Известия Гомельского государственного университета
имени Ф. Скорины, № 6 (135), 2022

УДК 519.872

О конструктивном описании времени обслуживания и входящего потока

Ю.В. Малинковский

Статья посвящена определенным особенностям конструктивного (параметрического) подхода при описании промежутков времени между моментами поступления запросов и длительностей их обслуживания.
Ключевые слова: длительности обслуживания, промежутки времени между моментами поступления запросов, конструктивное описание, системы массового обслуживания, составной прибор.

The paper deals with some properties of constructive (parametric) approach for description of time intervals between arrival moments and service times.
Keywords: service times, time intervals between arrival moments, constructive description, queueing systems, composite server.

Конструктивное или параметрическое задание длительностей обслуживания и промежутков времени между моментами поступления запросов часто встречается в современных литературных источниках по теории массового обслуживания. Однако в подавляющей массе таких источников изложение материала ведется на нечетком жаргонном языке, что создает определенные сложности для исследования, особенно у начинающих специалистов. Целью настоящей статьи является достаточно понятное научно-методическое обоснование конструктивного описания и преодоление жаргонных штампов в изложении.

Рассмотрим, например, систему массового обслуживания $M/M/s$ с простейшим входящим потоком с интенсивностью λ и s экспоненциальными приборами с интенсивностью обслуживания μ для каждого из них. Представим работающие приборы как единое целое – один прибор, который назовем составным. Чтобы отличать его от имеющихся в системе s приборов, будем называть их простыми приборами.

Пусть $\xi(t)$ – число запросов в системе в момент времени t, v_t – виртуальное время обслуживания запроса составным прибором, т. е. v_t – время обслуживания запроса, начинающееся в момент t при условии, что в этот момент в системе имеются запросы, т. е. $\xi(t) \neq 0$. Если v – время обслуживания запроса составным прибором, то в силу сказанного

$$B(x) = P\{v < x\} = P\{v_t < x \,/\, \xi(t) \neq 0\} = F_{v_t}(x \,/\, \xi(t) \neq 0).$$

Предположим, что в момент t работает составной прибор, тогда в силу отсутствия «памяти» у показательного распределения не имеет значения, сколько времени работают имеющиеся простые приборы.

Пусть $v_t(1), v_t(2), \ldots, v_t(s)$ – остаточные времена обслуживания запросов простыми приборами и пусть $x \geq 0$. Остаточное и виртуальное остаточные времена обслуживания составным прибором обозначим теми же буквами, что и полные времена обслуживания, но с тильдой.

Тогда

а) при $1 \leq n \leq s$

$$P\{\tilde{v} \geq x/\xi(t) = n\} = P\{\tilde{v}_t \geq x/\xi(t) = n\} = P\{\min(v_t(i_1), \ldots, v_t(i_n)) \geq x\} =$$
$$= P\{v_t(i_1) \geq t, \ldots, v_t(i_n) \geq x\} = e^{-\mu x} \ldots e^{-\mu x} = e^{-n\mu x},$$

т. е. условное распределение остаточного времени обслуживания запроса составным прибором $\tilde{v}$ при условии, что $\xi(t) = n$, – показательное с параметром $\mu(n) = n\mu$;

б) при $n > s$

$$P\{\tilde{v} \geq x/\xi(t) = n\} = P\{\tilde{v}_t \geq x/\xi(t) = n\} = P\{\tilde{v}_t \geq x/\xi(t) = n\} =$$
$$= P\{\min(v_t(1), \ldots, v_t(s)) \geq x\} = P\{v_t(1) \geq x, \ldots, v_t(s) \geq x\} =$$
$$= e^{-\mu x} \ldots e^{-\mu x} = e^{-s\mu x},$$

т. е. условное распределение остаточного времени обслуживания запроса составным прибором – показательное с параметром $\mu(n) = s\mu$. Объединяя а) и б) и учитывая, что если остаточное время жизни имеет показательное распределение с некоторым параметром, то и полное время жизни имеет показательное распределение с тем же параметром, получаем, что для $x \geq 0$

$$B(x / n) = P\{v < x / \xi(t) = n\} = 1 - e^{-\mu(n)x},$$

где

$$\mu(n) = \begin{cases} n\mu, & \text{если} \quad 1 \leq n \leq s \\ s\mu, & \text{если} \quad n > s \end{cases}. \qquad (1)$$

Здесь при $n = 0$ распределение случайной величины v_t не определено, так как обслуживание не может начинаться в момент, когда в системе нет заявок. Тем не менее, как будет видно из дальнейшего, удобно доопределить $\mu(n)$ в нуле, считая, что $\mu(0) = 0$.

Таким образом, условное распределение времени обслуживания запроса составным прибором – показательное с параметром $\mu(n)$, определяемым равенством (1).

Отметим, что конструктивное задание длительности обслуживания с помощью условного распределения времени обслуживания эквивалентно традиционному заданию, поскольку оба задания приводят к одному и тому же (в широком смысле) марковскому процессу. Это следует из того, что инфинитезимальные характеристики этих процессов одинаковы. Итак, конечномерные распределения процессов совпадают, а, следовательно, совпадают их стационарные распределения. Но при этом конструктивный подход удобнее, так как интенсивности переходов $n \to n-1$ (скачков вниз процесса $\xi(t)$) уже заданы с помощью (1), а при традиционном подходе их надо находить. К тому же при конструктивном подходе с помощью абстрактного задания $\mu(n)$ (не обязательно в виде (1)) можно описывать многие реальные и теоретически задаваемые системы массового обслуживания. При этом реальное распределение времени обслуживания, как правило, – не экспоненциальное.

Так, в рассмотренном примере, если выполнено условие эргодичности $\lambda < s\mu$, $\{p_n, n = 0,1,\ldots\}$ – стационарное распределение числа запросов в системе $M / M / s$, начальное распределение совпадает со стационарным, а $p_{>s} = \sum_{n=s+1}^{\infty} p_n$, то функция распределения времени обслуживания составным прибором v

$$B(x) = P\{v < x\} = P\{v_t < x / \xi(t) \neq 0\} = \sum_{n=1}^{\infty} P\{v_t < x / \xi(t) = n\} P\{\xi(t) =$$

$$= n / \xi(t) \neq 0\} = \sum_{n=1}^{\infty} \left[1 - e^{-\mu(n)x}\right] \frac{P\{\xi(t) = n\}}{P\{\xi(t) \neq 0\}} =$$

$$= \sum_{n=1}^{s} \frac{p_n}{1 - p_0} \left[1 - e^{-n\mu x}\right] + \frac{p_{>s}}{1 - p_0} \left[1 - e^{-s\mu x}\right].$$

Это – взвешенная линейная комбинация $s+1$ показательных распределений, т. е. безусловное распределение времени обслуживания фиктивным прибором (а не условное) является гиперэкспоненциальным.

Абстрагируемся теперь от конкретного вида $\mu(n)$. Рассмотрим систему, которую будем обозначать $M / M(n) / 1$, а время обслуживания задается параметрически. В однолинейную систему поступает простейший поток с интенсивностью λ. Условное распределение времени обслуживания при условии $\xi(t) = n\ (n \neq 0)$ является показательным с параметром $\mu(n)$, зависящим от числа заявок n в системе.

Лемма. *Для того чтобы условное распределение времени обслуживания при условии* $\xi(t) = n\ (n \neq 0)$ *являлось показательным с параметром* $\mu(n)$, *необходимо и достаточно, чтобы при* $\Delta t \downarrow 0$

$$P\{t \leq v < t + \Delta t / v \geq t, \xi(t) = n\} = \mu(n)\Delta t + o(\Delta t).$$

Доказательство. *Необходимость.* Пусть условное распределение v – показательное с параметром $\mu(n)$.Тогда

$$P\{t \le v < t+\Delta t \,/\, v \ge t, \xi(t)=n\} = \frac{P\{t \le v < t+\Delta t \,/\, \xi(t)=n\}}{P\{v \ge t \,/\, \xi(t)=n\}} =$$

$$= \frac{[1-e^{-\mu(n)(t+\Delta t)}]-[1-e^{-\mu(n)t}]}{1-[1-e^{-\mu(n)t}]} = 1-e^{-\mu(n)\Delta t} = \mu(n)\Delta t + o(\Delta t).$$

Достаточность. Пусть для $t \ge 0$ $P\{t \le v < t+\Delta t \,/\, v \ge t, \xi(t)=n\} = \mu(n)\Delta t + o(\Delta t).$

Если $B(t/n)$ – условная функция распределения времени обслуживания (при условии $\xi(t)=n$), то отсюда следует, что

$$\frac{B(t+\Delta t/n)-B(t/n)}{1-B(t/n)} = \mu(n)\Delta t + o(\Delta t).$$

Разделив обе части на Δt и устремив Δt к нулю, получим

$$\frac{B'(t/n)}{1-B(t/n)} = \mu(n),$$

откуда

$$1-B(t/n) = Ce^{-\mu(n)t}.$$

Поскольку $B(0/n)=0$, то $C=1$, откуда

$$B(t/n) = 1-e^{-\mu(n)t} \qquad для \quad t \ge 0.$$

Это означает, что условное распределение v при условии $\xi(t)=n$ – показательное с параметром $\mu(n)$.

Смысл леммы состоит в том, что если в данный момент времени прибор занимается обслуживанием запроса, а в системе находится n запросов, то вероятность того, что в течение ближайших Δt единиц времени ее обслуживание завершится, равна $\mu(n)\Delta t + o(\Delta t)$. Поэтому вероятность того, что обслуживание этого запроса не завершится в течение ближайших Δt единиц времени, равна $b_0(\Delta t) = 1-\mu(n)\Delta t + o(\Delta t)$. Отметим, что если $\xi(t) \ge 2$ в некоторый момент времени t, то вероятность того, что в промежутке времени $[t, t+\Delta t)$ будет обслужено не меньше двух запросов, есть $o(\Delta t)$ при $\Delta t \to 0$. Действительно, если в этом промежутке будет обслужено не меньше двух запросов, то найдется такое $c \in [t, t+\Delta t)$ в нем, что в каждом из промежутков $[t, t+c)$ и $[t+c, t+\Delta t)$ завершится обслуживание запроса, а по лемме вероятность этого события равна $[\mu(\xi(t))(c-t)+o(c-t)][\mu(\xi(c))(t+\Delta t-c)+o(t+\Delta t-c)] = o(\Delta t)$.

Составим уравнения Колмогорова для безусловных вероятностей состояний $P_n(t) = P\{\xi(t)=n\}$ процесса $\xi(t)$. По формуле полной вероятности при $\Delta t \to 0$

$$P_0(t+\Delta t) = P\{\xi(t+\Delta t)=0\} = \sum_{k=0}^{\infty} P\{\xi(t)=k\}P\{\xi(t+\Delta t)=0 \,/\, \xi(t)=k\} =$$

$$= P\{\xi(t)=0\}P\{\xi(t+\Delta t)=0 \,/\, \xi(t)=0\} + P\{\xi(t)=1\}P\{\xi(t+\Delta t) =$$

$$= 0 \,/\, \xi(t)=1\} + \sum_{k=2}^{\infty} P\{\xi(t)=k\}P\{\xi(t+\Delta t)=0 \,/\, \xi(t)=k\}.$$

Покажем, что последняя сумма есть $o(\Delta t)$. Действительно, эта сумма

$$\sum_{k=2}^{\infty} P\{\xi(t)=k,\, \xi(t+\Delta t)=0\} = P\{\xi(t)\ge 2,\, \xi(t+\Delta t)=0\} =$$

$$= P\{\xi(t)\ge 2\}P\{\xi(t+\Delta t)=0 \,/\, \xi(t)\ge 2\} \le P\{\xi(t+\Delta t)=0 \,/\, \xi(t)\ge 2\} \le o(\Delta t),$$

так как последняя вероятность не превосходит вероятности того, что за время Δt будет обслужено не меньше двух запросов (при условии, что сначала в системе было по крайней мере 2 запроса).

С учетом смысла леммы получим

$$P_0(t+\Delta t) = P_0(t)(1-\lambda\Delta t + o(\Delta t)) + P_1(t)(1-\lambda\Delta t + o(\Delta t))(\mu(1)\Delta t + o(\Delta t)) + 0(\Delta t),$$

откуда стандартным образом получаем

$$P_0'(t) = -\lambda P_0(t) + \mu(1)P_1(t).$$

Аналогично, для отличных от 0 состояний находим

$$P_n(t+\Delta t) = P_{n-1}(t)(\lambda\Delta t + o(\Delta t))(1-\mu(n-1)\Delta t + o(\Delta t)) + P_n(t)(1-\lambda\Delta t +$$
$$+o(\Delta t))(1-\mu(n)\Delta t + o(\Delta t)) + P_{n+1}(t)(1-\lambda\Delta t + o(\Delta t))(\mu(n+1)\Delta t + o(\Delta t)) + o(\Delta t),$$

откуда

$$P_n'(t) = -\lambda P_{n-1}(t) + (\lambda + \mu(n))P_n(t) + \mu(n+1)P_{n+1}(t), \qquad n = 1, 2, \ldots.$$

Итак, $\xi(t)$ – процесс размножения и гибели с постоянной интенсивностью рождения λ и переменной интенсивностью гибели $\mu(n)$.

Допустим, что процесс $\xi(t)$ эргодический, тогда существует предельное распределение $\{p_n\}$, являющееся единственным стационарным распределением. Выберем его в качестве начального распределения, тогда распределение цепи $\xi(t)$ в произвольный момент времени $\{P_n(t), n = 0,1,\ldots\}$ совпадет с ним и не будет зависеть от времени. Поэтому производные по времени обратятся в 0, а последние уравнения превратятся в следующие уравнения равновесия:

$$\lambda p_0 = \mu(1)p_1,$$
$$(\lambda + \mu(n))p_n = \lambda p_{n-1} + \mu(n+1)p_{n+1}.$$

Для процесса размножения и гибели уравнения равновесия эквивалентны уравнениям равновесия для вертикальных сечений

$$\lambda p_{n-1} = \mu(n)p_n, \qquad n = 1, 2, \ldots,$$

откуда легко находится стационарное распределение

$$p_n = p_0 \prod_{k=1}^{n} \frac{\lambda}{\mu(k)}, \tag{2}$$

где из условия нормировки

$$p_0 = \left(1 + \sum_{n=1}^{\infty} \prod_{k=1}^{n} \frac{\lambda}{\mu(k)}\right)^{-1}.$$

Заметим, что этим решением можно пользоваться не только для системы $M/M(n)/s$ с учетом (1), но и для всех других экспоненциальных систем с учетом конкретного вида $\mu(n)$. Из вида стационарного распределения заключаем, что для эргодичности необходимо, чтобы ряд

$$\sum_{n=1}^{\infty} \prod_{k=1}^{n} \frac{\lambda}{\mu(k)} \tag{3}$$

сходился. Это условие и достаточно для эргодичности по эргодической теореме Фостера. Действительно, из сходимости ряда (3) вытекает сходимость ряда

$$\sum_{n=1}^{\infty} (\lambda + \mu(m)) p_0 \prod_{k=1}^{n} \frac{\lambda}{\mu(k)}, \tag{4}$$

фигурирующего в эргодической теореме Фостера. На самом деле легко понять, используя признаки сравнения рядов с положительными членами, что ряды (3) и (4) сходятся или расходятся одновременно. Конечно, можно было необходимое и достаточное условие эргодичности получить из общего условия эргодичности для процессов размножения и гибели.

Отметим также, что аналогичным образом (конструктивно) можно задавать входящие потоки. Рассмотрим, например, систему $M(n)/M/1$, т. е. однолинейную систему с экспоненциальным прибором, в которую поступает рекуррентный поток, в котором условное распределение промежутков времени между моментами поступления запросов (при фиксированном числе запросов n в моменты поступления запросов в систему) – показательное с параметром $\lambda(n)$, зависящим от числа запросов n в момент поступления. Если, например, имеется единственный пуассоновский источник запросов с интенсивностью поступления λ, который перекрывается, когда число запросов в системе больше N (велико), то

$$\lambda(n)=\begin{cases}\lambda, & \text{если} \quad n \le N \\ 0, & \text{если} \quad n > N\end{cases}.$$

Это – так называемый ограниченный пуассоновский поток.

Заметим в заключение, что в литературе широкое распространение получили следующие жаргонные обороты:

а) время обслуживания запроса – показательное с параметром $\mu(n)$, зависящим от числа запросов n в системе (правильно говорить – условное распределение времени обслуживания запроса при фиксированном числе n запросов в системе – показательное с параметром $\mu(n)$);

б) в систему поступает пуассоновский поток с интенсивностью $\lambda(n)$, зависящей от числа n запросов в системе в моменты поступления запросов (правильно говорить – между моментами изменения состояний системы в систему поступает пуассоновский поток с интенсивностью, зависящей от состояния системы).

Применение этих жаргонных оборотов сокращает соответствующие фразы, но может привести к недоразумениям.

Литература

1. Гнеденко, Б. В. Введение в теорию массового обслуживания / Б. В. Гнеденко, И. Н. Коваленко. – М. : ЛКИ, 2013. – Изд. 6. – 400 с.

Гомельский государственный
университет имени Франциска Скорины

Поступила в редакцию 11.10.2022