放回到队列中, 直到服务员为其他顾客进行一轮服务后, 才可以再次得到服务. 当 δ 逐渐减小趋于零时, 轮循服务的极限就是处理器共享服务.

Pri 具有优先级 (priority) 的排队规则, 在此规则中, 顾客会根据优先级的高低被分成至少两个不同的类型, 当服务员完成一个顾客的服务时, 如果等待空间中有不同类型的顾客, 则高优先级的顾客优先得到服务. 当服务员正在为一个低优先级的顾客进行服务时, 如果有高优先级的顾客到达, 则服务规则又可以分成两种不同的情况: 一种情况是服务员继续为当前顾客服务完成后, 再为高优先级的顾客提供服务, 这被称为非抢占式 (non-preemptive) 服务; 另一种情况是服务员立即停下

当前的服务, 转而为高优先级顾客提供服务, 直到系统中无高优先级顾客时, 服务员再接着为刚才那个顾客提供服务, 这被称为抢占式 (preemptive) 服务.

排队系统一般分为等待制、损失制以及混合制三种类型, 当顾客到达系统时, 如果服务员全忙, 则顾客进入系统进行等待, 并按上述相应的规则接受服务, 此为等待制; 当顾客到达系统时, 如果服务员全忙, 则顾客离去, 并在其他地方寻求服务, 此为损失制; 有些服务系统为顾客准备了有限的等待空间, 当顾客到达时, 如果等待空间未满, 则进入系统接受或等待服务, 否则离开系统, 这就是混合制排队系统.

利用上述简短的缩写符号和规则就可以方便地表示一个排队系统了, 请见下例.

例 1.1.1　试叙述 "$GI/M/1/\infty/\infty/FIFO$" 所表示的排队系统.

根据排队系统的符号约定, 本例所表示的排队系统可以描述为:

(1) 顾客按照一个离散或连续时间的更新随机过程到达系统, 而且每次只到达一位顾客, 到达的时间间隔由一个一般分布所确定, 若我们把顾客编号为 $1, 2, \cdots$, 而且假设虚拟的第 0 个顾客在时间点 0 到达, 他/她不需要接受服务. 若 T_n 表示第 $n-1$ 个顾客和第 n 个顾客的到达时间间隔, 则 $\{T_n, n \geqslant 1\}$ 是一个独立同分布的随机变量序列, 相应的顾客到达系统的时间点为 $0, T_1, T_1 + T_2, \cdots$;

(2) 不同顾客的服务时间是独立同分布的指数分布随机变量, 而且与到达的过程相互独立;

(3) 系统中只有一个服务员;

(4) 等待空间为无穷;

(5) 顾客源的数量为无穷;

(6) 排队规则是**FIFO**, 即先到先服务.

在意义明确的情况下, 我们有时也仅用 "$GI/M/1$" 来表示上述排队模型.

结合上述例子, 对于本小节介绍的单节点排队模型, 可以用下面直观的图形来表示 (图 1.1).

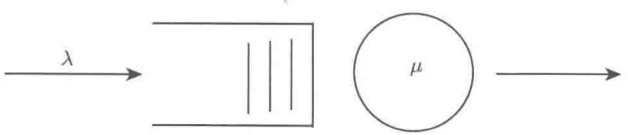


图 1.1　单节点排队模型示意图

图中, λ 表示单位时间内平均到达的顾客数, μ 表示单位时间内平均可服务的顾客数

注记 1.1.2　在随机服务系统中, 有时顾客是成批到达的, 此时, 我们通常在相应符号的右上角加上一个字母来表示, 例如, $M^X/G/1$ 表示顾客是按泊松过程成批

到达服务系统的, 而且在每一批次中, 顾客的数量所服从的分布用非负整数随机变量 X 来表示.

1.1.2 排队网络模型

上面介绍的是单节点的排队模型, 随着工程技术以及计算机网络的迅速发展, 排队网络 (queueing network) 也随处可见, 例如数据包在因特网上的传输, 在全中国或在全世界范围内经营的某物流系统等都可以看作是排队网络模型, 对于排队网络模型, 一般不容易用上面的简单记号来表示或描述, 但是可以通过绘制简单的图形来表示, 如图 1.2 所示.

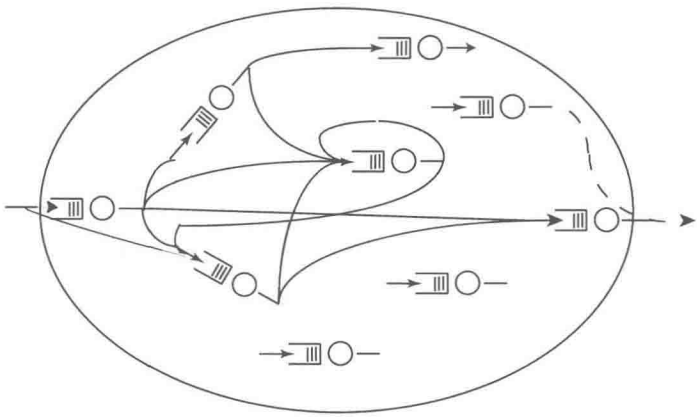


图 1.2 排队网络模型示意图

在上述模型中, 顾客在一个节点接受完服务以后并不一定立即离开系统, 而是可能转移到另外一个节点接受服务, 究竟转到哪一个节点, 可以用一个转移概率来描述, 我们将在后面的章节中给出排队网络模型更详细的介绍.

1.1.3 排队论所要研究的问题

排队论要研究的问题很多, 概括起来可以分成以下三类 ([180] 第 230 页).

(1) 性能分析问题 在实际问题中, 人们经常需要设计和构建满足一定需要的随机服务系统, 或者对正在运行的随机服务系统进行一定的优化, 要完成这些任务, 就必须先对服务系统进行性能方面的分析. 随机服务系统性能分析的主要指标包括: 一位顾客在接受服务之前需要等待的时间的分布, 在系统中逗留时间的分布, 在这里, 顾客在系统中的逗留时间是指其等待时间以及服务时间的和; 系统中 (包含, 或者不包含正在被接受服务的那位) 顾客人数的分布; 系统中工作量的分布, 工作量包括正在等待的那些顾客的服务时间的和, 再加上正在被服务的那位顾客的剩余服务时间; 服务员忙期的分布, 忙期是指服务员连续工作的时间; 系统的吞吐量

等. 在很多实际问题中, 由于系统结构的复杂性以至于无法进行严格的理论分析, 从工程应用的角度, 我们可以通过构建相应的仿真平台, 通过仿真近似给出系统的性能指标.

(2) 统计问题　在对一个随机服务系统进行研究或优化时, 首先遇到的一个问题是系统的参数是多少, 例如到达某个具体超市的顾客的到达规律是什么? 在理论分析阶段, 我们可以假设顾客的到达规律服从某种随机过程以及各个参数的具体值, 但是在实际问题中, 我们就不能做这样的无条件假设, 而是需要根据系统的实际情况, 利用观察得到的数据对其中的参数进行统计估计和推断, 利用观察得到的数据对某个估计值进行检验等.

(3) 优化问题　在优化问题中, 一方面可以通过改变系统模型的参数, 从理论方面得到系统的性能分析指标的变化, 进而对系统应该采用什么样的参数进行优化, 另一方面, 也可以通过引入适当的效用函数, 通过理论分析构造相应的优化问题, 通过对优化问题的求解反过来对系统参数进行优化等.

本书主要考虑第一类问题, 即在排队系统的参数已知的情况下, 对排队系统的性能指标进行研究和探讨.

1.2　若干概率分布

在一个普通的排队系统, 或者随机服务系统的预设参数中, 主要有两种类型的随机变量, 一类是描述顾客到达间隔的随机变量, 另一类是描述顾客服务时间的随机变量. 正是由于到达和服务的随机性, 才造成了随机服务系统 (诸如队列队长、系统队长、等待时间等) 性能指标的不确定性, 这也是学者们对排队系统进行深入研究的重要原因和意义之一. 为了更好的理解排队系统的知识, 我们在本节介绍几个在排队论中常用的随机变量.

1.2.1　指数分布

(1) 称随机变量 X 服从参数为 λ 的指数分布 (exponential distribution), 如果它的概率密度函数 (或等价的分布函数) 为

$$f(x)=\begin{cases}\lambda \mathrm{e}^{-\lambda x}, & x\geqslant 0,\\ 0, & \text{其他}.\end{cases}\qquad (F(x)=1-\mathrm{e}^{-\lambda x}, x\geqslant 0)$$

根据指数分布随机变量的密度函数, 很容易计算出 X 的数学期望和 Laplace-Stieltjes 变换 (LST) 为

$$EX=\frac{1}{\lambda},\quad \tilde{f}_X(s)=E\mathrm{e}^{-sX}=\int_0^{\infty}\mathrm{e}^{-sx}f(x)\mathrm{d}x=\frac{\lambda}{\lambda+s},\quad \mathrm{Re}s\geqslant 0.$$

注记 1.2.1 由上可知, 若指数分布随机变量的参数为 λ, 则它的数学期望就是 $\frac{1}{\lambda}$, 因此, 若 X 是参数为 λ 的指数分布随机变量, 也可以说 X 是均值为 $\frac{1}{\lambda}$ 的指数分布随机变量, 这是等价的. 另外, 指数分布有时也被称为负指数分布, 这可能是由于它的概率密度函数含有负指数的原因吧!

(2) 指数分布的一个非常重要的特性就是无记忆性, 无记忆性可以用条件概率表示为: 对于任意非负的 $s,t>0$, 有

$$P(X>s+t|X>s)=P(X>t). \tag{1.2.1}$$

举例来说, 如果用 X 表示某电子产品如日光灯的使用寿命, 并且 X 是指数分布, 则无记忆性是指: 如果知道一只日光灯已经使用了 s 这么长时间, 并且没有坏, 那么它还能用 t 这么长时间的概率与买来一只新的日光灯, 并问它能使用 t 这么长时间的概率是相同的, 这说明虽然这只灯已经使用了 s 这么长的时间, 但它还能用多久跟新灯是一样的, 仿佛它已经忘记了自己已经被使用了 s 这么长时间的历史, 这种性质被称为无记忆性. 可以证明 (证明留给读者完成), 若 X 是一个连续型非负随机变量, 且满足式 (1.2.1), 则它一定是一个指数分布随机变量, 换句话说在连续型随机变量中, 指数分布随机变量是唯一一个具有无记忆性的随机变量.

指数分布的无记忆性还可以通过非负随机变量的失效率函数来描述, 设 X 是一个非负连续型随机变量, 它的概率密度函数和分布函数分别为 $f(x)$ 和 $F(x)$, 则该分布的失效率 (风险率) 函数 $\lambda(t)$ 定义为

$$\lambda(t)=\frac{f(t)}{\bar{F}(t)}, \tag{1.2.2}$$

其中 $\bar{F}(t)=1-F(t)$, 实际上, 失效率函数可以这样来理解, 用 X 表示某元件的使用寿命, 假设该元件已经使用了 t 这么长时间, 则它在接下来的 Δt 时间内将要失效的条件概率为

$$\begin{aligned}P(t<X\leqslant t+\Delta t|X>t)&=\frac{P(t<X\leqslant t+\Delta t,X>t)}{P(X>t)}\\&=\frac{P(t<X\leqslant t+\Delta t)}{P(X>t)}\\&\approx\frac{f(t)\Delta t}{\bar{F}(t)}\\&=\lambda(t)\Delta t.\end{aligned}$$

在上式两边同时除以 Δt, 从而可见, 失效率函数实际上是指工作到时刻 t 的部件, 在接下来的单位时间内将要失效的概率强度. 容易证明, 若 X 是一个参数为 λ 的指数分布随机变量, 则其失效率函数为 $\lambda(t)\equiv\lambda$ 对于任意的 $t\geqslant 0$ 都成立.

正是由于指数分布的无记忆性, 使得若在排队系统中有关的参数可以用指数分布来刻画的话, 则对应的描述系统指标的随机过程在大部分情况下都具有无后效性, 或马氏性, 相应的性能就可以借助于经典的工具进行理论上的分析, 否则有关性能分析的进行是相当复杂和困难的.

(3) 相对于经典的指数分布而言, 下面的分布更具有一般性, 可以看成是经典指数分布的推广, 设 $0 \leqslant \alpha \leqslant 1$, 考虑如下的分布函数

$$F(x) = 1 - \alpha \mathrm{e}^{-\lambda x}, \quad x \geqslant 0.$$

考虑到 α 可以取不同的值, 这实际上是一个修正的指数分布函数族, 之所以使用这样一个称呼, 是因为随着 α 取值的不同, 该分布也会发生相应的变化, 具体地有以下情形.

如果 $\alpha = 0$, 则指数分布族退化成集中于原点的单点分布.

$$U_0(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geqslant 0. \end{cases}$$

如果 $\alpha = 1$, 则为标准的指数分布.

如果 $0 < \alpha < 1$, 因为 $F(0) = 1 - \alpha$, 即 $F(0) - F(0-) = 1 - \alpha$. 从而在 $(0, \infty)$ 上具有不完全概率密度函数 (即 $\int_0^{\infty} f(x)\mathrm{d}x < 1$)

$$f(x) = \alpha\lambda \mathrm{e}^{-\lambda x}, \quad x > 0.$$

此时, $F(x)$ 是一个混合型的分布, 对于该混合型的分布而言, 分布函数可以写成

$$F(x) = (1 - \alpha)U_0(x) + \alpha(1 - \mathrm{e}^{-\lambda x}), \quad x \geqslant 0.$$

对于修正的指数分布随机变量 X, 其数学期望以及 LST 变换为

$$EX = \frac{\alpha}{\lambda}, \quad \tilde{f}_X(s) = 1 - \alpha + \frac{\alpha\lambda}{\lambda + s}.$$

1.2.2　E_n 分布及 H_n 分布

在上一小节中, 我们已经学习了指数分布随机变量, 假设 $X_1, \cdots, X_n$ 是独立同分布于参数为 λ 的指数分布的随机变量, 其中 n 是一个大于或等于 1 的自然数, 则

$$X = X_1 + \cdots + X_n$$

就是一个参数为 λ 的 n 阶埃尔朗分布, 其概率密度函数为

$$f(x) = \begin{cases} \dfrac{\lambda^n}{(n-1)!} x^{n-1} \mathrm{e}^{-\lambda x}, & x \geqslant 0, \\ 0, & \text{其他}. \end{cases}$$

容易看出, 当 $n \equiv 1$ 时, 该分布就是标准的指数分布, 然而, 当 $n \neq 1$ 时, 它又可以看作是参数为 n 和 λ 的伽马分布, 并记为 $X \sim \Gamma(n, \lambda)$, 之所以如此, 是因为埃尔朗分布只是伽马分布的特例, 在伽马分布中, 参数 n 可以取成正实数, 一般记成 $\Gamma(\alpha, \beta)$, 其概率密度函数为

$$f(x) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\beta x}, & x \geqslant 0, \\ 0, & \text{其他}. \end{cases}$$

在伽马分布中, 通常称 α 为形状参数, β 为尺度参数, 而且 $\beta = 1$ 的伽马分布被称为是标准的伽马分布. 对于任意的 $Y \sim \Gamma(\alpha, \beta)$, 若令 $Z = \beta Y$, 则 $Z \sim \Gamma(\alpha, 1)$.

对于 $Y \sim \Gamma(\alpha, \beta)$, 直接计算表明

$$EY = \frac{\alpha}{\beta}, \quad \tilde{f}_Y(s) = \left(\frac{\beta}{\beta + s}\right)^\alpha, \quad \mathrm{Re}s \geqslant 0.$$

由此立知埃尔朗分布的相应特性: $EX = \dfrac{n}{\lambda}$, $\tilde{f}_X(s) = \left(\dfrac{\lambda}{\lambda + s}\right)^n$, $\mathrm{Re}s > 0$.

假设一位顾客来到一个随机服务系统, 他发现前面有 n 个服务台, 不同服务台的服务时间是独立同分布的指数随机变量, 如果他必须经过每个服务台的服务后才能离开系统, 则他的服务时间就是一个埃尔朗分布随机变量, 如果他以一定的概率选择某个服务台接受服务后即可离开系统, 则他的服务时间就是一个超指数分布随机变量, 因此, 如果说埃尔朗分布类似于一个串联系统的话, 那么下面介绍的超指数分布就类似于一个并联系统. 实际上, 从超指数分布的角度来看, 上述服务台的服务参数是可以不同的, 具体地, 假设 $X_1, \cdots, X_n$ 是 n 个独立的参数分别为 $\lambda_1, \cdots, \lambda_n$ 的指数分布随机变量, $\theta_1, \cdots, \theta_n$ 是 n 个非负实数, 且满足 $\theta_1 + \cdots + \theta_n = 1$, 则称以

$$f_X(x) = \sum_{k=1}^{n} \theta_k \lambda_k \mathrm{e}^{-\lambda_k x}, \quad x \geqslant 0$$

为概率密度函数的随机变量 X 为超指数分布随机变量. 若用上面的例子来解释超指数分布的话, X_k 就是第 k 个服务台的服务时间, 而 θ_k 就是顾客选择第 k 个服务台的概率, 其中 $1 \leqslant k \leqslant n$. 由定义易知

$$EX = \frac{\theta_1}{\lambda_1} + \cdots + \frac{\theta_n}{\lambda_n}, \quad DX = \frac{2\theta_1}{\lambda_1^2} + \cdots + \frac{2\theta_n}{\lambda_n^2} - \left(\frac{\theta_1}{\lambda_1} + \cdots + \frac{\theta_n}{\lambda_n}\right)^2.$$

超指数分布是混合密度 (mixture density) 的一种特例, 值得指出的是: 指数分布是几何分布的连续对应, 但是超指数分布并不是超几何分布的连续对应.

1.2.3　连续 PH 分布

PH 是 phase type 的简称, PH 分布也称为位相型分布, 从直观的角度来讲, 如果埃尔朗分布是个数确定的独立同分布的指数分布的和的分布的话, 那么连续位相型分布就是个数随机的独立的指数分布的和的分布, 因此从包含关系来说, 应该有

$$\{\text{指数分布}\} \subset \{\text{埃尔朗分布}\} \subset \{\text{连续 PH 分布}\}.$$

从严格的角度来说, 连续 PH 分布可以看成是一个时间参数连续带吸收状态的有限齐次马氏链吸收时间的分布. 具体地, 考虑一个时间参数连续的马氏链 $\{X_t, t \geqslant 0\}$, 其状态空间为 $\{1, \cdots, m, m+1\}(m \geqslant 1)$, 其中状态 $m+1$ 是吸收态, 过程的无穷小生成元 (或 Q 矩阵) 可以写成下列分块的形式

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{T} & \overline{\boldsymbol{T}^0} \\ \overline{0} & 0 \end{pmatrix},$$

其中 m 阶方阵 $\boldsymbol{T} = (T_{ij})_{m\times m}$ 满足 $T_{ii} < 0$, $T_{ij} \geqslant 0$, $i \neq j$, $1 \leqslant i, j \leqslant m$, $\overline{\boldsymbol{T}^0} = (T_1^0, \cdots, T_m^0)^{\mathrm{T}}$ 是非负列向量, 满足 $\boldsymbol{T}\bar{e} + \overline{\boldsymbol{T}^0} = \overline{0}$, 其中 $\bar{e}$ 表示全 1 列向量. 另设过程 $\{X_t, t \geqslant 0\}$ 的初始概率为 $(\bar{\alpha}, \alpha_{m+1})$, 其中 $\bar{\alpha} = (\alpha_1, \cdots, \alpha_m)$, $\bar{\alpha}\bar{e} + \alpha_{m+1} = 1$.

假设矩阵 $\boldsymbol{T}$ 具有较好的性质 (即 $|\boldsymbol{T}| \neq 0$), 使得状态 $1, \cdots, m$ 都是非常返态, 则有以下结论.

引理 1.2.2 ([163] 第 4 页引理 1.1)　*马氏链 $\{X_t, t \geqslant 0\}$ 从初始分布 $(\bar{\alpha}, \alpha_{m+1})$ 出发开始直到吸收于状态 $m+1$ 为止的时间有概率分布函数*

$$F(x) = 1 - \bar{\alpha}\exp(\boldsymbol{T}x)\bar{e}, \quad x \geqslant 0.$$

其中 $\exp(\boldsymbol{T}x) = \sum\limits_{k=0}^{\infty} \dfrac{x^k}{k!}\boldsymbol{T}^k$.

上述引理中吸收时间的分布也称为矩阵指数分布 (更多内容详见文献 [82]), 它是指数分布族从数值参数到矩阵参数的一种推广.

引理 1.2.3 ([163] 第 5 页引理 1.2)　*若用 a_j 表示吸收前过程在位相 (即状态)j 上的平均逗留时间, $j = 1, \cdots, m$, 则*

$$\bar{a} = (a_1, \cdots, a_m) = -\bar{\alpha}\boldsymbol{T}^{-1}.$$

定义 1.2.4 ([163] 第 5 页定义 1.1)　*取值于 $[0, +\infty)$ 上随机变量 X 的概率分布 $F(\cdot)$ 称为一个连续 PH 分布, 当且仅当它是一个有限状态马氏链吸收时间的分布, 此时称 $(\bar{\alpha}, \boldsymbol{T})$ 是它的 m 阶 PH 表示.*

对于连续 PH 分布, 有如下的结论 ([163] 第 5~6 页).

(1) $F(0) = 1 - \bar{\alpha}\bar{e} = \alpha_{m+1}$, 若 $\bar{\alpha}$ 是零向量, 则连续 PH 分布退化为单点分布 $U_0(x)$, 若 $0 < \bar{\alpha}\bar{e} < 1$, $F(x)$ 在 $x = 0$ 处集中一个概率质量 α_{m+1}, 并在 $(0, +\infty)$ 上