Об эргодических алгоритмах в системах случайного множественного доступа с обратной связью “успех-неуспех”

А.М.Тюрликов и С.Г. Фосс
Санкт-Петербургский государственный университет аэрокосмического приборостроения,
Институт Математики СО РАН, Новосибирск
и университет Хериот-Ватта, Эдинбург

27 августа 2009 г.

Рассматривается децентрализованная система случайного множественного доступа с двоичной обратной связью “успех–неуспех”. Вводится семейство алгоритмов (протоколов), называемых “алгоритмами с отложенными окнами”, и изучаются условия эргодичности на примере одного из них. Обсуждаются результаты численного моделирования и ряд возникающих при этом интересных вопросов и гипотез.

Ключевые слова: случайный множественный доступ, двоичная обратная связь, положительная возвратность, эргодичность.

1 Введение

В конце 70-х годов Цыбаков, Михайлов [1] и Капетанакис [2] впервые рассмотрели модель системы с бесконечным числом абонентов, обменивающихся между собой сообщениями по единому каналу связи, вход и выход которого доступны всем абонентам. В рамках этой модели был предложен алгоритм, позволяющий абонентам передавать данные с конечной средней задержкой при условии, что интенсивность входного потока ограничена некоторой величиной. Работа алгоритма основана на использовании так называемой полной троичной обратной связи. Полная троичная обратная связь предполагает, что абоненты постоянно наблюдают выход канала и по результатом наблюдений могут различить три ситуации - в канале нет передачи (ситуация "Пусто"), в канале передавалось сообщение от одного абонента (ситуация "Успех") и в канале произошло наложение сообщений от двух и более абонентов (ситуация "Конфликт"). После появления работ [1] и [2] сравнительно быстро были предложены и исследованы алгоритмы для двоичной обратной связи вида "Пусто-Непусто"и "Конфликт-Неконфликт". Наименее определенным является случай обратной связи "Успех-Неуспех"(У-НУ). При такой обратной связи абонент не может различать наложения сообщений от разных абонентов от отсутствия передачи. Различные алгоритмы, предложенные для такого вида обратной связи в работах [3], [4] и [5], обеспечивают устойчивое функционирование системы лишь за счет некоторых расширений модели системы типа введения специального тестирующего пакета и т.п. В 1992 в работе [6] была предложена идея алгоритма , который может обеспечивать устойчивую работу системы при обратной

связи У-НУ без расширения модели системы, но данная идея в этой работе не была доведена до уровня алгоритма. Затем в 1995 году в работе [7] было дано четкое описание одного такого алгоритма; выписаны уравнения баланса, которым должно удовлетворять стационарное распределение соответствующей цепи Маркова, и численно найдена граница пропускной способности этого алгоритма, т.е. такое значение интенсивности входного потока λ_0, что уравнения баланса имеют решение тогда и только тогда, когда $\lambda < \lambda_0$. Нам неизвестно о каких-то последующих публикациях работ, посвященных исследованию алгоритмов для случая обратной связи У-НУ.

В современных системах связи наряду с другими методами множественного доступа также может использоваться и режим случайного множественного доступа. Примерами таких систем являются сети, построенные на основе стандартов IEEE 802.11 и IEEE 802.16. В силу технических особенностей построения систем можно говорить о том, что используется обратная связь вида У-НУ. Особенно ярко это проявляется для стандарта IEEE802.16. Базовая станция, построенная согласно этому стандарту, не различает наложений сообщений от разных абонентских станций от отсутствия передачи. Известно, что алгоритмы, используемые на практике для такого вида обратной связи, не обеспечивают устойчивой работы для случая бесконечного числа абонентов(см. напр., [8]). Для существующих в настоящее время сетей, в которых за общий канал конкурирует относительно небольшое число абонентов, эта проблема не является существенной. Однако, при увеличении числа абонентов использование таких алгоритмов может привести к появлению существенных задержек. Таким образом, поиск алгоритмов, обеспечивающих устойчивую работу системы при обратной связи вида У-НУ, представляет не только теоретический, но практический интерес.

В настоящей работе вводится и описывается достаточно общий класс алгоритмов с двоичной обратной связью “Успех–Неуспех”, включающий алгоритм из работы [7]. Затем проводится корректный асимптотический анализ простейшего алгоритма из этого класса, сходного с алгоритмом из [7], см. Теорему 1 и ее следствия. Такой анализ можно проводить аналогичным образом и для других алгоритмов, и преимущество выбранного нами алгоритма состоит лишь в том, что он содержит наименьшее возможное число свободных параметров, что позволяет дополнить его обсуждение наглядным рисунком (см. параграф 5).

Работа построена следующим образом. В параграфе 2 вводится модель системы случайного множественного доступа и приводится описание алгоритма, выбранного для анализа; затем показывается, как этот алгоритм естественным образом обобщается на широкий класс алгоритмов для обратной связи У-НУ. Общим для всех алгоритмов из этого класса является то, что "просматриваются"временные интервалы, причем "удачные"интервалы просматриваются сразу по мере их возникновения, а "неудачные"отправляются в очередь, откуда извлекаются при появлении "удачного интервала"определенного вида. Следуя работе [7], этот класс назван классом алгоритмов с отложенными интервалами. В параграфе 3 на примере нашего алгоритма показывается как можно определять область устойчивости алгоритмов из класса алгоритмов с отложенными интервалами и указывается путь поиска оптимальных параметров алгоритма. При этом задача исследования устойчивости алгоритма сводится к задаче исследования условий положительной возвратности и эргодичности некоторой двумерной марковской цепи. В третьем параграфе формулируется ряд утверждений и теорем, описывающих свойства данной цепи. Доказательства приводятся в параграфе 4. В параграфе 5 приводится сравнительный численный анализ рассмотренного алгоритма и других алгоритмов, введенных в параграфе 2. Кроме того, в параграфе 5 обсуждается

ряд открытых вопросов. Приложение содержит известные вспомогательные утверждения, используемые в работе.

2 Модель системы и алгоритм доступа

2.1 Модель системы

Мы опишем и изучим вариант модели системы случайного множественного доступа, предложенной в [9]. В системе имеется бесконечное число абонентов и канал связи, вход и выход которого доступны всем абонентам. У абонентов возникают пакеты, которыми они обмениваются, используя канал связи. Предполагается, что все пакеты имеют одинаковую длину. Время передачи пакета принимается за единицу времени. Процесс поступления пакетов в систему образует однородный пуассоновский поток с интенсивностью λ (другими словами, интервалы времени между моментами поступления пакетов в систему являются независимыми случайными величинами, имеющими одно и то же экспоненциальное распределение с параметром (интенсивностью) λ и средним $1/\lambda$).

Мы наложим ряд предположений на функционирование канала связи и способ доступа вызовов к нему, сходные с условиями работы [9] и отличающиеся от них только предположением 3 (см. ниже).

Предположение 1. *Время передачи по каналу разделено на окна. Все окна имеют одинаковую длительность, равную времени передачи одного пакета. Окна пронумерованы целыми неотрицательными числами, окну с номером t соответствует интервал времени $[t, t+1)$* (краткости ради, в дальнейшем будем "окно с номером t" называть просто "окном t"). *Моменты разделения окон известны всем абонентам. Абонент может начинать передачу пакета только в начале очередного окна.*

Предположение 2. *В каждом окне может произойти одно из трех событий:*
– в окне передает один абонент (событие S *– success, успех)*;
– в окне не передает ни один абонент (событие E *– empty, пусто)*;
– в окне передают два или более абонентов (событие C *– collision, конфликт).*

Предположение 3. *Каждый из абонентов системы, наблюдая выход канала, к концу окна узнает, был в окне успех или нет. При этом, если успеха не было (т.е. произошел "неуспех"), то абоненты, не пытавшиеся передать сообщение в этом окне, не имеют возможности различить, что было: либо конфликт, либо пустое окно.* (В этом наша модель отлична от модели, предложенной в [9], где предполагалась возможность распознавания конфликта/пустого окна).

Обозначим через θ_i индикатор события $\{$ *в окне i произошел успех* $\}$, т.е. случайную величину, принимающую значение 1, если это событие произошло, и значение 0 в противном случае. Последовательность $\theta(t) = \{\theta_1, ...\theta_t\}$ будем называть *историей канала* к окну t. Предполагается, что все абоненты с начала работы системы постоянно следят за выходом канал и соответственно наблюдают одинаковую историю канала.

Предположение 4. *У абонента имеется буфер для хранения одного пакета.* Краткости ради, "пакет, возникший в момент x", будем называть "пакетом x". Через t_x обозначим целую часть числа x, т.е. такое целое число, что $t_x \le x < t_x + 1$.

Абонент запоминает пакет x в буфере и, начиная с окна $t_x + 1$, использует значение

x для принятия решения о том в каких окнах пакет x будет передаваться. Обозначим через $\nu_i^{(x)}$ индикатор события $\{$ *пакет x передавался в окне i* $\}$. Последовательность $\nu^{(x)}(t) = \{\nu_0^{(x)}, ...\nu_t^{(x)}\}$ будем называть *историей пакета x* к окну t. Если в окне t происходит передача пакета x, т.е. $\nu_t^{(x)} = 1$, и если других пакетов не передается, т.е. $\theta_t = 1$, то по завершении этого окна считается, что пакет x в успешно передан, и он удаляется из системы.

2.2 Алгоритм доступа

Следуя работе [9], *алгоритмом случайного множественного доступа* будем называть правило, в соответствии с которым каждый присутствующий в системе абонент в начале очередного окна t выбирает решение, передавать пакет x в окне t или нет, используя при этом общую для всех абонентов *историю канала* $\theta(t-1)$ и индивидуальную *историю пакета* $\nu^{(x)}(t-1)$. Это решение, вообще говоря, может быть как детерминированным, так и случайным.

В настоящей работе мы ограничимся рассмотрением алгоритмов, для которых правило принятия решения состоит из следующих двух этапов:
(1) сначала все абоненты на основе *истории канала* $\theta(t-1)$ к началу окна t одинаковым образом выбирают на временной оси некоторое множество $B(t)$ и число $p_t \in [0,1]$;
(2) после этого абоненты принимают индивидуальные решения о передаче пакета в окне t - если $x \in B(t)$, то пакет x передается в окне t с вероятностью p_t, если же x находится вне этого множества, то пакет не передается (с вероятностью единица).

При описании алгоритма будем пользоваться следующей терминологией. Моментам возникновения пакетов будем ставить в соответствия точки на временной оси - пакету x ставится в соответствие точка с координатой x. Если пакет успешно передан, то соответствующая точка удаляется. Если к началу окна t выбрано подмножество временной оси $B(t) = B$ и число $p_t = p$, то будем говорить, что множество B *просматривается с вероятностью p* в окне t . Для краткости при $p = 1$ будем говорить, что множество *просматривается* и только при $p < 1$ будем указывать, что множество *просматривается с вероятностью p*. Если множество B *просматривается* то при $\theta(t) = 1$ оно оказывается *просмотренным* , и *частично просмотренным* в противном случае (т.е. при $\theta(t) = 0$). Естественно, объединение просмотренных множеств является просмотренным множеством, т.е. если некоторое множество B является таковым к концу окна t, то это означает, что все пакеты, которые возникли в моменты времени из этого множества, получили успешную передачу в окне t или ранее, и покинули систему.

Используя выше введенную терминологию, опишем работу алгоритма. Она разбита на сеансы времени. Временная ось делится на (полу)интервалы длины $A + B$, где числа A и B являются параметрами алгоритма. Сеанс с номером 0 завершается в момент времени 0. Сеанс с номером k начинается с просмотра интервала $[(k-1)(A+B) + k(A+B))$. При $k \geq 1$ обозначим через s_k и e_k, моменты, соответственно, начала и окончания k-го сеанса. При завершении очередного сеанса (скажем, с номером $k-1$) следующий (k-ый) сеанс начинается сразу же ($s_k = e_{k-1}$), если $e_{k-1} \geq k(A+B)$. В противном случае происходит простой в течение $\lceil (k(A+B) - e_{k-1}) \rceil$ окон и после этого k-ый сеанс начинает работу, то есть $s_k = e_{k-1} + \lceil (k(A+B) - e_{k-1}) \rceil$. Мы обозначаем через $\lceil x \rceil$ наименьшее среди целых чисел, не меньших, чем x.

В каждом очередном сеансе, например с номером $k \geq 1$ выполняются следующие действия:

Шаг 1. В окне s_k+1 просматривается интервал $[(k-1)(A+B)+k(A+B))$, который ранее не просматривался и состоит из двух последовательных непересекающихся интервалов с длинами A и B, соответственно. В дальнейшем для краткости исходный интервал длины $A+B$ и интервалы с длинами A и B будем называть интервалом $A+B$, интервалом A и интервалом B соответственно.

Если $\theta(s_k+1)=1$ (т.е. происходит "Успех"), то интервал $A+B$ становится просмотренным, и сеанс заканчивается, иначе выполняется **Шаг 2**.

Шаг 2. В окне s_k+2 просматривается интервал A. Если $\theta(s_k+2)=0$ (происходит "Неуспех"), то весь интервал $A+B$ ставится в очередь отложенных интервалов и сеанс заканчивается. В противном случае интервал A становится просмотренным, из чего будет следовать, что интервал B непуст. При этом есть три возможности:

Шаг 2.1.Если отложенных интервалов нет в наличии, то *процедура просмотра непустого множества* применяется к интервалу B. По завершении работы *процедуры просмотра непустого множества* интервал B оказывается просмотренным (и удаляется из дальнейшего рассмотрения), и сеанс завершается.

Шаг 2.2. Если есть только один отложенный интервала, то интервал B объединяется с отложенным интервалом, и затем к этому объединению применяется *процедура просмотра непустого множества*, описанная ниже. По завершении работы *процедуры просмотра непустого множества* как интервал B, так и присоединенный к нему интервал оказываются просмотренными (и удаляются из последующего рассмотрения), и Сеанс завершается.

Шаг 2.3. Если есть более одного отложенного интервала, то интервал B объединяется с первыми двумя из отложенных интервалов, и затем к этому объединению применяется *процедура просмотра непустого множества*, описанная ниже. По завершении работы *процедуры просмотра непустого множества* как интервал B, так и присоединенные к нему два интервала оказываются просмотренными (и удаляются из последующего рассмотрения), и сеанс завершается.

Осталось описать *процедуру просмотра непустого множества.*

Применительно к некоторому множеству V, про которое известно, что оно не пусто, эта *процедура* состоит в выполнении следующих действий. Полагаем $V=V_0$.

Действие 1. Множество V_0 *просматривается с вероятностью единица.* Если при этом происходит "Успех" , то множество V_0 оказывается просмотренным и процедура завершена. В противном случае переходим к **Действию 2**.

Действие 2. Множество V_0 *просматривается с вероятностью* α. Если происходит "Неуспех", то **Действие 2** повторяется снова – до тех пор, пока "Успех" не произойдет. При этом "успешный" элемент, скажем $\{x\}$, удаляется. Полагаем $V_0:=V_0\setminus\{x\}$ и переходим снова к **Действию 1**.

Здесь $\alpha\in(0,1)$ - параметр *процедуры просмотра непустого множества.*

Нетрудно видеть, что описанный выше алгоритм задается пятью параметрами:
– длинами интервалов A и B;
– параметрами α_0,α_1 и α_2, которые используются в *процедуре просмотра непустого множества* на шагах **2.1**, **2.2** и **2.3**, соответственно. Индекс у параметра α показывает, сколько интервалов извлекается из очереди.

Ниже мы изучим этот алгоритм более подробно. В частности, мы рассмотрим следующие вопросы:

– При произвольной фиксации значений параметров, существуют ли значения интенсивности входного потока λ, при которых алгоритм обеспечивает устойчивую работу? И если такие значения существуют, то каковы они?
- При какой максимальной интенсивности входного потока λ может быть обеспечена устойчивая работа алгоритма? (Этот параметр естественно называть границей пропускной способности).

2.3 Класс алгоритмов доступа с отложенными интервалами

Описанный в 2.2 алгоритм естественно называть *алгоритмом с отложенными интервалами*. Предложим более общее описание алгоритмов такого вида.

Работа каждого алгоритма с отложенными интервалами состоит из последовательных сеансов. Прежде всего, фиксируются положительное целое число $N \geq 2$ и затем N положительных чисел $D_1, D_2, \ldots, D_N$. Во время сеанса выполняются следующие действия:

Шаг 1 алгоритма для сеанса с номером k начинается в окне $s_k + 1$ просмотром интервала длины $\sum_{i=1}^{N} D_i$, состоящего из N непересекающихся интервалов, которые ранее не просматривались, с длинами D_1, D_2 ... D_N, соответственно. Если $\theta(s_k) = 0$, то следующим выполняется **Шаг 2**, иначе сеанс заканчивается.

Шаг j (где $j < N$). В окне $s_k + j$ просматривается интервал длины $\sum_{i=1}^{N-j+1} D_i$, являющийся объединением первых $N - j + 1$ интервалов $D_1, \ldots, D_{N-j+1}$. Если $\theta(s_k + j) = 0$, то выполняется **Шаг j+1**. В противном случае интервал длины $\sum_{i=1}^{N-j+1} D_i$ оказывается просмотренным, из чего следует, что оставшийся интервал длины $\sum_{i=n-j+2}^{N} D_i$ непуст. Этот интервал объединяется с некоторым числом из отложенных интервалов (если такие имеются), и к полученному множеству применяется *процедура просмотра непустого множества*. Как способ выбора числа отложенных интервалов, так и параметр *процедуры просмотра непустого множества* являются параметрами алгоритма.

Шаг N аналогичен **Шагу j** при $j = N$ за исключением того, что при $\theta(s_k + N) = 0$ весь интервал $\sum_{i=1}^{N} D_i$ ставится в очередь отложенных интервалов и сеанс заканчивается.

Для описанного в предыдущем пункте 2.2 алгоритма

(а) $N = 2$ и $D_1 = A$ и $D_2 = B$;

(б) способ выбора части отложенных интервалов таков: если имеется q отложенных интервалов (которые, скажем, образуют очередь), то выбирается $\min(2, q)$ первых их них.

Заметим, что этот алгоритм можно назвать «самым простым из устойчивых алгоритмов в рассматриваемом классе». В конце пункта 3.3 мы покажем, что при $N = 2$ среди алгоритмов, при которых к рассматриваемому интервалу присоединяется не более одного из отложенных, устойчивых нет.

Для описанного в [7] алгоритма

(а) $N = 3$;

(б) если на шаге $j = 1, 2$ соответствующий подинтервал оказывается просмотренным, то оставшийся подинтервал объединяется ровно с одним (первым) из отложенных – если таковые имеются.

3 Асимптотический анализ алгоритма: условия положительности возвратности и эргодичности

3.1 Изменение масштаба времени

Изменим рассматриваемую модель, проведя масштабирование временной оси в λ раз. При этом длина окна становится равной λ, а процесс поступления заявок (абонентов) в систему становится пуассоновским с параметром 1. Удобство новой модели состоит в том, что новые длины окон $a = A\lambda$ и $b = B\lambda$ становятся свободными переменными, не связанными с параметром λ. Обозначим $L = a + b$.

Введем две характеристики системы: в момент времени t это
– длина непросмотренного интервала входного потока $W(t)$ и
– количество отложенных, частично просмотренных интервалов $Q(t)$.

Будет использовать те же символы s_k и e_k для моментов начала и окончания сеансов просмотра, но уже в новом масштабе времени.

Напомним, что при $k = 1, 2, \ldots$, сеанс с номером k начинается сразу же после окончания предыдущего $(k-1)$-го сеанса ($s_k = e_{k-1}$), если $W(e_{k-1}) \geq L$. В противном случае происходит простой в течение, скажем, i единиц времени, где i – наименьшее целое число, при котором $W(e_{k-1}) + i\lambda \geq L$, т.е. $i = \lceil (L - W(e_{k-1}))/\lambda \rceil$, при этом $s_k = e_{k-1} + ß\lambda$ и $W(s_k) = W(e_{k-1}) + i\lambda$.

За время T_k работы k-го сеанса ($k = 1, 2, \ldots$) длина непросмотренного интервала увеличивается на величину λT_k, т.е.

$$W(e_k) = W(s_k) - L + \lambda T_k,$$

и по завершении сеанса возможны три варианта:
– либо просмотренный интервал удаляется, при этом число отложенных интервалов не изменяется,

$$Q(e_k) = Q(e_{k-1}),$$

– либо просмотренный интервал удаляется вместе с единственным имеющимся отложенным интервалом,

$$Q(e_k) = Q(e_{k-1}) - 1,$$

– либо вместе с двумя отложенными,

$$Q(e_k) = Q(e_{k-1}) - 2,$$

– либо интервал просматривается лишь частично и добавляется к отложенным, т.е.

$$Q(e_k) = Q(e_{k-1}) + 1.$$

Мы сначала рассмотрим подпоследовательность случайных векторов во вложенные моменты окончаний сеансов,

$$(W_k, Q_k) := (W(e_k), Q(e_k)),$$

и отметим, что из приведенного выше построения и из марковости входного пуассоновского процесса следует, что эта последовательность образует однородную (по времени) цепь Маркова. Мы изучим условия ее возвратности/невозвратности (в зависимости от

параметров λ, a и b и дополнительных параметров $\alpha_0, \alpha_1, \alpha_2$). Отметим, что последовательность $(W(\lambda n), Q(\lambda n))$, вообще говоря, цепью Маркова не является. Мы покажем, что невозвратность вложенной цепи Маркова влечет аналогичное свойство последовательности $(W(\lambda n), Q(\lambda n))$; а также что из возвратности вложенной цепи следует (при выполнении одного технического условия), что последовательность $(W(\lambda n), Q(\lambda n))$ является регенерирующей и апериодической, что влечет существование ее стационарной версии и сходимость к ней в метрике полной вариации.

3.2 Вероятности событий в сеансе

Повторим, что в ходе работы сеанса может произойти одно из трех событий:

- сеанс завершается на **Шаге 1**, при этом длина очереди отложенных интервалов не меняется;

- сеанс завершается на **Шаге 2**, и в очередь отложенных интервалов добавляется один элемент.

- сеанс завершается на **Шаге 2**, и при этом из очереди длины q отложенных интервалов извлекается $\min(2, q)$ первых.

Обозначим через p_0, p_1 и p_- соответственно, вероятности выше описанных событий, и через X_a и X_b, соответственно, число точек пуассоновского потока интенсивностью единица на непересекающихся интервалах времени с длинами a и b.

На **Шаге 1** просматривается интервал длины $a + b$ и он становится просмотренным, если на нем находится только одна точка пуасоновского потока, т.е. если

$$X_a + X_b = 1,$$

Такое событие происходит с вероятностью

$$p_0 = \mathbf{P}(X_a + X_b = 1) = (a + b)e^{-a-b}.$$

Далее,

$$p_- = \mathbf{P}(X_a + X_b \neq 1, X_a = 1) = \mathbf{P}(X_a = 1, X_b \geq 1) = \mathbf{P}(X_a = 1)\mathbf{P}(X_b \geq 1) = ae^{-a}(1 - e^{-b})$$

и

$$p_1 = 1 - p_0 - p_- = 1 - be^{-a-b} - ae^{-a}.$$

3.3 Вложенная цепь Маркова

Как следует из предыдущего, если отложенные интервалы в системе имеются (скажем, их q), то после очередного сеанса их число либо убывает на $\min(q, 2)$ (с вероятностью p_-), либо возрастает на одно (с вероятностью p_1), либо остается неизменным (вероятность этого равна p_0). Если же отложенных интервалов нет ($q = 0$), то после сеанса либо появляется один интервал (с вероятностью p_1), либо нет (с вероятностью $p_- + p_0$).

Поэтому одномерная последовательность Q_n тоже образует цепь Маркова,

$$Q_{n+1} = \max(Q_n + \xi_n, 0), \tag{1}$$

где последовательность ξ_n состоит из независимых одинаково распределенных (н.о.р.) случайных величин,

$$\mathbf{P}(\xi_n = 1) = p_1, \quad \mathbf{P}(\xi_n = 0) = p_0, \quad \mathbf{P}(\xi_n = -2) = p_-.$$

У этой цепи существует стационарное распределение тогда и только тогда, когда $\mathbf{E}\xi_n < 0$, т.е. $h := p_1/2p_- < 1$. Обозначим это стационарное распределение через $\{\pi_i\}_{i\geq 0}$.

Последовательность (1) принято называть *целочисленным случайным блужданием с задержкой в нуле*. В нашем случае это блуждание является *непрерывным справа*, т.е. $\mathbf{P}(\xi_n \geq 2) = 0$. Отметим также, что непрерывными справа являются и все другие алгоритмы, описанные в п.2.2. Для таких блужданий известно (см., напр., [10], глава 11), что стационарное распределение является геометрическим, т.е.

$$\pi_i = \pi_0(1-\pi_0)^i, \quad i \geq 0,$$

и что число π_0 является единственным решением z уравнения $\sum \mathbf{P}(\xi_n = j)(1-z)^{-j} = 1$ в области $z \in (0,1)$. В нашем случае

$$\pi_0 = \frac{3-\sqrt{1+8h}}{2}.$$

Отметим также следующие хорошо известные факты (которые тоже можно найти, скажем, в [10]). Допустим, что $h < 1$. Если взять в качестве начального состояния $Q_0 = m \geq 0$ и обозначить

$$\tau^{(m)} = \min\{n \geq 1 \ : \ Q_n = 0 \mid Q_0 = m\},$$

то у этой случайной величины все степенные моменты конечны, $\mathbf{E}\left(\tau^{(m)}\right)^k < \infty$ при всех $k > 0$ и, более того, существует показательный момент, т.е. $\mathbf{E}e^{c\tau^{(m)}} < \infty$ при некотором $c = c_m > 0$. В частности, $\mathbf{E}\tau^{(0)} = 1/\pi_0$ и $\mathbf{E}\tau^{(m)} \leq C + m/(2p_- - p_1)$ при некотором C при каждом $m \geq 1$. Далее, эта цепь Маркова является геометрически эргодической, т.е. существуют абсолютная постоянная C и при любом $m \geq 0$ постоянная c_m такие, что при всех $n \geq 0$

$$\sup_k \left|\mathbf{P}(Q_n = k \mid Q_0 = m) - \pi_k\right| \leq c_m e^{-Cn}.$$

Отметим, что неравенство $h < 1$ может выполняться при некотором выборе параметров a и b. Действительно, оно эквивалентно такому:

$$2p_- - p_1 = 3ae^{-a} + (b-2a)e^{-a-b} - 1 > 0.$$

Последнее же неравенство имеет место, скажем, при $a = 1$ и $b = 2$.

Отметим также, что в более простом алгоритме, в котором разрешается извлекать не более одного из отложенных интервалов, вложенная цепь Маркова могла бы быть положительно возвратной только при $p_- > p_1$. Однако это эквивалентно неравенству

$$2ae^{-a} - ae^{-a-b} + be^{-a-b} > 1,$$

которое не имеет решений на множестве положительных вещественных чисел.

3.4 Просмотр непустого множества

Предположим, что мы знаем, что некоторое множество D непусто, но неизвестно количество элементов в нем. Предполагается, что это количество X случайно и имеет известное вероятностное распределение. Рассмотрим следующий алгоритм "идентификации" элементов множества.

На первом шаге всё множество просматривается с вероятностью единица. Если оказывается, что во множестве только один элемент, то процедура завершается.

В противном случае на каждом из следующих шагов каждый из элементов просматривается с некоторой вероятностью $\alpha \in (0,1)$ – до тех пор, пока не окажется ровно один такой. (Ясно, что число таких попыток случайно и имеет геометрическое распределение с параметром $r_{n,\alpha} = n\alpha(1-\alpha)^{n-1}$, если предположить, что $X = n$. При этом среднее число таких попыток равно $1/r_{n,\alpha}$).

Далее этот элемент вычеркивается, и процедура повторяется: сначала все оставшиеся элементы просматриваются с вероятностью единица. а затем – если их число больше единицы – несколько раз с вероятностью α, пока не будет просмотрен ровно один. Этот процесс продолжается до тех пор, пока все элементы не будут удалены.

Обозначим через $R_\alpha(X)$ длительность (т.е. число шагов) этого алгоритма. Тогда

$$\begin{aligned}
\mathbf{E}R_\alpha(X) &= \sum_{n=1}^{\infty} \mathbf{E}(R_\alpha(X) \mid X = n) \cdot \mathbf{P}(X = n) \\
&= \sum_{n=1}^{\infty} \mathbf{P}(X = n) \left(1 + \sum_{m=2}^{n} (1 + 1/r_{m,\alpha})\right) \\
&= \mathbf{E}X + \sum_{n=2}^{\infty} \mathbf{P}(X = n) \sum_{m=2}^{n} \frac{1}{r_{m,\alpha}}.
\end{aligned}$$

Заметим, что

$$\max_\alpha r_{m,\alpha} = r_{m,1/m} = (1 - 1/m)^{m-1}.$$

Поэтому при любом $\alpha \in (0,1)$ справедливо неравенство

$$\mathbf{E}R_\alpha(X) \geq \mathbf{E}X + \sum_{n=2}^{\infty} \mathbf{P}(X = n) \sum_{m=2}^{n} (1 - 1/m)^{-m+1}. \tag{2}$$

И так как $(1 - 1/m)^{-m+1} > 2$ при всех $m \geq 2$, то из неравенства (2) вытекает, в частности, простая нижняя оценка

$$\mathbf{E}R_\alpha(X) \geq 3\mathbf{E}X - 2. \tag{3}$$

3.5 Средняя длительность сеанса

Пусть по-прежнему $h < 1$. Пусть T – длительность типичного сеанса в стационарном режиме просмотра. Отметим, что T зависит от 5 параметров, $T = T(a, b, \alpha_0, \alpha_1, \alpha_2)$. Напомним, что отложенные интервалы образуют очередь, и обозначим через $Y^{(i)}$ количество требований в i-ом интервале. По построению, случайные величины Y^i, $i = 1, 2, \ldots$ независимы и одинаково распределены.

Повторим, что возможны 3 случая:

(1) с вероятностью p_0 в новом окне длины $a+b$ находится ровно одно требование – при этом $T=1$;

(2) с вероятностью p_1 последовательный просмотр окон $a+b$ и a показывает, что в каждом из них количество требований отлично от единицы – тогда $T=2$;

(3) если оказалось, что $X_{a+b}\neq 1$ и $X_a=1$ (вероятность такого исхода есть p_-), то длина сеанса равна $T=2+T_+$, где T_+ равно либо $R_{\alpha_0}(\widetilde{X}_b)$, если очередь отложенных интервалов пуста (т.е. с вероятностью π_0), либо $R_{\alpha_1}(Y^{(1)}+\widetilde{X}_b)$, если в очереди ровно один интервал (т.е. с вероятностью π_1), либо $R_{\alpha_2}(Y^{(1)}+Y^{(2)}+\widetilde{X}_b)$, если в очереди по крайней мере два отложенных интервала (т.е. с вероятностью $(1-\pi_0-\pi_1)$).

Здесь через $\widetilde{X}_b$ обозначена случайная величина, имеющая условное распределение $\mathbf{P}(\widetilde{X}_b\in\cdot)=\mathbf{P}(X_b\in\cdot\big|X_b\geq 1)$. Справедливы равенства

$$\begin{aligned}\mathbf{E}T &= p_0+2p_1+(2+\mathbf{E}T_+)p_- = 2-p_0+p_-\mathbf{E}T_+\\ &= 2-p_0+\left(E_0\pi_0+E_1\pi_1+E_2(1-\pi_0-\pi_1)\right)p_-,\end{aligned}$$

где

$$E_0=\mathbf{E}R_{\alpha_0}(\widetilde{X}_b)=\mathbf{E}\widetilde{X}_b+\sum_{m\geq 2}\frac{1}{r_{m,\alpha_0}}\mathbf{P}(\widetilde{X}_b\geq m),$$

$$E_1=\mathbf{E}R_{\alpha_1}(Y^{(1)}+\widetilde{X}_b)=\mathbf{E}Y^{(1)}+\mathbf{E}\widetilde{X}_b+\sum_{m\geq 2}\frac{1}{r_{m,\alpha_1}}\mathbf{P}(Y^{(1)}+\widetilde{X}_b\geq m),$$

$$E_2=\mathbf{E}R_{\alpha_2}(Y^{(1)}+Y^{(2)}+\widetilde{X}_b)=\mathbf{E}Y^{(1)}+\mathbf{E}Y^{(2)}+\mathbf{E}\widetilde{X}_b+\sum_{m\geq 2}\frac{1}{r_{m,\alpha_2}}\mathbf{P}(Y^{(1)}+Y^{(2)}+\widetilde{X}_b\geq m).$$

Ниже предлагаются выражения для математических ожиданий случайных величин $\widetilde{X}_b$ и $Y=_{st}Y^{(i)}$, в то время как явный вид выписанных выше сумм можно найти только численными методами. Имеем:

$$\mathbf{E}\widetilde{X}_b=\mathbf{E}(X_b\mid X_b\geq 1)=\frac{\mathbf{E}X_b}{\mathbf{P}(X_b\geq 1)}=\frac{b}{1-e^{-b}}.$$

$$\begin{aligned}\mathbf{E}Y &= \mathbf{E}(X_a+X_b\big|X_a\neq 1, X_a+X_b\neq 1)\\ &= \frac{1}{p_1}\mathbf{E}\left((X_a+X_b)\cdot(\mathbf{I}(X_a\geq 2)+\mathbf{I}(X_a=0,X_b\geq 2)\right)\\ &= \frac{1}{p_1}\left(a+b-ae^{-a}-be^{-a-b}-abe^{-a}\right).\end{aligned}$$

Здесь $\mathbf{I}$ – это индикаторная функция, $\mathbf{I}(B)=1$ если событие B происходит, и $\mathbf{I}(B)=0$ в противном случае.

3.6 Условия положительной возвратности и эргодичности

Напомним, что в стационарном режиме полностью просматривается от $N = 0$ до $N = 3$ интервалов длиной $a + b$, с соответствующими вероятностями

$$\mathbf{P}(N = 0) = p_1, \quad \mathbf{P}(N = 1) = p_0 + p_-\pi_0, \quad \mathbf{P}(N = 2) = p_-\pi_1, \quad \mathbf{P}(N = 3) = p_-(1 - \pi_0 - \pi_1). \tag{4}$$

При этом $\mathbf{E}N = 1$ и средняя суммарная длина выкинутых (полностью просмотренных) окон равна

$$L = (a + b)\mathbf{E}N = a + b$$

(что совершенно естественно, так как среднее число просмотренных окон не может быть ни меньше, ни больше числа новых окон, т.е. единицы). Далее, предположим, что цепь Маркова стартует из состояния $(W_0, 0)$, т.е. $Q_0 = 0$, и обозначим $\tau = \min\{n : \ Q_n = 0\}$. Пусть N_i –количество окон, просмотренных в сеансе i и T_i – длительность i-го сеанса. Тогда

$$\mathbf{E}\left(\sum_{i=1}^{\tau} N_i\right) = \mathbf{E}\tau\mathbf{E}N = \mathbf{E}\tau \quad \text{и} \quad \mathbf{E}\left(\sum_{i=1}^{\tau} T_i\right) = \mathbf{E}\tau\mathbf{E}T, \tag{5}$$

где T – длительность типичного сеанса в стационарном режиме, рассмотренная в предыдущем пункте.

Определение 1. Назовем цепь Маркова (W_n, Q_n) *возвратной*, если найдется ограниченное множество $A = \{W \le c_1, Q \le c_2\}$ такое, что
(1) $\tau_{(W,Q)} = \tau_{(W,Q)}(A) = \min\{n \ge 1 \ : \ (W_n, Q_n) \in A \mid W_0 = W, Q_0 = Q\} < \infty$ п.н. для любого начального значения (W, Q).
При этом марковская цепь является *положительно возвратной*, если
(2) $\sup_{(W,Q)\in A} \mathbf{E}\tau_{(W,Q)} < \infty$,
и *нулевой возвратной*, в противном случае.

Назовем марковскую цепь (W_n, Q_n) *невозвратной (transient)*, если $W_n + Q_n \to \infty$ п.н. при $n \to \infty$, при любом начальном условии $W_0 = W, Q_0 = Q$.

Замечание 1. Приведенные выше определения положительной и нулевой возвратности являются достаточно стандартным. Вообще говоря, существует несколько вариантов определения невозвратности цепи Маркова, и приведенное выше – самое ограничительное из них.

Определение 2. Естественно называть алгоритм просмотра (передачи сообщений) положительно/нулевым возвратным или невозвратным, если таковой является соответствующая ему марковская цепь (W_n, Q_n).

Теорема 1. *При пуассоновском входном потоке с интенсивностью λ и при произвольных $a > 0, b > 0$ и наборе вероятностей α описанный выше алгоритм является*
(а) положительно возвратным, если $2p_- > p_1$ и $\lambda < L/\mathbf{E}T$, и
(б) невозвратным, если либо $\lambda > L/\mathbf{E}T$, либо $2p_- < p_1$.

Замечание 2. Можно также показать, при выполнении условий (а) теоремы 1 множество A является положительно возвратным при *любом* выборе положительных чисел c_1, c_2.

Замечание 3. Нетрудно также показать, что если $p_- = p_1$ и $\lambda < L/\mathbf{E}T$, то описанная выше цепь Маркова будет иметь *нулевую возвратность*. Скорее всего, то же имеет место,

если $p_- > p_1$ и $\lambda = L/\mathbf{E}T$. Если это так, то теорема 1 может быть сформулирована и несколько по-иному: алгоритм является положительно возвратным только и только тогда, когда выполнены условия (а).

Замечание 4. Естественно называть отношение $L/\mathbf{E}T$ *скоростью* описанного алгоритма. Напомним, что это отношение зависит от 5 параметров.

Определение 2. Будем называть цепь Маркова $\{(W_n, Q_n)\}$ *эргодической* (и соответствующий алгоритм эргодическим), если у нее существует единственное стационарное распределение Π и, более того, при любом начальном условии (W_0, Q_0) распределения случайных векторов $\{(W_n, Q_n)\}$ с ростом n слабо сходятся к этому стационарному распределению; и *сильно эргодической* (соотв., сильно эргодическим), если к тому же имеет место сходимость в *метрике полной вариации*, т.е.

$$\sup |\mathbf{P}((W_n, Q_n) \in B) - \Pi(B)| \to 0, \quad n \to \infty,$$

где супремум берется по всех двумерным измеримым множествам B.

Отметим, что положительная возвратность цепи Маркова не гарантирует, вообще говоря, существования у нее (а также единственности) стационарного распределения.

Теорема 2. *Пусть $C = \sup L/\mathbf{E}T$, где супремум берется по всем возможным значениями пяти параметров, при которых $2p_- > p_1$ (естественно называть это число пропускной способностью рассматриваемого семейства алгоритмов).*
(а) Если $\lambda < C$, то можно указать параметры a,b и набор вероятностей α, при которых описанный нами алгоритм является сильно эргодическим. При этом исходный процесс $(W(t), Q(t))$ оказывается регенерирующим и апериодическим и, следовательно, существует собственное стационарное распределение, к которому распределения векторов $(W(t), Q(t))$ сходятся с ростом t в метрике полной вариации.
(б) При выполнении же противоположного строго неравенства $\lambda > C$ все рассматриваемые алгоритмы являются невозвратными.

Замечание 5. Можно относительно просто найти верхнюю оценку для числа C. А именно, воспользоваться (3 раза) нижней оценкой (2) и рассмотреть более простую задачу оптимизации по 2 параметрам.

4 Доказательства

4.1 Доказательство части (а) теоремы 1

Рассмотрим вложенные моменты начала тех сеансов k_n, в которых $Q_{k_n} = 0$ и назовем промежуток времени между двумя такими последовательными моментами циклом. Случайные величины $\{k_n - k_{n-1}\}$ являются независимыми одинаково распределенными, имеющими то же распределение, что и случайная величина τ, введенная в п.3.6.

Обозначим $\widetilde{W}_n = W_{k_n}$.

Покажем сначала, что цепь Маркова $\widetilde{W}_n$ является положительно возвратной. Для этого воспользуемся первой частью критерия Фостера (см. теорему 3 в Приложении). В этой части доказательства можно без ограничения общности положить $k_0 = 0$. При этом $\tau = k_1 - k_0$.

Обозначим

$$\Delta_x = \mathbf{E}(\widetilde{W}_1 \mid \widetilde{W}_0 = x) - x$$

и покажем, что

– Δ_x ограничено сверху одной и той же константой при всех x и

– $\limsup_{x\to\infty} \Delta_x < 0$.

Тогда применим критерий Фостера.

Действительно, суммарная длина просмотренных интервалов на этом цикле равна $L\sum_{i=1}^{\tau} N_i$, где, напомним, $L = a+b$ и N_i – количество просмотренных интервалов за i-ый сеанс, причем $\mathbf{E}\sum_{i=1}^{\tau} N_i = \mathbf{E}\tau$, в силу (5).

Так как суммарный прирост координаты W за время t есть λt, то ее прирост за время с момента начала первого цикла до момента начала второго цикла не меньше, чем $\lambda\sum_{i=1}^{\tau} T_i$, и не больше, чем

$$\lambda\sum_{i=1}^{\tau} T_i + (L+1)\sum_{i=1}^{\tau} \mathbf{I}(W_i^x < L),$$

где T_i есть длительность i-го сеанса, и верхний индекс x означает, что первый сеанс в цикле начинается с $W_0 = \widetilde{W}_0 = x$. Поэтому $W_i^x \geq x - 2iL$ при всех x и i и, значит,

$$0 \leq \sum_{i=1}^{\tau} \mathbf{I}(W_i^x < L) \leq \sum_{i=1}^{\tau} \mathbf{I}(x - iL < L) \leq \tau\mathbf{I}(x < 2L\tau + L),$$

где верхняя оценка $\tau\mathbf{I}(x < 2L\tau + L)$ стремится монотонно к нулю с ростом x почти наверное и в среднем (по теореме Лебега о монотонной сходимости). Поэтому с ростом x

$$\Delta_x \to \lambda\mathbf{E}\tau\mathbf{E}T - L\mathbf{E}\tau\mathbf{E}N = \mathbf{E}\tau\mathbf{E}T\,(\lambda - L/\mathbf{E}T) < 0.$$

Здесь N – количество просмотренных интервалов за один сеанс в стационарном режиме, имеющая распределение (4). И так как

$$\Delta_x \leq \mathbf{E}\left(\lambda\sum_{i=1}^{\tau} T_i + L\sum_{i=1}^{\tau} \mathbf{I}(W_I^x < L)\right) \leq \lambda\mathbf{E}\tau\mathbf{E}T + L\mathbf{E}\tau < \infty$$

при всех x, то цепь Маркова $\{\widetilde{W}_n\}$ положительно возвратна.

Докажем теперь положительную возвратность цепи Маркова (W_n, Q_n). Для этого воспользуемся первой частью обобщенного критерия Фостера (см. теорему 4 в Приложении).

Пусть $W_0 = W \geq 0$ и $Q_0 = m \in \{0, 1, 2, \ldots\}$. Тогда, с использованием обозначений из пункта 3.3, $\tau^{(m)}$ конечно п.н. и, более того, имеет конечное среднее. В этой части доказательства мы должны положить $k_0 = \tau^{(m)}$. Далее, $\widetilde{W}_0 = W_{\tau^{(m)}}$ и

$$\mathbf{E}\widetilde{W}_0 \leq W + \mathbf{E}\tau^{(m)}C = W + m\widetilde{C},$$

где $C = L + K_0$,

$$K_0 = 2 + \max(\mathbf{E}R_{\alpha_0}(\widetilde{X}_b, \mathbf{E}R_{\alpha_1}(Y^{(1)} + \widetilde{X}_b), \mathbf{E}R_{\alpha_2}(Y^{(1)} + Y^{(2)} + \widetilde{X}_b)) < \infty$$

и $\widetilde{C} = C/(2p_- - p_1)$. В качестве g возьмем функцию $g(w, m) = w + m$.

Положим $\widetilde{\mu} = \min\{n \geq 0 \ : \widetilde{W}_n \leq g_0\}$. Тогда, в силу доказанного ранее,

$$\mathbf{E}(\widetilde{\mu} \mid \widetilde{W}_0) \leq K(\widetilde{W}_0 + 1),$$

при некоторой абсолютной постоянной K. Следовательно, если обозначить

$$\gamma = \min\{n \ : \ W_n + Q_n \le g_0\},$$

то

$$\mathbf{E}(\gamma \mid W_0 = W, Q_0 = m) \le \mathbf{E}\tau^{(m)} + \mathbf{E}\left(\sum_1^{\widetilde{\mu}} z_i\right),$$

где z_i –длительность соответствующего цикла (эти циклы являются независимыми и одинаково распределенными с конечным средним). Поэтому найдется еще одна абсолютная постоянная $\widehat{K}$, такая что

$$\mathbf{E}(\gamma \mid W_0 = W, Q_0 = m) \le \widehat{K}(W + m + 1).$$

Следовательно, применима теорема 4, и ЦМ $\{W_n, Q_n\}$ является положительно возвратной.

4.2 Доказательство части (б) теоремы 1

Если $\mathbf{E}\xi_n = p_1 - 2p_- > 0$, то $\sum_1^n \xi_i \to \infty$ п.н., все Q_n положительны, начиная с некоторого номера и, по усиленному закону больших чисел,

$$\frac{Q_n}{n} \to \mathbf{E}\xi_1 > 0 \quad \text{п.н. при} \quad n \to \infty.$$

Пусть теперь $p_1 < 2p_-$ и $\lambda > L/\mathbf{E}T$. При этом циклы имеют по-прежнему конечное среднее $\mathbf{E}\tau$ и средняя суммарная длительность сеансов за один цикл равняется $\mathbf{E}\tau\mathbf{E}T$. Так как суммарный прирост первой координаты цепи Маркова за типичный цикл по-прежнему не меньше чем $\lambda\sum_1^\tau T_i$, то, опять-таки по усиленному закону больших чисел,

$$\liminf \frac{\widetilde{W}_n}{n} \ge \mathbf{E}\tau\mathbf{E}T(\lambda - L/\mathbf{E}T) > 0 \quad \text{п.н.}$$

Покажем, что и $W_n/n \to \infty$ п.н. Для этого воспользуемся стандартными рассуждениями из теории восстановления. Обозначим через τ_i длительность i-го цикла. Здесь случайные величины $\{\tau_i\}$ независимы в совокупности при $i \ge 1$ и одинаково распределены (с конечным средним) при $i \ge 2$. Положим $S_m = \sum_1^m \tau_i$. Пусть при $n \ge 1$

$$\eta_n = \min\{m \ : \ S_m \ge n\} \quad \text{и} \quad \chi_n = S_{\eta_n} - n.$$

При этом, как хорошо известно, $\chi_n/n \to 0$ п.н. Так как $W_{S_m} = \widetilde{W}_m$, то

$$\frac{W_n}{n} \ge \frac{\widetilde{W}_{\eta_n} - \lambda\chi_n}{n} = \frac{\widetilde{W}_{\eta_n}}{\eta_n} \cdot \frac{\eta_n}{n} - \lambda\frac{\chi_n}{n} \to \infty \quad \text{п.н.}$$

Действительно, так как $\eta_n \to \infty$ и $\eta_n/n \to 1/\mathbf{E}\tau > 0$ при $n \to \infty$, то из стремления к бесконечности $\widetilde{W}_n/n$ п.н. следует, что и $\widetilde{W}_{\eta_n}/\eta_n \to \infty$ п.н., и $W_n/n \to \infty$ п.н.

4.3 Доказательство части (а) теоремы 2

Мы всегда можем найти рациональные значения λ, a, b, c и набор вероятностей α, при которых $2p_- > p_1$ и $\lambda < L/\mathbf{E}T$. При таких параметрах цепь Маркова (W_n, Q_n) принимает значения на счетной решетке. Так как $\mathbf{P}(T = 1) > 0$ и $\mathbf{P}(T = 1, \tau = 1) > 0$, то цепь Маркова оказывается апериодичной и все ее состояния – сообщающимися. Поэтому применима вторая часть обобщенного критерия Фостера, и имеет место сходимость к стационарному распределению цепи Маркова (W_n, Q_n) в метрике полной вариации. Из тех же соображений вытекает, что случайная последовательность $(W(t), Q(t))$ является регенерирующей, причем длина цикла регенерации может принимать значение 1 с положительной вероятностью. Поэтому применима теорема 5, из чего следует утверждение (а).

4.4 Доказательство части (б) теоремы 2

Соотношение $W_n + Q_n \to \infty$ п.н. вытекает из соответствующего утверждения (б) теоремы 1.

5 Обсуждение результатов и открытые проблемы

5.1 Вычисление значения пропускной способности алгоритма .

Напомним, что при рассматриваемом нами алгоритме $\mathbf{E}T \equiv \mathbf{T}(a, b, \alpha_0, \alpha_1, \alpha_2)$ есть среднее число окон в сеансе при заданных значениях парамеров $a, b, \alpha_0, \alpha_1, \alpha_2$. Соответственно, вероятности $p_0 = p_0(a, b)$, $p_- = p_-(a, b)$ и $p_1 = p_1(a, b)$ зависят от параметров a и b .

По теореме 2 пропускная способность этого алгоритма определяется как решение следующей оптимизационной задачи: требуется *найти значение*

$C = \sup\{(a + b)/\mathbf{T}(a, b, \alpha_0, \alpha_1, \alpha_2)\}$

где супремум берется по всем возможным значениям параметров $\alpha_0, \alpha_1, \alpha_2$, лежащим в интервале $(0, 1)$, и неотрицательным значениям параметров a, b, для которых $p_1(a, b)/(2p_-(a, b)) < 1$.

Решение этой оптимизационной задачи может быть сведено к решению более простой оптимизационной задачи следующим образом.

Введем в рассмотрение функцию от переменных a и b

$$\varphi(a, b) = \max(a + b)/\mathbf{T}(a, b, \alpha_0, \alpha_1, \alpha_2),$$

где максимум берется по всем возможным значениям параметров $\alpha_0, \alpha_1, \alpha_2$, лежащим в интервале $(0, 1)$.

Функцию $\mathbf{T}(a, b, \alpha_0, \alpha_1, \alpha_2)$ можно представить в виде

$$\mathbf{T}(a, b, \alpha_0, \alpha_1, \alpha_2) = \gamma(a + b) + E0(a, b, \alpha_0) + E1(a, b, \alpha_1) + E2(a, b, \alpha_2), \tag{6}$$

где

$$\gamma(a + b) = 2 - p_0,$$

$$E0(a, b, \alpha_0) = (\mathbf{E}\widetilde{X}_b + \sum_{m \geq 2} \frac{1}{r_{m,\alpha_0}} \mathbf{P}(\widetilde{X}_b \geq m))\pi_0 p_-,$$

$$E1(a,b,\alpha_1) = (\mathbf{E}Y^{(1)} + \mathbf{E}\widetilde{X}_b + \sum_{m\geq 2} \frac{1}{r_{m,\alpha_1}} \mathbf{P}(Y^{(1)} + \widetilde{X}_b \geq m))\pi_1 p_-,$$

$$E2(a,b,\alpha_2) = (\mathbf{E}Y^{(1)} + \mathbf{E}Y^{(2)} + \mathbf{E}\widetilde{X}_b + \sum_{m\geq 2} \frac{1}{r_{m,\alpha_2}} \mathbf{P}(Y^{(1)} + Y^{(2)} + \widetilde{X}_b \geq m))(1-\pi_0-\pi_1) p_-.$$

Из равенства (6) следует, что для вычисления функции $\varphi(a,b)$ при фиксированных значениях a и b достаточно независимо минимизировать функции $E0(a,b,\alpha_0), E1(a,b,\alpha_1)$ и $E2(a,b,\alpha_2)$ по переменным α_0, α_1 и α_2 соответственно. Так как эти функции унимодальны, то эта минимизация быть выполнена численно с любой заданной точностью.

С использованием введенной выше функции $\varphi(a,b)$ вычисление пропускной способность сводится теперь к решению более простой оптимизационной задачи: требуется *найти значение*

$$C = \sup \varphi(a,b),$$

где супремум берется по всем возможным значениям неотрицательным значениям параметров a, b, для которых $p_1(a,b)/(2p_-(a,b)) < 1$.

Данная оптимизационной задача может быть решена численно. Значение пропускной способности вычисленное с точностью до четырех знаков после запятой равно $0,3098$ и достигается в точке $a \approx 0,651$ и $b \approx 1,18$.

Вычисление максимального значения функции φ оказывается очень трудоемким. Мы предполагаем, что можно получить достатосно просто хорошее приближение значения $C = \sup \varphi(a,b)$ так: достаточно найти среди пар (a,b) для которых $p_1(a,b)/(2p_-(a,b)) = 1$ такую пару, которая максимизирует значение $p_0 = (a+b)e^{-a-b}$, то есть вероятность успеха при просмотре исходного отрезка.

Пока мы не может сторого обсновать эту гипотезу и предложим взамен лишь некоторую иллюстрацию.

На рисунке 1 на плоскости (a,b) выделена область в которой $p_1(a,b)/(2p_-(a,b)) < 1$. Эту область будем называть *областью устойчивости очереди отложенных интервалов*, или кратко *областью устойчивости очереди*. Внутри этой области проведены три линии уровня функции $\varphi(a,b)$ при значениях $0,3$, $0,29$ и $0,28$. Заметим, что для функции $p_0(a,b) = (a+b)e^{-a-b}$ линии уровни - это прямые вида $b = D - a$, где D - константа, и значение функции $p_0(a,b)$ на линии уровня равно De^{-D}. Для любой точки внутри области устойчивости очереди выполнено неравенство $a + b > 1$, и при перемещении линий уровней $b = D - a$ к началу координат в пределах этой области значение De^{-D} возрастает. Верхняя грань для $p_0(a,b) = (a+b)e^{-a-b}$ в области устойчивости очереди равна значению фикции $p_0(a,b) = (a+b)e^{-a-b}$ в точке касания прямой $b = D - a$ и границы области $p_1(a,b)/(2p_-(a,b)) < 1$. Следует отметить, что значение функции $\varphi(a,b)$ в этой точке совпадает с верхней гранью для функции $\varphi(a,b)$ в области $p_1(a,b)/(2p_-(a,b)) < 1$ полученной численным способом.

5.2 Открытые проблемы

Описанные выше метод анализа и способ вычисления пропускной способности алгоритма применимы ко всем введенным выше алгоритмам с отложенными интервалами. Однако с увеличением числа интервалов, на которые разбивается исходный интервал, увеличивается

и число параметров алгоритма, и, как следствие, усложняется оптимизационная задача. Например, для описанного в [7] алгоритма исходный интервал разбивается не на два, а на три интервала. По аналогии с ранее рассмотренным алгоритмом длины этих интервалов обозначим за a, b c и введем в рассмотрение функцию $\varphi(a, b, c)$ – теперь уже от трех параметров. Задача поиска пропускной способности алгоритма сводится к поиску верхней грани этой функции в области устойчивости очереди. Численный анализ показывает, что пропускная способность этого алгоритма равна $0,318$, что уточняет ранее полученную в [7] нижнюю оценку. Как и в случае ранее рассмотренного алгоритма, область устойчивости очереди является выпуклой фигурой. Значение пропускной способности достигается в точке, в которой плоскость $a + b + c = D$ касается области устойчивости очереди. Как и в предыдущем случае, эта точка является ближайшей к началу координат. Поэтому и для алгоритма с разбиением исходного интервала на три подинтервала можно предположить, что, видимо, для вычисления пропускной способности достаточно на границе области устойчивости очереди найти такую тройку a, b и c, которая максимизирует вероятность успеха при просмотре исходного отрезка.

Что касается пропускной способности всего рассматриваемого класса алгоритмов с отложенными интервалами, то численный анализ показывает, что при увеличении числа интервалов, на которые разбивается исходный интервал и/или усложнении способа извлечения из очереди отложенных интервалов пропускная способность практически не увеличивается по сравнению с алгоритмом из [7]. Поэтому можно предположить, что пропускная способность всего класса алгоритмов с отложенными интервалами также приблизительна равна $0,318$.

Основная часть работы над этой статьей была осуществлена во время визита первого автора в университет Хериот-Ватта при финансовой поддержке гранта EURO-NF седьмой Европейской рамочной программы НТР. Авторы выражают благодарность программе за эту поддержку и университету Хериот-Ватта за гостеприимство.

Список литературы

[1] Б. С. Цыбаков, В. А. Михайлов. Свободный синхронный доступ в широковещательный канал с обратной связью. *Проблемы передачи информации*, 14(4):32–59, 1978.

[2] J. L. Capetanakis. Tree algorithms for packet broadcast channels. *IEEE Transactions on Information Theory*, 25(5):505–515, 1979.

[3] N. Mehravari, T. Berger. Poisson multiple-access contention with binary feedback. *IEEE Transactions on Information Theory*, 30(5):745–751, 1984.

[4] Б. С. Цыбаков, А. Н. Белояров. Случайный множественный доступ в канале с двоичной обратной связью вида «успех – не успех». *Проблемы передачи информации*, 26(3):67–82, 1990.

[5] Б. С. Цыбаков, А. Н. Белояров. Случайный множественный доступ в канале с двоичной обратной связью. *Проблемы передачи информации*, 26(4):83–97, 1990.

[6] B. Paris, B. Aazhang. Near-optimum control of multiple-access collision channels. *IEEE Transactions on Wireless Communications*, 40:1298–1308, 1992.

[7] A. Malkov, A. Turlikov. Random multiple access protocols for communication systems with "success-failure"feedback. *IEEE International Workshop on Information Theory*, 1:39, 1995.

[8] D. Aldous. Ultimate instability of exponential back-off protocol for acknowledgment based transmission control of random access communication channels. *IEEE Transactions on Information Theory*, 33(2):219–223, 1987.

[9] Б. С. Цыбаков, В. А. Михайлов. Случайный множественный доступ пакетов. Алгоритм дробления. *Проблемы передачи информации*, 16(4):65–79, 1980.

[10] А. А. Боровков. *Теория вероятностей.* Эдиториал УРСС, 1999.

[11] S. Foss, T. Konstantopoulos. An overview of some stochastic stability methods. *Journal of Operation Research Society Japan*, 47(4):275–303, 2004.

[12] A. Gut. *Stopped Random Walks. Limit Theorems and Applications.* Springer, 2009.

6 Приложение

Напомним ряд хорошо известных утверждений (первые два можно найти, напр., в обзоре [11]). Первое утверждение называется "критерием Мустафы-Фостера-Твиди" (часто говорят просто о "критерии Фостера").

Теорема 3. *Пусть $\{Z_n\}$ – однородная по времени цепь Маркова со значениями в измеримом пространстве $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$, и $g : \mathcal{Z}\ [0,\infty)$ – некоторая измеримая функция. Если при некоторых положительных числах C, g_0 и ε справедливы неравенства:*
(1) $\mathbf{E}(g(Z_1) \mid Z_0 = z) \le C$ *п.н. при всех z, таких что $g(z) \le g_0$;*
(2) $\mathbf{E}(g(Z_1) \mid Z_0 = z) \le -\varepsilon$ *п.н. при всех z, таких что $g(z) \ge g_0$;*
то множество $\{z\ :\ g(z) \le g_0\}$ является положительно возвратным и, более того, при любом z случайная величина

$$\mu_z = \min\{n \ge 1\ :\ g(Z_n) \le g_0 \mid Z_0 = z\}$$

имеет конечное среднее, причем

$$\mathbf{E}\mu_z \le g(z)/\varepsilon.$$

Если к тому же множество $\{z\ :\ g(z) \le g_0\}$ конечно и цепь Маркова неразложима и апериодична, то цепь Маркова эргодична, т.е. существует единственное стационарное распределение этой цепи и при любом начальном значении имеет место сходимость к этому стационарному распределению в метрике полной вариации.

Следующее утверждение естественно называть "обобщенным критерием Фостера" (см. также [11]) – оно применимо в более общих условиях, в частности, к приращениям цепи Маркова на интервалах случайной длины.

Теорема 4. *Пусть $\{Z_n\}$ – однородная по времени цепь Маркова со значениями в измеримом пространстве $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$, и $g : \mathcal{Z}\ [0,\infty)$ – некоторая измеримая функция. Пусть также ν_z –последовательность случайных моментов остановки, т.е. при каждом $z \in \mathcal{Z}$ рассматривается цепь Маркова $Z_0, Z_1, \ldots$, стартующая их начального состояния $Z_0 = x$, и для нее вводится случайная величина ν_z, обладающая свойством:*
при каждом $n = 0, 1, \ldots$ событие $\{\nu_z \le n\}$ принадлежит сигма-алгебре, порожденной случайными величинами $\{Z_0 = z, Z_1, \ldots, Z_n$.
Если при некоторых положительных числах C, c_1, c_2, g_0 и ε справедливы неравенства:
(1) $\mathbf{E}(g(Z_{\nu_z}) \mid Z_0 = z) \le C$ *п.н. при всех z, таких что $g(z) \le g_0$;*
(2) $\mathbf{E}(g(Z_{\nu_z}) \mid Z_0 = z) \le -\varepsilon$ *п.н. при всех z, таких что $g(z) \ge g_0$;*
(3) $\mathbf{E}\nu_z \le c_1 + c_2 g(z)$ *при всех $z \in \mathcal{Z}$;*
то множество $\{z\ :\ g(z) \le g_0\}$ является положительно возвратным и, более того, при любом z случайная величина

$$\mu_z = \min\{n \ge 1\ :\ g(Z_n) \le g_0 \mid Z_0 = z\}$$

имеет конечное среднее, причем

$$\mathbf{E}\mu_z \le g(z)/\varepsilon.$$

Если к тому же множество $\{z\ :\ g(z) \le g_0\}$ конечно и цепь Маркова неразложима и апериодична, то цепь Маркова эргодична, т.е. существует единственное стационарное распределение этой цепи и при любом начальном значении имеет место сходимость к этому стационарному распределению в метрике полной вариации.

Сформулируем теперь теорему о сходимости регенерирующих процессов в дискретном времени. Последовательность $\{Z_n\}$ называется *регенерирующей*, если можно указать такие целочисленные случайные моменты $S_0 = 0 \leq S_1 < S_2 < \ldots$, что случайные элементы $V_k = (S_k - S_{k-1}, Z_{S_{k-1}}, Z_{S_{k-1}+1}, \ldots, Z_{S_k-1})$ независимы в совокупности при $k \geq 1$ и одинаково распределены при $k \geq 2$.

Следующую теорему можно найти во многих книгах, где рассматриваются процессы восстановления (см., напр., [12]).

Теорема 5. *Если последовательность является регенерирующей, и если к тому же*
(а) $\mathbf{E}(S_2 - S_1) < \infty$ *и*
(б) наибольший общий делитель тех j*, при которых* $\mathbf{P}(S_2 - S_1 = j) > 0$*, равен единице, то распределения* Z_n *сходятся в метрике полной вероятности при* $n \to \infty$ *к предельному распределению* Π*, имеющему вид*

$$\Pi(B) = \frac{\mathbf{E}\left(\sum_{n=S_1}^{S_2-1} \mathbf{I}(Z_n \in B)\right)}{\mathbf{E}(S_2 - S_1)}.$$

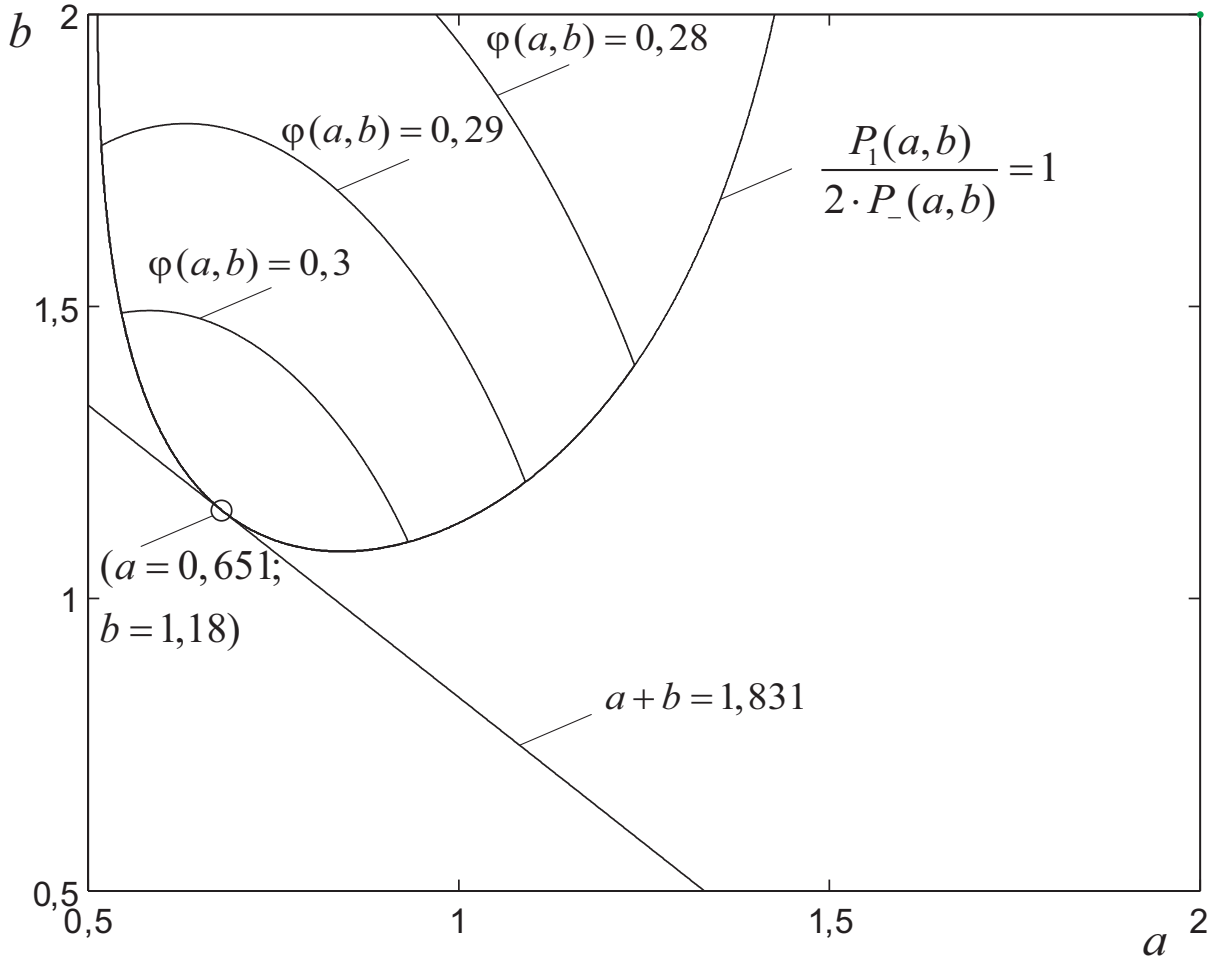

Рис. 1: Область устойчивости очереди отложенных интервалов