Курс лекций по математическим методам в Теории Массового Обслуживания

А.П. Серебровский

Для студентов МФТИ

Предисловие

Данный материал представляет собой конспект двухсеместрового курса лекций, читаемого автором студентам четвертого курса Московского физико-технического института (МФТИ), получающим базовую специальную подготовку по кафедре "Телекоммуникационные сети и системы"при Институте проблем передачи информации (ИППИ)РАН начиная с 2000/2001 учебного года. Курс лекций ориентирован на ознакомление с основными математическими методами, использующимися как при изучении классических систем массового обслуживания, так и в современных научных исследованиях, включающих приложения к системам и сетям передачи данных. Поскольку приходящие на кафедру студенты, как правило, имеют различную степень подготовки по классической теории вероятностей и теории случайных процессов, специальное внимание уделялось напоминанию некоторых фактов из этих дисциплин, знание которых непосредственно требовалось для понимания излагаемого материала. Настоящий конспект лекций отразил методический опыт, приобретенный автором в процессе его общения со студентами при уже неоднократном прочтении данного курса, и включил целый ряд изменений и дополнений, внесенных автором за это время.

Идея прочтения настоящего курса лекций студентам МФТИ, проходящим базовую подготовку при ИППИ РАН, возникла в процессе удивительно плодотворного (но, к несчастью, столь краткосрочного) общения автора с профессором Владимиром Калашниковым (имевшим богатейший собственный опыт чтения лекций по близкой тематике в МФТИ и на мехмате Московского Университета) в период создания при ИППИ специализированной внебюджетной лаборатории, превратившейся в настоящее время в Центр имени Владимира Калашникова. Вот почему материалы монографии профессора Калашникова В.В. по математическим методам в теории очередей составили основное ядро данного курса. Неоценимую дружескую поддержку и помощь в осмыслении второй части курса, связанной с приложениями к сетям передачи данных, оказал профессор Евсей Морозов, любезно поделившийся с автором своими материалами. Автор благодарен также организаторам Центра имени В.В. Калашникова при ИППИ и МФТИ академику Н.А. Кузнецову, М.А. Файнбергу и В.М. Седунову за понимание и постоянную поддержку в работе.

Алексей Серебровский

Основные обозначения и сокращения

с.в.	– случайная величина;
н.о.р.с.в.	– независимые одинаково распределенные с.в.;
ф.р.	– функция распределения;
$\mathbf{P}$	– вероятность;
$\stackrel{d}{=}$	– равенство по распределению;
$\mathbf{E}Y$	– математическое ожидание с.в. Y;
$\mathbf{E}(X\|A)$	– математическое ожидание с.в. X при условии события A;
$\mathbf{1}_A(x)$	$=\begin{cases}1, x \in A,\\ 0, \text{иначе};\end{cases}$
$\mathbf{1}(x)$	$=\begin{cases}1, x \geq 0,\\ 0, \text{иначе};\end{cases}$
$a \bigwedge b$	$=\min(a,b)$;
$a \bigvee b$	$=\max(a,b)$;
$(\centerdot)_+$	$=\max(0,\centerdot)$;
L-St.	– преобразование Лапласа-Стилтьеса;
$A * B(t)$	– свертка Стилтьеса ф.р. $A(t)$ и $B(t)$;
$A_*^k(t)$	– k-тая свертка Стилтьеса ф.р. $A(t)$;
$A_*^0(t)$	$= \mathbf{1}(t)$;
T_k	– момент времени прихода k-той заявки;
e_k	– k-тый интервал $(T_{k+1} - T_k)$ между моментами прихода соответствующих заявок;
$A(u)$	– общая ф.р. (для произвольного k) $\mathbf{P}(e_k \leq u)$ н.о.р.с.в. $\{e_k\}$;
a_k	– k-тый момент ф.р. A;
$a(s)$	– L-St. от ф.р. A;
$B(u)$	– общая ф.р. $\mathbf{P}(s_k \leq u)$ н.о.р.с.в. $\{s_k\}$ – длительностей обслуживания заявок;
b_k	– k-тый момент ф.р. B;
$b(s)$	– L-St. от ф.р. B;
$\mathcal{R}e\ s$	– действительная часть комплексной переменной s;
FIFO	– дисциплина обслуживания "первый-пришел-первым-ушел";
LIFO	– дисциплина обслуживания "последний-пришел-первым-ушел".

Оглавление

Глава 1

Основные понятия, терминология, примеры

Теория массового обслуживания (ТМО), являясь одной из ветвей теории вероятностей и случайных процессов, внесла и продолжает вносить заметный вклад в развитие самих этих дисциплин. Результатами ТМО и схожими математическими моделями пользуются во многих областях прикладной математики. Ориентированная на исследование приложений, ТМО развивается и как "чистая наука", и ряд ее результатов до сих пор не имеет примеров практического использования. Современная ТМО находит свое применение в таких разнообразных областях приложений как экстренная помощь, дорожное движение, компьютерные сети и системы и др. Существует достаточно широкий класс явлений реальной жизни, а также сложных систем, поведение которых во времени может быть вполне адекватно описано с помощью соответствующих математических моделей систем с очередями, изучаемых в ТМО. Но наиболее важные и современные примеры относятся к коммуникационным сетям. Именно эта область приложений ТМО стала главным источником столь стремительно возросшего интереса к этой науке в последнее время.

На протяжении всего данного курса лекций мы будем стараться изучить основные методы ТМО, а отдельные примеры конкретных систем будем рассматривать лишь в качестве иллюстрации использования этих методов. Вот почему ниже будут рассмотрены результаты далеко не всех исследованных к данному моменту моделей систем МО.

1.1 Системы МО в окружающей нас жизни

ТМО оперирует со стохастическими моделями, описывающими прохождение случайных потоков через устройства обслуживания (серверы) и, таким образом, применима всюду, где возникают такие понятия как очередь, ожидание, поломки и потери (отказы).

Пример 1. *Рассмотрим простейшую модель телефонной станции. Предположим, что $T_1, T_2, \ldots$ – моменты звонков, пронумерованные по порядку их поступления, пришедшие от абонентов с требованием их обслуживания, т.е. соединения с другими абонентами. Будем предполагать также, что станция имеет N линий связи и что любой звонок может использовать любую из имеющихся в данное время свободных линий, а если все линии заняты в момент прихода звонка, то он теряется, т.е. не обслуживается. Пусть i-тый звонок (если он не теряется) занимает некоторую линию в течение промежутка времени s_i. Вполне резонно считать, что последовательности $\{T_i\}$ и $\{s_i\}$ являются последовательностями с.в. с известными вероятностными характеристиками. Если при этом рассматривать их в качестве "входных"характеристик исследуемой системы (телефонной станции), то в процессе проектирования или реконструкции такой системы важно уметь определять или оценивать также и "выходные"характеристики ее работы, такие, например, как вероятностное распределение занятости линий, вероятность потери отдельного звонка, среднее время простоя оборудования (т.е. незанятости линий при отсутствии звонков) и др. Отметим, что в данном примере понятие очереди как таковой отсутствует благодаря предположению о потери входящего звонка в случае, когда система полностью занята.*

Пример 2. *В качестве простейшего примера системы с очередью рассмотрим следующую модель прибытия транспортных судов в порт, имеющий N погрузочно-разгрузочных терминалов. Пусть суда прибывают в порт в моменты $T_1, T_2, \ldots$, а необходимые времена их разгрузки или погрузки (т.е. обслуживания в порту) равны соответственно $s_1, s_2, \ldots$. Если в момент прибытия некоторого судна все терминалы оказываются занятыми другими судами, пришедшее судно встает на якорную стоянку и ожидает, пока хотя бы один из терминалов освободится. Можно предположить, что гавань имеет достаточные размеры (ёмкость), так что возникающая очередь может быть любой необходимой длины. Поступление судов на обслуживание, как правило, производится по порядку их прибытия. Такая дисциплина обслуживания носит специальное принятое в ТМО название FIFO (First-In-First-Out). По имеющимся значениям $\{T_i\}$ и $\{s_i\}$ можно в произвольный момент времени определить такие величины как "длина очереди", "время ожидания"в этой очереди для каждого судна на якорной стоянке и др.*

Если предположить, что ёмкость гавани конечна (ограничена), то будет возможно определить также ту часть судов, которые будут вынуждены покинуть порт по причине невозможности встать на якорную стоянку.

В этом примере предположение, что мы знаем $\{T_i\}$ и $\{s_i\}$ заранее уже не является таким неприемлемым, поскольку существует расписание движения судов и документы об объеме предстоящих погрузочно-разгрузочных работ. Но мы должны учитывать, что в реальной жизни метеоусловия на море и множество других факторов могут повлечь за

собой отклонения от имеющегося расписания. Поэтому и в этой модели вполне естественно считать величины $\{T_i\}$ и $\{s_i\}$ случайными. Но тогда и все интересующие нас "входные"характеристики рассматриваемой системы становятся случайными и встает задача определения таких величин как среднее время ожидания, средняя длина очереди и так далее.

Пример 3. *Предположим теперь, что в заводском цеху имеется M станков, каждый из которых может выйти из строя в некоторый случайный момент времени (т.е. потребовать ремонта). Любой сломавшийся станок может быть отремонтирован (обслужен) любыми из N рабочих-ремонтников. Каждый отремонтированный станок немедленно приступает к работе. Если $M > N$ и все рабочие уже заняты ремонтом, то в таком случае очередной вышедший из строя станок встает в очередь в ожидании обслуживания.*

Здесь, в отличие от предыдущих примеров, поток требований на обслуживание зависит от числа сломавшихся станков, а в случае, когда все станки выходят из строя, этот поток вообще прерывается. Случайный характер моментов наступления отказа и длительности времен ремонта затрудняют вычисление "выходных"характеристик подобных стстем обслуживания (таких как среднее число простаивающих станков; среднее количество станков, ожидающих ремонта; среднее число занятых ремонтников; среднего времени исправной работы станков, времен простоя и времен ожидания ремонта).

Пример 4. *Обратимся, наконец, к компьютерам. Их устройство и организация работы с ними порождают множество моделей систем с очередями. Рассмотрим пока лишь одну из них. Монитор, клавиатура, принтер – минимальная конфигурация, которая присутствует практически в любой компьютерной системе. Часто к одному компьютеру бывают подключены несколько удаленных терминалов. Каждый элемент такой системы с интервалами, задаваемыми программой, оператором или операционной системой, обращается к центральному процессору, который, естественно, не может одновременно обслуживать несколько устройств. Процессор вынужден распределять свои ресурсы и делает он это в соответствии с некоторыми правилами, задаваемыми, например, операционной системой. В этом примере мы имеем дело с несколькими входными потоками, требующими обслуживания от одного обслуживающего устройства – процессора. Аналогичная ситуация возникает и при обращении к серверу нескольких компьютеров, объединенных в сеть.*

Приведенные примеры демонстрируют лишь часть того широкого многообразия систем, которые являются сферой приложения ТМО. Все модели систем характеризуются наличием входящих материальных, информационных или иного рода потоков заявок (или требований) на обслуживание, где под последним понимается предоставление различных производимых в системе операций над этими потоками. Причем, если предположить к тому же, что различные заявки обслуживаются в зависимости от их приорите-

тов, то, естественно, возникнут различия и в соответствующих временах ожидания обслуживания.

1.2 Символика Д. Кендалла для обозначения моделей

Рассмотрев некоторые примеры систем МО, мы можем резюмировать следующее. **Определить систему МО – это значит задать:**

1. *входной поток заявок*, т.е. среднюю интенсивность их поступления и статистическую модель;

2. *механизм обслуживания*, т.е. определить, когда обслуживание допустимо, сколько заявок могут обслуживаться одновременно и каково распределение длительности обслуживания;

3. *дисциплину обслуживания*, т.е. порядок выбора на обслуживание очередной заявки из числа всех ожидающих.

Будем предполагать далее, что входящий в систему МО поток требований (или заявок) на обслуживание состоит из случайных моментов времени $T_1, T_2, \ldots$ последовательно пронумерованных по порядку их прихода в систему. Полагая $T_0 = 0$, этот же поток $\mathbf{T} = \{T_k\}_{k\geq 0}$ можно задать с помощью случайной последовательности $\mathbf{e} = \{e_k\}_{k\geq 0} = \{e_0, e_1, e_2, \ldots\}$ временных интервалов между последовательными моментами прихода требований, так что $T_{k+1} = T_k + e_k,\ k \geq 0$.

Механизм обслуживания будем задавать последовательностью временных интервалов $\mathbf{s} = \{s_i\}_{i>0} = \{s_1, s_2, s_3, \ldots\}$, затраченных на обслуживание соответствующих требований (или заявок).

Кроме того, во всех рассматриваемых далее в нашем курсе моделях систем с очередями мы будем предполагать, что случайные последовательности $\{e_k\}$ и $\{s_i\}$ являются независимыми.

Существует достаточно широкий спектр дисциплин обслуживания. Поступающие в систему МО требования могут обрабатываться либо в зависимости от порядка их поступления ($FIFO$, $LIFO$), либо могут подразделяться на различные приоритетные классы, образуя несколько очередей на обслуживание, либо пропорционально разделять между собой предоставляемое им общее время обслуживания. В каждом конкретном случае при описании системы МО вид дисциплины обслуживания будет указываться особо.

Общепринятая в настоящее время условная форма символьного обозначения моделей МО была предложена Д. Кендаллом и представляет собой, как правило, четыре символьные позиции, разделенные вертикальными черточками: **.** | **.** | **.** | **.** . При этом в первой позиции указывают вид входного потока, механизм обслуживания – во второй, в третьей позиции

указывается число серверов, работающих в системе параллельно, а четвертая позиция отражает ёмкость накопителя, т.е. допустимый размер очереди на обслуживание.

Таким образом, символ $\cdot \mid \cdot \mid N \mid M$ – означает, что модель системы содержит N серверов и M мест для ожидающих обслуживания.

Для обозначения вида входящего потока и механизма обслуживания используются идентичные символы:

M – (Марковский) т.е. с.в. $\{e_k\}$ или $\{s_i\}$ имеют экспоненциальную ф.р.
E_r – распределение Эрланга порядка r
D – детерминированное
G – распределение общего вида (General)
GI – с.в. $\{e_k\}$ или $\{s_i\}$ имеют функцию распределения общего вида (General) и являются н.о.р.с.в. (Independent).

1.3 Примеры алгебраического описания моделей

1. *Модель $G|G|1|\infty$ с дисциплиной обслуживания типа $FIFO$.*

Наиболее важной характеристикой модели является последовательность $\mathbf{w} = \{w_1, w_2, \ldots\}$ времён ожидания обслуживания (от момента прихода соответствующего требования в систему до его попадания в сервер). Нетрудно вывести рекуррентное уравнение для этих величин.

Рассмотрим k-тое требование, поступившее в систему в момент T_k. Оно ожидает в течение промежутка времени w_k, а затем обслуживается в течение s_k и только в момент $(T_k + w_k + s_k)$ покидает систему. В свою очередь, $(k+1)$-е требование поступает в момент $T_{k+1} = T_k + e_k$ и, если $e_k \le w_k + s_k$, то оно будет ожидать обслуживания в течение $w_{k+1} = w_k + s_k - e_k$. Если же $e_k > w_k + s_k$, то это означает, что $(k+1)$-е требование поступило в незанятую (уже освободившуюся) систему и, следовательно, ему вообще не придется ожидать, т.е. тогда $w_{k+1} = 0$. Если обозначить величиной w_0 время "прогрева"сервера в начале работы системы, т.е. время, в течение которого (начиная с $T_0 = 0$) сервер не может обслуживать поступающие требования, и доопределить $s_0 = 0$, то мы придем к следующему рекуррентному соотношению:

$$w_{k+1} = (w_k + s_k - e_k)_+ , \quad k \ge 0, \tag{1.1}$$

где использовано обозначение $(\cdot)_+ = \max(0, \cdot)$.

По индукции для любого $k \ge 1$ можно получить следующее решение уравнения (1.1):

$$\begin{aligned} w_k = \max\{0, (s_{k-1} - e_{k-1}), (s_{k-1} - e_{k-1}) + (s_{k-2} - e_{k-2}), \ldots, \\ (s_{k-1} - e_{k-1}) + (s_{k-2} - e_{k-2}) + \ldots + (s_1 - e_1), \\ (s_{k-1} - e_{k-1}) + (s_{k-2} - e_{k-2}) + \ldots + (s_1 - e_1) + (s_0 - e_0) + w_0\}. \end{aligned} \tag{1.2}$$

Рассмотрим далее ещё один процесс $\mathbf{V} = \{V(t)\}_{t\geq 0}$, равный времени, в течение которого сервер будет занят, обслуживая только заявки, поступившие до момента t. Траектории этого процесса непрерывны справа и имеют кусочнолинейную форму, причем $V(0) = w_0$, а в моменты T_k траектории претерпевают скачок на величину s_k. В промежутке между такими моментами тангенс угла наклона траектории $V(t)$ к оси времени равен (-1), за исключением тех точек, когда, достигнув нуля при некотором значении t (в момент полного освобождения сервера), функция $V(t)$ остается равной нулю вплоть до момента прихода очередного требования. В этой точке траектория вновь скачком возрастает на величину времени обслуживания прибывшего требования. Время ожидания для этого требования, очевидно, будет равно нулю, что полностью согласуется с данным выше определением: $\lim_{t\uparrow T_k} V(t) = w_k$.

2. *Модель* $(G|G|1|\infty) \to (G|1|\infty) \to \ldots \to (G|1|\infty)$.

Рассмотрим простейшую сеть с очередью (или модель многофазной системы с очередью в каждой фазе), работа которой представляет собой прохождение приходящего требования через последовательность из N серверов (или фаз обслуживания). Непосредственно перед каждым сервером имеется накопитель неограниченной ёмкости (∞).

Для вывода алгебраических соотношений, описывающих модель такой системы, положим $\mathbf{s_k} = (s_k(1), s_k(2), \ldots, s_k(N)),\ k \geq 1$ – N-мерный вектор времён обслуживания k-го требования на серверах $1, 2, \ldots, N$ соответственно. И пусть $w_k(j),\ 1 \leq j \leq N$ – полное время прохождения k-го требования через первые j фаз обслуживания. Рассмотрим также вектор $\mathbf{w_0}$, координаты которого имеют следующий смысл: $w_0(1)$ – время "прогрева"первого сервера, $w_0(2)$ – максимум из времён "прогрева"первого и второго серверов, $\ldots$, $w_0(N)$ – максимум из времён "прогрева"всех N рассматриваемых серверов.

Обозначим далее $\mathbf{w} = (\mathbf{w_0}, \mathbf{w_1}, \mathbf{w_2}, \ldots)$, где при каждом $k \geq 0$ $\mathbf{w_k} = (w_k(1), w_k(2), \ldots, w_k(N))$. Тогда можно написать следующие рекуррентные соотношения:

$$\begin{aligned} w_{k+1}(1) &= s_{k+1}(1) + (w_k(1) - e_k)_+; \\ w_{k+1}(j) &= s_{k+1}(j) + (w_k(j) - e_k) \vee w_{k+1}(j-1),\ 1 < j \leq N; \end{aligned}$$

в справедливости которых нетрудно убедиться самостоятельно.

1.4 Проблемы построения и исследования систем

Как было уже сказано выше, знание "входных"характеристик системы МО (например, последовательностей $\mathbf{e}$ и $\mathbf{s}$) позволяет находить некоторые важные "выходные"характеристики, такие как w_k – времена ожидания обслуживания k-тым требованием. Однако, такие решения все же нельзя в

полной мере рассматривать как решение задачи анализа системы с очередью до тех пор, пока мы не учитываем вероятностный характер последовательностей исходных данных. С учетом же их вероятностной природы задача анализа будет состоять в определении выходных характеристик системы, таких как поведение длины очереди $Q(t)$, среднего времени ожидания в очереди $\mathbf{E}w_k$ для k-го требования, вероятности потери требования в случае, когда размер буфера-накопителя ограничен (т.е. $N < \infty$), средних времён занятости и простоя серверов и др. В нашем курсе будут рассмотрены некоторые математические методы, позволяющие определять указанные характеристики. Тем не менее, следует понимать, что даже знание существующих методов, к сожалению не гарантирует успеха в решении любой практической проблемы, т.к. связи входных и выходных характеристик реальных систем слишком сложны и нелинейны. Недостаток информации о системе также является частой проблемой, возникающей при решении практических задач. Это приводит к необходимости решать задачи оценивания, сравнивать различные модели друг с другом, искать приближенные асимптотические решения. Искусство применения знаний по ТМО состоит в понимании того, что рассматриваемая модель реальной системы должна не только отвечать требованию простоты описания и поддаваться решению с помощью существующих математических методов, но и соответствовать имеющимся "сырым"данным о системе, или, другими словами, быть адекватной реальной системе. Впрочем, критический подход к получаемым результатам и внимательное отношение к адекватности выбираемых моделей является обычным требованием для практического применения любых разделов прикладной математики.

Часто целью исследования моделей системы МО является желание улучшить существующую систему путем некоторых изменений в ней. Следует помнить, что даже возникающие в системе перегрузки чаще зависят не столько от характеристик самой системы, сколько от нерегулярности ее работы. Решение о модернизации системы будет экономически оправдано, если, например, возникающие очереди приводят к недопустимому увеличению общего времени пребывания требований в системе, или, наоборот, если слишком малый поток заявок приводит к простою в течение большой доли времени дорогостоящего оборудования. При этом решение о модернизации системы может основываться как на интуиции, так и на теоретическом рассмотрении, а также и на экспериментальном исследовании (т.е. на моделировании, проведенном возможно даже не в полном масштабе системы МО).

Модификация систем с очередью может включать в себя следующие аспекты:

1. *Модификация входящих потоков:*

– изменить полную среднюю интенсивность поступления требований, например, исключая некоторые требования или классы требований;

– ввести управление моментами поступления отдельных требований в

соответствии с расписанием, предназначенным обычно для образования регулярных потоков;

– поощрять или штрафовать присоединение к очереди в зависимости от числа требований, уже находящихся в данный момент в системе.

2. *Модификация механизма обслуживания:*

– уменьшить среднюю длительность обслуживания;

– снизить коэфф. вариации длительности обслуживания;

– снижать длительность обслуживания при загрузках системы, превышающих среднедопустимую;

– изменить пропускную способность системы, например, увеличив число имеющихся серверов;

– увеличить общее время доступности обслуживающих приборов либо в среднем, либо в периоды повышенной нагрузки, сократив интервалы их "отдыха", "ремонта"и "регламентных проверок".

3. *Модификация дисциплины обслуживания:*

– предоставить приоритет (абсолютный или относительный) наиболее "важным"требованиям, т.е. требованиям, для которых стоимость единицы времени ожидания велика;

– предоставить приоритет тем требованиям, для которых априори ожидается малая длительность обслуживания;

– в многолинейных системах ввести или изменить порядок распределения требований по отдельным обслуживающим серверам.

Глава 2

Некоторые факты из теории вероятностей

Для успешного изучения моделей с очередями нам потребуются некоторые знания из теории вероятностей и теории случайных процессов и умение правильно применять их в конкретных ситуациях. Цель настоящей главы - освежить в памяти некоторые известные факты и ввести основные математические обозначения, которые будут использоваться при дальнейшем изложении материала.

Конечно, при конструировании вероятностных моделей систем, а особенно при попытках проверить их адекватность, очень важным является исследование достаточно тонких математических моментов, начиная с правильного выбора соответствующего вероятностного пространства. Однако, поскольку мы не будем затрагивать эти вопросы в рамках нашего курса, то и в данном обзоре мы в основном коснёмся лишь вычислительных аспектов.

2.1 Случайные величины и их распределения

В рамках теории вероятности мы, как правило, не можем различить элементарные события ω. Это в некоторой степени находит своё отражение в том, что мы рассматриваем события, которые являются подмножествами σ-алгебры $\mathcal{A}$ исходного вероятностного пространства $(\Omega,\ \mathcal{A},\ \mathbf{P})$, а не различные точки $\omega \in \Omega$. Но даже события могут оказаться ненаблюдаемыми. Вот почему на практике всегда выбирают некоторые определённые переменные с целью выразить модель в числах. Эти переменные могут также зависеть от "случая".

(I) Простейший (одномерный) вариант

Для начала предположим, что эти переменные действительны, т.е. принимают значения на действительной прямой $R^1 = (-\infty, \infty)$. Обозначим через $\mathcal{B}$ - борелевскую σ-алгебру на R^1 (т.е. набор всех всевозможных интер-

валов). Определим *случайную величину* (с.в.) ξ как измеримую функцию, переводящую $\Omega \to R^1$ и такую, что для любого борелевского множества $B \in \mathcal{B}$ множество $\xi^{-1}(B) = \{\omega : \xi(\omega) \in B \in \mathcal{B}\}$ принадлежит к $\mathcal{A}$ (т.е. измеримо).

Множества $\xi^{-1}(B), \quad B \in \mathcal{B}$, очевидно, также образуют σ-алгебру в Ω, которую мы обозначим $\sigma(\xi)$. Ясно, что $\sigma(\xi) \subset \mathcal{A}$ и, следовательно, каждая с.в. генерирует свою σ-алгебру в $\mathcal{A}$.

Если мы согласимся, что с.в. ξ наблюдаема, то мы должны допустить, что для наблюдателя все эти множества $\xi^{-1}(B), \quad B \in \mathcal{B}$ - различны и поэтому необходимо, чтобы все эти множества были *событиями* - но именно это и декларируется в определении с.в.. Но тогда мы можем определить вероятность $\mathbf{P}_\xi(B) = \mathbf{P}(\xi \in B)$, $B \in \mathcal{B}$ и рассмотреть другое вероятностное пространство $(R^1, \mathcal{B}, \mathbf{P}_\xi)$ в котором элементарными событиями будут сами значения с.в. ξ.

Это позволяет изучать ξ без обращения к исходному вероятностному пространству $(\Omega, \mathcal{A}, \mathbf{P})$. При этом набор вероятностей $\mathbf{P}_\xi(B), \ B \in \mathcal{B}$ называют *распределением с.в.* ξ .

Но если рассмотреть не весь набор $\mathcal{B}$, а только множества B вида $B = (-\infty, x)$, то это приведет к функции $F_\xi(x) = \mathbf{P}_\xi\Big((-\infty, x)\Big) = \mathbf{P}(\xi \leqslant x)$, которая называется *функцией распределения с.в.* ξ .

Функция распределения (ф.р.) обладает следующими свойствами:

(i) если $x_1 \leqslant x_2$, то $F_\xi(x_1) \leqslant F_\xi(x_2)$;
(ii) $\lim_{x\to-\infty} F_\xi(x) = 0, \quad \lim_{x\to\infty} F_\xi(x) = 1$;
(iii) $lim_{x\downarrow x_0} F_\xi(x) = F_\xi(x_0)$ для всех $-\infty < x_0 < \infty$ (непрерывность справа).

Для классификации ф.р. случайных величин заметим, что, как известно, любая ф.р. может быть представлена как сумма не более чем трёх компонент, а именно: дискретной, абсолютно непрерывной и сингулярной (см. [1]).

Для дискретных с.в. ξ, принимающих значения $\{x_k\}, \ (k = 0, 1, 2, \ldots)$ с вероятностями $\{p_k\}$ соответственно, ф.р.

$$F_\xi(x) = \mathbf{P}(\xi \leqslant x) = \sum_{k: x_k \leqslant x} p_k \tag{2.1}$$

имеет вид ступенчатой функции.

Абсолютно непрерывная ф.р. дифференцируема и для неё

$$F_\xi(x) = \int_{-\infty}^{x} f_\xi(u)\, du, \text{ где } f_\xi(x) = \frac{dF_\xi(x)}{dx} - \tag{2.2}$$

функция, называемая *плотностью распределения вероятностей.*

В сингулярном же случае функция $F_\xi(x)$ являясь непрерывной, имеет $\frac{dF_\xi(x)}{dx} = 0$ почти всюду, что означает, что множество, где $F_\xi(x)$ не имеет производной, имеет Лебегову меру ноль. В рамках нашего курса мы не будем иметь дело с подобными сингулярными случаями.

Для единообразного описания дискретных и непрерывных с.в. удобно использовать *интеграл Стилтьеса*, который определяется следующим образом: для неубывающей $F(x)$ и непрерывной функции $\varphi(x)$ рассмотрим две последовательности точек: $\{t_k\}$ и $\{\zeta_k\}$, такие, что $t_{k-1} < \zeta_k \leqslant t_k$ и сформируем следующую сумму

$$\sum_k \varphi(\zeta_k)\ [F(t_k)\ -\ F(t_{k-1})]$$

При условии $\max|t_k - t_{k-1}| \to 0$ эти суммы стремятся к пределу, не зависящему от последовательностей $\{t_k\}$ и $\{\zeta_k\}$. Этот предел и называют *интегралом Стилтьеса от функции φ по F* и обозначают

$$\int \varphi(x)\ dF(x) \tag{2.3}$$

Используя интеграл Стилтьеса, математическое ожидание (или среднее значение) некоторой с.в. ξ можно будет теперь представить в следующем виде:

$$\mathbf{E}\xi\ =\ \int_{-\infty}^{\infty} x\ dF_\xi(x)$$

Или (в общем случае) момент k-го порядка как

$$\mathbf{E}\xi^k\ =\ \int_{-\infty}^{\infty} x^k\ dF_\xi(x) \tag{2.4}$$

Замечание 2.1. *Если $F(x)$ - ф.р., имеющая плотность, то $dF(x) = f(x)\ dx$ и интеграл Стилтьеса автоматически переходит в обычный интеграл Римана.*

Однако, нужно запомнить, что

Замечание 2.2. *Плотность ф.р. может и не существовать, но интеграл Стилтьеса существует всегда!*

(II) Общий случай

Мы можем предположить, что ξ принимает не только действительные значения, а что её значения лежат в некотором пространстве более общего вида. Тогда все конструкции будут естественным обобщением только что рассмотренного одномерного случая.

Пусть ξ, например, принимает свои значения в некотором пространстве U с введённой в нем σ-алгеброй $\mathcal{U}$. Тогда ξ будет случайной величиной, если для любого множества $B \in \mathcal{U}$

$$\xi^{-1}(B) \;=\; \{\omega : \xi(\omega) \in B\} \in \mathcal{A},$$

а соответствующий набор вероятностей

$$\mathbf{P}_\xi(B) \;=\; \mathbf{P}(\xi \in B),\; B \in \mathcal{U}$$

будет называться *распределением вероятностей с.в. ξ* .

Если же $\xi = (\xi_1, \ldots, \xi_n)$ - с.в. со значениями в n-мерном Евклидовом пространстве R^n, то мы можем определить ф.р.

$$F_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n) \;=\; \mathbf{P}(\xi_1 \leqslant x_1, \ldots, \xi_n \leqslant x_n).$$

В этом случае можно использовать также следующую терминологию: распределения $\mathbf{P}_{\xi_1,\ldots,\xi_n}$ и ф.р. $F_{\xi_1,\ldots,\xi_n}$ - называются *совместными*. Зная совместные распределения (совместную ф.р.), можно определить любое маргинальное распределение (маргинальную ф.р.) следующим образом:
для любого случайного вектора $(\xi_{i_1}, \ldots, \xi_{i_k})$, $1 \leqslant i_1 < \ldots < i_k \leqslant n,\; k \geqslant 1$

$$\mathbf{P}(\xi_{i_1} \in B_1, \ldots, \xi_{i_k} \in B_k) \;=\; \mathbf{P}(\xi_1 \in B_1^*, \ldots, \xi_n \in B_n^*),$$

где $B_i^* = U$, если $i \neq \{i_1, \ldots, i_k\}$ и $B_{i_j} = B_j,\;\; 1 \leqslant j \leqslant k$ в противном случае; и соответственно

$$F_{\xi_{i_1},\ldots,\xi_{i_k}}(x_1, \ldots, x_k) \;=\; F_{\xi_1,\ldots,\xi_n}(x_1^*, \ldots, x_n^*),$$

где $x_i^* = \infty$, если $i \not\subseteq \{i_1, \ldots, i_k\}$ и $x_{i_j}^* = x_j,\; 1 \leqslant j \leqslant k$ в противном случае.

Однако, если мы знаем только маргинальные распределения (или маргинальные ф.р.) с.в. ξ , то этого не достаточно (в общем случае) для того, чтобы по ним восстановить совместное распределение (или совместную ф.р.). Действительно, если мы, например, знаем ф.р. F_1 и F_2 двух с.в. ξ_1 и ξ_2 соответственно, то они не определяют полностью их совместную ф.р. $F_{\xi_1,\xi_2}(x_1, x_2)$, поскольку можно предложить по меньшей мере две (различные!) функции в качестве их совместной ф.р. (т.е. имеющие F_1 и F_2 как свои маргинальные), а именно:

$$\begin{aligned} F^{(1)}_{\xi_1,\xi_2}(x_1, x_2) &= F_1\; F_2; \\ F^{(2)}_{\xi_1,\xi_2}(x_1, x_2) &= \min\{F_1,\; F_2\}. \end{aligned}$$

2.2 Способы описания случайных величин

Итак, случайные величины (с.в.) представляют собой некоторую измеримую функцию. Например, результат бросания одной игральной кости может быть описан как с.в., принимающая дискретные значения от 1 до 6. Количество запросов, поступивших за час в систему бронирования авиабилетов, или количество заявок на обслуживание, поступивших в компьютерную систему, также являются примерами случайных величин. Случайными величинами являются также и интервалы времени между прибытиями в компьютерную систему двух последовательных заявок $\{e_i\}$ или времена их обслуживания $\{s_i\}$. Последние две с.в. являются непрерывными, в то время как первые могли принимать только дискретные значения.

Дискретные с.в. описываются значениями, которые они могут принимать, и соответствующими вероятностями для каждого из этих значений. Таким образом, если возможными значениями с.в. X являются неотрицательные целые числа, то величины

$$p_k = \mathbf{P}(X = k),\ k = 0, 1, 2, \ldots$$

определяют соответствующую вероятность того, что с.в. X принимает значение k. Необходимыми являются следующие требования:

$$\mathbf{P}(X = k) \geqslant 0$$
$$\sum_{k=0}^{\infty} \mathbf{P}(X = k) \ = \ 1$$

Для нашего первого примера (с бросанием игральной кости) все возможные значения равновероятны, т.е.

$$\mathbf{P}(X = k) = \frac{1}{6},\ \ k = 1, 2, \ldots, 6.$$

Другими часто встречающимися примерами дискретных с.в. являются следующие:

♦ **С.в. в испытаниях Бернулли**. Рассмотрим случайный эксперимент, имеющий только два возможных исхода (как при бросании монеты), т.е. $k = 0, 1$. Для такой с.в. X

$$\mathbf{P}(X = 0) = 1 \ - \ p \ \text{ и } \ \mathbf{P}(X = 1) = p \ \text{ с } \ 0 \ < \ p \ < \ 1. \qquad (2.5)$$

♦ **Биноминальная с.в.**. Когда эксперимент с двумя исходами повторяют n раз и при этом все последовательные попытки являются независимыми (независимые испытания Бернулли), то для с.в. X, представляющей собой количество выпавших единиц в этих испытаниях, оказывается

$$\mathbf{P}(X = k) \ = \ C_n^k \, p^k (1 - p)^{n-k}, \ \text{ где } \ C_n^k \ - \qquad (2.6)$$

биномиальный коэффициент, равный как известно $\frac{n!}{k!(n-k)!}$.

♦ **Геометрическая с.в.**. Пусть теперь эксперимент с двумя возможными исходами повторяется несколько раз, а с.в. X представляет собой число попыток, потребовавшихся до первого выпадения единицы (включая и текущее испытание). Тогда

$$\mathbf{P}(X = k) \;=\; p\,(1-p)^{k-1}, \quad k = 1, 2, \ldots \tag{2.7}$$

♦ **Пуассоновская с.в.**. Для этой с.в. вероятность наступления k событий задается следующей формулой

$$\mathbf{P}(X = k) \;=\; \frac{\Lambda^k}{k!}\; e^{-\Lambda}, \quad \Lambda > 0, \quad k = 0, 1, 2, \ldots \tag{2.8}$$

где величина Λ называется параметром распределения Пуассона.

Последние две с.в. играют очень важную роль в ТМО и мы будем достаточно часто встречаться с ними в нашем курсе.

Непрерывные с.в. X, принимающие, например, все возможные неотрицательные значения $0 \leqslant x \;<\; \infty$, полностью описываются своей ф.р.

$$F(x) \;=\; \mathbf{P}\big(X \leqslant x\big), \tag{2.9}$$

которая задает для каждого x вероятность того, что с.в. X примет значение, не превосходящее этот x. Ясно, что $F(0) = 0$, а $\lim_{x\to\infty} F(x) = 1$. Для любых неотрицательных $x < y$ имеем:

$$\begin{aligned} F(x) \;&\leqslant\; F(y) \\ \mathbf{P}(x < X \leqslant y) \;&=\; F(y) \;-\; F(x) \end{aligned} \tag{2.10}$$

Заметим, что плотность распределения вероятности $f(x)$, если она существует, может быть определена для любого $0 \leqslant x < \infty$ как

$$f(x) \;=\; \lim_{\Delta x \to 0+} \frac{\mathbf{P}(x < X \leqslant x + \Delta x)}{\Delta x} \;=\; \frac{dF(x)}{dx} \tag{2.11}$$

$$f(0) = 0, \quad \int_0^\infty f(u)\, du = 1\ . \tag{2.12}$$

В абсолютно непрерывном случае $f(x)$ эквивалентна ф.р., поскольку так же позволяет полностью описать поведение рассматриваемой с.в..

Некоторые примеры с.в., имеющих абсолютно непрерывные ф.р., будут подробно рассмотрены немного ниже.

А пока введем еще несколько функций, допускающих эквивалентное описание поведения с.в., имеющих абсолютно непрерывную ф.р.

Заметим прежде всего, что если в качестве случайной величины X мы рассматриваем, например, длительность безотказной работы (время жизни)

некоторого прибора или устройства, то удобнее оперировать с дополнительной к ф.р. $F(x)$ функцией, а именно с функцией

$$\mathcal{F}(x) \;=\; \mathbf{P}(X > x) \;=\; 1 \;-\; F(x) \;=\; \int_x^\infty f(u)du. \qquad (2.13)$$

Поскольку в этом примере значение $F(x)$ задает вероятность того, что прибор отказал до момента x, в то время как $\mathcal{F}(x)$ будет определять вероятность того, что этот прибор не откажет до момента x, то эту новую функцию резонно назвать *функцией надёжности* ("survivor function").
Очевидно, что $\mathcal{F}(0) = 1,\ \mathcal{F}(\infty) = 0$ и $\mathcal{F}(x)$ является невозрастающей функцией x. При этом плотность распределения может легко быть определена и через функцию надежности, а именно

$$f(x) \;=\; -\,\mathcal{F}'(x).$$

Эквивалентной ко всем этим трем уже рассмотренным функциям является еще одна функция $r(x)$, которая называется *интенсивностью отказа* (или *интенсивностью отказа, зависящей от возраста*). Под эквивалентностью мы здесь понимаем тот факт, что каждая выбранная из рассматриваемых четырех функций, может быть однозначно выражена через любую из оставшихся трех.

Функция $r(x)$ определяется как предел отношения вероятности отказа некоторого элемента на интервале $(x, x + \Delta x]$ к длине этого интервала Δx при $\Delta x \to 0$ для элемента, о котором известно, что он не отказал до момента x:

$$r(x) \;=\; \lim_{\Delta x \to 0+} \frac{\mathbf{P}(x < X \leqslant x + \Delta x \mid X > x)}{\Delta x}. \qquad (2.14)$$

Так что величина $r(x)\Delta x$ определяет с точностью до $o(\Delta x)$ вероятность почти немедленного отказа элемента, проработавшего уже время x (т.е. находящегося в "возрасте"x).

Из данного определения нетрудно получить выражение функции $r(x)$ через уже рассмотренные ранее функции. Действительно, из (2.14) по определению условной вероятности с использованием (2.11) и (2.13), получим

$$\begin{aligned} r(x) \;&=\; \lim_{\Delta x \to 0} \frac{\mathbf{P}\Big(\{x < X \leqslant x + \Delta x\} \bigcap \{X > x\}\Big)}{\Delta x} \; \frac{1}{\mathbf{P}(X > x)} \\ = \lim_{\Delta x \to 0} \frac{\mathbf{P}\Big(x < X \leqslant x + \Delta x\Big)}{\Delta x} \frac{1}{\mathbf{P}(X > x)} &= \frac{F'(x)}{1 - F(x)} = -\frac{\mathcal{F}'(x)}{\mathcal{F}(x)} \qquad (2.15) \end{aligned}$$

Из последнего в частности вытекает, что

$$r(x) \;=\; \frac{f(x)}{1 \;-\; F(x)}\ , \quad \text{и, кроме того, } r(x) = -\frac{d}{dx}\Big(\ln \mathcal{F}(x)\Big), \qquad (2.16)$$

$$\mathcal{F}(x) \;=\; \exp\Big(-\int_0^x r(u)du\Big), \quad f(x) \;=\; r(x)\ \mathcal{F}(x). \qquad (2.17)$$

Вероятностный смысл $r(x)$ можно пояснить также, рассматривая распределение с.в. ξ_x - остаточного времени жизни некоторого прибора. Действительно, если ξ - с.в., характеризующая время жизни этого прибора, а x - момент времени, о котором известно, что $\xi > x$, то остаточное время жизни прибора (начиная отсчет от этого x) есть $\xi_x = \xi - x$. Но тогда ф.р. ξ_x можно выразить через ф.р. ξ следующим образом:

$$\begin{aligned} F_{\xi_x}(y) &= \mathbf{P}(\xi_x \leqslant y) = \mathbf{P}\Big(\xi - x \leqslant y \mid \xi > x\Big) \\ &= \mathbf{P}\Big(\xi \leqslant x + y \mid \xi > x\Big) = \frac{\mathbf{P}\Big(\{\xi \leqslant x+y\} \bigcap \{\xi > x\}\Big)}{\mathbf{P}(\xi > x)} \\ &= \frac{\mathbf{P}(x < \xi \leqslant x+y)}{1 - F_\xi(x)} = \frac{F_\xi(x+y) - F_\xi(x)}{\mathcal{F}_\xi(x)} \end{aligned} \tag{2.18}$$

Сравнивая полученное выражение с (2.15), видим, что

$$r(x) = \frac{d}{dy}\mathbf{P}(\xi_x \leqslant y)\Big|_{y=0} = F'_{\xi_x}(y)\Big|_{y=0} = f_{\xi_x}(0)\ , \tag{2.19}$$

то есть интенсивность отказа $r(x)$ равна значению в нуле плотности распределения остаточного времени жизни ξ_x.

2.3 Моменты с.в., теорема о вычислении моментов

В ТМО, как правило, рассматриваются неотрицательные случайные величины и поэтому для определения k-го момента с.в. ξ общая формула (2.4) приобретает вид

$$\mathbf{E}\xi^k = \int_0^\infty x^k\ dF_\xi(x) \tag{2.20}$$

Однако, более простое и удобное выражение для вычисления моментов дает следующая теорема:

Теорема 2.1. *Пусть ξ - неотрицательная с.в. с ф.р. $F_\xi(x)$. Тогда*
1) $\mathbf{E}\xi < \infty$ *тогда и только тогда, когда* $m_1 = \int_0^\infty [1 - F_\xi(x)]\ dx < \infty$. *Кроме того,* $m_1 = \mathbf{E}\xi$*;*
2) *Для любого* $s \geqslant 2$ *аналогично* $\mathbf{E}\xi^s < \infty$ *тогда и только тогда, когда* $m_s = s\int_0^\infty x^{s-1}[1 - F_\xi(x)]\ dx < \infty$. *Кроме того,* $m_s = \mathbf{E}\xi^s$.

Доказательство.
1) (I) (необходимость) Если $\mathbf{E}\xi = \int_0^\infty x\ dF_\xi(x) < \infty$, то $m_1 < \infty$ и $m_1 = \mathbf{E}\xi$.
Из существования несобственного интеграла $E\xi$ следует, что

$$\lim_{L\to\infty}\int_L^\infty x\ dF_\xi(x) = 0. \tag{2.21}$$

Но т.к. справедлива очевидная оценка

$$\int_L^\infty x\, dF_\xi(x) \geqslant L \int_L^\infty dF_\xi(x) = L\,(1 - F_\xi(L)), \tag{2.22}$$

то из (2.21) вытекает, что и

$$\lim_{L\to\infty} L\,(1 - F_\xi(L)) = 0. \tag{2.23}$$

Зафиксировав теперь некоторое произвольное число $L > 0$, по формуле интегрирования по частям получим:

$$\int_0^L [1 - F_\xi(x)]\, dx = L\,(1 - F_\xi(L)) + \int_0^L x\, dF_\xi(x). \tag{2.24}$$

Поскольку у обоих слагаемых в правой части последнего равенства существуют пределы при $L \to \infty$, то существует и предел в левой части этого равенства, причем равный m_1. Тем самым (с учетом (2.23)), необходимость доказана полностью.

(II) (достаточность) Если $m_1 < \infty$, то $\mathbf{E}\xi < \infty$ и $\mathbf{E}\xi = m_1$.
Из равенства (2.24) следует, что

$$\int_0^L x\, dF_\xi(x) \leqslant \int_0^L [1 - F_\xi(x)]\, dx \leqslant \int_0^\infty [1 - F_\xi(x)]\, dx = m_1 < \infty. \tag{2.25}$$

Таким образом, для любого $L > 0$

$$\int_0^L x\, dF_\xi(x) \leqslant m_1 < \infty, \tag{2.26}$$

что и означает существование $\mathbf{E}\xi < \infty$. Но тогда справедливо (2.21) и из (2.22) следует (2.23). Теперь, переходя к пределу по $L \to \infty$ в равенстве (2.24), получим $\mathbf{E}\xi = m_1$ и тем самым достаточность тоже доказана.

Замечание *Если ξ - дискретная с.в., принимающая значения $0, 1, 2, \ldots$ то по этой теореме $\mathbf{E}\xi < \infty$ тогда и только тогда, когда*

$$m_1 = \sum_{k=1}^{\infty} \mathbf{P}(\xi \geqslant k) < \infty, \tag{2.27}$$

причем $m_1 = \mathbf{E}\xi$.

2) Доказательство утверждения теоремы для произвольного $s \geqslant 2$ проводится совершенно аналогично рассмотренному случаю (т.е. тоже с использованием формулы интегрирования по частям).

2.4 Неравенства Чебышева, Йенсена, Ляпунова

Рассматриваемые ниже классические неравенства позволяют с помощью моментов случайных величин получать оценки соответствующих функций распределения, а также устанавливать полезные соотношения между моментами различного порядка.

Наиболее известными в теории вероятностей и ее приложениях являются **неравенства Чебышева**. Предположим, что $\xi \geqslant 0$ – неотрицательная с.в., такая, что $\mathbf{E}\xi < \infty$. Тогда для любого $x > 0$ справедливо неравенство

$$\mathbf{P}(\xi > x) \leqslant \frac{\mathbf{E}\xi}{x} \tag{2.28}$$

Действительно, для любого $x \geqslant 0$ имеем

$$\mathbf{E}\xi = \int_0^\infty u \, dF_\xi(u) \geqslant \int_x^\infty u \, dF_\xi(u) \geqslant x \, \mathbf{P}(\xi > x)$$

Совершенно аналогично показывается, что для неотрицательной с.в. ξ и любой неубывающей положительной функции $G(x), x \geqslant 0$, такой, что $\mathbf{E}G(\xi) < \infty$

$$\mathbf{P}(\xi > x) \leqslant \frac{\mathbf{E}G(\xi)}{G(x)} \tag{2.29}$$

Если теперь положить $G(\xi) = \mid \xi - \mathbf{E}\xi\mid^2$, то для любого $x > 0$ отсюда получим

$$\mathbf{P}\Big(\mid \xi - \mathbf{E}\xi\mid > x\Big) \leqslant \frac{D\xi}{x^2}, \tag{2.30}$$

где через $D\xi$ обозначен второй центральный момент с.в. ξ (не обязательно неотрицательной!), то есть $D\xi = \mathbf{E}\mid \xi - \mathbf{E}\xi\mid^2$, который, как известно, называется *дисперсией с.в.* ξ и в данном случае предполагается ограниченным ($D\xi < \infty$).
Все три неравенства (2.28)-(2.30) носят имя Чебышева.

Полезные соотношения между моментами с.в. могут быть получены также с помощью **неравенства Йенсена(Jensen)**. Пусть теперь $G(\mathbf{.})$ – просто некоторая выпуклая вниз функция (convex). Тогда для любой действительной с.в. ξ

$$\mathbf{E}G(\xi) \geqslant G(\mathbf{E}\xi), \tag{2.31}$$

где обе части неравенства предполагаются конечными.
Это неравенство носит имя Йенсена и вытекает из очевидного для любой выпуклой вниз функции соотношения

$$G(\xi) \geqslant G(y) \; + \; G^{'}(y) \; (\xi - y)$$

Достаточно вычислить математическое ожидание от обеих частей этого выражения, предварительно положив $y = \mathbf{E}\xi$. При этом в качестве $G'(y)$ можно взять любое значение производной в точке как слева, так и справа от y, если в самой этой точке производная не существует.

Если же функция $G(\cdot)$ – вогнутая вверх (concave) функция (при этом, очевидно, функция $[-G(\cdot)]$, будет уже выпуклой вниз), то тогда для любой действительной с.в. ξ будет справедливо обратное неравенство

$$\mathbf{E}G(\xi) \leqslant G(\mathbf{E}\xi),$$

т.е. неравенство Йенсена с обратным знаком.

Рассмотрим далее $0 < s < t < \infty$ и предположим, что момент с.в. $\mathbf{E}|\xi|^t < \infty$. Тогда и $\mathbf{E}|\xi|^s < \infty$, и между этими моментами справедливо следующее соотношение

$$\Big(\mathbf{E}|\xi|^t\Big)^s \geqslant \Big(\mathbf{E}|\xi|^s\Big)^t, \tag{2.32}$$

известное как **неравенство Ляпунова**.
Для доказательства этого факта воспользуемся очевидным равенством

$$|\,\xi\,|^t = \big(|\,\xi\,|^s\big)^{t/s} \tag{2.33}$$

Поскольку по условию $t > s$, то функция $G(u) = u^{t/s}$ оказывается выпуклой вниз и для нее справедливо неравенство Йенсена (2.31). Но тогда, вычисляя матожидание от обеих частей (2.33) и применяя (2.31), получим

$$\mathbf{E}|\,\xi\,|^t = \mathbf{E}\Big(|\,\xi\,|^s\Big)^{t/s} \geqslant \Big(\mathbf{E}|\,\xi\,|^s\Big)^{t/s}.$$

Искомое утверждение (2.32) получается при возведении в степень s крайних членов в обеих частях последнего неравенства.

2.5 Некоторые операции со случайными величинами

Пусть ξ_1 и ξ_2 – две независимые с.в. с ф.р. $F_1(\cdot)$ и $F_2(\cdot)$ соответственно. Рассмотрим некоторые часто встречающиеся операции с этими случайными величинами и напомним, как при этом преобразуются их функции распределения.

Сумма двух с.в.. Определим новую с.в. $\xi = \xi_1 + \xi_2$ и обозначим ее ф.р. через $F(\cdot)$. Тогда, как известно, для F справедливо следующее представление:

$$F(x) = \int_{-\infty}^{\infty} F_1(x-y)\, dF_2(y).$$

Если ξ_1 и ξ_2 – неотрицательные с.в., то этот интеграл преобразуется к следующему к виду

$$F(x) \;=\; \int_0^x F_1(x-y)\; dF_2(y) \quad x>0, \tag{2.34}$$

причем следует положить $F(x)=0$ при $x \leqslant 0$. Интеграл справа называется *сверткой Стилтьеса функций* F_1 *и* F_2 и обозначается $F \;=\; F_1 * F_2$.

Максимум двух с.в.. Определим теперь с.в. $\xi = \max\big(\xi_1, \xi_2\big)$. Тогда ф.р. такой с.в. ξ получим, используя независимость исходных с.в. ξ_1 и ξ_2:

$$F(x) \;=\; \mathbf{P}\Big(\{\xi_1 \leqslant x\} \bigcap \{\xi_2 \leqslant x\}\Big) \;=\; F_1(x)\; F_2(x). \tag{2.35}$$

Минимум двух с.в.. Если $\xi = \min\big(\xi_1, \xi_2\big)$. Тогда для $F(x)$ получим следующее выражение:

$$\begin{aligned} F(x) \;&=\; \mathbf{P}\Big(\{\xi_1 \leqslant x\} \bigcup \{\xi_2 \leqslant x\}\Big) \;=\; F_1(x) \;+\; F_2(x) \;-\; F_1(x)\; F_2(x) \\ &=\; 1 \;-\; \mathcal{F}_1(x)\; \mathcal{F}_2(x). \end{aligned} \tag{2.36}$$

Последнее равенство легко проверяется с помощью несложных преобразований.

2.6 Примеры непрерывных распределений вероятностей

♦ **Экспоненциальное распределение**. Среди абсолютно непрерывных функций распределений, используемых в ТМО, это распределение по праву занимает особое место и, как мы увидим ниже, многие классические результаты этой науки связаны со с. в., распределенными по экспоненциальному закону, т.е. имеющими ф.р. следующего вида:

$$F(x) \;=\; 1 \;-\; \exp(-\lambda x),\; \lambda>0,\; x \geqslant 0. \tag{2.37}$$

Предполагается, что $F(x)=0$ при $x<0$.
Плотность вероятности такого распределения, очевидно

$$f(x) \;=\; \lambda\; \exp(-\lambda x),\; x \geqslant 0 \;\Big(f(x)=0,\; x<0\Big). \tag{2.38}$$

Если с.в. ξ имеет экспоненциальную ф.р., то для нее легко можно вычислить первые два момента и дисперсию

$$\mathbf{E}\xi = 1/\lambda, \quad \mathbf{E}\xi^2 = 2/\lambda^2, \quad D\xi = 1/\lambda^2. \tag{2.39}$$

Для функции интенсивности отказов (см.(2.15)) этого распределения получим

$$r(x) \;=\; \frac{f(x)}{1 \;-\; F(x)} \;=\; \lambda. \tag{2.40}$$

А для ф.р. остаточного времени жизни ξ_x можно из (2.18) найти следующее выражение

$$\begin{aligned} \mathbf{P}(\xi_x \leqslant y) \;&=\; \frac{F(x+y) - F(x)}{1 \;-\; F(x)} \;=\; \frac{\exp(-\lambda x) - \exp(-\lambda(x+y))}{\exp(-\lambda x)} \\ &=\; 1 \;-\; \exp(-\lambda y) \;=\; F(y). \end{aligned} \tag{2.41}$$

Полученное нами равенство обычно интерпретируют как "свойство отсутствия памяти" у экспоненциального распределения (lack of memory). Это означает, что экспоненциальное распределение не имеет "последействия"(т.е. будущее поведение не зависит от прошлого).

Например, если время ожидания наступления некоторого события является с.в. с экспоненциальной ф.р., то по прошествии любого интервала времени (при условии, что за это время интересующее нас событие еще не произошло), начиная с этого момента новый отсчет времени, нам придется по-прежнему ожидать наступление события некоторое случайное время, не зависящее от величины уже прошедшего отрезка времени, т.е. оставшееся время ожидания будет иметь все ту же экспоненциальную ф.р.. Этот факт, конечно, не очень ободряет на рыбалке, если моменты поклевки распределены экспоненциально, или когда ожидаешь прихода очередного автобуса при экспотненциальном распределении интервалов их движения, но в ТМО он оказывается очень даже полезным при изучении различных моделей очередей.

♦ **Распределение Эрланга**. Пусть $\xi_1, \xi_2, \ldots, \xi_n$ – независимые, неотрицательные с.в., имеющие экспоненциальное распределение (2.37). Используя интеграл свертки (2.34), можно найти ф.р. суммы двух таких с.в.

$$\begin{aligned} \mathbf{P}(\xi_1 + \xi_2 \leqslant x) \;&=\; \int_0^x \Big(1 - e^{-\lambda(x-y)}\Big)\lambda\; e^{-\lambda y} dy \\ &=\; 1 - e^{-\lambda x} - \lambda x\; e^{-\lambda x}. \end{aligned} \tag{2.42}$$

По индукции можно показать, что для любого $n \geqslant 1$

$$\mathbf{P}(\xi_1 + \xi_2 + \ldots + \xi_n \leqslant x) \;=\; 1 \;-\; \sum_{k=0}^{n-1} \frac{(\lambda x)^k}{k!}\; e^{-\lambda x} \tag{2.43}$$

Нетрудно вычислить плотность такого распределения. Продифференцировав для этого по x ф.р. (2.43), получим

$$f(x) \;=\; \lambda\; \frac{(\lambda x)^{n-1}}{(n-1)!}\; e^{-\lambda x} \tag{2.44}$$

Ясно, что с.в. $\xi = \xi_1 + \xi_2 + \ldots + \xi_n$ будет иметь среднее значение $\mathbf{E}\xi = n/\lambda$. Если рассмотреть теперь "нормированную" с.в. $\widetilde{\xi} = \xi/n$ (чтобы получить $\mathbf{E}\,\widetilde{\xi} = 1/\lambda$), то тогда, очевидно

$$\mathbf{P}(\widetilde{\xi} \leqslant x) = \mathbf{P}(\xi_1 + \xi_2 + \ldots + \xi_n \leqslant nx) = 1 - \sum_{k=0}^{n-1} \frac{(n\lambda x)^k}{k!}\, e^{-n\lambda x} \equiv E_n(x) \quad (2.45)$$

Функция распределения (2.45) носит имя Эрланга и называется распределением Эрланга порядка n. Это распределение, как мы только что видели, возникло из распределения суммы независимых с.в., распределенных экспоненциально. Функция $E_n(x)$ является абсолютно непрерывной при любом $n < \infty$, однако при $n \to \infty$ стремится к вырождению в точке $x = 1/\lambda$.
Наряду с только что полученным распределением суммы неотрицательных независимых с.в. $\xi_1,\ \xi_2,\ \ldots\ \xi_n$, имеющих экспоненциальное распределение, представляет также интерес распределение целой с.в. ν_x, определяемой для некоторой фиксированной величины $x > 0$ следующим образом:

$$\nu_x \;=\; \begin{cases} 0, & \xi_1 > x; \\ max\{n : \xi_1 + \xi_2 + \ldots + \xi_n \leqslant x\}, & \text{иначе.} \end{cases} \quad (2.46)$$

Из этого определения непосредственно следует, что

$$\mathbf{P}(\nu_x = 0) \;=\; \mathbf{P}(\xi_1 > x) \;=\; e^{-\lambda x}. \quad (2.47)$$

А для $n \geqslant 1$, т.к. очевидно $\{\xi_1 + \ldots + \xi_{n+1} \leqslant x\} \subset \{\xi_1 + \ldots + \xi_n \leqslant x\}$, получим

$$\begin{aligned} \mathbf{P}(\nu_x = n) &= \mathbf{P}\Big(\{\xi_1 + \ldots + \xi_n \leqslant x\} \bigcap \{\xi_1 + \ldots + \xi_{n+1} > x\}\Big) \\ &= \mathbf{P}\Big(\xi_1 + \ldots + \xi_n \leqslant x\Big) \;-\; \mathbf{P}\Big(\xi_1 + \ldots + \xi_{n+1} \leqslant x\Big) \\ &= \frac{(\lambda x)^n}{n!} e^{-\lambda x}, \end{aligned} \quad (2.48)$$

где при последнем переходе было использовано (2.43).

Таким образом, мы получили, что с.в. ν_x имеет пуассоновское распределение (2.8) с параметром $\Lambda = (\lambda x)$.

♦ **Равномерное распределение**. Зафиксируем на действительной оси некоторый отрезок $[a, b]$, $a < b$. Абсолютно непрерывное распределение с функцией плотности вероятностей

$$f(x) \;=\; \begin{cases} 1/(b-a), & a \leqslant x \leqslant b; \\ 0, \text{ в других точках,} \end{cases} \quad (2.49)$$

называется *равномерным распределением вероятностей* на отрезке $[a, b]$. Ф.р. вероятностей при этом имеет следующий вид:

$$F(x) \;=\; \begin{cases} 0, & x < a; \\ (x-a)/(b-a), & a \leqslant x < b; \\ 1, & x \geqslant b. \end{cases} \quad (2.50)$$

Если с.в. ξ имеет равномерное на отрезке $[a,b]$ распределение, то

$$\mathbf{E}\xi = (b+a)/2, \quad \mathbf{E}\xi^2 = (b^2+ab+a^2)/3, \quad D\xi = (b-a)^2/12. \tag{2.51}$$

В качестве полезного упражнения рекомендуется проверить справедливость этих выражений путем непосредственных вычислений.

Рассмотрим снова с.в. ν_x (число независимых экспоненциально распределенных с.в. $\{\xi_i\}_{i\geqslant 1}$, сумма которых не превосходит некоторую фиксированную величину $x>0$) (см. (2.46)) и найдем условную ф.р. с.в. ξ_1 при условии, что $\nu_x = 1$, т.е. $\mathbf{P}\Big(\xi_1 \leqslant y \;\Big|\; \nu_x = 1\Big)$.
Для случая $y > x$, очевидно, $\mathbf{P}\Big(\xi_1 \leqslant y \;\Big|\; \nu_x = 1\Big) = 1$, т.к. событие $\{\nu_x = 1\} \equiv \{\xi_1 \leqslant x\} \bigcap \{(\xi_1+\xi_2) > x\} \subset \{\xi_1 \leqslant x\}$, а этот $x < y$.
Поэтому представляет интерес другой случай, когда $y \leqslant x$. По определению условной вероятности имеем:

$$\mathbf{P}\Big(\xi_1 \leqslant y \;\Big|\; \nu_x = 1\Big) = \frac{\mathbf{P}(\xi_1 \leqslant y,\ \nu_x = 1)}{\mathbf{P}(\nu_x = 1)} \equiv \frac{\mathbf{P}(\xi_1 \leqslant y,\ \xi_1+\xi_2 > x)}{\mathbf{P}(\nu_x = 1)}$$

Для нахождения числителя воспользуемся независимостью с.в. ξ_1 от ξ_2 и формулой полной вероятности, которую запишем в интегральной форме по всем возможным $\xi_1 = u$ от 0 до y:

$$\begin{aligned}\mathbf{P}(\xi_1 \leqslant y,\ \xi_1+\xi_2 > x) &= \int_0^y \mathbf{P}(\xi_2 > x-u)\ d\mathbf{P}(\xi_1 = u) \\ &= \int_0^y e^{-\lambda(x-u)}\lambda e^{-\lambda u}\ du = \lambda e^{-\lambda x} y.\end{aligned}$$

А знаменатель может быть найден из (2.48) при $n=1$ т.е. равен $\lambda x e^{-\lambda x}$.
Окончательно получаем, что

$$\mathbf{P}\Big(\xi_1 \leqslant y \;\Big|\; \nu_x = 1\Big) = \frac{y}{x}, \quad 0 \leqslant y \leqslant x, \tag{2.52}$$

т.е. найденное нами условное распределение экспоненциально распределенной с.в. ξ_1 – оказалось равномерным на $[0,x]$ распределением!

Связь равномерного и экспоненциального распределений оказывается еще более глубокой и мы сформулируем ее в общем виде следующим образом.

Пусть с.в. $\{\xi_i\}_{i\geqslant 1}$ распределены экспоненциально с параметром $\lambda > 0$. Обозначим $Y_1 = \xi_1,\ Y_2 = \xi_1+\xi_2,\ \ldots,\ Y_k = \sum_{i=1}^k \xi_i,\ \ldots,$ так что разности между двумя соседними Y_k и Y_{k-1} будут равны соответствующей с.в. ξ_k, и поэтому будут иметь экспоненциальное распределение.

Зафиксируем некоторый $x > 0$ и предположим, что на отрезке $[0,x]$ поместилось не более чем n таких Y-ков (т.е. для выбранного x оказалось $\nu_x = n$). Рассмотрим наряду с указанным набором $\{Y_1,\ Y_2, \ldots, Y_n\}$ другой набор из n с.в. $\{X_1,\ X_2, \ldots,\ X_n\}$, выбранных из равномерного на $[0,x]$

распределения, и обозначим $\{\eta^{(1)},\ \eta^{(2)},\dots,\ \eta^{(n)}\}$ их упорядоченную по возрастанию последовательность, так что $\eta^{(1)} = \min\{X_1,\ X_2,\dots,\ X_n\}$, а $\eta^{(n)} = \max\{X_1,\ X_2,\dots,\ X_n\}$. Эти величины $\eta^{(i)},\ 1 \leqslant i \leqslant n$ принято называть "упорядоченными статистиками" (the order statistics of a random sample).

Теорема 2.2. *Случайные величины $\{Y_1,\ Y_2,\dots,Y_n\}$, получаемые последовательным суммированием независимых экспоненциально распределенных с.в., при условии, что на отрезке $[0,x]$ их умещается ровно n штук (т.е. $\nu_x = n$), имеют точно такое же распределение, что и упорядоченные (по возрастанию) статистики – с.в. $\{\eta^{(1)},\ \eta^{(2)},\dots,\ \eta^{(n)}\}$, получаемые из набора независимых, равномерно распределенных на $[0,x]$ с.в. $\{X_1,\ X_2,\dots,\ X_n\}$.*

Доказательство. Будем сравнивать совместные плотности вероятностей интересующих нас распределений.

Так как все с.в. величины $X_1,\ X_2,\dots,\ X_n$ независимы, то их совместная плотность

$$f_n\Big(x_1,\dots,x_n\Big) = f(x_1)\cdots f(x_n) = \Big(\frac{1}{x}\Big)^n,\quad 0 \leqslant x_1 \leqslant\ \dots\ \leqslant x_n \leqslant x, \tag{2.53}$$

причем выражение в правой части не зависит от величин $x_1,\ \dots\ , x_n$.

Очевидно, что существует ровно $(n!)$ различных выборок из равномерного распределения, которые дадут один и тот же упорядоченный комплект, то есть ту же выборку для упорядоченных статистик (по числу $(n!)$ возможных перестановок-пермутаций). Поэтому

$$f_n\Big(\eta_1,\ \dots,\ \eta_n\Big) = \frac{n!}{x^n},\quad 0 \leqslant \eta_1 \leqslant\ \dots\ \leqslant \eta_n \leqslant x. \tag{2.54}$$

Заметим, что выражение, полученное в правой части (2.54), тоже не зависит от величин $\eta_1,\ \dots\ , \eta_n$, стоящих в левой части.

Далее для некоторых $0 \leqslant y_1 \leqslant y_2 \leqslant \dots \leqslant y_n \leqslant x$ рассмотрим совместное условное распределение

$$\begin{aligned}
&\mathbf{P}\quad \Big(Y_1 \leqslant y_1, Y_2 \leqslant y_2, \dots, Y_n \leqslant y_n \mid \nu_x = n\Big)\\
&= \frac{\mathbf{P}(Y_1 \leqslant y_1, Y_2 \leqslant y_2, \dots, Y_n \leqslant y_n,\ \nu_x = n)}{\mathbf{P}(\nu_x = n)}\\
&= \frac{\mathbf{P}\Big(\xi_1 \leqslant y_1, \xi_2 \leqslant (y_2 - y_1), \dots, \xi_n \leqslant (y_n - y_{n-1}), \xi_{n+1} > (x - y_n)\Big)}{\mathbf{P}(\nu_x = n)}\\
&= \frac{\mathbf{P}\Big(\xi_1 \leqslant y_1, \xi_2 \leqslant (y_2 - y_1), \dots, \xi_n \leqslant (y_n - y_{n-1})\Big)\mathbf{P}\Big(\xi_{n+1} > (x - y_n)\Big)}{\mathbf{P}(\nu_x = n)}.
\end{aligned}$$

Отсюда, воспользовавшись независимостью с.в. $\{\xi_i\}_{i\geqslant 1}$ и формулами (2.38), (2.47) и (2.48), нетрудно получить следующее выражение для плотности

вероятностей рассматриваемого условного распределения:

$$\begin{aligned} f_n & \left(y_1, y_2, \ldots, y_n \;\middle|\; \nu_x = n\right) \\ = & \frac{\left(\lambda e^{-\lambda y_1} \lambda e^{-\lambda(y_2-y_1)} \ldots \lambda e^{-\lambda(y_n-y_{n-1})}\right)\left(e^{-\lambda(x-y_n)}\right)}{\frac{(\lambda x)^n}{n!} e^{-\lambda x}} \\ = & \frac{n!}{x^n}. \end{aligned} \tag{2.55}$$

Полученный результат совпадает с (2.54), причем и в этом случае правая часть (2.55) не зависит от $y_1, y_2, \ldots, y_n$.
Доказательство теоремы завершено.

2.7 Преобразование Лапласа-Стилтьеса и производящая функция

Пусть $F(x),\ -\infty < x < \infty$ – некоторая функция действительного переменного, равная нулю при $x < 0$ и имеющая ограниченную вариацию на каждом отрезке $[0, T]$, то есть F может быть представлена как $F(x) = F_+(x) - F_-(x)$, где обе F_+ и F_- – неубывающие неотрицательные ограниченные функции при $x \geqslant 0$.

Предположим, что существуют действительные A и s_0 , такие, что

$$\left|F(x)\right| \leqslant A\, e^{s_0 x}, \quad x \geqslant 0, \tag{2.56}$$

и определим следующие две функции комплексного переменного s:

$$F_{LS}(s) = \int_0^\infty e^{-sx} dF(x), \tag{2.57}$$

$$F_L(s) = \int_0^\infty e^{-sx} F(x)\, dx, \quad \mathcal{R}e\ s > s_0. \tag{2.58}$$

Введенные таким образом функции называются *преобразованием Лапласа - Стилтьеса функции* $F(x)$ и *преобразованием Лапласа функции* $F(x)$ соответственно.

Если обозначить через $F_p(s)$ любую из функций $F_{LS}(s)$ или $F_L(s)$, то для обоих преобразований справедливы следующие утверждения:
(i) $F_p(s)$ - аналитическая функция при $\mathcal{R}e\ s > s_0$;
(ii) Если $F_{p1}(s)$ – преобразование функции $F_1(x)$, а $F_{p2}(s)$ – преобразование функции $F_2(x)$, и обе функции x удовлетворяют (2.56), а $F_{p1}(s) = F_{p2}(s)$ для $\mathcal{R}e\ s > s_0$, то и $F_1(x) = F_2(x)$ для всех тех x, которые являются точками непрерывности функций $F_1(x)$ и $F_2(x)$ одновременно.

Определенные выше преобразования связаны между собой очень простым соотношением:

$$F_{LS}(s) = s\, F_L(s). \tag{2.59}$$

Этот факт позволяет нам далее сформулировать основные свойства только для преобразования Лапласа-Стилтьеса.

(1) Если для некоторых действительных α и β функции $H(x)$, $F(x)$ и $G(x)$ связаны соотношением $H(x) = \alpha F(x) + \beta G(x)$, то и их преобразования оказываются связанными таким же соотношением, то есть

$$H_{LS}(s) \;=\; \alpha F_{LS}(s) \;+\; \beta G_{LS}(s).$$

(2) Если $G(x) \;=\; \int_0^x F(u)\; du$, то

$$G_{LS}(s) \;=\; \frac{F_{LS}(s)}{s}\;.$$

(3) Если $G(x) \;=\; \int_0^x e^{-\lambda u} dF(u)$, то

$$G_{LS}(s) \;=\; F_{LS}(s+\lambda).$$

(4) Пусть первая производная $G(x) \;=\; dF(x)/dx$ существует для любых $x \geqslant 0$, и при этом G – функция ограниченной вариации. Тогда

$$G_{LS}(s) \;=\; s\; F_{LS}(s) \;+\; s\; F(o).$$

(5) Пусть $H(x) \;=\; \int_0^x F(x-u)\; dG(u)$ – свёртка Стилтьеса (см. (2.34)) F и G. Тогда

$$H_{LS}(s) \;=\; F_{LS}(s)\; G_{LS}(s).$$

(6) Пусть $H(x) \;=\; \int_0^x F(x-u) G(u)\; du$ – интегральная свёртка функций F и G. Тогда

$$H_{LS}(s) \;=\; \frac{1}{s}\; F_{LS}(s)\; G_{LS}(s).$$

Зная преобразование L-St, можно с его помощью выяснить некоторые предельные свойства исходной функции, а именно

(7) Если существует $\lim_{x \downarrow 0} F(x)$, то

$$\lim_{x \downarrow 0} F(x) \;=\; \lim_{s \to \infty} F_{LS}(s),$$

а если существует $\lim_{x \to \infty} F(x)$, то

$$\lim_{x \to \infty} F(x) \;=\; \lim_{s \downarrow 0} F_{LS}(s).$$

Обратные утверждения, вообще говоря, неверны, однако, следующая теорема может рассматриваться как некая эрзац-замена обратных утверждений.

Теорема 2.3. *[**Таубера (Tauber)**] (Без доказательства).*
Пусть $F(x) \geqslant 0$, $\left|F_{LS}(s)\right| < \infty$ *для* $\Re e\ s > 0$ *и предположим, что существуют пределы (для* $\alpha > 0$*)*

$$\lim_{u\to\infty} u^{\alpha-1} F_{LS}(u) \ = \ f_* \ ,$$

или

$$\lim_{u\downarrow 0} u^{\alpha-1} F_{LS}(u) \ = \ f_{**} \ .$$

Тогда, соответственно,

$$\lim_{T\downarrow 0} T^{-\alpha} \int_0^T F(x)\ dx \ = \ f_* \ ,$$

или

$$\lim_{T\to\infty} T^{-\alpha} \int_0^T F(x)\ dx \ = \ f_{**} \ .$$

Приведем примеры преобразований L-St от некоторых наиболее часто встречающиеся в ТМО функций распределения:

(i) Если $F(x) \ = \ \mathbf{1}(x)$, то $F_{LS}(s) = 1$;

(ii) Если $F(x) \ = \ 1 - e^{-\lambda x}$, $\lambda > 0$, то $F_{LS}(s) \ = \ \lambda/(\lambda + s)$;

(iii) Если $F(x) \ = \ E_n(x) \ = \ 1 \ - \ \sum_{k=0}^{n-1} \frac{(n\lambda x)^k}{k!} e^{-n\lambda x}$, то $F_{LS}(s) \ = \ \left(\frac{n\lambda}{n\lambda+s}\right)^n$.

В заключение, сформулируем (без доказательства) ещё одну теорему, отражающую свойство непрерывности преобразования L-St, которая также окажется полезной в дальнейшем изложении.

Теорема 2.4. *[**О слабой сходимости**] (Без доказательства).*
Пусть $F^{(n)}(x)$, $n \geqslant 1$ *– последовательность ф.р. некоторой неотрицательной с.в., а* $F_{LS}^{(n)}(s)$ *– последовательность соответствующих преобразований L-St от этих функций. Тогда справедливы следующие утверждения:*
(i) *Если* $F^{(n)} \xrightarrow{w} F$, *то* $F_{LS}^{(n)}(s) \to F_{LS}(s)$ *для всех* s, $\Re e\ s > 0$.
(ii) *Соответственно, если* $F_{LS}^{(n)}(s)$ *сходится для любого* s, $\Re e\ s > 0$ *к функции* $F(s)$, *которая непрерывна при* $s = 0$, *то* $F(s)$ *есть преобразование L-St от функции* F *и* $F^{(n)} \xrightarrow{w} F$. *Причём в этом случае оказывается, что* $\lim_{x\to\infty} F(x) = 1$ *тогда и только тогда, когда* $\lim_{s\to 0} F(s) = 1$.

Преобразование Лапласа-Стилтьеса допускает следующую вероятностную интерпретацию. Пусть $F(x) \ = \ \mathbf{P}(\xi \leqslant x)$ – ф.р. некоторой неотрицательной с.в. ξ. Тогда $F_{LS}(s)$ можно представить в виде следующего математического ожидания:

$$F_{LS}(s) \ = \ \mathbf{E}\Big[e^{-s\xi}\Big]. \tag{2.60}$$

Отсюда следует, что существование k-го момента с.в. ξ эквивалентно существованию k-той производной функции $F_{LS}(s)$ при $s = 0$, и при этом справедлива следующая формула:

$$E\xi^k \;=\; (-1)^k \left.\frac{d^k F_{LS}(s)}{ds^k}\right|_{s=0} . \tag{2.61}$$

Действительно, если у ф.р. $F(x)$ существует плотность $f(x)$, то

$$F_{LS}(s) \;=\; \int_0^\infty e^{-sx} f(x)\, dx$$

Но тогда отсюда, разлагая экспоненту в степенной ряд, получим

$$\begin{aligned} F_{LS}(s) \;&=\; \int_0^\infty \Big(1 - sx + \frac{s^2x^2}{2!} - \frac{s^3x^3}{3!} + \dots\Big) f(x)\, dx \\ &=\; 1 \;-\; m_1\, s \;+\; m_2\, \frac{s^2}{2!} \;-\; m_3\, \frac{s^3}{3!} \;+\; \dots \;, \end{aligned}$$

где

$$m_k \;=\; \int_0^\infty x^k f(x)\, dx \;=\; \mathbf{E}\Big\{x^k\Big\} \;=\; (-1)^k \left.\frac{d^k F_{LS}(s)}{ds^k}\right|_{s=0} ,\; k \geqslant 1 \;. \tag{2.62}$$

Следовательно, значения моментов можно определять по разложению функции $F_{LS}(s)$ в степенной ряд, а именно:

$$m_1 = -\big(\text{ коэфф. при } s\ \big);\quad m_2 = \big(\text{ коэфф. при } \frac{s^2}{2!}\ \big) \quad \text{и т.д.}$$

Аналогично преобразованию L-St для ф.р. непрерывных с.в., удобным аппаратом исследования дискретных с.в. является аппарат *производящих функций*, с помощью которых также просто находятся моменты вероятностных распределений.
Пусть $\{a_n\}_{n\geqslant 0}$ – последовательность действительных чисел. Если функция

$$a(z) \;=\; \sum_{n=0}^{\infty} a_n z^n \tag{2.63}$$

сходится в некотором круге $|z|\leqslant z_0$ комплексной области, то тогда $a(z)$ называется **производящей функцией последовательности** $\{a_n\}$.

Определим следующую кусочно-непрерывную функцию $A(x)$ со скачками в целочисленных точках:

$$A(x) \;=\; \begin{cases} 0, & x \;<\; 0; \\ \sum_{j=0}^{n} a_j, & n \leqslant x \;<\; n+1, \quad n\geqslant 0. \end{cases} \tag{2.64}$$

Тогда производящая функция может быть представлена в следующем виде:

$$a(z) \;=\; \sum_{n=0}^{\infty} a_n z^n \;=\; \int_0^{\infty} z^x \; dA(x) \;=\; \int_0^{\infty} e^{x \ln z} \; dA(x), \qquad (2.65)$$

что означает ни что иное как то, что $a(z)$ – есть преобразование L-St функции $A(x)$ в точке $s \;=\; -\ln z$.
Это представление автоматически гарантирует тот факт, что производящая функция $a(z)$ единственным образом определяет свою исходную последовательность $\{a_n\}$.

Соотношение (2.65) позволяет выписать некоторые свойства производящих функций, вытекающие из рассмотренных выше свойств преобразования L-St. Пусть $a(z),\; b(z),\; c(z)$ – производящие функции последовательностей $\{a_n\},\;\; \{b_n\}$ и $\{c_n\}$ соответственно. Тогда

(1) Если $c_n = \alpha a_n + \beta b_n$ для любых n и некоторых действительных α и β, то

$$c(z) \;=\; \alpha \; a(z) \;+\; \beta \; b(z)$$

(2) Если $b_n \;=\; \sum_{j=0}^{n} a_j, \;\; n \geqslant 0,$
то

$$b(z) \;=\; \frac{a(z)}{1-z}$$

(3) Пусть $b_n \;=\; a_{n+1} \;-\; a_n, \;\;\; n \geqslant 0,$
тогда

$$b(z) \;=\; \frac{(1-z) \; a(z) \;-\; a_0}{z}$$

(4) Пусть $c_n \;=\; \sum_{j=0}^{n} a_j \; b_{n-j}, \;\;\; n \geqslant 0,$
тогда

$$c(z) \;=\; a(z) \; b(z).$$

Если $\{a_n\}$ – вероятностное распределение некоторой дискретной с.в. ξ, то производящую функцию можно рассматривать как следующее математическое ожидание:

$$a(z) \;=\; \mathbf{E} \; z^{\xi}$$

Тогда, соответственно,

$$\frac{d^k a(z)}{dz^k}\Big|_{z=1} \;=\; \mathbf{E}\Big[\xi(\xi-1)\ldots(\xi-k+1)\Big], \;\;\; k = 1,\; 2,\; \ldots \qquad (2.66)$$

Выражение в правой части определяет величину, называемую *k-тым факториальным моментом с.в. ξ*. Зная факториальные моменты, мы, очевидно, можем легко найти обычные моменты и наоборот.

В заключение, следует упомянуть одно весьма полезное *предельное свойство* производящей функции. Если последовательность $\{a_n\}$ такова, что $\lim_{n\to\infty} a_n = a < \infty$, то существует и следующий предел:

$$\lim_{z\uparrow 1}(1-z)\, a(z) = a$$

Обратное утверждение (в общем случае) не справедливо, однако, можно (наложив дополнительное ограничение) утверждать следующее.
Если $\lim_{z\uparrow 1}(1-z)\, a(z) = a < \infty$ и $\lim_{n\to\infty}[n(a_n - a_{n-1})] = 0$, то существует $\lim_{n\to\infty} a_n = a$.

Приведем несколько примеров производящих функций некоторых наиболее часто встречающиеся в ТМО дискретных распределений вероятностей:
(i) Если

$$a_n = C_N^n p^n q^{N-n}, \quad 0 \leqslant n \leqslant N, \quad p+q=1, \quad p \geqslant 0, \quad q \geqslant 0,$$

то

$$a(z) = (q+pz)^N;$$

(ii) Если

$$a_n = q(1-q)^n, \quad n \geqslant 0, \quad 0 < q < 1,$$

то

$$a(z) = \frac{q}{1-(1-q)z};$$

(iii) Если

$$a_n = \frac{\Lambda^n}{n!}\, e^{-\Lambda}, \quad n \geqslant 0,$$

то

$$a(z) = e^{-\Lambda(1-z)}.$$

Глава 3

Входящие потоки систем МО

Все поступающие в систему МО заявки (или требования) на обслуживание будем предполагать *гомогенными*, т.е. не имеющими никаких характеризующих их меток, кроме момента времени их поступления в систему. В приложениях к компьютерным сетям это означает, что нас не будет интересовать никакая информация о поступившем пакете кроме момента его прихода. Тогда *потоком гомогенных требований* естественно назвать упорядоченную по времени поступления последовательность случайных величин $\mathbf{T} = \{T_i\}_{i\geqslant 1}$, т.е. $0 < T_1 \leqslant T_2 \leqslant T_3 \leqslant \dots$.

3.1 Определение рекуррентного потока

Определение 3.1. *Поток* $\mathbf{T}$ *назовем* **рекуррентным**, *если последовательность с.в.* $\{e_i\}_{i\geqslant 0}$, *таких что* $e_k = T_{k+1} - T_k,\ k \geqslant 0,\ T_0 = 0$ *– является последовательностью независимых, одинаково распределенных с.в. (н.о.р.с.в.), имеющих общую ф.р.* $A(x) = \mathbf{P}(e_k \leqslant x),\ k \geqslant 0$.

Определение 3.2. *Поток* $\mathbf{T}$ *назовем* **рекуррентным с задержкой** *(dellayed recurrent), если все* $\{e_0, e_1,\ \dots\}$ *– независимы, но при этом с.в.* $\{e_1, e_2,\ \dots\}$ *имеют ф.р.* $A(x) = \mathbf{P}(e_k \leqslant x),\ k \geqslant 1$, *а с.в.* e_0 *имеет отличную от них ф.р.* $A_0(x) = \mathbf{P}(e_0 \leqslant x)$. *Эта с.в.* e_0 *естественно называется задержкой.*

Наличие потоков с задержкой вполне естественно, поскольку начало работы системы МО происходит независимо от имеющегося потока заявок и совсем не обязательно должно совпадать с моментом прихода очередной заявки. Множество реальных потоков рекуррентно по своей природе и они, как правило, оказываются потоками с задержкой. Первый после включения системы момент прихода заявки $T_1 = e_0$ может произойти лишь по окончании некоторого промежутка времени, необходимого, например, для

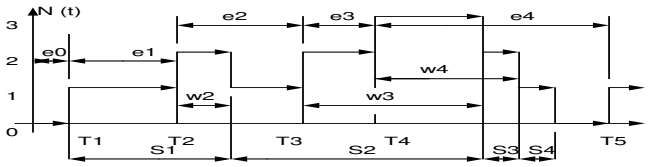


Рис. 3.1: Пример входящего потока.

"прогрева приборов". Таким образом чисто рекуррентный поток следует рассматривать как некую идеализацию реально наблюдаемых потоков.

Определение 3.3. *Поток гомогенных требований будем называть* **ординарным**, *если*

$$\mathbf{P}\big(T_1 < T_2 < T_3 < \dots\big) = 1.$$

Отсюда следует, что в таком потоке в любой отдельно взятый момент не может поступить более одного требования (с вер. 1).

Ниже, если не оговорено особо, будут рассматриваться только ординарные потоки.

Обозначим $N(t, t+x)$ – количество требований, приходящих в систему на интервале времени $[t, t+x]$ для некоторых $t, x \geqslant 0$. Эта с.в. оказывается очень важной характеристикой входных потоков.
Ниже для краткости мы часто будем также обозначать $N(x) = N(0, x)$.

Определение 3.4. *Поток* $\mathbf{T}$ *называется* **стационарным**, *если ф.р. величины* $N(t, t+x)$ *не зависит от* t *при любом* $x \geqslant 0$.

Свойство стационарности гарантирует неизменность статистического поведения потока во времени.

Как уже было отмечено выше, рекуррентный поток без задержки (задаваемый только ф.р. $A(\cdot)$) является частным случаем рекуррентного потока с задержкой (определяемого двумя ф.р. $A_0(\cdot)$ и $A(\cdot)$) при $A_0(x) = A(x)$. Поэтому будем рассматривать далее рекуррентный поток с задержкой, как более общий случай рекуррентного потока.

Для начала выведем в общем виде распределение с.в. $N(t) = N(0,t)$, равной числу заявок, поступивших на отрезке $[0,t]$. Ясно, что

$$N(t) = \max\{k:\ T_k \leqslant t,\ k \geqslant 0\}\ .$$

Как нетрудно заметить, события $\{N(t) \geqslant k\}$ и $\{T_k \leqslant t\}$ совпадают (т.е. эти события являются тождественными). Поэтому

$$\begin{cases} \mathbf{P}(N(t) \geqslant 0) = 1\ , \\ \mathbf{P}(N(t) \geqslant k) = \mathbf{P}(T_k \leqslant t) = A_0 * A_*^{k-1}(t), \quad k \geqslant 1\ , \end{cases} \tag{3.1}$$

где мы последовательно воспользовались правилом нахождения ф.р. для суммы с.в. (2.34) и тем, что по определению считается, что $A_*^0(t) = 1$. Отсюда легко следует, что

$$\begin{cases} \mathbf{P}(N(t) = 0) = \mathbf{P}(N(t) \geqslant 0) - \mathbf{P}(N(t) \geqslant 1) = 1 - A_0(t)\ , \\ \mathbf{P}(N(t) = k) = A_0 * A_*^{k-1}(t) - A_0 * A_*^k(t) = A_0 * A_*^{k-1} * (1 - A)(t), \quad k \geqslant 1\ , \end{cases} \tag{3.2}$$

Продвинуться дальше в упрощениях вида формул в общем случае, к сожалению, не удается. Мы будем, однако, возвращаться к этим формулам при рассмотрении конкретных примеров входных потоков.

3.2 Теорема об эквивалентных определениях пуассоновского потока

Вместе с требованием ординарности и *отсутствием последействия* (т.е. наличием взаимной независимости протекания потока на непересекающихся между собой интервалах времени) стационарность входит в определение *простейшего потока*, введенного Хинчиным в книге [2]. Там в частности показывается, что простейший поток (т.е. ординарный стационарный поток без последействия) существует и, более того, оказывается потоком, носящим имя Пуассона. В нашем же курсе удобнее будет рассмотреть несколько другое определение пуассоновского потока.

Рассмотрим следующие три утверждения, каждое из которых может быть использовано в качестве самостоятельного определения пуассоновского потока.

Утверждение 3.1. *Вероятность прихода ровно одной заявки на малом интервале $(t, t+\delta t)$ не зависит от t и равна*

$$\mathbf{P}\Big(N(t, t+\delta t) = 1\Big) = \lambda\ \delta t\ +\ o(\delta t), \tag{3.3}$$

а вероятность того, что ни одной заявки не придет на таком же интервале

$$\mathbf{P}\Big(N(t, t+\delta t) = 0\Big) = 1\ -\ \lambda\ \delta t\ +\ o(\delta t). \tag{3.4}$$

Утверждение 3.2. *Число заявок $N(0,t)$, приходящих в течение временного интервала длиной t, является с.в., имеющей пуассоновское распределение (2.8) с параметром $\Lambda = \lambda t$.*

Утверждение 3.3. *Временные интервалы $\{e_i\}_{i\geqslant 0}$ между двумя последовательными моментами прихода заявок имеют экспоненциальную ф.р. с параметром $\lambda > 0$.*

Оказывается, выполнение любого одного из этих трех утверждений влечет за собой справедливость и двух других. Этот результат можно сформулировать в следующем виде

Теорема 3.1. *Утверждения (3.1)-(3.3) эквивалентны. Любого из них достаточно для определения стационарного ординарного потока без последействия, называемого потоком Пуассона.*

Доказательство.

(3.1) $\Rightarrow$ (3.2):
Нужно найти вероятность прихода ровно n заявок на интервале времени $(0,t)$, зная вероятностные характеристики потока лишь на бесконечно малом отрезке времени длиной δt.
Обозначим искомую вероятность

$$p_n(t) \;=\; \mathbf{P}\Big(N(0,t) = n\Big). \tag{3.5}$$

Тогда из Утверждения 3.1 получим

$$p_1(\delta t) = \lambda\ \delta t + o(\delta t) \quad \text{и} \quad p_0(\delta t) = 1 - \lambda\delta t + o(\delta t). \tag{3.6}$$

Вероятность прихода двух и более заявок на интервале времени длиной δt обозначим через $\psi(\delta t)$ и с использованием (3.6) найдем, что

$$\begin{aligned} \psi(\delta t) &= \sum_{k=2}^{\infty} p_k(\delta t) \\ &= 1 - p_0(\delta t) - p_1(\delta t) = o(\delta t) \ \text{ при } \ \delta t \ \to\ 0, \end{aligned} \tag{3.7}$$

т.е. рассматриваемый поток оказывается ординарным.

Разделим теперь отрезок $(0,t)$ на m равных частей и положим $\delta t = \frac{t}{m}$. Устремляя $m \to \infty$ (чтобы при этом $\delta t \to 0$), мы довольно скоро получим $m > n$. Поскольку рассматриваемый поток ординарный, величина n конечна, а $m \to \infty$, то всегда найдется такое m_0, начиная с которого для всех $m > m_0$, моменты прихода всех n заявок окажутся по одному в различных интервалах длины $\delta t = \frac{t}{m}$. При этом ровно $(m-n)$ интервалов окажутся пустыми, т.е. не содержащими ни одного момента прихода заявки. Как известно, вероятность размещения n с.в. в m интервалах имеет биномиальное распределение (см. (2.6)). Поэтому распределение вероятности (3.5)

будем искать как предельную форму биномиального распределения. С учетом (3.6) для $\delta t = \frac{t}{m}$ получим

$$\begin{aligned} p_n(t) &= \lim_{m\to\infty} \Big(\frac{m}{n}\Big)\Big(p_1(t)\Big)^n\Big(p_o(t)\Big)^{m-n} \\ &= \lim_{m\to\infty} \Big(\frac{m}{n}\Big)\Big(\lambda\frac{t}{m}\Big)^n\Big(1-\lambda\frac{t}{m}\Big)^{m-n} \\ &= \lim_{m\to\infty} \frac{m(m-1)\dots(m-n+1)}{m^n}\frac{(\lambda t)^n}{n!}\Big(1-\frac{\lambda t}{m}\Big)^m\Big(1-\frac{\lambda t}{m}\Big)^{-n} \\ &= \frac{(\lambda t)^n}{n!}\exp(-\lambda t). \end{aligned} \tag{3.8}$$

И мы пришли к пуассоновскому распределению, т.е. показали, что при выполнении Утверждения 3.1 Утверждение 3.2 тоже справедливо.

Замечание 3.1. *Совершенно аналогично показывается, что распределение* $\mathbf{P}\big(N(t_0, t_0+t) = n\big)$ *также как и* $p_n(t)$ *является пуассоновским распределением, а значит не зависит от* t_0 *при любом* $t \geqslant 0$. *Последнее означает (см. Определение 3.4), что рассматриваемый поток является стационарным.*

(3.2) $\Rightarrow$ (3.1):

Чтобы доказать этот факт достаточно выписать для интервала времени длиной δt необходимые нам вероятности по формуле пуассоновского распределения (3.8), а затем воспользоваться разложением экспоненты в степенной ряд:

$$\begin{aligned} p_1(\delta t) &= \frac{(\lambda\ \delta t)}{1!}e^{-\lambda\delta t} = \lambda\ \delta t\Big(1-\lambda\ \delta t+\frac{(\lambda\ \delta t)^2}{2!}-\dots\ \Big) = \lambda\ \delta t + o(\delta t); \\ p_0(\delta t) &= e^{-\lambda\delta t} = 1-\lambda\ \delta t+\frac{(\lambda\ \delta t)^2}{2!}-\dots\ =\ 1-\lambda\ \delta t + o(\delta t). \end{aligned} \tag{3.9}$$

Тем самым мы показали, что при выполнении Утверждения 3.2 Утверждение 3.1 тоже оказывается справедливым.

Замечание 3.2. *В условиях Утверждения 3.2 поток также оказывается ординарным, поскольку из (3.9) следует, что и в этом случае (3.7) будет выполняться.*

(3.2) $\Rightarrow$ (3.3):

Очевидно, что любой временной интервал e_i между любыми двумя последовательными моментами прихода заявок представляет собой случайную величину. Для нахождения ее ф.р. воспользуемся тем фактом, что по определению функции плотности распределения вероятностей имеем

$$f(t)\ \delta t\ =\ \mathbf{P}\big(t < e_i \leqslant t+\delta t\big), \tag{3.10}$$

причем вероятность справа есть ни что иное, как вероятность первого прихода очередной заявки, а отсчет времени t включен в момент прихода предыдущей заявки. Но приход первой очередной заявки на интервале $(t, t+\delta t)$ означает, что ни одной заявки не пришло на интервале $(0,t)$ и ровно одна пришла за время δt. А тогда по формуле (3.8) имеем

$$\mathbf{P}\big(t < e_i \leqslant t+\delta t\big) = p_0(t)\ p_1(\delta t) = e^{-\lambda t}\ \frac{(\lambda\ \delta t)}{1!} e^{-\lambda \delta t} = \Big(\lambda\ e^{-\lambda t}\Big)\ \delta t + o(\delta t). \tag{3.11}$$

Непосредственно из (3.10) и (3.11) следует, что

$$f(t)\ =\ \lambda\ e^{-\lambda t}, \tag{3.12}$$

а это означает (см. (2.37) и (2.38)), что с.в. e_i имеют экспоненциальную ф.р. и тем самым справедливость Утверждения 3.3 полностью доказана.

(3.3) $\Rightarrow$ (3.2):

Для доказательство этого факта достаточно заметить, что с.в. $N(0,t)$ (в условиях Утверждения 3.3) совпадает с определенной в (2.46) с.в. ν_x, а она, как было показано в (2.47)-(2.48), имеет как раз требуемое в Утверждении 3.2 пуассоновское распределение.

Для завершения доказательства теоремы осталось лишь подчеркнуть, что отсутствие последействия у рассматриваемого нами потока, как уже было продемонстрировано ранее в (2.41), вытекает просто из самого содержания Утверждения 3.3. Поэтому из только что доказанной эквивалентности всех трех утверждений следует, что поток, определяемый любым из них, оказывается ординарным стационарным потоком без последействия, т.е. является потоком Пуассона. Тем самым, теорема доказана полностью.

3.3 Элементы теории восстановления

Исследование функции восстановления представлено довольно большим объемом литературы. Мы же коснемся в нашем курсе лишь вполне конкретных свойств этой функции, используемых для упрощения некоторых доказательств.

3.3.1 Представление для функции восстановления

Определение 3.5. *Для произвольного потока функция* $H(t) = E[N(t)]$, *равная среднему числу заявок, приходящих за промежуток времени* $[0,t]$, *называется* **функцией восстановления**.

Используя (3.1) и замечание (2.27) к Теореме 2.1, нетрудно получить следующее представление для функции восстановления произвольного рекуррентного потока:

$$H(t) \;=\; \sum_{k=1}^{\infty} \mathbf{P}(N(t)\geqslant k) \;=\; \sum_{k=1}^{\infty} A_0 * A_*^{k-1}(t) \;=\; A_0 * \sum_{k=0}^{\infty} A_*^k(t) \qquad (3.13)$$

Для рекуррентного потока без задержки обозначим функцию восстановления $H^0(t)$. Аналогичное (3.13) представление в этом случае, очевидно, примет следующий вид:

$$H^0(t) \;=\; \sum_{k=1}^{\infty} A_*^k(t) \qquad (3.14)$$

Функция восстановления, несмотря на кажущуюся банальность своего определения, как мы увидим ниже, играет заметную роль в ТМО.

3.3.2 Уравнения восстановления

Выделим в правой части (3.13) первый член суммы, соответствующий $k = 0$, и перепишем выражение для $H(t)$ следующим образом:

$$H(t) \;=\; A_0(t) \;+\; A_0 * \sum_{k=1}^{\infty} A_*^k(t)$$

Используя обозначение (3.14), получим соотношение:

$$H(t) \;=\; A_0(t) \;+\; A_0 * H^0(t) \qquad (3.15)$$

Преобразуя второй член в правой части этого соотношения по свойствам свертки и с учетом представления (3.13)

$$A_0 * H^0(t) = A_0 * \Big(A * \sum_{k=1}^{\infty} A_*^{k-1}(t)\Big) = A * A_0 * \sum_{k=0}^{\infty} A_*^k(t) = A * H(t)\ ,$$

получим, что справедливо также и следующее соотношение:

$$H(t) \;=\; A_0(t) \;+\; A * H(t)\ . \qquad (3.16)$$

Соотношения (3.15) и (3.16) носят название **уравнений восстановления**, хотя этот термин правильнее было бы отнести лишь к (3.16), поскольку соотношение (3.16) представляет собой следующее интегральное уравнение относительно неизвестной функции $H(t)$:

$$H(t) \;=\; A_0(t) \;+\; \int_0^t H(t-x)\ dA(x)\ . \qquad (3.17)$$

Однако формально можно и соотношение (3.15) представить в следующем "интегральном виде":

$$H(t) \;=\; \int_0^t \Big(1 \;+\; H^0(t-x)\Big)\ dA_0(x)\ . \qquad (3.18)$$

3.3.3 Теорема о единственности потока Пуассона

Нетрудно показать, что функция восстановления $H(t)$ для потока Пуассона равна λt, то есть что если поток поступающих заявок пуассоновский, то среднее число заявок, поступивших за промежуток времени $[0,t]$, пропорционально длительности t этого промежутка.

Поскольку поток Пуассона является рекуррентным потоком без задержки, то представляет интерес следующий вопрос:
"При каких условиях функция восстановления произвольного рекуррентного потока без задержки также будет линейной по времени, а именно, когда в общем случае будет выполняться равенство

$$H^0(t) = \lambda t, \tag{3.19}$$

где $\lambda > 0$ - некоторая положительная константа?"

Обозначим $h^0(s)$ и $a(s)$ – преобразования L-St функций $H^0(t)$ и $A(t)$ соответственно и применим это преобразование к обеим частям представления (3.14). Тогда по свойству преобразования свертки получим

$$h^0(s) \;=\; \sum_{k=1}^{\infty} a^k(s) \;=\; \frac{a(s)}{1-a(s)}\;, \tag{3.20}$$

где при подсчете суммы была использована формула для бесконечно убывающей геометрической прогрессии, поскольку с очевидностью

$$a(s) = \int_0^{\infty} e^{-sx}\; dA(x) < 1.$$

Применяя теперь преобразование L-St к обеим частям (3.19), получим равенство

$$h^0(s) \;=\; \frac{\lambda}{s}. \tag{3.21}$$

Сравнивая (3.20) и (3.21), заключаем, что

$$\frac{a(s)}{1-a(s)} \;=\; \frac{\lambda}{s}\;,$$

откуда можно найти

$$a(s) \;=\; \frac{\lambda}{\lambda+s}\;. \tag{3.22}$$

Но это означает что $A(x)$ – экспоненциальная ф.р.!
Из единственности обратного преобразования L-St. получаем, что тем самым нами доказана следующая теорема:

Теорема 3.2. *Единственным рекуррентным потоком без задержки, у которого среднее число требований, поступающих за промежуток времени $[0,t]$, растет линейно с ростом t, т.е. функция восстановления которого линейна по времени ($H^0(t) = \lambda t$) является поток Пуассона.*

3.3.4 Поток Пальма

Исследуем далее тот же вопрос о линейности по t функции восстановления, но в классе рекуррентных потоков с задержкой, т.е. пусть теперь

$$H(t) \;=\; \lambda t. \tag{3.23}$$

Обозначая $h(s)$ и $a_0(s)$ – преобразования L-St функций $H(t)$ и $A_0(t)$ соответственно, и применяя это преобразование к обеим частям представления (3.13), получим

$$h(s) \;=\; a_0(s)\sum_{k=0}^{\infty} a^k(s) \;=\; \frac{a_0(s)}{1-a(s)}\ . \tag{3.24}$$

Преобразование L-St от обеих частей (3.23) даст, в свою очередь, равенство

$$h(s) \;=\; \frac{\lambda}{s}. \tag{3.25}$$

Из (3.24) и (3.25) заключаем, что

$$\frac{a_0(s)}{1-a(s)} \;=\; \frac{\lambda}{s}\ ,$$

откуда сразу получается следующее соотношение:

$$a_0(s) \;=\; \frac{\lambda(1-a(s))}{s}\ . \tag{3.26}$$

По свойствам преобразования L-St это означает, что при условии (3.23) исходные функции распределения $A_0(\cdot)$ и $A(\cdot)$ тоже должны, с необходимостью, быть связаны соответствующим (3.26) соотношением следующего вида:

$$A_0(t) \;=\; \lambda\int_0^t \Big(1-A(x)\Big)\,dx\ . \tag{3.27}$$

Заметим, что интеграл справа в последнем соотношении определен при любой ф.р. $A(x)$, однако для корректности определения функции $A_0(t)$ мы должны потребовать, чтобы определяемая формулой (3.27) функция действительно была функцией распределения (см. (2.9)). Из вида формулы (3.27) ясно лишь, что функция $A_0(t)$ – неотрицательна, монотонна и $A_0(0) = 0$. Но для того, чтобы $A_0(t) \to 1$ при $t \to \infty$ мы должны, по крайней мере, потребовать существование несобственного интеграла

$$\int_0^{\infty} \Big(1-A(x)\Big)\,dx \;<\; \infty\ .$$

Вспоминая Теорему 2.1 (о вычислении моментов с.в.), заметим, что указанная выше величина представляет собой первый момент распределения $A(x)$.

Таким образом, для того, чтобы соотношение (3.27) задавало действительно ф.р. $A_0(t)$ мы должны дополнительно потребовать конечность первого момента распределения $A(x)$, т.е.

$$a_1 = \int_0^\infty \Big(1 - A(x)\Big)\, dx \; < \; \infty \; . \tag{3.28}$$

Но в таком случае константа λ – интенсивность потока уже не может оставаться произвольной, а с необходимостью должна быть равной $(1/a_1)$, чтобы удовлетворялось условие нормировки $A_0(\infty) = 1$.
И тогда соотношение (3.27) примет окончательно следующий вид:

$$A_0(t) \; = \; \frac{1}{a_1} \int_0^t \Big(1 - A(x)\Big)\, dx \; . \tag{3.29}$$

При этом условии наш искомый рекуррентный поток будет иметь следующую (линейную по времени) функцию восстановления:

$$H(t) \; = \; \frac{t}{a_1}. \tag{3.30}$$

И мы пришли к следующему:

Определение 3.6. *Среди рекуррентных потоков с задержкой, только те потоки, для которых выполняются условия (3.28) и (3.29), имеют линейную по времени функцию восстановления (3.30). Такие потоки называются* **потоками Пальма** *(C.Palm).*

Заметим, что поток Пуассона является в то же самое время и частным случаем потока Пальма с $A_0(t) = A(t) = 1 - e^{-\lambda t}$. Проверка выполнения необходимых требований рекомендуется в качестве самостоятельного упражнения.

3.3.5 Элементарная теорема восстановления Смита

Попробуем ещё расширить класс рассматриваемых рекуррентных потоков за счет ослабления требования о линейности функции восстановления (3.23), сохраняя это требование лишь в асимптотике. Существование таких потоков подтверждается следующей теоремой

Теорема 3.3. *[Смита (W. Smith)] Если рекуррентный поток с задержкой удовлетворяет следующим условиям*

$$\lim_{t\to\infty} A_0(t) \; = \; 1 \; , \tag{3.31}$$

$$a_1 \; = \; \int_0^\infty x \; dA(x) \; < \; \infty \; , \tag{3.32}$$

то существует предел

$$\lim_{t\to\infty} \frac{H(t)}{t} \; = \; \frac{1}{a_1} \; . \tag{3.33}$$

Доказательство.

Воспользуемся вторым интегральным представлением уравнений восстановления (в форме (3.18)). Будем также использовать тот факт, что ф.р. $0 \leqslant A_0(t) \leqslant 1$, и что обе функции $A_0(t)$ и $H^0(t)$ являются неубывающими функциями времени.

Сначала оценим функцию $H(t)$ сверху:

$$\begin{aligned} H(t) &= \int_0^t \Big(\mathbf{1} + H^0(t-u)\Big)\, dA_0(u) \\ &= A_0(t) + \int_0^t H^0(t-u)\, dA_0(u) \\ &\leqslant 1 + \int_0^t \sup_{0 \leqslant v \leqslant t} H^0(t-v)\, dA_0(u) \\ &\leqslant 1 + H^0(t) \int_0^t dA_0(u) \leqslant 1 + H^0(t) \cdot 1 \end{aligned} \tag{3.34}$$

Для получения нижней оценки функции $H(t)$ выберем произвольное число θ $(0 < \theta < 1)$, так чтобы

$$\begin{aligned} H(t) &\geqslant \int_0^t H^0(t-u)\, dA_0(u) \geqslant \int_0^{\theta t} H^0(t-u)\, dA_0(u) \\ &\geqslant \int_0^{\theta t} \inf_{0 \leqslant v \leqslant \theta t} H^0(t-v)\, dA_0(u) \geqslant H^0(t-\theta t)\, A_0(\theta t) \end{aligned} \tag{3.35}$$

Полученные оценки справедливы для любого рекуррентного потока с задержкой, в том числе и для потока Пальма.

Сначала предположим, что рассматривается поток Пальма, а значит $A_0(t)$ имеет специальную форму представления (3.29), и $H(t) = \frac{t}{a_1}$. Тогда из (3.34) следует, что

$$\frac{1 + H^0(t)}{t} \geqslant \frac{H(t)}{t},$$

откуда

$$\frac{H^0(t)}{t} \geqslant \frac{H(t)}{t} - \frac{1}{t} = \frac{1}{a_1} - \frac{1}{t} \to \frac{1}{a_1}, \quad \text{при } t \to \infty .$$

Тем самым мы установили, что

$$\liminf_{t \to \infty} \frac{H^0(t)}{t} \geqslant \frac{1}{a_1} . \tag{3.36}$$

Сделаем далее в (3.35) замену переменных $x = t - \theta t$.
Тогда

$$\frac{H^0(x)}{x} \leqslant \frac{H\Big(x(1-\theta)^{-1}\Big)}{x\ A_0\Big(\theta x(1-\theta)^{-1}\Big)}, \tag{3.37}$$

т.к. согласно замене переменных $t = \frac{x}{1-\theta} = x(1-\theta)^{-1}$.
Поскольку мы пока рассматриваем поток Пальма, то в (3.37)

$$H\Big(x(1-\theta)^{-1}\Big) \;=\; \frac{x(1-\theta)^{-1}}{a_1} \;\equiv\; \frac{x}{a_1(1-\theta)}$$

и, кроме того,

$$\lim_{x\to\infty} A_0\Big(\theta x(1-\theta)^{-1}\Big) \;=\; 1\ .$$

Но тогда

$$\limsup_{x\to\infty} \frac{H^0(x)}{x} \leqslant \frac{1}{a_1(1-\theta)}\ ,$$

а т.к. правая часть не зависит от x, и параметр θ может быть сколь угодно мал, отсюда окончательно получим

$$\limsup_{x\to\infty} \frac{H^0(x)}{x} \leqslant \frac{1}{a_1}\ , \tag{3.38}$$

Из (3.36) и (3.38) следует, что существует предел

$$\lim_{x\to\infty} \frac{H^0(x)}{x} \;=\; \frac{1}{a_1}\ , \tag{3.39}$$

Тем самым, теорема доказана для рекуррентных потоков без задержки, т.к. $H_0(t)$ – функция восстановления потока без задержки.

Наконец, предположим, что $H(t)$ – функция восстановления произвольного рекуррентного потока и выполнены лишь условия теоремы (3.31) и (3.32). Для этой функции, очевидно, также справедливы оценки (3.34) и (3.35), полученные нами для произвольного рекуррентного потока с задержкой.
Тогда для оценки нижнего предела из (3.34) с учетом (3.39) получим

$$\limsup_{t\to\infty} \frac{H(t)}{t} \leqslant \limsup_{t\to\infty} \Big(\frac{1}{t} \;+\; \frac{H^0(t)}{t}\Big) = \lim_{t\to\infty} \Big(\frac{1}{t} \;+\; \frac{H^0(t)}{t}\Big) \;=\; \frac{1}{a_1} \tag{3.40}$$

А из (3.35) вытекает оценка верхнего предела

$$\begin{aligned}\liminf_{t\to\infty} \frac{H(t)}{t} \;&\geqslant\; \lim_{t\to\infty} \frac{H^0(t(1-\theta))\ A_0(\theta t)}{t} \\ &=\; \lim_{t\to\infty} \frac{H^0(t(1-\theta))\ (1-\theta)\ A_0(\theta t)}{t\ (1-\theta)} \;=\; \frac{(1-\theta)}{a_1}\end{aligned}$$

Откуда вследствие произвольности θ можно заключить, что

$$\liminf_{t\to\infty} \frac{H(t)}{t} \geqslant \frac{1}{a_1} \tag{3.41}$$

Но (3.40) и (3.41) означают существование предела (3.33), что завершает доказательство теоремы.

3.4 Неоднородный поток Пуассона

Пусть теперь интенсивность потока меняется во времени, а в остальном его поведение остается прежним, т.е. удовлетворяет предположениям (3.3), (3.4), но с параметром $\lambda = \lambda(t)$.
Используя обозначение (3.5), предположим в качестве начальных следующие условия (н.у.) на $p_n(t),\ n\geqslant 0$: $p_0(0) = 1;\ p_n(0) = 0,\ n\geqslant 1$, то есть предположим, что в момент времени $t = 0$ в системе не было ни одной заявки.

Выпишем выражения для $p_n(t+\delta t),\ n\geqslant 0$ с точностью до членов $o(\delta t)$, используя свойства (3.6) и (3.7), которые для рассматриваемого потока будут отличаться лишь тем, что теперь $\lambda = \lambda(t)$.
Ясно, что для $n = 0$ отсутствие заявок в момент $(t+\delta t)$ возможно лишь тогда, когда их не было ни одной в момент t и ни одной заявки не пришло за промежуток δt, причем эти события независимы, поскольку соответствующие им интервалы времени не пересекаются. Поэтому имеем

$$p_0(t+\delta t) \ = \ p_0(t)\big(1 - \lambda(t)\ \delta t + o(\delta t)\big). \tag{3.42}$$

Для $n = 1$ в момент $(t+\delta t)$ возможны 2 варианта событий: либо одна заявка уже была к моменту t и ни одной не пришло за δt, либо не было ни одной до t и ровно одна заявка пришла за δt. Тогда

$$p_1(t+\delta t) \ = \ p_1(t)\big(1 - \lambda(t)\ \delta t + o(\delta t)\big) \ + \ p_0(t)\big(\ \lambda(t)\ \delta t + o(\delta t)\ \big). \tag{3.43}$$

В общем случае аналогично получим

$$p_n(t+\delta t) \ = \ p_n(t)\big(1 - \lambda(t)\ \delta t + o(\delta t)\big) \ + \ p_{n-1}(t)\big(\ \lambda(t)\ \delta t + o(\delta t)\ \big), \tag{3.44}$$

где n может быть любым от 0 до ∞.
После несложных преобразований можно написать следующую систему уравнений:

$$\begin{cases} \frac{p_0(t+\delta t)-p_0(t)}{\delta t} \ = \ -\lambda(t)p_0(t) \ + \ \frac{o(\delta t)}{\delta t} \\ \frac{p_1(t+\delta t)-p_1(t)}{\delta t} \ = \ \lambda(t)p_0(t) \ - \ \lambda(t)p_1(t) \ + \ \frac{o(\delta t))}{\delta t} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \frac{p_n(t+\delta t)-p_n(t)}{\delta t} \ = \ \lambda(t)p_{n-1}(t) \ - \ \lambda(t)p_n(t) \ + \ \frac{o(\delta t)}{\delta t} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases}$$

Устремляя $\delta t \to 0$, придем к системе дифференциальных уравнений:

$$\begin{cases} \frac{dp_0(t)}{dt} \ = \ -\lambda(t)p_0(t) \\ \frac{dp_1(t)}{dt} \ = \ \lambda(t)p_0(t) \ - \ \lambda(t)p_1(t) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \frac{dp_n(t)}{dt} \ = \ \lambda(t)p_{n-1}(t) \ - \ \lambda(t)p_n(t) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases} \tag{3.45}$$

Для решения этой системы рассмотрим производящую функцию (см. определение (2.63)) нашей последовательности $p_n(t)$

$$\pi(z,t) \ = \ \sum_{n=0}^{\infty} p_n(t)z^n. \tag{3.46}$$

Отметим здесь, что поскольку $p_n(t)$ зависели от t, то и соответствующая производящая функция стала функцией двух переменных.
Умножая каждое из уравнений системы (3.45) на z в соответствующей степени и складывая затем все уравнения почленно, получим

$$\sum_{n=0}^{\infty} z^n \frac{dp_n(t)}{dt} = \lambda(t)\ z \sum_{n=0}^{\infty} z^n p_n(t) - \lambda(t) \sum_{n=0}^{\infty} z^n p_n(t),$$

или

$$\frac{\partial \pi(z,t)}{\partial t} \ = \ \lambda(t)(z-1)\pi(z,t). \tag{3.47}$$

Решение этого уравнения можно найти следующим образом:

$$\int \frac{d\pi}{\pi} \ = \ (z-1) \int \lambda(t)\ dt\ ;$$

$$\left[\ln \pi(z,t)\right]_{t=0}^{t} \ = \ (z-1) \int_0^t \lambda(u)\ du\ ;$$

$$\frac{\pi(z,t)}{\pi(z,0)} \ = \ \exp\left[\ (z-1) \int_0^t \lambda(u)\ du\ \right]\ .$$

В соответствии с выбранными н.у. имеем $\pi(z,0) = 1$, поэтому искомое решение уравнения (3.47) примет окончательно следующий вид:

$$\pi(z,t) \ = \ e^{(z-1)\Lambda(t)}\ , \tag{3.48}$$

где

$$\Lambda(t) = \int_0^t \lambda(u)\ du\ . \tag{3.49}$$

Разлагая производящую функцию (3.48) в ряд по степеням z, получим

$$\pi(z,t) \ = \ e^{-\Lambda(t)} \Big[1 + \Lambda(t)\ z + \frac{\Lambda^2(t)}{2!}\ z^2 + \frac{\Lambda^3(t)}{3!}\ z^3 + \ \dots\ \Big]\ .$$

Откуда по определению производящей функции (3.46) следует

$$p_n(t) \ = \ \text{коэфф. при } z^n \ = \ \frac{\Lambda^n(t)}{n!}\ \ e^{-\Lambda(t)}\ . \tag{3.50}$$

Таким образом, мы пришли к следующему:

Определение 3.7. *Если входящий поток удовлетворяет предположениям (3.3), (3.4), но при этом его интенсивность λ меняется во времени, т.е. $\lambda = \lambda(t)$, то число заявок, приходящих к моменту t, имеет пуассоновское (см. (2.8)) распределение (3.50) с параметром $\Lambda = \Lambda(t)$, задаваемым формулой (3.49). Такой поток имеет специальное название* – **неоднородный поток Пуассона**.

Поскольку

$$\frac{\partial \pi(z,t)}{\partial z}\Big|_{z=1} \;=\; \sum_{n=1}^{\infty} n\; p_n(t)\; z^{n-1}\Big|_{z=1} \;=\; \mathbf{E}\big[N(t)\big] \;=\; \Lambda(t)\;, \qquad (3.51)$$

функция восстановления неоднородного пуассоновского потока есть

$$H(t) \;=\; \Lambda(t). \qquad (3.52)$$

Заметим, в заключение, что при $\lambda(t) = \lambda = const$ из (3.49) вытекает $\Lambda(t) = \lambda t$, и, следовательно, неоднородный пуассоновский поток становится обычным потоком Пуассона.

3.5 Некоторые свойства рекуррентных потоков

Зафиксируем некоторый момент времени t ($T_N \leqslant t < T_{N+1},\ N \geqslant 0$), и представим длину всего временного интервала между моментами прихода N-той и $(N+1)$-ой заявок в виде некоторой с.в. α_t, равной сумме двух случайных величин: $\alpha_t = \beta_t + \gamma_t$. Как мы вскоре увидим, поведение с.в. α_t отличается от известной нам ранее с.в. $e_N = T_{N+1} - T_N$ за счет того, что внутри временного интервала α_t содержится фиксированная точка t – момент времени, о котором точно известно, что до него пришло ровно N заявок, и что $(N+1)$-ая заявка к этому моменту еще не пришла.

Определение 3.8. *Случайная величина β_t, представляющая собой промежуток времени, прошедший с момента прихода N-той заявки до момента t, т.е.*

$$\beta(t) \;=\; \min_k\{(t - T_k) :\ T_k \leqslant t,\ k \geqslant 0\}, \qquad (3.53)$$

называется **дефектом потока**.

Определение 3.9. *Случайная величина γ_t, представляющая собой промежуток времени, равный остаточному времени, отсчитываемому от момента t до прихода ближайшей будущей заявки, т.е.*

$$\gamma(t) \;=\; \min_k\{(T_k - t) :\ T_k > t,\ k \geqslant 1\}, \qquad (3.54)$$

называется **эксцессом потока**.

Найдем формулы для функций распределения случайных величин эксцесса и дефекта сначала в общем виде для произвольного рекуррентного потока, а затем в качестве примера вычислим конкретный вид этих функций для стандартного пуассоновского потока.

По формуле полной вероятности запишем:

$$\mathbf{P}(\gamma_t \leqslant x) \;=\; \mathbf{P}(\gamma_t \leqslant x,\; N(t)=0) \;+\; \sum_{k=1}^{\infty} \mathbf{P}(\gamma_t \leqslant x,\; N(t)=k)$$

$$=\; \mathbf{P}(t < T_1 \leqslant t+x) \;+\; \sum_{k=1}^{\infty} \mathbf{P}\Big(\{\gamma_t \leqslant x\} \cap \{T_k \leqslant t\} \cap \{T_{k+1} > t\}\Big)$$

$$=\; \mathbf{P}(t < e_0 \leqslant t+x) \;+\; \sum_{k=1}^{\infty} \mathbf{P}\Big(\{T_k \leqslant t\} \cap \{t - T_k < e_k \leqslant t + x - T_k\}\Big), \quad (3.55)$$

где при последнем переходе было учтено, что $\gamma_t = T_{k+1} - t$ и $T_{k+1} - T_k = e_k$. Но случайные величины $T_k = \sum_{i=0}^{k-1} e_i$ и e_k независимы, а их распределения оба непрерывны, поэтому, полагая под знаком суммы в (3.55) произвольное допустимое значение $T_k = u,\; 0 \leqslant u \leqslant t$, и представляя вероятность $\mathbf{P}(\cdots)$ в следующем интегральном виде

$$\mathbf{P}(\cdots) = \int d\mathbf{P}_{(T_k, e_k)} = \int\int d\mathbf{P}_{(T_k)}\; d\mathbf{P}_{(e_k)} = \int_0^t dF_{T_k}(u) \int_{t-u}^{t+x-u} dA(u),$$

получим окончательно

$$\mathbf{P}(\gamma_t \leqslant x) \;=\; A_0(t+x) - A_0(t) + \sum_{k=1}^{\infty} \int_0^t \Big(A(t+x-u) - A(t-u)\Big)\, d\Big(A_0 * A_*^{k-1}(u)\Big)\,.$$

С учетом представления (3.13) для функции восстановления это выражение можно переписать следующим образом

$$\mathbf{P}(\gamma_t \leqslant x) \;=\; A_0(t+x) - A_0(t) \;+\; \int_0^t \Big(A(t+x-u) - A(t-u)\Big)\, dH(u)\,. \quad (3.56)$$

Это выражение справедливо для любого рекуррентного потока, поскольку при его выводе было использовано только предположение о независимости членов последовательности $\{e_i\}_{i \geqslant 0}$.

Найдем теперь ф.р. для дефекта β_t.

Заметим сначала, что по определению $\beta_t \leqslant t$ и тогда при всех $x \geqslant t$ сразу получаем $\beta \leqslant t \leqslant x \;\Rightarrow\; \mathbf{P}(\beta_t \leqslant x) = 1$ для $x \geqslant t$.

Поэтому далее будем рассматривать только $x < t$. Принимая во внимание тот факт, что в этом случае, т.е. при $x < t$: $\{N(t) = 0\} \equiv \{\beta_t = t\} \subset \{x < \beta_t\}$. Но тогда $\mathbf{P}(\beta_t > x,\; N(t) = 0) = 1$, и, следовательно, $\mathbf{P}(\beta_t \leqslant x,\; N(t) =$

$0) = 0$. С учетом этого по формуле полной вероятности запишем:

$$\begin{aligned}
\mathbf{P}(\beta_t \leqslant x)\Big|_{x<t} &= \sum_{k=1}^{\infty} \mathbf{P}(\beta_t \leqslant x,\ N(t) = k) \\
&= \sum_{k=1}^{\infty} \mathbf{P}\Big(\{\beta_t \leqslant x\} \cap \{T_k \leqslant t\} \cap \{T_{k+1} > t\}\Big) \\
&= \sum_{k=1}^{\infty} \mathbf{P}\Big(\{t - T_k \leqslant x\} \cap \{T_k \leqslant t\} \cap \{T_{k+1} - T_k > t - T_k\}\Big) \\
&= \sum_{k=1}^{\infty} \mathbf{P}\Big(\{t - x \leqslant T_k \leqslant t\} \cap \{e_k > t - T_k\}\Big) \\
&= \sum_{k=1}^{\infty} \int_{t-x}^{t} \Big(1 - A(t-u)\Big)\ d\Big(A_0 * A_*^{k-1}(u)\Big) \\
&= \int_{t-x}^{t} \Big(1 - A(t-u)\Big)\ dH(u)\ .
\end{aligned}$$

Объединяя оба полученных выражения, окончательно получим

$$\mathbf{P}(\beta_t \leqslant x) \ = \ \begin{cases} 1, & x \geqslant t\ ; \\ \int_{t-x}^{t} \Big(1 - A(t-u)\Big)\ dH(u), & x < t. \end{cases} \tag{3.57}$$

Видим, что ф.р. (3.57) с.в. β_t, в отличие от ф.р. (3.56) с.в. γ_t, имеет скачок в точке $x = t$.

Пример 1. *Подставим теперь в эти формулы конкретные выражения функций $H(u)$ и $A(u)$ для потока Пуассона, т.е.*

$$H(u) \ = \ \lambda u\ , \quad A(u) \ = \ 1 - e^{-\lambda u}$$

и учтем также, что у этого потока $A_0(t) = A(t)$. Получим

$$\begin{aligned}
\mathbf{P}(\gamma_t \leqslant x) &= (1 - e^{-\lambda(t+x)}) - (1 - e^{-\lambda t}) + \lambda \int_0^t \Big(e^{-\lambda(t-u)} - e^{-\lambda(t+x-u)}\Big)\ du \\
&= 1 - e^{-\lambda x}\ .
\end{aligned} \tag{3.58}$$

и

$$\mathbf{P}(\beta_t \leqslant x) \ = \ \begin{cases} 1, & x \geqslant t\ ; \\ 1 - e^{-\lambda x}, & x < t. \end{cases} \tag{3.59}$$

В заключение отметим, что эксцесс γ_t пуассоновского потока не зависит от t и имеет экспоненциальное распределение (3.58), а дефект β_t этого же потока оказывается зависящим от времени и в точке $x = t$ его ф.р. (3.59) претерпевает скачок, равный $e^{-\lambda t}$. Заметим также, что

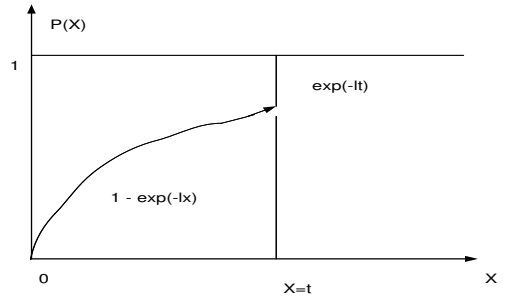


Рис. 3.2: Дефект β_t потока.

при $t \to \infty$ ф.р. дефекта теряет отмеченное вырождение и имеет предел, равный $1 - e^{-\lambda x}$.
Если подсчитать средние значения эксцесса и дефекта, то они окажутся соответственно равными

$$\mathbf{E}\ \gamma_t \ = \ \frac{1}{\lambda}, \quad \mathbf{E}\ \beta_t \ = \ \frac{1}{\lambda}\ (1 - e^{-\lambda t}),$$

и, следовательно, для с.в. $\alpha_t = \beta_t + \gamma_t$ получим

$$\mathbf{E}\ \alpha_t \ = \ \mathbf{E}\ \beta_t \ + \ \mathbf{E}\ \gamma_t \ = \ \frac{1}{\lambda}\ (2 - e^{-\lambda t}) \ > \ \frac{1}{\lambda} \ = \ \mathbf{E}\ e_i, \quad i \geqslant 0$$

для всех $t > 0$.
Однако, никакого парадокса в этом нет, поскольку отрезок α_t – не является просто интервалом между двумя последовательными моментами прихода заявок (как с.в. e_i), а представляет собой особенную с.в., т.к. содержит внутри себя фиксированную точку t. Очевидно, например, что чем длиннее мы выберем отрезок, тем с большей вероятностью точка t окажется внутри него, чего нельзя сказать об обычных с.в. e_i, $i \geqslant 0$.

Пример 2. *В качестве второго примера применения формул (3.56) и (3.57) найдем ф.р. эксцесса потока Пальма. Воспользуемся тем, что для потока Пальма справедливо (3.28) и (3.29), а функция восстановления*

$H(t)$ линейна по t (см. (3.30)). Тогда из (3.56) получим

$$\begin{aligned}
\mathbf{P}(\gamma_t \leqslant x) &= A_0(t+x) - A_0(t) \\
&+ \int_0^t \Big(1 - A(t-u)\Big)\, dH(u) - \int_0^t \Big(1 - A(t+x-u)\Big)\, dH(u) \\
&= \frac{1}{a_1}\int_0^{t+x} \Big(1 - A(u)\Big)\, du - \frac{1}{a_1}\int_0^t \Big(1 - A(u)\Big)\, du \\
&+ \frac{1}{a_1}\int_0^t \Big(1 - A(t-u)\Big)\, du - \frac{1}{a_1}\int_0^t \Big(1 - A(t+x-u)\Big)\, du \\
&= \frac{1}{a_1}\int_0^{t+x} \Big(1 - A(u)\Big)\, du - \frac{1}{a_1}\int_x^{t+x} \Big(1 - A(v)\Big)\, dv \\
&= \frac{1}{a_1}\int_0^{x} \Big(1 - A(u)\Big)\, du \;=\; A_0(x)\;, \qquad (3.60)
\end{aligned}$$

т.е. для любого потока Пальма ф.р. эксцесса не зависит от t и совпадает с ф.р. его задержки $A_0(x) = \mathbf{P}(e_0 \leqslant x)$.

Замечание 3.3. *Выше при вычислении интегралов мы дважды производили замену переменных: сначала $(t-u) = v$, а затем $(t+x-u) = v$.*

Представляет также интерес, используя (3.60), найти выражение для среднего значения эксцесса потока Пальма. Обозначим a_2 – второй момент ф.р. $A(\cdot)$, а σ_A^2 – дисперсию этого распределения (см. также общую формулу вычисления моментов m_s - пункт (2) Теоремы 2.1). Используя соотношение (3.29), получим

$$\begin{aligned}
\mathbf{E}\gamma_t &= \int_0^\infty x\; dA_0(x) \;=\; \frac{1}{a_1}\Big(\frac{2}{2}\Big)\int_0^\infty x\;(1 - A(x))\; dx \\
&= \frac{a_2}{2a_1} \;=\; \frac{(\sigma_A^2 + {a_1}^2)}{2a_1} \;=\; \frac{a_1}{2} \;+\; \frac{\sigma_A^2}{2a_1}\;. \qquad (3.61)
\end{aligned}$$

Видим, во-первых, что среднее значение эксцесса не меньше половины средней длительности интервала между приходами заявок, а во-вторых, оно зависит также и от второго момента распределения $A(\cdot)$, а это, например, означает, что эта величина будет увеличиваться при увеличении дисперсии σ_A^2 (!!!)

Таким образом, если идущая в реке рыба заглатывает опущенный рыбаком крючок в соответствии с требованиями потока Пальма, или если автобусы появляются на остановке по правилам этого потока, то вопрос: "Сколько времени (в среднем) еще осталось ожидать момента удачи?", если вы начали ожидание в некоторый случайный момент t, а точно знаете лишь среднюю характеристику a_1 этого потока – просто некорректен! Одной этой информации (см. (3.61)), оказывается, не достаточно для определения $\mathbf{E}\gamma_t$. Требуется еще знание величины среднего разброса

(дисперсии) длины интервалов между событиями интересующего вас потока, или (что фактически есть то же самое) знание величины второго момента a_2.
Только в единственном случае оказывается $\mathbf{E}\gamma_t = a_1/2$ *– когда* $\sigma_A = 0$, *т.е. когда интервалы детерминированы, а не случайны. Но это означало бы, что наш рыбак через каждые* $a_1/2$ *минут вытаскивает очередную рыбу, а не в среднем (усредненно за весь день) через* $a_1/2$ *минут. Как хорошо известно, ни рыбы, ни автобусы не бывают столь дисциплинированными.*

В качестве самостоятельного упражнения предлагается найти дисперсию эксцесса потока Пальма.

3.6 Стационарность рекуррентных потоков с задержкой

Ранее нами уже было дано определение стационарности потока (см. Определение 3.4). Мы также показали (при доказательстве Теоремы 3.1), что поток Пуассона является стационарным. Следующая ниже теорема отражает стационарное свойство всех рекуррентных потоков с задержкой.

Теорема 3.4. *Если* $\mathbf{T} = \{T_i\}_{i\geqslant 1}$, $T_0 = 0$ *– некоторый рекуррентный поток с задержкой, то он стационарен тогда и только тогда, когда он является потоком Пальма.*

Доказательство.
(I) (необходимость) *(Стационарность* $\Rightarrow$ *Поток Пальма)*

Пусть $\mathbf{T}$ – некоторый стационарный рекуррентный поток с задержкой. Покажем сначала, что тогда его функция восстановления ограничена, т.е. $H(t) = \mathbf{E}N(t) < \infty$ для любого t. Для этого в некоторой точке $t_0 > 0$, такой что $A(t_0) = \alpha < 1$, оценим

$$\begin{aligned} H(t_0) &= A_0 * \sum_{k=0}^{\infty} A_*^k(t_0) \\ &\leqslant A_0(t_0) \sum_{k=0}^{\infty} \Big(A(t_0)\Big)^k < A(t_0) \sum_{k=0}^{\infty} \alpha^k \leqslant \frac{A_0(t_0)}{1-\alpha} \leqslant \frac{1}{1-\alpha} \end{aligned} \quad (3.62)$$

Здесь мы, во-первых, воспользовались следующей очевидной оценкой для свертки Стилтьеса:

$$F * G(x) \equiv \int_0^x F(x-u)\, dG(u) \leqslant F(x) \int_0^x dG(u) = F(x)G(x),$$

затем применили известную формулу для суммы бесконечно убывающей геометрической прогрессии, и, во-вторых, использовали самую грубую, но простую и достаточную в данном случае оценку для ф.р. $A_0(t_0) \leqslant 1$.

Далее (вследствие стационарности) имеем:

$$\mathbf{E}\Big[N(t,\ t+t_0)\Big] \ = \ \mathbf{E}\Big[N(t_0)\Big] \ \text{ для любого } \ t>0.$$

Поэтому из (3.62)

$$H(kt_0) \leqslant \frac{k}{1-\alpha} \ \text{ для любого } \ k\geqslant 1,$$

и, т.к. для любого $t>0$ существует такой $k\geqslant 0$, что $t\leqslant kt_0$, тогда

$$H(t) \leqslant H(kt_0) \leqslant \frac{k}{1-\alpha} \ < \ \infty.$$

Таким образом, функция восстановления любого стационарного рекуррентного потока с задержкой ограничена.

Докажем теперь, что (при наших предположениях) функция восстановления такого потока должна быть линейной по t.
Возьмем величину $t=1$ и обозначим

$$\lambda \ = \ H(1) \ = \ \mathbf{E}\Big[N(0,1)\Big] \ = \ \mathbf{E}\Big[N(1)\Big] \ . \tag{3.63}$$

Как мы только что доказали, величина $\lambda<\infty$. Таким образом, для $t=1$ имеем

$$H(t) \ = \ H(1) \ = \ \lambda \ = \ \lambda\,1 \ = \ \lambda t \ (\text{при } \ t=1) \ . \tag{3.64}$$

Возьмем теперь величину $t=1/n$ для некоторого $n>0$. По свойству аддитивности функции $\mathbf{E}\Big[N(\cdot)\Big]$ и с использованием предположения о стационарности потока из (3.63) можно получить, что

$$\begin{aligned}\lambda=\mathbf{E}\Big[N(1)\Big] \ &= \ \mathbf{E}\Big[\sum_{k=0}^{n-1} N\Big(\frac{k}{n},\frac{k+1}{n}\Big)\Big] \\ &= \ \sum_{k=0}^{n-1}\mathbf{E}\Big[N\Big(0,\frac{1}{n}\Big)\Big] = n\ \mathbf{E}\Big[N\Big(\frac{1}{n}\Big)\Big] = n\ \mathbf{E}\Big[N(t)\Big] \ \text{при } t=1/n \ ,\end{aligned}$$

Откуда

$$H(t) \ = \ \mathbf{E}\Big[N(t)\Big] \ = \ \frac{\lambda}{n} \ = \ \lambda\,\frac{1}{n} \ = \ \lambda t \ \ (\text{при } t=1/n) \ . \tag{3.65}$$

Теперь возьмем величину $t=m$ для некоторого целого $m>0$. Тогда по предположению стационарности можно получить, что

$$\begin{aligned}H(t) \ &= \ \mathbf{E}\Big[\sum_{k=0}^{m-1} N(k,k+1)\Big] \\ &= \ \sum_{k=0}^{m-1}\mathbf{E}\Big[N(0,1)\Big] \ = \ m\ \lambda \ = \ \lambda t \ (\text{при } t=m) \ .\end{aligned} \tag{3.66}$$

Аналогично, для любого рационального $t > 0$, например, для $t = \frac{m}{n}$ получим

$$H(t) = \mathbf{E}\Big[N(\frac{m}{n})\Big] = \Big[\sum_{k=0}^{m-1} \mathbf{E}N(\frac{k}{n}, \frac{k+1}{n})\Big] = \sum_{k=0}^{m-1} \mathbf{E}\Big[N(\frac{1}{n})\Big]$$
$$= \sum_{k=0}^{m-1} \frac{\lambda}{n} = m\frac{\lambda}{n} = \lambda\frac{m}{n} = \lambda t \text{ (при } t = \frac{m}{n}) . \qquad (3.67)$$

И, наконец, если $t > 0$ – некоторое иррациональное число, то выберем тогда две последовательности рациональных чисел $\{\underline{t_i}\}_{i\geqslant 0}$ и $\{\overline{t_i}\}_{i\geqslant 0}$, таких что

$$\lim_{i\to\infty} \underline{t_i} = \lim_{i\to\infty} \overline{t_i} = t,$$

и при этом чтобы для любых i всегда сохранялось

$$\underline{t_i} \leqslant t \leqslant \overline{t_i}.$$

Отсюда, используя монотонность функции $H(t)$, получим

$$\lambda\underline{t_i} = H(\underline{t_i}) \leqslant H(t) \leqslant H(\overline{t_i}) = \lambda\overline{t_i}.$$

Устремляя теперь $i \to \infty$ получим, что равенство $H(t) = \lambda t$ справедливо теперь уже и для любых действительных $t > 0$.

Но мы знаем из предыдущих лекций, что рекуррентный поток с задержкой, у которого $H(t)$ линейна по t, называется потоком Пальма (См. Определение 3.6). Тем самым, первая часть теоремы (необходимость) доказана.

(II) (достаточность) *(Если поток является потоком Пальма, то тогда он стационарен.)*

Пусть теперь **T** – поток Пальма. Тогда используемая при определении стационарности потока с.в. $N(t,\ t+x)$, равная количеству заявок, приходящих за промежуток времени от момента t до $(t+x)$ (см. Определение 3.4), может зависеть только от эксцесса γ_t этого потока, от последовательности интервалов $\{e_i\}$ между приходами заявок, начиная с момента $(t+\gamma_t)$, и, конечно же, от величины x.

Но мы уже доказывали ранее (см. (3.60)), что ф.р. эксцесса потока Пальма совпадает с ф.р. задержки и не зависит от t. Кроме того, как известно, последовательность временных интервалов $\{e_i\}_{i\geqslant 1}$ между приходами заявок является последовательностью н.о.р.с.в. и, следовательно, никак не может зависеть от зафиксированного нами произвольного момента времени $t > 0$. А тогда с.в. $N(t,\ t+x)$ потока Пальма не зависит от t, что означает (по Определению 3.4) стационарность этого потока.

Таким образом, Теорема 3.4 доказана теперь полностью.

Доказанная нами Теорема 3.4 демонстрирует особую роль потоков Пальма среди всех других рекуррентных потоков с задержкой, а именно: только они являются стационарными!

3.7 Прореживание потоков

Сначала будет рассмотрена процедура геометрического прореживания (или просеивания) потока, позволяющая из произвольного рекуррентного потока получать поток Пуассона, а затем предложена другая процедура прореживания, с помощью которой, наоборот, из пуассоновского потока можно строить рекуррентный поток с желаемой ф.р. $A(\cdot)$.

3.7.1 Геометрическое просеивание, теорема Рени

Предположим, что работа некоторого устройства регулярно проверяется и временные интервалы между последовательными моментами проверок представляют собой последовательность н.о.р.с.в. $\{e_i\}_{i\geq 0}$ с общей ф.р. $A(\cdot)$. Пусть с вероятностью q это устройство может выйти из строя на интервале времени между последовательными проверками и тогда оно будет нуждаться в ремонте (замене) в ближайший следующий очередной момент проверки. В противном случае с вероятностью $(1-q)$ это устройство будет, возможно, лишь подрегулировано при очередной проверке. В любом случае, после каждой проверки устройство полностью восстанавливает свои первоначальные характеристики, причем предполагается, что как ремонт устройства, так и его регулировка производятся за нулевое время. Далее с вероятностью q это устройство вновь может выйти из строя до момента следующей проверки.

Рассмотрим случайный поток, состоящий только из моментов ремонта. Очевидно, этот поток тоже будет рекуррентным (вследствие рекуррентности исходного потока проверок и предположения о полном восстановлении характеристик устройства при ремонте). Обозначим ф.р. интервалов между ремонтами $A^{(re)}(x)$. Ясно, что каждый интервал между ремонтами состоит из ν интервалов между проверками, где ν – с.в., имеющая геометрическое распределение

$$\mathbf{P}(\nu = k) \;=\; q(1-q)^{k-1}, \quad k \geq 1. \tag{3.68}$$

Множитель $(1-q)^{k-1}$ означает, что ровно $k-1$ раз при последовательных проверках ремонт не требовался, а множитель q – что на интервале перед k-той проверкой произошла поломка, потребовавшая ремонта. Учитывая независимость поломок и проверок, и тот факт, что каждый интервал времени между последовательными ремонтами является суммой интервалов между проверками, ф.р. $A^{(re)}(x)$ может быть записана следующим образом

$$A^{(re)}(x) = \sum_{k=1}^{\infty} q(1-q)^{k-1} A_*^k(x), \tag{3.69}$$

а соответствующее преобразование L-St. от неё равно

$$a^{(re)}(s) = \sum_{k=1}^{\infty} q(1-q)^{k-1}\; a^k(s) = \frac{q\; a(s)}{1-(1-q)a(s)}. \tag{3.70}$$

Итак, если $\mathbf{T} = \{T_k\}_{k\geq 0},\ T_0 = 0$ – исходный рекуррентный поток, определяемый ф.р. $A(\cdot)$, а $\{\nu_k\}_{k\geq 0}$ – последовательность н.о.р.с.в., распределенных согласно (3.68), то

Определение 3.10. *Поток* $\mathbf{T}^{(re)} = \{T_k^{(re)}\}_{k\geq 0}$, *задаваемый формулами*

$$T_0^{(re)} = 0,\ T_k^{(re)} = T_{\nu_1+\nu_2+\ldots+\nu_k}, \tag{3.71}$$

называется прореженным потоком, полученным из исходного потока $\mathbf{T} = \{T_k\}_{k\geq 0}$ *с помощью геометрического просеивания.*

Обозначим далее $e_k^{(re)} = T_{k+1}^{(re)} - T_k^{(re)},\ k \geq 0$. Эти величины образуют последовательность н.о.р.с.в. $\{e_k^{(re)}\}_k \geq 0$ с ф.р. $\mathbf{P}(e_0^{(re)} \leq x) = A^{(re)}(x)$, определенной в (3.69).

Пусть $a_1 = \mathbf{E}e_0 = \int_0^\infty x\ dA(x) < \infty$. Тогда, очевидно, $\int_0^\infty x\ dA_*^k(x) = ka_1$ и мы можем, используя (3.69), вычислить первый момент распределения временных интервалов в прореженном потоке следующим образом:

$$\begin{aligned} a_1^{(re)} &= \int_0^\infty x\ dA^{(re)}(x) \\ &= \sum_{k=1}^\infty q(1-q)^{k-1} \int_0^\infty x\ dA_*^k(x) \\ &= \sum_{k=1}^\infty q(1-q)^{k-1}(ka_1) = a_1 q \sum_{k=1}^\infty k(1-q)^{k-1} \\ &= a_1 q\Big(1 + 2(1-q) + 3(1-q)^2 + \ldots\Big) = \frac{a_1 q}{q^2} = \frac{a_1}{q} \end{aligned} \tag{3.72}$$

Здесь при вычислении суммы $S = \big(1 + 2(1-q) + 3(1-q)^2 + \ldots\big)$ был использован тот факт, что для бесконечно убывающей (т.к. $(1-q) < 1$) геометрической прогрессии $S_1 = 1+(1-q)+(1-q)^2+\ldots = \frac{1}{q}$, и что искомая S представима в виде $S = S_1 + \big((1-q) + 2(1-q)^2 + \ldots\big) = S_1 + (1-q)S$, откуда $S = \frac{S_1}{q} = \frac{1}{q^2}$.

Но (3.72) означает, что $a_1^{(re)} \to \infty$, при $q \to 0$ даже при фиксированном значении первого момента a_1. Поэтому рассмотрим следующий "нормированный"поток $\mathbf{T}(q) = \{T_k(q)\}_{k\geq 0}$, который образуем с помощью замены масштаба шкалы времени:

$$T_k(q) = q\ T_k^{(re)},\ k \geq 0. \tag{3.73}$$

Если обозначить теперь

$$e_k(q) = T_{k+1}(q) - T_k(q),\ k \geq 0, \tag{3.74}$$

то можно рассмотреть ф.р.

$$A_q(x) = \mathbf{P}\big(e_k(q) \leq x\big) = A^{(re)}\big(\frac{x}{q}\big), \tag{3.75}$$

причем соответствующая ей функция L-St. будет равна

$$a_q(s) = \mathbf{E}\exp(-s\ e_0(q)) = a^{(re)}(qs) = \frac{q\ a(qs)}{1-(1-q)\ a(qs)}\ . \tag{3.76}$$

Теорема 3.5. *[**А. Рени**] Если вид ф.р. $A(\cdot)$ задан, а параметр $q \to 0$, то*

$$A_q(x) \to 1\ -\ \exp\big(-\frac{x}{a_1}\big), \textit{для любых } x \geq 0. \tag{3.77}$$

Доказательство.

Перепишем $a_q(s)$ из (3.76) следующим образом:

$$\begin{aligned} a_q(s) &= \frac{qa(qs)}{1-(1-q)\ a(qs)} = \frac{a(qs)}{\frac{1-a(qs)}{q} + a(qs)} \\ &= \frac{a(qs)}{\big(\frac{a(0)-a(qs)}{qs}\big)s + a(qs)}\ . \end{aligned} \tag{3.78}$$

Так как при $q \to 0$ выражение

$$\Big(\frac{a(0)-a(qs)}{qs}\Big) \to \Big(-\frac{da(s)}{ds}\Big)\Big|_{s=0} = a_1,$$

а функция $a(qs) \to 1$, то из (3.78) вытекает, что

$$\lim_{q\to 0} a_q(s) = \frac{1}{1+a_1 s} \text{ для любых s, } \mathcal{R}e\ s \geq 0. \tag{3.79}$$

Заметим, что справа в (3.79) получилось выражение, равное преобразованию L-St. от экспоненциальной ф.р., то есть от $\big(1-\exp(-\frac{x}{a_1})\big)$. Но преобразование L-St., как известно, обладает свойством непрерывности и поэтому из (3.79) следует (3.77), что и завершает доказательство.

□

Из этой Теоремы 3.5 [**А. Рени**] следует, что при $q \to 0$ предельный поток существует и является пуассоновским, т.е. что "нормированный" рекуррентный поток $\mathbf{T}(q)$ стремится к пуассоновскому потоку в том сиысле, что ф.р. его временных интервалов между приходами заявок (3.75) стремится к экспоненциальному распределению. А значит и все конечномерные распределения, связанные с этим потоком, сходятся в таком же смысле к соответствующим конечномерным распределениям потока Пуассона.

3.7.2 Построение потока с требуемой ф.р.

Предлагаемый ниже метод прореживания пуассоновского потока имеет не только теоретический смысл, но н действительно может быть использован в моделировании для генерации из исходного пуассоновского потока рекуррентных потоков с требуемыми свойствами (т.е. с заданной ф.р. $A(\cdot)$).

Пусть ф.р. $A(x)$ некоторой неотрицательной с.в. имеет плотность $a(x)$. Тогда для неё может быть определена следующая функция

$$r_A(x) = \frac{a(x)}{1 - A(x)},$$

введенная нами ранее как "интенсивность отказов" в (2.16), и при этом ф.р. $A(x)$, согласно (2.17), представима в виде

$$A(x) = 1 - \exp\big(-\int_0^x r_A(u)du\big). \tag{3.80}$$

Предположим сначала, что функция интенсивности отказов $r_A(x)$ ограничена сверху, т.е. существует такая $\lambda > 0$, что

$$\sup_x r_A(x) \le \lambda. \tag{3.81}$$

Рассмотрим пуассоновский поток $\mathbf{T} = \{T_1, T_2, \ldots\}$ с параметром λ из (3.81), полагая, как обычно, $T_0 = 0$, $e_j = T_{j+1} - T_j$, $j \ge 0$. Будем объявлять каждый очередной приход i-того требования из пуассоновского потока "успехом" или "неуспехом", разыгрывая каждый раз "успех" и "неуспех" с вероятностями $\frac{r_A(T_i)}{\lambda}$ и $\big(1 - \frac{r_A(T_i)}{\lambda}\big)$ соответственно (аналогично классической схеме независимых испытаний Бернулли).

Пусть $\kappa = \kappa(1)$ - номер первого момента, совпавшего с "успехом". Тогда если предположить, что вплоть до момента времени x пришло ровно j требований (т.е. $N(x) = j$), то (по определению числа κ) наступление события $\{T_\kappa > x\}$ будет эквивалентно утверждению, что приходы всех этих предыдущих j требований совпадали с выпадением "неуспеха". Вычислим сначала вероятность $p_{(i)}(x)$ - того, что каждый отдельно взятый такой момент $T_i = u$ совпадет с "неуспехом". Для этого вспомним, что по Теореме 2.2 последовательность моментов T_i из пуассоновского потока может рассматриваться как упорядоченная последовательность н.о.р.с.в., взятых из равномерного на $[0, x]$ распределения вероятностей (с плотностью $f(u) = \frac{1}{x}$). Но тогда

$$\begin{aligned} dp_{(i)}(u) &= \mathbf{P}(\text{"неуспех"})\ \mathbf{P}(u < T_i \le u + du) = \big(1 - \frac{r_A(u)}{\lambda}\big)\ \frac{1}{x}\ du, \\ p_{(i)}(x) &= \int_0^x \big(1 - \frac{r_A(u)}{\lambda}\big)\ \frac{1}{x}\ du = 1\ -\ \frac{1}{\lambda x}\int_0^x r_A(u)\ du. \end{aligned} \tag{3.82}$$

А вероятность того, что при приходе всех этих j требований ровно j раз подряд независимо выпадал "неуспех", будет равна

$$\Big(p_{(i)}(x)\Big)^j\ =\ \mathbf{P}\big(T_\kappa\ >\ x | N(x) = j\big)\ =\ \Big(1\ -\ \frac{1}{\lambda x}\int_0^x r_A(u)\ du\Big)^j. \tag{3.83}$$

Вспоминая, что с.в. $N(x)$ имеет пуассоновское распределение, отсюда по формуле полной вероятности имеем

$$\mathbf{P}(T_\kappa\ >\ x)\ =\ \sum_{j=0}^{\infty} \mathbf{P}\big(T_\kappa\ >\ x | N(x) = j\big)\ \frac{(\lambda x)^j}{j!}\ e^{-\lambda x}. \tag{3.84}$$

Подставляя сюда выражения из (3.83), и учитывая (3.82) и представление (3.80), получим

$$\begin{aligned}\mathbf{P}(T_\kappa > x) &= e^{-\lambda x}\sum_{j=0}^{\infty}\Big(p_{(i)}(x)\Big)^j\,\frac{(\lambda x)^j}{j!} = e^{-\lambda x}\sum_{j=0}^{\infty}\frac{\big(\lambda x p_{(i)}(x)\big)^j}{j!}\\ &= e^{-\lambda x}e^{\lambda x p_{(i)}(x)} = \exp\Big(-\lambda x(1-p_{(i)}(x))\Big)\\ &= \exp\Big(-\int_0^x r_A(u)\,du\Big) = 1 - A(x).\end{aligned}$$

Откуда сразу получаем, что интересующая нас ф.р. первого момента, совпавшего с "успехом", равна заданной нам ф.р. $A(x)$, т.е.

$$\mathbf{P}(T_\kappa \le x) = A(x). \tag{3.85}$$

Используя полученный результат, можно предложить следующую процедуру прореживания исходного пуассоновского потока для построения нового рекуррентного потока с желаемой ф.р. $A(x)$ его интервалов между последовательными моментами прихода требований. Мы определили выше с.в. $\kappa = \kappa(1)$ как первый момент исходного потока, совпавший с выпадением "успеха". Обозначим теперь $S_0 = 0$, $S_1 = T_{\kappa(1)}$, $\alpha_o = S_1 - S_0$. Тогда последовательность с.в.

$$T_i^{(1)} = T_{\kappa(1)+i} - S_1 = T_{\kappa(1)+i} - T_{\kappa(1)},\quad i \ge 0, \tag{3.86}$$

будет снова представлять собой ни что иное как пуассоновский поток с интенсивностью λ, который не зависит от величины $\alpha_o = S_1$, поскольку является продолжением исходного (стационарного) пуассоновского потока с переносом начала отсчета времени в момент $T_{\kappa(1)}$.

Далее для потока (3.86) совершенно аналогично определим снова номер первого момента, совпавшего с "успехом", обозначим его $\kappa(2)$ и положим

$$S_2 = S_1 + T^{(1)}_{\kappa(2)} \equiv T_{\kappa(1)+\kappa(2)},\quad \alpha_1 = S_2 - S_1.$$

Ясно, что α_0 и α_1 независимы и $\mathbf{P}(\alpha_1 \le x) = A(x)$.

Определим теперь

$$T_i^{(2)} = T_{\kappa(1)+\kappa(2)+i} - S_2,\ i \ge 0,$$

который снова окажется пуассоновским с интенсивностью λ, и пусть $\kappa(3)$ будет номером первого момента, совпавшего с "успехом" в этом потоке. Положим аналогично

$$S_3 = S_2 + T^{(2)}_{\kappa(3)} \equiv T_{\kappa(1)+\kappa(2)+\kappa(3)},\quad \alpha_2 = S_3 - S_2.$$

Очевидно, процедура может быть продолжена и далее, причем получаемая при этом последовательность с.в. $\{\alpha_0, \alpha_1, \alpha_2, \ldots\}$ будет последовательностью н.о.р.с.в. с общей ф.р. $A(x)$.

Тогда искомый рекуррентный поток есть последовательность моментов времени $0 = S_0 < S_1 < S_2 < S_3 < \dots$, а интервалы между этими моментами $\alpha_k = S_{k+1} - S_k,\ k \geq 0$, имеют общую ф.р. $A(x)$.

Замечание 3.4. *Для корректности данной процедуры при невыполнении условия (3.81) (т.е. когда величина $r_A(x)$ не ограничена) мы будем вынуждены менять интенсивность λ исходного пуассоновского потока на $\Lambda > \lambda$ всякий раз, когда $r_A(x)$ достаточно приблизится к текущему ограничивающему уровню.*

Замечание 3.5. *Частным случаем предложенного метода является процедура, позволяющая проредить имеющийся пуассоновский поток с интенсивностью Λ до пуассоновского потока с желаемой интенсивностью $\lambda < \Lambda$. Для этого всякий раз при генерации события более интенсивного потока мы должны разыгрывать "успех-неуспех"с вероятностями $\frac{\lambda}{\Lambda}$ и $\left(1-\frac{\lambda}{\Lambda}\right)$ соответственно. В случае "успеха" мы выбираем этот момент, а при "неуспехе" пропускаем и переходим к ожиданию очередного события из Λ-потока. Это действительно частный случай предложенного выше метода, поскольку при экспоненциальной ф.р. $A(x)$ соответствующая интенсивность отказа $r_A(x) \equiv \lambda$ (см. (2.40)).*

3.8 Суперпозиция потоков

Рассмотрим теперь "противоположную"к прореживанию процедуру сложения случайных потоков, приводящую в результате к пуассоновскому потоку. А именно, покажем, что суперпозиция (т.е. наложение) достаточно "редких"потоков (которые могут быть даже не рекуррентными) сходится к пуассоновскому потоку.

3.8.1 Постановка задачи, определения и обозначения

Выберем некоторое значение $n \geq 1$ и рассмотрим набор ровно из n независимых между собой произвольных случайных потоков (т.е. n произвольных последовательностей случайных моментов времени), каждый из которых обозначим

$$\mathbf{T}^{(n)}(i) = \{T_{i,k}^{(n)}\}_{k\geq 0},\quad T_{i,0}^{(n)} = 0,\quad 1 \leq i \leq n.$$

Определение 3.11. *Суперпозицией (наложением) потоков $\mathbf{T}^{(n)}(i)$ называется поток $\mathbf{T}^{(n)}$, представляющий собой упорядоченную коллекцию всех с.в. $T_{i,k}^{(n)},\ 1 \leq i \leq n,\ k \geq 0$ из указанного набора.*

При этом всякий раз, когда некоторые $m > 1$ с.в. из этой коллекции окажутся равными друг другу, мы будем сохранять в итоговом потоке все эти m одинаковых моментов, полагая соответствующие $(m-1)$ интервалов между ними равными нулю. (Ниже, после наложения некоторых специальных дополнительных ограничений, подобные ситуации будут автоматически исключены).

Пусть $e_0^{(n)},\ e_1^{(n)},\ e_2^{(n)}, \ldots$ – последовательность временных интервалов в потоке $\mathbf{T}^{(n)}$. Мы будем искать, при каких условиях последовательность потоков $\{\mathbf{T}^{(n)}\},\ n \geqq 1$ сходится к пуассоновскому потоку при $n \to \infty$, в том смысле, что все конечномерные ф.р. величин $(e_0^{(n)},\ e_1^{(n)},\ e_2^{(n)}, \ldots, e_k^{(n)})$ сходятся (для любого k) к соответствующей ф.р. пуассоновского потока. Вспоминая, что в пуассоновском потоке последовательность временных интервалов между моментами приходов требований – есть ни что иное как последовательность н.о.р.с.в. с экспоненциальной ф.р., заключаем, что нам достаточно будет найти условия, при которых для любых k и неотрицательных величин $x_0, x_1, \ldots, x_{k-1}, x$ будет гарантироваться, что

$$\lim_{n\to\infty} \mathbf{P}\big(e_k^{(n)} > x \mid e_0^{(n)} \leq x_0, e_1^{(n)} \leq x_1, \ldots, e_{k-1}^{(n)} \leq x_{k-1}\big) \ = \ \exp(-\lambda x). \quad (3.87)$$

Обозначим далее k - тый интервал времени в i - том потоке $(1 \ \leq \ i \ \leq \ n)$ через

$$e_{i,k}^{(n)} \ = \ T_{i,k+1}^{n} \ - \ T_{i,k}^{(n)}. \quad (3.88)$$

Все эти временные промежутки, вообще говоря, могут быть зависимыми, но мы будем далее использовать только две ф.р., связанные с этой последовательностью, а именно

$$A_{i,0}^{(n)}(x) \ = \ \mathbf{P}(e_{i,0}^{(n)} \leq x),\ 1 \leq i \leq n, \quad (3.89)$$

и

$$A_{i,1}^{(n)}(x) \ = \ \mathbf{P}(e_{i,0}^{(n)} + e_{i,1}^{(n)} \leq x),\ 1 \leq i \leq n, \quad (3.90)$$

которые представляют собой ф.р. моментов прихода первого и второго требований в каждом i - том потоке соответственно.

Ещё раз напомним, что мы предполагаем здесь все складываемые нами потоки $\mathbf{T}^{(n)}(1),\ \mathbf{T}^{(n)}(2),\ \ldots, \mathbf{T}^{(n)}(n)$ независимыми при каждом фиксированном значении n.

Мы упоминали уже, что будем рассматривать суперпозицию так называемых "редких"потоков. Формально это означает, что будут рассматриваться такие последовательные наборы из n потоков, у которых для любого $i,\ 1 \leq i \leq n$, и для любого $x > 0$ вероятность следующего события:
$\mathbf{P}$(хотя бы одно требование поступает на $[0, x]$ в i-том потоке)$= A_{i,0}^{(n)}(x) \to 0$, при $n \to \infty$.
Очевидно, для того, чтобы все складываемые потоки были "редкими", достаточно потребовать выполнения условия

$$\alpha_n(x) \ \equiv \ \max_{1\leq i\leq n} A_{i,0}^{(n)}(x) \ \to \ 0 \quad (3.91)$$

для любого $x > 0$ при $n \to \infty$.
Далее до конца данного раздела мы будем всюду предполагать, что условие (3.91) выполняется.

3.8.2 Теорема Григелиониса

Лемма 3.1. *Пусть (3.91) выполняется и каждый* $\mathbf{T}^{(n)}(i)$, $1 \le i \le n$, *состоит только из одного требования. Тогда* $\mathbf{T}^{(n)}$ *сходится к пуассоновскому потоку с параметром* λ *тогда и только тогда, когда для любого* $x > 0$

$$\lim_{n\to\infty} \prod_{i=1}^{n} \big(1 - A_{i,0}^{(n)}(x)\big) \;=\; \exp(-\lambda x). \tag{3.92}$$

Доказательство.
1. *(необходимость). Если* $\mathbf{T}^{(n)}$ *сходится к пуассоновскому потоку, то справедливо (3.92).*

Действительно, если каждый исходный поток $\mathbf{T}^{(n)}(i)$, $1 \le i \le n$, состоит только из одного требования, то результирующий поток $\mathbf{T}^{(n)}$ будет состоять точно из n требований. Обозначим временные интервалы этого потока через $e_0^{(n)}$, $e_1^{(n)}, \ldots, e_{n-1}^{(n)}$. Так как все исходные потоки были независимы между собой, то

$$\mathbf{P}\big(e_0^{(n)} > x\big) = \prod_{i=1}^{n} \mathbf{P}\big(e_{i,0}^{(n)} > x\big) = \prod_{i=1}^{n} \big(1 - A_{i,0}^{(n)}(x)\big), \tag{3.93}$$

(поскольку если минимальное больше x, то и все остальные – тоже больше). Но если $\mathbf{T}^{(n)}$ сходится к пуассоновскому потоку, то для любого x

$$\lim_{n\to\infty} \mathbf{P}\big(e_0^{(n)} > x\big) = \exp(-\lambda x),$$

и поэтому (3.92) выполняется.

2. *(достаточность). Пусть теперь (3.91) и (3.92) выполнены. Докажем, что тогда (3.87) выполняется, т.е.* $\mathbf{T}^{(n)}$ *сходится к пуассоновскому потоку.*

Для $k = 0$ равенство (3.87) вытекает из (3.92) и (3.93) непосредственно (т.к. в этом случае выражение под знаком предела в (3.87) есть просто левая часть (3.93)).

Пусть далее k – некоторое целое число $(0 < k \le n)$ и возьмем следующий набор произвольных неотрицательных величин $(x, x_o, x_1, \ldots, x_{k-1})$. Тогда по определению условной вероятности

$$\begin{aligned} &\mathbf{P} \quad \big(e_k^{(n)} > x | e_0^{(n)} \le x_0, e_1^{(n)} \le x_1, \ldots, e_{k-1}^{(n)} \le x_{k-1}\big) \\ &= \quad \frac{\mathbf{P}\big(e_k^{(n)} > x, e_0^{(n)} \le x_0, e_1^{(n)} \le x_1, \ldots, e_{k-1}^{(n)} \le x_{k-1}\big)}{\mathbf{P}\big(e_0^{(n)} \le x_0, e_1^{(n)} \le x_1, \ldots, e_{k-1}^{(n)} \le x_{k-1}\big)} \equiv \frac{A}{B}, \end{aligned} \tag{3.94}$$

где последнее тождество означает ни что иное как введение кратких обозначений для вероятностей, указанных в числителе и знаменателе этой формулы, причём нетрудно заметить, что $B = A|_{x=0}$.

Для упрощения обозначений будем опускать (до конца доказательства) верхний индекс (n) и нижний индекс 0 в обозначении ф.р. $A_{i,0}^{(n)}$, так что $A_{i,0}^{(n)} \equiv A_i$.

Пусть одна из возможных реализаций n с.в. – моментов прихода требований (по одному в каждом из n исходных потоков) дала при построении суперпозиции этих потоков следующую последовательность k, $(k < n)$ индексов $(i_1, i_2, \ldots, i_k)$, выбранных из возможного набора индексов $(1, 2, \ldots, n)$, и соответствующих первым k расположенным по возрастанию моментам прихода требований результирующего потока. Ясно, что все остальные $(n-k)$ моментов прихода требований в получающемся при такой суперпозиции потоке расположатся после момента $T_{i_k} + x$, поскольку $e_k^{(n)} > x$. Так как по условию исходные потоки $T^{(n)}(i)$, $1 \le i \le n$ независимы, то число возможных различных случайных выборок k индексов подобного типа, согласно правилам комбинаторики (выбор без возвращений и с учётом порядка), равно числу $(\mathbf{n!C_n^k})$ штук.

Чтобы найти величину вероятности, обозначенной через A в (3.94), заметим, что событие, стоящее в скобках под знаком этой вероятности, есть ни что иное как одновременное наступление следующих событий:
$\{e_0^{(n)} \le x_0\}$, что означает для $u_1 \equiv T_{i_1}$ выполнение $0 < u_1 \le x_0$;
$\{e_1^{(n)} \le x_1\}$, что означает для $u_2 \equiv T_{i_2}$ выполнение $u_1 \le u_2 \le u_1 + x_1$;
$\ldots$;
$\{e_{k-1}^{(n)} \le x_{k-1}\}$, то есть для $u_k \equiv T_{i_k}$ выполнение $u_{k-1} \le u_k \le u_{k-1} + x_{k-1}$.
А так как все остальные моменты (не равные $T_{i_1}, T_{i_2}, \ldots, T_{i_k}$)) должны наступать позже, чем $T_{i_k} + x \equiv u_k + x$, то для любых $j \ne \{i_1, i_2, \ldots, i_k\}$ должно выполняться $T_j > u_k + x$.

Таким образом, с учётом всего вышесказанного, имеем

$$A = \sum_{(i_1,i_2,\ldots,i_k)} \int_0^{x_0} dA_{i_1}(u_1) \int_{u_1}^{u_1+x_1} dA_{i_2}(u_2) \ldots$$
$$\ldots \int_{u_{k-1}}^{u_{k-1}+x_{k-1}} \prod_{j \ne \{i_1,i_2,\ldots,i_k\}} (1 - A_j(u_k + x))\, dA_{i_k}(u_k), \qquad (3.95)$$

где сумма берётся по всем $(\mathbf{n!C_n^k})$ возможным индексным вариантам.

Стоящее в последнем выражении произведение можно переписать в следующем, более понятном виде:

$$\prod_{j \ne \{i_1,i_2,\ldots,i_k\}} (1 - A_j(u_k + x)) \equiv \frac{\prod_{j=1}^{n}(1 - A_j(u_k + x))}{\prod_{m=1}^{k}(1 - A_{i_m}(u_k + x))}. \qquad (3.96)$$

Отметим также, что в соответствии с введёнными нами обозначениями справедливо неравенство

$$u_k + x \;\equiv\; T_{i_k} + x \;\leq\; x_0 + x_1 + \ldots + x_{k-1} + x. \tag{3.97}$$

Поскольку (как ф.р.) все $A_i(\cdot) \;<\; 1,\;\; (1 \leq i \leq n)$, и аналогичное ограничение справедливо и для введенной в (3.91) неубывающей функции $\alpha_n(\cdot)$, т.е.

$$\alpha_n(x + x_0 + x_1 + \ldots + x_{k-1}) \;<\; 1, \tag{3.98}$$

то для произведения, стоящего в знаменателе (3.96), можно выписать следующую оценку

$$\Big(1 - \alpha_n(x + x_0 + x_1 + \ldots + x_{k-1})\Big)^k \;\leq \prod_{m=1}^{k} (1 - A_{i_m}(u_k + x)) \;\leq\; 1. \tag{3.99}$$

Для оценки произведения, стоящего в числителе (3.96), определим функцию

$$\varepsilon_n(x) \equiv \sup_{0<u\leq x} \Big| \prod_{j=1}^{n} \big(1 - A_j(u)\big) \;-\; \exp(-\lambda u) \Big| \tag{3.100}$$

и обозначим

$$\delta_n(u) \;\equiv\; \varepsilon_n(u)\;\exp(\lambda u). \tag{3.101}$$

Вследствие справедливости условия (3.92) очевидно, что обе эти функции убывают к нулю при $n \to \infty$.

Если в определении функции $\varepsilon_n(x)$ рассмотреть $0 \;<\; u \;\leq\; u_k + x$, то будет справедливо неравенство

$$\varepsilon_n(u_k + x) \geq \Big| \prod_{j=1}^{n} \big(1 - A_j(u_k + x)\big) \;-\; \exp\big(-\lambda(u_k + x)\big)\Big|, \tag{3.102}$$

так как "sup" (по своему определению) не меньше любого из значений стоящей под его знаком функции.

Далее, учитывая (3.97) и определение (3.101), мы можем из (3.102) получить следующее

$$\begin{aligned}
&\Big| \quad \prod_{j=1}^{n} \big(1 - A_j(u_k + x)\big) \;-\; \exp\big(-\lambda(u_k + x)\big) \Big| \\
&\leq \quad \varepsilon_n(x + x_0 + x_1 + \ldots + x_{k-1}) \\
&= \quad \delta_n(x + x_0 + x_1 + \ldots + x_{k-1})\;\exp\big(-\lambda(x + x_0 + x_1 + \ldots + x_{k-1})\big) \\
&\leq \quad \delta_n(x + x_0 + x_1 + \ldots + x_{k-1})\;\exp\big(-\lambda(u_k + x)\big),
\end{aligned}$$

где при последнем сравнении экспонент мы ещё раз использовали (3.97).

Отсюда, снимая знак модуля, окончательно получается следующая оценка для произведения, стоящего в числителе правой части (3.96):

$$\begin{aligned} &\Big(\quad 1-\delta_n(x+x_0+x_1+\ldots+x_{k-1})\Big)\ \exp\Big(-\lambda(u_k+x)\Big) \\ \leq\ &\prod_{j=1}^{n}(1-A_j(u_k+x))\ \leq \\ &\Big(\quad 1+\delta_n(x+x_0+x_1+\ldots+x_{k-1})\Big)\ \exp\Big(-\lambda(u_k+x)\Big). \end{aligned} \tag{3.103}$$

Если теперь обозначить (сравни с (3.95))

$$\begin{aligned} C \quad = \quad & \sum_{(i_1,i_2,\ldots,i_k)} \int_0^{x_0} dA_{i_1}(u_1) \int_{u_1}^{u_1+x_1} dA_{i_2}(u_2)\ldots \\ & \ldots \quad \int_{u_{k-1}}^{u_{k-1}+x_{k-1}} \exp(-\lambda u_k)\ dA_{i_k}(u_k), \end{aligned}$$

то с учетом (3.96),(3.99) и (3.103) окончательно получим следующую оценку для вероятности A:

$$\begin{aligned} C \quad & \exp(-\lambda x)\ \Big(1-\delta_n(x+x_0+x_1+\ldots+x_{k-1})\Big) \\ \leq \quad & A\ \leq \\ C \quad & \exp(-\lambda x)\ \frac{\Big(1+\delta_n(x+x_0+x_1+\ldots+x_{k-1})\Big)}{\Big(1-\alpha_n(x+x_0+x_1+\ldots+x_{k-1})\Big)^k}\ . \end{aligned} \tag{3.104}$$

Аналогично может быть оценена и вероятность, обозначенная нами буквой B в (3.94). Но т.к. формально $B = A|_{x=0}$, то из (3.104) мы сразу получим

$$\begin{aligned} C \quad & \Big(1-\delta_n(x_0+x_1+\ldots+x_{k-1})\Big) \\ \leq \quad & B\ \leq \\ C \quad & \frac{\Big(1+\delta_n(x_0+x_1+\ldots+x_{k-1})\Big)}{\Big(1-\alpha_n(x_0+x_1+\ldots+x_{k-1})\Big)^k}\ . \end{aligned} \tag{3.105}$$

Поскольку условие (3.92) выполнено, то, как уже было замечено выше, из определений функций (3.100) и (3.101) следует, что при $n \to \infty$ обе эти функции стремятся к нулю. Но тогда неравенства (3.104), (3.105), и условие (3.91), которое тоже согласно предположению выполнено, дают

$$\lim_{n\to\infty} \frac{A}{B} \ = \ \exp(-\lambda x). \tag{3.106}$$

□

Рассмотрим, наконец, общий случай, когда исходные потоки $T^{(n)}(i),\ i \le n$ могут иметь любое количество поступающих требований (конечное или счетное).

Теорема 3.6. *[**Григелиониса**] Пусть условия (3.91) и (3.92) выполнены. Предположим, кроме того, что для любого $x > 0$ (см. обозначение (3.90))*

$$\beta_n(x) \ = \ \sum_{i=1}^{n} A_{i,1}^{(n)}(x) \ \to \ 0, \ \textit{при } n \to \infty. \tag{3.107}$$

Тогда $T^{(n)}$ сходится к пуассоновскому потоку с параметром $\lambda > 0$.

Доказательство.

Вначале выясним смысл дополнительного условия (3.107)). Для этого рассмотрим следующее событие:

$$S(x) = \{\text{хотя бы один из } T^{(n)}(i) \text{ имеет на } [0,x] \text{ два или более требований}\}. \tag{3.108}$$

Очевидно,

$$S(x) \ \subseteq \ \bigcup_{i=1}^{n} \{T_{i,2}^{(n)} \le x\} \ = \ \bigcup_{i=1}^{n} \{e_{i,0}^{(n)} + e_{i,1}^{(n)} \le x\}. \tag{3.109}$$

Сразу же оценим (довольно грубо) вероятность этого события, принимая во внимание (3.90), (3.107) и (3.109):

$$\mathbf{P}(S(x)) \ \le \ \sum_{i=1}^{n} \mathbf{P}\Big(e_{i,0}^{(n)} + e_{i,1}^{(n)} \le x\Big) \ = \ \beta_n(x). \tag{3.110}$$

Значит, условие (3.107) требует устранения (в пределе при $n \to \infty$) такой возможности, чтобы хотя бы в одном из исходных потоков $T^{(n)}(i),\ 1 \le i \le n$, имело бы место поступление двух или более требований на любом фиксированном интервале времени $[0,x]$.

Итак, пусть снова $T^{(n)}$ будет обозначать суперпозицию всех исходных потоков $T^{(n)}(i),\ 1 \le i \le n$, а через $T^{(n,1)}$ обозначим тоже суперпозицию, но построенную специальным образом: в неё включим только первые моменты прихода требований из каждого исходного потока, так что поток $T^{(n,1)}$ будет содержать ровно n моментов

$$T_1^{(n,1)}, T_2^{(n,1)}, \ldots, T_n^{(n,1)}.$$

Как обычно, положим

$$T_0^{(n,1)} \ = \ 0, \ \text{а} \ e_k^{(n,1)} \ = \ T_{k+1}^{(n,1)} \ - \ T_k^{(n,1)}, \ \ 0 \le k \le n-1.$$

Обозначим через $\overline{S}(x)$ – событие, дополнительное к событию S, определенному в (3.108), то есть заключающееся в том, что любой i-тый исходный поток состоит точно из одного требования на $[0,x]$. По только что доказанной Лемме 3.1 выполнение условий (3.91) и (3.92) достаточно для сходимости к пуассоновскому потоку суперпозиции таких потоков, каждый из которых состоит только из одного требования.

Утверждение теоремы обобщает утверждение достаточности условий Леммы 3.1 на случай произвольных исходных потоков, для которых выполняется также и требование (3.107). Поэтому доказательство будет заключается в установлении того факта, что (при выполнении условий теоремы) для любого k, $(1 \le k \le n)$ и для любых неотрицательных величин $x,\ x_0,\ x_1,\ \ldots,\ x_{k-1}$ выполняется следующее предельное равенство (сравни с (3.87), (3.94) и (3.106)):

$$\begin{aligned}
\lim_{n\to\infty} \quad & \mathbf{P} \quad \left(e_k^{(n)} > x | e_0^{(n)} \le x_0, e_1^{(n)} \le x_1, \ldots, e_{k-1}^{(n)} \le x_{k-1}\right) \\
& = \lim_{n\to\infty} \frac{\mathbf{P}\left(e_k^{(n)} > x, e_0^{(n)} \le x_0, e_1^{(n)} \le x_1, \ldots, e_{k-1}^{(n)} \le x_{k-1}\right)}{\mathbf{P}\left(e_0^{(n)} \le x_0, e_1^{(n)} \le x_1, \ldots, e_{k-1}^{(n)} \le x_{k-1}\right)} \\
& = \exp(-\lambda x). \qquad (3.111)
\end{aligned}$$

По формуле полной вероятности мы можем написать, что

$$\begin{aligned}
& \mathbf{P} \quad \left(e_k^{(n)} > x,\ e_0^{(n)} \le x_0, \ldots, e_{k-1}^{(n)} \le x_{k-1}\right) \ = \\
& \mathbf{P} \quad \left(e_k^{(n)} > x,\ e_0^{(n)} \le x_0, \ldots, e_{k-1}^{(n)} \le x_{k-1},\ S(x + x_0 + \ldots + x_{k-1})\right) \\
+ \quad & \mathbf{P} \quad \left(e_k^{(n)} > x,\ e_0^{(n)} \le x_0, \ldots, e_{k-1}^{(n)} \le x_{k-1},\ \overline{S}(x + x_0 + \ldots + x_{k-1})\right) \\
& \qquad (3.112)
\end{aligned}$$

и аналогично

$$\begin{aligned}
& \mathbf{P} \quad \left(e_k^{(n,1)} > x,\ e_0^{(n,1)} \le x_0, \ldots, e_{k-1}^{(n,1)} \le x_{k-1}\right) \ = \\
& \mathbf{P} \quad \left(e_k^{(n,1)} > x,\ e_0^{(n,1)} \le x_0, \ldots, e_{k-1}^{(n,1)} \le x_{k-1},\ S(x + x_0 + \ldots + x_{k-1})\right) \\
+ \quad & \mathbf{P} \quad \left(e_k^{(n,1)} > x,\ e_0^{(n,1)} \le x_0, \ldots, e_{k-1}^{(n,1)} \le x_{k-1},\ \overline{S}(x + x_0 + \ldots + x_{k-1})\right), \\
& \qquad (3.113)
\end{aligned}$$

причем

$$\begin{aligned}
& \mathbf{P} \quad \left(e_k^{(n)} > x,\ e_0^{(n)} \le x_0, \ldots, e_{k-1}^{(n)} \le x_{k-1},\ \overline{S}(x + x_0 + \ldots + x_{k-1})\right) \\
= \quad & \mathbf{P} \quad \left(e_k^{(n,1)} > x,\ e_0^{(n,1)} \le x_0, \ldots, e_{k-1}^{(n,1)} \le x_{k-1},\ \overline{S}(x + x_0 + \ldots + x_{k-1})\right), \\
& \qquad (3.114)
\end{aligned}$$

так как при выполнении события $\overline{S}$ все исходные потоки имеют ровно по одному требованию и

$$e_j^{(n)} \equiv e_j^{(n,1)}, \quad 1 \le j \le n-1.$$

Но тогда из (3.112), (3.113) и (3.114), учитывая (3.110) и условие (3.107), нетрудно получить следующую последовательность утверждений:

$$\begin{aligned}
&\Big| \quad \mathbf{P} \quad \Big(e_k^{(n)} > x,\ e_0^{(n)} \le x_0, \ldots, e_{k-1}^{(n)} \le x_{k-1}\Big) \ - \\
&\qquad \mathbf{P} \quad \Big(e_k^{(n,1)} > x,\ e_0^{(n,1)} \le x_0, \ldots, e_{k-1}^{(n,1)} \le x_{k-1}\Big) \ \Big| \\
=&\Big| \quad \mathbf{P} \quad \Big(e_k^{(n)} > x,\ e_0^{(n)} \le x_0, \ldots, e_{k-1}^{(n)} \le x_{k-1},\ S(x + x_0 + \ldots + x_{k-1})\Big) \ - \\
&\qquad \mathbf{P} \quad \Big(e_k^{(n,1)} > x,\ e_0^{(n,1)} \le x_0, \ldots, e_{k-1}^{(n,1)} \le x_{k-1},\ S(x + x_0 + \ldots + x_{k-1})\Big) \ \Big| \\
&\qquad \le \quad \mathbf{P}\Big(S(x + x_0 + \ldots + x_{k-1})\Big) \\
&\qquad \le \quad \beta_n(x + x_0 + \ldots + x_{k-1}) \ \to \ 0 \text{ при } n \to \infty,
\end{aligned}$$

справедливую для любых неотрицательных величин $(x, x_0, \ldots, x_{k-1})$.

Полученный результат вместе с доказанной Леммой 3.1 устанавливает справедливость (3.111), что и заканчивает доказательство теоремы. □

ПРИМЕР.

Рассмотрим некоторый поток Пальма T, задаваемый, как обычно, двумя ф.р. $A(x)$ и $A_0(x) = a_1^{-1}\int_0^x (1 - A(u))\, du$, где $a_1 = \int_0^\infty x\, dA(x) < \infty$. Но дополнительно потребуем, чтобы

$$\lim_{x \downarrow 0} A(x) \ = \ 0. \tag{3.115}$$

Возьмем теперь n независимых потоков Пальма такого типа, произведем в них замену масштаба шкалы времени и проверим, будет ли суперпозиция полученных таким образом "редких" потоков сходиться к пуассоновскому потоку с параметром $\lambda = \frac{1}{a_1}$ при $n \to \infty$.

Определим каждый из исходных потоков $T^{(n)}(i),\ 1 \le i \le n$ с помощью ф.р. $A(\frac{x}{n})$ и $A_0(\frac{x}{n})$. Это означает, что мы выбираем n сходных потоков, устроенных аналогично потоку T. Ясно, что при указанной замене времени среднее значение интервала между последовательными приходами требований в каждом рассматриваемом потоке станет na_1 вместо a_1. Переходя к нашим прежним обозначениям, имеем:

$$A_{i,0}^{(n)}(x) \ = \ A_0(\frac{x}{n}) \ = \ \frac{1}{a_1}\int_0^{\frac{x}{n}} (1 - A(u))\, du, \tag{3.116}$$

$$A_{i,1}^{(n)}(x) \ = \ A_0 * A(\frac{x}{n}). \tag{3.117}$$

Условие (3.91) выполняется, т.к.

$$\begin{aligned} \alpha_n(x) &= \max_{1\le i\le n} A_{i,0}^{(n)}(x) = A_0(\frac{x}{n}) \\ &= \frac{1}{a_1}\int_0^{\frac{x}{n}} (1 - A(u))\, du \le \frac{x}{na_1} \to 0 \end{aligned} \tag{3.118}$$

для любого $x > o$ при $n \to \infty$.
Из (3.115) и (3.116), очевидно, следует, что

$$1 - A_{i,0}^{(n)}(x) = 1 - A_0(\frac{x}{n}) = 1 - \frac{x}{na_1} + o(\frac{1}{n}).$$

Поэтому

$$\prod_{i=1}^{n}\Big(1 - A_{i,0}^{(n)}(x)\Big) = \Big(1 - \frac{x}{na_1} + o(\frac{1}{n})\Big)^n \to \exp\Big(-\frac{x}{a_1}\Big)$$

при $n \to \infty$, т.е. и условие (3.92) тоже выполняется.
Проверим, наконец, выполняется ли дополнительное условие (3.107). Из (3.117) и (3.118) вытекает следующая оценка:

$$A_{i,1}^{(n)}(x) \le A_0(\frac{x}{n})\, A(\frac{x}{n}) \le \frac{x}{na_1}\, A(\frac{x}{n}).$$

Поэтому

$$\beta_n(x) = \sum_{i=1}^{n} A_{i,1}^{(n)}(x) \le \sum_{i=1}^{n} \frac{x}{na_1}\, A(\frac{x}{n}) = \frac{x}{a_1}\, A(\frac{x}{n}).$$

Но по условию (3.115) ф.р. $A(\frac{x}{n}) \to 0$ при $n \to \infty$, и, следовательно, (3.107) выполняется.

Тем самым все требования Теоремы 3.6 [**Григелиониса**] выполнены и, следовательно, суперпозиция рассматриваемых в данном примере потоков Пальма действительно сходится к пуассоновскому потоку с интенсивностью $\lambda = \frac{1}{a_1}$.

Глава 4

Элементарные методы теории МО

В этой главе мы, наконец, переходим к исследованию самих систем массового обслуживания. Вполне естественно начать с рассмотрения простейших методов, позволяющих непосредственно получать результаты для некоторых конкретных моделей систем МО.

4.1 Система $M_\lambda|M_\mu|1|\infty$ в установившемся режиме

Будем предполагать, что изучаемая система работает с дисциплиной обслуживания FIFO. В соответствии с символикой Д. Кендалла, система имеет один обслуживающий сервер, неограниченную емкость бункера-накопителя и экспоненциальное распределение времен обслуживания $\{s_i\}$ с параметром μ, а на ее вход поступает пуассоновский поток с интенсивностью λ.

Рассмотрим вероятностные характеристики работы системы на достаточно малом временном интервале $(t, t+\delta t)$, причем выписывая вероятности мы будем опускать все члены, представляюие собой величины с порядком малости $o(\delta t)$. В таком случае вероятность того, что за время δt в систему придет ровно одна заявка, будет равна $\lambda\delta t$, а того, что ни одной заявки не поступит в систему за этот же промежуток времени – $(1-\lambda\delta t)$. Аналогично, вероятность того, что на этом интервале времени закончится идущий процесс обслуживания заявки, будет равен $\mu\delta t$, а того, что идущее обслуживание не закончится на этом интервале – $(1-\mu\delta t)$, тоже с точностью до членов $o(\delta t)$.

Вероятность того, что в произвольно взятый момент времени t в системе находится ровно n заявок (одна - в сервере и $(n-1)$ - в очереди

ожидания обслуживания) при условии, что в начальный момент $t = 0$ в системе не было ни одной заявки, обозначим

$$p_n(t) = \mathbf{P}\big(N(t) = n \big| N(0) = 0\big), \quad n = 0, 1, 2, \ldots \tag{4.1}$$

Ясно, что в такой системе в момент $(t+\delta t)$ не будет ни одной заявки лишь если в ней и не было ни одной заявки в момент t и при этом ни одной не пришло за время δt, либо была одна заявка в момент t, но ее обслуживание закончилось за промежуток времени δt и при этом ни одной новой заявки не пришло за тот же δt. Таким образом, соответствующие вероятности можно связать следующим равенством:

$$p_0(t + \delta t) = p_0(t)(1 - \lambda\delta t) + p_1(t)\mu\delta t(1 - \lambda\delta t) \tag{4.2}$$

Аналогично,

$$p_1(t + \delta t) = p_0(t)\lambda\delta t + p_1(t)(1 - \lambda\delta t)(1 - \mu\delta t) + p_2(t)(1 - \lambda\delta t)\mu\delta t, \tag{4.3}$$

то есть в системе будет ровно одна заявка в момент времени $(t + \delta t)$, если либо в ней не было ни одной в момент t, но ровно одна пришла за промежуток δt; либо в ней была одна заявка в момент t, а за δt процесс ее обслуживания не закончился и ни одной новой не пришло; либо были две в момент t , но для одной из них процесс обслуживания закончился и при этом ни одной новой заявки за δt не пришло.
Для произвольного $n > 1$, рассуждая совершенно аналогично, можно записать

$$p_n(t+\delta t) = p_{n-1}(t)\lambda\delta t(1-\mu\delta t)+p_n(t)(1-\lambda\delta t)(1-\mu\delta t)+p_{n+1}(t)(1-\lambda\delta t)\mu\delta t. \tag{4.4}$$

При $\delta t \to 0$ из (4.2)-(4.4) получим следующую систему дифференциальных уравнений:

$$\begin{cases} dp_0(t)/dt = -\lambda p_0(t) + \mu p_1(t) \\ dp_1(t)/dt = \lambda p_0(t) - (\lambda + \mu)p_1(t) + \mu p_2(t) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ dp_n(t)/dt = \lambda p_{n-1}(t) - (\lambda + \mu)p_n(t) + \mu p_{n+1}(t) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases} \tag{4.5}$$

Решить эти уравнения достаточно трудно, и, к тому же, не представляет особого интереса, поскольку чаще интересуются поведением системы в установившемся режиме, т. е. при $t \to \infty$. Как будет указано в следующей главе, при условии $\varrho = \lambda/\mu < 1$ у системы дифференциальных уравнений (4.5) существует установившееся решение, то есть существуют следующие пределы

$$\lim_{t\to\infty} p_n(t) = p_n, \quad \lim_{t\to\infty} \frac{dp_n(t)}{dt} = 0, \quad n = 0, 1, 2, \ldots.$$

Величина ϱ, равная среднему числу новых заявок, поступающих в течение среднего времени обслуживания одной заявки, носит название *интенсивность трафика.*

Итак, при условии $\lambda < \mu$ система дифференциальных уравнений (4.5) преобразуется в следующую систему алгебраических уравнений

$$\begin{cases} 0 = -\lambda p_0 + \mu p_1 \\ 0 = \lambda p_0 - (\lambda + \mu) p_1 + \mu p_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ 0 = \lambda p_{n-1} - (\lambda + \mu) p_n + \mu p_{n+1} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases} \tag{4.6}$$

Решение системы может быть легко найдено просто последовательно разрешая ее уравнения. Однако нам для дальнейшего исследования будет полезно сначала найти производящую функцию $P(z)$ последовательности p_n (см. (2.63)). Поэтому будем решать систему (4.6) методом производящей функции. Для этого умножим каждое из уравнений этой системы на $z^0, z^1, z^2, \ldots$ соответственно, а затем сложим их все почленно. Получим

$$0 = \lambda z \sum_{n=0}^{\infty} p_n z^n - \lambda \sum_{n=0}^{\infty} p_n z^n - \mu \sum_{n=1}^{\infty} p_n z^n + \frac{\mu}{z} \sum_{n=1}^{\infty} p_n z^n,$$

или

$$\lambda z P(z) - \lambda P(z) - \mu\Big(P(z) - p_0\Big) + \frac{\mu}{z}\Big(P(z) - p_0\Big) = 0.$$

Из последнего нетрудно найти выражение для производящей функции

$$P(z) = \frac{p_0}{1 - \lambda z/\mu} = \frac{p_0}{1 - \varrho z} \tag{4.7}$$

Но по определению производящей функции и условию нормировки должно выполняться

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

А тогда из (4.7) вытекает, что

$$p_0 = 1 - \varrho \ , \tag{4.8}$$

и, следовательно,

$$P(z) = \frac{1 - \varrho}{1 - \varrho z} = (1 - \varrho)[1 + \varrho z + \varrho^2 z^2 + \ldots] \ , \tag{4.9}$$

откуда (снова по определению производящей функции) получим

$$p_n = (1 - \varrho)\varrho^n, \quad n = 0, 1, 2, \ldots \ . \tag{4.10}$$

Заметим, что установившиеся значения вероятностей $p_n,\ n = 0, 1, 2, \ldots$ означают как вероятности нахождения в системе соответствующего количества n заявок в некоторый момент времени t (достаточно удаленный), так и соответствующую долю времени (при достаточно большой длительности наблюдения T), в течение которой система (суммарно) находилась в состоянии, когда в ней было ровно n заявок. Обе интерпретации пригодны, верны и полезны.

Например, доля времени простоя системы обслуживания (когда система пуста, т. е. в ней нет ни одной заявки) равна p_0, а доля времени занятости системы – есть остаток $(1 - p_0)$, который в рассматриваемой системе, согласно (4.8), оказывается равным ϱ.

4.1.1 Среднее число и дисперсия заявок в системе

Среднее число заявок, находящихся в системе, принято называть *загрузкой системы МО* и обозначать латинской буквой L (от английского слова "Load"). Мы можем определить среднее число заявок, находящихся в рассматриваемой системе (при установившемся режиме ее работы), с использованием вероятностей (4.10) по хорошо известной формуле

$$\mathbf{E}[N] = \sum_{n=0}^{\infty} np_n,$$

либо с помощью производящей функции (4.9), поскольку из определения производящей функции нетрудно получить, что

$$P^{'}(z)\Big|_{z=1} = \sum_{n=1}^{\infty} np_n = \mathbf{E}[N]. \tag{4.11}$$

Отсюда и из (4.9) найдем

$$L = \mathbf{E}[N] = \frac{\varrho\ (1-\varrho)}{(1-\varrho z)^2}\Big|_{z=1} = \frac{\varrho}{1-\varrho} = \frac{\lambda}{\mu - \lambda}\ . \tag{4.12}$$

Для вычисления дисперсии числа заявок тоже воспользуемся свойством (2.66) производящей функции, согласно которому вторая производная производящей функции в точке $z = 1$ равняется второму факториальному моменту, или

$$P^{''}(1) = \frac{2\ \varrho^2}{(1-\varrho)^2} = \mathbf{E}[N(N-1)] = \mathbf{E}[N^2] - \mathbf{E}[N].$$

Поскольку, как известно,

$$D_N = Var[N] = \mathbf{E}[N^2] - \Big(\mathbf{E}[N]\Big)^2,$$

то после подстановок и преобразований получим следующее выражение для дисперсии

$$D_N = \frac{2\ \varrho^2}{(1-\varrho)^2} + \frac{\varrho}{1-\varrho} - \frac{\varrho^2}{(1-\varrho)^2} = \frac{\varrho}{(1-\varrho)^2}\ . \tag{4.13}$$

4.1.2 Длина очереди

Вспомним, что у рассматриваемой системы имеется одно сервисное устройство, не способное одновременно обслуживать более, чем одну заявку. Поэтому, если система не пуста и в ней находится ровно $n\ \ (n = 1, 2, \ldots)$ заявок, то одна из них обслуживается, а остальные $(n-1)$ ожидают в очереди.

Вычислим среднюю длину этой очереди, обозначив ее величину L_q

$$L_q = \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty} n\ p_n - \sum_{n=1}^{\infty} p_n = L - (1-p_0) = L - \varrho \tag{4.14}$$

Или, подставив сюда выражение (4.12) для L,

$$L_q = \frac{\varrho}{1-\varrho} - \varrho = \frac{\varrho^2}{(1-\varrho)} = \frac{\lambda^2}{\mu(\mu-\lambda)} \tag{4.15}$$

Заметим, что полученная нами величина L_q – длина очереди оказалась равной $(\mathbf{E}[N] - \varrho)$, а не $(\mathbf{E}[N] - 1)$, как можно было бы предположить!

4.1.3 Длительность ожидания обслуживания

Если в системе (при прибытии очередной заявки) не оказывается ни одной заявки ни в очереди, ни в сервере, то она сразу же попадает в сервер, то есть длительность ожидания в очереди для нее оказывается равной нулю. Как мы уже знаем, в установившемся режиме работы вероятность такой ситуации равна p_0.

Во всех остальных случаях очередная приходящая заявка встает в очередь ожидающих обслуживания и будет ожидать в ней до тех пор, пока все пришедшие раньше нее заявки не закончат обслуживание (в соответствии с рассматриваемой нами дисциплиной обслуживания FIFO).

Предположим, что пришедшая заявка застала в системе ровно $n,\ n \geq 1$ ранее пришедших заявок, и пусть время ожидания начала обслуживания для этой заявки представляет собой некоторую случайную величину X с плотностью распределения $\varphi(x)$. Будем искать выражение для этой плотности, используя следующее известное определение

$$\mathbf{P}\Big(x\ \leq\ X\ \leq\ x+\delta x\Big)\ =\ \varphi(x)\ \delta x \tag{4.16}$$

В соответствии с вышесказанным, в рассматриваемом случае длительность ожидания обслуживания X (принимающая свои значения $x > 0$) начинает свой отсчет в момент прибытия очередной заявки и заканчивается в момент попадания этой заявки в сервер, или в момент окончания обслуживания последней из предшествовавших ей n заявок. Но тогда вероятность события, указанного в левой части (4.16), представляет собой сумму по всем возможным значениям n вероятностей событий, заключающихся в том, что в системе находилось ровно n заявок, ровно $(n-1)$ из них были обслужены за интервал времени $[0, x]$ и одна заявка (последняя из этих n) закончила свое обслуживание на интервале времени δx:

$$\mathbf{P}\Big(x \;\le\; X \;\le\; x + \delta x\Big) \;=\; \sum_{n=1}^{\infty} p_n \;\frac{(\mu x)^{n-1}\; e^{-\mu x}}{(n-1)!}\;\mu\;\delta x \qquad (4.17)$$

Подставляя сюда выражение для p_n из (4.10), производя преобразования и сворачивая сумму, получим для плотности $\varphi(x)$ (которая, согласно (4.16), есть коэффициент при δx) следующее выражение

$$\varphi(x) \;=\; \lambda\,\Big(1 - \frac{\lambda}{\mu}\Big)\; e^{-(\mu-\lambda)x}\;, \quad x > 0\;. \qquad (4.18)$$

Стоит еще раз подчеркнуть, что плотность вероятности (4.18) была получена нами в предположении $x > 0$. А случаи, когда $x = 0$, имеют вероятность p_0. В качестве упражнения рекомендуется самим проверить справедливость указанной нормировки, а именно

$$p_0 \;+\; \int_0^{\infty} \varphi(x)\; dx \;=\; 1$$

Используя найденную плотность вероятности $\varphi(x)$, нетрудно вычислить *среднее время ожидания в очереди*, обозначаемое обычно W_q:

$$W_q = \mathbf{E}[X] = \int_0^{\infty} x\;\varphi(x)\;dx \;=\; \frac{\lambda}{\mu(\mu-\lambda)} \qquad (4.19)$$

4.1.4 Полное время пребывания в системе

Полное время пребывания в системе складывается из времени обслуживания и времени ожидания в очереди, причем для тех заявок, которые приходят в пустую систему, последнее, как известно, равно нулю.

4.2 Доказательство формулы Литтла

Пусть $T = 1/\lambda$ – есть среднее значение промежутка времени между двумя последовательными приходами заявок в систему, L – среднее количество заявок, находящихся в системе в некоторый произвольный момент времени, а W – среднее значение полного времени пребывания отдельной заявки в системе. Оказывается, при выполнении некоторых, вполне естественных условий (которые выполняются для всех применяемых на практике моделей систем МО), эти три величины связаны между собой следующим соотношением

$$W = TL, \quad \text{или} \quad L = \lambda W. \tag{4.20}$$

Это соотношение было доказано Литтлом (John D.C.Little) в 1960 году и имеет с тех пор общепринятое название *формула Литтла*. Формула очень полезна, поскольку иногда при исследовании моделей МО бывает проще определить (или оценить) L, чем W, а иногда – наоборот. Зная же одну из этих двух важных характеристик работы системы, мы по формуле Литтла автоматически получаем и вторую. Ниже мы рассмотрим доказательство этой формулы, следуя публикации ее автора [Little J.].

Отметим сначала, что Литтлу удалось доказать свою формулу в самых общих предположениях о системе с очередью. Считается, что математически описан процесс обслуживания, при котором каждая заявка прибывает в систему, проводит там некоторое время (ожидание плюс обслуживание), а затем покидает эту систему. Другими словами, предполагается, что некоторым образом генерируются следующие три вероятностные процесса:

$\{n_t,\ -\infty < t < \infty\}$ – количество заявок в системе в момент t ;

$\{w_i,\ -\infty < i < \infty\}$ – время, проведенное в системе i-той заявкой ;

$\{e_i,\ -\infty < i < \infty\}$ – время между прибытием i-той и (i+1)-ой заявок .

Эти процессы определены на некотором пространстве Ω и любой его точке $\omega \in \Omega$ соответствуют следующие функция и две последовательности:

$$n_t(\omega), \quad \{w_i(\omega)\}, \quad \{e_i(\omega)\} \ ,$$

представляющие собой конкретные реализации трех упомянутых выше вероятностных процессов. Причем случайные величины n_t, w_i и e_i принимают лишь неотрицательные значения. Моменты прихода заявок в систему обозначаются t_i и определяются рекуррентно

$$t_{i+1}(\omega) = t_i(\omega) + e_i(\omega).$$

До конца этого раздела мы для удобства будем предполагать, что

$$t_1(\omega) \geq 0, \quad t_0(\omega) < 0.$$

Для задания основного процесса обслуживания, следуя автору, будем использовать индикаторную функцию

$$u(x) = \begin{cases} 1, \text{для } x \geq 0; \\ 0, \text{для } x < 0. \end{cases} \tag{4.21}$$

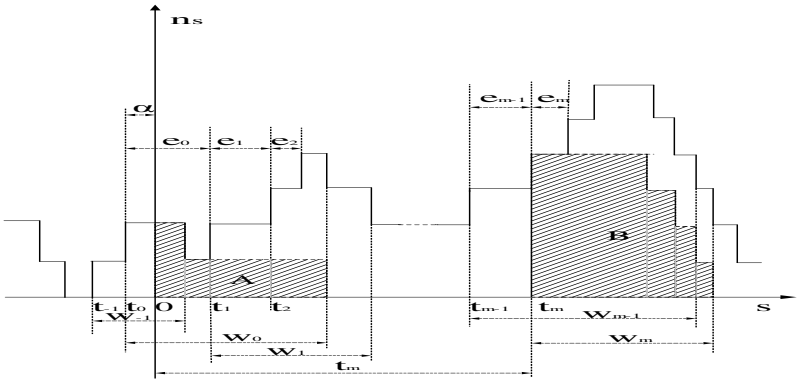


Рис. 4.1: Пример реализации n_s

Тогда для любой точки ω можно записать

$$n_t \;=\; \sum_{j=-\infty}^{\infty} u(t-t_j)\cdot u(t_j+w_j-t). \tag{4.22}$$

Это означает, что в любой момент времени t из всех возможных заявок в системе находятся лишь те из них, у которых время прихода было меньше (или равно) t, и при этом соответствующее время ухода из системы – больше (или равно) t, т.е. принадлежащие множеству $\{t \geq t_j\} \cap \{t \leq t_j + w_j\}$.
Далее нам также потребуется функция $v(x)$, являющаяся интегралом от индикатора $u(x)$, а именно

$$v(x) \;=\; \begin{cases} x, \text{для } x > 0; \\ 0, \text{для } x \leq 0. \end{cases} \tag{4.23}$$

Для некоторого фиксированного ω на Рис.4.1 изображена часть реализации процесса $n_t(\omega)$ в виде графика функции n_s от s. Для конкретности на рисунке в качестве дисциплины обслуживания принята дисциплина FIFO (т.е. считается, что убытие заявок из системы происходит в том же порядке, что и их прибытие). В дальнейшем же изложении и при доказательстве результата конкретный тип дисциплины обслуживания вообще никоим образом использоваться не будет.

Данный рисунок удобен для понимания интегрирования функции (4.22)

$$\int_0^{t_m} n_s\, ds \;=\; \sum_{j=1}^{m} w_j \;+\; \sum_{j\leq 0} v(w_j+t_j) \;-\; \sum_{j\leq m} v(w_j+t_j-t_m). \tag{4.24}$$

Площадь под кривой n_s, подсчитанная от 0 до t_m, оказалась равной сумме времен пребывания в системе заявок, пришедших внутри этого отрезка интегрирования (то есть вплоть до момента t_m включительно) с учетом двух поправок на его концах. Эти поправки, которые соответствуют последним двум суммам в правой части (4.24), равны соответственно площадям заштрихованных на рисунке фигур A и B.

Для проведения усреднения нам потребуются следующие величины

$$\begin{aligned} L_m(\omega) &= \frac{1}{t_m}\int_0^{t_m} n_s(\omega)\, ds\ , \\ W_m(\omega) &= \frac{1}{m}\sum_{j=1}^{m} w_j(\omega)\ , \\ T_m(\omega) &= \frac{1}{m}\sum_{j=0}^{m-1} e_j(\omega)\Big(=\frac{t_m+\alpha}{m}\Big)\ . \end{aligned} \tag{4.25}$$

Наряду с ними рассмотрим также и их пределы

$$\begin{aligned} L(\omega) &= \lim_{t\to\infty}\frac{1}{t}\int_0^{t} n_s(\omega)\, ds\ , \\ W(\omega) &= \lim_{m\to\infty}\frac{1}{m}\sum_{j=1}^{m} w_j(\omega)\ , \\ T(\omega) &= \lim_{m\to\infty}\frac{1}{m}\sum_{j=0}^{m-1} e_j(\omega). \end{aligned} \tag{4.26}$$

Лемма 4.1. *Если каждый из процессов n_t, w_i и e_i является строго стационарным и имеет конечное среднее значение, а процесс e_i , к тому же, метрически транзитивен и его среднее значение $T \equiv 1/\lambda > 0$, то (с вероятностью 1) пределы (4.26) существуют, являются конечными и удовлетворяют соотношению*

$$W(\omega) \;=\; T(\omega)\cdot L(\omega). \tag{4.27}$$

Доказательство.

Существование и конечность пределов (4.26) непосредственно следуют из эргодических теорем для строго стационарных процессов (см., например, [Doob J.L.]).

Чтобы подсчитать пределы в (4.25), заметим сначала, что $m \to \infty$ влечет за собой и $t_m \to \infty$, с вероятностью 1. Действительно, вследствие справедливости эргодической теоремы, предположения о метрической транзитивности случайного процесса e_i и того, что среднее значение этого процесса отлично от нуля, имеем:

$$\frac{1}{T_m(\omega)} \;\to\; \frac{1}{T(\omega)} \;=\; \frac{1}{T} \;<\; \infty, \quad \text{с вероятностью 1.} \tag{4.28}$$

А тогда с использованием (4.25) получаем

$$\frac{1}{T_m(\omega)} \;=\; \frac{m}{t_m+\alpha} \;\to\; \lim\frac{m}{t_m}\;, \quad \text{с вероятностью 1.} \tag{4.29}$$

Из (4.28) и (4.29) следует, что

$$\lim\frac{m}{t_m} \;<\; \infty, \quad \text{с вероятностью 1.}$$

Но это и означает, что сам факт $m \to \infty$ влечет за собой с необходимостью и $t_m \to \infty$, с вероятностью 1.

Используя обозначения (4.25), и заменяя затем соответствующий интеграл результатом его вычисления из (4.24), рассмотрим следующую разность

$$\begin{aligned}
W_m \;& - T_m \cdot L_m \\
=\;& \frac{1}{m}\sum_{j=1}^{m} w_j \;-\; \Big(\frac{t_m+\alpha}{m}\Big)\cdot L_m \\
=\;& \frac{1}{m}\sum_{j=1}^{m} w_j - \frac{1}{m}\Big(1+\frac{\alpha}{t_m}\Big)\int_0^{t_m} n_s\; ds \\
=\;& \frac{1}{m}\sum_{j=1}^{m} w_j - \frac{1}{m}\Big(\sum_{j=1}^{m} w_j + \sum_{j\le 0} v(w_j+t_j) - \sum_{j\le m} v(w_j+t_j-t_m)\Big) - \frac{\alpha}{m}L_m \\
=\;& \frac{1}{m}\sum_{j\le m} v(w_j+t_j-t_m) - \frac{1}{m}\sum_{j\le 0} v(w_j+t_j) - \frac{\alpha L_m}{m}.
\end{aligned}$$

Поскольку в предположениях Леммы периоды занятости и простоя при работе системы чередуются, то в сумме, стоящей в середине последней строки, содержится лишь конечное число членов (моменты прихода соответствующих заявок находятся левее точки начала интервала усреднения) и эта сумма не зависит от величины m, а числитель последнего члена ограничен, то мы можем заключить, что оба последних члена с вероятностью 1 стремятся к нулю при $m \to \infty$, а тогда с вероятностью 1

$$W(\omega) - T(\omega)\cdot L(\omega) \;=\; \lim_{m\to\infty}\frac{1}{m}\sum_{j\le m} v(w_j+t_j-t_m) \;\ge\; 0\;. \tag{4.30}$$

Если теперь рассмотреть интервал $(t_{-m}, 0)$ и определить средние L_{-m}, W_{-m} и T_{-m} совершенно аналогично тому, как мы это проделали выше, то есть определить, например,

$$L_{-m} \;=\; \frac{1}{(-t_{-m})}\int_{t_{-m}}^{0} n_s(\omega)\; ds\;,$$

то вследствие симметрии эргодических теорем относительно времени, можно получить

$$W(\omega)-T(\omega)\cdot L(\omega) \;=\; -\lim_{m\to\infty}\frac{1}{m}\sum_{j\le -m} v(w_j+t_j-t_{-m}) \;\le\; 0 \text{ с вероятностью 1.}$$

А это вместе с (4.30) означает ни что иное как

$$W(\omega) \;=\; T(\omega)\cdot L(\omega)\;, \quad \text{с вероятностью 1,}$$

что и требовалось установить. Тем самым, Лемма доказана полностью.

Как уже неоднократно указывалось выше, из условия метрической транзитивности процесса e_i следует, что $T(\omega)=T$, и поэтому утверждение (4.27) Леммы 4.1 переходит автоматически в

$$W(\omega) \;=\; T\cdot L(\omega)\;, \quad \text{с вероятностью 1.} \tag{4.31}$$

Для доказательства формулы Литтла нам осталось обозначить $L = \mathbf{E}\{n_0\}$, $W=\mathbf{E}\{w_0\}$ и указать, что из эргодических теорем также следует, что для почти всех ω пределы (4.26) представляют собой условные математические ожидания, например

$$L(\omega)=\mathbf{E}\{n_0|\mathcal{F}_a\},\; W(\omega)=\mathbf{E}\{w_0|\mathcal{F}_b\},$$

где $\mathcal{F}_a$ и $\mathcal{F}_b$ - борелевские поля инвариантных подмножеств соответствующих процессов.

Усреднение равенства (4.31) по всему пространству Ω дает (по определению условных вероятностей) формулу Литтла в ее окончательном виде (4.20).

4.3 Метод условно-пуассоновского потока

При исследовании входящих потоков систем массового обслуживаниянами нами были, в частности для пуассоновского потока, доказаны следующие факты:

(**1**) вероятность того, что ровно k заявок поступит на временном отрезке $[0,x]$, равна $\frac{(\lambda x)^k}{k!}e^{-\lambda x},\; k\geqslant 0$;

(**2**) при условии, что ровно $k,\; k\geqslant 1$ заявок поступает на временном отрезке $[0,x]$, соответствующие времена прихода заявок могут быть рассмотрены

как упорядоченная выборка из k независимых с.в., равномерно распределенных на $[0, x]$.

В некоторых случаях, когда приходящие в систему заявки обслуживаются независимо друг от друга, указанные выше свойства позволяют эффективно получать необходимые "глобальные"характеристики систем с очередью, начиная рассмотрение с единичной заявки, и учитывая свойства (**1**) и (**2**). Такой метод исследования систем называют *методом условного пуассоновского потока*. Рассмотрим этот метод на примере исследования предстационарного и стационарного поведения системы $M_\lambda|GI|\infty$.

4.3.1 Распределение числа обслуживаемых заявок в системе $M_\lambda|GI|\infty$

Итак, заявки, поступающие на вход системы, образуют пуассоновский поток с интенсивностью λ. Система содержит бесконечное число обслуживающих приборов (серверов), причем каждая заявка обслуживается только на одном из этих приборов. Обслуживание заявки начинается сразу же по ее поступлению, т.к. в такой системе всегда найдется не занятый сервер.

Длительность обслуживания любой заявки на любом сервере представляет собой с.в., имеющую произвольную ф.р., например, $B(\cdot)$, не зависящую от длительности обслуживания других заявок из их потока. Поэтому имеет смысл применить метод условного пуассоновского потока.

Обозначим число заявок, находящихся на обслуживании в системе в некоторый момент времени t, (число занятых серверов) через $Q(t)$. При этом будем предполагать, что

$$Q(0) \;=\; 0\;. \tag{4.32}$$

Будем искать распределение с.в. $Q(t),\ t>0$, а именно

$$q_j(t) \;=\; \mathbf{P}\big(Q(t)=j\big),\ j\geqslant 0\;. \tag{4.33}$$

Пусть $\nu(t)$ – число заявок, поступивших в систему за период времени $[0,t]$. Поскольку входной поток пуассоновский, имеем

$$\mathbf{P}(\nu(t)=k) \;=\; \frac{(\lambda t)^k}{k!}e^{-\lambda t}\;. \tag{4.34}$$

Учитывая свойство (**2**), сначала найдем вероятность $\wp$ того, что некоторая отдельно взятая заявка, поступившая в систему в момент u $(0\leqslant u\leqslant t)$, не закончит процесс обслуживания к моменту t. Для этого применим интегральный аналог формулы полной вероятности.

Итак, единичная заявка поступает в систему в промежутке времени $[u, u+du],\ u<t$ с вероятностью du/t (как равномерно распределенная на $[0,t]$ случайная величина) и за оставшееся (до момента t) время $(t-u)$

её обслуживание не будет завершено с вероятностью $[1 - B(t-u)]$. Следовательно,

$$\wp \;=\; \int_0^t [1 - B(t-u)]\frac{du}{t} \;=\; \frac{1}{t}\int_0^t [1 - B(v)]\; dv \;, \tag{4.35}$$

где последнее равенство было получено при замене переменной интегрирования u на $v = t - u$.

Далее предположим, что из k заявок, поступивших в систему на временном интервале $[0, t]$, ровно j штук $(j \leqslant k)$ всё ещё находятся в системе вплоть до момента t, а остальные $(k - j)$ штук уже покинули её по окончании обслуживания. Тогда условная вероятность такого события, очевидно, будет равна

$$\mathbf{P}(Q(t) = j \mid \nu(t) = k) \;=\; C_k^j\; \wp^j\; (1-\wp)^{k-j},\; k \geqslant j \;. \tag{4.36}$$

Очевидно также, что

$$\mathbf{P}(Q(t) = j \mid \nu(t) = k) \;=\; 0 \;,\; \text{при } k < j \;, \tag{4.37}$$

так как согласно (4.32) количество заявок, всё ещё находящихся на обслуживании в момент t, не может превышать общее число заявок, поступивших в систему вплоть до этого момента.

С учетом всего вышесказанного для вероятности $q_j(t)$ получим

$$\begin{aligned}
q_j(t) \;&=\; \sum_{k=0}^{\infty} \mathbf{P}\big(Q(t) = j,\; \nu(t) = k\big) \\
&=\; \sum_{k=j}^{\infty} \mathbf{P}(Q(t) = j \mid \nu(t) = k)\; \mathbf{P}(\nu(t) = k) \\
&=\; \sum_{k=j}^{\infty} C_k^j\; \wp^j\; (1-\wp)^{k-j}\, \frac{(\lambda t)^k}{k!}\; e^{-\lambda t} \\
&=\; \frac{(\lambda t \wp)^j}{j!}\; e^{-\lambda t}\; \sum_{k=j}^{\infty} \frac{(\lambda t(1-\wp))^{k-j}}{(k-j)!} \\
&=\; \frac{(\lambda t \wp)^j}{j!}\; e^{-\lambda t}\; e^{\lambda t(1-\wp)} \;=\; \frac{(\Lambda(t))^j}{j!}\; e^{-\Lambda(t)} \;,
\end{aligned} \tag{4.38}$$

где

$$\Lambda(t) \;=\; \lambda t \wp \;=\; \lambda \int_0^t (1 - B(u))\; du \;. \tag{4.39}$$

Видим, что для каждого t распределение с.в. $Q(t),\; t > 0$ является пуассоновским с параметром $\Lambda(t)$.

Исследуем полученный результат.

Если у распределения $B(\cdot)$ существует первый момент

$$b_1 \;=\; \int_0^\infty u\; dB(u) \;=\; \int_0^\infty (1-B(u))\; du\;, \tag{4.40}$$

то при $t \to \infty$ из (4.39) получим $\Lambda(t) \to \Lambda = \lambda b_1$, то есть предельное распределение с.в. $Q(\infty)$ тоже оказывается пуассоновским с параметром Λ.

Однако, если $b_1 = \infty$, то в таком случае для каждого j получим, что при $t \to \infty$ вероятности $q_j(t) \to 0$. А это означает, что для любого $N > 0$

$$\lim_{t\to\infty} \mathbf{P}\big(Q(t)\geqslant N\big) \;=\; 1\;. \tag{4.41}$$

Последнее означает, что в указанных условиях количество заявок, занимающих систему обслуживания, будет стремится к бесконечности, не смотря на то, что в этой системе имеется бесконечное число обрабатывающих серверов.

4.3.2 Выходной поток системы $M_\lambda|GI|\infty$

Используя рассуждения, аналогичные использовавшимся в предыдущем разделе, рассмотрим теперь выходной поток из этой системы обслуживания.

Обозначим через $k(t,t+x)$ – количество заявок, покидающих систему (по окончании обслуживания) в течение временного интервала $[t,t+x]$, и найдем распределение этой случайной величины

$$r_j(t,t+x) \;=\; \mathbf{P}\big(k(t,t+x)=j\big),\; j\geqslant 0\;. \tag{4.42}$$

Очевидно, что для каждой отдельно взятой заявки, покидающей систему после окончания обслуживания на интервале $[t,t+x]$ должны с необходимостью выполняться следующие условия: момент u ее прихода в систему может быть любым из интервала $0\leqslant u < t+x$, а длительность ее обслуживания s удовлетворять условию $(t-u)_+ \leqslant s \leqslant (t+x-u)$, где использовано объявленое во введении обозначение $(t-u)_+ = \max(0,t-u)$.
Следовательно, вероятность d того, что одиночная заявка покинет систему в промежутке времени $[t,t+x]$ с учетом свойств (**1**) и (**2**) будет равна

$$\begin{aligned} d \;&=\; \int_0^{t+x} \Big(B(t+x-u) \;-\; B(t-u)\Big)\frac{du}{t+x} \\ &=\; \frac{1}{t+x}\Big(\int_0^{t+x} B(t+x-u)\; du \;-\; \int_0^{t} B(t-u)\; du\Big) \\ &=\; \frac{1}{t+x}\Big(\int_0^{t+x} B(u)\; du \;-\; \int_0^{t} B(u)\; du\Big) \\ &=\; \frac{1}{t+x}\int_t^{t+x} B(u)\; du\;. \end{aligned} \tag{4.43}$$

Таким образом,

$$\mathbf{P}\Big(k(t,t+x)=j \mid \nu(t+x)=k\Big) \;=\; C_k^j\; d^j(1-d)^{k-j},\; k\geqslant j\;. \tag{4.44}$$

Повторяя преобразования, совершенно аналогичные проделанным выше при выводе формулы (4.38), получим для искомого распределения следующее выражение

$$r_j(t,t+x) \;=\; \frac{\big(\Lambda(t,t+x)\big)^j}{j!}\; e^{-\Lambda(t,t+x)}\;, \tag{4.45}$$

где обозначено

$$\Lambda(t,t+x) \;=\; \lambda\int_t^{t+x} B(u)\;du\;. \tag{4.46}$$

Ясно, что при $t\to\infty$, для любой выбранной величины $x>0$

$$\lim_{t\to\infty}\Lambda(t,t+x) \;=\; \lambda x\;, \tag{4.47}$$

так как $B(u)\;\to\;1$ при $u\to\infty$.
Заметим, что предел в (4.47) существует независимо от существования первого вомента b_1 у ф.р. $B(\cdot)$.
Оказывается также, что с.в. $k(t,t+x)$ и с.в. $k(t+x,t+x+y)$ при $t\to\infty$ стремятся быть независимыми, то есть

$$\begin{aligned}\lim_{t\to\infty}\;\; & \mathbf{P}\;\;\Big(k(t,t+x)=j,\;k(t+x,t+x+y)=l\Big)\\ &= \lim_{t\to\infty}\Big[\mathbf{P}\Big(k(t,t+x)=j\Big)\;\mathbf{P}\Big(k(t+x,t+x+y)=l\Big)\Big]\\ &= \frac{(\lambda x)^j}{j!}\;e^{-\lambda x}\;\frac{(\lambda y)^l}{l!}\;e^{-\lambda y}\;.\end{aligned} \tag{4.48}$$

И это заканчивает доказательство того факта, что выходной поток такой системы при $t\to\infty$ стремится к пуассоновскому потоку.

4.4 Метод построения точек восстановления

Рассмотрим еще один метод исследования систем МО, который тоже может быть отнесен к элементарным методам ТМО. С его помощью мы найдем средние значения длительности периодов занятости и простоя и среднее число заявок, обслуживаемых за период занятости, для более сложных моделей систем $M_\lambda|GI|1|\infty$ и $GI|M_\mu|1|\infty$.

4.4.1 Длительность периода занятости системы $M_\lambda|GI|1|\infty$

Пусть $\{\tau_i,\ i \geq 1\}$ – последовательность периодов занятости системы, а $\{\vartheta_i,\ i \geq 1\}$ – последовательность периодов простоя, так что период ϑ_1 следует за τ_1, а ϑ_2 – за τ_2 и т.д., образуя последовательность $\{(\tau_i, \vartheta_i),\ i \geq 1\}$ – повторяющихся "циклов" работы системы.

Будем предполагать, что ф.р. времен обслуживания заявок $B(x)$ удовлетворяет следующему условию

$$B(0) \neq 1 \tag{4.49}$$

В противном случае, все времена обслуживания были бы нулевыми, а значит, и все периоды занятости оказались бы равными нулю с вероятностью единица.

Так как входящий поток рассматриваемой системы – пуассоновский, то каждый период простоя ϑ_i, начинающийся с момента окончания соответствующего периода занятости, и длящийся до момента прихода в систему очередной заявки, открывающей следующий период занятости, представляет собой ни что иное как остаточное время ожидания прихода очередной заявки в пуассоновском потоке. Поэтому, аналогично тому, как мы уже выясняли ранее для величины эксцесса γ_t пуассоновского потока (см. (3.58)), здесь имеем

$$\mathbf{P}(\vartheta_i \leq x) = 1 - \exp(-\lambda x) \tag{4.50}$$

Причем с.в. $\vartheta_i,\ i \geq 1$ – не зависимы между собой и не зависят от с.в. $\tau_i,\ i \geq 1$.

В свою очередь, с.в. τ_i зависит от времен прихода и времен обслуживания лишь тех заявок, которые обслуживаются в течение именно этого периода занятости, но не зависит от времен прихода и от длительности обслуживания ни предыдущих, ни последующих заявок. Таким образом, последовательность с.в. $\{\tau_i,\ i \geq 1\}$ – также является последовательностью независимых одинаково распределенных с.в., как и образующие ее независимые между собой последовательности н.о.р.с.в. $\{e_i\}$ и $\{s_i\}$ (интервалов между приходами заявок и времен их обслуживания соответственно).

Обозначим ф.р. такого периода занятости τ

$$\Pi(x) = \mathbf{P}(\tau \leq x), \tag{4.51}$$

и будем искать параметры этого распределения, используя преобразование L-St. $\pi(s)$ от ф.р. $\Pi(x)$ (см. (2.60))

$$\pi(s) = \mathbf{E}\{\exp(-s\ \tau)\}. \tag{4.52}$$

Пусть s_0 – с.в., равная времени обслуживания заявки, начинающей период занятости τ. Ясно, что $\tau \geq s_0$. Причем $\tau = s_0$ только в том случае, когда за это время обслуживания s_0 в систему не пришло ни одной новой заявки. Если же за это время обслуживания s_0 в систему пришла только

одна заявка, то, очевидно, $\tau = s_0 + \tau(1)$, где $\tau(1)$ – интервал времени, начинающийся в момент начала обслуживания этой новой прибывшей заявки и продолжающийся до того времени, пока система не перейдет к периоду простоя. Но этот период $\tau(1)$, таким образом, определяется точно так же как мы определили выше с.в. τ. Следовательно, $\tau(1)$ имеет ту же самую функцию распределения (4.51), что и τ, и, кроме того, с.в. $\tau(1)$ не зависит от s_0.

Начальную точку временного интервала $\tau(1)$ называют *точкой восстановления.*

Предположим теперь, что за время обслуживания s_0 в систему пришло $k \geq 1$ заявок. Ясно, что и в этом случае мы можем представить

$$\tau = s_0 + \tau(1) + \ldots + \tau(k), \tag{4.53}$$

причем все случайные слагаемые справа являются независимыми с.в., а все $\tau(i),\ 1 \leq i \leq k$, имеют ту же ф.р., что и τ. Для лучшего понимания этого факта заметим, что если в течение всего периода занятости τ было обслужено $m \geq k$ заявок, то длительность этого периода занятости системы равна сумме времен обслуживания всех этих m заявок и никоим образом не зависит от действующей в системе дисциплины обслуживания (например, FIFO или LIFO). Но тогда легко можно представить себе следующий порядок обслуживания: по окончании периода обслуживания s_0 начинается обслуживание $k-$ той заявки (т.е. последней из пришедших), открывающей период $\tau(k)$, который следует за s_0 и не зависит от s_0. За этот период времени кроме этой заявки должны будут быть обслужены и все новые заявки, пришедшие в систему после прихода $k-$ той заявки (если таковые появятся). Затем, наконец, приходит очередь на обслуживание $(k-1)-$ ой заявки, т.е. начинается период $\tau(k-1)$, не зависящий от обеих с.в. s_0 и $\tau(k)$, и длящийся до момента начала обслуживания $(k-2)-$ ой заявки и так далее. Окончательно, мы приходим к представлению (4.53).

Начальную точку каждого такого интервала $\tau(k),\ k \geqslant 1$ называют *точкой восстановления (restoration point)*, поскольку все с.в. $\tau(k),\ k \geqslant 1$ имеют ту же, что и с.в. τ, ф.р. $\Pi(x)$, и также не зависят от s_0.

Таким образом, если предположить, что $N(t)-$ число заявок, приходящих в систему из пуассоновского потока за интервал времени $[0, t]$, то

$$\tau = s_0 + \tau(1) + \ldots + \tau(N(s_0)), \tag{4.54}$$

где $N(s_0)-$ с.в., равная числу заявок, пришедших за $[0, s_0]$.

Еще раз напомним, что все эти $\tau(i),\ i \geq 1-$ независимые о.р.с.в. не зависящие от s_0.

Поскольку число членов в правой части (4.54) само по себе также является случайной величиной, то ясно, что поиск среднего значения такого периода является далеко не простой задачей.

Для решения этой задачи воспользуемся методом преобразования L-St. Подставляя (4.54) в (4.52), используя формулу полной вероятности и неза-

висимость с.в. $\tau(i)$ от s_0 и от $N(s_0)$, получим

$$\begin{aligned}
\pi(s) &= \mathbf{E}\exp\Big(-s\big(s_0+\tau(1)+\ldots+\tau(N(s_0))\big)\Big)\\
&= \int_0^\infty \mathbf{E}\Big\{\exp\Big(-s\big(x+\tau(1)+\ldots+\tau(N(x))\big)\Big)\Big|s_0=x\Big\}d\,B(x)\\
&= \int_0^\infty e^{-sx}\sum_{k=0}^\infty \mathbf{E}\Big\{\exp\Big(-s\big(\tau(1)+\ldots\\
&+ \tau(k)\big)\Big)\Big|s_0=x, N(x)=k\Big\}\frac{(\lambda x)^k}{k!}\,e^{-\lambda x}d\,B(x)\\
&= \int_0^\infty e^{-sx}\sum_{k=0}^\infty \mathbf{E}\Big\{\exp\Big(-s\big(\tau(1)+\ldots+\tau(k)\big)\Big)\Big\}\frac{(\lambda x)^k}{k!}\,e^{-\lambda x}d\,B(x)\\
&= \int_0^\infty e^{-sx-\lambda x}\sum_{k=0}^\infty \pi^k(s)\frac{(\lambda x)^k}{k!}\,d\,B(x)\\
&= \int_0^\infty e^{-\big(s+\lambda-\lambda\pi(s)\big)x}d\,B(x)\\
&= b\big(s+\lambda-\lambda\pi(s)\big),
\end{aligned}$$

где $b(\,\cdot\,)-$ преобразование L-St. от ф.р. $B(x)$.

Следовательно, $\pi(s)$ удовлетворяет следующему функциональному уравнению:

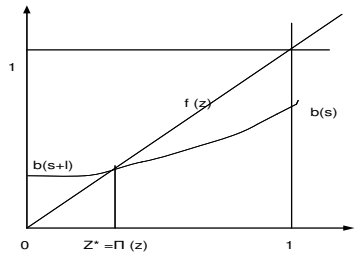
$$\pi(s) \;=\; b\big(s+\lambda-\lambda\pi(s)\big)\;. \tag{4.55}$$

Далее мы будем интересоваться только теми решениями (4.55), которые имеют вероятностный смысл. Заметим, в частности, что если $\mathcal{R}e\ s>0$, то $|\pi(s)|<1$, т.к. $\pi(s)-$ преобразование L-St. от ф.р. $\Pi(x)$ (см. (4.52)).
При исследовании полученного уравнения (4.55) положим $\pi(s)=z$, перепишем его в виде

$$z \;=\; b\big(s+\lambda-\lambda z\big)\;, \tag{4.56}$$

и зададим себе следующий вопрос: "Сколько решений имеет уравнение (4.56) на множестве $|z|<1$ при условии, что $\mathcal{R}e\ s>0$?"

Для ответа на этот вопрос рассмотрим следующие две функции комплексного переменного z: $f(z)=z$ и $g(z)=b\big(s+\lambda-\lambda z\big)$.
Ясно, что они обе являются аналитическими при $|z|\leqslant 1$ и $\mathcal{R}e\ s>0$, и, кроме того, на границе $|z|=1$ у одной из них $|f(z)|\equiv|z|=1$, а у другой $|g(z)|<1$, то есть на границе $|z|=1$ выполняется условие $|g(z)|<|f(z)|$ теоремы Руше, известной в курсе ТФКП. А из этой теоремы следует, что в таком случае функции $f(z)$ и $(f(z)-g(z))$ должны иметь одинаковое число нулей. Поскольку функция $f(z)$ имеет на множестве $|z|\leqslant 1$ лишь единственный нуль (равный нулю!), отсюда вытекает, что наше уравнение (4.56) имеет единственное решение на множестве $|z|<1$, $\mathcal{R}e\ s>0$, а, следовательно, это



решение с необходимостью совпадает с $\pi(s)-$ преобразованием L-St. от ф.р. $\Pi(x)$.

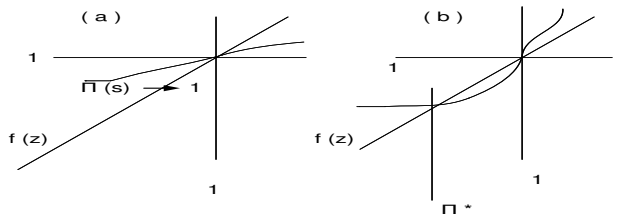
Для изучения свойств этого решения снова рассмотрим уравнение (4.56) на $0 \leqslant z \leqslant 1$ и обозначим $z^* = \pi(s)$ его единственный корень. Как видно из рисунка, этот корень – есть точка пересечения прямой $f(z) = z$ и выпуклой вниз возрастающей функции $g(z)$. Утверждение о характере поведения функции $g(z)$ легко получается из следующих формул

$$
\begin{aligned}
g(z) &= b(s+\lambda-\lambda z) = \int_0^\infty e^{-x(s+\lambda-\lambda z)}\, dB(x) \\
g(0) &= b(s+\lambda) < b(s) = g(1), \quad b(s) \to 1, \text{ при } s \to 0 \\
g'(z) &= \lambda \int_0^\infty x\, e^{-x(s+\lambda-\lambda z)}\, dB(x) \geqslant 0 \\
g''(z) &= \lambda^2 \int_0^\infty x^2\, e^{-x(s+\lambda-\lambda z)}\, dB(x) \geqslant 0 \\
g'(1) &= \lambda \int_0^\infty x\, e^{-xs}\, dB(x), \quad g'(1)\big|_{s=0} = \lambda \int_0^\infty x\, dB(x) = \lambda b_1
\end{aligned}
$$

Ясно, что при $s \to 0$ кривая $g(z)$ будет сдвигаться выше и, следовательно, корень $z^* = \pi(s)$ будет сдвигаться вправо и вверх, то есть будет расти.

При этом, очевидно, возможны два случая:

(**а**) если $g'(1)\big|_{s=0} = \lambda b_1 \leqslant 1$, то $\pi(s) \to 1$ при $s \to 0$ и тогда предельная кривая $g(z)\big|_{s=0}$ окажется целиком выше $f(z)$;



(**b**) если $g^{'}(1)\big|_{s=0} = \lambda b_1 \ > \ 1$, то при $s \to 0 \ \ \pi(s) \to \pi^* < 1$, а это означает, что соответствующая ф.р. $\Pi(x)$ при $x \to \infty$ имеет тот же предел $\pi^* < 1$, то есть с вероятностью $(1 - \pi^*)$ возможен факт существования бесконечно долгого периода занятости рассматриваемой системы обслуживания. Этот результат вполне естественен, так как соотношение $\lambda b_1 > 1$ означает, что более одной заявки поступает (в среднем) за среднее время обслуживания одной заявки. Отметим также, что введенная нами величина π^* является решением следующего уравнения

$$\pi^* \ = \ b\big(\lambda - \lambda\pi^*\big) \ , \tag{4.57}$$

в которое переходит (при $s \to 0$) уравнение (4.55).

При $\lambda b_1 \leqslant 1$ период занятости с вероятностью 1 имеет конечную длительность и, следовательно, имеет смысл поставить вопрос о нахождении моментов распределения $\Pi(x)$. В качестве примера найдем первый момент π_1, то есть среднее значение длительности периода занятости. Для этого продифференцируем (4.55) по правилам дифференцирования неявной функции:

$$\pi^{'}(s) \ = \ b^{'}(s + \lambda - \lambda\pi(s))(1 - \lambda\pi(s)). \tag{4.58}$$

Устремляя в обеих частях этого выражения $s \to 0$ и вспоминая, что пределы $b^{'}(0) = -b_1, \ \ \pi^{'}(0) = -\pi_1$, получим для случая $\lambda b_1 < 1$

$$\pi_1 \ = \ \frac{b_1}{1 - \lambda b_1}. \tag{4.59}$$

А для $\lambda b_1 = 1$ или $b_1 = \infty$ (тогда $b^{'}(s) \to -\infty$ при $s \to 0$) из (4.58) следует, что функция $\pi^{'}(s)$ не может иметь конечный предел при $s \to 0$. В обоих этих случаях средняя величина длительности периода занятости системы оказывается равной ∞.

Аналогичный прием (с дальнейшим дифференцированием неявной функции) может быть применен и для вычисления моментов $\Pi(x)$ более высокого порядка. **Задача нахождения второго момента этого распределения предлагается в качестве домашнего упражнения.**

4.4.2 Среднее число заявок, обслуживаемых за период занятости в $M_\lambda|GI|1|\infty$

Пусть в течение некоторого периода занятости τ система обслуживает ровно κ заявок. Ясно, что $\kappa \geqslant 1$ (иначе в этот период система находилась бы в простое). Нас будут интересовать параметры распределения $\mathbf{P}(\kappa = j),\ j \geq 1$ этой случайной величины.

Из рассуждений, совершенно аналогичных приведенным в предыдущем разделе и приведшим к (4.54), в данном случае можно заключить, что

$$\kappa \ = \ 1 \ + \ \kappa(1) \ + \ \ldots \ + \ \kappa(N(s_0)), \tag{4.60}$$

где $\kappa(i)-$ число заявок, обслуженных системой за соответствующий период времени $\tau(i)$. При этом последовательность $\{\kappa(i),\ i \geqslant 1\}$ представляет собой последовательность независимых, одинаково распределенных с.в., не зависящих от s_0 и распределенных так же как с.в. κ.

Рассмотрим производящую функцию

$$K(z) \ = \ \sum_{j=1}^{\infty} z^j \ P(\kappa = j) \tag{4.61}$$

последовательности вероятностей $P(\kappa = j)$ и, используя (4.60), по формуле полной вероятности получим:

$$\begin{aligned}
K(z) \ &= \ \mathbf{E}z^{\kappa} \ = \ \mathbf{E}z^{[1+\kappa(1)+\ldots+\kappa(N(s_0))]} \\
&= \ \int_0^{\infty} \mathbf{E}\Big(z^{[1+\kappa(1)+\ldots+\kappa(N(x))]}\Big| s_0 = x\Big)\ dB(x) \\
&= \ z\int_0^{\infty} \sum_{i=0}^{\infty} \mathbf{E}\Big(z^{[\kappa(1)+\ldots+\kappa(i)]}\Big| s_0 = x, N(x) = i\Big)\frac{(\lambda x)^i}{i!}\ e^{-\lambda x}\ dB(x) \\
&= \ z\int_0^{\infty} e^{-\lambda x} \sum_{i=0}^{\infty} \frac{(K(z))^i(\lambda x)^i}{i!}\ dB(x) \ = \ z\int_0^{\infty} e^{-(\lambda-\lambda K(z))x}\ dB(x) \\
&= \ z\ b\big(\lambda - \lambda K(z)\big)\ .
\end{aligned} \tag{4.62}$$

В наших преобразованиях (при вычислении условного матожидания) мы воспользовались независимостью всех случайных величин $\kappa(i)$ от s_0.

Полученное уравнение

$$K(z) \;=\; z\, b\big(\lambda - \lambda K(z)\big) \tag{4.63}$$

исследуется аналогично уравнению (4.55), а при $z \to 1$ оно переходит в

$$K^* \;=\; b\big(\lambda - \lambda K^*\big) \tag{4.64}$$

которое совпадает с (4.57).
Поэтому при $\lambda b_1 > 1$ существует $\lim_{z\to 1} K(z) = K^* < 1$, а это означает, что с вероятностью $(1 - K^*) = (1 - \pi^*)$ количество заявок, обслуживаемых за период занятости рассматриваемой системы может быть бесконечно большим.
Если же $\lambda b_1 \leqslant 1$, то $K(1) = \mathbf{P}(\kappa < \infty) = 1$ и можно найти моменты исследуемого распределения, для чего следует продифференцировать (4.63) по z по правилам дифференцирования неявных функций. Для первого момента по свойствам производящих функций $\mathbf{E}\kappa = K^{'}(1)$. Кроме того, как известно, $b(0) = 1, \;\; b^{'}(0) = -b_1$. Тогда для $\lambda b_1 < 1$ существует

$$\mathbf{E}\kappa \;=\; \frac{1}{1 - \lambda b_1}. \tag{4.65}$$

Любые моменты более высокого порядка для этого распределения могут быть вычислены аналогичным образом.

4.4.3 Периоды занятости и простоя в системе $GI|M_\mu|1|\infty$

Для этой системы последовательность временных интервалов $\{e_i\}_{i\geqslant 0}$ между моментами прихода заявок так же, как и в системе, только что рассмотренной нами выше, представляет собой последовательность н.о.р.с.в., имеющих одинаковую ф.р., но которая теперь уже не экспоненциальная, а просто ф.р. общего вида – $A(x)$. При этом последовательность н.о.р.с.в. $\{s_i\}_{i\geqslant 1}$, представляющая собой последовательность времен обслуживания поступивших заявок, напротив, имеет теперь экспоненциальную функцию распределения, то есть

$$B(x) \;=\; 1 - e^{-\mu x}. \tag{4.66}$$

В связи с этим, мы далее теперь будем использовать уже известный нам факт, что остаточное время обслуживания заявки имеет то же самое экспоненциальное распределение, что и само время обслуживания.

Снова будем обозначать $\{\tau_i\}_{i\geqslant 1}$ и $\{\vartheta_i\}_{i\geqslant 1}$ – последовательности чередующихся между собой периодов занятости и простоя системы.

Можно заметить, что последовательность $\{(\tau_i + \vartheta_i), \quad i\geqslant 1\}$ состоит из независимых, одинаково распределенных случайных величин, и, более того, пара случайных величин $(\tau_i,\ \vartheta_i)$ при $i\geqslant 1$ – образует последовательность независимых пар с.в. (причем этот факт остается справедливым даже для системы $GI|GI|1|\infty$!).

Будем далее называть эти пары *циклами.*

Но случайные величины τ_i и ϑ_i, составляющие i-тую пару, в общем случае могут быть зависимыми между собой. Поэтому для полного описания последовательности циклов $\{(\tau_i,\ \vartheta_i),\quad i\geqslant 1\}$ нам потребуется рассмотреть совместное распределение пары $(\tau,\ \vartheta)$:

$$\Pi(x,y)\ =\ \mathbf{P}(\tau\leqslant x,\ \vartheta\leqslant y). \tag{4.67}$$

Если теперь ввести в рассмотрение новую пару с.в. $(\tau(i),\ \vartheta(i))$, которая определяется точно так же как и пара $(\tau,\ \vartheta)$, но при условии, что заявка, открывающая своим приходом период занятости $\tau(i)$, застает ровно i заявок, уже находящихся в системе, то тогда для рассматриваемой нами системы каждый момент прихода очередной заявки будет являться *точкой восстановления.*

Будем называть такую пару $(\tau(i),\ \vartheta(i))$ *i-циклом.*

Обозначим совместную ф.р. пары с.в. $(\tau(i),\ \vartheta(i))$

$$\Pi_i(x,y)\ =\ \mathbf{P}(\tau(i)\leqslant x,\ \vartheta(i)\leqslant y),\quad i\geqslant 0 \tag{4.68}$$

и определим следующее событие

$$S_k(t)\ =\ \big\{\text{точно } k \text{ заявок было обслужено в течение } [0,t]\big\} \tag{4.69}$$

Таким образом, обычный, интересующий нас цикл работы системы $(\tau,\ \vartheta)$, будет в этих обозначениях 0-циклом, а его ф.р. $\Pi(x,y)=\Pi_0(x,y)$.

В соответствии с введенными обозначениями работу рассматриваемой нами системы можно теперь описать следующим образом. Если некоторая заявка, приходящая в систему, например, в момент времени T_m, застает в системе ровно i заявок, то она открывает цикл $(\tau(i),\ \vartheta(i))$. Тогда следующая за ней очередная заявка, приходящая в систему в момент $T_{m+1}=T_m+e_m$, либо закрывает цикл, если оказывается $e_m>\sum_{j=1}^{i+1}s_j=\tau(i)$ и тогда $\vartheta(i)=e_m-\tau(i)$ (то есть при этом происходит событие $S_{i+1}(\tau(i))$), либо открывает $(i+1-k)$-тый цикл, если происходит одно из событий $S_k(e_m)$, где $0\leqslant k\leqslant i$.

Тогда по формуле полной вероятности получим

$$\begin{aligned}\Pi_i(x,y)\ &=\ \mathbf{P}\Big(\tau(i)\leqslant x,\ \vartheta(i)\leqslant y\Big)=\mathbf{P}\Big(\tau(i)\leqslant x,\ \vartheta(i)\leqslant y,\ S_{i+1}(\tau(i))\Big)\\ &+\ \sum_{k=0}^{i}\mathbf{P}\Big(\tau(i)\leqslant x,\ \vartheta(i)\leqslant y,\ S_k(e_m)\Big),\quad i\geqslant 0.\ .\end{aligned} \tag{4.70}$$

Очевидно, что

$$\begin{aligned}
&\mathbf{P} \quad \Big(\tau(i)\leqslant x,\ \vartheta(i)\leqslant y,\ S_{i+1}(\tau(i))\Big) \\
&= \mathbf{P}\Big(0\leqslant\tau(i)\leqslant x,\ 0\leqslant e_m-\tau(i)\leqslant y,\ \tau(i)=\sum_{j=1}^{i+1}s_j\Big) \\
&= \mathbf{P}\Big(0\leqslant\sum_{j=1}^{i+1}s_j\leqslant x,\ \sum_{j=1}^{i+1}s_j\leqslant e_m\leqslant y+\sum_{j=1}^{i+1}s_j,\Big) \\
&= \int_0^x \mathbf{P}\Big(u\leqslant e_m\leqslant y+u\Big)\mathbf{P}\Big(u\leqslant\sum_{j=1}^{i+1}s_j\leqslant u+du\Big) \\
&= \int_0^x \Big(A(y+u)-A(u)\Big)\,dB_*^{i+1}(u),\ , \qquad (4.71)
\end{aligned}$$

где

$$B_*^{i+1}(u) \;=\; 1 \;-\; e^{-\mu u}\sum_{j=0}^{i}\frac{(\mu u)^j}{j!}$$

– есть распределение Эрланга порядка $(i+1)$ и, как мы уже показывали ранее,

$$dB_*^{i+1}(u) \;=\; \mu\frac{(\mu u)^i}{i!}\,e^{-\mu u}\,du\ . \qquad (4.72)$$

В свою очередь, нетрудно подсчитать, что

$$\begin{aligned}
&\mathbf{P} \quad \Big(\tau(i)\leqslant x,\ \vartheta(i)\leqslant y,\ S_k(e_m)\Big) \\
&= \mathbf{P}\Big(e_m+\tau(i+1-k)\leqslant x,\ \vartheta(i+1-k)\leqslant y,\ S_k(e_m)\Big) \\
&= \int_0^x \mathbf{P}\big(S_k(u)\big)\ \mathbf{P}\Big(\tau(i+1-k)\leqslant x-u,\ \vartheta(i+1-k)\leqslant y\Big)\ dP(e_m=u) \\
&= \int_0^x \mathbf{P}\big(S_k(u)\big)\ \Pi_{i+1-k}\big(x-u,y\big)\ dA(u), \qquad (4.73)
\end{aligned}$$

где следует учесть, что

$$\mathbf{P}\big(S_k(u)\big) \;=\; \frac{(\mu u)^k}{k!}\,e^{-\mu u} \qquad (4.74)$$

– есть распределение Пуассона.
Подставляя в (4.70) выражения (4.71), (4.72), (4.73) и (4.74), получим для $i\geqslant 0$

$$\begin{aligned}
\Pi_i(x,y) \;&=\; \mu\int_0^x \frac{(\mu u)^i}{i!}\,e^{-\mu u}\Big(A(y+u)-A(u)\Big)\,du \\
&+\; \sum_{k=0}^{i}\int_0^x \frac{(\mu u)^k}{k!}\,e^{-\mu u}\,\Pi_{i+1-k}(x-u,y)\,dA(u). \qquad (4.75)
\end{aligned}$$

Обозначим

$$\pi_i(s,y) \;=\; \int_0^\infty e^{-sx}\; d_x\Pi_i(x,y),\;\; i\geqslant 0 \tag{4.76}$$

– преобразование L-St. (по x) от ф.р. $\Pi_i(x,y)$, и далее

$$\pi(z,s,y) \;=\; \sum_{i=0}^{\infty} z^i\;\pi_i(s,y) \tag{4.77}$$

– производящую функцию последовательности $\pi_i(s,y), i\geqslant 0$.
Заметим также, что входящее под знаком суммы в (4.75) выражение

$$\int_0^x \frac{(\mu u)^k}{k!}\; e^{-\mu u}\;\Pi_{i+1-k}(x-u,y)\; dA(u)$$

– есть ни что иное как свертка (по Стилтьесу) функций $\Pi_{i+1-k}(\cdot,y)$ и

$$F(x) \;=\; \int_0^x \frac{(\mu u)^k}{k!}\; e^{-\mu u}\; dA(u).$$

Производя с учетом этого преобразование (4.76) в обеих частях (4.75), получим для $i\geqslant 0$

$$\begin{aligned}\pi_i(s,y) \;&=\; \mu\int_0^\infty \frac{(\mu u)^i}{i!}\; e^{-(s+\mu)u}\Big(A(y+u)-A(u)\Big)\; du\\ &+\; \sum_{k=0}^{i}\pi_{i+1-k}(s,y)\int_0^\infty \frac{(\mu u)^k}{k!}\; e^{-(s+\mu)u}\; dA(u).\end{aligned} \tag{4.78}$$

Чтобы перейти в (4.78) к производящей функции (4.77), рассмотрим сначала следующие равенства

$$\begin{aligned}&\sum_{i=0}^{\infty} z^i \sum_{k=0}^{i}\pi_{i+1-k}(s,y)\frac{(\mu u)^k}{k!} \;=\; \sum_{k=0}^{\infty} z^k\frac{(\mu u)^k}{k!}\sum_{i=k}^{\infty} z^{i-k}\;\pi_{i+1-k}(s,y)\\ &=\; \sum_{k=0}^{\infty} z^k\frac{(\mu u)^k}{k!}\sum_{m=0}^{\infty} z^m\;\pi_{m+1}(s,y) \;=\; e^{\mu z u}\frac{1}{z}\Big(\pi(z,s,y)\;-\;\pi_0(s,y)\Big),\end{aligned} \tag{4.79}$$

которые получены с использованием легко проверяемого правила перестановки порядка суммирования в двойных суммах

$$\sum_{i=0}^{\infty}\sum_{k=0}^{i} a_{ik} \;=\; \sum_{k=0}^{\infty}\sum_{i=k}^{\infty} a_{ik} \tag{4.80}$$

С учетом этого получим

$$\begin{aligned}\pi(z,s,y) \;&=\; \mu\int_0^\infty e^{-(s+\mu-\mu z)u}\Big(A(y+u)-A(u)\Big)\; du\\ &+\; \frac{1}{z}\; a(s+\mu-\mu z)\Big(\pi(z,s,y)-\pi_0(s,y)\Big),\end{aligned} \tag{4.81}$$

где

$$a(s) \;=\; \int_0^\infty e^{-su}\; dA(u)$$

– преобразование L-St. от ф.р. $A(\cdot)$.

Обозначим далее

$$R(z,s,y) \;=\; \mu \int_0^\infty e^{-(s+\mu-\mu z)u}\Big(A(y+u)-A(u)\Big)\; du \tag{4.82}$$

Тогда из уравнения (4.81) можно выразить

$$\pi(z,s,y) \;=\; \frac{zR(z,s,y)-a(s+\mu-\mu z)\pi_0(s,y)}{z-a(s+\mu-\mu z)} \tag{4.83}$$

Заметим, что, если $\mathcal{R}e\; s \;>\; 0$ и $|z| \;<\; 1$, то функция $\pi(z,s,y)$ должна быть ограниченной (как дважды преобразованная последовательность функций распределения). Однако ранее мы уже показывали, что у выражения, аналогичного стоящему в знаменателе (4.83), имеется единственный корень $z^* = z^*(s)$ как раз на указанном множестве $\mathcal{R}e\; s \;>\; 0,\; |z| \;<\; 1$. Все эти факты можно согласовать лишь потребовав, чтобы и числитель выражения (4.83) тоже обращался в нуль при $z = z^*$. Но тогда с необходимостью из (4.83) получим

$$\pi_0(s,y) \;=\; \frac{z^*(s)R(z^*(s),s,y)}{a(s+\mu-\mu z^*(s))} \;=\; R(z^*(s),s,y), \tag{4.84}$$

т.к. корень знаменателя (4.83) есть решение уравнения (при некотором s)

$$z^*(s) \;=\; a(s+\mu-\mu z^*(s)). \tag{4.85}$$

Рассмотрим двойное преобразование L-St, т.е.

$$\begin{aligned}\pi(s_1,s_2) \;&=\; \mathbf{E}\,\{e^{-s_1\tau-s_2\vartheta}\}\\ &=\; \int_0^\infty e^{-s_2y}\; d_y\pi_0(s_1,y)\\ &=\; \int_0^\infty e^{-s_2y}\; d_yR(z^*(s_1),s_1,y).\end{aligned} \tag{4.86}$$

Преобразуем теперь выражение, стоящее в правой части (4.86), заменяя для удобства (чтобы не писать индексы) параметры $s_1 = s,\; s_2 = \rho,\; z^*(s_1) = z$.

Считая далее $\Re e\ \rho\ >\ 0$, получим:

$$\begin{aligned}
&\int_0^\infty \ e^{-\rho y}\ d_y R(z,s,y) \\
&= \ \mu\int_0^\infty \Big(e^{-(s+\mu-\mu z)u}\int_0^\infty e^{-\rho y} d_y\big(A(y+u)-A(u)\big)\Big)\ du \\
&= \ \mu\int_0^\infty \Big(e^{-(s+\mu-\mu z)u}\int_u^\infty e^{-\rho(x-u)} dA(x)\Big)\ du \\
&= \ \mu\int_0^\infty \Big(e^{-\rho x}\int_0^x e^{-(s+\mu-\mu z-\rho)u} du\Big)\ dA(x) \\
&= \ \frac{\mu}{s+\mu-\mu z-\rho}\Big(\int_0^\infty \big(e^{-\rho x}-e^{-(s+\mu-\mu z)x}\big)\ dA(x)\Big) \\
&= \ \frac{\mu\big(a(\rho)-a(s+\mu-\mu z)\big)}{s+\mu-\mu z-\rho}\ ,
\end{aligned} \tag{4.87}$$

где была использована следующая формула смены порядка интегрирования:

$$\int_0^\infty \Big(\int_u^\infty f(x,u)\ dx\Big)\ du\ =\ \int_0^\infty \Big(\int_0^x f(x,u)\ du\Big)\ dx\ , \tag{4.88}$$

аналогичная по смыслу формуле (4.80).
Из (4.86), (4.87) и (4.85) окончательно получим функцию:

$$\pi(s_1,s_2)\ =\ \mu\frac{a(s_2)\ -\ z^*(s_1)}{s_1+\mu-\mu z^*(s_1)-s_2}\ , \tag{4.89}$$

которая представляет собой преобразование L-St от интересующей нас ф.р. (4.67)

Прежде, чем анализировать полученное нами решение, вспомним, что выражение $z^*(s_1)$ является корнем уравнения (4.85), которое совершенно аналогично уже исследовавшемуся ранее уравнению (4.56), и мы можем по аналогии выписать два следующих случая:

(**a**) если

$$\mu a_1 \leqslant 1, \tag{4.90}$$

то $z^*(s_1)\to 1$ при $s_1\to 0$;

(**b**) если

$$\mu a_1\ >\ 1, \tag{4.91}$$

то при $s_1\to 0\ \ z^*(s_1)\to z^*(0)=z_0<1$, где величина z_0 является решением уравнения

$$z_0\ =\ a\big(\mu-\mu z_0\big)\ , \tag{4.92}$$

в которое (при $s_1\to 0$) переходит (4.85).
Символом a_1 мы как всегда обозначили первый момент ф.р. $A(\cdot)$.

Напомним также, что по свойствам преобразования L-St справедлива формула нахождения моментов (2.62) и, кроме того, выполняются предельные соотношения для поведения ф.р. и преобразования L-St этой функции (см. свойство (7)). Воспользуемся этими фактами для исследования существования моментов у интересующего нас распределения (4.67), а затем и при вычислении самих этих моментов с использованием вида (4.89) полученного нами двойного преобразования L-St от этой ф.р.

Для этого сначала подсчитаем первую производную по s функции $z^*(s)$, используя уравнение (4.85) и применяя правила дифференцирования неявной функции. Получим:

$$\frac{dz^*(s)}{ds} \;=\; \frac{da(u)}{du}\Big|_{u=s+\mu-\mu z^*(s)} \Big(1 \;-\; \mu\, \frac{dz^*(s)}{ds}\Big)\;, \tag{4.93}$$

откуда

$$\frac{dz^*(s)}{ds} \;=\; \frac{a^{'}(s+\mu-\mu z^*(s))}{1 \;+\; \mu a^{'}(s+\mu-\mu z^*(s))}\;. \tag{4.94}$$

Кроме того, по свойствам преобразования L-St имеем

$$a(0) \;=\; 1, \quad a^{'}(0) \;=\; -a_1. \tag{4.95}$$

И тогда для первого случая (4.90), используя (4.89) и применяя правило Лопиталя при вычислении предела, можно подсчитать, что

$$\begin{aligned}
\mathbf{P}(\tau<\infty) \;&=\; \lim_{s\to 0}\mathbf{E}\{e^{-s\tau}\} = \lim_{s\to 0}\pi(s,0) = \lim_{s\to 0}\frac{\mu(1-z^*(s))}{s+\mu-\mu z^*(s)} \\
&=\; -\lim_{s\to 0}\mu\Big(\frac{dz^*(s)}{ds}\Big/\Big(1-\mu\frac{dz^*(s)}{ds}\Big)\Big) \\
&=\; -\mu\lim_{s\to 0}a^{'}(s+\mu-\mu z^*(s)) = \mu a_1 \;<\; 1.
\end{aligned} \tag{4.96}$$

Аналогично можно вычислить

$$\begin{aligned}
\mathbf{P}(\vartheta<\infty) \;&=\; \lim_{s\to 0}\mathbf{E}\{e^{-s\vartheta}\} = \lim_{s\to 0}\pi(0,s) = \lim_{s\to 0}\frac{\mu(a(s)-z^*(0))}{\mu-\mu z^*(0)-s} \\
&=\; \lim_{s\to 0}\mu\frac{a(s)-1}{(-s)} = \mu\lim_{s\to 0}\frac{a^{'}(s)}{-1} = \mu a_1\;.
\end{aligned} \tag{4.97}$$

А это означает, что в случае, когда $\mu a_1 \leqslant 1$, случайные величины τ - период занятости и ϑ - период простоя не имеют правильной функции распределения, поскольку

$$\text{при } s_1\to 0,\; s_2\to 0 \quad \pi(s_1,s_2)\to \mathbf{P}(\vartheta<\infty,\tau<\infty) = \pi(0,0) = \mu a_1, \tag{4.98}$$

т.е. существует не равная нулю вероятность, с которой эти величины могут принимать бесконечно большие значения. Однако, при условии конечности с.в. τ, с.в. ϑ оказывается правильной, так как

$$\mathbf{P}(\vartheta<\infty|\tau<\infty) \;=\; 1\;, \tag{4.99}$$

что вытекает из (4.98) и (4.96).

Для исследования второго случая (см. (4.91)) нетрудно показать, что

$$\pi(s_1, s_2) \ \to \ 1 \ \text{ при } s_1 \to 0, \ s_2 \to 0, \tag{4.100}$$

если $\mu a_1 \ > \ 1$. Это означает, что если система не перегружена, т.е. $\mu > 1/a_1$, то периоды занятости и простоя оказываются правильными с.в. и, следовательно, в этом случае имеет смысл ставить вопрос о нахождении моментов соответствующих ф.р.
Однако прежде, чем перейти непосредственно к вычислению этих моментов, обозначим для удобства

$$z_k \ = \ \frac{d^{(k)} z^*(s)}{ds^k}\Big|_{s=0}, \ k \geqslant 1 \tag{4.101}$$

и предварительно подсчитаем несколько таких величин дифференцируя (4.85) по правилам дифференцирования неявной функции (аналогично (4.93)). Тогда, как нетрудно проверить, получим

$$z_1 \ = \ \frac{a^{'}(\mu - \mu z_0)}{1 + \mu a^{'}(\mu - \mu z_0)} \ ; \tag{4.102}$$

$$z_2 \ = \ \frac{a^{''}(\mu - \mu z_0)}{\big(1 + \mu a^{'}(\mu - \mu z_0)\big)^3}; \tag{4.103}$$

. . . . . .

где в качестве предела $z^*(s)$ в нуле подставлен $z^*(0) = z_0$ - корень уравнения (4.92).

И, наконец, сами первые два момента

$$\mathbf{E}\tau \ = \ -\frac{d}{ds}\pi(s,0)\big|_{s=0} \ = \ \frac{1}{\mu(1-z_0)}; \tag{4.104}$$

$$\mathbf{E}\tau^2 \ = \ \frac{d^2}{ds^2}\pi(s,0)\big|_{s=0} \ = \ \frac{2(1-\mu z_1)}{\mu^2(1-z_0)^2}; \tag{4.105}$$

$$\mathbf{E}\vartheta \ = \ -\frac{d}{ds}\pi(0,s)\big|_{s=0} \ = \ \frac{\mu a_1 - 1}{\mu(1-z_0)}; \tag{4.106}$$

$$\mathbf{E}\vartheta^2 \ = \ \frac{d^2}{ds^2}\pi(0,s)\big|_{s=0} \ = \ \frac{\mu^2 a_2(1-z_0) - 2(\mu a_1 - 1)}{\mu^2(1-z_0)^2}; \tag{4.107}$$

$$\mathbf{E}(\tau\vartheta) \ = \ \frac{\partial^2 \pi(s_1,s_2)}{\partial s_1 \partial s_2}\Big|_{s_1=s_2=0} \ = \ \frac{\mu a_1 - \mu z_1(\mu a_1 - 1) - 2}{\mu^2(1-z_0)^2} \ . \tag{4.108}$$

Проверить правильность приведенных результатов рекомендуется в качестве полезного упражнения.

Если положить в формуле (4.89) $s_1 = s_2 = s$, то мы получим вид преобразования L-St функции распределения с.в. $(\tau + \vartheta)$ - длительности цикла работы системы, а именно

$$\pi(s,s) \ = \ \mathbf{E}\{e^{-s(\tau+\vartheta)}\} \ = \ \frac{a(s) - z^*(s)}{1 - z^*(s)} \ . \tag{4.109}$$

Нетрудно показать, что моменты этой случайной величины равны соответственно

$$\mathbf{E}(\tau+\vartheta) = \frac{a_1}{1-z_0}; \tag{4.110}$$

$$\mathbf{E}(\tau+\vartheta)^2 = \frac{a_2(1-z_0)-2a_1z_1}{(1-z_0)^2}\ . \tag{4.111}$$

Четко и внимательно производя дифференцирование и проявив достаточно внимания и усидчивости, можно вывести не только любую из приведенных формул, но и найти любые другие моменты указанных выше функций распределений.

4.4.4 Количество заявок, обслуживаемых в $GI|M_\mu|1|\infty$ в течение периода занятости

Пусть $K-$ число заявок, обслуженных за период занятости системы. Ясно, что $K \geq 1$. Наша цель – найти параметры распределения $\mathbf{P}(K=j),\ j \geq 1$ этой случайной величины.

Для решения задачи рассмотрим вспомогательные случайные величины $K(i),\ i \geq 0$, которые представляют собой количество заявок, обслуживаемых в течение соответствующего периода занятости системы $\tau(i)$, введенного в предыдущем разделе. Из определения $\tau(i)$ очевидно, что $K(i) \geq i+1$. Кроме того, $K(0) \stackrel{d}{=} K$.

Обозначим

$$q_i(j) = \mathbf{P}\big(K(i) = i+1+j\big),\ \ j \geq 0\ . \tag{4.112}$$

Ясно, что при $j=0$ с вероятностью $q_i(0)$ происходит событие $\big\{K(i)=i+1\big\}$, что соответствует рассмотренному в предыдущем разделе случаю, когда за промежуток времени e_n между приходом заявки, открывшей цикл $(\tau(i),\vartheta(i))$, и моментом прихода следующей за ней очередной заявки, система успевает обслужить все имеющиеся в ней заявки. Обозначая величину $\sum_{k=1}^{i+1} s_i = \tau(i) = u$ и учитывая, что $e_n > u$, по формуле полной вероятности получим

$$\begin{aligned} q_i(0) &= \int_0^\infty P(e_n > u)\ dB_*^{i+1}(u) \\ &= \int_0^\infty \big(1-A(u)\big)\ dB_*^{i+1}(u) \\ &= \int_0^\infty \mu\frac{(\mu u)^i}{i!}e^{-\mu u}\big(1-A(u)\big)\ du \end{aligned} \tag{4.113}$$

Если же обслуживание не заканчивается до прихода очередной заявки, то с момента T_{n+1} ее прихода начинается новый период занятости системы, а именно период $\tau(i+1-m)$, где число $m,\ 0 \leq m \leq i$ означает количество заявок, которые система успела обслужить за период времени e_n. Так как

при этом оказывается $\tau(i) = e_n + \tau(i+1-m)$, а за e_n было обслужено ровно m заявок, то для того, чтобы за период $\tau(i)$ было в целом обслужено $(i+1+j)$ заявок, мы должны обслужить за период $\tau(i+1-m)$ ровным счетом еще $(i+1-m+j)$ штук. Но тогда из-за независимости рассматриваемых событий на не перескающихся временных интервалах и требования одновременности их выполнения, получим

$$\mathbf{P}\big(K(i) = i+1+j\big) = \mathbf{P}\big(m \text{ штук за } e_n = u\big) \cdot \mathbf{P}\big(K(i+1-m) = i+1-m+j\big).$$

Откуда по формуле полной вероятности можно записать

$$q_i(j) \ = \sum_{m=0}^{i} \int_0^\infty \frac{(\mu u)^m}{m!}\ e^{-\mu u}\ dA(u)\ q_{i+1-m}(j-1)\ , \quad j \geq 1\ . \tag{4.114}$$

Рассмотрим далее производящую функцию последовательности $q_i(j)$:

$$\widehat{q}_i\ (z) \ = \sum_{j=0}^{\infty} z^j\ q_i(j) \tag{4.115}$$

Выражения (4.113) и (4.114) позволяют получить для (4.115) следующее представление

$$\begin{aligned} \widehat{q}_i\ (z) \ &= \ q_i(0) + \sum_{j=1}^{\infty} z^j \sum_{m=0}^{i} q_{i+1-m}(j-1) \int_0^\infty \frac{(\mu u)^m}{m!}\ e^{-\mu u}\ dA(u) \\ &= \ q_i(0) + z \sum_{m=0}^{i} \widehat{q}_{i+1-m}\ (z) \int_0^\infty \frac{(\mu u)^m}{m!}\ e^{-\mu u}\ dA(u),\ i \geqslant 0 \end{aligned} \tag{4.116}$$

где значение вероятности $q_i(0)$ определено в (4.113).

Определим теперь двойное преобразование – следующую функцию комплексных переменных

$$Q(z,w) \ = \sum_{i=0}^{\infty} w^i\ \ \widehat{q}_i\ (z), \tag{4.117}$$

которая, конечно, сходится в области $|z| \leqslant 1, \ \ |w| < 1$.

Умножая обе части равенства (4.116) на w^i, суммируя по i от 0 до ∞, и

используя (4.117), (4.113), получим

$$\begin{aligned}
Q(z,w) &= \sum_{i=0}^{\infty} w^i \int_0^{\infty} \mu \frac{(\mu u)^i}{i!}\, e^{-\mu u}\big(1-A(u)\big)\, du \\
&+ z\sum_{i=0}^{\infty} w^i \sum_{m=0}^{i} \widehat{q}_{i+1-m}\,(z) \int_0^{\infty} \frac{(\mu u)^m}{m!}\, e^{-\mu u}\, dA(u) \\
&= \mu\int_0^{\infty} e^{-(\mu-\mu w)u}\big(1-A(u)\big)\, du \\
&+ z\int_0^{\infty} \sum_{m=0}^{\infty} w^m \frac{(\mu u)^m}{m!} \sum_{i=m}^{\infty} w^{i-m}\, \widehat{q}_{i+1-m}\,(z)\, e^{-\mu u}\, dA(u) \\
&= \mu\int_0^{\infty} e^{-(\mu-\mu w)u} du - \mu\int_0^{\infty} e^{-(\mu-\mu w)u} A(u)\, du \\
&+ z\int_0^{\infty} e^{-(\mu-\mu w)u} dA(u) \frac{1}{w}\Big(Q(z,w) - \widehat{q}_0\,(z)\Big) \\
&= \frac{\mu}{\mu-\mu w}\Big(1-a(\mu-\mu w)\Big) \\
&+ \frac{z}{w}\, a(\mu-\mu w)\Big(Q(z,w) - \widehat{q}_0\,(z)\Big) \qquad (4.118)
\end{aligned}$$

Используя (4.118), можно получить для функции $Q(z,w)$ следующее выражение:

$$Q(z,w) = \frac{\frac{w}{1-w}\big(1-a(\mu-\mu w)\big) - za(\mu-\mu w)\,\widehat{q}_0\,(z)}{w - za(\mu-\mu w)} \qquad (4.119)$$

Так как по своему определению функция $Q(z,w)$ должна быть ограниченной в области $|z|\leqslant 1,\ \ |w|<1$, а знаменатель (4.119), как мы уже знаем, имеет единственный корень $w^*=w^*(z)$, являющийся решением уравнения

$$w^* = z\, a(\mu-\mu w^*)\,, \qquad (4.120)$$

то и числитель (4.119) тоже должен обращаться в нуль при $w=w^*(z)$. Следовательно, мы можем написать, что

$$\widehat{q}_0\,(z) = \frac{w^*(1-a(\mu-\mu w^*))}{(1-w^*)za(\mu-\mu w^*)} = \frac{1-a(\mu-\mu w^*))}{(1-w^*)} = \frac{z-w^*}{z(1-w^*)} \qquad (4.121)$$

Возвращаясь к определению (4.115) производящих функций $\widehat{q}_i\,(z),\ i\geqslant 0$ последовательностей $q_i(j),\ j\geqslant 0$, получим для интересующего нас случая $i=0$ следующее выражение:

$$\widehat{q}_0\,(z) = \sum_{j=0}^{\infty} z^j\, q_0(j) = \sum_{j=0}^{\infty} z^j\, P\big(K=j+1\big) = \frac{1}{z}\sum_{j=0}^{\infty} z^{j+1}\, P\big(K=j+1\big) \qquad (4.122)$$

Из (4.122) и (4.121) можно получить для производящей функции $\varphi(z)$ последовательности вероятностей $P(K=j)$ следующее окончательное выражение:

$$\varphi(z)=\sum_{j=1}^{\infty} z^j\ P(K=j)=z\ \widehat{q}_0\ (z)=\frac{z-w^*}{1-w^*} \tag{4.123}$$

Как мы уже знаем, единственный корень $w^*=w^*(z)$ уравнения (4.120) имеет следующую асимптотику при $z\to 1$:

$$\lim_{z\to 1} w^*\ =\ \begin{cases}1,\ \text{если } \mu a_1\leqslant 1;\\ z_0(<1),\ \text{если } \mu a_1>1.\end{cases} \tag{4.124}$$

При $\mu a_1\leqslant 1$ мы можем найти предел $\varphi(z)$ при $z\to 1$ по правилу Лопиталя, принимая во внимание, что

$$\frac{dw^*(z)}{dz}=\frac{a(\mu-\mu w^*(z))}{1+\mu z a'(\mu-\mu w^*(z))}\ \to\ \frac{1}{1-\mu a_1},\ \text{при } z\to 1\ . \tag{4.125}$$

Действительно, из (4.123) и (4.125) имеем:

$$\lim_{z\to 1}\varphi(z)=\lim_{z\to 1}\frac{1-dw^*(z)/dz}{-dw^*(z)/dz}=\mu a_1\leqslant 1\ . \tag{4.126}$$

Если же $\mu a_1>1$, то

$$\lim_{z\to 1}\varphi(z)\ =\ 1, \tag{4.127}$$

и, следовательно, в этом случае с.в. K конечна с вероятностью 1. Таким образом,

$$\mathbf{P}(K<\infty)\ =\ \begin{cases}1,\ \text{если } \mu a_1>1;\\ \mu a_1,\ \text{если } \mu a_1\leqslant 1.\end{cases}$$

В случае, когда $\mu a_1>1$ имеет смысл говорить о моментах распределения с.в. K, например, используя (4.123) и (4.125), можно вычислить

$$\mathbf{E}K\ =\ \sum_{j=1}^{\infty} j\ \mathbf{P}(K=j)\ =\ \varphi'(1)\ =\ \frac{1}{1-w^*(1)}, \tag{4.128}$$

где $w^*(1)=\lim_{z\to 1} w^*(z)$ (см. (4.124)).

Окончательное выражение для первого момента распределения с.в. K примет вид:

$$\mathbf{E}K\ =\ \begin{cases}1/(1-z_0),\ \mu a_1>1;\\ \infty,\ \mu a_1=1.\end{cases} \tag{4.129}$$

Любые моменты более высокого порядка для этого распределения могут быть вычислены аналогично.

Глава 5

Процесс рождения и гибели в приложении к ТМО

Когда мы интересуемся одномерными характеристиками систем МО, такими как длина очереди или количество заявок, находящихся в системе, то в предположении о пуассоновости входного потока мы получаем процессы, которые могут иметь лишь единичные скачки, связанные либо с приходом новой заявки, либо с уходом заявки из очереди на обслуживание или вообще из системы по окончании обслуживания соответственно. Похожие случайные процессы исследовались и до возникновения ТМО во многих приложениях, особенно в биологии, откуда и получили свое название как *процессы рождения и гибели*. Ниже мы рассмотрим некоторые результаты этих исследований, а затем их конкретные интерпретации в применении к системам МО.

5.1 Нахождение стационарных вероятностей

Рассмотрим марковский процесс $\{\xi(t), t\geqslant 0\}$ с непрерывным временем и дискретным множеством состояний $\{0, 1, \ldots\}$. Будем предполагать, что переходные вероятности этого процесса отличны от нуля только для соседних состояний, и при $h \to 0$

$$\begin{aligned}
\mathbf{P}\Big(\xi(t+h) = n+1 \mid \xi(t) = n\Big) &= \lambda_n\ h\ +\ o(h)\ , \\
\mathbf{P}\Big(\xi(t+h) = n-1 \mid \xi(t) = n\Big) &= \mu_n\ h\ +\ o(h)\ , \\
\mathbf{P}\Big(\xi(t+h) = n \mid \xi(t) = n\Big) &= 1\ -\ (\lambda_n + \mu_n)\ h\ +\ o(h)), \quad (5.1)
\end{aligned}$$

где константами $\lambda_n\geqslant 0$ и $\mu_n\geqslant 0$ обозначены интенсивности, называемые интенсивностями рождения и гибели соответственно, которые эависят от состояния n, а константа $\mu_0 = 0$.

Анализируя все возможные переходы между состояниями при $h \to 0$, нетрудно вывести следующие дифференциальные уравнения для вероятностей $P_n(t) = \mathbf{P}(\xi(t) = n)$:

$$\begin{aligned} P_0^{'}(t) &= -\lambda_0 P_0(t) + \mu_1 P_1(t) , \\ P_n^{'}(t) &= \lambda_{n-1} P_{n-1}(t) - (\lambda_n + \mu_n) P_n(t) + \mu_{n+1} P_{n+1}(t) , \quad n \geqslant 1 . \end{aligned} \tag{5.2}$$

Нас будут интересовать стационарные (установившиеся) вероятности

$$P_n = \lim_{t\to\infty} P_n(t), \quad \sum_{n\geqslant 0} P_n = 1 . \tag{5.3}$$

А так как стационарность предполагает, что $\lim_{t\to\infty} P_n^{'}(t) = 0$, то для их нахождения нам потребуется решить следующую систему алгебраических уравнений, легко получаемую из (5.2)

$$\begin{aligned} -\lambda_0 P_0 + \mu_1 P_1 &= 0, \\ \lambda_{n-1} P_{n-1} - (\lambda_n + \mu_n) P_n + \mu_{n+1} P_{n+1} &= 0 , \quad n \geqslant 1 . \end{aligned} \tag{5.4}$$

Для решения этой системы заметим, что замена переменных

$$u_n = \mu_n P_n - \lambda_{n-1} P_{n-1}, \quad n \geqslant 1 \tag{5.5}$$

приводит систему (5.4) к следующему виду

$$\begin{aligned} u_1 &= 0 , \\ u_{n+1} - u_n &= 0 , \quad n \geqslant 1 , \end{aligned} \tag{5.6}$$

которая, очевидно, имеет единственное решение $u_n \equiv 0, \ n \geqslant 1$.
Откуда нетрудно получить

$$P_n = \frac{\lambda_{n-1}}{\mu_n} P_{n-1} = \ldots = \frac{\lambda_0 \cdots \lambda_{n-1}}{\mu_1 \cdots \mu_n} P_0 . \tag{5.7}$$

Если обозначить

$$\pi_n = \frac{\lambda_0 \cdots \lambda_{n-1}}{\mu_1 \cdots \mu_n} , \quad n \geqslant 1 , \text{ и положить } \pi_0 = 1 , \tag{5.8}$$

то с учетом условия нормировки $\sum_{n\geqslant 0} P_n = 1$, окончательно получим

$$P_n = \pi_n P_0 , \quad P_0 = \Big(1 + \sum_{n\geqslant 1} \pi_n\Big)^{-1} = \Big(\sum_{n\geqslant 0} \pi_n\Big)^{-1} . \tag{5.9}$$

Видим, что если у процесса рождения и гибели существуют предельные вероятности (5.3) отличные от нуля, то должен сходиться ряд

$$\sum_{j=0}^{\infty} \pi_j < \infty . \tag{5.10}$$

Из теории марковских процессов известно, что для процесса со счетным числом состояний среднее значение времени перехода

$$\mathbf{E}\,\tau_{n,n+1} \;=\; \frac{1}{\lambda_n \pi_n}\sum_{j=0}^{n}\pi_j, \quad n\geqslant 0\;, \tag{5.11}$$

где

$$\tau_{n,n+1} \;=\; \inf\Big\{t:\; X(t)=n+1,\;\; X(0)=n\Big\}, \quad n\geqslant 0\;. \tag{5.12}$$

При этом процесс будет регулярным, то есть с вероятностью единица время его ухода на бесконечность окажется велико, если для любого $x>0$

$$\lim_{n\to\infty}\mathbf{P}\big(\tau_{0,n} \;>\; x\big) \;=\; 1\;, \tag{5.13}$$

что для процесса рождения и гибели равносильно требованию

$$\sum_{n=0}^{\infty}\frac{1}{\lambda_n\pi_n}\sum_{j=0}^{n}\pi_j \;=\; \infty \tag{5.14}$$

Условия (5.10) и (5.14) оказываются необходимыми и достаточными условиями существования и единственности стационарного распределения для процесса рождения и гибели и называются *критерием Карлина-МакГрегора (S.Karlin, J.L. McGregor).*

Напомним еще раз, что первое условие гарантирует, в частности, что стационарное значение $P_0\neq 0$, а значит и P_n для конечных значений n тоже будут ненулевыми.

5.2 Основные характеристики модели $M_\lambda|M_\mu|m|n$

Сначала покажем, что случайный процесс $\nu=\{\nu(t),\; t\geqslant 0\}$, описывающий количество заявок, находящихся в системе $M_\lambda|M_\mu|m|n$, является процессом рождения и гибели, и найдем для него соответствующие интенсивности "рождения" и "гибели".

Вследствие марковского свойства "отсутствие памяти" и независимости входного потока от времен обслуживания, для переходных вероятностей между событиями $\{\nu(t)=k\;\}$, $(0\leqslant k\leqslant m+n)$ при малом $h\to 0$ для вероятностей переходов из начального состояния имеем

$$\begin{aligned}\mathbf{P}\big(\nu(t+h)=0\;\big|\;\nu(t)=0\big) \;&=\; e^{-\lambda h} \;=\; 1-\lambda h+o(h)\;;\\ \mathbf{P}\big(\nu(t+h)=1\;\big|\;\nu(t)=0\big) \;&=\; 1\;-\;e^{-\lambda h} \;=\; \lambda h+o(h)\;.\end{aligned} \tag{5.15}$$

Для произвольных значений k получим интенсивность обслуживания, соответствующую случаю k, $(k\leqslant m)$ параллельно работающих серверов, причем

ненулевые вероятности переходов в соседние состояния будут (в зависимости от конкретного значения k) определяться не одинаково, а именно

(**I**) если $1 \leqslant k \leqslant m$, то

$$\begin{aligned}
\mathbf{P}\big(\nu(t+h)=k-1 \;\big|\; \nu(t)=k\big) &= e^{-\lambda h}\big(1-e^{-k\mu h}\big) = k\mu h + o(h)\ ; \\
\mathbf{P}\big(\nu(t+h)=k \;\big|\; \nu(t)=k\big) &= e^{-\lambda h}\ e^{-k\mu h} \ = \ 1-(\lambda+k\mu)h+o(h)\ ; \\
\mathbf{P}\big(\nu(t+h)=k+1 \;\big|\; \nu(t)=k\big) &= \big(1-e^{-\lambda h}\big)\ e^{-k\mu h} = \lambda h + o(h)\ ; \qquad (5.16)
\end{aligned}$$

(**II**) если же $m \leqslant k < m+n$, эти три вероятности будут соответственно равны

$$m\mu h \ + \ o(h)\ ; \quad 1-(\lambda+m\mu)h+o(h)\ ; \quad \lambda h + o(h)\ . \qquad (5.17)$$

Кроме того, поскольку переход из состояния $\{\nu = m+n\}$ в следующее за ним состояние $\{\nu = m+n+1\}$ невозможен, мы должны положить $\lambda_{m+n} = 0$. Все другие переходные вероятности (при переходах не в соседние состояния) оказываются равными $o(h),\ h \to 0$.

Видим, что рассматриваемый процесс $\nu(t)$ действительно является процессом "рождения и гибели"с интенсивностями

$$\begin{aligned}
\lambda_k &= \lambda\ , \quad \text{при}\ \ 0 \leqslant k < m+n\ ; \\
\mu_k &= \begin{cases} k\mu, \ \text{при}\ 1 \leqslant k \leqslant m\ ; \\ m\mu, \ \text{при}\ m \leqslant k \leqslant m+n\ . \end{cases} \qquad (5.18)
\end{aligned}$$

Соответствующие дифференциальные уравнения (см. 5.2)) примут теперь следующий вид

$$\begin{aligned}
P_0^{'}(t) &= -\lambda P_0(t) + \mu P_1(t)\ ; \\
P_k^{'}(t) &= \lambda P_{k-1}(t) - (\lambda + k\mu)P_k(t) + (k+1)\mu P_{k+1}(t)\ ,\ 1 \leqslant k \leqslant m\ ; \\
P_k^{'}(t) &= \lambda P_{k-1}(t) - (\lambda + m\mu)P_k(t) + m\mu P_{k+1}(t)\ ,\ m \leqslant k \leqslant m+n-1\ ; \\
P_{m+n}^{'}(t) &= \lambda P_{m+n-1} \ - \ m\mu P_{m+n}(t)\ . \qquad (5.19)
\end{aligned}$$

В рассматриваемом случае условия стационарности (5.10) и (5.14) оказываются автоматически выполненными, поскольку у нас $\lambda_{m+n} = 0$. Следовательно, стационарные вероятности (5.3) существуют и могут быть найдены как решения следующей системы алгебраических уравнений

$$\begin{aligned}
\lambda P_0 &= \mu P_1\ ; \\
(\lambda + k\mu)P_k &= \lambda P_{k-1} + (k+1)\mu P_{k+1}\ ,\ 1 \leqslant k \leqslant m\ ; \\
(\lambda + m\mu)P_k &= \lambda P_{k-1} + m\mu P_{k+1}\ ,\ m \leqslant k \leqslant m+n-1\ ; \\
m\mu P_{m+n} &= \lambda P_{m+n-1}\ , \qquad (5.20)
\end{aligned}$$

откуда окончательно получим

$$P_k = \begin{cases} \frac{\varrho^k}{k!}\ P_0, \ \text{при}\ \ 0 \leqslant k \leqslant m\ ; \\ \frac{\varrho^k}{m!\ m^{k-m}}\ P_0, \ \text{при}\ \ m \leqslant k \leqslant m+n\ , \end{cases} \qquad (5.21)$$

где мы обозначили

величину *интенсивности трафика*

$$\lambda/\mu \;=\; \varrho\;, \tag{5.22}$$

а значение P_0, найденное, как всегда, из условия нормировки, равно

$$P_0 = \Big(\sum_{i=0}^{m}\frac{\varrho^i}{i!} + \frac{\varrho^m}{m!}\sum_{i=1}^{n}\big(\frac{\varrho}{m}\big)^i\Big)^{-1}\;. \tag{5.23}$$

Вероятность P_{m+n}, вычисленная согласно (5.21), носит специальное наименование – *вероятность потери*, так как в этом состоянии система полна и ей некуда принять очередную приходяшую заявку. Величина этой вероятности равняется также той доле времени (от общего времени работы рассматриваемой системы), в течение которой заявки, приходящие в систему, теряются.

5.3 Некоторые частные случаи этой модели

Многие известные в ТМО марковские модели систем МО являются частными случаями рассмотренной выше системы $M_\lambda|M_\mu|m|n$. Приведем примеры некоторых из них.

1) $\mathbf{n = 0}$ – Модель Эрланга телефонной станции с потерями.

Для вероятностей состояний такой системы из (5.21) и (5.23) получим

$$P_k \;=\; \frac{\varrho^k/k!}{1+\varrho+\ldots+\varrho^m/m!}\;,\quad k \;=\; 0,1,\ldots,m\;. \tag{5.24}$$

При $k = m$ из (5.24) вытекает знаменитая формула Эрланга нахождения вероятности потерь для модели телефонной станции. Эта формула впервые позволила оценить долю времени, когда в телефонной станции с имеющимися m линиями связи будут теряться очередные приходящие звонки (вследствие полной занятости имеющихся линий).

2) $\mathbf{m = \infty,\ n = 0}$ – Queue with "self-service".

Клиенты обслуживают себя сами при экспоненциальном распределении времени обслуживания с интенсивностью μ. При $m = \infty$ предыдущая модель трансформируется в $M_\lambda|M_\mu|\infty$ со стационарным распределением

$$P_k \;=\; \frac{\varrho^k}{k!}\,e^{-\varrho}\;,\quad k \;=\; 0,1,\ldots \tag{5.25}$$

Эта модель является очень хорошей аппроксимацией реальных систем МО с большим числом серверов. Отметим, что при $m \to \infty$ (5.24) всегда переходит в (5.25), т.к. экспотенциальный ряд сходится при любом ϱ.

3) $\mathbf{n} = \infty$ – Система с бесконечным размером буфера-накопителя.

Условия (5.10) и (5.14) существования стационарного распределения (условия стабильности очереди) в этом случае будут выполнены, если величина интенсивности трафика (5.22) будет удовлетворять условию

$$\varrho \ < \ m \ . \tag{5.26}$$

При этом стационарные вероятности (5.21) и (5.23) примут вид

$$\begin{aligned} P_k &= \begin{cases} \frac{\varrho^k}{k!} \ P_0, \text{ при } \ 0 \leqslant k \leqslant m \ ; \\ \frac{\varrho^k}{m! \ m^{k-m}} \ P_0, \text{ при } k \geqslant m \ , \end{cases} \\ P_0 &= \Big(\sum_{i=0}^{m-1} \frac{\varrho^i}{i!} + \frac{\varrho^m}{m!} \sum_{i=0}^{\infty} \big(\frac{\varrho}{m}\big)^i \Big)^{-1} \ . \end{aligned} \tag{5.27}$$

В следующем разделе мы исследуем выходной поток этой системы и получим интересный и очень важный результат (Теорема Бёрке), в соответствии с которым этот поток оказывается пуассоновским и имеет ту же интенсивность, что и входящий поток.

4) $\mathbf{m = 1}, \ \mathbf{n = \infty}$ – Простейшая система $\mathbf{M}_\lambda|\mathbf{M}_\mu|\mathbf{1}|\infty$.

Для этого случая условие существования и сам вид стационарного распределения легко получаются из выражений (5.26) и (5.27) предыдущего примера при $m = 1$ и имеют следующий вид

$$P_k \ = \ \varrho^k(1-\varrho), \quad k \geqslant 0 \ , \quad \varrho \ < \ 1 \ . \tag{5.28}$$

Этот результат нам хорошо знаком, поскольку мы его уже получали ранее при независимом рассмотрении этой простейшей модели.

5) $\mathbf{m = 1}, \ \mathbf{n = N}$ – Система с ограниченным буфером $\mathbf{M}_\lambda|\mathbf{M}_\mu|\mathbf{1}|\mathbf{N}$.

Процесс $\nu(t)$ в этой системе представляет собой процесс рождения и гибели, у которого

$$\mu_1 = \mu_2 = \ldots = \mu_N = \mu,$$

$$\lambda_0 = \lambda_1 = \lambda_2 = \ldots = \lambda_{N-1} = \lambda,$$

а остальные

$$\lambda_N = \lambda_{N+1} = \ldots = 0 \ .$$

Здесь стационарные вероятности имеют вид

$$P_k = \begin{cases} \varrho^k P_0, \text{ для } \ k = 0, 1, 2, \ldots, N \ ; \\ 0, \text{ для } k \ > \ N \ , \end{cases}$$

с условием нормировки

$$\sum_{j=0}^{N} P_j \ = \ 1 \ .$$

Откуда окончательно получим

$$P_k \;=\; \frac{\varrho^k(\varrho-1)}{\varrho^{N+1}-1}\;, \quad k=0,1,2,\ldots,N\;. \tag{5.29}$$

Результат, естественно, справедлив при любом ϱ, так как рассматриваемый ряд суммирования имеет конечное число $N+1$ членов. Очевидно также, что при $\varrho \;<\; 1$ мы можем устремить $N \to \infty$ и получить тогда выражение (5.28).

6) – Простейшая модель с "пугающей"очередью.

Здесь предполагается, что пришедшая в систему $M_\lambda|M_\mu|1|\infty$ заявка с вероятностью f_n немедленно покинет систему, если до неё там уже находилось n заявок, либо останется в системе (встанет в очередь на обслуживание) с вероятностью $(1-f_n)$. Заметим, что при $f_n \equiv 0$ мы получим обычную систему $M_\lambda|M_\mu|1|\infty$, а если

$$f_n = \begin{cases} 0, \text{для } 0\leqslant n\leqslant N; \\ 1, \text{для } n > N\;, \end{cases}$$

то получим систему $M_\lambda|M_\mu|1|N$.

В модели с "пугающей" очередью, очевидно,

$$\lambda_i \;=\; \lambda(1-f_i),\; i\geqslant 0\;; \quad \mu_i \;=\; \mu\;,\; i\geqslant 1\;.$$

Поэтому здесь

$$\pi_j \;=\; \varrho^j \prod_{i=0}^{j-1}(1-f_i)\;, \quad j\geqslant 1\;, \quad \varrho \;=\; \frac{\lambda}{\mu}\;,$$

а для существования предельных (стационарных) вероятностей необходимо потребовать, чтобы

$$\sigma \;=\; 1 \;+\; \sum_{j=1}^{\infty}\pi_j \;<\; \infty\;.$$

Тогда для искомых вероятностей получим следующие выражения

$$P_0 \;=\; \frac{1}{\sigma}\;, \quad P_j \;=\; \frac{\pi_j}{\sigma}\;, \quad j\geqslant 1\;.$$

В качестве интересного примера рассмотрим один частный случай модели с "пугающей" очередью, когда вероятность $(1-f_n)$ присоединиться к очереди размера $n\geqslant 1$ падает по мере увеличения очереди, например, как (f/n), где константа f ограничена $0 \;<\; f \;<\; 1$. Поскольку эта вероятность при $k \to \infty$ стремится к нулю, предельные (стационарные) вероятности будут существовать при любых значениях λ и μ.

Так как в этом случае

$$\prod_{i=0}^{j-1}(1-f_i) \;=\; \frac{f^{j-1}}{(j-1)!}\;,\quad j\geqslant 1\;,$$

а

$$\sigma \;=\; 1 \;+\; \varrho\; e^{\varrho f}\;,$$

окончательно получим следующий вид стационарных вероятностей

$$P_0 \;=\; \frac{1}{1+\varrho\; e^{\varrho f}}\;,\quad P_j \;=\; \frac{\varrho^j\; f^{j-1}}{(1+\varrho\; e^{\varrho f})(j-1)!}\;,\quad j\geqslant 1\;.$$

Этот пример полезен для практических приложений.

7) – Модель $\mathbf{M}_\lambda|\mathbf{M}_\mu|\mathbf{N}|\infty$ с "нетерпеливыми" ожидающими.

Предполагается, что каждая эаявка, поступающая в систему и встающая в очередь ожидания обслуживания, будет находиться в ней не более, чем некоторое случайное время, имеющее ф.р. $1-e^{-\gamma x}$. Если за это время заявка не успевает поступить на обслуживание, она покидает очередь, а значит и саму систему.

Для процесса рождения и гибели, описывающего такую модель, следует ввести следующие интенсивности переходов

$$\lambda_i \equiv \lambda\;,\quad i\geqslant 0\;;\quad \mu_i = \begin{cases} i\mu\;, & i\leqslant N\;;\\ N\mu+(i-N)\gamma\;, & i>N\;. \end{cases}$$

Для такой модели получим

$$\pi_j = \begin{cases} \frac{1}{j!}(\frac{\lambda}{\mu})^j\;, & j\leqslant N\;;\\ \lambda^j\Big(N!\mu^N\prod_{i=N+1}^{j}\big(N\mu+(i-N)\gamma\big)\Big)^{-1}\;, & j>N\;. \end{cases}$$

Условие существования предельных (стационарных) вероятностей будет выполнено при $\gamma>0$ для любой пары $\lambda>0,\ \mu>0$, а сами эти вероятности определяются из следующих выражений

$$P_j = \begin{cases} \frac{P_0}{j!}(\frac{\lambda}{\mu})^j\;, & j\leqslant N\;;\\ P_0\lambda^j\Big(N!\mu^N\prod_{i=N+1}^{j}\big(N\mu+(i-N)\gamma\big)\Big)^{-1}\;, & j>N\;, \end{cases}$$

где

$$P_0 = \Big(\sum_{j=0}^{N}\frac{1}{j!}(\frac{\lambda}{\mu})^j + \frac{1}{N!}(\frac{\lambda}{\mu})^N\sum_{k=1}^{\infty}\lambda^k\big(\prod_{r=1}^{k}(N\mu+r\gamma)\big)^{-1}\Big)^{-1}\;.$$

5.4 Выходной поток системы $M_\lambda|M_\mu|m|\infty$. Теорема Бёрке

Как мы видели выше, для системы $M_\lambda|M_\mu|m|\infty$, у которой входящий поток имеет интенсивность λ, а время обслуживания в каждом из m серверов подчиняется экспоненциальному закону распределения с параметром μ, стационарная вероятность P_n того, что в системе находится ровно n заявок, может быть получена как решение системы уравнений для процесса рождения и гибели с интенсивностями рождения $\lambda_n \equiv \lambda$, а гибели $\mu_n = n\mu \ \ (1\leqslant n\leqslant m)$, или $\mu_n = m\mu \ \ (n > m)$.
Мы могли заметить, что вероятности P_n удовлетворяют рекуррентным соотношениям

$$P_n = \begin{cases} \frac{\lambda}{n\mu} P_{n-1}, & 1\leqslant n\leqslant m \ ; \\ \frac{\lambda}{m\mu} \ P_{n-1}, & n > m \ . \end{cases} \tag{5.30}$$

Но вероятность P_n может также означать, например, и вероятность того, что закончившая обслуживание заявка, покидая систему, оставляет в ней ровно n заявок (при условии установившегося режима работы системы).

Воспользуемся этим результатом при изучении выходного потока рассматриваемой системы. Оказывается справедливой следующая теорема

Теорема 5.1. *[Бёрке (P.J. Burke)] Выходящий из системы $M_\lambda|M_\mu|m|\infty$ поток обслуженных заявок оказывается пуассоновским, причем той же интенсивности $\lambda > 0$, что и входящий в неё пуассоновский поток.*

Доказательство.

Пусть с.в. τ обозначает временной интервал между двумя последовательными моментами ухода обслуженных заявок из нашей системы. Определим

$$S_n(t) \ = \ \mathbf{P}\big(N(t) = n, \ \tau > t\big) \ , \tag{5.31}$$

где $N(t)$ - есть состояние системы в момент времени t, так что до момента t еще не произошло ни одного из следующих "уходов".
Тогда на интервале времени $(t, \ t+\delta t)$, на котором не может быть следующих "уходов", в первом приближении по δt допустима лишь следующая "активность"

$$\begin{aligned} S_0(t+\delta t) &= (1-\lambda\delta t) \ S_0(t) \ ; \\ S_n(t+\delta t) &= \lambda\delta t \ S_{n-1}(t) + (1-\lambda\delta t)(1-n\mu\delta t) \ S_n(t), \ \ 1\leqslant n\leqslant m \ ; \\ S_n(t+\delta t) &= \lambda\delta t \ S_{n-1}(t) + (1-\lambda\delta t)(1-m\mu\delta t) \ S_n(t), \ \ n > m \ . \end{aligned} \tag{5.32}$$

Откуда нетрудно получить систему уравнений

$$\begin{aligned} \frac{dS_0(t)}{dt} &= -\lambda \ S_0(t) \ ; \\ \frac{dS_n(t)}{dt} &= \lambda \ S_{n-1}(t) - (\lambda + n \ \mu) \ S_n(t), \ \ 1\leqslant n\leqslant m \ ; \\ \frac{dS_n(t)}{dt} &= \lambda \ S_{n-1}(t) - (\lambda + m \ \mu) \ S_n(t), \ \ n > m \ . \end{aligned} \tag{5.33}$$

Очевидно, эти дифференциально-разностные уравнения относительно $S_n(t)$ нам нужно решать с начальными условиями (см. (5.30))

$$S_n(0) \;=\; P_n \;, \tag{5.34}$$

поскольку в момент ухода заявки, от которого мы ведем отсчет τ, в системе находилось ровно n заявок.
Как легко проверяется подстановкой в (5.33), следующие формулы

$$S_0 \;=\; P_0\, e^{-\lambda t}, \;\; S_n(t) \;=\; P_n\, e^{-\lambda t} \tag{5.35}$$

дают решение, удовлетворяющее указанным начальным условиям.

Действительно, если, например, при $1\leqslant n\leqslant m$ в левой части уравнения имеем $(-\lambda P_n e^{-\lambda t})$, то в правой части получаем $[\lambda P_{n-1}e^{-\lambda t} - \lambda P_n e^{-\lambda t} - n\mu P_n e^{-\lambda t}]$, а это и есть то же самое выражение, поскольку из (5.30) следует $\lambda P_{n-1} = n\mu P_n$ при $1\leqslant n\leqslant m$. Случай $n > m$ рассматривается аналогично.

Таким образом, окончательно получаем

$$S_n(t) \;=\; P_n\, e^{-\lambda t}, \;\; n\geqslant 0, \;\; (t < \tau) \;. \tag{5.36}$$

Если теперь рассмотреть все возможные состояния $n\geqslant 0$, то из определения совместного распределения (5.31) получим

$$\mathbf{P}(\tau > t) \;=\; \sum_{n=0}^{\infty} S_n(t) \;=\; e^{-\lambda t}\sum_{n=0}^{\infty} P_n \;=\; e^{-\lambda t} \;. \tag{5.37}$$

А следовательно

$$\mathbf{P}(\tau\leqslant t) \;=\; 1 \;-\; e^{-\lambda t} \;. \tag{5.38}$$

Тем самым мы показали, что распределение отдельного интервала между двумя последовательными уходами заявок из нашей системы является экспоненциальным распределением. А мы уже знаем, что если интервалы между последовательными моментами имеют экспоненциальное распределение, то сами эти моменты образуют пуассоновский поток.

Оказывается, что эти временные интервалы между последовательными моментами ухода заявок из системы по окончании обслуживания не зависят от состояния системы до момента их ухода, а также от предыдущих временных интервалов между такими же последовательными уходами. Это легко показать, если снова рассмотреть двумерное распределение (5.31), но теперь на интервале ухода следующей очередной заявки.
Тогда

$$\mathbf{P}\big(N(t+\delta t) = n,\; t < \tau\leqslant t+\delta t\big) = \begin{cases} S_{n+1}(t)\cdot(n+1)\mu\delta t, \;\; (n+1)\leqslant m \;; \\ S_{n+1}(t)\cdot m\mu\delta t, \;\; (n+1) > m \;, \end{cases} \tag{5.39}$$

то есть когда в системе вплоть до момента t находилась $(n+1)$ заявка, а затем одна ушла.

Если переписать теперь выражение в левой части в терминах τ, а в правой части подставить выражение S_{n+1} через P_{n+1} согласно (5.36) и учесть, что там соответственно

$$P_{n+1} \;=\; \frac{\lambda P_n}{(n+1)\mu}\;,\quad m\mu P_{n+1} \;=\; \lambda P_n,$$

то получим выражение

$$\mathbf{P}\big(N(\tau+0)=n,\;\; t<\tau<t+\delta t\big) \;=\; P_n\cdot\;\lambda\, e^{-\lambda t}\;\delta t\;. \qquad (5.40)$$

А это с учетом (5.38) означает ни что иное как

$$\mathbf{P}\big(N(\tau{+}0)=n,\;\; t<\tau<t{+}\delta t\big) \;=\; \mathbf{P}\big(N(\tau{+}0)=n\big)\cdot\mathbf{P}\big(t<\tau<t{+}\delta t\big)\;, \quad (5.41)$$

и тем самым доказана независимость τ и $N(\tau)$.
Таким образом, мы доказали, что число заявок в системе в любой момент времени не зависит от состояния системы в предыдущие моменты ухода из нее заявок, а также и от предыдущих временных интервалов между такими последовательными уходами. Этот факт завершает исследование выходного потока нашей системы.
Теорема доказана полностью. □

Для еще лучшего понимания сути доказанной теоремы рассмотрим в заключение этого раздела выходной поток более простой системы $M_\lambda|M_\mu|1|\infty$.

Ясно, что если обслуженная заявка покидает непустую систему, то интервал между моментом ее ухода и моментом ухода следующей за ней заявки равен времени обслуживания этой следующей заявки, то есть имеет ф.р. $(1-e^{-\mu t})$. Но если обслуженная заявка, покидая систему, оставляет систему пустой, то этот интервал будет равен сумме двух случайных величин: остаточного времени X - до прихода в систему следующей заявки и времени Y - длительности ее обслуживания. Как мы уже хорошо знаем, первая величина имеет ф.р. $(1-e^{-\lambda t})$, а вторая - $(1-e^{-\mu t})$.

Если мы обозначим $Z=X+Y$, то для плотности вероятности распределения этой с.в. Z по правилу свертки получим

$$\begin{aligned} f_Z(z) \;&=\; \int_0^z f_X(x)\; f_Y(z-x)\; dx \\ &=\; \int_0^z \lambda e^{-\lambda x}\;\mu e^{-\mu(z-x)}\; dx \;=\; \lambda\mu\; e^{-\mu z}\int_0^z e^{(\mu-\lambda)x}\; dx \\ &=\; \frac{\lambda\mu}{\mu-\lambda}\; e^{-\mu z}\Big[e^{(\mu-\lambda)z} \;-\; 1\Big] \;=\; \frac{\lambda\mu}{\mu-\lambda}\; e^{-\lambda z} \;-\; \frac{\lambda\mu}{\mu-\lambda}\; e^{-\mu z}\;. \end{aligned}$$

Но тогда, если обозначить $\varphi(t)$ - плотность вероятности распределения интервалов времени между моментами любых двух последовательных уходов заявок из системы, то, учитывая обе рассмотренные выше возможности,

получим

$$\begin{aligned} \varphi(t) &= P_0 \left[\frac{\lambda\mu}{\mu-\lambda}\, e^{-\lambda t} - \frac{\lambda\mu}{\mu-\lambda}\, e^{-\mu t} \right] \\ &+ (1-P_0)\, \mu\, e^{-\mu t}\,, \end{aligned} \tag{5.42}$$

где $P_0 = 1 - \varrho = 1 - \lambda/\mu = (\mu-\lambda)/\mu$.
Откуда после подстановки окончательно получим

$$\varphi(t) = \lambda\, e^{-\lambda t} - \lambda\, e^{-\mu t} + \lambda\, e^{-\mu t} = \lambda e^{-\lambda t}\,, \quad t \geqslant 0\,. \tag{5.43}$$

Это и есть то, что дает нам теорема Бёрке, а именно, что выходной поток оказывается пуассоновским и имеет ту же интенсивность λ, что и входящий в систему поток.

Глава 6

Основы теории марковских сетей

До сих пор мы рассматривали системы МО более или менее изолированными друг от друга. Однако, во многих практических приложениях, например, в телекоммуникационных системах, мы наблюдаем ситуации, когда выход одной системы с очередью оказывается входом другой системы, или когда выходы двух и более систем в той или иной комбинации образуют вход в некоторую новую систему, а ее выход затем расщепляется на несколько каналов, каждый из которых, в свою очередь, оказывается входом для последующих систем МО, и т.д. и т.п.
Такие соединения наблюдаются и в производственных процессах, когда выпускаемые изделия, перед тем как стать готовой продукцией, проходят, например, несколько стадий обработки, сборки, настройки, а затем встают в очередь на реализацию.
В системах электронной коммуникации и передающих сетях сообщения сперва кодируются, преобразуясь в электронный сигнал, затем поступают в передающее устройство, после чего через точки передающей сети попадают, наконец, в приемное устройство для последующей обработки в устройствах декодирования.
К непосредственному рассмотрению таких соединений систем МО мы и приступаем в нашем дальнейшем изложении (см. [10]).

6.1 Слияние и расщепление пуассоновских потоков

Для начала рассмотрим слияние (лат. – merging) двух независимых пуассоновских потоков и покажем, что в результате такого слияния получится пуассоновский поток суммарной интенсивности.

Пусть $X(t)$ – число заявок, приходящих к моменту t из пуассоновского

потока с интенсивностью λ_1, а $Y(t)$ – число заявок к моменту t, приходящих из независимого от $X(t)$ пуассоновского потока с интенсивностью λ_2. Тогда, если $Z(t) = X(t) + Y(t)$, то число заявок $Z(t)$ в образованном при таком слиянии объединенном потоке оказывается таким же, как и число заявок, приходящих к моменту t из пуассоновского потока с интенсивностью $(\lambda_1 + \lambda_2)$.

Действительно, так как сливающиеся потоки – пуассоновские, то

$$\begin{aligned} \mathbf{P}(X(t) = x) &= \frac{(\lambda_1 t)^x \ e^{-\lambda_1 t}}{x!} \ , \\ \mathbf{P}(Y(t) = y) &= \frac{(\lambda_2 t)^y \ e^{-\lambda_2 t}}{y!} \ . \end{aligned}$$

Но тогда распределение суммы этих потоков, как свертка указанных двух распределений, с очевидностью, будет равна

$$\begin{aligned} \mathbf{P}(Z(t) = z) &= \sum_{x=0}^{z} \mathbf{P}(X(t) = x) \ \mathbf{P}(Y(t) = z - x) \\ &= \sum_{x=0}^{z} \frac{(\lambda_1 t)^x \ e^{-\lambda_1 t}}{x!} \ \frac{(\lambda_2 t)^{z-x} \ e^{-\lambda_2 t}}{(z-x)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)t}}{z!} \sum_{x=0}^{z} \frac{z!}{x! \ (z-x)!} (\lambda_1 t)^x \ (\lambda_2 t)^{z-x} \\ &= \frac{\big[(\lambda_1 + \lambda_2)t\big]^z}{z!} \ e^{-(\lambda_1+\lambda_2)t} \ , \quad z = 0, 1, 2, \dots \ , \end{aligned}$$

где при последнем преобразовании выражений была использована известная формула Бинома Ньютона.
Полученный результат легко распространяется на сумму трех и более независимых потоков.

При рассмотрении расщепления (лат. – splitting) пуассоновского потока предположим, что лишь доля p заявок, выходящих из некоторой системы МО (типа $M_\lambda|M_\mu|1|\infty$, или $M_\lambda|M_\mu|m|\infty$, или $M_\lambda|M_\mu|\infty$), направляется во вторую систему МО, а остальная доля $(1-p)$ из них уходит, не заходя на второе обслуживание. Либо можно просто предположить, что выходящий поток заявок $Z(t)$ направляется далее, разделившись на два потока $X(t)$ и $Y(t)$ в указанной выше пропорции.
Математически это означает, что с вероятностью p заявка из потока $Z(t)$ попадает в ветвь $X(t)$, а остальная доля $(1-p)$ заявок – в $Y(t)$. Тогда ясно, что

$$\mathbf{P}\Big(X(t) = m \ \Big| Z(t) = n\Big) \ = \ C_n^m p^m (1-p)^{n-m} \ , \quad n \geqslant m \ ,$$

и тогда

$$\mathbf{P}\Big(X(t) = m\Big) \ = \ \sum_{n=m}^{\infty} \mathbf{P}\Big(X(t) = m \ \Big| Z(t) = n\Big) \ \mathbf{P}\Big(Z(t) = n\Big) \ .$$

Подставляя сюда соответствующие выражения вероятностей, получим

$$\begin{aligned}\mathbf{P}\Big(X(t)=m\Big) &= \sum_{n=m}^{\infty} \frac{n!}{m!\ (n-m)!}\ p^m(1-p)^{n-m}\ \frac{(\lambda t)^n}{n!}\ e^{-\lambda t} \\ &= \frac{(p\lambda t)^m\ e^{-\lambda t}}{m!} \sum_{n=m}^{\infty} \frac{\big[\lambda t(1-p)\big]^{n-m}}{(n-m)!} \\ &= \frac{(p\lambda t)^m\ e^{-\lambda t}}{m!}\ e^{\lambda t(1-p)} \ = \ \frac{(p\lambda t)^m}{m!}\ e^{-p\lambda t}\ ,\end{aligned}$$

а это означает, что $X(t)$ имеет пуассоновское распределение с параметром $(p\lambda)$. Вторая же ветвь расщепленого потока – $Y(t)$, очевидно, будет иметь интенсивность $(1-p)\lambda$.
Этот результат тоже легко распространяется на случаи, когда пуассоновский поток расщепляется не на два, а на большее число потоков.

В качестве упражнения предлагается провести оба рассмотренных в данном разделе доказательства с использованием производящей функции пуассоновского потока.

6.2 Процессы, используемые при моделировании сетей

Итак, мы переходим от изучения отдельных систем МО к исследованию поведения их соединений друг с другом. Будем предполагать, что эти соединения образуют некую сетевую структуру, а сами системы МО, входящие в эту структуру, будем называть далее *узлами* такой сети.

Пусть сеть состоит из m узлов, пронумерованых, например, $1, 2, \ldots, m$, где $m < \infty$. Между этими узлами определенным образом движутся некоторые дискретные элементы, модули или пакеты, являющиеся "клиентами" , обслуживаемыми в узлах сети. Например, в компьютерных или телекоммуникационных сетях узлом может обозначаться компьютер, некоторый файл с данными, или станция коммутации; а дискретными элементами могут быть блоки данных, сообщения (т.е. пакеты, состоящие из нескольких блоков), телефонные звонки или трансакции (элементарные целостные операции над данными, например, запрос, удаление или модификация записи в базе данных, или обновление файла – в файловых системах). В производственных сетях узлом может быть рабочая станция, зона складирования, точка контроля, источник запросов, или станция автоматически управляемых перевозчиков-погрузчиков; а элементами – часть или группа частей или запросов на продукт, или сообщения. В таких сетях случайность или рандомизация присутствует в процедуре обслуживания или в маршрутизации элементов между узлами и обусловливается либо свойствами движущихся в сети элементов, либо свойствами узлов, либо и тех и других сразу.

Изменение состояния такой сети представляют стохастическим процессом непрерывного времени $\{X(t):\ t\geqslant 0\}$, состояния которого описываются вектором $x = (x_1, x_2, \ldots, x_m)$ в конечном или бесконечном пространстве состояний $\mathbb{E}$, где величина x_j означает число элементов в узле j этой сети.

Сеть называется *замкнутой*, если полное число элементов в ней $|x| = x_1 + x_2 + \ldots + x_m$ всегда сохраняется $|x| = \nu = const$. Для замкнутой сети $\mathbb{E} = \{x:\ |x| = \nu\}$.

Иначе, сеть называется *открытой*, причем различают *открытые сети с конечной емкостью* ν – если $\mathbb{E} = \{x:\ 0\leqslant|x|\leqslant\nu\}$, и *открытые сети с неограниченной емкостью* – когда $\mathbb{E} = \{x:\ 0\leqslant|x| < \infty\}$.

Будем предполагать далее, что рассматриваемый сетевой процесс X является скачкообразным марковским процессом, или марковской цепью с непрерывным временем. Тогда его вероятностные распределения будут полностью определяться переходной интенсивностью (интенсивностью переходов между его состояниями), которая определяется следующим образом

$$q(x,y) = \lim_{t\downarrow 0} t^{-1}\ \mathbf{P}\{X_t = y \mid X_0 = x\}\ ,\quad y \neq x\ , \tag{6.1}$$

$$q(x,x) \equiv 0\ . \tag{6.2}$$

Будем предполагать также, что сетевой процесс X является регулярным, то есть что он не может совершить бесконечное число скачков за конечный временной интервал.

Кроме того, чтобы избежать вырождений, будем предполагать, что сетевой процесс X не имеет поглощающих состояний.

Все это гарантирует достижимость из произвольного допустимого состояния любого другого допустимого состояния за конечное число переходов.

Чтобы смоделировать реальную сеть таким процессом, нам необходимо будет в дальнейшем перевести функциональные свойства узлов сети и правила маршрутизации движущихся в сети элементов в специфический вид функции q интенсивности переходов.

Сначала, однако, займемся рассмотрением некоторых общих свойств используемых процессов. Поскольку X – марковский скачкообразный процесс, то время его пребывания в любом состоянии распределено экспоненциально. Таким образом, как только X оказывается в некотором состоянии x, он будет оставаться в этом состоянии в течение времени, экспоненциально распределенного с интенсивностью

$$q(x) = \sum_y q(x,y)\ , \tag{6.3}$$

а после окончания пребывания в этом состоянии x сетевой процесс перейдет скачком в некоторое другое состояние y с вероятностью

$$p(x,y) = \frac{q(x,y)}{q(x)}\ . \tag{6.4}$$

Такие перескоки, чередуясь с пребываниями в состояниях, могут продолжаться неограниченно. Получающаяся в результате последовательность состояний, посещаемых процессом X, образует марковскую цепь с переходными вероятностями $p(x,y)$.

Следует отметить, что в практических приложениях стандартный путь определения переходных интенсивностей $q(x,y)$ обычно состоит сначала в определении (или оценивании) интенсивностей $q(x)$ экспоненциального распределения времен пребывания в состояниях x и вероятностей переходов $p(x,y)$, а затем уже в вычислении самих величин $q(x,y) = q(x)p(x,y)$.

Покажем, что введенное нами описание (6.1)-(6.4) сетевого процесса ничуть не умаляет общности. Так, например, процесс, определенный немного иначе, довольно легко может быть сведен к "стандартному", рассмотренному выше, описанию следующим образом.

Пусть X – процесс, заданный на счетном пространстве состояний $\mathbb{E}$, и такой, что последовательность состояний, которые он посещает, тоже образует марковскую цепь с переходной вероятностью $\widetilde{p}\ (x,y)$, но вероятность возможного перехода из состояния x обратно в него же $\widetilde{p}\ (x,x) \geqslant 0$, а время самого пребывания в любом таком состоянии x до следующего перехода имеет экспоненциальное распределение с интенсивностью $\lambda(x)$.

Очевидно, последовательность "различных" состояний, посещаемых таким процессом X, образует марковскую цепь со следующей переходной вероятностью

$$p(x,y) \;=\; \frac{\widetilde{p}\ (x,y)}{1 - \widetilde{p}\ (x,x)}\ , \quad y \neq x\ , \tag{6.5}$$

$$p(x,x) \;\equiv\; 0\ . \tag{6.6}$$

Причем, если $\widetilde{p}\ (x,x) > 0$, то "полное" время пребывания такого процесса в состоянии x – есть сумма всех последовательных времен пребывания, распределенных экспоненциально с параметром (интенсивностью) $\lambda(x)$. В этом случае мы наблюдаем ситуацию, когда на выходе из состояния x поток с интенсивностью $\lambda(x)$ расщепляется на два потока: доля $\widetilde{p}\ (x,x)$ возвращается обратно в x, а доля $(1 - \widetilde{p}\ (x,x))$ направляется в некоторое другое состояние $y \neq x$. Поэтому полное время пребывания в состоянии x распределено экспоненциально с параметром

$$q(x) \;=\; \lambda(x)(1 - \widetilde{p}\ (x,x))\ . \tag{6.7}$$

Но тогда X – есть марковский процесс с переходной интенсивностью

$$q(x,y) \;=\; q(x)\ p(x,y) \;=\; \lambda(x)\ \widetilde{p}\ (x,y)\ , \tag{6.8}$$

что и заканчивает приведение его к "стандартному" описанию.

Для открытых сетей удобно ввести в рассмотрение понятие узла "0", объединяющего в себе всё, что находится вне рассматриваемой сети. Причем размер популяции в этом нулевом узле, являющемся лишь "источником", либо "приёмником"элементов, движущихся в сети, никак не будет

отражаться в векторе состояния сети, которым по-прежнему будет являться вектор $x = (x_1, x_2, \ldots, x_m)$. Однако множеством узлов M открытой сети будем далее считать множество $\{0, 1, 2, \ldots, m\}$, в отличие от множества узлов $\{1, 2, \ldots, m\}$ для случая замкнутой сети.

Мы будем рассматривать сетевой процесс лишь с перемещениями единичных элементов (или одноэлементными перемещениями), так что типичным переходом для процесса X будет перескок, вызванный перемещением одного элемента из некоторого узла $j \in M$ в другой узел $k \in M$. Конкретно, если сетевой процесс X находится в состоянии $x \in \mathbb{E}$, и единичный элемент (например, пакет) перемещается из узла $j \in M$ в другой узел сети $k \in M$, то следующее состояние этой сети будет $T_{jk}x$, где T_{jk} – оператор, действующий на x таким образом, что $(T_{jk}x)$ – это вектор, у которого j-я координата на единицу меньше, а k-я – на единицу больше, чем соответствующие координаты у вектора x. Мы можем формально написать, что

$$T_{jk}x \;=\; x \;-\; e_j \;+\; e_k \;, \quad e_0 = 0 \;, \tag{6.9}$$

где $e_i, \;\; i = 1, 2, \ldots, m$ – единичный вектор, у которого единица стоит на i-том месте, а остальные координаты – нули.

Экспоненциальность распределения времен пребывания в состоянии x естественно сформулировать тогда следующим образом: для каждой пары $j, k \in M$ предполагается, что время до следующего "возможного"перехода пакета из j в k, или смена состояния сети x на новое состояние $T_{jk}x$ распределено экспоненциально с интенсивностью $q(x, T_{jk}x)$ и все такие времена – независимы. Конкретная форма задания этих величин $q(x, T_{jk}x)$ будет, естественно, зависеть от конкретного вида модели сети, которую мы будем далее рассматривать.

Тогда время τ пребывания процесса X в состоянии x, которое, очевидно, является минимумом из всех этих независимых возможных времен ожидания единичного перескока, будет также распределено экспоненциально, а интенсивность этого распределения будет, в соответствие с (2.36), равна

$$q(x) \;=\; \sum_{j \in M} \sum_{k \in M} q(x,\; T_{jk}x) \;, \tag{6.10}$$

причем сумма берется по всем j и k из M, если специально не оговорено что-либо иное.

Более того, величина

$$\frac{q(x,\; T_{jk}x)}{q(x)} \tag{6.11}$$

представляет собой вероятность того, что такой скачок произойдет именно за счет перемещения единичного элемента из данного узла j в конкретно указанный узел k.

Заметим, что более сложные описания сетей рассматривают также переходы типа многоэлементных перемещений, с возможностью перехода из x в $x + a - d$, где векторы $a = (a_1, a_2, \ldots, a_m)$ и $d = (d_1, d_2, \ldots, d_m)$ – определяют число приходящих и уходящих элементов в соответствующих узлах

сети. В этом случае естественным допущением будет предположение, что если сеть находится в состоянии x, то время до следующего потенциально возможного перехода в состояние $x-d+a$ распределено экспоненциально с интенсивностью $q(x, x-d+a)$, и что эти времена являются независимыми для всех возможных векторов d и a.

6.3 Некоторые свойства скачкообразных процессов

Пусть $\{X_t\}_{t\geqslant 0}$ – описанный выше скачкообразный марковский процесс на счетном пространстве состояний $\mathbb{E}$ с заданными переходными интенсивностями $q(x, y)$.
Положительная мера π на $\mathbb{E}$ называется *инвариантной мерой* для X (или для q), если она удовлетворяет следующему балансному уравнению

$$\pi(x)\sum_y q(x,y) \;=\; \sum_y \pi(y)q(y,x)\ , \quad x\in\mathbb{E}\ , \tag{6.12}$$

причем, если процесс X является неприводимым и положительно рекуррентным, то существует единственная положительная вероятностная мера π, удовлетворяющая этому уравнению. В таком случае процесс X называют *эргодическим* процессом, а π – стационарным или равновесным распределением X. Иногда (для простоты) для эргодического процесса, найдя его инвариантную меру, следующий шаг – ее нормировку (чтобы получить само стационарное распределение) не проводят.
Отметим также, что если X – эргодический процесс, то его стационарное распределение π является и его предельным распределением, в том смысле, что

$$\lim_{t\to\infty} \mathbf{P}\big(X_t = x\big) \;=\; \pi(x) \tag{6.13}$$

Вспомним, что стохастический процесс является стационарным, если его конечномерные распределения инвариантны относительно любых сдвигов во времени.
Так как наш X – марковский процесс, то для его стационарности необходимо и достаточно, чтобы

$$\mathbf{P}\big(X_t = x\big) \;=\; \pi(x),\ \text{для любых x и t.} \tag{6.14}$$

Множество ценовых параметров и параметров эффективности, определяющих свойства изучаемых марковских процессов, обычно выражают в терминах некоторых специальных функционалов. Предположим, в связи с этим, что некоторая величина (например, цена или производительность) изменяется непрерывно с интенсивностью $f(x)$ в единицу времени всякий раз, как

процесс X находится в состоянии x. Тогда полное изменение этой величины за интервал времени $[0,t]$ будет равно

$$\int_0^t f(X_s)\ ds \tag{6.15}$$

Бывают интересны также и характеристики, связанные с переходами процесса. Предположим при этом, что $h(x,y)$ – некоторая ценовая величина, связанная с переходами процесса X из x в y. Тогда полная стоимость переходов на интервале $(0,t]$ будет

$$\sum_n h(X_{\tau_{n-1}},\ X_{\tau_n})\ \mathbf{1}(\tau_n \in (0,t])\ , \tag{6.16}$$

где $0 \equiv \tau_0 < \tau_1 < \tau_2 < \ldots$ – моменты скачков процесса X. Здесь следует отметить, что X_{τ_n} – есть значение процесса X в момент n-го скачка, а функция $\mathbf{1}(\cdot)$ – обозначает индикатор, равный 1 или 0 в зависимости от верности или ложности утверждения в скобках.

Эргодическая теория марковских процессов устанавливает, что предельное среднее значение таких функционалов существует, и, более того, эти пределы совпадают с математическим ожиданием функционалов в случае, когда процесс X стационарен. Этот результат удобно сформулировать следующим образом

Теорема 6.1. *[**закон больших чисел**] Если марковский процесс X – эргодический со стационарным распределением π, то с вероятностью 1 (w.p.1)*

$$\lim_{t\to\infty} t^{-1} \int_0^t f(X_s)\ ds = \sum_x \pi(x) f(x)\ , \tag{6.17}$$

$$\lim_{t\to\infty} t^{-1} \sum_n h\big(X_{\tau_{n-1}}, X_{\tau_n}\big)\mathbf{1}\big(\tau_n \in (0,t]\big) = \sum_x \pi(x) \sum_y q(x,y) h(x,y)\ , \tag{6.18}$$

при условии, что указанные суммы существуют.
Эти предельные выражения остаются справедливыми даже после замены случайных функций в них на их соответствующие математические ожидания.
Более того, если X – стационарный процесс, то данные пределы являются соответствующими математическими ожиданиями:

$$\mathbf{E}\int_0^1 f(X_s)\ ds\ ,\quad \mathbf{E}\sum_n h(X_{\tau_{n-1}},\ X_{\tau_n})\ \mathbf{1}(\tau_n \in (0,1])\ . \tag{6.19}$$

Эта теорема позволяет подробнее исследовать некоторые полезные свойства процесса X.

Например, среднее число переходов процесса X из области A в B в единицу времени – есть ни что иное как

$$\begin{aligned}\mathsf{\Pi Q}(A,B) &\equiv \sum_{x\in A}\pi(x)\sum_{y\in B}q(x,y)\\ &= \lim_{t\to\infty}t^{-1}\sum_{n}\mathbf{1}\big(X_{\tau_{n-1}}\in A, X_{\tau_n}\in B, \tau_n\in(0,t]\big)\ . \quad (6.20)\end{aligned}$$

Эта же величина будет являться математическим ожиданием числа таких переходов за единичный временной интервал, если X – стационарный процесс.

Величину $\mathsf{\Pi Q}(A,B)$ называют *вероятностным потоком* между A и B. Тогда, в частности, $\mathsf{\Pi Q}(x,y)$ – есть средняя или равновесная интенсивность переходов из состояния x в y (при этом $q(x,y)$ называют *инфинитезимальной интенсивностью переходов*).

Уравнение *глобального баланса* (6.12) в свете всего вышесказанного (в терминах вероятностного потока) может быть переписано следующим образом:

$$\mathsf{\Pi Q}(x,\mathbb{E}) \ = \ \mathsf{\Pi Q}(\mathbb{E},x)\ , \quad x\in\mathbb{E}\ . \qquad (6.21)$$

Это означает, что при равновесии (в стационарном режиме) среднее число переходов в единицу времени из состояния x во все другие возможные состояния равно среднему числу переходов из всевозможных других, отличных от x, состояний в это состояние x. Проще (короче) говоря, интенсивность потока, исходящего из состояния x, равна интенсивности потока, входящего в x.

Суммируя уравнения глобального баланса для всех состояний $x\in A$, получим более общее балансовое уравнение

$$\mathsf{\Pi Q}(A,\mathbb{E}) \ = \ \mathsf{\Pi Q}(\mathbb{E},A)\ , \qquad (6.22)$$

или, исключив из обеих частей величину $\mathsf{\Pi Q}(A,A)$,

$$\mathsf{\Pi Q}(A,A^C) \ = \ \mathsf{\Pi Q}(A^C,A)\ , \qquad (6.23)$$

где A^C – обозначает дополнительное к A (или комплиментарное A) множество.

Таким образом, мы получили, что интенсивность потока из A равна интенсивности потока в A, что мы и должны были, естественно, ожидать для рассматриваемой стабильной системы.

Число заходов процесса X в A на интервале $(0,t]$ (называемое также *числом посещений*), может быть вычислено следующим образом:

$$N_A(t) \ \equiv \ \sum_{n}\mathbf{1}\big(X_{\tau_{n-1}}\in A^C,\ X_{\tau_n}\in A,\ \tau_n\in(0,t]\big)\ . \qquad (6.24)$$

Откуда (с вер. 1) интенсивность потока $\lambda(A)$, с которой процесс X входит в A, может быть определена через число посещений за единицу времени

$$\lambda(A) \ \equiv \ \lim_{t\to\infty}t^{-1}N_A(t)\ . \qquad (6.25)$$

Следовательно,

$$\lambda(A) \;=\; \Pi\mathsf{Q}(A^C,\; A) = \sum_{x\in A^C} \pi(x) \sum_{y\in A} q(x,y)\;. \tag{6.26}$$

Эта интенсивность связана также и с моментом T_n – n-го попадания процесса X в A. По закону больших чисел для точечных процессов очевидно, что

$$\lim_{n\to\infty} n^{-1} T_n \;=\; \lambda^{-1}(A) \;\text{ с вер. 1}\;. \tag{6.27}$$

Другой интересной величиной, характеризующей поведение процесса X, является среднее время его пребывания в A (или время ожидания), определяемое следующим образом:

$$W(A) \;\equiv\; \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} W_i(A) \;\text{ с вер. 1}\;, \tag{6.28}$$

где $W_i(A)$ – есть время, проведенное процессом X в A на его i-тый визит. Формулу для вычисления этой величины задает следующая Теорема

Теорема 6.2. *Если процесс X – эргодический со стационарным распределением π, то предельная величина $W(A)$ существует и*

$$W(A) \;=\; \lambda^{-1}(A) \sum_{x\in A} \pi(x)\;. \tag{6.29}$$

Доказательство.
Интересующее нас утверждение является простым следствием (6.17) из Теоремы 6.1 и равенства (6.27).
Действительно,

$$\begin{aligned} W(A) \;&=\; \lim_{n\to\infty} n^{-1} \int_0^{T_{n+1}} \mathbf{1}\big(X_t \in A\big)\; dt \\ &=\; \lim_{n\to\infty} n^{-1} T_{n+1} \cdot \lim_{n\to\infty} T_{n+1}^{-1} \int_0^{T_{n+1}} \mathbf{1}\big(X_t \in A\big)\; dt \\ &=\; \lambda^{-1}(A) \sum_{x\in A} \pi(x) \;\text{ с вер. 1}\;. \end{aligned}$$

Теорема доказана. □

6.4 Тандем, как простейший пример сети

Рассмотрим открытую сеть, состоящую из m последовательно соединенных узлов $1, 2, ..., m$ и предположим, что движущиеся в ней элементы поступают на вход узла 1 в виде пуассоновского потока с интенсивностью λ,

последовательно обслуживаются во всех узлах по порядку их подключения, а затем покидают эту систему. Будем предполагать также, что каждый узел представляет собой единичный сервер, обслуживающий элементы по одному в соответствии с дисциплиной $FIFO$, и что время обслуживания любого элемента в узле j распределено экспоненциально с параметром μ_j, причем независимо от входного процесса и от работы других серверов сети. Если по прибытию в узел элемент застает сервер занятым, то он встает в очередь и ожидает своего обслуживания. Таким образом, каждый узел представляет собой некоторую модель $M_\lambda|M_{\mu_j}|1|\infty$ системы МО со входом λ и одиночным сервером μ_j, причем предполагается, что все $\varrho_j = (\lambda/\mu_j) < 1,\ j = 1, 2, \ldots, m$. Состояние сети представляется вектором $x = (x_1, x_2, \ldots, x_m)$ на множестве $\mathbb{E} \equiv \{x : |x| < \infty\}$, где x_j – обозначает число элементов в узле j. Такая сеть называется *тандемом* и, как мы увидим далее, представляет собой простейший пример открытой сети Джексона.

Пусть X_t описывает состояние сети в момент t. Тогда сетевой процесс $X = \{X_t : t \geqslant 0\}$ представляет собой процесс в открытой сети с неограниченной ёмкостью, который эволюционирует во времени следующим образом: попав в состояние x, процесс X_t остается в этом состоянии до прихода нового элемента в узел 1, или до окончания обслуживания в каком-либо из узлов. Другими словами, типичными переходами в этой сети являются либо переход из x в $T_{01}x = x + e_1$ (приход элемента в узел 1 извне сети), либо переход из x в $T_{j,j+1}x = x - e_j + e_{j+1}$ (завершение обслуживания в $j-$том узле), возможный лишь при условии, что $x_j \geqslant 1$. Здесь, кроме того, предполагается, что узел $m + 1 \equiv 0$, т.е. что оператор перехода $T_{j,j+1} = T_{m0}$ при $j = m$. Из всего выше сказанного ясно, что переходные интенсивности рассматриваемого процесса будут равны

$$q(x, T_{01}x) = \lambda, \quad q(x, T_{j,j+1}x) = \mu_j \mathbf{1}(x_j \geqslant 1),$$

причем интенсивности всех других переходов тождественно равны 0.

Уравнение глобального баланса (6.12), которому должна удовлетворять инвариантная мера такого процесса, очевидно, примет здесь следующий вид

$$\pi(x) \sum_{j=0}^{m} q(x, T_{j,j+1}x) \ = \ \sum_{j=1}^{m+1} \pi(T_{j,j-1}x) q(T_{j,j-1}x, x) \mathbf{1}(x_j \geqslant 1),\ x \in \mathbb{E}, \quad (6.30)$$

где в правой части указаны только все те состояния, из которых возможен переход в состояние x за счет ненулевых переходных интенсивностей.

Вследствие независимости работы каждого из узлов рассматриваемой сети друг от друга можно предположить, что стационарное распределение рассматриваемого сетевого процесса будет равно произведению стационарных распределений всех его отдельных узлов, т.е. что

$$\pi(x) \ = \ \prod_{j=1}^{m} (1 - \varrho_j)(\varrho_j)^{x_j},\ \text{где}\ \varrho_j = \frac{\lambda}{\mu_j}\ ,\ \Big(\varrho_j < 1,\ j = 1, 2, \ldots, m\Big). \quad (6.31)$$

Чтобы проверить это предположение, заметим, что $\pi(x)$, определенное такой формулой, удовлетворяет следующей системе уравнений:

$$\begin{aligned} \pi(x)q(x,T_{01}x) &= \pi(T_{0m}x)q(T_{0m}x,x), \quad \big(j=0,\ m+1=0\big), \\ \pi(x)q(x,T_{j,j+1}x) &= \pi(T_{j,j-1}x)q(T_{j,j-1}x,x)\mathbf{1}(x_j\geqslant 1), \quad 1\leqslant j\leqslant m. \end{aligned} \tag{6.32}$$

Действительно, выпишем и сравним левую и правую части первого равенства в системе уравнений (6.32)

$$\begin{aligned} \pi(x)\ q(x,T_{01}x) &= \pi(x)\ \lambda\ ; \\ \pi(T_{0m}x)\ q(T_{0m}x,x) &= \big(\pi(x)\ \varrho_m\big)\big(\mu_m\big) \\ &= \pi(x)\ \lambda\ . \end{aligned}$$

Аналогично для второго равенства в (6.32) получим:

$$\begin{aligned} \pi(x)\ q(x,T_{j,j+1}x) &= \pi(x)\ \mu_j\ \mathbf{1}(x_j\geqslant 1), \quad 1\leqslant j\leqslant m\ ; \\ \pi(T_{j,j-1}x)\ q(T_{j,j-1}x,x)\ \mathbf{1}(x_j\geqslant 1) &= \big(\pi(x)\frac{\mathbf{1}(x_j\geqslant 1)}{\varrho_j}\varrho_{j-1}\big)\big(\mu_{j-1}\big)\big(\mathbf{1}(x_{j-1}\geqslant 1)\big) \\ &= \pi(x)\ \mu_j\ \mathbf{1}(x_j\geqslant 1), \quad 1\leqslant j\leqslant m\ . \end{aligned}$$

Если теперь почленно просуммировать все уравнения системы (6.32), то получим, что $\pi(x)$ в виде (6.31) удовлетворяет и уравнению (6.30) глобального баланса рассматриваемой сети. Значит, $\pi(x)$ является инвариантной мерой для рассматриваемого нами сетевого процесса X_t.

Заметим также, что каждое из уравнений системы (6.32) представляет собой уравнение, называемое *уравнением локального баланса* в соответствующем узле сети, которое означает, что среднее количество перемещений элементов за единичное время из узла j, выводящее сеть из некоторого состояния x, равно среднему числу перемещений элементов в этот же узел j, приводящих сеть в то же состояние x. Ясно, что для сети типа тандем такие перемещения на выходе из узла j возможны лишь в узел $(j+1)$, а при входе в узел j - только из узла $(j-1)$ при всех $j=1,2,\ldots,m$.

Мы только что показали (на примере сети типа тандем), что выполнение условий локального баланса во всех узлах сети является достаточным условием существования глобального баланса в сети. Этот факт оказывается справедливым и для других типов сетей, однако, обратное утверждение в общем случае не верно. Мультипликативный вид (6.31) стационарного распределения означает, что в этой сети каждый узел работает независимо от остальных, т.е. как отдельно взятая изолированная система МО.

ПРИМЕР (Тандем в стационарном режиме).
Тандем состоит только из двух узлов типа $M_\lambda|M_\mu|1|\infty$,
$\big(M=\{0,1,2\}; X=\{x_1,x_2\}; \mathbb{E}=\{x: x_1\geq 0, x_2\geq 0\}; \varrho_1=\lambda/\mu_1<1, \varrho_2=\lambda/\mu_2<1; \pi(x)=(1-\varrho_1)(1-\varrho_2)\varrho_1^{x_1}\varrho_2^{x_2}\ \big)$.
1) Найти долю времени, которую процесс X проводит в подмножестве состояний $A=\{x: x_1<x_2\}$, $\big(\pi(A)-?\big)$;

2) Найти среднее время пребывания в этом подмножестве состояний $\big(W(A)-?\big)$.

Р Е Ш Е Н И Е :

$\pi(A) = \sum_{x\in A}\pi(x)$;

$W(A) = \frac{1}{\lambda(A)}\sum_{x\in A}\pi(x) = \frac{\pi(A)}{\lambda(A)}$, где $\lambda(A)$ – поток в состояние A из состояния A^c: $\big(A^c = \mathbb{E}\setminus A; A^c\bigcup A = \mathbb{E}, A^c\bigcap A = \emptyset\big)$.

Пусть $x_1 = n_1, x_2 = n_2$.

Если $n_2 > n_1$, то $n_2 = n_1 + s; s = 1, 2, ...$

$$\begin{aligned}\pi(A) &= \sum_{n_1,n_2}\pi(n_1,n_2)1(n_1<n_2) = \sum_{n_1,s}(1-\varrho_1)(1-\varrho_2)\varrho_1^{n_1}\varrho_2^{n_1+s}\\ &= (1-\varrho_1)(1-\varrho_2)\sum_{n_1=0}^{\infty}(\varrho_1\varrho_2)^{n_1}\sum_{s=1}^{\infty}\varrho_2^s\\ &= (1-\varrho_1)(1-\varrho_2)\frac{1}{(1-\varrho_1\varrho_2)}\frac{\varrho_2}{(1-\varrho_2)}\\ &= \frac{(1-\varrho_1)\varrho_2}{1-\varrho_1\varrho_2}.\end{aligned}$$

$$\lambda(A) = \sum_{x\in A^c}\sum_{y\in A}\pi(x)q(x,y) = \sum_x \pi(x)1(x\in A^c)\sum_y q(x,y)1(y\in A),$$

где для $x\in\mathbb{E}$:

$$q(x,y) = \begin{cases} q(x,T_{jk}x), & y = T_{jk}x\in\mathbb{E}, \quad (j\neq k)\in M\ ;\\ 0, & \text{иначе}\ ,\end{cases}$$

В примере:

$$\begin{cases} q(x,T_{01}x) = \lambda;\\ q(x,T_{12}x) = \mu_1 1(x_1\geq 1);\\ q(x,T_{20}x) = \mu_2 1(x_2\geq 1);\\ 0\ , \quad \text{иначе}\ .\end{cases}$$

Разберемся с переходами:

(I) $q(x,T_{12}x) = \mu_1 1(x_1\geq 1)$, $x\in A^c$, $T_{12}x\in A$,

$A^c = \{(n_1,n_2) : n_1\geq n_2\}$, $\quad A = \{(n_1,n_2) : n_1 < n_2\}$.

Поток из A^c в A состоит только из двух следующих переходов:

$n_1 > n_2$ и $n_1 = n_2$.

До перехода

$n_1 = n_2+1$, $n_2 = n_2$; $\qquad$ $n_1 = n_2$, $n_2 = n_2$;

после перехода

$\widehat{n}_1 = n_2$, $\widehat{n}_2 = n_2+1$; $\qquad$ $\widehat{n}_1 = n_2-1$, $\widehat{n}_2 = n_2+1$.

В остальных случаях перехода $A^c \longrightarrow A$ не получается.

(II) $q(x, T_{20}x) = \mu_2 1(x_2 \geq 1)$, $x \in A^c$, $T_{20}x \in A$.

$A^c = \{(n_1, n_2) : n_1 \geq n_2\}$

$T_{20}x = x - e_2 + e_0$, т.е. $(n_1, n_2) \longrightarrow (n_1, n_2 - 1)$.

Если $x \in A^c$, то и $T_{20}x \in A^c$.

В результате проведенного анализа получаем:

$$\begin{aligned}
\lambda(A) &= (1-\varrho_1)(1-\varrho_2)\mu_1 \sum_{n_2=0}^{\infty} \varrho_1^{n_2+1}\varrho_2^{n_2} + (1-\varrho_1)(1-\varrho_2)\mu_1 \sum_{n_2=1}^{\infty} \varrho_1^{n_2}\varrho_2^{n_2} \\
&= (1-\varrho_1)(1-\varrho_2)\mu_1 \Big[\varrho_1 \sum_{n=0}^{\infty} (\varrho_1\varrho_2)^n + \sum_{n=1}^{\infty} (\varrho_1\varrho_2)^n\Big] \\
&= (1-\varrho_1)(1-\varrho_2)\mu_1 \Big(\frac{\varrho_1}{1-\varrho_1\varrho_2} + \frac{\varrho_1\varrho_2}{1-\varrho_1\varrho_2}\Big) \\
&= \frac{\lambda(1-\varrho_1)(1-\varrho_2)(1+\varrho_2)}{1-\varrho_1\varrho_2} = \frac{\lambda(1-\varrho_1)(1-\varrho_2^2)}{1-\varrho_1\varrho_2} \ .
\end{aligned}$$

Найдем теперь $W(A)$ – среднее время пребывания в состоянии A.

$$\begin{aligned}
W(A) &= \frac{\pi(A)}{\lambda(A)} = \frac{(1-\varrho_1)\varrho_2}{(1-\varrho_1\varrho_2)} \frac{(1-\varrho_1\varrho_2)}{\lambda(1-\varrho_1)(1-\varrho_2^2)} \\
&= \frac{\varrho_2}{\lambda(1-\varrho_2^2)} = \frac{1}{\mu_2(1-\frac{\lambda^2}{\mu_2^2})} = \frac{\mu_2}{\mu_2^2-\lambda^2}
\end{aligned}$$

Итак,

$$\pi(A) = \frac{(1-\varrho_1)\varrho_2}{1-\varrho_1\varrho_2}, \qquad W(A) = \frac{\varrho_2}{\lambda(1-\varrho_2^2)} = \frac{\mu_2}{\mu_2^2-\lambda^2}.$$

6.5 Сетевые процессы Джексона и Виттла

Мы показали, что скачкообразный марковский процесс $\{X_t : t \geqslant 0\}$ может описывать поведение некоторой сети, состоящей из m узлов, с движущимися между ними элементами, которым одномоментно разрешены лишь единичные перемещения. Координаты этого процесса отражают количество единичных элементов в каждом из m узлов сети. Основным допущением, которое мы уже приняли ранее, было предположение о том, что всякий раз, когда сеть попадает в некоторое состояние x, промежуток времени до следующего перемещения единичного элемента, например, из узла j в узел k, связанного с переходом сети из состояния x в другое состояние $T_{jk}x = x - e_j + e_k$, имеет экспоненциальное распределение.

Для дальнейшей конкретизации и детализации этой модели сети будем предполагать, что параметр экспоненциального распределения (его интенсивность) имеет следующий вид

$$\lambda_{jk}\phi_j(x) \ , \tag{6.33}$$

где $\lambda_{jk} \geqslant 0$ с $\lambda_{jj} = 0$, а $\phi_j(x) > 0$, за исключением случая $\phi_j(x) = 0$, когда $x_j = 0$ и $j \neq 0$.

Такое предположение об экспоненциальном распределении времен ожидания перемещений, очевидно, будет автоматически выполняться при следующих условиях:

(**I**) Каждый раз, когда сеть попадает в состояние x, промежуток времени до выхода из него за счет перемещения единичного элемента из данного узла j распределено экспоненциально с параметром $\phi_j(x)$.

(**II**) Всякий уходящий из некоторого узла j элемент направляется в узел k с вероятностью λ_{jk}, которая не зависит ни от чего другого, кроме данных j и k.

Следуя общепринятой договоренности о том, что λ_{jk} может представлять собой либо маршрутную вероятность, либо ненулевую интенсивность выбора узлов j и k (как интенсивности в процессах рождения-гибели), мы будем называть λ_{jk} либо *маршрутной $j - to - k$ интенсивностью*, либо *маршрутной таксой (rate)*. (Кроме того, рассматривая λ_{jk} как переходную вероятность некоторого скачкообразного марковского процесса с непрерывным временем, можно описать движение отдельно взятого единичного элемента по множеству M узлов сети и получить искусственный маршрутный процесс, отличный от сетевого процесса).

Величину $\phi_j(x)$ мы будем называть либо *сервисной ставкой (rate) за обслуживание*, либо *интенсивностью выхода из узла j* (по окончании обслуживания), в момент, когда сеть находится в состоянии x. Если сеть является открытой, то единичные элементы могут приходить в нее, например, через узел j в виде системно-зависимого пуассоновского потока с интенсивностью $\lambda_{0k}\phi_0(x)$, где $\phi_0(x)$ будет тогда являться *входной интенсивностью извне сети* (или из нулевого узла). Если $\phi_0(x) \equiv 1$, то приходящий извне сети в соответствующие ее узлы пуассоновский поток будет независимым пуассоновским процессом с интенсивностями $\lambda_{01}, \ldots, \lambda_{0m}$, причем нулевая интенсивность входа для какого-либо узла будет означать, что конфигурацией сети не предусмотрен вход в сеть извне через данный узел. Таким образом, мы можем индивидуально рассматривать величины λ_{jk} и $\phi_j(x)$ (как таксу, ставку или интенсивность), хотя каждая из них является лишь частью составной интенсивности (6.33). Мы будем далее называть эти величины "маршрутной" и "сервисной" интенсивностями соответственно, хотя каждая из них может иметь и отличную от этой интерпретацию.

В соответствии со всеми сделанными предположениями, сетевой процесс X – это марковский процесс с одноэлементными перемещениями и интенсивностью переходов

$$q(x,y) = \begin{cases} \lambda_{jk}\phi_j(x), & \text{если } y = T_{jk}x \in \mathbb{E} \text{ для } j \neq k \text{ из } M\ ; \\ 0 & \text{иначе .} \end{cases} \tag{6.34}$$

Иногда мы будем изображать эту переходную интенсивность в более компактном виде $q(x, T_{jk}x) = \lambda_{jk}\phi_j(x)$, предполагая, конечно, что $T_{jk}x \in \mathbb{E}$.

В дополнение к предположению об экспоненциальном распределении временных интервалов до перескоков, мы будем также предполагать, что сервисные интенсивности сбалансированы специальным образом.

Определение 6.1. *Сервисные интенсивности ϕ_j сетевого процесса X называются* **Φ-сбалансированными**, *если для некоторой положительной функции Φ (определенной на $\mathbb{E}$ и такой, что для каждого $x \in \mathbb{E}$ и для любых j, k из M, с $T_{jk}x \in \mathbb{E}$) эти сервисные интенсивности удовлетворяют следующему равенству*

$$\Phi(x)\phi_j(x) \;=\; \Phi(T_{jk}x)\phi_k(T_{jk}x)\;. \tag{6.35}$$

Как мы увидим ниже, условие (6.35) является естественным требованием, при выполнении которого формулы, задающие стационарное распределение сетевого процесса, приобретают более простой вид.

В завершение описания моделей сетевых процессов дадим теперь строгое определение двух из них, выделяющихся среди различных других тем, что их стационарные распределения могут быть записаны в замкнутой форме.

Определение 6.2. *Марковский сетевой процесс X с интенсивностью переходов (6.34) и Φ-сбалансированными (6.35) сервисными интенсивностями называется процессом Виттла (P.Whittle).*
Этот процесс является процессом Джексона (J.Jackson), если его сервисные интенсивности $\phi_j(x) = \phi_j(x_j)$, оказываются функциями только x_j для каждого $j = 1, \ldots, m$, а в случае открытой сети и $\phi_0(\cdot) \equiv 1$.

Процессы Джексона и Виттла удобно рассматривать параллельно, поскольку их поведение имеет достаточно много общих черт. Подчеркнем еще раз, что у процессов Джексона сервисная интенсивность в узле j является функцией x_j – количества элементов, находящихся только в этом узле, указывая на независимость работы узлов друг от друга.
Нетрудно проверить, что такие интенсивности обслуживания балансируются (в смысле определения (6.35)) следующей положительной функцией

$$\Phi(x) \;=\; \prod_{j=1}^{m}\prod_{n=1}^{x_j}\phi_j^{-1}(n)\;. \tag{6.36}$$

Проверку этого факта рекомендуется произвести в качестве домашнего упражнения.

В процессах Виттла сервисные интенсивности узлов являются системно-зависимыми (зависят от всего вектора x), указывая на зависимость работы узлов друг от друга. Следует упомянуть также, что сети Джексона носят свое название исторически по имени автора, впервые представившего эту

модель в печати в 1957 году. Имя Виттла употребляется в названии сетей в связи с тем, что именно этот автор внес наиболее значимый вклад в исследование модели сети с системно-зависимыми интенсивностями переходов.

Мы будем предполагать далее, что рассматриваемый сетевой процесс X является либо процессом Виттла, либо процессом Джексона, определенных выше. Мы будем особо указывать на тип процесса только в тех случаях, когда конкретный результат будет относится только к процессам Джексона. Сначала рассмотрим некоторые примеры путей, сервисов и маршрутизацию в процессах Виттла X. Так как этот процесс является марковским, каждое его время пребывания в состоянии x распределено экспоненциально с интенсивностью

$$\sum_j \sum_k q(x, T_{jk}x) \;=\; \sum_j \phi_j(x) \sum_k \lambda_{jk} \;. \tag{6.37}$$

К тому же, когда сеть находится в состоянии x, время до "потенциального" перехода в ее другое состояние за счет ухода единичного элемента из узла j (минимум из всех возможных времен уходов в узлы $k \neq j$) распределено экспоненциально с интенсивностью $\phi_j(x) \sum_k \lambda_{jk}$. Это следует из того факта, что минимум из набора независимых экспоненциально распределенных случайных величин распределен экспоненциально с интенсивностью, равной сумме всех представленных в этом наборе интенсивностей. Этот факт легко может быть получен из (2.36).

По окончании времени пребывания в состоянии x, сетевой процесс скачком переходит в состояние $T_{jk}x \in \mathbb{E}$ с вероятностью

$$p_{jk} \;=\; \frac{q(x, T_{jk}x)}{\sum_l q(x, T_{jl}x)} \;=\; \frac{\lambda_{jk}}{\sum_l \lambda_{jl}} \;, \quad j, k \in M \;. \tag{6.38}$$

Отметим, что вероятность (6.38) не зависит от $\phi_j(x)$ и от состояния x. Матрица $||p_{jk}||$ – является матрицей марковской цепи с диагональными элементами $p_{jj} = 0$. Мы будем называть p_{jk} *путевыми вероятностями* процесса X. Эти вероятности представляют собой условные вероятности того, что единичный элемент переместится из узла j в узел k, вычисленную при условии, что он действительно покинул узел j. Так как выражение (6.38) не зависит от x, мы можем рассматривать все единичные элементы, покидающие узел j, как независимо и одинаково маршрутизируемые в соответствии с заданными вероятностями $p_{jk},\ k \in M$.

Легко показать, что наше предположение $\lambda_{jj} = 0$ вовсе не исключает из рассмотрения возможности возврата единичного элемента обратно в узел j сразу же по окончании его обслуживания в нем. Для учета этой ситуации нужно лишь провести небольшую модификацию, то есть преобразования, совершенно аналогичные уже рассматривавшимся нами выше в примере (6.5)-(6.8) при описании функционирования узлов с "обратной связью". Действительно, для процесса Виттла X, для которого предполагается, что как только он попадает в состояние x, время до следующего перехода в

другое состояние за счет ухода элемента из узла j распределено экспоненциально с интенсивностью $\phi_j(x)$, если существует отличная от нуля вероятность немедленного возврата ушедшего из j элемента обратно в j, то есть $\widetilde{p}_{jj} \geqslant 0$, то его переходная вероятность должна быть просто определена как $q(x, T_{jk}x) = \widetilde{p}_{jk}\phi_j(x)$, а интенсивность экспоненциального времени его пребывания в состоянии x будет тогда равна $\sum_j \phi_j(x)(1 - \widetilde{p}_{jj})$.

Рассмотрим теперь более подробно некоторые типы интенсивности обслуживания, а также их интерпретации.
Будем говорить, что узел j содержит s экспоненциальных серверов с интенсивностью μ_j, если

$$\phi_j(x_j) \;=\; \mu_j \;\min\{x_j, s\}\;, \quad x_j \geqslant 1\;. \tag{6.39}$$

Это обычно относится к случаю, когда имеется s, $1 \leqslant s \leqslant \infty$ независимых, параллельно работающих одиночных серверов, а их времена обслуживания (например, времена, требующиеся для обработки одного пакета) независимы и распределены экспоненциально с интенсивностью μ_j. Такой узел работает независимо от других узлов сети в том смысле, что его интенсивность обслуживания не зависит от x_k, $k \neq j$. При этом допустимы самые различные дисциплины обслуживания, поскольку конкретные элементы (пакеты) в рассматриваемой нами модели неразличимы друг от друга (все они гетерогенны, т.е. одинаковые). При этом случай, когда $\phi_j(x_j) = \mu_j x_j$ (когда $s = \infty$) означает, что узел j представляет собой просто точку задержки, причем для каждого пакета его конкретная задержка является независимой.

Другой интерпретацией интенсивности обслуживания ϕ_j является представление с её помощью схемы разделения времени процессора (time shearing), при которой элементы обрабатываются следующим образом: в любой момент, когда в узле j присутствуют x_j элементов, время до очередного убытия некоторого i-го элемента из этого узла распределено экспоненциально с интенсивностью $\mu_i(x_j) > 0$ такой, что

$$\sum_{i=1}^{x_j} \mu_i(x_j) \;=\; \phi_j(x_j)\;. \tag{6.40}$$

Это означает, что все элементы (из имеющихся в узле x_j штук) одновременно получают обслуживание, причем каждый i-тый получает только свою долю μ_i этого обслуживания. В такой модели возможно описать и эголитарное (равноправное) разделение времени обслуживания, для чего следует положить

$$\mu_i(x_j) \;=\; \frac{\phi_j(x_j)}{x_j}\;. \tag{6.41}$$

При этом каждый клиент (единичный элемент, или пакет) получит одинаковую для всех них долю от $\phi_j(x_j)$, а полная интенсивность выхода из узла,

вне зависимости от конкретного правила обслуживания, все равно будет равна $\phi_j(x_j)$. Нужно иметь в виду, что функция ϕ_j при разделении процессорного времени обслуживания может принимать любую форму. Например, $\phi_j(x_j) = \mu_j$ может быть долей интенсивности при процессорном разделении, а может и просто представлять собой интенсивность единичного сервера, работающего с дисциплиной обслуживания типа $FIFO$.

В заключение, приведем критерий неприводимости сетевого процесса X. Напомним, что путевой процесс, соответствующий сетевому процессу X, является марковским процессом с переходными интенсивностями λ_{jk}. Последовательность состояний, которые посещает этот путевой процесс, образует марковскую цепь, переходными вероятностями которой служат маршрутные вероятности p_{jk}, равные

$$p_{jk} \;=\; \frac{\lambda_{jk}}{\sum_{k'} \lambda_{jk'}} \;. \tag{6.42}$$

Теорема 6.3. *Процессы Джексона и Виттла неприводимы тогда и только тогда, когда соответствующий путевой процесс является неприводимым.*

Доказательство.

Сначала предположим, что X является неприводимым. Чтобы доказать, что тогда соответствующий ему путевой процесс также оказывается неприводимым, достаточно показать, что для любых фиксированных $j \neq k$ из M существует последовательность $j_1, j_2, \ldots, j_l$ в M такая, что

$$\lambda_{jj_1}\lambda_{j_1j_2}\cdots\lambda_{j_lk} \;>\; 0 \;. \tag{6.43}$$

Выберем некоторые x и $\widetilde{x}$ в $\mathbb{E}$ такие, чтобы x_j и $\widetilde{x}_k$ были положительными. Неприводимость процесса X гарантирует существование некоторой последовательности $j_1, j_2, \ldots, j_l$ в M такой, что следующие состояния сетевого процесса:

$$x,\; x^1 = T_{jj_1}x,\; \ldots,\; x^l = T_{j_{l-1}j_l}x^{l-1},\; \widetilde{x} = T_{j_lk}x^l, \tag{6.44}$$

образуют реально осуществимый путь из состояния x в $\widetilde{x}$, а тогда с необходимостью

$$q(x, x^1)q(x^1, x^2)\cdots q(x^l, \widetilde{x}) \;>\; 0 \;. \tag{6.45}$$

Практический выбор этих состояний предполагает сначала нахождение таких j_1 и j_l, чтобы $q(x, x^1)$ и $q(x^l, \widetilde{x})$ были положительны, с последующим выбором $x^2, \ldots, x^{l-1}$ таких, чтобы $x^1, x^2, \ldots, x^{l-1}, x^l$ были бы реально осуществимым путем из состояния x^1 в x^l и т.д.
Поскольку все включенные в (6.45) величины ϕ_j-тые должны быть положительными, то из (6.45) с необходимостью следует желаемое (6.43).

Теперь, наоборот, предположим, что путевой процесс (марковская цепь) является неприводимой. Зафиксируем некоторые x и $\widetilde{x}$, $x \neq \widetilde{x}$ в $\mathbb{E}$. Выберем

теперь любые $j \neq k$ из M такие, чтобы их координаты x_j и $\widetilde{x}_k$ были положительными и выделим последовательность $j_1, \ldots, j_l$ в M, так чтобы удовлетворялось (6.43). Это возможно вследствие сделанного предположения. Далее рассмотрим состояния сетевого процесса X, определенные в (6.44). Тогда (6.43) и положительность величин ϕ_j-тых автоматически приведет к выполнению (6.45). А это означает, что процесс X может достичь любое состояние $\widetilde{x}$ из произвольного своего начального состояния x, что означает его неприводимость и завершает доказательство.
Теорема доказана полностью. □

6.6 Нахождение стационарных распределений

В дополнение к введенным выше обозначениям, рассмотрим еще положительную инвариантную меру $\omega_j, \quad j \in M$, удовлетворяющую следующим *уравнениям маршрутного баланса* или *уравнениям трафика*:

$$\omega_j \sum_{k \in M} \lambda_{jk} \;=\; \sum_{k \in M} \omega_k \lambda_{kj} \;, \quad j \in M \;. \tag{6.46}$$

Для открытой сети примем дополнительное соглашение $\omega_0 = 1$, что упростит вид некоторых выражений.

Существование этой инвариантной меры гарантируется тем, что множество M – конечное и путевой процесс не имеет неустановившихся (промежуточных) состояний.

Для замкнутой сети возможно желание нормировать ω, чтобы получить вероятностное распределение. И тогда мы получили бы инвариантное распределение для путевых (маршрутных) интенсивностей λ_{jk}, а также и для маршрутных вероятностей (6.42).

Следующие результаты описывают равновесное поведение сетевого процесса Джексона. Для процесса Джексона мы предполагаем, что матрица $||\lambda_{jk}||$ неприводима. Как мы доказали выше (Теорема 6.3), в этом случае и сам сетевой процесс X тоже будет неприводимым.

Теорема 6.4. *Если X – сетевой процесс замкнутой сети Джексона, то он является эргодическим, а его стационарное распределение имеет вид*

$$\pi(x) \;=\; C \prod_{j=1}^{m} \omega_j^{x_j} \prod_{n=1}^{x_j} \phi_j^{-1}(n) \;, \quad x \in \mathbb{E} = \{x : |x| = \nu\} \;, \tag{6.47}$$

где предполагается, что величины ω_j удовлетворяют уравнению трафика (6.46), а константа нормировки C определяется формулой

$$C^{-1} \;=\; \sum_{x \in \mathbb{E}} \prod_{j=1}^{m} \omega_j^{x_j} \prod_{n=1}^{x_j} \phi_j^{-1}(n) \;. \tag{6.48}$$

Теорема 6.5. *Если X – процесс открытой сети Джексона с конечной емкостью ν, то для него тоже справедливы утверждения Теоремы 6.4, но с $\mathbb{E} = \{x : |x| \leqslant \nu\}$.*

Теорема 6.6. *Если X – процесс открытой сети Джексона с неограниченной емкостью ν, то он имеет инвариантную меру в форме (6.47) с $\mathbb{E} = \{x : |x| < \infty\}$. Этот процесс будет положительно рекуррентным тогда и только тогда, когда*

$$C_j^{-1} \equiv \sum_{x_j=0}^{\infty} \omega_j^{x_j} \prod_{n=1}^{x_j} \phi_j^{-1}(n) \; < \; \infty \; , \quad j = 1, \ldots, m \; . \tag{6.49}$$

В этом случае его стационарное распределение будет иметь вид

$$\pi(x) \; = \; \pi_1(x_1) \cdots \pi_m(x_m) \; , \quad x \in \mathbb{E} \; , \tag{6.50}$$

где

$$\pi_j(x_j) \; = \; C_j \; \omega_j^{x_j} \prod_{n=1}^{x_j} \phi_j^{-1}(n) \; . \tag{6.51}$$

Все три сформулированные выше теоремы для сетевого процесса Джексона являются непосредственным следствием справедливости более общего результата, который мы докажем для сетевого процесса Виттла, у которого, как известно, интенсивности обслуживания $\phi_j = \phi_j(x)$ – системнозависимы.

Теорема 6.7. *Инвариантной мерой процесса Виттла X является*

$$\pi(x) \; = \; \Phi(x) \prod_{j=1}^{m} \omega_j^{x_j} \; , \quad x \in \mathbb{E} \; , \tag{6.52}$$

где для ω_j выполняется (6.46), а $\Phi(x)$ представляет собой балансирующую функцию интенсивностей обслуживания $\phi_j(x)$ в соответствии с определением (6.35). Указанная мера π, кроме того, удовлетворяет следующим **уравнениям частичного баланса**

$$\pi(x) \sum_{k \in M} q(x, T_{jk}x) \; = \; \sum_{k \in M} \pi(T_{jk}x) \; q(T_{jk}x, x) \; , \quad j \in M, \quad x \in \mathbb{E} \; . \tag{6.53}$$

Доказательство.
Так как рассматриваемый сетевой процесс Виттла X является скачкообразным марковским процессом с одноэлементными перемещениями, его инвариантная мера должна удовлетворять следующему уравнению глобального баланса

$$\pi(x) \sum_{j \in M} \sum_{k \in M} q(x, \; T_{jk}x) \; = \; \sum_{j \in M} \sum_{k \in M} \pi(T_{jk}x) \; q(T_{jk}x, \; x) \; , \quad x \in \mathbb{E} \; . \tag{6.54}$$

Как нетрудно заметить, это уравнение представляет собой сумму (по j) уравнений (6.53). Поэтому любая мера, удовлетворяющая (6.53), будет удовлетворять и (6.54), то есть окажется инвариантной мерой для сетевого процесса Виттла X.

Таким образом, для доказательства теоремы нам достаточно показать, что мера π из (6.52) удовлетворяет уравнению (6.53). Для этого зафиксируем некоторое $j \in M$ и любое состояние $x \in \mathbb{E}$. Если при этом оказалось $x_j = 0$, то обе части уравнения (6.53) окажутся равными нулю, так как в этом случае для любого k состояние $T_{jk}x \notin \mathbb{E}$ (см. (6.34)).

Поэтому далее предположим, что $x_j > 0$. С использованием определения величин q и ω_j преобразуем левую часть (6.53) следующим образом

$$\begin{aligned} \pi(x) \sum_{k \in M} q(x, T_{jk}x) &= \pi(x)\ \phi_j(x) \sum_{k \in M} \lambda_{jk} \\ &= \pi(x)\ \phi_j(x)\ \omega_j^{-1} \sum_{k \in M} \omega_k\ \lambda_{kj} \end{aligned} \tag{6.55}$$

Заметим далее, что выражение (6.52) для $\pi(x)$ и свойство $\Phi-$баланса (6.35) позволяют нам написать следующую цепочку равенств

$$\begin{aligned} \pi(x)\phi_j(x)\omega_j^{-1}\omega_k &= \Phi(x) \prod_{i=1}^{m} \omega_i^{x_i} \phi_j(x) \omega_j^{-1} \omega_k \\ &= \Phi(T_{jk}x)\phi_k(T_{jk}x)\ \omega_1^{x_1} \cdots \omega_j^{x_j - 1} \cdots \omega_k^{x_k + 1} \cdots \omega_m^{x_m} \\ &= \pi(T_{jk}x)\ \phi_k(T_{jk}x)\ , \quad k \in M\ . \end{aligned}$$

Подставляя последнее равенство в (6.55), получим окончательно, что

$$\pi(x) \sum_{k \in M} q(x, T_{jk}x) = \sum_{k \in M} \pi(T_{jk}x)\ \phi_k(T_{jk}x)\ \lambda_{kj} = \sum_{k \in M} \pi(T_{jk}x)\ q(T_{jk}x, x)\ . \tag{6.56}$$

Тем самым, теорема доказана. □

Литература

[1] Vladimir V. Kalashnikov (Калашников В.В.) (1994) *Mathematical methods in Queueing Theory*. Kluwer Acad. Publ., Dordrecht. (ISBN 0 7923 2568 0).

[2] Хинчин А.Я. (1963) *Работы по математической теории массового обслуживания.* - М.: Физматгиз.

[3] Клейнрок Л. (L. Kleinrok) (1979) *Теория массового обслуживания I ,* - М.: Машиностроение.

[4] Гнеденко Б.В., Коваленко И.Н. (1987) *Введение в теорию массового обслуживания.* - М.: Наука.

[5] Кокс Д.Р., Смит У.Л. (D.R. Cox and W.L. Smith) (1966) *Теория очередей.* - М.: Мир.

[6] Ивченко Г.И., Каштанов В.А., Коваленко И.Н. (1982) *Теория массового обслуживания.* - М.: Высшая школа.

[7] Brian D. Bunday (1996) *An Introduction to Queueing Theory.* - Bristol, UK: J.W. Arrowsmith Ltd. (ISBN 0 340 66239 5).

[8] Gunter Bolch et.al. (1998) *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications.* - John Wiley & Sons Inc. NY. (ISBN 0 471 19366 6).

[9] Evsei Morozov (2001) *Elements of Queueing Theory with Application to Communication Networks.* - Institute for Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences.

[10] Richard Serfozo (1999) *Introduction to Stohastic Networks.* - Springer-Verlag, N.Y. (ISBN 0-387-98773-8).

(serebrovski@mail.ru) (495)-602-79-37